# Data Mining
# zur Unterstützung betrieblicher Entscheidungsprozesse

Inaugural-Dissertation

zur Erlangung des akademischen Grades

eines Doktors der Wirtschafts- und Sozialwissenschaften

(Dr. rer. pol.)

der Universität Dortmund

vorgelegt von

**Dipl.-Oec. Christoph Tillmanns**

aus Dortmund

# Inhaltsverzeichnis

# Abkürzungsverzeichnis

| | |
|---|---|
| CBR | Case Based Reasoning |
| DNF | Disjunktive Normalform |
| FOL | First Order Logic |
| HTML | Hypertext Markup Language |
| KNF | Konjunktive Normalform |
| KDD | Knowledge Discovery in Databases |
| OLAP | Online Analytical Processing |
| PC | Personal Computer |
| XML | Extensible Markup Language |

# Symbolverzeichnis

| | |
|---|---|
| $\Diamond$ | Ende einer Definition |
| $\varnothing$ | Leere Menge |
| $\Delta\bullet$ | Veränderung der Variablen $\bullet$ |
| $\hat{\mu}_i$, $\mu_i$ | (Geschätzter) Erwartungswert der Ergebnisse der $i$-ten Alternative ($i = 1, ..., hmax$) |
| $A$ | Menge der beobachteten Variablen eines Realsystems |
| $A_{BR}$ | Menge der Attribute, für die ein Joinpfad zur Bezugsrelation, $BR$, definiert ist |
| $A^C$ | Menge der potentiell in die Lösung aufzunehmenden erklärenden Attribute |
| $A^D$ | Menge der potentiell in die Lösung aufzunehmenden zu erklärenden Attribute$AL, AL'$Mengen von weiter zu untersuchenden Interessantheitsgrad-Lösungs-Paaren („Auswahlliste“) |
| $a, a', R.a$ | Attribut (aus der Relation $R$) |
| $acmax$ | Anzahl der erklärenden Attribute in der Trainingsmenge = $\|C\|$ |
| $admax$ | Anzahl der zu erklärenden Attribute in der Trainingsmenge = $\|D\|$ |
| $amax$ | Anzahl der Attribute in der Trainingsmenge (ohne die Schlüsselattribute) |
| $a(\bullet)$ | Attributwert des Objektes $\bullet$ bezüglich Attribut $a$ |
| $B$ | Benutzermodell |
| $Bmax$ | Anzahl Regeln im Benutzermodell |
| $BR$ | Bezugsrelation |
| $C$ | Menge von erklärenden Attributen („Condition“) in der aktuellen Lösung |
| $c_g$ | Akzeptanzwert („Temperatur“, „Wasserstand“, ...) in Generation $g$ |
| $CP$ | Menge der Cutting Points aller Attribute in der aktuellen Lösung |
| $CP(\bullet)$ | Menge der Cutting Points zum Attribut $\bullet$ in der aktuellen Lösung |
| $CP^V(\bullet)$ | Menge der vordefinierten Cutting Points zum geordneten Attribut $\bullet$ |
| $cpmax(\bullet)$ | Index des höchsten Cutting Points zu Attribut $\bullet$ in der aktuellen Lösung |
| $cpmax^V(\bullet)$ | Index des höchsten vorgegebenen Cutting Points zu Attribut $\bullet$ |
| $c_\bullet$ | Konzept bezüglich des Terms $\bullet$ |
| $D$ | Menge von zu erklärenden Attributen („Decision“) in der aktuellen Lösung |

| | |
|---|---|
| $DB(BR,A)$ | Datenbasis für das in Kapitel 5 entwickelte Data-Mining-Verfahren bezüglich der Bezugsrelation, $BR$, und der aktuellen Attributmenge, $A$ |
| $dk\ (dk_i)$ | direkte Kosten (der $i$-ten Handlungsalternative) |
| $DM^{KNF}(C,D)$ | Menge aller Datenmuster in konjunktiver Normalform mit der Menge der zu erklärenden Attribute, $D$, und der Menge erklärender Attribute, $C$ |
| $dmax(i)$ | Anzahl von Klassen in der Domäne $dom(D_i)$ |
| $dom(\bullet)$ | Domäne von $\bullet$ |
| $E,\ E^+$ | Menge von gerichteten Kanten (Kapitel 2) bzw. Menge von Handlungsergebnissen (Kapitel 3 ff.) |
| $EB$ | Entscheidungsbaum |
| $ER(\bullet)$ | Relative Erfolgsrelevanz des Planungsobjektes $\bullet$ |
| $ES$ | Menge von Entscheidungssituationen |
| $e,\ e^+,\ e(\bullet)$ | Handlungsergebnis (eines Objektes $\bullet$) |
| $er\ (er_i)$ | Erlöse (der $i$-ten Handlungsalternative) |
| $f^W$ | Wirkungsfunktion eines Entscheidungsmodells |
| $f^Z$ | Zielerreichungsfunktion eines Entscheidungsmodells |
| $G$ | Graph (Kapitel 2) bzw. Anzahl Iterationen der Suche (Kapitel 5 und 6) |
| $\vec{G}$ | Gerichteter Baum |
| $g$ | aktuelle Generation im Suchprozeß |
| $gmax$ | maximale Anzahl von Generationen, bis das Suchverfahren abbricht |
| $gmax^{oV}$ | maximale Anzahl von Generationen ohne Verbesserung des Zielfunktionswertes, bis das Suchverfahren abbricht |
| $H,\ H^*$ | Menge von Handlungsalternativen |
| $H_0,\ H_1$ | Hypothesen |
| $h^+,\ h^-,\ h^*,\ h',\ h$ | Handlungsalternativen |
| $hmax$ | Anzahl der Handlungsalternativen in einer Alternativenmenge |
| $I_\bullet$ | Ununterscheidbarkeitsrelation bezüglich der Attributmenge $\bullet$ |
| $IG,\ IG_{O^T}$ | Interessatheitsgrad einer Datenmuster-Menge (bezüglich $O^T$) |
| $ig,\ ig^{alt},\ ig^{neu},\ ig^{local\text{-}best},\ ig_{O^T}$ | Interessatheitsgrade eines Datenmusters (bezüglich $O^T$) |
| $ig^{min}$ | minimal zu erreichender Interessantheitsgrad |
| $IV$ | Menge der inneren Knoten eines Entscheidungsbaums |
| $Kl,\ Kl^{neu},\ Kl_\bullet$ | Klauseln in konjunktiver Normalform (mit Attribut $\bullet$) |
| $Kl^{KNF}(\bullet)$ | Menge aller Klauseln in konjunktiver Normalform mit Attributmenge $\bullet$ |

| | |
|---|---|
| $Klmax$ | Anzahl der Klauseln eines Terms |
| $Ko$ | Konklusion einer Regel |
| $Komax\_max$ | maximal erlaubte Anzahl von Klauseln in einer Konklusion |
| $L$ | Lösungsraum für das Data Mining als Optimierungsproblem |
| $L'$ | Lösungsraum für das Data Mining als Suchproblem |
| $L^*$ | Menge der Zielzustände im Lösungsraum |
| $LV$ | Menge der Blätter eines Entscheidungsbaums |
| $M$ | Anzahl Datenmuster im Ergebnis eines Data-Mining-Verfahrens |
| $M^{allgemein}$ | Das allgemeinste mögliche Modell |
| $M^{speziell}$ | Das speziellste mögliche Modell |
| $M_\bullet$ | Aus der Objektmenge $\bullet$ generiertes Modell, z.B. $M_{O^T}$ |
| $M^\bullet$ | Funktionales Data-Mining-Modell (entweder $M^{Ent}$ oder $M^{Pro}$) |
| $M^{Bes}$ | Beschreibungsmodell |
| $M^{Ent}$ | Entscheidungsmodell |
| $M^{Erk}$ | Erklärungsmodell |
| $M^{Pro}$ | Prognosemodell |
| $N$ | in Kapitel 3: Menge der Nutzwerte<br>sonst: Anzahl von Beobachtungen (Datensätzen) in einer Objektmenge |
| $\boldsymbol{N}$ | Menge der natürlichen Zahlen |
| $n^*, n^+$ | Nutzwerte |
| $N(s,Tr)$ | Nachbarschaft einer Lösung, $s$, die durch Anwendung einer Transformation aus $Tr$ erreichbar ist |
| $O, O^N, O^I, O'$ | Objektmengen |
| $O^E$ | Evaluierungsmenge (Stichprobe) |
| $ON_\bullet(X)$ | Obere Näherung an eine Objektmenge $X$ bezüglich der Attributmenge $\bullet$ |
| $O^{neu}$ | Menge neuer Objekte, auf die ein Modell angewendet wird |
| $O^T$ | Trainingsmenge (Stichprobe) |
| $P(\bullet)$ | Wahrscheinlichkeit für das Eintreten eines Ereignisses $\bullet$ |
| $Pop$ | Anzahl der Individuen in einer Population |
| $Pot(\bullet)$ | Potenzmenge (Menge aller Teilmengen) von $\bullet$ |
| $POS_C(D)$ | Positive Region bezüglich der Attributmengen $C$ und $D$ |
| $Pr$ | Prämisse einer Regel |
| $Prmax$ | Anzahl Klauseln in der Prämisse einer Regel |

| | |
|---|---|
| $Prmax\_max$ | maximal erlaubte Anzahl von Klauseln in einer Prämisse |
| $\boldsymbol{R}$ | Menge der reellen Zahlen |
| $R$ | Relation |
| $R.A$ | Menge der Attribute der Relation $R$ |
| $R.a$ | Attribut $a$ aus der Relation $R$ |
| $R(\bullet)$ | Rang des Ordinalwertes $\bullet$ |
| $RM$ | Menge von Relationen |
| $Rmax$ | Anzahl von Relationen in einer Relationenmenge |
| $S, S(C, Prmax\_max)$ | Suchraum für das Data-Mining-Verfahren (Menge aller zulässigen Rough-Set-Lösungen) |
| $S^*$ | Menge der Zielzustände im Suchraum |
| $SA$ | Menge der Schlüsselattribute |
| $s, s^{alt}, s^*, s^{neu}, s^{TOP}, s^{lokal\text{-}best}$ | Lösungen im Suchraum (Datenmuster bzw. Datenmuster-Menge) |
| $samax$ | Anzahl der Schlüsselattribute $= \|SA\|$ |
| $Te(\bullet), Te$ | Term in konjunktiver Normalform bezüglich Attributmenge $\bullet$ |
| $Te^{KNF}(\bullet)$ | Menge aller Terme in KNF bezüglich Attributmenge $\bullet$ |
| $TOP$ | Menge der besten Lösungen (Rekorde) und ihrer Bewertungen |
| $Tr$ | Menge möglicher Transformationen (Züge, Suchschritte) |
| $U$ | Menge der Umweltsituationen |
| $UN_\bullet(X)$ | Untere Näherung an eine Objektmenge $X$ bzgl. der Attributmenge $\bullet$ |
| $u, u^*$ | Umweltsituationen |
| $umax$ | Anzahl der Umweltsituationen in einer Menge von Umweltsituationen |
| $V$ | Menge von Knoten |
| $V^P(\bullet)$ | Menge der Vorgängerknoten von Knoten $\bullet$ |
| $V^S(\bullet)$ | Menge der Nachfolgerknoten von Knoten $\bullet$ |
| $W^\bullet$ | Wahrscheinlichkeitsdichtefunktion für Handlungalternative $\bullet$ |
| $w$ | ein Attributwert |
| $w^E$ | Gewicht für die Erfolgsrelevanz |
| $wmax$ | Anzahl der Werte in einer Wertemenge |
| $Z, Z^+$ | Menge von Zielbeiträgen |
| $z, z^+, z(\bullet)$ | Zielbeitrag (eines Objektes $\bullet$) |
| $zf$ | Zielfunktion eines Entscheidungsmodells |

# Abbildungsverzeichnis

# Tabellenverzeichnis

# Definitionsverzeichnis

# 1 Einleitung

Das Data Mining ist in den letzten Jahren verstärkt ein Thema in Forschung und Praxis geworden. Die Motivation, die Forscher und Praktiker dazu treibt, sich mit diesem Thema auseinanderzusetzen, wird in Abschnitt 1.1 vorgestellt. Der folgende Abschnitt 1.2 umreißt kurz den Stand der Data-Mining-Forschung und leitet daraus den offenstehenden Forschungsbedarf ab. Durch Eingrenzung des Forschungsbedarfs können in Abschnitt 1.3 die Ziele dieser Untersuchung formuliert werden. Darauf aufbauend beschreibt Abschnitt 1.4 den Gang der weiteren Untersuchung.

## 1.1 Motivation des Data-Mining-Ansatzes in Forschung und Praxis

In der Unternehmenspraxis sind unter dem Schlagwort „*Data Warehouse*“ zunehmende Bestrebungen zur Integration verschiedener interner und externer Datenquellen zu beobachten.[1] Eine integrierte Datenhaltung stellt eine wesentliche Voraussetzung für die Analyse von Zusammenhängen zwischen Markt- und Unternehmensdaten dar. Auf solche Analysen sind die betrieblichen Informationssysteme häufig nur unzureichend ausgerichtet. So wurden, bevor das Data Mining Einzug in die analytischen Informationssysteme nahm, Analysen häufig nur durch vordefinierte Standardmodelle und -methoden unterstützt, wie z.B. deskriptive statistische Verfahren, Kennzahlsysteme, Break-Even-Analysen, ABC-Analysen oder Ampelfunktionen. Die daraus gewonnenen Informationen gehen in die Standardberichte des betrieblichen Reportings ein. Neben den Standardberichten gibt es Ad-hoc-Auswertungen, welche durch schließende statistische Verfahren, OLAP[2]-Operationen und andere Sonderrechnungen unterstützt werden. Dabei ist allen Auswertungsmethoden gemein, daß der Anteil an intellektueller Eigenleistung sehr hoch ist; das analytische Informationssystem übernimmt jeweils nur Hilfsfunktionen wie einen schnellen Datenzugriff, vordefinierte Lösungsverfahren oder eine visuelle Aufbereitung der Ergebnisse. Die eigentliche Modellkonstruktion wird dem Analytiker aufgebürdet. Lediglich einige wenige Parameter, wie etwa die Koeffizienten

---

[1] Vgl. TOTOK (1997), S. 4.

[2] „*OLAP*“ steht für „*Online Analytical Processing*“ und betont die analytische im Gegensatz zur transaktionsorientierten Datenverarbeitung (*OLTP, Online Transaction Processing*) im Bereich der operativen DV-Systeme.

bei der Regressionsanalyse, werden durch eine automatisierte Methode bestimmt. Die geringe Unterstützung des Analytikers bei der Modellkonstruktion selbst kann dazu führen, daß der Analytiker relevante Zusammenhänge des Realsystems übersieht oder verkennt und nicht oder falsch im Modell abbildet.

Gerade in praxisnahen Veröffentlichungen hat die Behauptung für Aufsehen gesorgt, es gäbe Verfahren, die selbständig *interessante* und „*ungeahnte*“ Zusammenhänge in großen Datenbeständen aufdecken. Es heißt, diese Verfahren würden auch geschäftsrelevante Fragen beantworten, die gar nicht gestellt wurden.[3]

Solche Aussagen müssen bei ernsthafter Auseinandersetzung mit dem Thema relativiert werden. Hierzu genügt es an dieser Stelle, Data-Mining-Verfahren als Systeme zu verstehen, die in der Lage sind, zu einem feststehenden Modelltyp eine große Menge von Modellinstanzen zu generieren und anhand empirischer Daten zu testen.[4] Wenn dieser Modelltyp nur allgemein genug definiert werden und nahezu beliebige Realzusammenhänge approximieren kann, so mag die einzelne, durch ein Data-Mining-Verfahren ermittelte Aussage als Instanz dieses Modelltyps für den Benutzer durchaus unbekannt sein.

Die Grundidee, unbekannte Zusammenhänge aufzuspüren, ist für viele Unternehmen von großem Wert. Unterstellt man, daß solche Zusammenhänge erstens wettbewerbsrelevant und zweitens auch für die Wettbewerber unbekannt sind, kann durch das Data Mining ein komparativer Konkurrenzvorteil entstehen. Diese Voraussetzungen liegen u.a. häufig gerade dort vor, wo große *Kunden*datenbestände erfaßt und verwaltet werden. Anwendungspotentiale ergeben sich beispielsweise in der Ermittlung typischer Warenkörbe im Einzelhandel, der Abschätzung von Forderungsausfällen im Kreditkartengeschäft, der Analyse von Vertragsstornierungen bei Versicherern, der Prognose der Zeitpunkte von Automobilkäufen oder der Selektion von Kunden für Direktmarketingaktionen.

Die Aussicht, sich durch das Entdecken wettbewerbsrelevanten Wissens Konkurrenzvorteile zu verschaffen, motiviert ein großes Interesse in der Praxis. Die Forschung wird neben der praktischen Relevanz durch die Komplexität der Aufgabenstellung motiviert.

---

[3] Vgl. KRAHL/WINDHEUSER/ZICK (1998), S. 12.

[4] Vgl. GEBHARDT (1994), S. 9.

So erfordert die Entwicklung „intelligenter“ Data-Mining-Verfahren ein Zusammenspiel aus Methoden der Statistik und der künstlichen Intelligenz mit Datenbanktechnologien und betriebswirtschaftlicher Modellbildung. Diese Herausforderungen werden im nächsten Abschnitt näher beleuchtet.

## 1.2 Stand der Forschung und Bedarf weiterer Forschungsaktivitäten

Das Data Mining hat sich aus den Forschungsbereichen des Maschinellen Lernens und der Statistik entwickelt, wobei die meisten Beiträge dem erstgenannten Forschungsbereich zuzuordnen sind. Aktuellere Entwicklungen sind auch durch Einflüsse aus dem Bereich der Datenbanken geprägt. Aufgrund dieser Historie sind die meisten Beiträge zum Data Mining technisch orientiert und beziehen sich auf die Konzeption von bestimmten Teilkomponenten von Data-Mining-Verfahren. Welcher Forschungsbedarf im technischen Bereich besteht, kann erst nach der Betrachtung der Teilkomponenten existierender Data-Mining-Verfahren ermittelt werden.[5] Vorgreifend sei hier schon bemerkt, daß das Potential und die Notwendigkeit für weitere Verbesserungen im technischen Bereich als vergleichsweise gering einzustufen sind. Eine Ausnahme stellt der Datentransfer zwischen der Unternehmensdatenbank und dem Data-Mining-Programm dar. Der Datenzugriff bildet einen Engpaß, dessen Abbau zu den modernen „Grand Challenges“, den großen Herausforderungen an heutige Hochleistungsrechner, gezählt wird.[6]

Bisher fehlt in der Data-Mining-Literatur eine Systematik zur Unterstützung betriebswirtschaftlicher Entscheidungen. Teilweise werden zwar betriebliche – seltener auch betriebswirtschaftliche – Anwendungsbereiche genannt. Aber nirgendwo werden diese Anwendungsbereiche analysiert und daraus Vorgehensmodelle zur Problemlösung mit Data-Mining-Verfahren abgeleitet. Ein solches Vorgehensmodell müßte die Anwendungsprobleme kategorisieren und für jede Problemkategorie einen Lösungsweg vorschlagen. Würden dabei nur die existierenden Data-Mining-Verfahren betrachtet, so könnten einige betriebswirtschaftlich relevante Problemkategorien nicht gelöst werden, da die existierenden Verfahren ihre Modelle nach einem festen, häufig nicht

---

[5] Dies wird in Kapitel 4 nachgeholt.

[6] Vgl. CAP (1998), S. 53.

austauschbaren Gütekriterium generieren und bewerten. Das Gütekriterium unterscheidet sich i.d.R. erheblich von den Kriterien, die Modellaussagen aus betriebswirtschaftlicher Sicht interessant machen. Der Begriff der Interessantheit ökonomischer Aussagen muß daher operational definiert werden, so daß er in ein Data-Mining-Verfahren integriert werden kann. Außerdem muß ein direkter Bezug zwischen verschiedenen betriebswirtschaftlichen Problemstellungen und den anzuwendenden Bewertungsvorschriften hergestellt werden. Gerade dies leistet kein einziger Forschungsbeitrag in hinreichendem Maße.

Zusammengefaßt bestehen im Data Mining folgende Forschungsbedarfe:

- ⇨ Gewährleistung eines effizienten Datenzugriffs;
- ⇨ Entwicklung eines Vorgehensmodells zur Lösung betriebswirtschaftlicher Probleme per Data Mining – wobei das Vorgehensmodell insbesondere die Operationalisierung des Interessantheitsbegriffs für die jeweilige Problemstellung umfassen müßte;
- ⇨ Entwicklung von Data-Mining-Verfahren, die sich an diesem Vorgehensmodell orientieren.

## 1.3 Ziele und Grenzen der Untersuchung

Vor dem Hintergrund der dargestellten Forschungsbedarfe ist zu klären, in welchen Grenzen sie sich im Rahmen dieser Arbeit erfüllen lassen. Dabei sollen folgende Grenzen schon im voraus gezogen werden:

- ⇨ Die **Datenzugriffskomponente** wird nur aus Sicht des Analyseverfahrens auf der Client-Seite betrachtet und nicht etwa aus Sicht des Datenbanksystems auf der Server-Seite, obwohl auch dort Effizienzsteigerungspotentiale liegen.[7]
- ⇨ Die **Entwicklung eines Vorgehensmodells** orientiert sich an dem Informationsbedarf betriebswirtschaftlicher Entscheidungen. Das Vorgehensmodell soll dabei *von bereits existierenden Lösungsverfahren unabhängig* sein, um die Gefahr zu vermeiden, daß die Problemstellungen an die impliziten Restriktionen der Verfahren angepaßt werden müssen. Außerdem besteht Anlaß zu der Vermutung, daß für

[7] Vgl. zur Effizienzsteigerung auf Seiten des Datenbankservers HOLSHEIMER ET AL. (1995), S. 151 ff.

bestimmte betriebswirtschaftliche Problemstellungen kein geeignetes Data-Mining-Verfahren existiert.

Als Bestandteil des Vorgehensmodells wird die Interessantheit von Aussagen nur insoweit operationalisiert, wie aufwendige manuelle Benutzereingaben in Grenzen gehalten werden können. Damit werden insbesondere hochgradig interaktive Verfahren sowie Ansätze, die umfangreiches Hintergrundwissen in maschinell verarbeitbarer Form voraussetzen, ausgegrenzt.

⇨ Die **Entwicklung eines Data-Mining-Verfahrens** soll dazu dienen, eine der identifizierten Problemklassen, für die bisher noch keine annehmbaren Lösungsansätze existieren, durch ein geeignetes Verfahren lösen zu können. Dabei werden als Grenzen die Ressourcenbeschränkungen eines einzelnen Personalcomputers gezogen. Fokussiert wird das sog. „***Data Mining i.e.S.***“, d.h., der Teilprozeß, der sich mit der eigentlichen Suche nach interessanten Datenmustern befaßt. Die vor- und nachgelagerten Schritte, das sog. „***Preprocessing***“ und „***Postprocessing***“, werden nur teilweise betrachtet.

Das Ziel dieser Arbeit besteht darin, zu untersuchen, *wie das Data Mining zur Generierung entscheidungsunterstützender Modelle eingesetzt werden kann und die Ergebnisse dieser Untersuchung in ein Data-Mining-Verfahren zu integrieren*. Damit bestehen die Einzelziele dieser Arbeit darin,

1. ein Vorgehensmodell zur Lösung betriebswirtschaftlicher Problemstellungen zu konzipieren (**Ziel 1: „Lösungsschema konzipieren“**),
2. ein Data-Mining-Verfahren zur Lösung der Problemstellungen zu entwickeln (**Ziel 2: „Lösungsverfahren entwickeln“**) und
3. durch Anwendung des Verfahrens auf eine betriebswirtschaftliche Problemstellung dessen praktische Eignung zu evaluieren (**Ziel 3: „Verfahren evaluieren“**).

## 1.4 Aufbau der Untersuchung

Das Data Mining wurde, wie gesagt, durch technische Entwicklungen vorangetrieben. Erst nach und nach wurde es in der Praxis aufgenommen und verbreitet. An dieser Entwicklung orientiert sich auch der Aufbau dieser Untersuchung und damit der Aufbau der

der schriftlichen Arbeit (vgl. Abbildung 1-1). So arbeitet Kapitel 2 die überwiegend technischen Grundlagen des Data Mining heraus. Nach einer Einführung in grundlegende Begriffe des Data Mining werden in Abschnitt 2.2 die Arbeitsweise und die Teilkomponenten von Data-Mining-Verfahren vorgestellt. Die Brücke zu den Anwendungen von Data-Mining-Verfahren schlägt Abschnitt 2.3 mit der Aufstellung von Eignungskriterien für Data-Mining-Anwendungen, die sich aus den zuvor dargestellten technischen Grundlagen ableiten lassen.

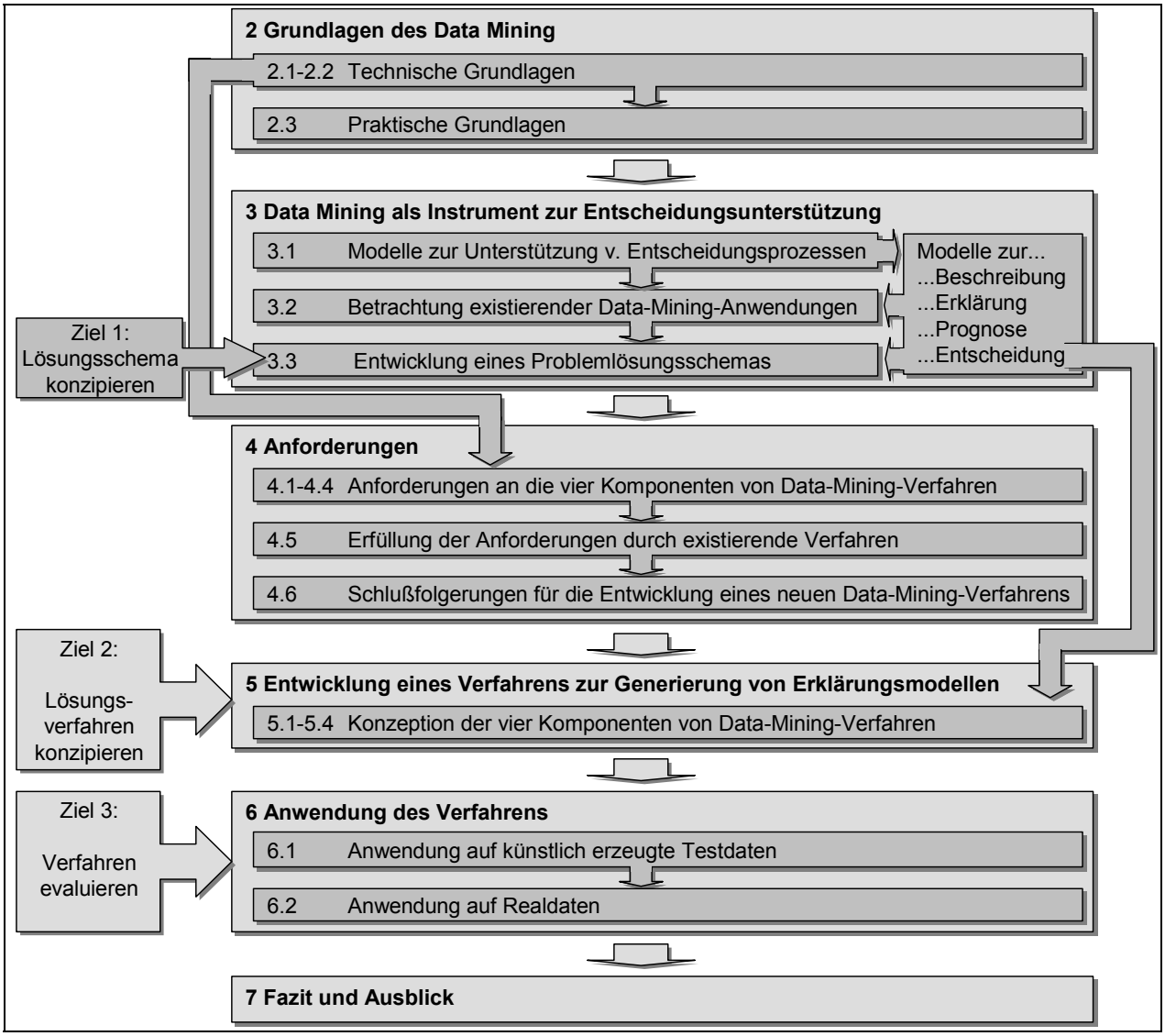



**Abbildung 1-1: Einzelziele und Aufbau der Arbeit**[8]

In Kapitel 3 werden die Unterstützungspotentiale für betriebswirtschaftliche Entscheidungsprozesse betrachtet. Abschnitt 3.1 nimmt eine Einführung in die modellgestützte

[8] Die Pfeile stellen inhaltliche Abhängigkeiten dar, die im Text erläutert werden.

Entscheidungsfindung vor. Dabei werden mit den Beschreibungs-, Erklärungs-, Prognose- und Entscheidungsmodellen vier Modelltypen vorgestellt, die in der folgenden Untersuchung zur differenzierten Betrachtung verwendet werden. Abschnitt 3.2 untersucht existierende Anwendungen des Data Mining und ordnet diese in die Phasen betriebswirtschaftlicher Entscheidungsprozesse ein. Diese Untersuchung ist notwendig, um daraus in Abschnitt 3.3 ein allgemeingültiges Problemlösungsschema für betriebswirtschaftliche Anwendungen abzuleiten (**Ziel 1**).

Sowohl aus dem Vorgehensmodell als auch aus den technischen Grundlagen des Data Mining ergeben sich Anforderungen an die Komponenten von Data-Mining-Verfahren. Um ein Lösungsverfahren entwickeln zu können (**Ziel 2**), werden diese Anforderungen in Kapitel 4 zusammengestellt und deren Erfüllung durch existierende Data-Mining-Verfahren analysiert. Hier wird auch der in Abschnitt 1.2 nur kurz angerissene Stand der Forschung detailliert. Aus dieser Untersuchung werden in Abschnitt 4.6 Schlußfolgerungen für die Entwicklung eines neuen Data-Mining-Verfahrens gezogen.

Zumindest zur Generierung von Entscheidungsmodellen kann dann in Kapitel 5 ein Verfahren entwickelt werden. In den vier Abschnitten werden die vier Komponten des Data-Mining-Verfahrens konzipiert.

Das Evaluierungsziel (**Ziel 3**) schließlich wird in Kapitel 6 erfüllt. In Abschnitt 6.1 wird das entwickelte Verfahren anhand von künstlich erzeugten Testdaten analysiert, verbessert und die Auswirkungen verschiedener Parametereinstellungen untersucht. Die hier gewonnenen Erkenntnisse werden in Abschnitt 6.2 zur Anwendung des Verfahrens auf Realdaten genutzt.

Die Untersuchung schließt in Kapitel 7 mit einer Zusammenfassung und einem Ausblick auf anknüpfende Untersuchungen.

# 2 Grundlagen des Data Mining

In diesem Kapitel werden die Grundlagen des Data Mining eingeführt. Hierzu zählen grundlegende Begriffe, Begriffsabgrenzungen und -einordnungen, welche in Abschnitt 2.1 behandelt werden. Abschnitt 2.2 führt in die Arbeitsweise von Data-Mining-Verfahren ein und stellt Komponenten vor, die jedes Data-Mining-Verfahren umfaßt. Abschnitt 2.3 schließlich bildet eine Brücke zu dem nachfolgenden anwendungsorientierten Kapitel und stellt Voraussetzungen für die Anwendung von Data-Mining-Verfahren dar.

## 2.1 Grundlegende Begriffe und Einordnung des Data Mining

Dieser Abschnitt dient der Einführung grundlegender Begriffe, Begriffsabgrenzungen und -einordnungen. Abschnitt 2.1.1 stellt eine Definition des Data-Mining-Begriffs dar, die sich in der Literatur weitgehend durchgesetzt hat und charakterisiert darauf aufbauend das Data Mining als Anwendung induktiver Lernverfahren. In Abschnitt 2.1.2 werden Schwächen dieser klassischen Definition aufgezeigt und eine verbesserte Begriffsdefinition vorgeschlagen. Abschnitt 2.1.3 differenziert vorläufig die verschiedenen Problemstellungen im Data Mining. Abschnitt 2.1.4 grenzt Data-Mining- von verwandten Verfahren ab. Die begriffliche Einordnung des Data Mining in den Knowledge Discovery in Databases Process erfolgt in Abschnitt 2.1.5.

### 2.1.1 Klassisches Verständnis des Data Mining als Anwendung induktiver Lernverfahren

In der Literatur hat sich im wesentlichen die Begriffsdefinition von FAYYAD, PIATETSKY-SHAPIRO und SMYTH durchgesetzt. Diese Autoren verstehen unter ***Data Mining*** die *Anwendung spezieller Algorithmen unter akzeptablen Laufzeitrestriktionen, die als Output eine Menge von Datenmustern liefern*.[9] Dabei bezeichnet ein ***Datenmuster*** einen *Ausdruck in einer formalen Sprache, der eine Objektmenge beschreibt, ohne die einzelnen*

[9] Vgl. FAYYAD/PIATETSKY-SHAPIRO/SMYTH (1996), S. 9.

*Datenobjekte einfach nur aufzulisten*.[10] Die Datenobjekte beschreiben Objekte eines Realsystems durch ausgewählte Merkmale, und die Datenmuster werden dazu verwendet, reale Zusammenhänge zwischen den beobachteten Merkmalen zu modellieren. Die Beschreibung durch ein Datenmuster muß nicht unbedingt leicht verständlich und interpretierbar sein. Vielmehr kann es sich hierbei prinzipiell um beliebig komplexe Funktionen handeln, wie z.B um ein neuronales Netz.

Beschränkt man die Betrachtung auf wenige Merkmale und grenzt dabei relevante Einflußfaktoren aus, so unterliegt der reale Zusammenhang einem Zufallseinfluß, dem sog. „***Datenrauschen***" oder „***Datenschmutz***".

*Beispielsweise könnten reale Zusammenhänge zwischen soziodemographischen Kundenmerkmalen und dem Einkaufsverhalten von Kunden bestehen. In Abbildung 2-1 sind zu einer Menge von Kunden die beobachteten Variablen „Alter", „Einkommen" und „Käufer" eines bestimmten Produktes eingetragen. Im linken Teil der Abbildung ist der unbekannte wahre Zusammenhang zwischen diesen drei Variablen durch eine gestrichelte Linie angedeutet. Man erkennt, daß nicht alle beobachteten Kunden in dem markierten Bereich Käufer des Produktes sind und nicht alle Kunden außerhalb dieses Bereiches Nichtkäufer. Diese Unterschiede zwischen den Beobachtungsdaten und dem wahren Zusammenhang werden als „Datenrauschen" bezeichnet; sie lassen vermuten, daß es andere, unbeobachtete Variablen gibt, von denen die Variable „Käufer" abhängt.*

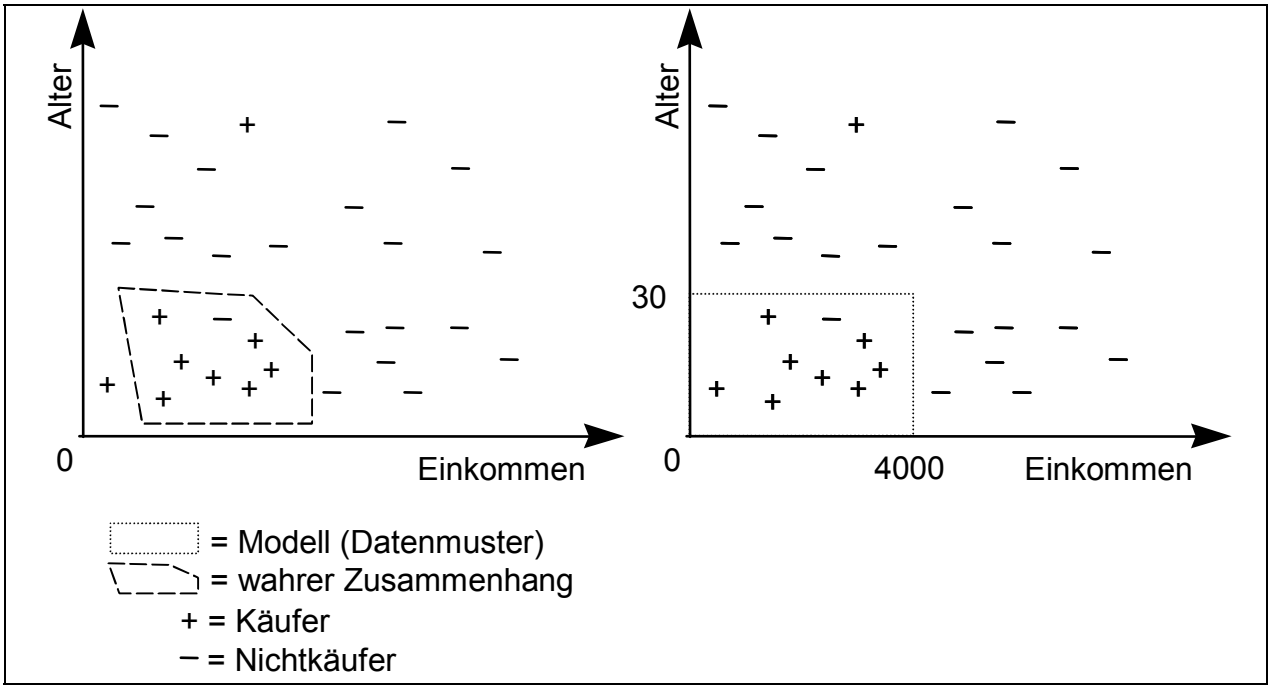



***Abbildung 2-1:*** ***Wahrer Wirkungszusammenhang (links) und Modell (rechts) der Käufer von Produkt B***

[10] Vgl. FAYYAD/PIATETSKY-SHAPIRO/SMYTH (1996), S. 7.

Induktiv ermittelte Modelle zeichnen sich dadurch aus, daß sie aus einer endlichen Anzahl von Beobachtungen erstellt werden. Nun kommt für die Modellierung des realen Systems eine unendlich große Anzahl möglicher Modelle in Frage. Daher werden induktiv erstellte Modelle i.d.R. – auch unter Vernachlässigung des Datenrauschens – nicht mit dem wahren Zusammenhang übereinstimmen. Der Unterschied wird hier als „***Modellfehler***“ bezeichnet.[11]

*Im rechten Teil der Abbildung 2-1 ist durch die gepunktete Linie ein Modell angedeutet, das den realen Zusammenhang approximiert. Das Modell ist ein Datenmuster, das auf der Grundlage der Beobachtungen, also induktiv, entstanden ist. Dabei können die einzelnen Kunden in einer regelorientierten Sprache beschrieben werden, z.B.:*

*WENN Alter ∈ [0;30] UND Einkommen ∈ [0;4000] DANN Käufer.*

*Bei Vergleich mit dem linken Teil der Abbildung erkennt man den Unterschied zwischen Modell und Realzusammenhang. Die unregelmäßige Form des wahren Zusammenhangs läßt sich allein auf Grundlage der Beobachtungen nicht rekonstruieren, da gerade entlang der gestrichelten Linie nur wenige Beobachtungen vorliegen.*

Wie gesagt kommt für die Modellierung des realen Systems eine unendlich große Anzahl möglicher Modelle in Frage. Bei dem Schluß von einzelnen Beobachtungen auf ein allgemeingültiges Modell ist ein sog. „***inductive leap***“[12] („induktiver Schluß“) durchzuführen, d.h. es müssen Annahmen getroffen werden, wie das Realsystem beschaffen sein könnte.

*In dem Beispiel wurde nur nach Regeln gesucht, die einen rechteckigen Ausschnitt aus dem durch die Variablen „Alter“ und „Einkommen“ aufgespannten Raum beschreiben. Auch hiervon existieren noch unendlich viele Möglichkeiten, da Alter und Einkommen stetige Wertebereiche besitzen. Aus diesem Grunde wurde nur nach gut lesbaren Alters- (z.B. 0, 10, 20, ..., 100) und Einkommensgrenzen (z.B. 0, 1.000, 2.000, ..., 1.000.000) innerhalb abgeschlossener Wertebereiche gesucht. Jetzt sind die Wertebereiche endlich, aber immer noch existieren viele in Frage kommende Rechtecke. Also mußten weitere Voraussetzungen getroffen werden, die die Suche nach dem wahren deterministischen Zusammenhang einschränken. Daher wurde nach demjenigen „Rechteck“ gesucht, welches das größte Verhältnis von Käufern des Produktes B zu allen Kunden innerhalb des Rechtecks aufweist und dabei eine gewisse Mindestanzahl von Kunden einschließt. Die Suche erfolgte vollständig über alle unter den getroffenen Voraussetzungen möglichen Rechtecke in einer durch das Suchverfahren determinierten Reihenfolge. Wären mehrere Optima gefunden worden, so wäre das erste entdeckte als Lösung ausgegeben worden.*

Die getroffenen Voraussetzungen, welche die Suche nach dem wahren deterministischen Zusammenhang leiten, werden als „***bias***“[13] bezeichnet. Hierzu zählen:[14]

---

[11] Quantifiziert wird der Modellfehler in Abschnitt 2.2.4.1.

[12] BRISCOE/CAELLI (1996), S. 90

[13] BRISCOE/CAELLI (1996), S. 90.

[14] Vgl. UTGOFF (1986), S. 107 und BRISCOE/CAELLI (1996), S. 89 ff.

- ⇨ der Lösungsraum (im Beispiel: Rechtecke, vorgegebene Intervallgrenzen),
- ⇨ die Bewertung der Modelle im Lösungsraum (im Beispiel: Anteil der korrekt klassifizierten an allen Kunden im Rechteck, Anzahl von Kunden im Rechteck) und
- ⇨ das Suchverfahren (im Beispiel: Reihenfolge und Abbruch der Suche).

Die in maschinelle Lernsysteme implementierten Versionen dieser Annahmen können als „'intelligente' Komponenten" bezeichnet werden. Sie werden im folgenden kurz erläutert und mit Verweisen auf die entsprechenden Abschnitte, in denen sie ausführlicher dargestellt werden, versehen.

Man schränkt die Menge der in Frage kommenden Modelle auf einen bestimmten **Lösungsraum** ein. Es kann – wie in dem obigen Beispiel gesehen – sein, daß der wahre deterministische Zusammenhang nicht einmal Element des Lösungsraums ist.[15] Da der wahre deterministische Zusammenhang unbekannt ist, muß anhand des beobachteten Systemverhaltens abgeschätzt werden, wie gut ein Modell aus dem Lösungsraum den wahren deterministischen Zusammenhang approximiert. Das beobachtete Systemverhalten ist durch eine Datenbasis gegeben. Die Systemstruktur wird in ihrem Grundtyp durch den Modellierer vorgegeben. Er läßt dabei eine Menge von Strukturvariablen (im Beispiel: die Koordinaten des Rechtecks) offen, die durch das Data-Mining-Verfahren instanziiert werden. Bei der Abschätzung des Modellfehlers besteht das Problem, daß immer nur die Übereinstimmung des realen und des modellierten System*verhaltens* bewertet werden kann. Die reale System*struktur* ist unbekannt. Hinzu kommt, daß das modellierte Systemverhalten nur anhand endlich vieler Beobachtungen des Realsystems bewertet werden kann. Damit kann ein Modell zwar falsifiziert oder endlich oft bestätigt, aber niemals mit letzter Gewißheit nachgewiesen werden.[16] Dieses Problem tritt beim induktiven Lernen regelmäßig auf und wird als „***Induktionsproblem***"[17] bezeichnet.

Nun stellt die Güte der Approximation nicht das einzige Bewertungskriterium für Datenmuster dar. So fordern FAYYAD, PIATETSKY-SHAPIRO und SMYTH, Datenmuster sollten

---

[15] Der Modelltyp und der dadurch definerte Lösungsraum werden in Abschnitt 2.2.2 ausführlich behandelt.

[16] Vgl. CHMIELEWICZ (1979), S. 100 ff.

[17] Vgl. TIETZEL (1985), S. 104.

nicht nur valide, sondern auch neuartig, potentiell nützlich und verständlich sein.[18] Diese Kriterien evaluiert i.d.R. eine **Bewertungsfunktion**.[19]

Die letztgenannte intelligente Komponente, das **Suchverfahren**, umfaßt (außer der Bewertungsfunktion) alle algorithmischen Aspekte eines Data-Mining-Verfahrens.[20]

### 2.1.2 Definition und Zielsetzung des Data Mining

Es erscheint zweckmäßig, die im Abschnitt zuvor genannte Definition des Data-Mining-Begriffs anzupassen, da ein entscheidender Aspekt fehlt: die Evaluation der Datenmuster, ohne die kein Data-Mining-Algorithmus auskommen kann. Ohne eine automatische Bewertung und Filterung der besten Datenmuster würde der Benutzer, wie im Abschnitt zuvor erläutert wurde, mit einer unüberschaubar großen Anzahl von Datenmustern konfrontiert. Damit kommt man zu folgender Definition:

**Definition 2-1: Data Mining**

*„Data Mining"* bezeichnet die Anwendung autonomer Algorithmen auf eine Trainingsmenge, $O^T$, die als Output eine Menge von interessanten Datenmustern, $M_{O^T} = \{s_{O_1}, \ldots, s_{O_M}\}$ (mit $O_1, \ldots, O_M \subseteq O^T$), liefern. Dabei stehen:

$M_{O^T}$ für die aus $O^T$ generierte Datenmustermenge,

$M$ für die Anzahl der gefundenen Datenmuster,

$O_i$ für die durch das $i$-te Datenmuster beschriebene Objektmenge und

$s_{O_i}$ für das $i$-te Datenmuster selbst (mit $i = 1, \ldots, M$). ◊

Die in der Data-Mining-Definition verwendeten Terme „Trainingsmenge", „Datenmuster" und „interessant" werden im folgenden genauer bestimmt.

**Definition 2-2: Trainingsmenge**

Gegeben seien eine Grundgesamtheit von Planungsobjekten[21], $O$, eine Menge von beobachteten Attributen, $A = \{a_1, \ldots, a_{amax}\}$, und eine Menge von Attributen, $SA = \{a_{amax+1}, \ldots,$

[18] Vgl. FAYYAD/PIATETSKY-SHAPIRO/SMYTH (1996), S. 6.

[19] Die Bewertungskomponente wird in Abschnitt 2.2.4 behandelt.

[20] Die Suchverfahrenkomponente wird in Abschnitt 2.2.2.3.4 behandelt.

$a_{amax+samax}\}$, die zusammen als Schlüsselattribute dienen. Letztere besitzen keine inhaltliche Bedeutung, sondern identifizieren die Datenobjekte aus der Grundgesamtheit. Als „*Trainingsmenge*“ bezeichnet man eine Menge von Planungsobjekten, $O^T = \{o_1, ..., o_N\}$, mit:

$O^T \subseteq O;$

$o_i = (a_1(o_i), ..., a_{amax+samax}(o_i));$

$i = 1, ..., N;$

$a_j: \quad O \rightarrow dom(a_j);$

$\quad o \rightarrow a_j(o);$

$j = 1, ..., amax+samax.$

Dabei bezeichnen $a_j(o)$ den Wert des Attributes $a_j$ für das Datenobjekt $o$, $N$ die Anzahl der Planungsobjekte, $amax$ die Anzahl der beobachteten und $samax$ die Anzahl der Schlüsselattribute. ◊

**Definition 2-3: Datenmuster**

Gegeben seien dieselben Voraussetzungen wie in der Definition zuvor. Dann ist ein *Datenmuster*, $s_{O'} \in L'$, ein Modell aus einem Lösungsraum[22], $L'$, das eine Teilmenge, $O' \subseteq O^T$, aus der Trainingsmenge, $O^T$, beschreibt, ohne die Objekte aus $O'$ einfach nur aufzulisten. ◊

Wenn im folgenden die Datenmenge $O'$, auf die sich ein Datenmuster, $s_{O'}$, bezieht, nicht von Bedeutung ist, dann wird statt $s_{O'}$ einfach nur $s$ geschrieben, d.h. es gilt $s = s_{O'}$.

Nachzutragen bleibt noch die Definition der Interessantheitsbewertung. Unterschieden werden sollen hier die Bewertung eines einzelnen Datenmusters und die Bewertung einer Menge von Datenmustern.

---

[21] Der Begriff des Planungsobjektes wird hier bereits im Hinblick auf die Anwendung des Data Mining in der Unternehmensplanung verwendet.

[22] Der Begriff des Lösungsraums wird in Definition 2-15 formal definiert. Dies kann erst in Abschnitt 2.2.2.2 erfolgen, da hierzu ein konkreter Modelltyp eingeführt werden muß.

**Definition 2-4: Interessantheitsgrad eines Datenmusters**

Gegeben seien mit $L'$ ein Lösungsraum, mit $O$ die Grundgesamtheit und mit $\boldsymbol{R}$ die Menge der reellen Zahlen. Allgemein wird davon ausgegangen, daß der Interessantheitsgrad eines Datenmusters, $s \in L'$, durch eine Funktion

$$ig: Pot(O) \times L' \rightarrow \boldsymbol{R},$$
$$(O^T; s) \rightarrow ig(O^T; s),$$

quantifiziert werden kann. ◊

Die Interessantheitsfunktion greift während der Modellgenerierung auf die Trainingsmenge zu, was durch den Parameter $O^T$ symbolisiert wird. Da $O^T$ während des Trainings i.d.R. konstant bleibt, kann der Parameter $O^T$ auch weggelassen werden, so daß gilt: $ig(O^T; s) = ig(s)$.

**Definition 2-5: Interessantheitsgrad einer Menge von Datenmustern**

Gegeben seien mit $L'$ ein Lösungsraum, mit $O^T$ eine Trainingsmenge und mit $\boldsymbol{R}$ die Menge der reellen Zahlen. Der Interessantheitsgrad einer Menge von Datenmustern, $M_{O^T}$, ist definiert als:

$$IG: Pot(O) \times Pot(L') \rightarrow \boldsymbol{R}$$
$$(O^T; M_{O^T}) \rightarrow IG(O^T; M_{O^T})$$

◊

Im folgenden kann auch hier der Parameter $O^T$ weggelassen werden, so daß gilt: $IG(O^T; M_{O^T}) = IG(M_{O^T})$.

Die Bewertung einer Datenmustermenge, $M_{O^T} = \{s_{O_1}, \ldots, s_{O_M}\}$, hängt funktional von der Bewertung der einzelnen Datenmuster ab: $IG(M_{O^T}) = f\left(ig(s_{O_1}), \ldots, ig(s_{O_M})\right)$.

### 2.1.3 Differenzierung der Problemstellungen im Data Mining

Im Data Mining werden zwei grundsätzlich verschiedene Problemstellungen verfolgt. Diese Problemstellungen sollen nun formal definiert werden.

**Definition 2-6: Data Mining als Suchproblem** [23]

$O$ sei die Grundgesamtheit aller Objekte und $\boldsymbol{R}$ die Menge der reellen Zahlen. Gesucht sind alle interessanten Datenmuster aus einem Lösungsraum, $L'$. *Data Mining als Suchproblem* läßt sich als 5-Tupel $(L^*, L', ig, O^T, ig^{min})$ definieren mit:

| | |
|---|---|
| $L^* := \{s \in L' \mid ig(O^T; s) \geq ig^{min}\}$ | gesuchte Menge der Datenmuster, die mindestens vom Interessantheitsgrad $ig^{min}$ sind (auch „***Lösungsmenge***“ genannt); |
| $L'$ | Lösungsraum; |
| $ig$ | Interessantheitsgrad eines Datenmusters; |
| $O^T \subseteq O$ | Trainingsmenge; |
| $ig^{min} \in \boldsymbol{R}$ | minimal erwünschter Interessantheitsgrad; ◊ |

Das Suchproblem ist gelöst, sobald die Lösungsmenge, $L^*$, vollständig bestimmt wurde.

**Definition 2-7: Data Mining als Optimierungsproblem**

$O$ sei die Grundgesamtheit aller Objekte und $\boldsymbol{R}$ die Menge der reellen Zahlen. Gesucht ist die interessanteste Menge von Datenmustern. Es sei $L$ der Lösungsraum, der alle möglichen Datenmuster-Mengen umfaßt, d.h. $L = Pot(L')$. *Data Mining als Optimierungsproblem* ist als 4-Tupel $(L^*, L, IG, O^T)$ definiert mit:

| | |
|---|---|
| $L^* := \{M^* \in L \mid IG(O^T; M^*) = \max_{M \in L} IG(O^T; M)\}$ | Menge der optimalen Modelle, $M^*$; |
| $L$ | Lösungsraum; |
| $IG$ | Interessantheitsgrad einer Datenmuster-Menge; |
| $O^T \subseteq O$ | Trainingsmenge; ◊ |

Das Optimierungsproblem ist gelöst, sobald eine Lösung aus der Lösungsmenge, $L^*$, ermittelt wurde. Würde man dagegen etwa jede Datenmustermenge ausgeben, die einen gewissen Mindestinteressantheitsgrad aufweist, so müßte aus allen Ausgabemengen intellektuell die beste ausgewählt werden, wofür wieder ein eigenes Auswahlkriterium benötigt würde. Doch man geht davon aus, daß alle Aspekte der Interesantheit

[23] Vgl. zu dieser Definition MANNILA (1997), S. 43.

quantifiziert und durch das Data-Mining-Verfahren optimiert werden können, so daß eine anschließende manuelle Auswahl aus einer Ergebnismenge entfallen kann.

Unabhängig davon, ob das Data Mining als Such- oder als Optimierungsproblem verstanden wird, sind für die Modellbildung die Trainingsmenge $O^T$, der Lösungsraum $L'$ bzw. $L$ und die Interessantheitsbewertung $ig$ bzw. $IG$ festzulegen. Im weiteren Verlauf wird Definition 2-7 Verwendung finden, da nur die Betrachtung einer Menge von Datenmustern die folgenden Fragen beantworten kann:

- ⇨ Wie sollen Objekte bei der Bewertung berücksichtigt werden, die von mehreren Datenmustern simultan erfaßt werden – insbesondere wenn deren Aussagen sich widersprechen?
- ⇨ Wie sollen Objekte bei der Bewertung berücksichtigt werden, die von keinem Datenmuster erfaßt wurden?
- ⇨ Wie soll die Anzahl der Datenmuster in dem Gesamtmodell in die Bewertung einfließen?

*Man betrachte die folgende Regelmenge zur Prognose der Laufzeiten von Versicherungsverträgen:*

1. `WENN Startalter ≤ 25 UND Beruf = Angestellter`
   `DANN Vertragslaufzeit ∈ [1; 3].`
2. `WENN Geschlecht = männlich UND Beruf = Angestellter`
   `DANN Vertragslaufzeit ∈ [6; 10].`

*Die folgenden Punkte können nur in die Bewertung der ganzen Regelmenge einfließen:*

- ⇨ *Sehr viele Kunden werden durch diese Regeln simultan erfaßt. So können für einen Kunden widersprüchliche Konklusionen ausgesprochen werden. Beispielsweise wird für männliche Angestellte, die mit 25 oder weniger Jahren ihren ersten Versicherungsvertrag abgeschlossen haben, ausgesagt, daß ihre voraussichtliche Vertragslaufzeit zwischen 1 und 3 und zwischen 6 und 10 Jahren liegt.*
- ⇨ *Außderdem werden durch diese Regelmenge keine Kunden erfaßt, die nicht im Angestelltenverhältnis beschäftigt sind, und auch keine weiblichen Angestellten sowie Angestellte, deren Startalter über 25 Jahren liegt.*
- ⇨ *Sollen die Regeln manuell interpretiert werden, so ist ein Modell mit zwei Datenmustern schnell zu überblicken, ein Modell mit 200 Datenmustern (ohne weitere Strukturierungsregeln) dagegen kaum noch.*

Entsprechende Ansätze in der Literatur[24], die Data Mining als Suchproblem definieren und Datenmuster einzeln anstelle von Datenmuster-Mengen bewerten, müssen als ungeeignet eingestuft werden.

[24] Vgl. beispielsweise KRABS (1994), S. 25 ff. oder MANNILA (1997), S. 43.

Eine andere, an die Lösungsmethoden angelehnte Differenzierung der Problemstellungen im Data Mining unterscheidet folgende Aufgabenbereiche:[25]

- ⇨ ***Klassifikation***: Erlernen einer Funktion, $O \rightarrow \{c_1,...,c_{cmax}\}$, die einem neuen, nicht zum Trainieren verwendeten Objekt, $o \in O^{neu},\ O^{neu} \cap O^T = \varnothing,\ O^{neu}, O^T \subset O$, eine der vordefinierten Klassen $c_1,...,c_{cmax}$ zuordnet;
- ⇨ ***Prognose***: Erlernen einer Funktion, $O \rightarrow R',\ R' \subseteq \boldsymbol{R}$ ($\boldsymbol{R}$: Menge der reellen Zahlen), die einem neuen Objekt einen Prognosewert aus einem Wertebereich, $R'$, mit kardinalem Skalenniveau zuordnet;
- ⇨ ***Clusterung***: Zuordnung von $N$ Objekten, $O^T = \{o_1,...,o_N\}$, zu $cmax$ Gruppen: $O^T \rightarrow \{1,...,cmax\},\ 1 < cmax < N$ ;
- ⇨ ***Zusammenfassung***: Erlernen einer kompakten Beschreibung für eine Teilmenge von Daten;[26]
- ⇨ ***Abhängigkeitsmodellierung***: Erlernen einer Funktion vom Typ $dom(a_{i(1)}) \times ... \times dom(a_{i(n)}) \rightarrow dom(a_{i(n+1)})$, die signifikante Abhängigkeiten zwischen Attributen beschreibt (mit $A = \{a_1,...,a_{amax}\};\ \forall k,j \in \{1,...,n+1\}: i(j) \in \{1,...,amax\};\ k \neq j \Rightarrow i(k) \neq i(j)$ );
- ⇨ ***Änderungs- und Abweichungserkennung***: Entdecken von signifikanten Datenänderungen von Attributwerten, $a_i(o)$, bezüglich früherer oder normativer Werte, $a_j(o)$, mit $j \neq i,\ j,i \in \{1,...,amax\}$.

Diese Differenzierung wurde hier wiedergegeben, da sie sehr weit verbreitet ist. Dabei sind folgende Aspekte problematisch:

- ⇨ Die Clusterung selbst stellt nur eine Zuordnung von Objekten zu Gruppen (Segmenten) dar. Zur Verwendung dieser Clusterung ist – wie bereits die Definition des Data Mining fordert – immer auch eine Beschreibung der Segmente durch die Attribute der Trainingsmenge, $A$, erforderlich. Die Identifizierung und Beschreibung von Segmenten soll hier als „***Segmentierung***“ bezeichnet werden.

---

[25] Vgl. FAYYAD/PIATETSKY-SHAPIRO/SMYTH (1996), S. 85. Eine Übersicht über alternative Differenzierungsansätze gibt SÄUBERLICH (2000), S. 41.

[26] Die Art der Zusammenfassung kann sehr unterschiedlich sein, z.B. das Ergebnis einer Generalisierung nach ESTER und SANDER (VGL. ESTER/SANDER (2000), S. 190 f.). Sie verstehen darunter das Ersetzen aller Attributwerte, $a(o)\ \forall o \in O^T$, für ein bestimmtes Attribut, $a$, durch den übergeordneten Wert, $Vorgänger(a(o))$, in einem gegebenen Konzeptbaum. Vgl. zum Begriff des Konzeptbaums Abschnitt 2.2.3.2.

⇨ Die Zusammenfassung wird i.d.R. als Generalisierung von Einzelfällen verstanden. Damit stellt sie eher einen möglichen Operator für Data-Mining-Verfahren als eine eigene Analyseaufgabe dar.[27] Denn von der Analyseaufgabe hängt ab, *welche* Einzelfälle zusammengefaßt und beschrieben werden. Weiterhin ist anzuzweifeln, ob überhaupt sinnvolle Anwendungen für die Generalisierung einer Gruppe von Einzelfällen existieren.

*Beispielsweise könnte man alle Kunden zusammenfassen und beschreiben, die ein bestimmtes Produkt kaufen. Im Rahmen sog. „Cross Selling"-Aktionen könnte man dann solchen Kunden mit ähnlichen Beschreibungen, die das Produkt bisher nicht gekauft haben, ein entsprechendes Angebot unterbreitet. Eine derartige Vorgehensweise macht allerdings keinen Sinn, da man den Kauf bzw. Nichtkauf eines Produktes ebenfalls als Merkmal in der Datenbasis vorsehen könnte. Wenn dann der Unterschied zwischen Käufern und Nichtkäufern durch Einflußfaktoren begündet ist, welche in der Datenbasis erfaßt sind, dann sollte man eher diese Einflußgrößen als erklärende und den Kauf bzw. Nichtkauf als zu erklärende Größe definieren und die Abhängigkeiten zwischen diesen Größen modellieren. Dies wäre dann ein Anwendungsfall für die o.g. „Abhängigkeitsmodellierung".*

⇨ Die Änderungs- und Abweichungserkennung allein ist noch kein Data Mining, wenn die Abweichungen nur *erkannt* werden. Werden sie aber *erkannt und durch eine generelle Aussage beschrieben*, so kann diese Aufgabe als Spezialfall der Abhängigkeitsmodellierung betrachtet werden, da eine Wertänderung bzw. -abweichung eine spezielle Variable darstellt, deren Abhängigkeiten von speziellen Einflußfaktoren offengelegt werden sollen, z.B.:

*WENN Region = ... UND Vertreter = ... UND Produkt = ... UND Zeitraum = ... DANN Plan-Ist-Umsatzabweichung = hoch.*

⇨ Die Differenzierung der genannten Data-Mining-Aufgaben beruht auf verschiedenen Kriterien:

→ So besteht der Unterschied zwischen Klassifikation und Prognose lediglich in dem Skalenniveau der abhängigen Variable. Dieser Unterschied erscheint aus betriebswirtschaftlicher Sicht unbedeutend, so daß diese beiden Aufgabenstellungen im folgenden zu der übergeordneten Aufgabe „***Prognose i.w.S.***" zusammengefaßt werden.

→ Die Differenzierung zwischen Segmentierung (Clusteridentifikation und -beschreibung) und Zusammenfassung (im Sinne einer Generalisierung) besteht

[27] Der Generalisierungsoperator wird durch Definition 2-40 bis Definition 2-42 eingeführt

allein darin, daß bei ersterer mehrere Gruppen und bei letzterer nur eine Gruppe gebildet und beschrieben werden.

→ Die Abgrenzung dieser Gruppierungsaufgaben zur Abhängigkeitsmodellierung besteht darin, daß nur bei letzterer zwischen abhängigen und unabhängigen Größen unterschieden werden kann.

→ Der Unterschied der Abhängigkeitsmodellierung zur Prognose i.w.S. wiederum besteht darin, daß nur bei letzterer eine Anwendung des Modells auf neue Planungsobjekte vorgesehen ist.

Teilweise wird auch die *Identifikation von Ausreißern* zum Data Mining gezählt.[28] Dabei handelt es sich jedoch um eine bloße Identifikation von Einzelfällen ohne die im Data Mining geforderte Beschreibung von Objektmengen.

Als Fazit kann man festhalten, daß vorhandene Differenzierungsansätze für die Aufgabenstellungen im Data Mining weniger geeignet sind. Vorläufig sollen hier aufgrund der vorstehenden Argumentation die Aufgaben „Prognose“, „Abhängigkeitsmodellierung“, „Zusammenfassung“ und „Segmentierung“ differenziert und jeweils als Optimierungsproblem betrachtet werden. Eine geeignetere Differenzierung im Hinblick auf betriebswirtschaftliche Problemstellungen wird in Kapitel 3 vorgenommen.[29]

### 2.1.4 Abgrenzung des Data Mining zu verwandten Gebieten

Im folgenden soll der Begriff des Data Mining zu folgenden verwandten Gebieten abgegrenzt werden:

⇨ zum fallbasierten Schließen (Case Based Reasoning, CBR);

⇨ zur explorativen statistischen Datenanalyse.

Die Grundidee des **fallbasierten Schließens** besteht darin, ein Problem zu lösen, indem man sich an ähnliche Probleme aus der Vergangenheit erinnert, deren Lösung an das aktuelle Problem anpaßt und den neuen Problemfall einschließlich Lösung zur

---

[28] Vgl. SÄUBERLICH (2000), S. 41.

[29] Im dritten Kapitel wird die Aufgabe der „Zusammenfassung“ entfallen, die Aufgabe der „Prognose“ wird übernommen, die „Segmentierung“ wird als „reine Beschreibung“ bezeichnet, die „Abhängigkeitsmodellierung“ wird in „Erklärung“ umbenannt, und es wird eine bisher ungenannte Aufgabe, die „Entscheidung“, hinzukommen.

weiteren Verwendung speichert.[30] Eine Problem im fallbasierten Schließen kann in der Suche nach einer für einen gegebenen Fall optimalen Prognose i.w.S. bestehen. Dieses Problemfeld überschneidet sich mit den Anwendungsbereichen des Data Mining. In der Methodik der Problemlösung sind jedoch folgende Unterschiede auszumachen:

- ⇨ Beim Data Mining wird die Lösung eines Problemfalls durch Anwendung eines allgemeingültigen Modells erzeugt, das zuvor aus den Falldaten induziert wurde. Das Modell findet in der Anwendungsphase sehr schnell zu einer Lösung und kann unabhängig von dem Data-Mining-Verfahren, durch das es erzeugt wurde, kommerziell vertrieben werden.[31] Beim fallbasierten Schließen findet keine Generalisierung statt.[32] Die Falldaten werden zum Zeitpunkt der Anwendung mit allen Einträgen in dem Fallspeicher verglichen, ähnliche Fälle extrahiert und deren Lösung an das aktuelle Problem angepaßt. Die Umgehung der Generalisierung von Einzelfällen hat den Vorteil, daß das CBR auch in Domänen mit vielen Ausnahmen von generellen Zusammenhängen und mit *„missing values“* angewendet werden kann.[33] Der Vorteil des Data Mining besteht darin, daß die generalisierten Datenmuster Informationen über Problemlösungen darstellen, die unabhängig von einem einzelnen Problemfall von Nutzen sein können, z.B. bei der Planung von Maßnahmen für Kundengruppen.
- ⇨ Bei Data Mining ist keine Anpassung an das aktuelle Problem vorgesehen. Entweder stimmen die Merkmale des aktuellen mit denen der vergangenen Probleme so stark überein, daß frühere Lösungen unverändert übernommen werden können, oder das Problem kann nicht gelöst werden. Diesem Nachteil steht der Vorteil gegenüber, daß kein zusätzliches Hintergrundwissen zur Lösungsadaption formalisiert werden muß.
- ⇨ Beim Data Mining ist keine explizite Aufnahme aktueller Problemfälle und -lösungen in den Fallspeicher vorgesehen. Ein CBR-System ist durch die veränderliche Fallbasis nicht so statisch wie ein Modell, das einmal induziert wurde und die

---

[30] Vgl. LENZ (1994), S. 9.

[31] Vgl. LENZ/AURIOL/MANAGO (1998), S. 57.

[32] Vgl. LENZ/AURIOL/MANAGO (1998), S. 58.

[33] Vgl. LENZ/AURIOL/MANAGO (1998), S. 58.

Problemlösung unveränderlich modelliert.[34] Dies kann einen Vorteil bei der Anwendung in einem dynamischen Umfeld darstellen.[35]

Die beiden letztgenannten Punkte – die Lösungsadaption und die Erweiterung des Fallspeichers um neue Fälle – können bei der Entwicklung von Data-Mining-Verfahren integriert werden, z.B. durch ein intellektuell aufgestelltes Modell zur Lösungsadaption und durch ein inkrementelles[36] Lernverfahren. Damit stellen diese Punkte keine unüberwindbaren Unterschiede zwischen Data-Mining- und CBR-Verfahren dar. Der Verzicht auf die Induktion eines Modells stellt dagegen einen fundamentalen methodischen Unterschied von CBR- gegenüber Data-Mining-Verfahren dar.

Zu den klassischen Verfahren der **explorativen statistischen Datenanalyse** gehören u.a. die Faktoranalyse, die Clusteranalyse, die Regressionsanalyse, die Diskriminanzanalyse, die Varianzanalyse und die Kontingenzanalyse. Auch viele modernere Verfahren lassen sich diesen klassischen Verfahren zuordnen. Ohne näher auf die einzelnen Verfahren eingehen zu können, sei hier nur festgehalten, daß die beiden erstgenannten Verfahren zu den struktur*entdeckenden* und die übrigen zu den struktur*prüfenden* Verfahren gezählt werden. Das Data Mining beschäftigt sich mit der *Entdeckung* unbekannter Strukturen, weswegen häufig regelbasierte Modelle oder neuronale Netze zum Einsatz kommen, da diesen Modelltypen eine universelle Approximationsfähigkeit unbekannter Systemstrukturen zugeschrieben wird.[37] Die Varianz-, Kontingenz-, Regressions- und die Diskriminanzanalyse sind dagegen auf die *Prüfung* von Modellen eines vorgegebenen Typs beschränkt, z.B. auf lineare Funktionen. Sie suchen die optimale Modellinstanz und verwenden dabei stets alle vorhandenen Variablen. Außerdem erfolgt die Optimierung anhand von Zielvorstellungen, die erheblich von ökonomischen Zielen abweichen können. Zwar ist es beispielsweise bei der Diskriminanzanalyse möglich, die Klassifikation neuer Objekte so vorzunehmen, daß ein ökonomisches Gütemaß optimiert wird.[38] Allerdings besteht zum Zeitpunkt der Klassifikation das Modell bereits,

---

[34] Vgl. LENZ/AURIOL/MANAGO (1998), S. 59.

[35] Vgl. LENZ/AURIOL/MANAGO (1998), S. 57.

[36] „***Inkrementell***“ ist ein Lernverfahren dann, wenn es nach bereits erfolgter Modellgenerierung neue Fälle berücksichtigen kann, indem es das generierte Modell anpaßt, so daß kein erneutes Lernen („***Relearning***“) notwendig ist (vgl. HOLSHEIMER/SIEBES (1994), S. 40).

[37] Vgl. zur Approximationsfähigkeit neuronaler Netze DÜSING (1997), S. 121 ff. und zu regelbasierten Modellen Abschnitt 2.2.2.1.

[38] Vgl. BACKHAUS ET AL. (2000), S. 184.

so daß es sich hier nur um eine „Nachoptimierung“ handelt.[39] Würde dagegen bereits während der Modellgenerierung eine Erfolgsgröße optimiert, so würde ein Modell generiert, das besonders erfolgswirksame Zusammenhänge besonders gut approximiert.[40] Auch erlauben einige statistische Verfahren eine unterschiedliche Gewichtung einzelner Datensätze oder Outputs, so daß tatsächlich Modelle erzeugt werden können, die besonders erfolgswirksame Zusammenhänge besonders gut approximieren. Doch um ein ökonomisch optimales Modell zu generieren, muß die Zielvorstellung des Entscheidungsträgers optimiert werden. Hierzu müssen dessen Einzelziele und Risikopräferenzen integriert werden, und das Modell müßte in jeder Entscheidungssituation die optimale Entscheidung ausgeben.[41] Dies leisten die bekannten statistischen Verfahren nicht.

Bei der *Faktoranalyse* wird – anders als im Data Mining – kein Modell generiert, das eine Menge von Planungsobjekten beschreibt. Hier geht es um die Entdeckung von unabhängigen Faktoren, auf die die beobachteten Merkmale hindeuten, die aber nicht direkt durch diese gemessen werden.[42] Und auch bei der *Clusteranalyse* wird kein Modell generiert, das eine Menge von Planungsobjekten beschreibt. Es findet lediglich eine Zuordnung der Objekte zu Clustern statt. Die eigentliche Beschreibung der Cluster findet häufig durch eine anschließende Diskriminanzanalyse oder logistische Regression statt. Diese wurden oben bereits von dem Data Mining abgegrenzt. Hinzu kommt noch, daß bei diesem Vorgehen in zwei Schritten zuerst eine an der Clusteridentifikation orientierte Zielfunktion (i.d.R. eine Abstandsfunktion) optimiert wird, wobei immer *alle* Merkmale in die Zielfunktion einfließen.[43] Im zweiten Schritt wird dann eine ganz andere Zielfunktion optimiert, die die Diskriminierung der im ersten Schritt identifizierten Cluster durch ein zu konstruierendes Modell eines vorgegebenen Typs bewertet. Auch hier fließen wieder alle Merkmale in die Zielfunktion ein. Neben dem Nachteil, daß

---

[39] Auf die damit verbundenen Nachteile wird u.a. in Abschnitt 3.2.5 eingegangen. Abschnitt 3.3.2 zeigt auf, wie ökonomische Ziele in das Data Mining integriert werden können.

[40] Die Berücksichtigung von ökonomischen Zielen in der Modellbildungsphase wird in der statistischen Literatur sogar teilweise als nachteilig beschrieben, da bei einer Änderung von Preisen oder Kosten ein Relearning erfolgen muß (vgl. z.B. BONNE (2000), S. 56). Dieser Nachteil fällt weniger ins Gewicht als der Vorteil, erfolgsrelevante Zusammenhänge besonders gut zu approximieren.

[41] Auf diese Problemauspekte wird in Kaptel 3 ausführlich eingegangen.

[42] Vgl. AMBROSI (1985), S. 10 f.

[43] Vgl. BACKHAUS/WEIBER (1989), S. 55.

immer alle Merkmale in dem Modell vorkommen müssen, bringt die schrittweise Optimierung verschiedener Zielfunktionen den Nachteil mit sich, daß bei der ersten Optimierung (der Clusterung) die Beschreibungsfähigkeit der generierten Cluster durch den Typ der Diskriminanzfunktion nicht berücksichtigt wird, denn die spielt erst im zweiten Schritt eine Rolle – dann existieren die Cluster aber schon.[44] Die Anwendung klassischer

Clusteranalyseverfahren hat darüber hinaus den Nachteil, daß man sich auf wenige Variablen beschränken muß.

*Dies kann am Beispiel der Kassenbonanalyse demonstriert werden.[45] Will man die herkömmlichen Clusterverfahren anwenden, so ist man dazu gezwungen, eine Tabelle der folgenden Form aufzubauen (vgl. Tabelle 2-1).*

| ***Bonnr.*** | ***Artikel 1*** | ***Artikel 2*** | ***Artikel 3*** | ***Artikel 4*** | ***Artikel 5*** | ***Artikel 6*** | ***Artikel 7*** | ***...*** | ***Artikel n*** |
|---|---|---|---|---|---|---|---|---|---|
| *1* | *1* | *0* | *0* | *0* | *1* | *0* | *1* | *...* | *0* |
| *2* | *0* | *0* | *0* | *0* | *0* | *0* | *1* | *...* | *0* |
| *3* | *0* | *0* | *0* | *1* | *1* | *0* | *1* | *...* | *0* |
| *4* | *1* | *0* | *0* | *0* | *0* | *0* | *0* | *...* | *1* |
| *...* | *...* | *...* | *...* | *...* | *...* | *...* | *...* | *...* | *...* |

***Tabelle 2-1: Datenbasis für eine Clusteranalyse zur Identifizierung typischer Warenkörbe***

*Jeder Kassenbon bildet eine Zeile in der Tabelle. Jeder potentiell nachgefragte Artikel bildet eine Spalte. Eine Zelle, $z_{ji}$, enthält eine $1$, falls Kassenbon Nr. $j$ Artikel Nr. $i$ enthält und eine $0$, falls nicht. Der Vorteil dieser Modellierung ist darin zu sehen, daß auf diese Datenbasis die weit verbreiteten Clusteranalyseverfahren angewendet werden können, welche die Cluster aufgrund der Abstände der Objekte bilden. Der „Abstand" zwischen den zwei Kassenbons (Zeilen $j$ und $k$) läßt sich beispielsweise als Durchschnitt der $n$ absoluten Differenzen $|z_{ji} - z_{ki}|$ $(i = 1, ..., n)$ berechnen:*

$$d(j,k) = \frac{1}{n}\sum_{i=1}^{n} | z_{ji} - z_{ki} | \cdot$$

*Der Nachteil dieser Modellierung liegt darin, daß die Tabelle i.d.R. sehr breit wird. Wenn dies die Leistungsfähigkeit der zur Verfügung stehenden Datenvorverarbeitungs- und -auswertungsprogramme übersteigt, so muß man sich notgedrungen auf bestimmte Artikel beschränken. Dadurch werden aber möglicherweise wichtige Zusammenhänge übersehen.*

Hinzu kommt als potentielles Abgrenzungskriterium zu den statistischen Verfahren, daß im Data Mining die Notwendigkeit einer engen Kopplung mit der Datenbank gefordert wird.[46] Da dies aber erstens gegenwärtig von kaum einem Data-Mining-Verfahren erfüllt wird und zweitens prinzipiell auch für statistische Verfahren realisierbar wäre, wurde dieses Kriterium hier nicht mit aufgeführt.

---

[44] Vgl. MICHALSKI/STEPP (1983b), S. 396.

[45] Vgl. zu diesem Beispiel FISCHER (1993), S. 117 ff.

[46] Vgl. IMIELINSKI/MANNILA (1996), S. 59.

Zusammengefaßt weisen die im Data Mining angewendeten Algorithmen folgende *abgrenzende Merkmale* zu den betrachteten Verfahren auf:

- ⇨ **Generalisierung**: Im Data Mining werden im Gegensatz zum CBR aus Einzelfällen generelle Modelle induziert.
- ⇨ **Integration von Musteridentifikation und -beschreibung**: Das Data Mining integriert im Gegensatz zur Cluster- und Diskriminanzanalyse die Identifikation und die Beschreibung von Datenmustern.
- ⇨ **Universelle Approximation**: Im Data Mining werden im Gegensatz zu den strukturprüfenden Verfahren Modelle mit universellen Approximationsfähigkeiten generiert. Insbesondere muß man sich im Data Mining nicht a-priori auf wenige Variablen oder Zusammenhänge beschränken.
- ⇨ **Beliebige Zielfunktionen**: Während bei den klassischen statistischen Verfahren ökonomische Ziele durch Gewichtungen einzelner Datensätze oder Outputs angenähert werden, besteht im Data Mining die Möglichkeit, bereits während der Modellbildung direkt eine „beliebige“[47] ökonomische Zielfunktion zu optimieren und – in noch zu spezifizierenden Fällen – automatisch ein Modell zu generieren, daß in einer gegebenen Situation die optimale Entscheidung ausgibt.

### 2.1.5 Einordnung des Data Mining in den “Knowledge Discovery in Databases Process”

Im engeren Sinne umfaßt das Data Mining als Funktion nur die eigentliche Ausführung eines autonomen Suchverfahrens. Vor und nach dieser Ausführung des Suchverfahrens sind noch eine Reihe von Tätigkeiten durchzuführen, die Interaktionen mit einem oder mehreren Benutzern voraussetzen. Der gesamte Auswertungsprozeß wird üblicherweise als „***Knowledge Discovery in Databases (KDD) Process***“ bezeichnet. Er wird in Abbildung 2-2 als ereignisgesteuerte Prozeßkette dargestellt.[48]

[47] Die „Beliebigkeit“ der integrierbaren Zielfunktionen ist insoweit eingeschränkt, daß die Zielfunktion nur auf die vorhandene Datenbasis sowie eventuell vorgegebene Parameter zugreifen kann.

[48] Die Notation ereignisgesteuerter Prozeßketten wird in Anhang B erläutert.

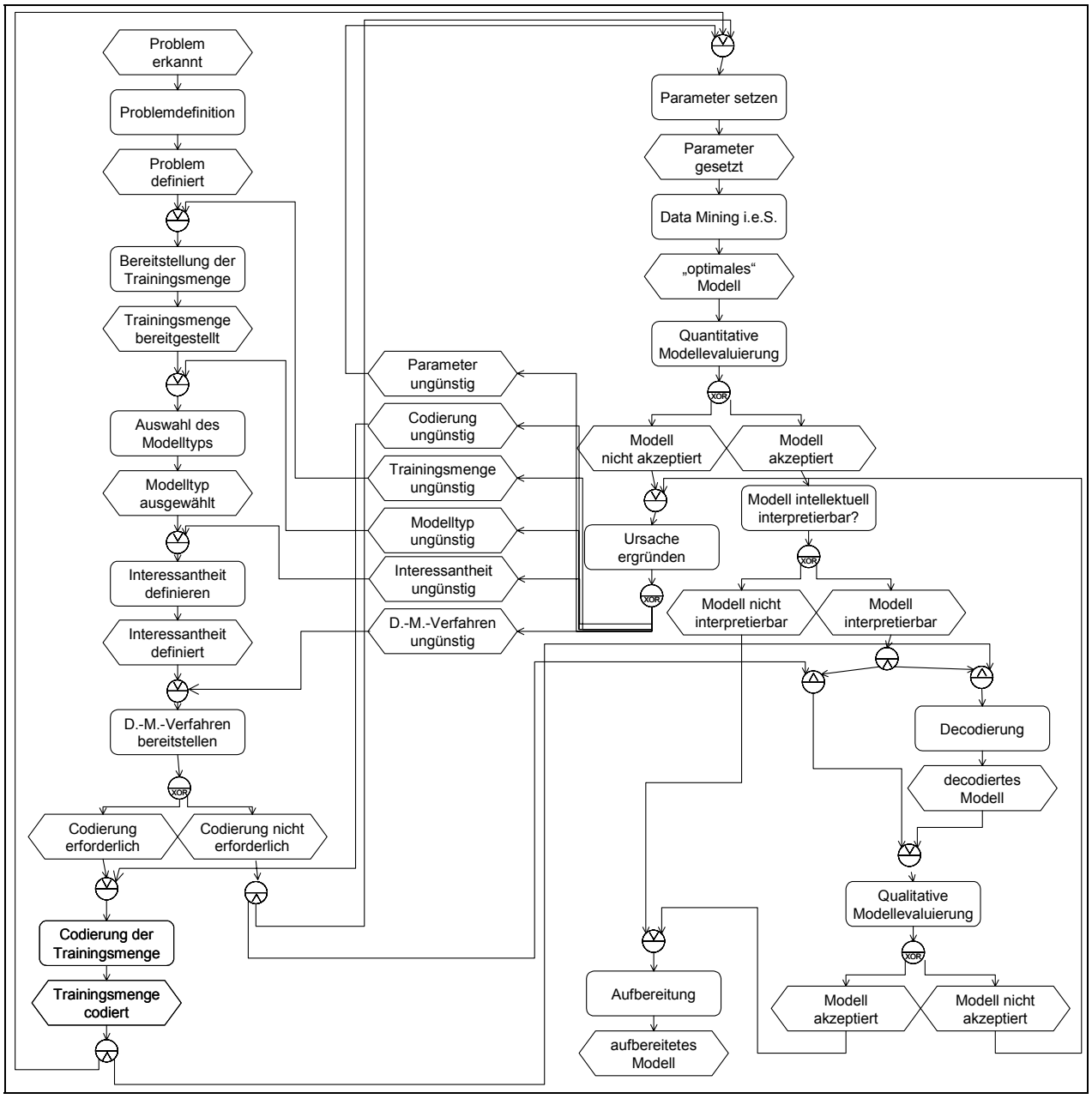



**Abbildung 2-2: Der Knowledge Discovery in Databases Process**

Der KDD-Prozeß umfaßt folgende Tätigkeiten:[49]

⇨ **Problemdefinition**: Wird ein Problem erkannt, das potentiell durch Data-Mining-Verfahren lösbar ist, so muß zunächst das Problem genauer definiert werden. Als Ergebnis muß ein definiertes Problem vorliegen, welches das abzubildende Realsystem, die Zieldefinition sowie eine grobe Klassifikation des Problemtyps umfaßt.

[49] Vgl. FAYYAD/PIATETSKY-SHAPIRO/SMYTH (1996), S. 9 ff., BRACHMAN/ANAND (1996), S. 42 ff., KAFKA (1999), S. 42 ff. und CHAPMAN ET AL. (2000), S. 13 ff. Eine vergleichende Übersicht verschiedener Prozeßmodelle für das KDD gibt SÄUBERLICH (vgl. SÄUBERLICH (2000), S. 22 ff.).

⇨ **Bereitstellung der Trainingsmenge**: Ausgehend von der Problemstellung müssen die für relevant befundenen Attribute und Datenobjekte aus den Datenbeständen des Unternehmens extrahiert werden. I.d.R müssen hierzu diverse Datentransformationen durchgeführt werden, wie z.B. Projektionen, Verbund-Operationen, Selektionen, Gruppierungen, Aggregationen oder die Definition sonstiger abgeleiteter Attribute. Mit der Selektion der Datenobjekte legt man den maximal möglichen Stichprobenumfang für das spätere Data Mining fest.

Voraussetzung für das Data Mining ist eine Datenbasis in ausreichender Qualität. Um dies zu gewährleisten, muß die Datenbasis eventuell manuell bereinigt werden. Hierzu gehört eine Beseitigung von Tippfehlern und Inkonsistenzen. Bei der Zusammenführung verschiedener Datenquellen müssen darüber hinaus eventuell Daten vereinheitlicht werden.

*Beispielsweise können in verschiedenen Datenquellen die Angaben des Geschlechts einmal als „männlich" bzw. „weiblich" und einmal als „m" bzw. „w" codiert sein.*

⇨ **Auswahl des Modelltyps**: Je nach Problemstellung und Datenbasis muß die Wahl für einen bestimmten Modelltyp getroffen werden.[50]

⇨ **Interessantheit definieren**: Erst jetzt kann ein Maß für die Bewertung der Interessantheit definiert werden, da sich die Bewertung immer an dem konkreten Problemtyp und an dem ausgewählten Modelltyp orientiert.

⇨ **Data-Mining-Verfahren bereitstellen**: Wenn Problemstellung, Trainingsmenge, Datenbasis und Interessantheitsmaß feststehen, dann kann ein diesen Anforderungen genügendes Verfahren gesucht werden. Zur Zeit muß für gehobene Anforderungen zumeist ein entsprechendes Verfahren eigens entwickelt werden.

⇨ **Codierung der Trainingsmenge**: Bestimmte Verfahren setzen eine spezielle Codierung der Trainingsdaten voraus. Neuronale Netze beispielsweise benötigen häufig Daten mit den Wertebereichen $\{0,1\}$ oder $\boldsymbol{R} \cap [0;1]$.[51] Fuzzy-Verfahren setzten u.U. unscharfe Inputdaten voraus. Und auch Data-Mining-Verfahren, die auf den klassischen statistischen Verfahren beruhen, setzen bestimmte Skalenniveaus voraus, so daß eventuell Skalentransformationen notwendig sind. Wenn das gewählte

[50] Die Auswahl des Modelltyps wird in Abschnitt 2.2.2.1 genauer betrachtet.

[51] Vgl. LACKES/MACK/TILLMANNS (1998), S. 254.

Data-Mining-Verfahren keine besonderen Anforderungen an die Codierung stellt, kann dieser Schritt entfallen.

⇨ **Parameter setzen**: Bevor das Data-Mining-Verfahren gestartet werden kann, müssen i.d.R. bestimmte Parameter gesetzt werden, die den Modelltyp (und damit den Lösungsraum) eingrenzen, das Interessantheitsmaß spezifizieren und die zu startende Instanz des Lösungsverfahrens definieren.

⇨ **Data Mining i.e.S.**: Das Data Mining i.e.S. umfaßt die Ausführung eines autonomen Verfahrens. Als Ergebnis liegt ein bezüglich des definierten Interessantheitsmaßes „optimales" Modell vor – wobei „optimal" als „beste gefundene Lösung" zu verstehen ist, da Data-Mining-Verfahren i.d.R. Heuristiken[52] und keine exakten Verfahren sind.

⇨ **Quantitative Modellevaluierung**: Bei der quantitativen Modellevaluierung wird das „optimale" Modell auf eine Menge von sog. „Validierungs-" oder „Testdaten", $O^E \subset O$, angewendet, die zu den Trainingsdaten disjunkt sind, d.h. $O^E \cap O^T = \varnothing$. Dabei wird ein vorher festzulegendes Evaluierungskriterium ausgewertet. Anschließend ist aufgrund des Kriteriums und eventuell aufgrund von Vergleichen mit den Evaluierungsergebnissen von Benchmarks zu entscheiden, ob das Modell akzeptiert wird oder nicht.

⇨ **Ursache ergründen**: Kommt man zu dem Schluß, daß das „optimale" Modell bezüglich des Validierungskriteriums oder bezüglich seiner praktischen Verwendbarkeit zu schlecht ist, so kommen dafür verschiedene Ursachen in Frage: Möglicherweise waren nur die Modell-, Interessantheits- oder Verfahrensparameter schlecht gewählt, so daß diese variiert und ein neuer Data-Mining-Lauf gestartet werden sollte. Dabei ist zu beachten, daß dieselbe Validierungsmenge – wie in Abschnitt 2.2.4.2 diskutiert werden wird – streng genommen nur ein einziges Mal verwendet werden darf.

Auch eine ungünstige Codierung kann schlechte Ergebnisse erzeugen, so daß eventuell die Codierung angepaßt werden muß. Bezüglich der gewählten Parameter und Codierung sollten Sensitivitätsanalysen durchgeführt werden. Bei der Verwendung von Verfahren, die sehr sensitiv auf die gewählte Skalierung reagieren, sind

---

[52] Vgl. zum Begriff der Heuristik VOß/FIEDLER/GREISTORFER (2000), S. 554.

u.U. häufige Rücksprünge zur Codierungsphase notwendig. Dies kann den Aufwand von KDD-Studien enorm erhöhen.

Bei einem schlechten Validierungsergebnis muß möglicherweise ein anderes Data-Mining-Verfahren oder ein anderes Interessantheitsmaß verwendet werden. Noch ungünstiger ist die Situation, falls bereits der Modelltyp schlecht gewählt war oder falls sich herausstellt, daß die Trainingsdaten zu stark verrauscht sind.

- ⇨ **Modell intellektuell interpretierbar**: Es muß unterschieden werden, ob ein Modell intellektuell interpretiert werden soll oder als „black box“ z.B. für Prognosezwecke verwendet werden soll. Im erstgenannten Fall muß es eventuell decodiert und qualitativ evaluiert werden – im zweitgenannten Fall können diese Schritte entfallen.
- ⇨ **Decodierung**: Eine Decodierung der Ergebnisse aus der Data-Mining-Phase ist nur dann erforderlich, wenn zuvor eine Codierung erfolgte und das Modell manuell interpretiert werden soll. Dann müssen die codierten Ergebnisse in eine lesbare Form transformiert werden.
- ⇨ **Qualitative Modellevaluierung**: Interpretierbare Modelle – insbesondere solche, die nicht auf neue Daten angewendet werden – sollten bezüglich bereits vorhandenen Wissens evaluiert werden, um zu garantieren, daß die ermittelten Strukturen tatsächlich vorliegen und nicht in die Daten „hineingelesen“[53] wurden.
- ⇨ **Aufbereitung**: Die akzeptierten Modelle werden nach verschiedenen Kriterien geordnet, selektiert, verdichtet und visualisiert, so daß sie dem Management präsentiert werden können. Werden die präsentierten Ergebnisse für weiterverwendbar erachtet, ist zu entscheiden, in welcher Form die Ergebnisse dem Endanwender verfügbar gemacht werden sollen, z.B. als Bericht, als Aktionsplan, als Monitor oder als Spezifikation einer KDD-Applikation.[54] Ein ***Bericht*** stellt lediglich die Analyseergebnisse dar und erläutert sie. Ein ***Aktionsplan*** umfaßt bereits eine auf Grundlage der Analyseergebnisse getroffene Entscheidung und Anweisungen zu deren Umsetzung. Ein ***Monitor*** ist ein Trigger in der Datenbank, der bei bestimmten Ereignissen eine Aktion auslöst, z.B. eine Alarmmeldung, daß eine bestimmte Kennzahl einen

---

[53] Vgl. BUHMANN (1998), S. 37.

[54] Vgl. BRACHMAN/ANAND (1996), S. 47 f.

Planwert unterschreitet, wenn zuvor herausgefunden wurde, daß diese Kennzahl einen wichtigen Einflußfaktor auf eine Erfolgsgröße darstellt. „***KDD***-“ oder auch „***Data-Mining-Applikationen***“ unterstützen im Gegensatz zu Data-Mining-Verfahren die Anwendung und nicht die Generierung der Modelle. KDD-Applikationen werden von Analytikern so konfiguriert, daß sie von der Fachabteilung ohne Methodenkenntnisse eingesetzt werden können.[55] Dazu benutzen sie zur Definition der Aufgabenstellung und zur Präsentation der Ergebnisse die für den Endanwender vertraute Sprache aus seinem Geschäftsbereich.[56]

*Vordefinierte Objektmengen beispielsweise werden als „Stammkunden“, „Yuppies“, „gute Kunden“, „Kernartikel“, „Cash Cows“ oder „erfolgreiche Außendienstmitarbeiter“ bezeichnet. Data-Mining-Applikationen unterstützen eine konkrete fachliche Problemstellung, wie z.B. die Zielkundenselektion im Rahmen von Direct-Mail-Kampagnen.*

Die Schritte von der Auswahl des Modelltyps bis zur Decodierung können auch als „***Data Mining im weiteren Sinne (i.w.S.)***“ und der gesamte KDD-Prozeß als „***Data Mining im weitesten Sinne***“ bezeichnet werden. Es existieren Ansätze zur Automatisierung des gesamten KDD-Prozesses.[57] Diese Untersuchung fokussiert das Data Mining i.e.S. und bezieht einige Aspekte des Data Mining i.w.S. mit ein.

## 2.2 Arbeitsweise und Komponenten eines Data-Mining-Verfahrens

Bereits in Abschnitt 2.1.1 wurden die intelligenten Komponenten eines Data-Mining-Verfahrens, die die Annahmen über das betrachtete Realsystem modellieren, eingeführt. Diese Komponenten werden in den Abschnitten 2.2.2 bis 2.2.4 detaillierter vorgestellt. Zuvor gibt Abschnitt 2.2.1 einen Überblick über die Arbeitsweise von Data-Mining-Verfahren.

### 2.2.1 Überblick über die Arbeitsweise eines Data-Mining-Verfahrens

Die meisten Data-Mining-Verfahren arbeiten nach demselben Prinzip, das sich wie folgt zusammenfassen läßt (vgl. Abbildung 2-3):

---

[55] Vgl. BACHMAN/ANAND (1996), S. 55.

[56] Vgl. BRACHMAN ET AL. (1996), S 44.

[57] Vgl. ENGELS (1999), S. 85 ff.

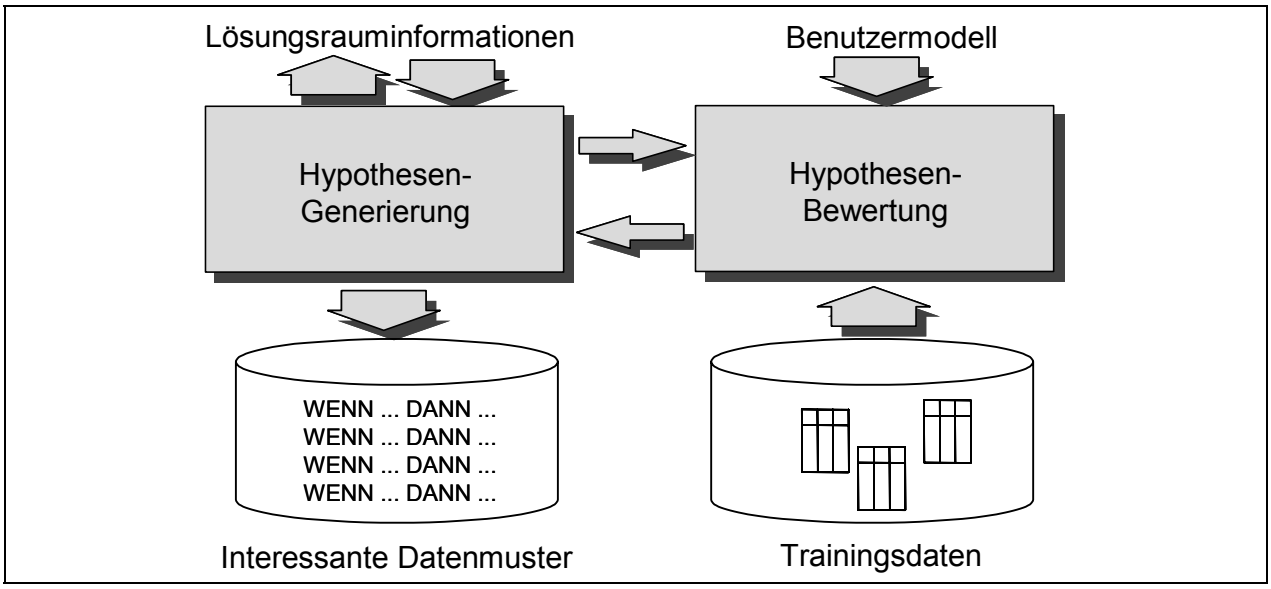



**Abbildung 2-3: Arbeitsweise eines Data-Mining-Verfahrens**[58]

Die Suche nach interessanten Datenmustern erfolgt in einem Wechsel aus Hypothesengenerierung und -bewertung. **Hypothesen werden generiert**, indem von der aktuellen Position im Lösungsraum, welche in Abbildung 2-3 den Lösungsrauminformationen entnommen wird, benachbarte Lösungen (Hypothesen) besucht werden, die anschließend an die Hypothesenbewertung weitergeleitet werden. Die **Bewertung der Hypothesen** erfolgt anhand der Trainingsdaten, welche aus einer Datenquelle in die Bewertungskomponente eingelesen werden. Außerdem können in die Bewertung Informationen aus einem Benutzermodell einfließen, z.B. Präferenzen des Benutzers bezüglich bestimmter Arten von Datenmustern oder bereits vorhandenes Wissen des Benutzers. Die Bewertungen der Hypothesen werden an die Hypothesengenerierungskompenente zurückgegeben, wo z.B. Statistiken über besonders gut bewertete Hypothesen zu den Lösungsrauminformationen gespeichert werden. Die interessanten Datenmuster werden abgespeichert, um sie später dem Benutzer präsentieren zu können.

Angemerkt sei noch, daß hier, wie aus Abbildung 2-3 hervorgeht, unter einem Benutzermodell nur diejenigen benutzerbezogenen Informationen verstanden werden, die sich auf die Bewertung beziehen. In Abgrenzung hierzu können auch gewisse Lösungsrauminformationen benutzerbezogen sein.

---

[58] In Anlehnung an: BREITNER/LOCKEMANN/SCHLÖSSER (1998), S. 45. Die Pfeile stellen Informationsflüsse dar, welche im Text erläutert werden.

*Ein Beispiel für solche Lösungsrauminformationen wären vom Benutzer vorgegebene Aussagenschablonen der Form:*

*`Vertreter x hat in Region y z Prozent des Umsatzes verloren.`*

*Die Variablen dieser Aussagenschablonen werden durch das Data-Mining-Verfahren so instanziiert, daß sich interessante Aussagen ergeben. Interessante Aussagen sind beispielsweise solche, die hohe Umsatzverluste, `z`, beschreiben.*

### 2.2.2 Die Repräsentation von Modellen im Data Mining

Die erste zu behandelnde Komponente von Data-Mining-Verfahren ist die Wissensrepräsentationskomponente, die den Lösungsraum definiert. Zunächst werden in Abschnitt 2.2.2.1 Kriterien zur Auswahl einer Repräsentationsform eingeführt. Anschließend wird in Abschnitt 2.2.2.2 die in der weiteren Arbeit verwendete Repräsentationsform einer Regel in konjunktiver Normalform eingeführt. Abschnitt 2.2.2.3 stellt verschiedene Möglichkeiten zur Strukturierung einer Menge solcher Regeln vor.

#### *2.2.2.1 Kriterien zur Auswahl einer Repräsentationsform*

In Abschnitt 2.1.1 wurde bereits darauf hingewiesen, daß ein intelligentes System, das ein Modell eines Realitätsauschnittes erlernt, bestimmte Voraussetzungen über die Beschaffenheit des Realsystems treffen muß. Zu diesen Voraussetzungen gehörten u.a. Annahmen über den Typ des realen deterministischen Zusammenhangs und damit über den Typ des Modells.

Durch die Wahl eines Repräsentationsformalismus‘ zur Beschreibung von Objektmengen wird die Menge potentiell identifizierbarer Muster vorweg auf diejenigen Muster eingeschränkt, die durch den Formalismus beschreibbar sind. Wenn der Modelltyp auf falschen oder vereinfachten Annahmen über das Realsystem basiert, kann es vorkommen, daß der gesuchte wahre Zusammenhang nicht einmal Element des Lösungsraums ist.[59] Daher stellen die *Annahmen des Analytikers über die Beschaffenheit des Realsystems* das wichtigste Kriterium zur Auswahl eines Modelltyps dar. Hierzu sind Annahmen über potentiell interessante Zusammenhangstypen unvermeidbar, wie das folgende Beispiel zeigt:

---

[59] Dies wurde bereits an Abbildung 2-1 (S. 9) verdeutlicht.

*Gegeben sei eine relationale Datenbank mit den folgenden beiden Tabellen:*[60]

| ***Person*** | | | | |
|---|---|---|---|---|
| ***Name*** | ***Alter*** | ***Geschlecht*** | ***Einkommen*** | ***Kunde*** |
| *Ann* | *32* | *w* | *10.000* | *ja* |
| *Joan* | *53* | *w* | *1.000.000* | *ja* |
| *Mary* | *27* | *w* | *20.000* | *nein* |
| *Jane* | *55* | *w* | *20.000* | *ja* |
| *Bob* | *30* | *m* | *100.000* | *ja* |
| *Jack* | *50* | *m* | *200.000* | *ja* |

| ***verheiratet*** | |
|---|---|
| ***Mann*** | ***Frau*** |
| *Bob* | *Ann* |
| *Jack* | *Jane* |

***Tabelle 2-2: Eine Beispiel-Datenbank mit zwei Tabellen***

*Ein Data-Mining-Verfahren, das Regeln in konjunktiver Normalform (KNF)*[61] *darstellt, könnte aus dieser Datenbank beispielsweise die folgenden beiden Muster produzieren:*

1. `WENN Einkommen ∈ [100.000; 1.000.000]`
   `DANN Kunde = ja,`
2. `WENN Geschlecht = w UND Alter ∈ [32; 55]`
   `DANN Kunde = ja.`

*Regeln in konjunktiver Normalform bestehen, grob gesagt, sowohl im WENN- als auch im DANN-Teil aus UND-Verknüpfungen von (Attribut, Operator, Wertemenge)-Tupeln. Sie dürfen – im Gegensatz zu Verfahren auf Basis der Prädikatenlogik – keine Variablen enthalten. Damit sind sie nicht in der Lage, sog. „**multirelationale Muster**“*[62] *zu repräsentieren, die sich auf mehrere Relationen beziehen, wie z.B.:*

1. `WENN verheiratet(`$X_1$`,`$X_2$`)`
   `UND Einkommen(`$X_1$`) ∈ [100.000; 1.000.000]`
   `DANN Kunde(`$X_2$`) = ja,`
2. `WENN verheiratet(`$X_1$`,`$X_2$`)`
   `UND Kunde(`$X_1$`) = ja`
   `DANN Kunde(`$X_2$`) = ja.`

Will man möglichst wenig Vorwissen über mögliche Zusammenhänge voraussetzen, so muß man sich für eine möglichst allgemeine, *mächtige Wissensrepräsentation mit universeller Approximationfähigkeit* entscheiden.

*In der Vergangenheit wurden beispielsweise folgende Modelltypen mit breiter Approximationsfähigkeit verwendet: semantische Netze*[63]*, Bayes-Netze*[64]*, possibilistische Netze*[65]*, neuronale Netze*[66]*, Neuro-*

---

60 Das Beispiel wurde entnommen aus: DZEROSKI (1996), S. 118 f.

61 Die exakte Notation von Regeln in konjunktiver Normalform wird im nächsten Abschnitt eingeführt.

62 KNOBBE/SIEBES/VAN DER WALLEN (1999), S. 378

63 Vgl. HOLSHEIMER/SIEBES (1994), S. 45.

64 Vgl. HAN/KAMBER (2001), S. 299 ff.

65 Vgl. BORGELT/KRUSE/LINDNER (1998), S. 13.

66 Vgl. LACKES/MACK (2000), S. 21 ff.

*Fuzzy-Systeme[67], Frames[68], die Prädikatenlogik erster Ordnung[69], Entscheidungsbäume[70], Modellbäume[71], Entscheidungslisten[72] oder Regeln in diskunktiver[73] oder konjunktiver Normalform[74].*

Neben den Fähigkeiten, beliebige Zusammenhänge zu approximieren sowie multirelationale Muster darzustellen, gehört zu einer mächtigen Wissensrepräsentation auch die Fähigkeit, Merkmale *verschiedener Skalentypen* zu repräsentieren, *ohne eine besondere Codierung* zu erfordern. Codierungsprozeduren, wie sie etwa bei Cluster-, Regressions-, Varianz-, Kontingenz- oder Diskriminanzanalysen regelmäßig notwendig sind, sind nicht eindeutig, und die Ergebnisse können sehr sensitiv auf verschiedene Codierungen reagieren. Damit treiben Codierungen und die notwendigen Sensitivitätsanalysen den Aufwand für die durchzuführenden Analysen in die Höhe. Wenn bei der Codierung eine Transformation des Skalenniveaus vorgenommen wird, so sind damit weitere Probleme verbunden. Denn die Verringerung des Skalenniveaus bedeutet einen Informationsverlust, und die Erhöhung des Skalenniveaus bedeutet eine „Hineininterpretation“ von Zusatzinformationen in die Daten, die aus den vorliegenden Daten selbst nicht zu rechtfertigen sind.[75]

Im folgenden wird anhand von Abbildung 2-4 eine mögliche Zuordnung von Modelltypen zu Aufgabenbereichen dargestellt. Dabei werden – wie schon in Abschnitt 2.1.3 – die Aufgaben der Prognose, Abhängigkeitsmodellierung, Zusammenfassung und Segmentierung unterschieden.

---

67 Vgl. WITTMANN (2000), S. 88 ff.

68 Vgl. HOLSHEIMER/SIEBES (1994), S. 46.

69 Vgl. HOLSHEIMER/SIEBES (1994), S. 44.

70 Vgl. HOLSHEIMER/SIEBES (1994), S. 42.

71 Vgl. QUINLAN (1992), S. 343.

72 Vgl. HOLSHEIMER/SIEBES (1994), S. 43.

73 Vgl. HOLSHEIMER/SIEBES (1994), S. 41.

74 Vgl. HOLSHEIMER/SIEBES (1994), S. 41.

75 Vgl. BACKHAUS ET AL. (2000), S. XX.

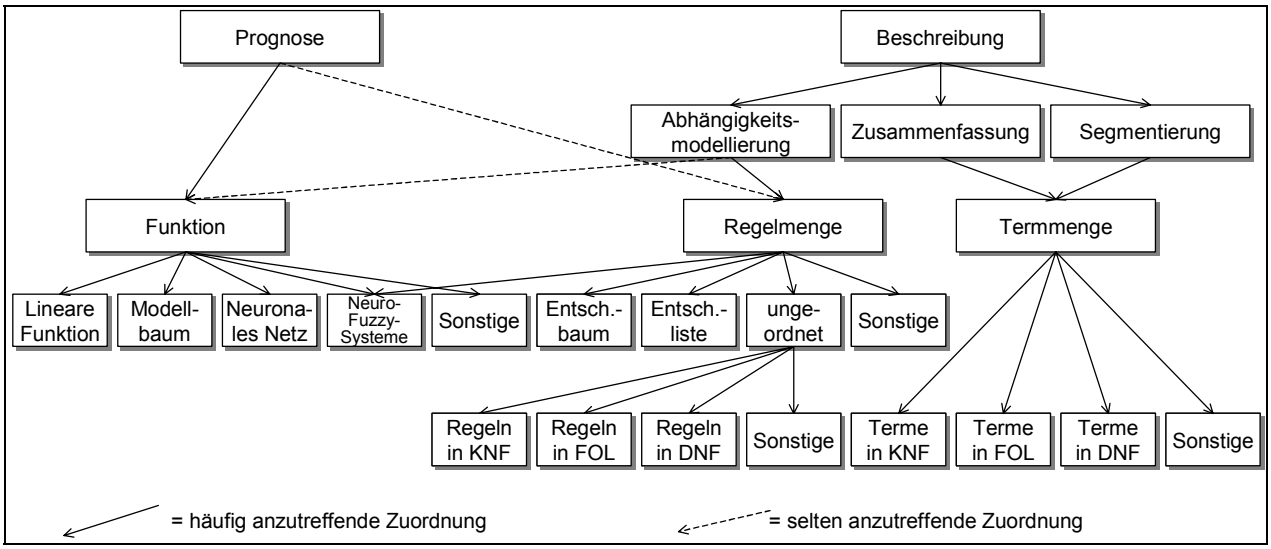



**Abbildung 2-4: Zuordnung von Aufgabenbereichen zu Modelltypen des Data Mining**

Während bei dem Aufgabenbereich der Prognose ein *möglichst exakter Prognosewert* erforderlich ist, spielt bei den eher beschreibenden Aufgabebereichen der Abhängigkeitsmodellierung, Zusammenfassung und Segmentierung die *Verständlichkeit* des Modells für den Anwender eine entscheidende Rolle. Um für den Anwender verständlich zu sein, sollte die Sprache *visuelle Darstellungsmittel benutzen* oder *nahe an die natürliche Sprache angelehnt* sein. Die visuellen Wissensrepräsentationen sind immer auf relativ wenige Dimensionen[76] begrenzt und werden daher hier nicht weiter betrachtet. Ansonsten erfüllen die Anforderung der Verständlichkeit die Wissensrepräsentation als Regelmenge (z.B. in Form von Entscheidungsbäumen, Entscheidungslisten, Neuro-Fuzzy-Systemen oder ungeordneten Mengen von Regeln) oder als Termmenge. Während in einer Regelmenge jede Regel aus zwei Termen besteht – der Prämisse und der Konklusion – entfällt diese Unterscheidung in einen WENN- und einen DANN-Teil bei reinen Termmengen. Einzelne Terme und Regeln können u.a. in konjunktiver Normalform (KNF), diskjunktiver Normalform (DNF) oder Prädikatenlogik erster Ordnung (First Order Logic, FOL) repräsentiert werden. Weiterhin ließen sich alle Typen von Termen nach der verwendeten Logik in scharfe und Fuzzy-Terme unterscheiden – dies wurde aus Platzgründen nicht mit in die Abbildung aufgenommen. Entscheidungsbäume und

[76] Auch mit ausgefeilteren visuellen Techniken, wie z.B. den parallelen Koordinaten (vgl. KEIM/KRIEGEL (1996), S. 927), lassen sich im Vergleich zum Anspruch des Data Mining relativ wenige Dimensionen darstellen.

-listen sind, wie noch gezeigt wird, spezielle Anordnungen von Regeln in KNF.[77] Auch bei diesen Repräsentationsformen sind sowohl scharfe als unscharfe Varianten denkbar.

Regelmengen können, sofern man dafür sorgt, daß sie einen eindeutigen und präzisen Output liefern, auch zur Prognose eingesetzt werden. Dies gilt insbesondere für Neuro-Fuzzy-Systeme, in denen Fuzzy-Regelmengen so auf ein neuronales Netz abgebildet werden, daß sie durch leistungsfähige neuronale Lernverfahren adaptiert werden können.[78] Im Gegensatz zu reinen neuronalen Netzen wird das Wissen hier symbolisch und nicht subsymbolisch repräsentiert und kann somit interpretiert werden. Wie noch genauer diskutiert wird, sollten prognostische und beschreibende Zielsetzungen strikt getrennt werden, so daß der Sinn solcher – mehrere Zielsetzungen vermischenden – Verfahren in Frage gestellt werden muß.

Häufiger werden zur Prognose spezielle Funktionstypen, wie z.B. lineare Funktionen, Modellbäume, neuronale Netze oder Neuro-Fuzzy-Systeme, verwendet. Diese Funktionstypen zeichnen sich nicht durchgängig durch ihre manuelle Interpretierbarkeit, sondern vielmehr durch ihre guten Prognoseeigenschaften und leistungsfähige Lernverfahren aus.

Je komplexer die Zusammenhänge sind, die durch formale Ausdrücke darstellbar sein sollen, desto ausdrucksstärker muß auch die formale Sprache sein, die zur Beschreibung der Muster verwendet wird. Doch mit der Ausdrucksfähigkeit wächst auch die *Komplexität der Algorithmen* zur Erzeugung von Ausdrücken in dieser Sprache, insbesondere auch deren Laufzeitkomplexität.[79] Hinzu kommt, daß komplexe Ausdrücke schwer interpretierbar werden können. Diese Ausführungen machen deutlich, daß die Existenz vorhandener, einfacher Algorithmen für die Wahl des Modelltyps ebenfalls eine Rolle spielt. Für jeden der genannten Modelltypen existieren spezielle Verfahren, die Modelle dieses Typs erlernen können. So kann die lineare Regression lineare Funktionen erlernen, Modellbaumverfahren erzeugen Modellbäume, neuronale Lernverfahren trainieren neuronale Netze, Baumverfahren erzeugen Entscheidungsbäume,

---

[77] Entscheidungsbäume werden in Abschnitt 2.2.2.3.2 eingeführt, Entscheidungslisten in Abschnitt 2.2.2.3.3.

[78] Vgl. WITTMANN (2000), S. 129 ff.

[79] Vgl. HOLSHEIMER/SIEBES (1994), S. 45.

KNF-Regellerner generieren Regeln in konjunktiver Normalform usw. gemäß Abbildung 2-4. Jeder Verfahrensgruppe gehören verschiedene konkrete Lösungsverfahren an.

*Beispielsweise gehört das Verfahren AQ15[80] zur Gruppe der KNF-Regellerner, ID3[81] zur Gruppe der Entscheidungsbaumverfahren und CN2[82] zur Gruppe der Entscheidungslistenverfahren.*

Jedes Verfahren benutzt eine ganz spezielle Art der Interessantheitsbewertung, die nicht so ohne weiteres austauschbar sind. Damit eignen sich diese Verfahren jeweils nur für eine ganz spezielle Aufgabenstellung.

Die Alternative zur Verwendung dieser Data-Mining-Verfahren besteht darin, ein neues Data-Mining-Verfahren mit einem eigenen Modelltyp und einer eigenen Interessantheitsbewertung zu konzipieren. Dazu bietet es sich an, eine allgemeinverwendbare Suchstrategie (auch „***Meta-Heuristik***“[83] genannt) zu benutzen.

*Zu diesen zählen beispielsweise Hill Climbing[84], Simulated Annealing[85], Threshold Accepting[86], Record-To-Record-Travel[87], der Great-Deluge-Algorithmus[88], genetische Algorithmen[89], Evolutionsstrategien[90], Scatter Search[91] und Tabu Search[92].*

Zwar hat jede Meta-Heuristik bestimmte Stärken und Schwächen – grundsätzlich aber kann jede dieser Strategien für alle Modelltypen eingesetzt werden.

Zusammengefaßt bestimmen folgende Kriterien die Auswahl einer Wissensrepräsentation:

⇨ die Annahmen des Analytikers über das Realsystem;

---

80 Vgl. MICHALSKI ET AL. (1986), S. 1041 ff.

81 Vgl. QUINLAN (1986), S. 87 ff.

82 Vgl. CLARK/NIBLETT (1989), S. 267 ff.

83 VOß/FIEDLER/GREISTORFER (2000), S. 556

84 Vgl. HOLSHEIMER/SIEBES (1994), S. 32.

85 Vgl. HOLSHEIMER/SIEBES (1994), S. 35.

86 Vgl. DUECK/SCHEUER (1990), S. 162.

87 Vgl. DUECK (1993), S. 87.

88 Vgl. DUECK (1993), S. 87.

89 Vgl. HOLSHEIMER/SIEBES (1994), S. 33 ff.

90 Vgl. RECHENBERG (1973), S. 19 ff. und SCHWEFEL (1981), S. 104 ff.

91 Vgl. GLOVER/LAGUNA (1997), S. 314 ff.

92 Vgl. GLOVER (1989), S. 192 ff.

- ⇨ die Mächtigkeit der Wissensrepräsentation (u.a.: universelle Approximationsfähigkeiten, Darstellung multirelationaler Datenmuster, Eignung für beliebige Skalentypen ohne besondere Codierung);
- ⇨ der Aufgabenbereich (u.a.: Prognosepräzision vs. Verständlichkeit der Beschreibung);
- ⇨ die Komplexität des Lösungsraums und der Lernverfahren.

#### *2.2.2.2 Die Repräsentation einzelner Datenmuster in konjunktiver Normalform*

In praxisnahen Systemen hat sich die Regeldarstellung in konjunktiver Normalform (KNF) bewährt.

*Ein Beispiel für ein Datenmuster in KNF stellt die folgende Regel dar:*
*WENN Alter ∈ [0;30] UND Einkommen ∈ [0;4000] DANN Käufer = +.*
*Abbildung 2-5 zeigt, daß durch diese Regel ein Ausschnitt aus dem Merkmalsraum beschrieben wird, der senkrecht zu den Achsen verläuft.*

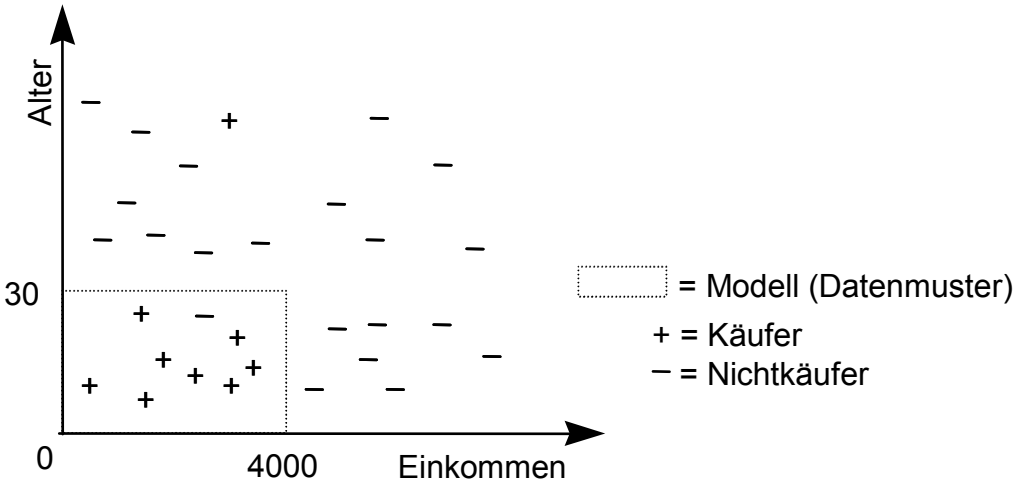



**Abbildung 2-5: Datenmuster in konjunktiver Normalform (KNF)**

Die Regeldarstellung in KNF weist folgende **Vorteile** gegenüber anderen Repräsentationsformen auf:

- ⇨ Regeln in konjunktiver Normalform sind leicht verständlich.
- ⇨ Die konjunktive Normalform ist eine Sprache von angemessener Komplexität. Einerseits können Regeln in KNF viele Typen möglicher Zusammenhänge des Realsystems approximieren. Andererseits ist der durch diese Sprache definierte Lösungsraum nicht zu groß für ein entsprechendes Lernverfahren.

⇨ Die Operationen zur Konstruktion, Bewertung und Modifikation von Regeln in konjunktiver Normalform können einfach und effizient implementiert werden.

⇨ Regelmodelle beschreiben ein Realsystem durch beliebige Merkmale. Dadurch müssen die Daten nicht unbedingt vor dem eigentlichen Data Mining noch codiert werden.

Diesen Vorteilen stehen folgende **Nachteile** gegenüber:

⇨ Multirelationale Datenmuster können mit Regelmodellen in konjunktiver Normalform nicht dargestellt werden.

⇨ Regeln in konjunktiver Normalform beschreiben, wie an Abbildung 2-5 gezeigt wurde, Ausschnitte aus dem Merkmalsraum, die senkrecht zu den Achsen verlaufen. Andere Typen von Zusammenhängen können lediglich approximiert werden.

Aufgrund der genannten Vorteile wird für diese Untersuchung die Regeldarstellung in konjunktiver Normalform ausgewählt. Der erstgenannte Nachteil kann, wie in Kapitel 5 gezeigt wird, durch einfache Erweiterungen umgangen werden.

Dieser Abschnitt führt alle zur Repräsentation einzelner Datenmuster in KNF notwendigen Begriffsdefinitionen ein. Solche Datenmuster setzen sich aus einzelnen Klauseln zusammen, so daß zunächst dieser Begriff zu definieren ist. Dabei sollen nominale und nichtnominale Klauseln differenziert werden:

**Definition 2-8: Nominale Klausel**

Gegeben sei die Menge der beobachteten Attribute, $A$, und die Menge der natürlichen Zahlen, $\mathbf{N}$. Dann ist eine *nominale Klausel,* $Kl_a$, eine Disjunktion (ODER-Verknüpfung) von Attribut-Wert-Bedingungen der Form:

$Kl_a = ((a = w_1) \vee \ldots \vee (a = w_{wmax}))$ mit

$a \in A;$

$a$ nominal;

$w_i \in dom(a)$;

$i = 1, \ldots, wmax$;

$wmax \in \mathbf{N}$.

Eine nominale Klausel, $Kl_a$, läßt sich auch schreiben als: $Kl_a = (a \in \{w_1, ..., w_{wmax}\})$. Falls das Attribut $a$ für einen Kontext unwichtig ist, wird anstelle von $Kl_a$ auch nur $Kl$ geschrieben. ◊

Überträgt man diese Definition auf kardinale oder ordinale Attribute, so ist die entsprechende Klausel wie folgt definiert:

**Definition 2-9: Nichtnominale Klausel, geordnete Klausel**

Gegeben sei eine Attributmenge, $A$. Dann ist eine *nichtnominale (geordnete) Klausel,* $Kl_a$, eine Attribut-Wert-Bedingung der Form:

$Kl_a = (a \in [w^{ug}; w^{og}])$ mit
$a \in A;$
$a$ kardinal oder ordinal;
$w^{ug} \leq w^{og}$ und
$w^{ug}, w^{og} \in dom(a)$.
Anstelle von $Kl_a$ wird auch hier $Kl$ geschrieben. ◊

Mit diesen beiden Klauseltypen läßt sich nun die Menge aller Klauseln in KNF definieren:

**Definition 2-10: Menge aller Klauseln in KNF**

$A$ sei die Menge der beobachteten Attribute und $A'$ eine Teilmenge davon, $A' \subseteq A$. Die *Menge aller bezüglich $A'$ möglichen Klauseln in konjunktiver Normalform* sei mit $Kl^{KNF}(A')$ bezeichnet und wie folgt definiert:

$Kl^{KNF}(A') := \{Kl_a \mid a \in A', Kl_a$ ist eine nominale oder nichtnominale Klausel in konjunktiver Normalform$\}$. ◊

Mit dem Begriff der Klausel kann der darauf aufsetzende Begriff des Terms definiert werden:

**Definition 2-11: Term in KNF**

$A$ sei die Menge der beobachteten Attribute, $A'$ eine Teilmenge davon, $A' \subseteq A$, und $\boldsymbol{N}$ die Menge der natürlichen Zahlen. Gegeben sei weiterhin mit $Kl^{KNF}(A')$ die Menge aller möglichen Klauseln bezüglich der Attributmenge $A' \subseteq A$. Eine Konjunktion (UND-

Verknüpfung) von Klauseln aus $Kl^{KNF}(A')$ heißt „*Term konjunktiver Normalform*“ und ist wie folgt definiert:

$Te(A') = (Kl_1 \wedge ... \wedge Kl_{Klmax});$
$Kl_1, ..., Kl_{Klmax} \in Kl^{KNF}(A');$
$Klmax \in \mathbf{N}.$
Anstelle von $Te(A')$ wird auch $Te$ geschrieben. ◊

Mit dem Begriff des Terms läßt sich nun die Menge aller Terme in KNF definieren:

**Definition 2-12: Menge aller Terme in KNF**

$A$ sei die Menge der beobachteten Attribute. Die *Menge aller bezüglich $A$ möglichen Terme in konjunktiver Normalform* sei mit $Te^{KNF}(A)$ bezeichnet und wie folgt definiert:

$Te^{KNF}(A) := \{Te(A') \mid A' \subseteq A;\ Te(A')$ ist ein Term in konjunktiver Normalform$\}$. ◊

Bei der Definition eines Datenmusters in KNF sind zwei Varianten zu unterscheiden:

1. Muster in Regelform, wobei die Prämisse, $Pr$, und die Konklusion, $Ko$, jeweils Terme in konjunktiver Normalform darstellen: `WENN` $Pr$ `DANN` $Ko$;
2. Muster, die ein (Teil-)Segment namens *Segmentname* definieren und aus einer Disjunktion von Termen in KNF, $Pr_1 \vee ... \vee Pr_{Prmax}$, bestehen: $(Segment = <Segmentname>) := Pr_1 \vee ... \vee Pr_{Prmax}$.

Muster in Regelform beschreiben gerichtete Abhängigkeiten (Folgerungen), während die zweitgenannte Musterform ungerichtete (gegenseitige) Abhängigkeiten beschreibt, wie sie für die Ergebnisse von konzeptionellen Clusterverfahren[93] typisch sind. Die Variable *Segment* stellt dabei insofern eine Besonderheit dar, daß sie nicht der Trainingsmenge entstammt (d.h. $Segment \notin A$); ihre Werte werden erst nach Ermittlung der Segmente durch den Benutzer oder durch ein intelligentes Verfahren definiert. Die Variablenwerte haben die Aufgabe, die ermittelten Segmente sinnvoll zu bezeichnen.

*Beispielsweise würde ein Muster der Form*
*(Segment = Junge_Niedrigverdiener) := ((Alter ∈ [0;30]) UND (Einkommen ∈ [0;4000]) ODER (Beruf = Auszubildender)*

[93] Vgl. zum zu dieser Verfahrensgruppe: MICHALSKI/STEPP (1983), S. 331 ff.

*aussagen, daß sich junge Niedrigverdiener durch die besagten Alters- und Einkommensintervalle oder durch ihren Beruf (Auszubildender) auszeichnen. Es kann keine jungen Niedrigverdiener mit anderen Eigenschaften geben. Diese definitorische Musterform läßt sich durch mehrere Regeln ausdrücken:*
*(Segment = Junge_Niedrigverdiener) := ((Alter ∈ [0;30]) UND (Einkommen ∈ [0;4000])) ODER (Beruf = Auszubildender)*
*⇔*
*WENN (Alter ∈ [0;30]) UND (Einkommen ∈ [0;4000]) DANN (Segment = Junge_Niedrigverdiener),*
*WENN (Beruf = Auszubildender) DANN (Segment = Junge_Niedrigverdiener),*
*WENN (Segment = Junge_Niedrigverdiener) DANN ((Alter ∈ [0;30]) UND (Einkommen ∈ [0;4000])) ODER (Beruf = Auszubildender) .*
*Auf die letztgenannte Regel (welche ohnehin nicht der konjunktiven Normalform genügt, da sie eine Disjunktion enthält) kann dann verzichtet werden, wenn man davon ausgeht, daß eine gegebene Regelmenge vollständig ist, d.h., daß es keine weitere Regel gibt, die junge Niedrigverdiener charakterisiert (wie etwa: WENN Beruf = Schüler DANN Segment = Junge_Niedrigverdiener).*

Das Beispiel soll verdeutlichen, daß sich ein Segment (unter der genannten Annahme der Vollständigkeit) auch durch Regeln in KNF beschreiben läßt. Daher genügt es, ein Datemnuster als Regel zu definieren:

**Definition 2-13: Datenmuster in KNF**

Es sei $Segment$ eine spezielle Variable, deren Domäne beliebige Zeichenketten umfaßt. Weiter sei $A$ die Menge der beobachteten Attribute, $D$ sei eine Menge von zu erklärenden, $C$ eine Menge von erklärenden Attributen mit $D \in \{\{Segment\},D'\}$, $D' \subset A$, $C \subset A$, $C \cap D = \emptyset$, $C,D \neq \emptyset$. Gegeben seien mit $Te^{KNF}(D)$ und $Te^{KNF}(C)$ die Menge aller Terme in konjunktiver Normalform bezüglich $D$ bzw. $C$. Dann bezeichnet ein „*Datenmuster in konjunktiver Normalform*“ die Folgerung:

$Pr(C) \rightarrow Ko(D)$ oder kurz: $Pr \rightarrow Ko$ mit:
$Pr = Pr(C) \in Te^{KNF}(C)$;
$Ko = Ko(D) \in Te^{KNF}(D)$. ◊

Damit läßt sich die Menge aller Datenmuster in KNF definieren:

**Definition 2-14: Menge aller Datenmuster in KNF**

Es gelten dieselben Voraussetzungen wie in der Definition zuvor. Die *Menge aller bezüglich $C$ und $D$ möglichen Datenmuster in konjunktiver Normalform* sei mit $DM^{KNF}(C,D)$ bezeichnet und wie folgt definiert:

$DM^{KNF}(C,D) :=$ $\{Pr(C) \rightarrow Ko(D) \mid Pr(C) \rightarrow Ko(D)$ ist ein Datenmuster in KNF$\} \cup$

$\{Pr(C) \rightarrow Ko(\{Segment\}) \mid Pr(C) \rightarrow Ko(\{Segment\})$ ist ein Datenmuster in KNF$\}$ ◊

Jetzt sind alle zur Bestimmung des Modelltyps notwendigen Begriffe eingeführt, so daß der Begriff des Lösungsraums als Menge aller – unter den getroffenen Voraussetzungen – möglichen Modelle des bestimmten Typs definiert werden kann:

**Definition 2-15: Lösungsraum von Datenmuster-Mengen in KNF**

Es gelten dieselben Voraussetzungen wie in den Definition zuvor. Dann stellt der *Lösungsraum von Datenmuster-Mengen in konjunktiver Normalform bezüglich* $C$ *und* $D$, $L$, einen Ausschnitt aus der Menge aller möglichen Datenmuster-Mengen dar:

$L \subseteq Pot(DM^{KNF}(C,D))$. ◊

Datenmuster und Terme in konjunktiver Normalform sowie Klauseln stellen rein sprachliche Ausdrücke dar, die noch nichts über deren Erfülltheit für bestimmte Datenobjekte aussagen. Die Erfülltheit wird durch die Definition eines Konzeptes in das definitorische Gerüst aufgenommen:

**Definition 2-16: Konzept**[94]

Gegeben sei eine Grundgesamtheit von Objekten, $O$, und ein Term in konjunktiver Normalform, $Te$. Dann stellt ein *Konzept,* $c_{Te}$, eine Abbildung dar, die jedem möglichen Datenobjekt, $o \in O$, einen scharfen Erfülltheitsgrad von $0$ (Term $Te$ nicht erfüllt) oder $1$ (Term $Te$ erfüllt) zuordnet:

$c_{Te}$: $O \rightarrow \{0,1\}$;

$o \rightarrow c_{Te}(o)$. ◊

Jetzt benötigt man noch Vorschriften, die besagen, wann ein Term, $Te$, durch ein Objekt, $o$, erfüllt ist und wann nicht.

[94] Vgl. KEARNS/VAZIRANI (1994), S. 8.

**Definition 2-17: Erfülltheit eines Terms durch ein Objekt**

$Te = (Kl_1 \wedge ... \wedge Kl_{Klmax})$ sei ein Term in konjunktiver Normalform mit den Klauseln $Kl_1, ..., Kl_{Klmax}$. Weiterhin sei $o$ ein Objekt aus der Grundgesamtheit, $O$. Man sagt genau dann, *Objekt $o$ erfülle Term $Te$*, wenn $o$ alle Klauseln erfüllt, d.h. wenn gilt:

$\forall i = 1, ..., Klmax$: $o$ erfüllt $Kl_i$. ◊

Hier ist wieder zwischen nominalen und nichtnominalen Klauseln zu unterscheiden:

**Definition 2-18: Erfülltheit einer nichtnominalen Klausel durch ein Objekt**

$Kl = (a \in [w^{ug}; w^{og}])$ sei eine nichtnominale Klausel mit dem kardinalen oder ordinalen Attribut $a$ und den Intervallgrenzen $w^{ug}$ und $w^{og}$. Weiterhin sei $o$ ein Objekt aus der Grundgesamtheit, $O$, mit der Ausprägung $a(o)$ bezüglich des Attributes $a$. Man sagt genau dann, *Objekt $o$ erfülle die nichtnominale Klausel $Kl$*, wenn gilt:

$w^{ug} \leq a(o) \leq w^{og}$. ◊

**Definition 2-19: Erfülltheit einer nominalen Klausel durch ein Objekt**

$Kl = (a \in \{ w_1, ..., w_{wmax} \})$ sei eine nominale Klausel mit dem nominalen Attribut $a$ und den Werten $w_1, ..., w_{wmax}$. Weiterhin sei $o$ ein Objekt aus der Grundgesamtheit, $O$, mit der Ausprägung $a(o)$ bezüglich des Attributes $a$. Man sagt genau dann, Objekt $o$ *erfülle die nominale Klausel* $Kl$, wenn gilt:

$a(o) \in \{ w_1, ..., w_{wmax} \}$. ◊

Nun kann für ein Datenmuster, $s_{O'} = (Pr \rightarrow Ko)$, auch der bisher übergangene Zusammenhang zwischen der Beschreibung, $(Pr \rightarrow Ko)$, und den beschriebenen Daten, $O'$, formalisiert werden. Es gilt:

$O' = \{o \in O^T \mid c_{Pr}(o) = 1\}$.

Läßt man nun als Modelle beliebige Regelmengen aus dem Lösungsraum $L$ zu, so können die einzelnen Regeln eines Modells *widersprüchliche Aussagen* darstellen. Dieses Problem spielt weiter unten bei der Auswahl einer geeigneten Struktur für Regelmengen eine wichtige Rolle, so daß es im folgenden genauer beleuchtet werden soll.

*Zur Verdeutlichung der Problematik widersprüchlicher Aussagen betrachte man die folgende Regelmenge:*

```
WENN Einkommen = mittel UND Alter = hoch DANN Käufer von A = ja.
```

*WENN Beruf = Beamter UND Region = West DANN Käufer von A = nein.*
*Diese Regelmenge führt zu einem Widerspruch, falls für einen aus der Region West stammenden Beamten hohen Alters und mit mittlerem Einkommen entschieden werden soll, ob er zu den Käufern von A zählt oder nicht.*

Allgemein läßt sich ein Widerspruch wie folgt definieren:

**Definition 2-20: Widerspruch**

$L$ sei ein Lösungsraum von Datenmuster-Mengen in KNF, $O$ die Grundgesamtheit an Datenobjekten. Zwei Aussagen aus $L$, $(Pr \rightarrow Ko)$, $(Pr' \rightarrow Ko') \in L$ seien genau dann widersprüchlich, wenn es Objekte gibt, die beide Prämissen, aber maximal eine Konklusion erfüllen:

$$\exists o \in O: (c_{Pr}(o) = c_{Pr'}(o) = 1) \wedge (c_{Ko}(o) = 1) \Rightarrow (c_{Ko'}(o) = 0) \wedge (c_{Ko'}(o) = 1) \Rightarrow (c_{Ko}(o) = 0). \quad \Diamond$$

Grundsätzlich gibt es zwei Möglichkeiten, an das Problem widersprüchlicher Aussagen heranzugehen:

- ⇨ Man sorgt dafür, daß widersprüchliche Aussagen erst gar nicht generiert werden.
- ⇨ Man stellt eine Verarbeitungsvorschrift bereit, die Widersprüche auflösen kann und für ein gegebenes Datenobjekt zu einer eindeutigen Aussage kommt. Eine solche Verarbeitungsvorschrift sei hier als „*Akkumulation*“ bezeichnet:

**Definition 2-21: Akkumulation**

Gegeben sei mit $L$ ein Lösungsraum, eine Grundgesamtheit, $O$, sowie eine Attributmenge, $D \subset A$, mit der Domäne $dom(D)$, welche die möglichen Outputs darstellt, die durch ein Modell erzeugt werden können. Dann bezeichnet man als „Akkumulation“ eine Funktion, die einem Objekt, $o$, bezüglich einer Datenmuster-Menge, $M_{O^T}$, einen eindeutigen Output aus $dom(D)$ zuordnet:

$$\begin{aligned} Akkumulation: \quad & L \times O \quad \rightarrow dom(D), \\ & (M_{O^T}, o) \quad \rightarrow Akkumulation(M_{O^T}, o). \end{aligned} \quad \Diamond$$

Bestehen potentiell in einer Datenmuster-Menge, $M_{O^T}$, widersprüchliche Konklusionen, so erhält man erst zusammen mit der Akkumulationsfunktion das gewünschte Modell, das die deterministischen Wirkungszusammenhänge aus dem betrachteten Realsystem approximiert. Letzteres wird hier im Gegensatz zur Datenmuster-Menge $M_{O^T}$ mit

$M^\bullet$ symbolisiert, wobei das Symbol $\bullet$ später durch konkrete Modelltyp-Bezeichner ersetzt wird. Die Akkumulationsfunktion bleibt während der Lösungssuche konstant.

*Als einfache Akkumulationsfunktion für eine Regelmenge in KNF mag beispielsweise die Vorschrift dienen, auf ein Datenobjekt nur die jeweils vertrauenswürdigste[95] Regel anzuwenden. Komplexere Verarbeitungsvorschriften führen die unterschiedlichen Regeln durch Gewichtung mit dem in die jeweilige Regel gesetzten Vertrauen zu einer Aussage zusammen.[96]*

$M^\bullet$ wird auch als „*funktionales Data-Mining-Modell*" bezeichnet, da es eine eindeutige Abbildung von Planungsobjekten auf die möglichen Outputs darstellt:

**Definition 2-22: Funktionales Data-Mining-Modell**
Gegeben sei mit $O$ eine Grundgesamtheit, mit $M_{O^T}$ eine Datenmuster-Menge, mit $dom(D)$ eine Menge möglicher Outputs und mit $Akkumulation$ eine Akkumulationsfunktion gemäß Definition 2-21. Dann ist ein *funktionales Data-Mining-Modell*, $M^\bullet$, definiert als Funktion, die jedem Objekt, $o \in O$, einen eindeutigen Output, $M^\bullet(o) \in dom(D)$, zuordnet:

$M^\bullet$: $O \rightarrow dom(D)$,
$o \rightarrow M^\bullet(o)$ mit $M^\bullet(o) := Akkumulation(M_{O^T}, o)$. ◊

### *2.2.2.3 Die Struktur von Datenmuster-Mengen*

Nachdem im Abschnitt zuvor mit den Regeln in konjunktiver Normalform eine Repräsentationsform für einzelne Datenmuster eingeführt wurde, wird im folgenden die Frage behandelt, wie eine Menge von Datenmustern organisiert werden kann. Die einfachste Möglichkeit besteht darin, keine besondere Strukturierung vorzunehmen und die Datenmuster als ungeordnete Liste zu repräsentieren (vgl. Abschnitt 2.2.2.3.1). Die Abschnitte 2.2.2.3.2, 2.2.2.3.3 und 2.2.2.3.4 stellen mit den Entscheidungsbäumen, den Entscheidungslisten und den Rough-Sets drei höher strukturierte Organisationsformen für eine Menge von Regeln in KNF vor.

---

[95] Als Maß für das Vertrauen in einer Regel kann die Sicherheit gemäß Definition 2-52 verwendet werden.

[96] Ein Beispiel für eine solche Verarbeitungsvorschrift ist die typischerweise in Fuzzy-Controllern realisierte Fuzzy-Akkumulation einschließlich der anschließenden Defuzzyfizierung (vgl. ZIMMERMANN (1993), S. 98 ff.).

#### 2.2.2.3.1 Die Strukturierung von Datenmustern als ungeordnete Liste

Die einfachste Möglichkeit, eine Menge von Datenmustern zu organisieren, bietet die Organisationsform einer ungeordneten Liste, welche wie folgt definiert werden kann:

**Definition 2-23: Ungeordnete Liste von Datenmustern**

Es sei $Segment$ eine spezielle Variable, deren Domäne beliebige Zeichenketten umfaßt. Weiter sei $A$ die Menge der beobachteten Attribute, $D$ sei eine Menge von zu erklärenden, $C$ eine Menge von erklärenden Attributen mit $D \in \{\{Segment\}, D'\}$, $D' \subset A$, $C \subset A$, $C \cap D = \emptyset$, $C, D \neq \emptyset$. Gegeben seien mit $Te^{KNF}(D)$ und $Te^{KNF}(C)$ die Menge aller Terme in konjunktiver Normalform bezüglich $D$ bzw. $C$ und mit $\mathbf{N}$ die Menge der natürlichen Zahlen. Dann ist eine *ungeordnete Liste,* $M_{O^T}$, wie folgt definiert:

$M_{O^T} = \{(Pr_1 \rightarrow Ko_1), ..., (Pr_M \rightarrow Ko_M)\}$ mit

$Pr_j \in Te^{KNF}(C)$;

$Ko_j \in Te^{KNF}(D)$;

$j \in \{1, ..., M\}$;

$M \in \mathbf{N}$. ◊

Dem **Vorteil** der *einfachen Generierung* einer solchen Liste stehen folgende **Nachteile** gegenüber:

- ⇨ Falls die Anzahl der generierten Datenmuster, $M$, sehr groß ist, ist die ungeordnete Liste sehr *unübersichtlich*. Der Betrachter verliert den Überblick, welche Datensätze durch ein Datenmuster, durch mehrere Datenmuster oder durch überhaupt kein Datenmuster abgedeckt werden.
- ⇨ Ungeordnete Listen von Datenmustern können *widersprüchliche Aussagen* enthalten.

#### 2.2.2.3.2 Die Strukturierung von Datenmustern als Entscheidungsbaum

Eine bekannte Möglichkeit, eine Menge von Regeln in KNF zu organisieren, bieten die Entscheidungsbäume. In der Literatur haben sich vielfältige Erscheinungsformen von Entscheidungsbäumen herausgebildet, welche sich in ihrer Ausdrucksstärke und Komplexität unterscheiden. In ihrer ursprünglichen, ausdrucksschwächsten und einfachsten

Form wurden sie von QUINLAN[97] entwickelt. Diese Entscheidungsbäume zeichnen sich dadurch aus, daß sie als Beschriftung der inneren Knoten nur diskrete Attribute, als Beschrftung der Blätter nur diskrete Klassen und als Beschriftung der Kanten jeweils nur einen diskreten Attributwert zulassen. Außerdem müssen alle von einem inneren Knoten ausgehenden Kanten disjunkte Werte aufweisen, welche zusammen den gesamten Wertebereich des entsprechenden Attributs abdecken. Diese Form der Entscheidungsbäume soll im folgenden genau definiert werden. Komplexere Formen werden hier nicht benötigt, da Entscheidungsbäume später[98] nicht zur Konzeption eines Modelltyps herangezogen werden.

**Definition 2-24: Baum**[99]

Gegeben sei ein Paar, $G = (V,E)$, bestehend aus einer nichtleeren Menge, $V$, und einer nichtleeren Menge, $E \subseteq V \times V$, von geordneten Paaren aus $V \times V$. Die Elemente von $V$ heißen „***Knoten***“, die Elemente von $E$ heißen „***Kanten***“. Für einen Knoten, $v \in V$, heißt $V^P(v) := \{v' \mid v' \in V; \exists e \in E: e=(v',v)\}$ „***Menge der Vorgängerknoten***“ und $V^S(v) := \{v' \mid v' \in V; \exists e \in E: e=(v,v')\}$ „***Menge der Nachfolgerknoten*** von $v$“. Es gebe einen ausgezeichneten Knoten, $r \in V$, ohne Vorgänger, d.h.: $V^P(r)=\emptyset$. Dieser Knoten wird als „***Wurzel***“ bezeichnet. Alle anderen Knoten haben genau einen Vorgänger, d.h.: $\forall v \in V-r: |V^P(v)|=1$. Es bezeichnen $LV := \{v \in V \mid V^S(v)=\emptyset\}$ die „***Menge der Blätter***“ und $IV := V-LV$ die „***Menge aller inneren Knoten***“. Dann heißt $\vec{G}$ *„gerichteter Baum“*. ◊

Sowohl die Knoten als auch die Kanten von Entscheidungsbäumen werden beschriftet. Hierzu dienen die folgenden beiden Definitionen:

**Definition 2-25: Knotenbeschriftung**

Es gelten dieselben Voraussetzungen wie in Definition 2-23. Sei weiterhin $G = (V,E)$ ein Baum und $dom(D)$ eine endliche Menge von Klassen. Dann liefert die Abbildung:

$$\textit{Knotenbeschriftung}: \quad V \rightarrow D \cup C,$$

$$\textit{Knotenbeschriftung}(v) := \begin{cases} \textit{Attribut}(v), v \in IV \\ \textit{Klasse}(v), v \in LV \end{cases} \text{ mit}$$

---

[97] Vgl. QUINLAN (1986), S.86.

[98] Die Konzeption des Modelltyps erfolgt in Abschnitt 5.2.

[99] Vgl. NEUMANN/MORLOCK (1993), S. 183f.

$$\begin{aligned} &\textit{Attribut}: && IV \rightarrow C, \\ & && iv \rightarrow Attribut(iv), \\ &\textit{Klasse}: && LV \rightarrow dom(D), \\ & && l \rightarrow Klasse(l) \end{aligned}$$

die *Knotenbeschriftung* des Knotens $v \in V$. Ist $v$ ein innerer Knoten, d.h. $v \in IV$, so wird dem Knoten ein erklärendes Attribut, $a \in C$, zugeordnet. Ist $v$ ein Blatt, d.h. $v \in LV$, so wird ihm eine Klasse zugewiesen. ◊

**Definition 2-26: Kantenbeschriftung**

Es gelten dieselben Voraussetzungen wie in der vorstehenden Definition. Dann liefert die Abbildung:

$$\begin{aligned} &\textit{Kantenbeschriftung}: && E \rightarrow dom(a), \\ & && e \rightarrow Kantenbeschriftung(e) \text{ mit} \\ & && a = Knotenbeschriftung(v_1), \\ & && dom(a) = \{w_1, \ldots, w_{wmax}\}, \\ & && e = (v_1, v_2), \\ & && v_1, v_2 \in V \end{aligned}$$

die *Kantenbeschriftung* einer Kante, $e \in E$. ◊

Mit den eingeführten Begriffen kann nun ein Entscheidungsbaum definiert werden:

**Definition 2-27: Entscheidungsbaum**

Ein *Entscheidungsbaum*, $EB$, ist gegeben durch das 3-Tupel $EB = (G, Knotenbeschriftung, Kantenbeschriftung)$ mit

| | |
|---|---|
| $G$ | ein Baum gemäß Definition 2-24, |
| $Knotenbeschriftung$ | die Beschriftung der Knoten gemäß Definition 2-25, |
| $Kantenbeschriftung$ | die Beschriftung der Kanten gemäß Definition 2-26, |

wenn für alle inneren Knoten, $iv \in IV$, gilt:

$$\forall v_1, v_2 \in V^S(iv),\ v_1 \neq v_2:\ Kantenbeschriftung(iv, v_1) \neq Kantenbeschriftung(iv, v_2);$$

$$\left\{ Kantenbeschriftung(iv, v) \;\middle|\; v \in V^S(iv) \right\} = dom(a) \text{ mit } a = Knotenbeschriftung(iv).$$

D.h. alle von einem inneren Knoten, $iv$, ausgehenden Kanten besitzen disjunkte Wertebereiche und decken zusammen den gesamten Wertebereich des Attributs, $a$, ab, mit

dem der Knoten $iv$ beschriftet ist. In den Blättern müssen nicht alle Werte der zu erklärenden Größe auftreten.

$v^P(v)$ sei der eindeutige Vorgänger eines Knotens, $v$, d.h.: $v^P(v) \in V^P(v)$, $v \in V - \{r\}$. Der $i$-te Pfad zwischen dem $i$-ten Blatt, $l_i$, und der Wurzel, $r$, verbindet die folgenden Knoten:

$l_i, v^P(l_i), v^P(v^P(l_i)), ..., r.$

Zur Vereinfachung der Schreibweise ohne den Index $i$ sei $v_{Prmax+1} := l_i$, $v_{Prmax} := v^P(l_i)$, $v_{Prmax-1} := v^P(v^P(l_i)), ..., v_1 := r.$ Diese Knotenfolge stellt zusammen mit den entsprechenden Kanten eine Regel, $(Pr \rightarrow Ko)$, dar, für die gilt:

$$Ko = (Entscheidung = Knotenbeschriftung(v_{Prmax+1}));$$

$$Pr = Kl^{a_1} \wedge ... \wedge Kl^{a_{Prmax}};$$

$$Kl^{a_j} = (a_j = w_j);$$

$$a_j = Knotenbeschriftung(v_j);$$

$$w_j = Kantenbeschriftung(v_j, v_{j+1});$$

$$j = 1, ..., Prmax;$$

$$Prmax \in \mathbf{N}. \qquad \Diamond$$

*Abbildung 2-6 zeigt ein Beispiel für einen Entscheidungsbaum, mit dessen Hilfe aus gegebenen Merkmalen für einen Kunden bestimmt werden kann, welcher Telefontyp für diesen Kunden der attraktivste ist. Der Pfad von der Wurzel bis zu dem am weitesten rechts eingezeichneten Blatt beispielsweise verbindet die Knoten:* `Nutzung des WWW, Altersgruppe, Einfaches Gerät.` *Diese Knotenfolge liest sich zusammen mit den entsprechenden Kanten wie folgt als Regel:*

```
WENN Nutzung des WWW = Ja
UND Altersgruppe = 45 und älter
DANN Entscheidung = Einfaches Gerät.
```

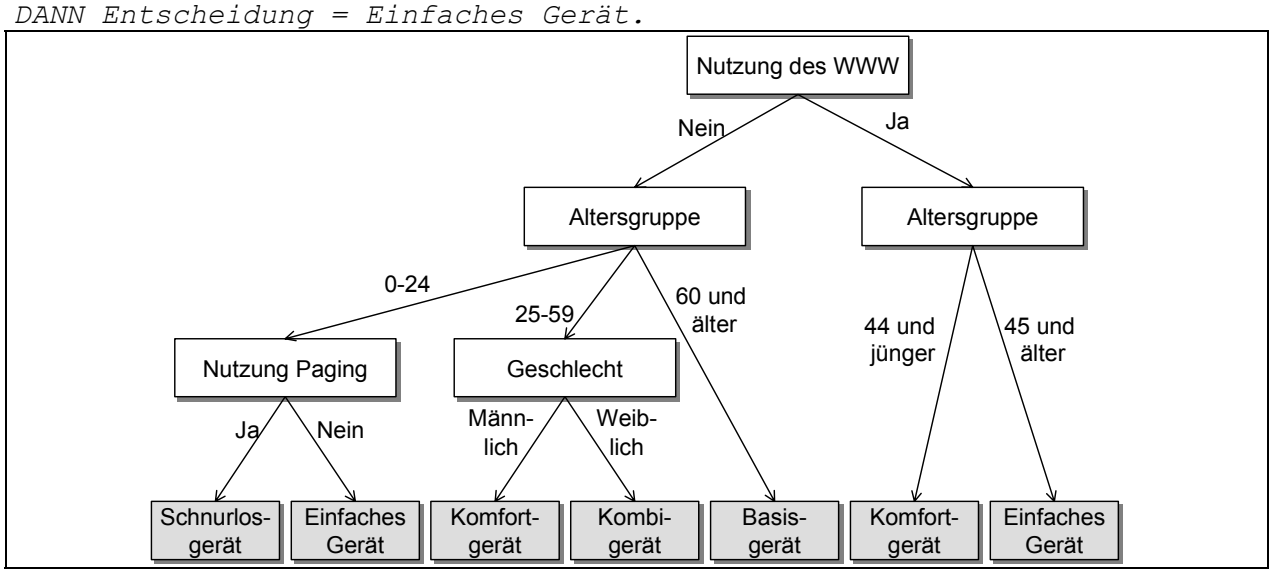



***Abbildung 2-6: Beispiel für einen Entscheidungsbaum***

**Vorteilhaft** gegenüber ungeordneten Regellisten sind Entscheidungsbäume bezüglich folgender Aspekte:

⇨ Die Organisation von Datenmustern als Entscheidungsbaum vermeidet widersprüchliche Aussagen für gegebene Datenobjekte, indem jedem Datenobjekt genau ein Pfad von der Wurzel bis zu einem Blatt und damit eine eindeutige Entscheidung zugeordnet wird. Somit wird keine Akkumulationsfunktion gemäß Definition 2-21 benötigt.

⇨ Außerdem sind Entscheidungsbäume übersichtlicher als ungeordnete Listen von Datenmustern. Allerdings können auch Entscheidungsbäume so groß werden, daß sie kaum noch ohne technische Hilfen zu überblicken sind.

Entscheidungsbäume sind ebenso einfach zu generieren wie ungeordnete Listen von Regeln. Sie weisen jedoch auch **Nachteile** gegenüber ungeordneten Regelmengen auf:

⇨ Entscheidungsbäume sind weniger ausdrucksfähig als ungeordnete Regelmengen. Jeder Entscheidungsbaum kann als Regelmenge, aber nicht jede Regelmenge als Entscheidungsbaum dargestellt werden. Das Attribut, welches die Wurzel des Baumes beschriftet, muß in allen durch den Baum repräsentierten Regeln vorkommen. Dies gilt nicht nur für den gesamten Entscheidungsbaum, sondern auch für seine Teilbäume.

*Beispielsweise muß in Abbildung 2-6 u.a. jede Regel das Attribut „`Nutzung des WWW`“ umfassen. Nicht durch Entscheidungsbäume dargestellt werden können Mengen von Regeln mit verschiedenen Attributen in den Prämissen, z.B. die folgende Regelmenge:*

```
WENN Ausbildung = Schüler DANN Entscheidung = Einfaches Gerät.
WENN Nutzung des WWW = Ja UND Altersgruppe = 44 und jünger DANN Ent-
scheidung = Komfortgerät.
```

⇨ Ein Entscheidungsbaum kann ein Datenobjekt nur dann klassifizieren, wenn alle Ausprägungen der Attribute des Datenobjektes bekannt sind, welche auf mindestens einem Pfad des Entscheidungsbaums auftreten.

⇨ Für jedes im Entscheidungsbaum vorkommende Attribut muß von einem inneren Knoten, $iv$, für jede mögliche Ausprägung des Attributes genau eine gerichtete Kante zu einem Nachfolger verzweigen. Dies macht nur dann Sinn, wenn die Wertemenge $dom(iv)$ eine geringe Zahl von Ausprägungen umfaßt. Daher müssen stetige Wertemengen in einige wenige Intervalle zerlegt werden, die dann als diskrete

Werte aufgefaßt werden. Gruppierungen mehrerer Attributwerte an einer Kante sind nicht zugelassen. Diese Einschränkungen sind durch die klassischen Algorithmen zur Generierung von Entscheidungsbäumen bedingt und sind zur reinen Wissensrepräsentation nicht unbedingt erforderlich. Durch Erweiterungen der Wissensrepräsentation könnten einige Einschränkungen aufgehoben werden – die Algorithmen werden dann allerdings wesentlich komplexer.[100]

#### 2.2.2.3.3 Die Strukturierung von Datenmustern als Entscheidungsliste

Eine weitere Möglichkeit, eine Menge von Regeln in KNF zu organisieren, bieten die sog. „Entscheidungslisten". Eine Entscheidungsliste ist wie folgt definiert:[101]

**Definition 2-28: Entscheidungsliste**

Es gelten dieselben Voraussetzungen wie in Definition 2-23. Für die Menge der zu erklärenden Variablen gelte: $D = \{D_1, ..., D_{admax}\}$. Weiter sei $dom(D_i) = \{d_{i,1}, ..., d_{i,dmax(i)}\}$ für $i = 1, ..., admax$ und $\mathbf{N}$ die Menge der natürlichen Zahlen. Eine ***Entscheidungsliste,*** $M_{O^T}$, ist eine geordnete Liste von $M$ Regeln:

$M_{O^T} = ((Pr_1 \rightarrow Ko_1), ..., (Pr_M \rightarrow Ko_M))$ mit

$Pr_j \in Te^{KNF}(C);$

$Ko_j \in Te^{KNF}(D);$

$j = 1, ..., M;$ ($j$: Nummer der Regel)

$M \in \mathbf{N};$

$Ko_j = (D_1 = d_{1,k(j)}) \wedge ... \wedge (D_{admax} = d_{admax,k(j)});$

$k(j) \in \{1, ..., dmax(i)\};$ ($k(j)$: Nummer des Wertes in der $j$-ten Konklusion)

$i = 1, ..., admax;$ ($i$: Nummer des zu erklärenden Attributes)

wobei die letzte Regel, $(Pr_M \rightarrow Ko_M)$, eine Default-Regel mit der Prämisse $Pr_M = wahr$ ist. Zur Entscheidungsliste gehören $admax$ Akkumulationsfunktionen, die einem gegebenen Datenobjekt, $o \in O$, einen eindeutigen Output zuordnen:

$$\begin{aligned} Akkumulation_i: \; & L \times O \rightarrow dom(D_i), \\ & (M_{O^T}, o) \rightarrow Akkumulation_i(M_{O^T}, o) := d_{i,k(j^*)} \end{aligned}$$

---

[100] Vgl. beispielsweise QUINLAN (1993), S. 25 f. zu Entscheidungsbäumen, die kardinale Wertebereiche durch mehrere binäre Splits aufteilen, oder QUINLAN (1993), S. 63 ff. zur Gruppierung von diskreten Attributwerten.

[101] Vgl. RIVEST (1987), S. 234.

mit $j^* = \min_{j \in \{1,\ldots,M \mid c_{Pr_j}(o)=1\}} j$ und $i = 1, \ldots admax$. ◊

Einem Datenobjekt, $o \in O$, wird derjenige Output, $c_{i,k(j^*)}$, zugeordnet, dessen Prämisse, $Pr_{j^*}$, die erste erfüllte Prämisse in der Entscheidungsliste ist. Ob eine Prämisse, $Pr_j$, erfüllt ist, wird durch ein Konzept, $c_{Pr_j}$, gemäß Definition 2-16 bestimmt. Damit wird erreicht, daß einem Datenobjekt bei Erfülltheit mehrerer Prämissen keine widersprüchlichen Outputs zugeordnet werden.

Eine Entscheidungsliste kann wie folgt interpretiert werden:

WENN $Pr_1$ DANN $Ko_1$
SONST WENN $Pr_2$ DANN $Ko_2$
SONST ...
SONST WENN $Pr_{M-1}$ DANN $Ko_{M-1}$
SONST $Ko_M$.

*Beispielsweise könnte der Baum in Abbildung 2-5 wie folgt als Entscheidungsliste dargestellt werden:*

```
WENN    Nutzung des WWW = Ja UND Altersgruppe = 45 und älter
DANN    Entscheidung = Einfaches Gerät
SONST   WENN    Nutzung des WWW = Ja
        DANN    Entscheidung = Komfortgerät
        SONST   WENN    Altersgruppe = 60 und älter
                DANN    Entscheidung = Basisgerät
                SONST   WENN    Altersgruppe = 25-59
                        UND     Geschlecht = Weiblich
                        DANN    Entscheidung = Kombigerät
                        SONST   WENN Altersgruppe = 25-59
                                DANN Entscheidung = Komfortgerät
                                ...
```

Es existieren mehrere Möglichkeiten, einen Baum als Entscheidungsliste darzustellen. Ob die im Beispiel gewählte Möglichkeit sinnvoll ist, hängt von der Allgemeinheit der Regeln ab. Die Regeln innerhalb einer Entscheidungsliste sind so angeordnet, daß Regeln, die Einzelfälle (Ausnahmen von den allgemeineren Regeln) beschreiben, am Anfang und allgemeingültige Regeln am Ende der Liste stehen. Die allgemeinste Regel, deren Prämisse „$Pr_M = wahr$“ immer erfüllt ist, steht an der letzten Position in der Liste, wodurch gewährleistet ist, daß für jedes Datenobjekt, $o \in O$, eine Entscheidung getroffen werden kann.

Eine einfache Heuristik zur Generierung von Entscheidungslisten könnte so vorgehen, daß zunächst nach einer Regel gesucht wird, welche möglichst viele Objekte abdeckt

(unter gewissen Nebenbedingungen, die die Korrektheit der Regel sicherstellen). Diese würde dann den Anfang der Entscheidungsliste bilden. Als nächstes würde unter den noch nicht abgedeckten Objekten wieder diejenige Regel gesucht, welche die meisten Objekte abdeckt, und an den Anfang der Liste plaziert usw. Sobald keine Regel mehr gefunden werden kann, welche genügend Objekte abdeckt, wird die beschriebene Schleife abgebrochen und die Default-Regel bestimmt. Für die Konklusion der Default-Regel, $Ko_M = (D_1=d_{1,M}) \wedge ... \wedge (D_{admax}=d_{admax,M})$, wählt man jeweils (für $i = 1, ..., admax$) denjenigen Output $d_{i,k(M)}$, der unter den Objekten, die durch $Pr_M$ und nicht durch $Pr_1 \vee ... \vee Pr_{M-1}$ abgedeckt werden, am häufigsten vorkommt:

$$\max_{k(M)=1,...,dmax(i)} | O^T [Pr_M \wedge \neg(Pr_1 \vee ... \vee Pr_{M-1}) \wedge (D_i = d_{i,k(M)})] | .$$

Diese Ausführungen sollen verdeutlichen, daß Entscheidungslisten ähnlich einfach zu generieren sind wie Entscheidungsbäume oder ungeordnete Regellisten. Ihre Ausdrucksstärke übersteigt die der Entscheidungsbäume, bei denen das Attribut, welches die Wurzel eines (Teil-)Baumes beschriftet, in allen durch den (Teil-)Baum repräsentierten Regeln vorkommen muß. Entscheidungslisten können beliebige, nicht widersprüchliche Regelmengen darstellen, sind aber wesentlich unübersichtlicher als Entscheidungsbäume.

**Vorteilhaft** gegenüber ungeordneten Regellisten ist zum einen, daß widersprüchliche Aussagen durch die Strukturierung der Regeln vermieden werden. Zum anderen erhält man durch die Ordnung der Regeln Zusatzinformation über die Allgemeinheit der Aussagen. **Nachteilig** daran ist, daß eine Regel, $(Pr_i \rightarrow Ko_i)$, nicht für sich allein interpretiert werden darf, sondern nur als SONST-Teil der vorgelagerten Regeln, $(Pr_1 \rightarrow Ko_1), ..., (Pr_{i-1} \rightarrow Ko_{i-1})$.[102] Dieser Nachteil kann umgangen werden, indem man die Entscheidungsliste auflöst und als ungeordnete Regelliste schreibt. Damit würde aus der $i$-ten Regel der Entscheidungsliste die vollständige Regel: $(\neg Pr_1 \wedge ... \wedge \neg Pr_{i-1}) \wedge Pr_i \rightarrow Ko_i$. Auf diese Weise würde eine Entscheidungsliste allerdings noch unübersichtlicher als sie ohnehin schon ist.

[102] Vgl. HOLSHEIMER/SIEBES (1994), S. 43.

#### 2.2.2.3.4 Die Strukturierung von Datenmustern als Rough-Set-Regelmenge

Sog. „*Rough Sets*“ (engl.: „Grobe Mengen“) bilden die Grundlage für die Organisation von Regeln in Tabellenform. Es existiert eine eigene Theorie der Rough Sets, die weit über die hier benötigten Elemente hinausgeht. In diesem Abschnitt werden nur Begriffe eingeführt, die in der folgenden Arbeit verwendet werden.[103] Die Einführung der Begriffe orientiert sich an folgendem Beispiel:

*Gegeben sei die Trainingsmenge $O^T = \{o_1, o_2, o_3, o_4, o_5, o_6, o_7, o_8\}$ mit der Attributmenge $A = \{Artikelgruppe, Kundengruppe, Region, Umsatz\}$ gemäß Tabelle 2-3.*[104]

| Objekt | Artikelgruppe | Kundengruppe | Region | Umsatz |
|---|---|---|---|---|
| $o_1$ | Tourenräder | Fachhandel | Süd | niedrig |
| $o_2$ | Cityräder | Discounter | West | hoch |
| $o_3$ | Mountainbikes | Fachhandel | Nord | mittel |
| $o_4$ | Tourenräder | Fachhandel | Nord | niedrig |
| $o_5$ | Tourenräder | Fachhandel | Süd | hoch |
| $o_6$ | Mountainbikes | Kaufhaus | Nord | mittel |
| $o_7$ | Mountainbikes | Discounter | West | mittel |
| $o_8$ | Cityräder | Discounter | West | hoch |

***Tabelle 2-3: Beispiel für eine Trainingsmenge, $O^T$***

Der Rough-Set-Ansatz verfolgt das Ziel, die Anzahl der Attribute derart zu reduzieren, daß viele Objekte durch dieselbe Wertekombination beschrieben werden können. Dies ist tendenziell eher der Fall, wenn erstens nur wenige Attribute betrachtet werden und wenn zweitens die Domänen der Attribute nur wenige diskrete Werte umfassen.

*Betrachet man beispielsweise nur die Attribute $Kundengruppe$ und $Region$, so werden die Objekte innerhalb der Mengen $\{o_1,o_5\}$, $\{o_2,o_7,o_8\}$, $\{o_3,o_4\}$, $\{o_6\}$ jeweils durch dieselbe Wertekombination beschrieben. Die Objekte innerhalb dieser Mengen sind bezüglich der Attribute $Kundengruppe$ und $Region$ ununterscheidbar.*

Ob zwei Objekte bezüglich der betrachteten Attribute unterscheidbar sind, ermittelt man nach folgender Vorschrift:

**Definition 2-29: Ununterscheidbarkeitsrelation**

Gegeben sei eine Objektmenge, $O^T$, und eine Attributmenge, $A$. Eine Teilmenge, $\ddot{A} \subseteq A$, bestimmt eine binäre Relation $I_{\ddot{A}}$, die „*Ununterscheidbarkeitsrelation*“ genannt wird.

---

[103] Vgl. zu den eingeführten Begriffen: PAWLAK (1998), S. 29 f.

[104] Das Beispiel wurde entnommen aus: BISSANTZ (1996), S. 69. Es wurde für die Zwecke dieser Arbeit angepaßt.

Zwei Objekte, $o,o' \in O^T$, sind genau dann ununterscheidbar bezüglich $\ddot{A}$, d.h. $I_{\ddot{A}}(o,o')$, wenn gilt:

$\forall a \in \ddot{A}: a(o) = a(o')$. ◊

Damit läßt sich der folgende Begriff definieren:

**Definition 2-30: Äquivalenzklasse**

Gegeben sei eine Objektmenge, $O^T$, und eine Ununterscheidbarkeitsrelation, $I_{\ddot{A}}$. Eine auf $I_{\ddot{A}}$ basierende *Äquivalenzklasse* zu einem Objekt, $o \in O^T$, ist eine Teilmenge der gesamten Objektmenge, $O^T$, die alle Objekte enthält, die zu $o$ ununterscheidbar sind:

$\ddot{A}(o) := \{o' \in O^T \mid I_{\ddot{A}}(o,o')\}$. ◊

*Beispielsweise ergibt sich aus der Attributmenge $\ddot{A} = \{Artikelgruppe, Kundengruppe, Region\}$ für das Objekt $o_1$ die Äquivalenzklasse $\ddot{A}(o_1) = \{o_1, o_5\}$ (vgl. Tabelle 2-4).*

| Objekt | Artikelgruppe | Kundengruppe | Region | Umsatz |
|---|---|---|---|---|
| $o_1$ | Tourenräder | Fachhandel | Süd | niedrig |
| $o_5$ | Tourenräder | Fachhandel | Süd | hoch |

***Tabelle 2-4: Die Äquivalenzklasse $\ddot{A}(o_1)$***

Eine Äquivalenzklasse, $\ddot{A}(o)$, induziert einen Term in konjunktiver Normalform:

**Definition 2-31: Durch eine Äquivalenzklasse induzierter Term**

Gegeben sei eine Objektmenge, $O^T$, eine Attribut-Teilmenge, $\ddot{A} = \{\ddot{a}_1, ..., \ddot{a}_{\ddot{a}max}\} \subseteq A$, und eine Äquivalenzklasse, $\ddot{A}(o)$ mit $o \in O^T$. Dann bestimmt man den *durch die Äquivalenzklasse $\ddot{A}(o)$ induzierten Term* $T_{\ddot{A}}$ durch die Attributwerte, $\ddot{a}_1(x), ..., \ddot{a}_{\ddot{a}max}(x)$, eines beliebigen Objektes, $x$, aus der Äquivalenzklasse:

$T_{\ddot{A}} := (Kl_1 \wedge ... \wedge Kl_{\ddot{a}max})$ mit
$Kl_i := ( \ddot{a}_i = \ddot{a}_i(x) )$;
$x \in \ddot{A}(o)$;
$i = 1, ..., \ddot{a}max$. ◊

*Beispielsweise induziert die Äquivalenzklasse aus Tabelle 2-4, $\ddot{A}(o_1) = \{o_1, o_5\}$, den folgenden Term:*

```
Artikelgruppe = Tourenräder UND
Kundengruppe = Fachhandel UND
Region = Süd.
```

Der Begriff der *Rough Set* kommt nun dadurch zustande, daß eine interessante Objektmenge, wie z.B. die Menge aller Datensätze mit hohem Umsatz, durch zwei andere Mengen „grob“ approximiert wird. Diese beiden Mengen sind wie folgt definiert:

**Definition 2-32: Obere und untere Näherung**

Gegeben sei eine Objektmenge, $O^T$, und eine Attributmenge, $A$. Als „*untere Näherung,* $UN_{Ä}(X)$, einer Objektmenge $X \subseteq O^T$ bezüglich einer Attributmenge $Ä \subseteq A$“ bezeichnet man die Menge aller Objekte, deren Äquivalenzklassen komplett in $X$ enthalten sind:

$$UN_{Ä}(X) := \{o \in O^T \mid Ä(o) \subseteq X\}.$$

Als „*obere Näherung*“, $ON_{Ä}(X)$, bezeichnet man die Menge aller Objekte, deren Äquivalenzklassen zumindest ein Objekt mit $X$ gemeinsam haben:

$$ON_{Ä}(X) := \{o \in O^T \mid Ä(o) \cap X \neq \varnothing\}. \qquad \lozenge$$

*Im Beispiel sei die Menge aller Objekte mit hohem Umsatz,* $X = \{o_2, o_5, o_8\}$*, durch die Attributmenge* $Ä$ *zu beschreiben. Hierzu selektiert man aus* $O^T$ *die Menge der Objekte mit hohem Umsatz,* $X$*, bildet die Äquivalenzklassen* $Ä(o)$*,* $\forall o \in O^T$ *und erhält anschließend die folgende obere und untere Näherung:*
$UN_{Ä}(\{o_2, o_5, o_8\}) = \{o_2, o_8\}$ *(vgl. Tabelle 2-5),*
$ON_{Ä}\{o_2, o_5, o_8\}) = \{o_1, o_2, o_5, o_8\}$ *(vgl. Tabelle 2-6).*

| Objekt | Artikelgruppe | Kundengruppe | Region | Umsatz |
|---|---|---|---|---|
| $o_2$ | Cityräder | Discounter | West | hoch |
| $o_8$ | Cityräder | Discounter | West | hoch |

***Tabelle 2-5: Die untere Näherung*** $UN_{\{Artikelgruppe, Kundengruppe, Region\}}(o_2, o_5, o_8)$

| Objekt | Artikelgruppe | Kundengruppe | Region | Umsatz |
|---|---|---|---|---|
| $o_1$ | Tourenräder | Fachhandel | Süd | niedrig |
| $o_5$ | Tourenräder | Fachhandel | Süd | hoch |
| $o_2$ | Cityräder | Discounter | West | hoch |
| $o_8$ | Cityräder | Discounter | West | hoch |

***Tabelle 2-6: Die obere Näherung*** $ON_{\{Artikelgruppe, Kundengruppe, Region\}}(o_2, o_5, o_8)$

Damit kann man den Begriff der *Rough Set* formal definieren:

**Definition 2-33: Grobe Menge (Rough Set)**

Gegeben sei eine Objektmenge, $O^T$, und eine Attributmenge, $A$. Eine Objektmenge $X \subseteq O^T$ ist genau dann eine „*Grobe Menge*“ („*Rough Set*“) bezüglich einer Attributmenge $Ä \subseteq A$, falls sich ihre obere und untere Näherung unterscheiden:

$$ON_{Ä}(X) - UN_{Ä}(X) \neq \varnothing. \qquad \lozenge$$

Die untere Näherung einer Objektmenge, $X$, bezüglich der Attribute $\ddot{A}$ kann als Regel gelesen werden, deren Prämisse Klauseln mit den Attributen aus $\ddot{A}$ umfaßt und deren Konklusion die Objektmenge $X$ beschreibt. Diese Regeln sind widerspruchsfrei in dem Sinne, daß es keinen Datensatz in der Trainingsmenge gibt, der die Prämisse, nicht aber die Konklusion erfüllt.

*Beispielsweise kann die untere Näherung der Objekte mit hohem Umsatz aus Tabelle 2-5 wie folgt als Regel gelesen werden:*

```
WENN Artikelgruppe = Cityräder
UND Kundengruppe = Discounter
UND Region = West
DANN Umsatz = hoch.
```

Es bleibt noch zu zeigen, wie aus einer Objektmenge eine Regelmenge induziert werden kann. Hierzu partitioniert man zunächst die Objektmenge nach folgender Vorschrift:

**Definition 2-34: Partition**

Gegeben sei eine Objektmenge, $O^T$, und eine Attributmenge, $A$. Dann ist $O^T/\ddot{A}$ die Menge aller Äquivalenzklassen in $O^T$, die sich durch die Attributmenge $\ddot{A} \subseteq A$ ergibt. $O^T/\ddot{A}$ wird „*Partition*“ genannt und läßt sich formal wie folgt definieren:

$$O^T/\ddot{A} := \{\ddot{A}(o) \mid o \in O^T\}. \qquad \Diamond$$

Diese Definition kann nun dazu genutzt werden, aus der Trainingsmenge, $O^T$, eine Regelmenge zu induzieren. Wie bisher soll die Menge der erklärenden Variablen mit $C$, die Menge der zu erklärenden Variablen mit $D$ bezeichnet werden. Zunächst ist die Trainingsmenge nach $D$ zu partitionieren.

*Im Beispiel ergibt sich für $D = \{Umsatz\}$ die in Tabelle 2-7 dargestellte Partition $O^T/D = \{\{o_1,o_4\}, \{o_2,o_5,o_8\}, \{o_3,o_6,o_7\}\}$.*

| Objekt | Artikelgruppe | Kundengruppe | Region | Umsatz |
|---|---|---|---|---|
| $o_1$ | Tourenräder | Fachhandel | Süd | niedrig |
| $o_4$ | Tourenräder | Fachhandel | Nord | niedrig |
| $o_2$ | Cityräder | Discounter | West | hoch |
| $o_5$ | Tourenräder | Fachhandel | Süd | hoch |
| $o_8$ | Cityräder | Discounter | West | hoch |
| $o_3$ | Mountainbikes | Fachhandel | Nord | mittel |
| $o_6$ | Mountainbikes | Kaufhaus | Nord | mittel |
| $o_7$ | Mountainbikes | Discounter | West | mittel |

***Tabelle 2-7: Die Partition $O^T/\{Umsatz\}$***

Hat man die Trainingsmenge nach den abhängigen Variablen, $D$, partitioniert, so bildet man für alle Klassen, $\forall X \in O^T/D$, die untere Näherung, $UN_C(X)$.

*Im Beispiel ergeben sich für $C$ = {Artikelgruppe, Kundengruppe, Region} und $D$ = {Umsatz} die folgenden unteren Näherungen für die Äquivalenzklassen aus der Partition $O^T/D = \{\{o_1,o_4\}, \{o_2,o_5,o_8\}, \{o_3,o_6,o_7\}\}$:*
*$UN_C(\{o_1,o_4\}) = \{o_4\}$;*
*$UN_C(\{o_2,o_5,o_8\}) = \{o_2, o_8\}$;*
*$UN_C(\{o_3,o_6,o_7\}) = \{o_3,o_6,o_7\}$.*

Damit kann die Menge aller Objekte, die widerspruchsfrei einer Klasse zugeordnet werden können, wie folgt definiert werden:

**Definition 2-35: Positive Region**

Gegeben seien mit $O^T$ eine Objektmenge, mit $A$ die Menge der beobachteten Attribute, mit $D$ eine Menge von zu erklärenden und mit $C$ eine Menge von erklärenden Attributen, wobei gilt: $D,C \subset A$, $C \cap D = \emptyset$, $C,D \neq \emptyset$. Dann heißt $POS_C(D)$ „*positive Region*“ mit:

$$POS_C(D) = \bigcup_{X \in O^T/D} UN_C(X).$$ ◊

*Beispielsweise ergibt die Menge aller Datensätze, die mit den erklärenden Attributen $C$ := {Artikelgruppe, Kundengruppe, Region} widerspruchsfrei einer Umsatzklasse zugeordnet werden können, die in Tabelle 2-8 abgebildete positive Region:*

$POS_C(D) = UN_C(\{o_1,o_4\}) \cup UN_C(\{o_2,o_5,o_8\}) \cup UN_C(\{o_3,o_6,o_7\})$
$= \{o_4\} \cup \{o_2, o_8\} \cup \{o_3,o_6,o_7\}.$

| Objekt | Artikelgruppe | Kundengruppe | Region | Umsatz |
|---|---|---|---|---|
| $o_4$ | Tourenräder | Fachhandel | Nord | niedrig |
| $o_2$ | Cityräder | Discounter | West | hoch |
| $o_8$ | Cityräder | Discounter | West | hoch |
| $o_3$ | Mountainbikes | Fachhandel | Nord | mittel |
| $o_6$ | Mountainbikes | Kaufhaus | Nord | mittel |
| $o_7$ | Mountainbikes | Discounter | West | mittel |

***Tabelle 2-8: Die positive Region $POS_{\{Artikelgruppe, Kundengruppe, Region\}}(\{Umsatz\})$***

Aus der positiven Region kann wie folgt eine Menge widerspruchsfreier Regeln induziert werden:

**Definition 2-36: Durch eine positive Region induzierte Regelmenge**

Es gelten dieselben Voraussetzungen wie in der Definition zuvor. Dann bestimmt man die *durch die positive Region $POS_C(D)$ induzierte Regelmenge* $M_{O^T}$ wie folgt:

$M_{O^T} := \{(Pr \rightarrow Ko) \mid Pr$ ist der durch $C(o)$ induzierte Term;

$Ko$ ist der durch $D(o)$ induzierte Term;

$o \in POS_C(D)$ }. ◊

*Beispielsweise kann aus Tabelle 2-8 folgende Regelmenge induziert werden:*

```
WENN Artikelgruppe = Cityräder
UND Kundengruppe = Discounter
UND Region = West
DANN Umsatz = hoch.

WENN Artikelgruppe = Mountainbikes
UND Kundengruppe = Fachhandel
UND Region = Nord
DANN Umsatz = mittel.

WENN Artikelgruppe = Tourenräder
UND Kundengruppe = Fachhandel
UND Region = Nord
DANN Umsatz = niedrig.

WENN Artikelgruppe = Mountainbikes
UND Kundengruppe = Kaufhaus
UND Region = Nord
DANN Umsatz = mittel.

WENN Artikelgruppe = Mountainbikes
UND Kundengruppe = Discounter
UND Region = West
DANN Umsatz = mittel.
```

**Vorteilhaft** gegenüber den zuvor dargestellten Organisationsformen für Regelmengen sind Rough-Set-Regelmengen aufgrund ihrer übersichtlichen Darstellung in Tabellenform und ihrer effizienten Generierbarkeit, die daraus resultiert, daß aus einer Trainingsmenge und einer Menge von Attributen unmittelbar eine Regelmenge induziert werden kann.

**Nachteilig** im Vergleich zu den übrigen Formen sind Rough-Set-Regelmengen aufgrund ihrer eingeschränkten Approximationsfähigkeit. Diese rührt daher, daß erstens alle Regeln dieselben Attribute umfassen. Zweitens stellt es eine starke Einschränkung dar, wenn man fordert, daß die von einer Regel abgedeckten Objekte widerspruchsfrei sein müssen. Drittens gilt der vorgestellte Ansatz nur für diskrete Domänen.

Alle genannten Nachteile können durch Erweiterungen behoben werden. Zur Behebung des ersten Nachteils lassen sich Algorithmen auf eine Rough-Set-Regelmenge anwenden, die überflüssige Klauseln aus der Regelmenge entfernen.[105] Die beiden anderen Nachteile werden in Abschnitt 5.2.2 diskutiert.

[105] Vgl. SHAN ET AL. (1995), S. 267 f.

### 2.2.3 Die Suche nach Datenmustern und Datenmuster-Mengen im Data Mining

Bereits in Abschnitt 2.1.1 wurde hergeleitet, daß zur induktiven Modellbildung neben einem Lösungsraum weitere „intelligente“ Komponenten benötigt werden. Noch nicht dargestellt wurden die algorithmischen Komponenten eines lernenden Systems. Hierzu zählen neben der in Abschnitt 2.2.4 darzustellenden Bewertungskomponente:

- ⇨ die Bestimmung des Ausgangspunktes für die Suche,
- ⇨ die möglichen Operationen zur Suche im Lösungsraum,
- ⇨ die Auswahl der zu testenden Operationen,
- ⇨ die Entscheidung über die Akzeptanz der getesteten Operationen,
- ⇨ die Auswahl der durchzuführenden aus den akzeptierten Operationen sowie
- ⇨ die Kriterien, die den Abbruch der Suche determinieren.

Diese algorithmischen Komponenten werden in dem folgenden Abschnitt 2.2.3.1 im Rahmen einer Kontrollstruktur vorgestellt, die so generell ist, daß sie für die meisten Data-Mining-Verfahren gilt.[106] Die daran anschließenden Abschnitte 2.2.3.2 bis 2.2.3.7 beschreiben die genannten algorithmischen Komponenten im einzelnen.

#### *2.2.3.1 Grundlagen und generelle Kontrollstruktur von Suchverfahren*

Jede Art von Problemlösung, so auch das Data Mining, kann als Suche in einem Lösungsraum, $L$, aufgefaßt werden, wobei gemäß einer Auswahlstrategie eine Folge von Suchoperationen durchgeführt wird, die einen Anfangszustand in einen Zielzustand überführt.[107] Die Bestimmung des Anfangszustandes wird im folgenden Abschnitt behandelt, das Erreichen eines Zielzustandes in Abschnitt 2.2.3.7. An dieser Stelle erfolgt die Definition der zur Überführung von Anfangs- in Zielzustände dienenden Suchoperationen:

---

[106] Auch die Beschreibung des in Abschnitt 5.4 vorzustellenden Verfahrens kann sich an dieser Kontrollstruktur orientieren.

[107] Vgl. JOEREßEN/SEBASTIAN (1998), S. 42.

**Definition 2-37: Transformation/Suchoperation/Suchschritt/Zug**

Gegeben sei ein Lösungsraum, $L$. Dann ist eine *Transformation*, $tr$, (oder auch „*Suchoperation*“, „*Suchschritt*“, „*Zug*“) eine Abbildung, die jedem Punkt im Lösungsraum, $s \in L$, einen „benachbarten“ Punkt im Lösungsraum, $tr(s) \in L$, zuordnet:

$$tr: \quad L \rightarrow L,$$
$$s \rightarrow tr(s).$$ ◊

Mit Hilfe der vorstehenden Definition kann man den Begriff der „Nachbarschaft“ abgrenzen:

**Definition 2-38: Nachbarschaft**[108]

$Tr$ sei die Menge der erlaubten Transformationen und $s$ eine Lösung. Dann ist die *Nachbarschaft* der Lösung, $N(s,Tr)$, die Menge der Lösungen, die von $s$ aus durch Anwendung einer Transformation, $tr \in Tr$, erreicht werden kann:

$$N(s,Tr) := \{s' \mid s' = tr(s);\ tr \in Tr\}.$$ ◊

Nach der Stärke der Transformation unterscheidet man globale und lokale Suchschritte:

- ⇨ **Lokale Suchschritte** sind auf die Nachbarschaft der aktuellen Lösung beschränkt, welche i.d.R. im Vergleich zum Lösungsraum sehr klein definiert ist. Sie dienen insbesondere dazu, innerhalb eines kleinen Ausschnitts aus dem Lösungsraum die beste Lösung, ein lokales Optimum, zu ermitteln.
- ⇨ **Globale Suchschritte** können nicht nur die Nachbarschaft der aktuellen Lösung erreichen, sondern alle Lösungen aus $L$. Sie dienen insbesondere dazu, den Einflußbereich lokaler Optima zu verlassen und in einen neuen Ausschnitt des Lösungsraums vorzudringen.

Die Verwendung lokaler und globaler Suchschritte wird im Rahmen lokaler und globaler Suchstrategien in Abschnitt 2.2.3.4 verdeutlicht.

[108] Vgl. REEVES (1993), S. 5.

Im Zusammenhang mit Suchverfahren verwendet man i.d.R. den Begriff des Suchraums anstelle des Lösungsraums aus Definition 2-15. Denn der Suchraum stellt eine durch das Suchverfahren bestimmte Teilmenge des Lösungsraums dar. Die Abgrenzung wird aus folgender Definition ersichtlich:

**Definition 2-39: Suchraum**

$Tr$ sei eine Menge von Transformationen, $L$ der Lösungsraum, $s_0 \in L$ eine Anfangslösung und $\mathbf{N}$ die Menge der natürlichen Zahlen. Dann ist der Suchraum als Menge derjenigen Lösungen des Lösungsraumes definiert, die von $s_0$ aus durch Transformationen aus $Tr$ erreichbar sind:

$S = \{s_n \in L \mid tr_1(s_0)=s_1, tr_2(s_1)=s_2, ..., tr_n(s_{n-1})=s_n\}$;
$tr_i \in Tr$;
$i = 0, ..., n$;
$n \in \mathbf{N} \cup \{0\}$. ◊

Im folgenden soll anhand von Abbildung 2-7 ein allgemeiner Algorithmus vorgestellt werden, der zur Durchforstung des Suchraums eingesetzt werden kann.[109]

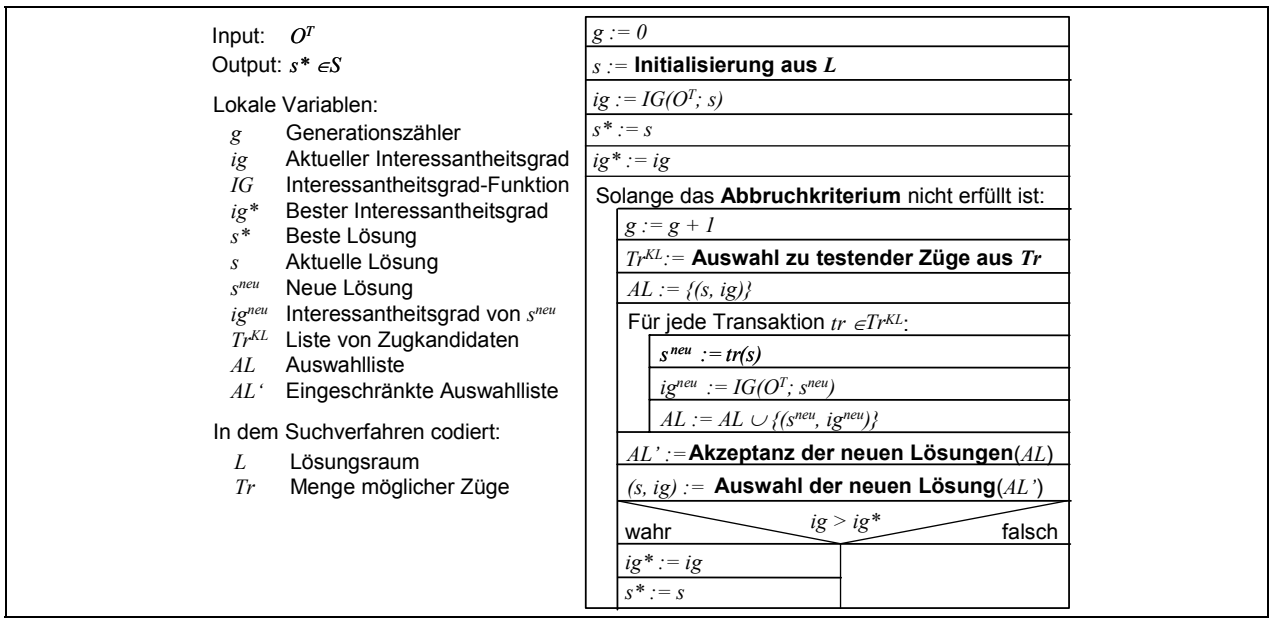



**Abbildung 2-7: Allgemeine Kontrollstruktur eines Data-Mining-Suchverfahrens**

[109] Auch VAESSENS, AARTS und LENSTRA haben eine generelle Kontrollstruktur für Suchverfahren vorgestellt (vgl. VAESSENS/AARTS/LENSTRA (1992), S. 67 ff.). Diese ist im Vergleich zu der hier vorgestellten Kontrollstruktur noch allgemeiner, da sie auch mehrphasige Suchverfahren einschließt, welche verschiedene Suchstrategien kombinieren. Diese Erweiterung könnte hier relativ leicht durchgeführt werden, bringt aber keine neuen Erkenntnisse. Dafür wird hier der Suchprozeß genauer zerlegt als bei VAESSENS, AARTS und LENSTRA, da dort nicht zwischen der Akzeptanz und der Auswahl neuer Lösungen unterschieden wird.

In die gezeigte, allgemein gehaltene Kontrollstruktur lassen sich die meisten Verfahren einordnen, die Data Mining als Optimierungsproblem betrachten. Dies gilt auch für Verfahren, die nicht mit einer einzelnen aktuellen Lösung arbeiten, sondern mit einer sog. „Population“, wie z.B. genetische Algorithmen[110] oder Evolutionsstrategien[111]. Um diese in das Schema einordnen zu können, muß $s$ einen Vektor von individuellen Lösungen, $s = (s_1, ..., s_{Pop})'$, darstellen. Entsprechend muß $ig = (ig_1, ..., ig_{Pop})'$ einen Vektor der dazugehörigen Interessantheitsgrade darstellen (mit $s_i \in S$; $ig_i \in \mathbf{R}$; $i = 1, ..., Pop$; $Pop$: Anzahl Individuen in der Population, $ig_i$: Interessantheitsgrad des Modells $s_i$).

Die in Abbildung 2-7 fettgedruckten Komponenten von Data-Mining-Verfahren werden in den Abschnitten 2.2.3.3 bis 2.2.3.7 behandelt. Zuvor werden im nächsten Abschnitt grundlegende Suchoperationen für Regeln in KNF eingeführt.

#### *2.2.3.2 Operationen zur Suche im Suchraum*

Im folgenden sollen mit der Generalisierung und der Spezialisierung zwei Typen von Suchoperationen vorgestellt werden, die in vielen Data-Mining-Verfahren eine zentrale Stellung einnehmen.[112] Die Generalisierung und die Spezialisierung lassen sich sowohl bezüglich eines Terms als auch bezüglich einer Klausel in konjunktiver Normalform definieren:

**Definition 2-40: Generalisierung eines Terms**

Es seien $A$ die Menge der beobachteten Attribute und $Te^{KNF}(A)$ die Menge aller möglichen Terme in konjunktiver Normalform bezüglich dieser Attribute. Die *Generalisierung eines Terms*, $Te = (Kl_1 \wedge ... \wedge Kl_{Klmax})$, bezüglich der Stelle $i \in \{1, ..., Klmax\}$ ist eine Operation, die die $i$-te Klausel aus $Te$ streicht:

$$\begin{aligned} Termgeneralisierung^i: \quad Te^{KNF}(A) &\rightarrow Te^{KNF}(A), \\ Te &\rightarrow Termgeneralisierung^i(Te) \\ &:= (Kl_1 \wedge ... \wedge Kl_{i-1} \wedge Kl_{i+1} \wedge ... \wedge Kl_{Klmax}) \end{aligned}$$

mit $Klmax > 1$. ◊

[110] Vgl. GOLDBERG (1989), S. 1 ff.

[111] Vgl. RECHENBERG (1973), S. 19 ff. und SCHWEFEL (1981), S. 104 ff.

[112] Weitere Suchoperationen werden in MICHALSKI (1994), S. 27 ff., vorgestellt.

**Definition 2-41: Generalisierung einer nominalen Klausel**

Es seien $A$ die Menge der beobachteten Attribute und $Kl^{KNF}(A)$ die Menge aller möglichen Klauseln in konjunktiver Normalform bezüglich dieser Attribute. Die *Generalisierung einer nominalen Klausel*, $Kl = (a \in \{w_1, ..., w_{wmax}\})$, bezüglich eines nominalen Wertes, $w^{neu} \in dom(a)$, ist eine Operation, die die Wertemenge $\{w_1, ..., w_{wmax}\}$ um den Wert $w^{neu}$ erweitert:

$$\begin{aligned} Generalisierung^{w^{neu}}: \quad & Kl^{KNF}(A) \quad \rightarrow \quad Kl^{KNF}(A), \\ & Kl \qquad\quad \rightarrow \quad Generalisierung^{w^{neu}} := (a \in \{w_1, ..., w_{wmax}, w^{neu}\}). \end{aligned}$$ ◊

**Definition 2-42: Generalisierung einer nichtnominalen Klausel**

Es gelten dieselben Voraussetzungen wie in Definition 2-41. Die *Generalisierung einer nichtnominalen Klausel*, $Kl = (a \in [w^{ug}; w^{og}])$, mit den neuen Intervallgrenzen $ug \in dom(a)$ und $og \in dom(a)$ ist eine Operation, die der Klausel die neuen Intervallgrenzen genau dann zuordnet, wenn sich dadurch das Intervall vergrößert:

$$\begin{aligned} Generalisierung^{ug,og}: \quad & Kl^{KNF}(A) \quad \rightarrow \quad Kl^{KNF}(A), \\ & Kl \qquad\quad \rightarrow \quad Generalisierung^{ug,og}(Kl) := (a \in [ug'; og']) \end{aligned}$$

mit $ug' := \begin{cases} ug & \text{falls } ug \le w^{ug}; \\ w^{ug} & \text{sonst;} \end{cases}$

$og' := \begin{cases} og & \text{falls } og \ge w^{og}; \\ w^{og} & \text{sonst.} \end{cases}$ ◊

Nach der Generalisierung werden i.d.R. durch den Term bzw. durch die Klausel mehr Trainingsobjekte abgedeckt als zuvor. Es gilt:

$|O^T[Termgeneralisierung^i(Te)]| \ge |O^T[Te]|$ bzw.

$|O^T[Generalisierung^{w^{neu}}(Kl)]| \ge |O^T[Kl]|$ bzw.

$|O^T[Generalisierung^{ug,og}(Kl)]| \ge |O^T[Kl]|$.

Den umgekehrten Effekt erzielt man durch die entsprechenden Spezialisierungszüge:

**Definition 2-43: Spezialisierung eines Terms**

Es gelten dieselben Voraussetzungen wie in Definition 2-40. Die *Spezialisierung eines Terms*, $Te = (Kl_1 \wedge ... \wedge Kl_{Klmax})$, bezüglich einer Klausel $Kl^{neu} \in Kl^{KNF}(A)$ ist eine Operation, die den Term $Te$ um die Klausel $Kl^{neu}$ erweitert:

$$Termspezialisierung^{Kl^{neu}}: \; Te^{KNF}(A) \rightarrow Te^{KNF}(A),$$
$$Te \rightarrow Termspezialisierung^{Kl^{neu}}(Te) := (Kl_1 \wedge ... \wedge Kl_{Klmax} \wedge Kl^{neu}).$$ ◊

**Definition 2-44: Spezialisierung einer nominalen Klausel**

Es gelten dieselben Voraussetzungen wie in Definition 2-41. Die *Spezialisierung einer nominalen Klausel*, $Kl = (a \in \{w_1, ..., w_{wmax}\})$, bezüglich der Stelle $i \in \{1, ..., wmax\}$ ist eine Operation, die den $i$-ten Wert aus der Wertemenge $\{w_1, ..., w_{wmax}\}$ streicht:

$$Spezialisierung^{i}: \; Kl^{KNF}(A) \rightarrow Kl^{KNF}(A),$$
$$Kl \rightarrow Spezialisierung^{i}(Kl) := (a \in \{w_1, ..., w_{i-1}, w_{i+1}, ..., w_{wmax}\}).$$

mit $wmax > 1$. ◊

**Definition 2-45: Spezialisierung einer nichtnominalen Klausel**

Es gelten dieselben Voraussetzungen wie in Definition 2-41. Die *Spezialisierung einer nichtnominalen Klausel*, $Kl = (a \in [w^{ug}; w^{og}])$, mit den neuen Intervallgrenzen $ug \in dom(a)$ und $og \in dom(a)$ ist eine Operation, die der Klausel die neuen Intervallgrenzen genau dann zuordnet, wenn sich dadurch das Intervall verkleinert:

$$Spezialisierung^{ug,og}: \; Kl^{KNF}(A) \rightarrow Kl^{KNF}(A),$$
$$Kl \rightarrow Spezialisierung^{ug,og}(Kl) := (a \in [ug'; og']);$$

mit $ug' := \begin{cases} ug & \text{falls } ug \geq w^{ug}; \\ w^{ug} & \text{sonst;} \end{cases}$

$og' := \begin{cases} og & \text{falls } og \leq w^{og}; \\ w^{og} & \text{sonst.} \end{cases}$ ◊

Nach der Spezialisierung werden i.d.R. durch den Term bzw. durch die Klausel weniger Trainingsobjekte abgedeckt als zuvor. Es gilt:

$|O^T[Termspezialisierung^{Kl^{neu}}(Te)]| \leq |O^T[Te]|$ bzw.

$|O^T[Spezialisierung^{i}(Kl)]| \leq |O^T[Kl]|$ bzw.

$|O^T[Spezialisierung^{ug,og}(Kl)]| \leq |O^T[Kl]|$.

Die konkreten Werte, die bei den klauselorientierten Generalisierungs- und Spezialisierungszügen zu streichen bzw. hinzuzufügen sind, ergeben sich häufig aus einem

vorgegebenen Konzeptbaum[113]. Ein ***Konzeptbaum*** für ein Attribut, $a$, ist ein Baum gemäß Definition 2-24, in dem jeder Knoten eine Wertemenge, $WM \subseteq dom(a)$, und jede Kante eine partielle Ordnung zwischen den verbundenen Wertemengen repräsentiert (vgl. Abbildung 2-8). Die Wurzel bildet die Wertemenge $ALLE = dom(a)$. Eine Generalisierung entspricht dann dem Übergang von einem Knoten, $v_2$, über eine Kante, $(v_1,v_2)$, zu seinem Vorgänger, $v_1$. Analog entspricht eine Spezialisierung dem Übergang von einem Knoten, $v_1$, über eine Kante, $(v_1,v_2)$, zu seinem Nachfolger, $v_2$.

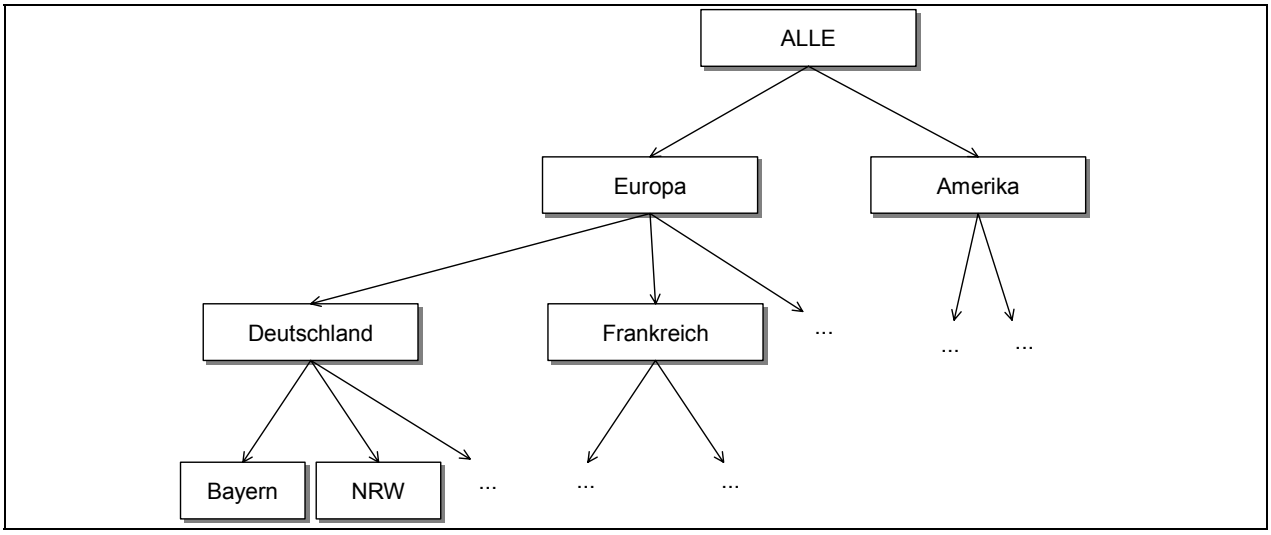



***Abbildung 2-8: Konzeptbaum für das Attribut „Region“***

Will man ohne einen solchen Konzeptbaum auskommen, um den damit verbunden manuellen Definitionsaufwand zu vermeiden, so muß das Suchverfahren die entsprechenden Werte und Klauseln völlig frei variieren. Diese Freiheitsgrade führen zu einem wesentlich größeren Suchraum als bei Verwendung eines Konzeptbaumes.

### *2.2.3.3 Initialisierung der Suche*

Dem Suchalgorithmus obliegt die Aufgabe, eine Ausgangslösung, $s \in S$, für den Suchprozeß zu generieren. Im Data Mining sind zwei Ansätze zur Initialisierung der Suche weit verbreitet:[114]

⇨ **Datenbasierte Initialisierung**: Als erste Datenmuster-Menge wird das *speziellste mögliche Modell*, $M^{speziell}$, aus der Trainingsmenge, $O^T$, erzeugt.

[113] Vgl. ESTER/SANDER (2000), S. 190 f.

[114] Vgl. HOLSHEIMER/SIEBES (1994), S. 29 f.

⇨ **Modellbasierte Initialisierung**: Als erste Datenmuster-Menge wird das *allgemeinste mögliche Modell*, $M^{allgemein}$, aus den Domänen der betrachteten Attribute, $A$, erzeugt.

Zwischen diesen beiden Extremen sind Kompromißlösungen denkbar, wie z.B. die zufällige Generierung eines Modells mittlerer Spezialisierung und mittlerer Allgemeinheit. Die eingeführten Begriffe werden im folgenden formal definiert.

**Definition 2-46: Das speziellste mögliche Modell**

Es sei $O^T$ eine Trainingsmenge mit $N$ Datenobjekten und der Attributmenge $A = \{a_1, ..., a_{amax}\}$ gemäß Definition 2-2. Weiterhin gebe es eine ausgezeichnete Teilmenge von zu erklärenden Attributen, $D \subset A$, $D = \{a_{amax-acmax+1}, ..., a_{amax-1}, a_{amax}\}$. Dann besteht das *speziellste mögliche Modell*, $M^{speziell}$, aus genau einer Regel, $s_i$, pro Datenobjekt, $o_i$:

$M^{speziell} := \{s_1, ..., s_N\}$ mit

$s_i := ((Kl_{1,i} \wedge ... \wedge Kl_{amax-acmax,i}) \rightarrow (Kl_{amax-acmax+1,i} \wedge ... \wedge Kl_{amax,i}))$;

$$Kl_{j,i} := \begin{cases} (a_j = a_j(o_i)) & \text{falls } a_j \text{ nominal;} \\ (a_j \in [a_j(o_i); a_j(o_i)]) & \text{sonst;} \end{cases}$$

$j = 1, ..., amax$ und

$i = 1, ..., N$. ◊

Zur Verdeutlichung dieser Definition möge das folgende Beispiel dienen.

*Gegeben seien die Trainingsdaten aus Tabelle 2-9. $A = \{Name, Alter, Geschlecht, Kunde\}$ seien die Attribute; die zu erklärende Attribut-Teilmenge sei $D = \{Kunde\}$.*

| *Name* | *Alter* | *Geschlecht* | *Kunde* |
|---|---|---|---|
| *Ann* | *32* | *w* | *ja* |
| *John* | *53* | *m* | *ja* |
| *Mary* | *27* | *w* | *nein* |
| *James* | *55* | *m* | *ja* |

***Tabelle 2-9: Eine Beispiel-Trainingstabelle***

*Das komplexeste mögliche Modell bezüglich des zu erklärenden Attributs „Kunde" lautet:*

```
WENN Name = Ann UND Alter ∈ [32; 32] UND Geschlecht = w
DANN Kunde = ja;
WENN Name = John UND Alter ∈ [53; 53] UND Geschlecht = m
DANN Kunde = ja;
WENN Name = Mary UND Alter ∈ [27; 27] UND Geschlecht = w
DANN Kunde = nein;
WENN Name = James UND Alter ∈ [55; 55] UND Geschlecht = m
DANN Kunde = ja.
```

Wenn es keine ausgezeichnete Menge von zu erklärenden Attributen, $C$, gibt, dann muß, wie die nächste Definition zeigt, in der Konklusion der $i$-ten Regel ein eigenes Segment, $c_i$, definiert werden (mit $i = 1, ..., N$).

**Definition 2-47: Das speziellste mögliche Modell (ohne zu erklärende Attribute)**

Es gelten dieselben Voraussetzungen wie in Definition 2-46 mit $D=\varnothing$. Dann besteht das *speziellste mögliche Modell*, $M^{speziell}$, aus $N$ Datenmustern, die für jeden Datensatz, $o_i$, ein eigenes Segment, $c_i$, definieren:

$M^{speziell} := \{s_1, ..., s_N\}$ mit

$s_i := ((Kl_{1,i} \wedge ... \wedge Kl_{amax,i}) \rightarrow (\texttt{Segment} = c_i));$

$$Kl_{j,i} := \begin{cases} (a_j = a_j(o_i)) & \text{falls } a_j \text{ nominal;} \\ (a_j \in [a_j(o_i); a_j(o_i)]) & \text{sonst;} \end{cases}$$

$j = 1, ..., amax$ und

$i = 1, ..., N$. ◊

Analog kann das allgemeinste mögliche Modell definiert werden:

**Definition 2-48: Das allgemeinste mögliche Modell**

Es gelten dieselben Voraussetzungen wie in Definition 2-46. Dann besteht das *allgemeinste mögliche Modell*, $M^{allgemein}$, aus einem Datenmuster, das alle Datensätze abdeckt und derselben allumfassenden Klasse, $dom(D)$, zuordnet:

$M^{allgemein} := \{((Kl_1 \wedge ... \wedge Kl_{amax}) \rightarrow (C \in dom(D)))\}$ mit:

$Kl_j := (a_j \in dom(a_j));$

$j = 1, ..., amax$. ◊

*Gegeben seien wieder die Trainingsdaten aus Tabelle 2-9. Dann lautet das allgemeinste mögliche Modell bezüglich des zu erklärenden Attributes „Kunde“:*

```
WENN Name ∈ {Ann, John, Mary, James}
UND Alter ∈ [0; 100]
UND Geschlecht ∈ {w, m}
DANN Kunde ∈ {ja, nein}.
```

**Definition 2-49: Das allgemeinste mögliche Modell (ohne zu erklärende Attribute)**

Es gelten dieselben Voraussetzungen wie in Definition 2-46 mit $D=\varnothing$. Dann besteht das *allgemeinste mögliche Modell ohne Vorgabe von zu erklärenden Attributen*,

$M^{allgemein}$, aus einem einzigen Datenmuster, das alle Datensätze abdeckt und einer einzigen, allumfassenden Klasse, $c_1$, zuordnet:

$M^{allgemein} := \{((Kl_1 \wedge ... \wedge Kl_{amax}) \rightarrow (\texttt{Segment} = c_i))\}$ mit:

$Kl_j := (a_j \in dom(a_j));$

$j = 1, ..., amax.$ ◊

### *2.2.3.4 Die Auswahl der zu testenden Suchoperationen*

Die Strategie zur Auswahl der zu testenden Suchoperationen ist eine wichtige Determinante der Leistungsfähigkeit von Suchstrategien. Zur genaueren Betrachtung der Auswahlstrategien sollen diese hier gemäß Abbildung 2-9 differenziert werden.

- Auswahl zu testender Suchoperationen
  - Verwendung globaler und lokaler Suchschritte
    - Lokale Suchstrategien
    - Globale Suchstrategien
    - Hybride Suchstrategien
  - Eingeschlagene Richtung im Suchraum
    - Top-Down
    - Bottom-Up
    - Gemischt
  - Erstellung einer Kandidatenliste
    - Strategien ohne Berücksichtigung der Suchhistorie
      - Erstellung einer vollständigen Liste aller Kandidaten
      - Erstellung einer Liste durch zufällig-gleichverteilte Auswahl von $K$ Kandidaten
      - Erstellung einer Liste durch kennzahlgesteuerte Auswahl von $K$ Kandidaten
    - Strategien mit Berücksichtigung der Suchhistorie
      - Tabu-Strategien
      - Erlernen von guten Schrittweiten
      - Erlernen von Bestandteilen guter Lösungen
      - Erlernen von vielversprechenden Teilmengen des Suchraums

**Abbildung 2-9: Strategien zur Auswahl zu testender Suchoperationen**

Zunächst werden die Auswahlstrategien nach folgenden Kriterien unterschieden:

⇨ nach der Verwendung globaler bzw. lokaler Suchschritte,

⇨ nach der eingeschlagenen Richtung im Suchraum und

⇨ nach der Strategie, gemäß der eine Menge von Suchoperationen, $Tr$, als Kandidaten für den nächsten Suchschritt (die sog. „***Kandidatenliste***“, $Tr^{KL}$) erstellt wird.

Diese Klassifizierungskriterien werden im folgenden genauer betrachtet.

Die **Differenzierung globaler und lokaler Suchschritte** wurde bereits in Abschnitt 2.2.3.1 angesprochen. Hier sollen lokale, globale und hybride Suchstrategien unterschieden werden:

- ⇨ **Lokale Suchstrategien** seien Strategien mit ausschließlich lokalen Suchschritten. Sie sind in der Lage, die Eigenschaften der lokalen Umgebung zu berücksichtigen. Beispielsweise können sie die Richtung des steilsten Anstiegs der Lösungsgüte einschlagen und über einige Iterationen beibehalten. Allerdings laufen sie Gefahr, lokale Optima nicht überwinden zu können.
- ⇨ **Globale Suchstrategien** seien Strategien mit ausschließlich globalen Suchschritten. Sie haben keine Probleme mit lokalen Optima, nutzen aber auch keine Informationen der lokalen Umgebung.
- ⇨ **Hybride Suchstrategien** kombinieren globale und lokale Suchschritte miteinander. Meist erfolgt dies durch alternierende Ausführung zweier Phasen: In einer ***Intensivierungsphase*** werden lokale Suchschritte dazu verwendet, innerhalb eines kleinen Ausschnitts aus dem Suchraum die beste Lösung, ein lokales Optimum, zu ermitteln. In einer anschließenden ***Diversifizierungsphase*** werden globale Suchschritte ausgeführt, um den Einflußbereich lokaler Optima zu verlassen und in einen neuen Ausschnitt des Suchraums vorzudringen. Das richtige Zusammenspiel von Intensivierung und Diversifizierung wird als entscheidender Faktor für die Leistungsfähigkeit von Suchstrategien erachtet.[115]

Nach dem zweiten o.g. Klassifizierungskriterium für Suchstrategien, der **eingeschlagenen Richtung im Suchraum**, unterscheidet man:[116]

- ⇨ Top-Down-Strategien,
- ⇨ Bottom-Up-Strategien und
- ⇨ gemischte Strategien.

**Top-Down-Strategien** setzen voraus, daß die Initialisierung modellbasiert vorgenommen wurde. Sie beginnen beim allgemeinsten möglichen Modell und spezialisieren es Schritt für Schritt.[117]

---

[115] Vgl. VOß/FIEDLER/GREISTORFER (2000), S. 564.

[116] Vgl. HOLSHEIMER/SIEBES (1994), S. 29 f.

*Führte man in dem Beispiel aus dem Abschnitt zuvor Spezialisierungsoperationen durch, welche die Wertemengen der Attribute in der Prämisse und der Konklusion verkleinern, dann würden nicht mehr alle Datenobjekte erfaßt. Für die nicht mehr erfaßten Datenobjekte können dann neue Datenmuster eingeführt werden. Das Streichen des Wertes „John" und dessen Aufnahme in ein neues Datenmuster würde beispielsweise zu folgendem Modell führen:*

```
WENN Name ∈ {Ann, Mary, James}
UND Alter ∈ [27; 55]
UND Geschlecht ∈ {w, m}
DANN Kunde ∈ {ja, nein};

WENN Name = John
UND Alter ∈ [53; 53]
UND Geschlecht = m
DANN Kunde = ja.
```

**Bottom-Up-Strategien** setzen voraus, daß die Initialisierung datenbasiert vorgenommen wurde. Sie beginnen beim speziellsten möglichen Modell und generalisieren es Schritt für Schritt.[118]

*Führte man in dem Beispiel aus dem Abschnitt zuvor Generalisierungsoperationen durch, welche die Wertemengen der Klauseln in der Prämisse vergrößern oder ganze Klauseln streichen, dann würden mehrere identische Datenmuster entstehen, von denen dann jeweils nur noch eines benötigt würde. Das Streichen des Attribut „Name" aus allen Datenmustern und das Vergrößern der Alters-Intervalle für Regeln, welche die männlichen Kunden abdecken, würde beispielsweise zu folgender Menge von Datenmustern führen:*

```
WENN Alter ∈ [32; 32] UND Geschlecht = w
DANN Kunde = ja;

WENN Alter ∈ [27; 27] UND Geschlecht = w
DANN Kunde = nein;

WENN UND Alter ∈ [53; 55] UND Geschlecht = m
DANN Kunde = ja.
```

Beide Ansätze, Top-Down- und Bottom-Up-Strategien, durchforsten den Suchraum im gesamten Suchprozeß in derselben Richtung, die durch die Initialisierung festgelegt wird. Dadurch wird der Suchraum relativ schnell durchquert, aber nicht besonders gut abgedeckt. **Gemischte Strategien** lassen dagegen Richtungswechsel zu, so daß vergangene Suchschritte rückgängig gemacht werden können.

---

[117] Beispiele für Top-Down-Strategien sind der zur Generierung von Entscheidungsbäumen eingesetzte Algorithmus „ID3" (vgl. QUINLAN (1986), S. 87 ff.), der zur Generierung von ungeordneten Regelmengen eingesetzte Algorithmus „AQ15" (vgl. MICHALSKI ET AL. (1986), S. 1041 ff.) und der zur Generierung von Entscheidungslisten eingesetzte Algorithmus „CN2" (vgl. CLARK/NIBLETT (1989), S. 267 ff.).

[118] Beispiele für Bottom-Up-Strategien sind die meisten attributorientierten Verfahren (vgl. ESTER/SANDER (2000), S. 206 ff.).

Geht man von einer gemischten Strategie mit den in Abschnitt 2.2.3.2 vorgeschlagenen Generalisierungs- und Spezialisierungsoperationen aus, so muß konkretisiert werden, mit welchen Parametern diese Operationen ausgeführt werden sollen. Die möglichen Werte für diese Parameter, also die zu generalisierenden bzw. zu spezialisierenden Attribute und Attributwerte, bilden die Kandidaten für den nächsten Suchschritt. Die Strategien zur **Erstellung einer Kandidatenliste** unterscheidet man in:

⇨ Strategien ohne Berücksichtigung der Suchhistorie und

⇨ Strategien mit Berücksichtigung der Suchhistorie.

Bei den **Strategien ohne Berücksichtigung der Suchhistorie** differenziert man die Art und Weise der Erstellung einer Kandidatenliste weiter:

⇨ Die **Erstellung einer vollständigen Liste aller möglichen Kandidaten** ist sehr zeitaufwendig und wird daher bei großen Suchräumen i.d.R. nur dann praktiziert, wenn reine Top-Down- oder Bottom-Up-Strategien eingesetzt werden, so daß der Suchraum relativ schnell durchquert werden kann. So werden beispielsweise bei den meisten Entscheidungsbaumverfahren für jeden Spezialisierungsschritt alle noch zur Verfügung stehenden Attribute in die Kandidatenliste aufgenommen.

⇨ Die **Erstellung einer Liste durch zufällig-gleichverteilte Auswahl von $K$ Kandidaten** kommt dann in Frage, wenn die Erstellung einer vollständigen Kandidatenliste zu aufwendig erscheint und keine intelligentere, kennzahlgesteuerte Strategie zur Verfügung steht. Sie birgt, falls $K$ zu klein gewählt wird, die Gefahr, daß wichtige Bereiche des Suchraums erst gar nicht beschritten werden.

⇨ Die **Erstellung einer Liste durch kennzahlgesteuerte Auswahl von $K$ Kandidaten** ist der zufälligen Auswahl immer dann vorzuziehen, wenn sinnvolle Kennzahlen zur Verfügung stehen und deren Berechnung nicht zu aufwendig erscheint.

*Eine solche Kennzahl könnte beispielsweise ein durch den Benutzer vorgegebener Relevanzwert für jedes Attribut sein.*

Die **Strategien mit Berücksichtigung der Suchhistorie** besitzen einen Speicher, in dem sie Informationen über die bisher durchgeführten Suchschritte und deren Erfolg sammeln. Diese Strategien erlernen bestimmte Eigenschaften des Suchraums. Bei genauerer Betrachtung lassen sich u.a. folgende Ansätze differenzieren:

⇨ Tabu-Strategien,

⇨ das Erlernen von guten Schrittweiten,

⇨ das Erlernen von Bestandteilen guter Lösungen und

⇨ das Erlernen einer vielversprechenden Teilmenge des Suchraums.

**Tabu-Strategien** speichern Informationen über zuletzt beschrittene Lösungen oder zuletzt durchgeführte Operationen in einer sog. „***Tabu-Liste***".[119] Die in einer Tabu-Liste gespeicherten Operationen bzw. Lösungen sind „tabu", d.h. die Operationen dürfen nicht durchgeführt bzw. die Lösungen nicht erneut getestet werden, um Zyklen[120] in der Problemlösung zu vermeiden. Aus der Kandidatenliste, $Tr^{KL}$, werden diejenigen Transformationen gestrichen, die tabu sind bzw. die zu Lösungen in der Tabu-Liste führen. Die Tabu-Liste muß bei den meisten Problemen nicht besonders groß sein, denn wenn von einer Lösung ausgehend eine gewisse Anzahl von Zügen durchgeführt wurde, dann wird die Wahrscheinlichkeit, daß diese Lösung noch einmal besucht wird, als gering eingeschätzt.

Das **Erlernen von guten Schrittweiten** stellt eine Idee dar, die den bekannten *Evolutionsstrategien*[121] entstammt. Diese stellen ein lokales Suchverfahren dar, das die Weiten der Schritte durch den Suchraum an die lokalen Gegebenheiten des Suchraums anpaßt. Schrittweiten lassen sich bei stetigen Wertebereichen sinnvoll definieren. Als Gegenstück zu den Schrittweiten wurden für kombinatorische Problemstellungen Mutationsraten für diskrete Variablenwerte eingeführt. Diese können aber nicht erlernt werden, wenn sich keine interpretierbaren Abstände zwischen den diskreten Werten definieren lassen.[122]

Das **Erlernen von Bestandteilen guter Lösungen** stellt eine Idee dar, die den bekannten *genetischen Algorithmen*[123] entstammt. Durch die Ausführung von sog.

---

[119] Vgl. GLOVER (1989), S. 192.

[120] Man sagt, die Suche gerate in einen ***Zyklus***, wenn immer wieder dieselben Folgen von Lösungen besucht werden. Am gravierendsten wirkt sich ein deterministischer Zyklus auf das Suchergebnis aus, der nicht mehr durchbrochen werden kann. Aber auch ein stochastischer Zyklus, dem die Suche nur mit geringer Wahrscheinlichkeit entkommen kann, kann das Suchergebnis negativ beeinflussen.

[121] Vgl. RECHENBERG (1973), S. 19 ff. und SCHWEFEL (1981), S. 104 ff.

[122] Vgl. BÄCK/SCHÜTZ (1995), S. 38.

[123] Vgl. GOLDBERG (1989), S. 1 ff.

„Selektionsoperationen“ werden gute Lösungen mit überdurchschnittlicher Wahrscheinlichkeit aus einem Lösungspool ausgewählt. Durch die Ausführung von sog. „Rekombinationsoperatoren“ werden die selektierten Lösungen zu Nachfolgerlösungen rekombiniert. Dadurch wird erreicht, daß sich bestimmte Kombinationen von Informationen über überdurchschnittlich gute Individuen exponentiell vermehren.[124] Diese als „***building blocks***“ bezeichneten Lösungsinformationen werden zu Lösungen zusammengesetzt. Einerseits reduziert dies die Dimension des Suchproblems gegenüber der Kombination der elementaren Lösungsinformationen. Andererseits funktioniert die Optimierung nur dann:

1. wenn sich relevante Lösungsinformationen auch als building blocks repräsentieren lassen und
2. wenn das Zusammensetzen von building blocks tatsächlich zu guten Lösungen führt.[125]

Kritisiert wird an genetischen Algorithmen vor allem, daß aufgrund von Punkt 1 zusätzlich zu dem eigentlichen Optimierungs- noch ein Repräsentationsproblem gelöst werden muß und daß aufgrund von Punkt 2 gute Lösungen, die sich aus unterdurchschnittlich guten Lösungsblöcken zusammensetzen, durch genetische Algorithmen nur mit verschwindend geringer Wahrscheinlichkeit gefunden werden können.[126] Diese Probleme versucht das auf derselben Grundidee basierende *Scatter Search*[127] zu umgehen, indem es erstens zuläßt, daß die Lösungen in ihrer problemabhängigen, „natürlichen“ Form repräsentiert werden und indem es zweitens flexiblere Arten der Rekombination von guten Lösungen gestattet.[128]

Das **Erlernen einer vielversprechenden Teilmenge des Suchraums** stellt einen Ansatz dar, der nicht nur Lösungen, $s \in S$, im Verlauf der Suche variiert, sondern auch die betrachtete Teilmenge des Suchraums, $S_i^T \subset S$.[129] Stellt $S_i^T$ den aktuell (in Phase $i$) betrachteten Suchraumausschnitt dar, so wird in dieser Phase die beste Lösung aus

---

[124] Vgl. GOLDBERG (1989), S. 33.

[125] Vgl. GOLDBERG (1989), S. 41 ff.

[126] Vgl. HEISTERMANN (1994), S. 62 ff.

[127] Vgl. zum Scatter Search GLOVER/LAGUNA (1997), S. 314 ff.

[128] Vgl. VOß/FIEDLER/GREISTORFER (2000), S. 554.

[129] Vgl. UTGOFF (1986), S. 113 ff.

diesem Ausschnitt, $s_i^* \in S_i^T$, gesucht. Nach einem bestimmten Kriterium wird festlegt, wann zu einem neuen Suchraumausschnitt, $S_{i+1}^T \subset S$, übergegangen werden soll. Dann kann der beste Suchraumausschnitt, $S_{i^*}^T$, als derjenige ermittelt werden, in dem – über alle Phasen, $i = 1, ..., I$, betrachtet – die beste Lösung, $s^* \in (S_1^T \cup \ldots \cup S_I^T)$, gefunden wurde.[130] Dieses Vorgehen stellt nicht nur eine Kandidatenlisten-Strategie dar, sondern auch eine Möglichkeit, die Suche nach dem optimalen Modell für zukünftige Data-Mining-Analysen auf einen kleineren Suchraum einzugrenzen, indem man dann $S := S_{i^*}^T$ definiert.

#### *2.2.3.5 Die Entscheidung über die Akzeptanz der neuen Lösungen*

Die Entscheidung über die Akzeptanz der neuen Lösungen in der Auswahlliste, $AL$, erfolgt aufgrund eines Akzeptanzkriteriums. Dieses Kriterium wird auf jeden Eintrag in der Auswahlliste, $al \in AL$, angewendet und entscheidet, ob er akzeptiert oder verworfen wird. Die akzeptierten Lösungen werden gemäß Abbildung 2-7 in die eingeschränkte Auswahlliste $AL'$ übernommen und kommen als potentielle Lösungen für die nächste Generation in Frage.

Man unterscheidet Akzeptanzkriterien in:

⇨ Verschlechterungen verbietende Strategien und

⇨ Verschlechterungen zulassende Strategien.

**Verschlechterungen verbietende Strategien** akzeptieren eine neue Lösung, $s^{neu}$, nur dann, wenn sie mindestens so interessant wie die alte Lösung ist, d.h. das Kriterium lautet: $ig^{neu} \geq ig^{alt}$. Dieses Akzeptanzkriterium impliziert das folgende Abbruchkriterium: Das Verfahren muß terminieren, wenn in der aktuellen Nachbarschaft, $N(s, Tr)$, keine Verbesserung mehr gefunden werden kann. Dies führt bei Erreichen eines lokalen Optimums zu einem vorzeitigen Abbruch des Verfahrens.[131]

**Verschlechterungen zulassende Strategien** akzeptieren eine neue Lösung, $s^{neu}$, auch dann, wenn $ig^{neu} < ig^{alt}$ ist – allerdings nur dann, wenn sich die Verschlechterung $ig^{alt} - ig^{neu}$ „im Rahmen hält“. Für die Entscheidung, ob eine Verschlechterung noch im

[130] Vgl. UTGOFF (1986), S. 112.

[131] Vgl. VOß/FIEDLER/GREISTORFER (2000), S. 556.

Rahmen liegt oder nicht, seien in Tabelle 2-10 einige ausgewählte Akzeptanzkriterien (mit den Bezeichnungen der Suchalgorithmen, denen sie entnommen wurden) angeführt.

| Algorithmus | Akzeptanzwahrscheinlichkeit für $s^{neu}$ | Ergänzungen |
|---|---|---|
| **Simulated Annealing**[132] | $e^{\frac{ig^{neu}-ig^{alt}}{c_g}}$ falls $ig^{neu} \le ig^{alt}$<br>1 sonst | $c_g$: „Temperatur“ in Generation $g$, wobei die „Temperatur“ fällt, d.h. es gilt:<br>$c_{g+1} \le c_g$,<br>$c_g \in \{r \mid r \in \mathbf{R}, r>0\}$, $g = 0, ..., gmax$. |
| **Threshold Accepting**[133] | $1$ falls $ig^{neu} - ig^{alt} > -c_g$<br>$0$ sonst | $c_g$: „Threshold“ in Generation $g$, wobei der „Threshold“ fällt, d.h. es gilt:<br>$c_{g+1} \le c_g$,<br>$c_g \in \{r \mid r \in \mathbf{R}, r>0\}$, $g = 0, ..., gmax$. |
| **Great Deluge**[134] | $1$ falls $ig^{neu} > c_g$<br>$0$ sonst | $c_g$: „Wasserstand“ in Generation $g$, wobei der „Wasserstand“ wie folgt steigt:<br>$c_{g+1} := c_g + c$,<br>$c, c_0 \in \{r \mid r \in \mathbf{R}, r>0\}$; $g = 0, ..., gmax$. |
| **Record-To-Record Travel**[135] | $1$ falls $ig^{neu} > Rekord_g - c$<br>$0$ sonst | $Rekord_g$: bester Interessantheitsgrad, der in den Generationen $0$ bis $g$ erreicht wurde;<br>$c \in \{r \mid r \in \mathbf{R}, r>0\}$; $g = 0, ..., gmax$. |

**Tabelle 2-10: Akzeptanzkriterien für neue Lösungen**

Beim **Simulated Annealing** wird in Anlehnung an den Abkühlungsprozeß von Kristallen ein sog. „Temperatur“-Parameter, $c_g$, so eingestellt, daß er von Generation zu Generation kleiner wird oder gleich bleibt. Je kleiner die Temperatur ist, desto geringer ist die Akzeptanzwahrscheinlichkeit c.p. (für eine feste Lösungsverschlechterung, $ig^{alt}-ig^{neu}$). Eine Lösungsverbesserung wird immer akzeptiert.

Beim **Threshold Accepting** wird eine Verschlechterung des Interessantheitsgrades deterministisch immer dann akzeptiert, wenn sie über einer Schranke, $c_g$, liegt. Wie zuvor sinkt auch hier $c_g$ mit fortschreitender Suche, so daß die Akzeptanz einer schlechteren Lösung tendenziell immer seltener wird.

---

[132] Vgl. KIRKPATRICK/GELATT/VECCHI (1983), S. 672 f.

[133] Vgl. DUECK/SCHEUER (1990), S. 162.

[134] Vgl. DUECK (1993), S. 87.

[135] Vgl. DUECK (1993), S. 87.

Beim **Great Deluge** wird der Interessantheitsgrad der aktuellen Lösung nicht, wie bei den beiden Ansätzen zuvor, mit dem der letzten Lösung verglichen, sondern direkt mit einem vorgegebenen Parameter, $c_g$. Die neue Lösung wird deterministisch dann akzeptiert, wenn ihre Bewertung über $c_g$ liegt, wobei die Anforderung $c_g$ von Generation zu Generation um einen konstanten Wert, $c$, ansteigt. Damit sind nur $c$ und ein Startwert $c_0$ vorzugeben, so daß das Verfahren wesentlich leichter handzuhaben ist als die beiden zuzuvor genannten.

Beim **Record-To-Record-Travel** kommt man sogar mit nur einem Parameter, $c$, aus. Eine Lösung wird deterministisch dann akzeptiert, wenn ihr Interessantheitsgrad um nicht mehr als $c$ Einheiten unter dem bisher besten Interessantheitsgrad liegt. Hier wird also die aktuelle Bewertung nicht, wie beim Simulated Annealing und beim Threshold Accepting, mit der jeweils letzten verglichen, sondern mit der besten überhaupt erreichten.

Während die beiden erstgenannten Ansätze, wie gesagt, aufgrund der großen Anzahl einzustellender Parameter schwer handzuhaben sind, muß die Entscheidung zwischen den beiden letztgenannten Akzeptanzkriterien im Zweifel experimentell getroffen werden.

#### *2.2.3.6 Die Auswahl der neuen Lösung*

Nach der Einschränkung der Menge der Lösungen auf eine verkleinerte Auswahlliste, $AL'$, gilt es, aus $AL'$ eine Lösung auszuwählen, die in die nächste Generation übernommen wird. Dabei kann eine Lösung, $s$, wie gesagt, durchaus eine sog. „Population" einzelner Lösungen, $s = (s_1, ..., s_{Pop})$', darstellen (mit $Pop$: Anzahl der einzelnen Lösungen in der Population), so daß mehrere Lösungswege parallel beschritten werden können.

Für die Auswahl der neuen Lösung (bzw. der neuen Population) sind zwei sinnvolle Alternativen zu nennen:

⇨ die deterministische Auswahl einer Lösung,

⇨ die stochastische Auswahl einer Lösung.

Die **deterministische Auswahl einer Lösung** selektiert die Lösung mit dem besten Interessantheitsgrad aus der eingeschränkten Auswahlliste, $AL'$.

Die **stochastische Auswahl einer Lösung** selektiert eine Lösung aus der eingeschränkten Auswahlliste, $AL'$, mit einer bestimmten Wahrscheinlichkeit. Nach der Art der Bestimmung dieser Wahrscheinlichkeit können folgende Varianten unterschieden werden:

- ⇨ Die Auswahlwahrscheinlichkeit ist über die Lösungen der Auswahlliste, $AL'$, identisch verteilt. Die gleichberechtigte Auswahl verringert die Wahrscheinlichkeit, daß der Suchprozeß in einen Zyklus gerät, aus dem er nur noch mit geringer Wahrscheinlichkeit entkommt.
- ⇨ Die Auswahlwahrscheinlichkeit ist eine monoton steigende Funktion des Interessantheitsgrades der Lösungen in der Auswahlliste, $AL'$. Diese Variante bevorzugt bessere Lösungen, kann aber eher dazu führen, daß der Suchprozeß in einen stochastischen Zyklus gerät.

#### *2.2.3.7 Kriterien zum Abbruch der Suche*

Eigentlich ist der Abbruch durch das Erreichen eines Zielzustandes (beim Data Mining als Optimierungsproblem nach Definition 2-7) bzw. aller Zielzustände aus der Lösungsmenge $L^*$ (beim Data Mining als Suchproblem nach Definition 2-6) genau bestimmt. Da Data-Mining-Verfahren keine exakten Lösungsverfahren sind, können sie nicht feststellen, ob dieses Kriterium erfüllt ist. Die einzige Ausnahme stellt der Sonderfall dar, daß der gesamte Suchraum durchforstet werden konnte. Da dieser Fall bei Problemstellungen realistischer Größenordnungen nicht anzutreffen ist, wird im Zusammenhang mit heuristischen Suchstrategien ein Ersatz-Abbruchkriterium benötigt. Derartige Ersatz-Abbruchkriterien sind beispielsweise:

- ⇨ das Erreichen einer vorgegebenen Anzahl von $gmax$ Generationen;
- ⇨ das Verstreichen einer vorgegebenen Anzahl von $gmax^{oV}$ Generationen ohne Verbesserung der Lösungsgüte;
- ⇨ das Erreichen oder Überschreiten eines vorgegebenen Interessantheitsniveaus, $ig^{min}$;
- ⇨ das Absinken des Grenznutzens der Suche, $\Delta ig/\Delta g$, unter ein vorgegebenes Niveau, $n^{min}$. Dabei stellt $\Delta ig$ die Veränderung des besten erreichten Interessantheitsgrades über $\Delta g$ Generationen dar.

### 2.2.4 Die Bewertung von Datenmustern und Datenmuster-Mengen

Die Notwendigkeit zur Bewertung von Modellen im allgemeinen wurde bereits in Abschnitt 2.1.1 verdeutlicht. Die Bewertung von Datenmustern der Form $(Pr \rightarrow Ko)$ erfolgt u.a. anhand der Objekte in der Trainingsmenge, $O^T$. Um daraus die benötigten Teilmengen bestimmen zu können, werden hier zunächst der Selektions- und der Projektionsoperator definiert:[136]

**Definition 2-50: Projektion**

Gegeben sei eine Objektmenge, $O'$, und eine Attributmenge, $A'$. Dann ist die *Projektion auf eine Attributmenge*, $B=\{b_1, ..., b_{bmax}\} \subseteq A'$, definiert als:

$$O'[B] := \{(b_1(o), ..., b_{bmax}(o)) \mid o \in O'\}. \qquad \Diamond$$

**Definition 2-51: Selektion**

Gegeben sei eine Objektmenge, $O'$, ein Term in konjunktiver Normalform, $Te$, und ein entsprechendes Konzept gemäß Definition 2-16, $c_{Te}$. Dann ist die *Selektion der Objekte*, für die $Te$ gilt, definiert als:

$$O'[Te] := \{o \in O \mid c_{Te}(o)=1\}. \qquad \Diamond$$

Vor allem werden die folgenden Selektionen aus $O^T$ benötigt:

| | | |
|---|---|---|
| $O^T[Pr]$ | $= \{o \in O^T \mid c_{Pr}(o)=1\}$ | die Menge der Objekte aus $O^T$, die die Prämisse, $Pr$, erfüllen; |
| $O^T[Ko]$ | $= \{o \in O^T \mid c_{Ko}(o)=1\}$ | die Menge der Objekte aus $O^T$, die die Konklusion, $Ko$, erfüllen; |
| $O^T[Pr \wedge Ko]$ | $= \{o \in O^T \mid c_{Pr \wedge Ko}(o)=1\}$ | die Menge der Objekte aus $O^T$, die Prämisse und Konklusion erfüllen. |

Die wichtigsten in der Literatur definierten oder verbal genannten Bewertungskonzepte sind in Abbildung 2-10 zusammengestellt.[137] Der Abbildung kann entnommen werden, in welchem Abschnitt die Interessantheitsfacetten und ihre Teilkomponenten behandelt

---

[136] Diese Definitionen erfolgen in Anlehnung an SCHLAGETER/STUCKY (1983), S. 139.

[137] Vgl. CLARK/NIBLETT (1989), S. 262, S. 270 zur Korrektheit, Reliabilität und Einfachheit, KRABS (1994), S. 42 ff. zur Stärke des Zusammenhangs, CHMIELEWICZ (1979), S. 129 ff. zur Reliabilität, zum Informationsgehalt und zur Neuheit, MICHALSKI/STEPP (1983a), S. 344 f. zur Genauigkeit, Homogenität und Heterogenität und MÜLLER/HAUSDORF/SCHNEEBERGER (1998), S. 255 ff. zur Nützlichkeit.

und größtenteils formal definiert werden und welche mathematischen Symbole die Facetten bezeichnen. Ein hochgestelltes $^{M}$ markiert Facetten, die ein ganzes Modell bewerten. Die übrigen Facetten bewerten einen einzelnen Term oder ein einzelnes Datenmuster. Teilweise haben die zu definierenden Symbole keine weitere Bedeutung, sondern dienen nur dazu, die Definition einer übergeordneten Interessantheitsfacette übersichtlicher zu gestalten. Solche Beziehungen zwischen zwei Facetten sind durch einen gestrichelten Pfeil gekennzeichnet.

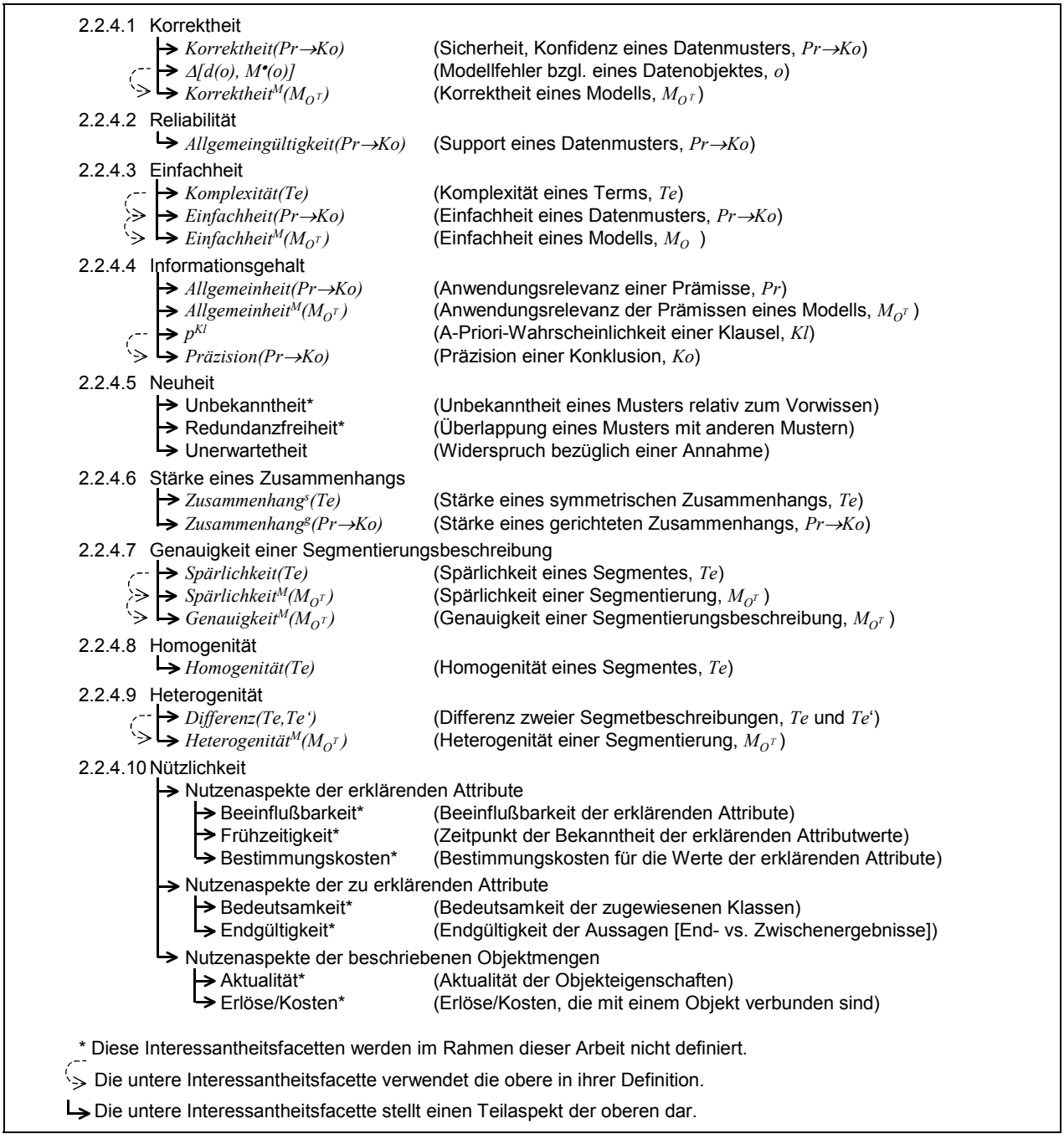



**Abbildung 2-10: Facetten der Interessantheit**

### *2.2.4.1 Die Bewertung der Korrektheit*

Die *Korrektheit* gibt an, inwieweit ein induktiv erlerntes Modell den Beobachtungen der Trainingsmenge entspricht. Sie kann für einzelne Datenmuster und für die gesamte Datenmuster-Menge definiert werden.

**Definition 2-52: Korrektheit (Sicherheit, Konfidenz) eines Datenmusters**

$A$ sei die Menge der beobachteten Attribute, $D$ sei eine Menge von zu erklärenden, $C$ eine Menge von erklärenden Attributen mit $D,C \subset A,\ C \cap D = \emptyset,\ C,D \neq \emptyset$. $DM^{KNF}(C,D)$ sei die Menge der möglichen Datenmuster in KNF bzgl. $C$ und $D$ und $\boldsymbol{R}$ die Menge der reellen Zahlen. Dann ist die *Korrektheit (Sicherheit, Konfidenz)* eines Datenmusters definiert als:[138]

$$\begin{aligned} &\textit{Korrektheit}: \quad DM^{KNF}(C,D) \quad \rightarrow \{r \mid r \in \boldsymbol{R},\ 0 \leq r \leq 1\} \\ &\qquad\qquad (Pr \rightarrow Ko) \qquad \rightarrow Korrektheit(Pr \rightarrow Ko) := \frac{|\, O^T\,[Pr \wedge Ko]\,|}{|\, O^T\,[Pr]\,|}. \end{aligned} \qquad \Diamond$$

Die Korrektheit kann als bedingte relative Häufigkeit interpretiert werden, daß ein Objekt die Konklusion erfüllt, wenn bekannt ist, daß es die Prämisse erfüllt. Bei einer Korrektheit von $1$ sagt man, die Prämisse, $Pr$, sei eine „***hinreichende***“[139] Bedingung für das Vorliegen der Konklusion, $Ko$.

Um die Korrektheit eines Datenmusters auf ein Modell erweitern zu können, wird zunächst folgende Definition benötigt, welche den Fehler zwischen dem Output eines Modells und dem tatsächlichen Wert der Outputvariablen quantifiziert. Beschränkt man den Output zunächst auf eine Variable, $d$, so kann der Modellfehler bezüglich eines Datenobjekts, $\Delta[d(o),\ Akkumulation(M_{O^T},\ o)]$, in Abhängigkeit des Skalenniveaus der abhängigen Variable, $d$, wie folgt definiert werden:

**Definition 2-53: Modellfehler bezüglich eines Datenobjekts**

Es sei $d$ eine zu erklärende Variable mit dem Wertebereich $dom(d)$ und $M^{\bullet}$ ein funktionales Data-Mining-Modell gemäß Definition 2-22. Dann betrage der *Modellfehler*, $\Delta[d(o),\ M^{\bullet}(o))]$, bezüglich eines Datenobjekts, $o \in O$:

---

[138] Vgl. YAO/ZHONG (1999), S. 484.

[139] Vgl. HOLSHEIMER/SIEBES (1994), S. 28. „***Notwendig***“ ist eine Prämisse, $Pr$, für das Vorliegen einer Konklusion, $Ko$, genau dann, wenn $|O^T[Pr \wedge Ko]| \,/\, |O^T[Ko]| = 1$.

$$\Delta[d(o), M^{\bullet}(o)] := \begin{cases} \left(d(o) - M^{\bullet}(o)\right)^2 & \text{falls } d \text{ kardinal ist;} \\ 1 & \text{falls } d \text{ nominal ist und } d(o) = M^{\bullet}(o); \\ 0 & \text{falls } d \text{ nominal ist und } d(o) \neq M^{\bullet}(o); \\ \left(R(d(o)) - R(M^{\bullet}(o))\right)^2 & \text{falls } d \text{ ordinal ist;} \end{cases}$$

wobei $R(\bullet)$ den Rang eines Ordinalwertes, $\bullet$, darstellt. ◊

Bei der Verwendung des Rangs wird unterstellt, daß die Abstände zwischen aufeinanderfolgenden Rangwerten identisch sind. Dies ist zwar gängige Praxis, aber streng genommen nur für Intervallskalen zulässig – ordinale Skalen zeichnen sich dadurch aus, daß keine Abstände definiert sind.[140] Streng genommen bleibt zur Berechnung des Modellfehlers nur die Möglichkeit, ordinale wie nominale Skalen zu behandeln und damit einen Informationsverlust in Kauf zu nehmen.

Will man sich nicht auf eine Outputvariable beschränken und $D = \{d_1, ..., d_{admax}\}$ zulassen, so müssen die Modellfehler für jede Outputvariable, $\Delta[d_1(o), M^{\bullet}(o))], ..., \Delta[d_{admax}(o), M^{\bullet}(o))]$, berechnet, gewichtet und aufaddiert werden, um ein Maß des Gesamtfehlers bezüglich eines Datenobjektes zu erhalten. Dies geschieht in der folgenden Definition:

**Definition 2-54: Korrektheit eines Modells**

Gegeben seien eine Trainingsmenge, $O^T = \{o_1, ..., o_N\}$, eine Attributmenge $A$, eine Menge zu erklärender Attribute, $D \subset A$, ein Lösungsraum, $L$, die Menge der reellen Zahlen, $\boldsymbol{R}$, und eine Abstandsfunktion, $\Delta$. Für die zu erklärenden Attribute, $D = \{d_1, ..., d_{admax}\}$, existieren die Gewichte $w_1, ..., w_{admax} \in \boldsymbol{R}$ und je eine Akkumulationsfunktion gemäß Definition 2-21, $Akkumulation_j$ $(j = 1, ..., admax)$. Dann ist die *Korrektheit eines Modells*, $M_{O^T} \in L$, als Negation des mittleren gewichteten Modellfehlers definiert:[141]

$$\begin{aligned} Korrektheit^M: \quad L &\rightarrow \{r \mid r \in \boldsymbol{R},\ r \leq 0\} \\ M_{O^T} &\rightarrow Korrektheit^M(M_{O^T}) := \\ &\quad -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{admax} w_j \cdot \Delta[d_j(o_i),\ Akkumulation_j(M_{O^T},\ o_i)]. \end{aligned}$$ ◊

[140] Vgl. BACKHAUS ET AL. (2000), S. XIX.

[141] Vgl. in der ungewichteten Variante: FRIEDMAN (1994), S. 7. Der Modellfehler geht negiert in die Korrektheit ein, um mit Definition 2-7 konsistent zu bleiben, wo das Data Mining als Maximierungsproblem definiert wurde.

*Man betrachte Tabelle 2-11 als Beispiel zur Berechnung der Modellkorrektheit. In der Kopfzeile wurden aus Platzgründen folgende Abkürzungen verwendet:*

$Akku_1 := Akkumulation_1(M_{O^T}, o_i)$ ;

$Fehler_1 := \Delta[d_1(o_i), Akkumulation_1(M_{O^T}, o_i)]$ ;

$Akku_2 := Akkumulation_2(M_{O^T}, o_i)]$ ;

$Fehler_2 := \Delta[d_2(o_i), Akkumulation_2(M_{O^T}, o_i)]$ ;

*Als Gewichte wurden $w_1 = 1/1000$ und $w_2=1$ vorgegeben. Eine angemessene Gewichtung zu finden ist problematisch, da eine „natürliche" Gewichtung i.d.R. fehlt. Beispielsweise ist es zumeist unangemessen, das Gewicht der $i$-ten Outputvariable, $w_i$, durch $w_i := 1/emax_i$ zu bestimmen (mit $emax_i$: maximal möglicher Fehler bei der $i$-ten Outputvariable; im Beispiel mit $dom(d_1) = \{r \mid r \in \mathbf{R}, 0 \le r \le 100\}$: $emax_1 = 100^2 = 10.000$), da die resultierenden Werte fast immer viel zu klein wären (im Beispiel etwa: $25/10000 = 0{,}0025$; $100/10000 = 0{,}01$) – gerade im Vergleich zu Fehlerwerten von $0$ oder $1$ bei nominalen Attributen.*

| Objekt | $d_1$ | $Akku_1$ | $Fehler_1$ | $d_2$ | $Akku_2$ | $Fehler_2$ | $1/1000 \cdot Fehler_1 + 1 \cdot Fehler_2$ |
|---|---|---|---|---|---|---|---|
| $o_1$ | 10 | 15 | $5^2$ | A | A | 0 | 25/1000 + 0 |
| $o_2$ | 20 | 0 | $20^2$ | A | B | 1 | 400/1000 + 1 |
| $o_3$ | 70 | 60 | $10^2$ | B | B | 0 | 100/1000 + 0 |
| $o_4$ | 90 | 85 | $5^2$ | A | A | 0 | 25/1000 + 0 |
| | | | | | | $-\frac{1}{4}\sum_{i=1}^{4}\sum_{j=1}^{2} w_j \cdot \Delta[d_j(o_i), Akkumulation_j(M_{O^T}, o_i)]$ = | – (0,025 + 1,4 + 0,1 + 0,025) / 4 = –0,3875 |

***Tabelle 2-11: Berechnung der Modellkorrektheit***

### *2.2.4.2 Die Bewertung der Reliabilität*

Die *Reliabilität* gibt an, inwieweit ein induktiv erlerntes Modell neuen Beobachtungen, die nicht der Trainingsmenge angehören, entspricht – m.a.W.: inwieweit es allgemeine Gültigkeit besitzt. Die Reliabilität testet man i.d.R., indem man das Modell anhand einer Menge von Testdaten, die dieselbe Struktur wie die Trainingsdaten (vgl. Definition 2-2) besitzt, validiert.[142] Die quantitative Validierung erfolgt dann durch die Berechnung der Korrektheit, wobei in Definition 2-54 die Trainingsmenge, $O^T$, durch die Evaluierungsmenge, $O^E$, zu ersetzen ist. Diese Evaluierungsmenge darf erst nach Abschluß der Data-Mining-Phase und nur ein einziges Mal verwendet werden.[143] Würden anhand der Evaluierungsdaten immer wieder Verfahrens- oder Parameteranpassungen vorgenommen, so würden Informationen genutzt, die eigentlich unbekannt sind. Auf diese Weise könnte nicht verhindert werden, daß das Verfahren bzw. die Parameterkonfiguration an die aktuelle Datensituation „überangepaßt" würde – man spricht hier auch von dem

[142] Vgl. KEARNS/VAZIRANI (1994), S. 7.

[143] Vgl. PODDIG (1999), S. 432.

Problem des „***Overfitting***“ oder „***Overlearning***“.[144] Damit würde die Evaluierungs- zur Trainingsmenge umfunktioniert.

Ein Ansatz, schon während der Trainingsphase (Data Mining i.e.S.) mit einer gewissen Wahrscheinlichkeit reliable Modelle zu erzeugen, besteht darin, die Allgemeingültigkeit der an der Generierung des Modell-Outputs beteiligten Datenmustern zu prüfen.[145] Diese ist wie folgt definiert:

**Definition 2-55: Allgemeingültigkeit (Support) eines Datenmusters**

Es gelten dieselben Voraussetzungen wie in Definition 2-52. Dann ist die *Allgemeingültigkeit* (der *Support*) eines Datenmusters definiert als:[146]

$$\begin{aligned} &\textit{Allgemeingültigkeit:} \quad DM^{KNF}(C,D) \quad \rightarrow \{r \mid r \in \mathbf{R},\ 0 \le r \le 1\} \\ &\qquad\qquad\qquad\qquad (Pr \rightarrow Ko) \qquad \rightarrow Allgemeingültigkeit(Pr \rightarrow Ko) := \frac{|O^T[Pr \wedge Ko]|}{|O^T|}. \end{aligned} \qquad \Diamond$$

Mit der Anzahl der Beobachtungen, $|O^T[Pr \wedge Ko]|$, steigt die erreichbare Güte der Approximation des Realsystems durch das induzierte Modell.[147]

*Dies kann wie folgt verdeutlicht werden. Abbildung 2-11 zeigt links Beobachtungen, die folgenden Zusammenhang vermuten lassen:*

$WENN\ Alter \in [0;30]\ UND\ Einkommen \in [0;4000]\ DANN\ Käufer = +.$

*Der rechte Teil der Abbildung verdeutlicht, daß bei Verdoppelung der Anzahl der Trainingsobjekte möglicherweise folgender Zusammenhang induziert werden könnte:*

$WENN\ Alter \in [0;30]\ UND\ Einkommen \in [1000;4000]\ DANN\ Käufer = +.$

*Durch die links unten in dem Beobachtungsraum neu hinzugekommenen Objekte wurde erkannt, daß die ehemals modellierte untere Einkommensgrenze von $0$ falsch war, und die $0$ wurde durch eine $1000$ ersetzt. Natürlich ist auch nach der Verdoppelung der Beobachtungen noch nicht sichergestellt, daß der induzierte Zusammenhang mit der unteren Einkommensgrenze von $1000$ richtig ist, aber die Wahrscheinlichkeit ist höher als beim ersten Induktionsversuch mit der Hälfte der Beobachtungen.*

---

[144] Vgl. HOLTE (1993), S. 67. Mit „Overfitting“ bzw. „Overlearning“ bezeichnet man die Neigung bestimmter Lernverfahren, wie z.B. einfacher Entscheidungsbaumalgorithmen, Trainingsdaten „auswendig“ zu lernen, indem sie beispielsweise eine Regel pro Datensatz bilden. Ein derartiges Modell klassifiziert die Trainingsdaten zwar gut, ist aber nicht reliabel. PODDIG schlägt vor, eine zweite Validierungsmenge einzuführen und wenige Male für Parameteranpassungen zu verwenden. Die andere Validierungsmenge dient dann ausschließlich zum abschließenden Test auf Reliabilität. Fällt dieser Test schlecht aus, muß nicht nur das Modell, sondern die gesamte Studie verworfen werden. (Vgl. PODDIG (1999), S. 431 ff.)

[145] Vgl. FREITAS (2000), S. 67.

[146] Vgl. YAO/ZHONG (1999), S. 484.

[147] Vgl. FRIEDMAN (1994), S. 10.

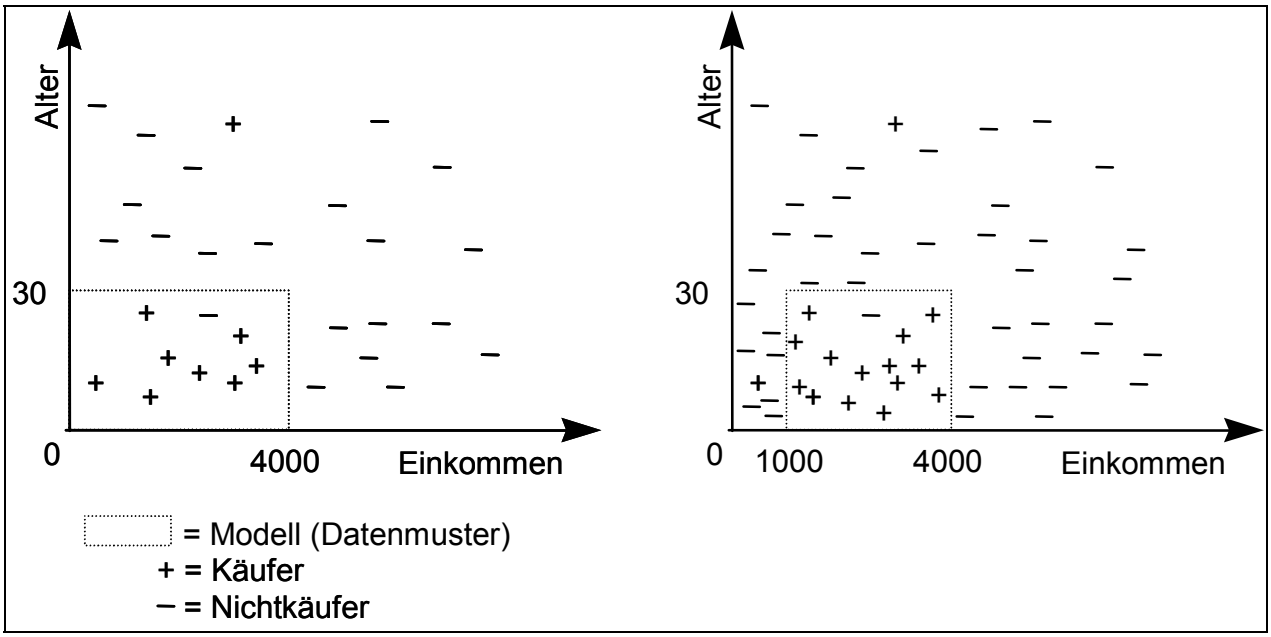



***Abbildung 2-11: Verbesserung der Induktion durch Verdoppelung der Anzahl an Beobachtungen***

### *2.2.4.3 Die Bewertung der Einfachheit*

Das Streben nach möglichst einfachen Modellen wird in der Literatur unter der Bezeichnung „***Occam's razor***" diskutiert. Gegenstand der Diskussionen sind folgende Effekte der Einfachheit:

⇨ Einfache Modelle, z.B. Entscheidungsbäume mit wenigen Knoten, sind tendenziell kompakt und leicht verständlich. So lautet die erste diskutierte Version von „Occam's razor":

*Stehen zwei gleich reliable Modelle zur Auswahl, so sollte das einfachere gewählt werden, da die Einfachheit selbst ein anzustrebendes Ziel darstellt.*[148]

⇨ Die zweite diskutierte – so nicht richtige – Version von „Occam's razor" lautet:

*Stehen zwei gleich korrekte Modelle zur Auswahl, so sollte das einfachere gewählt werden, da es wahrscheinlicher ist, daß dieses Modell einen wahren Sachverhalt der Realität abbildet.*[149]

Die Überlegung hinter dieser Aussage lautet wie folgt: In einem einfachen Modell umfasse jedes Datenmuster tendenziell mehr Objekte als dies bei einem komplexen

[148] Vgl. DOMINGOS (1998), S. 37.

[149] Vgl. DOMINGOS (1998), S. 37.

Modell der Fall wäre. Damit sei die Allgemeingültigkeit jedes einzelnen Datenmusters tendenziell höher als bei einem komplexen Modell. Die Forderung nach Einfachheit beuge dem bereits erwähnten „Overfitting“ vor.

Diese Version von „Occam’s razor“ ist nur unter der Voraussetzung korrekt, daß man aufgrund von Vorwissen glaubt, *daß das Realsystem in der benutzen Sprache tatsächlich durch ein einfaches Modell repräsentiert werden kann*.[150] Existieren unterschiedliche syntaktische Repräsentationsmöglichkeiten für dieselben semantischen Zusammenhänge, so kann diese zweite Version von „Occam’s razor“ auf triviale Weise erfüllt werden, indem man dem korrekteren Modell die syntaktisch einfachere Repräsentationsmöglichkeit zuweist.[151]

Aufgrund dieser Überlegungen wird hier mit der Einfachheit nur die in der ersten Version von „Occam’s razor“ genannte Zielsetzung verfolgt.

Die Einfachheit eines Datenmusters wird zumeist auf syntaktischer Ebene definiert, d.h. sie ist abhängig von der Syntax des verwendeten Repräsentationsformalismus’. Aus diesem Grund beziehen sich die folgenden Definitionen auf eine konkrete Sprache, und zwar auf Datenmuster in konjunktiver Normalform. Die Einfachheit eines Datenmusters, $Pr \rightarrow Ko$, setzt sich aus der Komplexität der Prämisse und der Konklusion zusammen. Dabei ist die Komplexität des Terms $Pr$ bzw. $Ko$ wie folgt definiert:

**Definition 2-56: Komplexität eines Terms in konjunktiver Normalform**

$A$ sei die Menge der Attribute in der Trainingsmenge und $Te^{KNF}(A)$ die Menge aller Terme in konjunktiver Normalform mit diesen Attributen. Dann ist die *Komplexität* eines Terms aus $Te^{KNF}(A)$, $Te = (Kl_1 \wedge ... \wedge Kl_{Klmax})$, wie folgt definiert.

$$\begin{aligned} Komplexität: \quad Te^{KNF}(A) &\rightarrow \{r \mid r \in \mathbf{R},\ 0 \le r \le 1\}; \\ Te &\rightarrow Komplexität(Te) := Klmax. \end{aligned}$$ ◊

Die Komplexität der einzelnen Klauseln muß nicht genauer differenziert werden, wenn man – wie dies in der weiteren Untersuchung geschieht – folgende Annahmen trifft:

[150] Vgl. DOMINGOS (1998), S. 41.

[151] Vgl. DOMINGOS (1998), S. 41.

⇨ Für nominale Klauseln, $(a = WM)$, wird festgesetzt, daß die Wertemenge $WM$ genau einen Wert enthält: $(a = w),\ w \in dom(a)$.

⇨ Nichtnominale Klauseln geben, wie bereits definiert wurde, ein Intervall an und sind damit von ähnlicher Einfachheit wie die vereinfachten nominalen Klauseln, welche genau einen Wert umfassen.

Aus der Komplexität der Prämisse und der Konklusion kann nun die Komplexität eines Datenmusters bestimmt werden:

**Definition 2-57: Einfachheit eines Datenmusters in konjunktiver Normalform**

Es gelten dieselben Voraussetzungen wie in Definition 2-54. $\boldsymbol{N}$ sei die Menge der natürlichen Zahlen, und $Prmax\_max$ und $Komax\_max$ seien Konstanten, die die maximal in der Prämisse bzw. in der Konklusion erlaubte Anzahl von Klauseln spezifiziert. Dann ist die *Einfachheit* eines Datenmusters definiert als:

$$
\begin{aligned}
Einfachheit:\quad DM^{KNF}(C,D) \;&\rightarrow\; \{r \mid r \in \boldsymbol{R},\ 0 \le r \le 1\};\\
(Pr \rightarrow Ko) \;&\rightarrow\; Einfachheit(Pr \rightarrow Ko) :=\\
&\qquad 1 - \frac{Komplexität(Pr) + Komplexität(Ko) - 2}{Prmax\_max + Komax\_max - 2};\\
&\qquad Prmax\_max,\ Komax\_max \in \boldsymbol{N};\\
&\qquad Prmax\_max > 1.
\end{aligned}
$$

◊

Die Konstante 2 wird jeweils subtrahiert, da sowohl die Konklusion als auch die Prämisse jeweils aus mindestens einer Klausel besteht.

*Beispielsweise ergeben sich für die Vorgaben $Prmax\_max := 5,\ Komax\_max := 2$ und die Komplexitätswerte aus Tabelle 2-12 die ebenfalls in der Tabelle aufgeführten Werte für den Zähler, den Nenner und den gesamten Ausdruck zur Berechnung der Einfachheit.*

| ***Komplexität(Pr)*** | *1* | *2* | *3* | *4* | *5* | *1* | *2* | *3* | *4* | *5* |
|---|---|---|---|---|---|---|---|---|---|---|
| ***Komplexität(Ko)*** | *1* | *1* | *1* | *1* | *1* | *2* | *2* | *2* | *2* | *2* |
| ***Komplexität(Pr) + Komplexität(Ko) – 2*** | *0* | *1* | *2* | *3* | *4* | *1* | *2* | *3* | *4* | *5* |
| ***Prmax_max + Komax_max – 2*** | *5* | *5* | *5* | *5* | *5* | *5* | *5* | *5* | *5* | *5* |
| ***Einfachheit(Pr→Ko)*** | *1* | *4/5* | *3/5* | *2/5* | *1/5* | *4/5* | *3/5* | *2/5* | *1/5* | *0* |

***Tabelle 2-12: Beispiele für die Berechnung der Einfachheit eines Datenmusters***

Überträgt man die Einfachheit auf die Bewertung einer Menge von Datenmustern, so empfielt sich folgende Definition:

**Definition 2-58: Einfachheit einer Menge von Datenmustern**

$A$ sei die Menge der Attribute in der Trainingsmenge, $L$ der Lösungsraum und $O^T$ die Trainingsmenge. Dann ist die *Einfachheit* einer Datenmuster-Menge, $M_{O^T}$ = *{(Pr1→Ko1), ..., (PrM→KoM)}*, definiert als:

$$Einfachheit^M: \quad L \rightarrow \{r \mid r \in \mathbf{R},\ 0 \le r \le 1\};$$

$$M_{O^T} \rightarrow Einfachheit^M(M_{O^T}) := 1 - \frac{1}{\left|\bigcup_{i=1}^{M} O^T[Pr_i]\right|} \cdot \sum_{i=1}^{M} \left(1 - Einfachheit(Pr_i \rightarrow Ko_i)\right). \quad \Diamond$$

Je mehr Regeln das Modell umfaßt, desto mehr Werte werden aufaddiert, so daß aufgrund des negativen Vorzeichens die Einfachheit umso kleiner wird. Die Einfachheit ist auf *[0;1]* normiert, da durch die (leicht modifizierte) Komplexität des komplexesten möglichen Modells dividiert wird. Das komplexeste mögliche Modell wurde in Definition 2-46 bzw. Definition 2-47 so konstruiert, daß für jedes Objekt eine eigene Regel existiert, so daß seine Komplexität genau der Anzahl der Objekte in der Trainingsmenge, $|O^T|$, entspricht. Da nicht unbedingt jedes Objekt durch eine Regel abgedeckt werden muß, wird hier nicht durch $|O^T|$, sondern durch die Anzahl der Objekte dividiert, die durch mindestens eine Regelprämisse erfasst werden:

$$\left|\bigcup_{i=1}^{M} O^T[Pr_i]\right|.$$

*Die Konsequenzen der getroffenen Definitionen lassen sich am besten anhand eines Beispiels verdeutlichen. Es seien drei Modelle zu bewerten, die inhaltlich identisch sind:*

*1.)* *WENN Alter ∈ [0;30] UND Einkommen ∈ [0;4000] DANN Käufer = +.*

*2.)* *WENN Alter ∈ [0;15] UND Einkommen ∈ [0;4000] DANN Käufer = +;*
*WENN Alter ∈ (15;30] UND Einkommen ∈ [0;4000] DANN Käufer = +.*

*3.)* *WENN Alter ∈ [0;15] UND Einkommen ∈ [0;2000] DANN Käufer = +;*
*WENN Alter ∈ (15;30] UND Einkommen ∈ [0;2000] DANN Käufer = +;*
*WENN Alter ∈ [0;15] UND Einkommen ∈ (2000;4000] DANN Käufer = +;*
*WENN Alter ∈ (15;30] UND Einkommen ∈ (2000;4000] DANN Käufer = +.*

*Weiter seien gegeben: Prmax_max := 2, Komax_max := 1 und $|O^T[Alter \in [0;30] \text{ UND } Einkommen \in [0;4000]]| = 100$. Jede einzelne Regel besitzt dann eine Einfachheit von $1-(2+1-2)/(2+1-2) = 0$. Damit ergeben sich für die erste Regelmenge eine Einfachheit von $1-1/100\cdot(1-0) = 99/100$, für die zweite eine Einfachheit von $1-1/100\cdot((1-0) + (1-0)) = 98/100$ und für die dritte eine Einfachheit von $1-1/100\cdot((1-0) + (1-0) + (1-0) + (1-0)) = 96/100$. Diese drei Werte unterscheiden sich nur wenig, so daß die Division durch die Komplexität des komplexesten möglichen Modells etwas überzogen erscheinen mag. Doch die Alternative zu dieser Definition der Einfachheit bestünde darin, gar keine Normierung vorzunehmen, da kein besserer Vergleichswert als der des komplexesten möglichen Modells zur Verfügung steht. Dann aber*

*könnte die Einfachheit nicht, wie beabsichtigt, durch Gewichtung mit anderen Interessantheitsfacetten zu einem Maximierungsziel zusammengeführt werden.*

#### *2.2.4.4 Die Bewertung des Informationsgehaltes*

Der ***Informationsgehalt*** einer Aussage gibt an, wieviele denkbare Sachverhalte durch die Aussage ausgeschlossen werden.[152]

*Man vergleiche beispielsweise die folgenden Aussagen:*

1. *WENN ein Kassenbon Cola und Bier enthält, DANN enthält er auch Chips;*
2. *WENN ein Kassenbon Pepsi-Cola und Krombacher Bier und Gerolsteiner Mineralwasser enthält, DANN enthält er auch Chips;*
3. *WENN ein Kassenbon Cola und Bier enthält, DANN enthält er auch Chips oder Salzstangen.*

*Da die erste Aussage mehr denkbare Sachverhalte ausschließt als die folgenden beiden Aussagen, sind deren Informationsgehalte geringer. So schließt Aussage 1 alle Kassenbons aus, die zwar Cola und Bier enthalten, aber keine Chips. Aussage 2 schließt dagegen nur diejenigen Kassenbons aus, die zwar ganz spezielle Cola- und Biersorten, aber keine Chips enthalten. Und Aussage 3 schließt nur diejenigen Bons aus, die zwar Cola und Bier enthalten, aber weder Chips noch Salzstangen. Aussage 1 besitzt eine allgemeinere Prämisse als Aussage 2 und eine präzisere Konklusion als Aussage 3.*

Das Beispiel macht deutlich, daß sich der Informationsgehalt aus zwei Komponenten zusammensetzt, die hier als *Allgemeinheit* und *Präzision* bezeichnet werden. Diese beiden Kriterien können wie folgt definiert werden:

**Definition 2-59: Allgemeinheit (Anwendungsrelevanz) eines Datenmusters**

Es gelten dieselben Voraussetzungen wie in Definition 2-52. Dann ist die *Allgemeinheit* bzw. *Anwendungsrelevanz* eines Datenmusters definiert als:

$$\begin{aligned} \textit{Allgemeinheit:} \quad & DM^{KNF}(C,D) \rightarrow \{r \mid r \in \mathbf{R},\ 0 \le r \le 1\}; \\ & (Pr \rightarrow Ko) \rightarrow Allgemeinheit(Pr \rightarrow Ko) := |O^T[Pr]| \,/\, |O^T|. \end{aligned} \qquad \Diamond$$

Die Allgemeinheit eines Datenmusters kann als relative Häufigkeit dafür, daß die Prämisse erfüllt ist, interpretiert werden. Übertragen auf eine Menge von Datenmustern kann deren Allgemeinheit als Anteil der durch mindestens eine Regel abgedeckten Objekte, relativ zu der Anzahl aller Objekte definiert werden:

---

[152] Vgl. CHMIELEWICZ (1979), S. 124.

**Definition 2-60: Allgemeinheit (Anwendungsrelevanz) einer Datenmuster-Menge**

Es gelten dieselben Voraussetzungen wie in Definition 2-58. Dann ist die *Allgemeinheit (Anwendungsrelevanz)* einer Datenmuster-Menge, $M_{O^T} = \{(Pr_1 \rightarrow Ko_1), \ldots, (Pr_M \rightarrow Ko_M)\}$, definiert als:

$$\begin{aligned} Allgemeinheit^M: \quad L &\rightarrow \{r \mid r \in \mathbf{R},\ 0 \le r \le 1\}; \\ M_{O^T} &\rightarrow Allgemeinheit^M(M_{O^T}) := \frac{1}{|O^T|}\left|\bigcup_{i=1}^{M} O^T[Pr_i]\right|. \end{aligned}$$ ◊

Nun zur zweiten Komponente des Informationsgehaltes, der Präzision. In dem einleitenden Beispiel wurde bereits gezeigt, daß eine Regel umso präziser ist, je spezieller die Konklusion ist, denn je spezieller die Konklusion ist, desto mehr denkbare Sachverhalte werden durch sie ausgeschlossen.

Zunächst wird die Wahrscheinlichkeit eingeführt, daß ein Sachverhalt durch eine (Konklusions-)Klausel erfaßt wird. Diese wird als „A-priori-Wahrscheinlichkeit“ bezeichnet und meint die Wahrscheinlichkeit ohne Kenntnis der Verteilung der Attributwerte über die Datenobjekte (bei Unterstellung einer Gleichverteilung). Sie ist wie folgt definiert:

**Definition 2-61: A-priori-Wahrscheinlichkeit für die Erfüllung einer Klausel**

Gegeben sei die Menge der natürlichen Zahlen, $\mathbf{N}$, sowie eine Klausel germäß Definition 2-8 bzw. Definition 2-9:

$$Kl^a = \begin{cases} (a \in \{w_1, \ldots, w_{wmax}\}) & \text{falls } a \text{ nominal;} \\ (a \in [w^{ug}; w^{og}]) & \text{sonst;} \end{cases}$$

$w_1, \ldots, w_{wmax} \in dom(a);$

$w^{ug} \le w^{og};\ w^{ug}, w^{og} \in dom(a);$

$wmax \in \mathbf{N}.$

Dann ist die *a-priori-Wahrscheinlichkeit* dafür, daß ein Datensatz die Klausel $Kl_a$ erfüllt, wie folgt definiert:

$$p^{Kl^a} := \begin{cases} \dfrac{wmax}{|dom(a)|} & \text{falls } a \text{ nominal;} \\ \dfrac{w^{og} - w^{ug}}{\max\limits_{w \in O^T[a]} w \ - \min\limits_{w \in O^T[a]} w} & \text{sonst.} \end{cases}$$ ◊

Damit kann nun die Präzision eines Datenmusters definiert werden als A-priori-Wahrscheinlichkeit, daß ein beliebiges Datenobjekt *nicht* von der Konklusion des Datenmusters erfaßt wird:

**Definition 2-62: Präzision eines Datenmusters**[153]
Es gelten dieselben Voraussetzungen wie in Definition 2-52. Weiterhin gebe $p^{Kl_i}$ die a-priori-Wahrscheinlichkeit dafür an, daß die $i$-te Klausel der Konklusion eines Datenmusters, $(Pr \rightarrow Ko) \in DM^{KNF}(C,D),\ Ko = (Kl_1 \wedge ... \wedge Kl_{Komax})$, erfüllt ist ($i = 1, ..., Komax$). Dann ist die *Präzision* wie folgt definiert:

$$\begin{aligned} Präzision:\quad & DM^{KNF}(C,D) \rightarrow \{r \mid r \in \mathbf{R},\ 0 \le r \le 1\}; \\ & (Pr \rightarrow Ko) \rightarrow Präzision(Pr \rightarrow Ko) := 1 - p^{Kl_1} \cdot \ldots \cdot p^{Kl_{Komax}}. \end{aligned}$$ ◊

### *2.2.4.5 Die Bewertung der Neuheit*

Die ***Neuheit*** sollte messen, wie neu die mit einer generierten Aussage verbundene Information für den Benutzer ist. Hier stellt sich die grundsätzliche Frage, ob die Suche nach vordefinierten Modelltypen überhaupt zu neuen Erkenntnissen führen kann. Eine Voraussetzung für einen Erkenntniszuwachs besteht darin, daß der Lösungsraum so groß ist, daß er potentiell neuartige Informationen umfaßt. Dies geschieht vor allem durch das Zulassen einer großen Anzahl von Attributen, die potentiell in einer Lösung vorkommen können. Zum anderen läßt man viele Freiheitsgrade in den Beziehungen der Attribute untereinander zu. Man gibt also den Modelltyp für das Data Mining grob, d.h. mit vielen Variablen, vor und optimiert die Werte dieser Variablen.

Bei der Durchforstung großer Suchräume werden i.d.R. sehr viele triviale oder bereits bekannte Aussagen produziert. Diese Aussagen müssen bei Anwendung eines Neuigkeitsmaßes eine schlechte Bewertung erhalten. Doch ein derartiges Maß läßt sich nur schwer implementieren, da sowohl die Trivialität von Aussagen als auch die Kenntnis des Benutzers gegenüber automatisch ermittelten Zusammenhängen nur durch eine inhaltliche Interpretation der Aussagen geprüft werden können. Daher wird für die

---

[153] Diese Definition ist angelehnt an die Definition des sog. „Fit"-Wertes nach MICHALSKI/STEPP (1983b), S. 398. Ein Maß für die Präzision einer Datenmuster-Menge wird in dieser Arbeit nicht benötigt, da in denjenigen Problemstellungen, in denen die Präzision eines Modells eine Rolle spielt, funktionale Modelle gemäß Definition 2-22 verwendet werden, so daß die Präzision durch eine genügend präzise Definition des Wertebereichs $dom(C)$ gewährleistet werden kann.

Interessantheitsbewertung teilweise auf Ersatzkriterien zurückgegriffen. So zerlegen MÜLLER, HAUSDORF und SCHNEEBERGER die Neuheit in die Facetten *Unbekanntheit, Redundanzfreiheit* und *Unerwartetheit*:[154]

- ⇨ Die ***Unbekanntheit*** mißt, wieviel Wissen bezüglich eines ermittelten Zusammenhangs bereits vorliegt. Die Auswertung dieser Facette erfordert damit ein Zugriff des Data-Mining-Verfahrens auf Vorwissen aus dem Benutzermodell.
- ⇨ Die ***Redundanzfreiheit*** gibt an, inwieweit sich eine Aussage mit bereits vorliegenden Aussagen überdeckt.

  *Beispielsweise überlappen sich die folgenden Aussagen:*

  1. `WENN Einkommen ∈ [3.000; 4.000] DANN Käufer von Produkt B = ja.`
  2. `WENN Einkommen ∈ [3.500; 4.500] DANN Käufer von Produkt B = ja.`
- ⇨ Die ***Unerwartetheit*** quantifiziert die Abweichung zwischen einer gemessenen Kennzahl über eine Aussage, wie z.B. $Korrektheit(Pr \rightarrow Ko)$, und der Erwartung des Benutzers, z.B. $E(Korrektheit(Pr \rightarrow Ko))$.

  Würde man, wie MÜLLER, HAUSDORF und SCHNEEBERGER unterstellen, a-priori für jede mögliche Aussage die Erwartung des Benutzers festlegen, so stünde dieser Aufwand in keinem Verhältnis zu dem Nutzen der Unerwartetheitsbewertung. Eher praktikabel wäre die Ermittlung des Erwartungswertes $E(Korrektheit(Pr \rightarrow Ko))$ aus früheren Data-Mining-Analysen. Damit wäre man auf den Vergleich identischer Aussagen über die Zeit beschränkt.

Während im Abschnitt 5.2.1 dieser Arbeit gezeigt wird, daß die Redundanzfreiheit leicht durch die Definition des Modelltyps gewährleistet werden kann und während die Unbekanntheit in Abschnitt 3.3.2.3 nachgereicht wird, soll an dieser Stelle ein Ansatz zur Formalisierung der Unerwartetheit nach PADMANABHAN und TUZHILIN widergegeben werden:[155]

**Definition 2-63: Unerwartetheit bezüglich einer Annahme**

Es gelten dieselben Voraussetzungen wie in Definition 2-52. Eine Regel, $(Pr \rightarrow Ko) \in DM^{KNF}(C,D)$, ist genau dann *unerwartet bezüglich einer Annahme*, $Pr^A \rightarrow Ko^A$, falls gilt:

---

[154] Vgl. MÜLLER/HAUSDORF/SCHNEEBERGER (1998), S. 255 f.

[155] Vgl. PADMANABHAN/TUZHILIN (1998), S. 95.

$Ko \wedge Ko^A = falsch;$

$Allgemeinheit(Pr \wedge Pr^A \rightarrow Ko) \geq min\_Allgemeinheit;$

$Korrektheit(Pr \wedge Pr^A \rightarrow Ko) \geq min\_Korrektheit;$

$min\_Allgemeinheit, min\_Korrektheit \in \{r \mid r \in \mathbf{R}, 0 \leq r \leq 1\}.$ ◊

*Beispielsweise wurde in einem Supermarkt bei einer Analyse bezüglich der Annahme*
`WENN Kinder = ja DANN Getränketyp = normale Getränke`
*die dieser Annahme widersprechende Aussage*
`WENN Kinder = ja UND Werbetyp = große Anzeige`
`DANN Getränketyp = Diätgetränke`
*gefunden.*[156]

Dieser Ansatz stellt zwar kein quantitatives Maß für die Unerwartetheit dar, kann aber als zusätzliche Nebenbedingung für extrahierte Regeln dienen. Dabei müssen die Erwartungen des Analytikers vorab in ein Benutzermodell eingegeben werden. Aufgrund der erforderlichen Vorgaben ist dieser Ansatz nur für Untersuchungen mit wenigen, sehr gezielt zu formulierenden Annahmen praktikabel und wird hier nicht weiter verfolgt.

#### *2.2.4.6 Die Bewertung der Stärke eines Zusammenhangs*

Die Stärke eines Zusammenhängs drückt aus, wie sehr sich mehrere Größen beeinflussen. Unterschieden werden sollen hier ungerichtete (symmetrische) und gerichtete (unsysmmetrische) Zusammenhänge. Bei einem symmetrischen Zusammenhang beeinflussen sich die beteiligten Größen gegenseitig. Bei einem gerichteten Zusammenhang werden abhängige und unabhängige Größen unterschieden. Für beide Typen von Zusammenhängen soll im folgenden je ein Zusammenhangsmaß definiert werden:

**Definition 2-64: Stärke eines symmetrischen Zusammenhangs**[157]

$A$ sei die Menge der beobachteten Attribute. $Te^{KNF}(A)$ sei die Menge aller Terme in konjunktiver Normalform mit diesen Attributen und $O^T$ die Trainingsmenge. Dann läßt sich die *Stärke eines ungerichteten Zusammenhangs* zwischen den Klauseln einer Segmentbeschreibung, $Te = (Kl_1, ..., Kl_{Klmax}) \in Te^{KNF}(A)$, wie folgt definieren:

$Zusammenhang^s: \quad Te^{KNF}(A) \quad \rightarrow \{r \mid r \in \mathbf{R}, 0 \leq r \leq 1\}$

---

[156] Vgl. PADMANABHAN/TUZHILIN (1998), S. 99.

[157] HILBERT definiert ein ähnliches Interessantheitsmaß, das allerdings nur die Differenz der entsprechenden Wahrscheinlichkeiten darstellt (vgl. HILBERT (1998), S. 83 ff.). Die hier eingeführte Division durch die relative Häufigkeit des Terms $Te$ bewirkt, daß nun die relative anstelle der absoluten Zunahme der relativen Häufigkeiten gemessen wird.

$$Te \rightarrow Zusammenhang^{s}(Te)$$

$$:= \begin{cases} 1 - \dfrac{\prod_{i=1}^{Klmax} \dfrac{|O^T[Kl_i]|}{|O^T|}}{\dfrac{|O^T[Te]|}{|O^T|}} & \text{falls} \prod_{i=1}^{Klmax} \dfrac{|O^T[Kl_i]|}{|O^T|} \leq \dfrac{|O^T[Te]|}{|O^T|}; \\ 0 & \text{sonst.} \end{cases} \quad \Diamond$$

Interpretiert man die in der Definition verwendeten relativen Häufigkeiten als Wahrscheinlichkeiten, so quantifiziert das Zusammenhangsmaß die relative Zunahme der Wahrscheinlichkeit für das Eintreten des Terms, $P(Te)$, im Vergleich zu der Wahrscheinlichkeit, die man bei Unabhängigkeit der einzelnen Klauseln erwarten würde, $P(Kl_1) \cdot \ldots \cdot P(Kl_{Klmax})$.

*Beispielsweise sei die in Tabelle 2-13 aufgeführte Trainingsmenge gegeben. Dann betragen die relativen Häufigkeiten für die Ereignisse* $Alter \in [0;30]$, $Einkommen \in [0;4000]$ *und* $Käufer = +$*:*
$|O^T[Alter \in [0;30]]| / |O^T| = 4/10 = 0{,}4;$
$|O^T[Einkommenr \in [0;4000]]| / |O^T| = 6/10 = 0{,}6;$
$|O^T[Käufer = +]| / |O^T| = 4/10 = 0{,}4.$
*Damit würde man bei Unabhängigkeit der einzelnen Ereignisse eine gemeinsame relative Häufigkeit von* $0{,}4 \cdot 0{,}6 \cdot 0{,}4 = 0{,}096$ *erwarten. Tatsächlich liegt diese relative Häufigkeit mit*
$|O^T[(Alter \in [0;30]) \wedge (Einkommenr \in [0;4000]) \wedge (Käufer = +)]| / |O^T| = 3/10 = 0{,}3$
*wesentlich über dem erwarteten Wert, was einen Zusammenhang zwischen den Ereignissen vermuten läßt. Die Stärke dieses (symmetrischen) Zusammenhangs beträgt:* $1 – 0096/0{,}3 = 0{,}68$.

| Objekt | Alter | Einkommen | Käufer |
|---|---|---|---|
| $o_1$ | 20 | 2200 | + |
| $o_2$ | 50 | 1240 | – |
| $o_3$ | 23 | 6485 | – |
| $o_4$ | 28 | 2745 | + |
| $o_5$ | 58 | 2673 | – |
| $o_6$ | 36 | 5743 | + |
| $o_7$ | 61 | 7584 | – |
| $o_8$ | 45 | 6846 | – |
| $o_9$ | 31 | 2640 | – |
| $o_{10}$ | 22 | 2040 | + |

***Tabelle 2-13: Trainingsmenge zur Berechnung der Stärke des Zusammenhangs***

Auch das Kriterium zur Messung der *Stärke eines gerichteten Ursache-Wirkungs-zusammenhangs* zwischen Zufallsvariablen baut auf dem Konzept der stochastischen

Unabhängigkeit auf. Für stochastisch unabhängige Zufallsereignisse, $Pr$ und $Ko$, gilt: $P(Ko|Pr) = P(Ko)$.[158] Interpretiert man den Unterschied zwischen $P(Ko|Pr)$ und $P(Ko)$ als Stärke der Abhängigkeit des Ereignisses $Ko$ von dem Ereignis $Pr$, so läßt sich eine Bewertungsfunktion als relative Zunahme der entsprechenden relativen Häufigkeiten definieren:[159]

**Definition 2-65: Stärke eines gerichteten Zusammenhangs**

Es gelten dieselben Voraussetzungen wie in Definition 2-52. Dann ist die *Stärke des gerichteten Zusammenhangs* eines Datenmusters, $(Pr \rightarrow Ko) \in DM^{KNF}(C,D)$, definiert als:

$$Zusammenhang^g: \quad DM^{KNF}(C,D) \rightarrow \{r \mid r \in \mathbf{R},\ 0 \le r \le 1\}$$

$$(Pr \rightarrow Ko) \rightarrow Zusammenhang^g(Pr \rightarrow Ko)$$

$$:= \begin{cases} 1 - \dfrac{\dfrac{|O^T[Ko]|}{|O^T|}}{\dfrac{|O^T[Pr \wedge Ko]|}{|O^T[Pr]|}} & \text{falls } \dfrac{|O^T[Ko]|}{|O^T|} \le \dfrac{|O^T[Pr \wedge Ko]|}{|O^T[Pr]|}; \\ 1 - \dfrac{1 - \dfrac{|O^T[Ko]|}{|O^T|}}{1 - \dfrac{|O^T[Pr \wedge Ko]|}{|O^T[Pr]|}} & \text{sonst.} \end{cases} \quad \Diamond$$

Im erstgenannten Fall gibt diese Interessantheitsfacette an, wieviel Prozent die relative Häufigkeit für das Eintreten der Konklusion, $Ko$, zunimmt, falls die Prämisse, $Pr$, eintritt. Nimmt die relative Häufigkeit für das Eintreten der Konklusion, $Ko$, ab, so kann man das Datenmuster als *negative Regel*

WENN $Pr$ DANN NICHT $Ko$

interpretieren, und an die Stelle von $(Pr \rightarrow Ko)$ tritt die Regel $(Pr \rightarrow \neg Ko)$.

Damit man überhaupt von einer Regel sprechen kann, sollte

$$Zusammenhang^{UW}(Pr \rightarrow Ko) \ge min\_Zusammenhang^{UW}$$

gelten (mit $min\_Zusammenhang^{UW} \in \{r \in \mathbf{R} \mid 0 < r \le 1\}$).

*Gegeben sei wieder die Trainingsmenge aus Tabelle 2-13. Zu bewerten sei der durch die folgende Regel repräsentierte gerichtete Zusammenhang:*

*WENN Alter $\in$ [0;30] UND Einkommen $\in$ [0;4000] DANN Käufer = +.*

---

[158] Vgl. BAMBERG/BAUR (1998), S. 88 f.

[159] Vgl. HILBERT (1998), S. 109, der wie oben die absolute Differenz zwischen den entsprechenden Wahrscheinlichkeiten als Zusammenhangsmaß definiert.

*Dann beträgt die relative Häufigkeit für das Ereignis $Käufer = +$ ohne weitere Annahmen:*
$|O^T[Käufer = +]| / |O^T| = 4/10 = 0{,}4.$
*Ist dagegen bekannt, daß das Alter zwischen $0$ und $30$ und das Einkommen zwischen $0$ und $4000$ liegt, so beträgt die relative Häufigkeit für das bedingte Ereignis $(Käufer = +) \mid ((Alter \in [0;30]) \wedge (Einkommenr \in [0;4000]))$:*
$|O^T[(Alter \in [0;30]) \wedge (Einkommenr \in [0;4000]) \wedge (Käufer = +)]| / |O^T[(Alter \in [0;30]) \wedge (Einkommenr \in [0;4000])]| = 3/3 = 1.$
*Dieser Wert liegt weit über der relativen Häufigkeit des unbedingten Ereignisses, so daß sich ein gerichteter Zusammenhang vermuten läßt. Dieser wird mit der Särke $1–0{,}4/1 = 0{,}6$ bewertet.*

### *2.2.4.7 Die Bewertung der Genauigkeit einer Segmentierungsbeschreibung*

Durch die Zielsetzung, eine Segmentierung möglichst genau beschreiben zu wollen, vermeidet man die Beschreibung von Teilsegmenten, die nur spärlich mit Objekten besetzt sind.

*Abbildung 2-12 zeigt den Beobachtungsraum zu den beobachteten Variablen „Geschlecht“ und „Einkommen“ mit ihren Ausprägungen. Die Punkte in diesem Raum stellen Käufer und Nichtkäufer eines bestimmten Produktes dar. Die Regionen der Männer mit hohem Einkommen und der Frauen mit niedrigem Einkommen sind extrem spärlich (nämlich gar nicht) besetzt. Man vergleiche nun die folgenden beiden Beschreibungen der Nichtkäufer:*
*1.)* `Einkommen = mittel UND Geschlecht = männlich.`
*2.)* `Einkommen ∈ {mittel,hoch} UND Geschlecht = männlich.`
*Beide Beschreibungen decken ausschließlich Nichtkäufer und keine Käufer ab, unterscheiden sich aber in ihrer Genauigkeit. So ist der zweite Term unangemessen allgemein, da er auch eine nicht besetzte Region des Merkmalsraums umfaßt. Der erste Term dagegen beschreibt genau die Nichtkäufer und umfaßt keine überflüssigen Regionen des Merkmalsraums.*

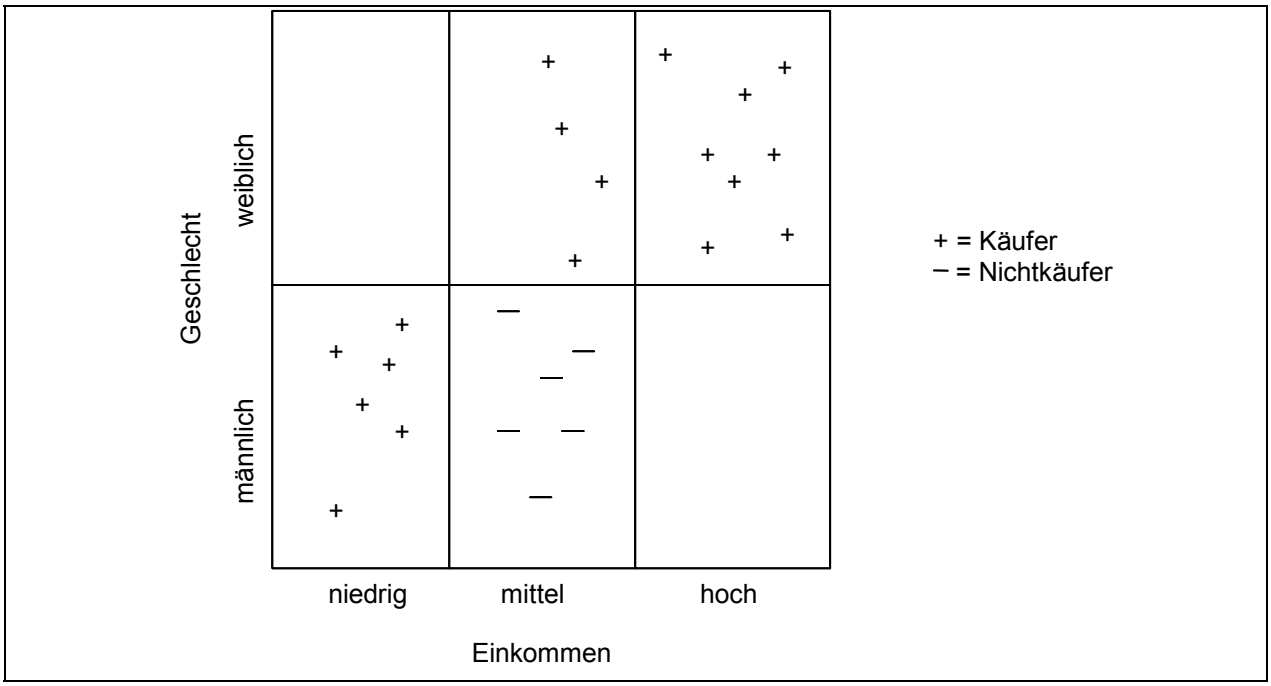



***Abbildung 2-12: Spärliche Besetzung zweier Regionen im Beobachtungsraum***

Theoretisch können $|dom(a_1)| \cdot ... \cdot |dom(a_{amax})|$ unterscheidbare[160] Objekte in dem durch die Attribute $A = \{a_1,...,a_{amax}\}$ aufgespannten Beobachtungsraum existieren. Diese Berechnung ist nur für endliche Wertebereiche sinnvoll. Ein durch $Te$ beschriebenes Segment ist spärlich besetzt, wenn es relativ wenige unterscheidbare Datenobjekte umfaßt. Die Anzahl der beobachteten und durch $Te$ beschriebenen Objekte erhält man durch Selektion der Objekte, die den Term, $Te$, erfüllen, aus der Trainingsmenge, $O^T$:

$|O^T[Te]|$.

Da die Trainingsobjekte auch durch die identifizierenden Schlüsselattribute, $SA$, charakterisiert sind, können in $O^T[Te]$ Objekte mit denselben Eigenschaften mehrfach auftreten. Die dadurch verursachte Mehrfach-Zählung kann durch eine Projektion auf die Nicht-Schlüsselattribute umgangen werden:

$|O^T[Te][a_1,...,a_{amax}]|$.

*Betrachtet man wieder die zwei Terme aus dem obigen Beispiel, so werden durch beide Terme jeweils sechs Objekte beschrieben, denn es gilt:*
$|O^T[(Einkommen = mittel) \wedge (Geschlecht = männlich)]| = 6;$
$|O^T[(Einkommen \in \{mittel, hoch\}) \wedge (Geschlecht = männlich)]| = 6.$
*Projiziert man die Trainingsmenge auf die Attribute Einkommen und Geschlecht, so sind die sechs Objekte nicht mehr unterscheidbar, da sie alle dasselbe Einkommen und dasselbe Geschlecht aufweisen, d.h. es gilt:*
$|O^T[(Einkommen = mittel) \wedge (Geschlecht = männlich)][Einkommen, Geschlecht]| = 1;$
$|O^T[(Einkommen \in \{mittel, hoch\}) \wedge (Geschlecht = männlich)][Einkommen, Geschlecht]| = 1.$
*Während beide Terme genau ein differenzierbares Objekt beschreiben, unterscheiden sie sich in der Anzahl der Objekte, die durch ihre Domänen potentiell differenzierbar wären, denn diese beträgt beim ersten Term: 1 (Männer mit mittlerem Einkommen), beim zweiten: 2 (Männer mit mittlerem Einkommen sowie Männer mit hohem Einkommen). Die Spärlichkeit eines Segmentes soll nun so definiert werden, daß sie im ersten Fall 0 beträgt (d.h. keine der beschriebenen Regionen ist unbesetzt) und im zweiten Fall den Wert 0,5 annimmt (d.h. die Hälfte der beschriebenen Regionen ist unbesetzt).*

Nach diesen Überlegungen wird die Spärlichkeit eines Segmentes in Anlehnung an MICHALSKI und STEPP wie folgt definiert:[161]

**Definition 2-66: Spärlichkeit eines Segmentes**

Es sei $A = \{a_1,...,a_{amax}\}$ die Menge der beobachteten Attribute (ohne die Schlüsselattribute) mit endlichen Wertebereichen, $dom(a_i)$ für $i = 1, ..., amax$. Weiterhin seien $Te^{KNF}(A)$ die Menge aller möglichen Terme in konjunktiver Normalform bezüglich dieser Attribute

[160] „Unterscheidbar“ bezieht sich gemäß Definition 2-29 auf die Merkmale, $A$, nicht auf die Schlüssel, $SA$.

[161] Vgl. MICHALSKI/STEPP (1983a), S. 339.

und $O^T$ die Menge der beobachteten Objekte. Dann ist die *Spärlichkeit* eines Segmentes $O^T[Te]$, $Te \in Te^{KNF}(A)$, wie folgt definiert.

$$\begin{aligned}
\textit{Spärlichkeit}: \quad & Te^{KNF}(A) \rightarrow \{r \mid r \in \mathbf{R},\ 0 \le r \le 1\} \\
& Te \rightarrow \textit{Spärlichkeit}(Te) := 1 - \frac{|O^T[Te][a_1,...,a_{amax}]|}{\prod_{a \in A - A'} |dom(a)| \cdot \prod_{a \in A'} |WM_a|} \text{ mit} \\
& A' = \{a'_1,...,a'_{Klmax}\}; \\
& A' \subseteq A; \\
& Te = (Kl_{a'_1} \wedge ... \wedge Kl_{a'_{Klmax}}); \\
& Kl_{a'_j} = (a'_j \in WM_{a'_j}); \\
& j = 1,...,Klmax.
\end{aligned}$$ ◊

Die Spärlichkeit stellt die relative Anzahl der unbeobachteten Objekte in dem durch $Te$ charakterisierten Segment dar. Relativiert wird diese Anzahl durch die Anzahl aller unterscheidbaren Objekte in diesem Segment. Dabei können die Objekte sowohl durch Attribute aus der Segmentbeschreibung, $A'$, als auch durch die übrigen Attribute, $A–A'$, unterschieden werden. MICHALSKI und STEPP interpretieren diesen Wert als *Grad der Generalisierung* der Segmentbeschreibung $Te$.[162]

*Beispielsweise ergeben sich für Abbildung 2-12, $A$ = {Geschlecht, Einkommen} und das Segment* `Einkommen ∈ {mittel, hoch}:`
$|O^T[Einkommen \in \{mittel, hoch\}]| = |\{(weiblich, mittel), (weiblich, hoch), (männlich, mittel)\}| = 3$,
$|dom(Geschlecht)| = 2$,
$|WM_{Einkommen}| = |\{mittel, hoch\}| = 2$
*und somit eine Spärlichkeit von $1 – 3/(2 \cdot 2) = 1/4$, und für das Segment* `Geschlecht = weiblich:`
$|O^T[Geschlecht = weiblich]| = |\{(weiblich, mittel), (weiblich, hoch)\}| = 2$,
$|dom(Einkommen)| = 3$,
$|WM_{Geschlecht}| = |\{weiblich\}| = 1$
*und somit eine Spärlichkeit von $1 – 2/(3 \cdot 1) = 1/3$.*

Da eine Lösung im Data Mining eine Menge von Datenmustern darstellt, ist die o.g. Definition der Spärlichkeit auf eine Menge von Segmenten zu übertragen. Dies geschieht ebenfalls in Anlehnung an MICHALSKI und STEPP:[163]

[162] Vgl. ebenda.

[163] Vgl. MICHALSKI/STEPP (1983a), S. 339.

**Definition 2-67: Spärlichkeit einer Segmentierung**

Es sei $L$ der Lösungsraum und $O^T$ die Menge der beobachteten Objekte. Jedes Element des Lösungsraums, $M_{O^T} = \{(Pr_1 \rightarrow Ko_1),...,(Pr_M \rightarrow Ko_M)\}$, sei eine Menge von Datenmustern, deren Prämissen eine Segmentierung beschreiben. Dann ist die Spärlichkeit der so charakterisierten Segmentierung wie folgt definiert.

$$\begin{aligned} Spärlichkeit^M: \quad L &\rightarrow \{r \mid r \in \mathbf{R},\ 0 \le r \le 1\}; \\ M_{O^T} &\rightarrow Spärlichkeit^M(M_{O^T}) := \frac{1}{M} \sum_{(Pr \rightarrow Ko) \in M_{O^T}} Spärlichkeit(Pr). \end{aligned} \qquad \Diamond$$

Die Spärlichkeit bezieht sich nur auf die Prämissen der Datenmuster. Sie eignet sich damit zur Bewertung von rein beschreibenden Modellen ohne abhängige Variablen.

Die Spärlichkeit der Segmentierung kann nun herangezogen werden, um die Genauigkeit der Beschreibung einer Menge von Objekten, $O^T$, zu messen:[164]

**Definition 2-68: Genauigkeit einer Segmentierungsbeschreibung**

Es gelten dieselben Voraussetzungen wie in Definition 2-67. Dann ist die Genauigkeit einer Segmentierungsbeschreibung, $M_{O^T} \in L$, wie folgt definiert.

$$\begin{aligned} Genauigkeit^M: \quad L &\rightarrow \{r \mid r \in \mathbf{R},\ 0 \le r \le 1\} \\ M_{O^T} &\rightarrow Genauigkeit^M(M_{O^T}) := 1 - Spärlichkeit^M(M_{O^T}). \end{aligned} \qquad \Diamond$$

### *2.2.4.8 Die Bewertung der Homogenität*

Eine Zielsetzung konventioneller Clusterverfahren besteht darin, einem Cluster (Segment) möglichst ähnliche Objekte zuzuordnen, wobei die Ähnlichkeit in Abhängigkeit von den Objekteigenschaften berechnet wird.[165] Alternativ dazu läßt sich die Homogenität eines Segmentes auch unter Einbezug der Segmentbeschreibung, $Te \in Te^{KNF}(A)$, definieren. Dabei setzt man voraus, daß die Segmentbeschreibung auf genügend Objekte zutrifft. So mißt STEPP die Homogenität der Objekte eines Segmentes als Anzahl der gemeinsamen Eigenschaften der Objekte in diesem Segment:[166]

---

[164] Vgl. MICHALSKI/STEPP (1983a), S. 344. Dort wird die Genauigkeit allerdings zwischen – 1und 0 gemessen.

[165] Vgl. MICHALSKI/STEPP (1983b), S. 396.

[166] Vgl. STEPP (1984), S. 49.

$Homogenität(Te) := Klmax$ mit $Te = (Kl_1 \wedge ... \wedge Kl_{Klmax})$.

Dieser Ansatz ist auf der einen Seite plausibel, da eine Segmentbeschreibung, welche viele Klauseln umfaßt, tendenziell ein homogeneres Segment beschreibt als eine Beschreibung mit wenigen Klauseln.

*Beispielsweise ist das Segment, das durch den Term*
`(Geschlecht = weiblich) UND (Einkommen = hoch) UND (Alter = hoch) UND (regionale Lage = Speckgürtel) UND (Beruf = Hausfrau)`
*beschrieben wird, homogener als ein Segment, das nur durch die folgenden Eigenschaften beschrieben wird:*
`(Geschlecht = weiblich) UND (Einkommen = hoch).`

STEPPs Konzeption der Homogenität berücksichtigt nicht, daß die einzelnen Klauseln einer Segmentbeschreibung ganz unterschiedlich präzise formuliert sein können.

*Beispielsweise würde in dem Segment, das durch den Term*
`(Geschlecht ∈ {weiblich, männlich}) UND (Einkommen = hoch)`
*beschrieben wird, die erstgenannte Klausel überhaupt keine Einschränkung der erfaßten Personen bedeuten und damit auch keinen Beitrag zur Homogenität dieses Segmentes leisten.*

Eine Beurteilung der Homogenität sollte demnach eher analog zur Bewertung der Präzision gemäß Definition 2-62 erfolgen. Damit kommt man zu folgender Definition:

**Definition 2-69: Homogenität eines Segmentes**

Es seien $A$ die Menge der beobachteten Attribute, $Te^{KNF}(A)$ die Menge aller möglichen Terme in konjunktiver Normalform bezüglich dieser Attribute und $\boldsymbol{R}$ die Menge der reellen Zahlen. Weiterhin gebe $p^{Kl_i}$ die a-priori-Wahrscheinlichkeit[167] dafür an, daß die $i$-te Klausel eines Terms $Te = (Kl_1 \wedge ... \wedge Kl_{Klmax}) \in Te^{KNF}(A)$, erfüllt ist ($i = 1, ..., Klmax$). Dann ist die *Homogenität* eines Segmentes, das durch $Te$ charakterisiert wird, wie folgt definiert:

$$\begin{aligned} Homogenität: \quad Te^{KNF}(A) &\rightarrow \{r \mid r \in \boldsymbol{R},\ 0 \le r \le 1\} \\ Te &\rightarrow Homogenität(Te) := 1 - p^{Kl_1} \cdot \ldots \cdot p^{Kl_{Komax}}; \\ &\quad \text{mit } Te = (Kl_1 \wedge ... \wedge Kl_{Klmax}). \end{aligned}$$ ◊

*Mit der vorgestellten Definition würde beispielsweise das obige Segment*
`(Geschlecht ∈ {weiblich, männlich}) UND (Einkommen = hoch)`

[167] Vgl. Definition 2-61.

*als weniger homogen bewertet, da die A-Priori-Wahrscheinlichkeit, daß das Geschlecht entweder männlich oder weiblich ist, eins beträgt. Bei drei möglichen Werten für das Einkommen ergäbe sich eine Homogenität von* $1-1\cdot1/3 = 2/3$. *Dagegen würde das Segment*

`(Geschlecht = weiblich) UND (Einkommen = hoch)`

*eine Homogenität von von* $1-1/2\cdot1/3 = 5/6$ *erhalten.*

#### *2.2.4.9 Die Bewertung der Heterogenität*

Konventionelle Clusterverfahren zielen darauf ab, möglichst unterschiedliche Cluster (Segmente) zu erzeugen, damit die Abgrenzung der Cluster möglichst trennscharf ist.[168] Dabei ergeben sich die Unterschiede der Cluster durch die Eigenschaften der Objekte, welche die Cluster bilden. Alternativ dazu läßt sich die Heterogenität einer Segmentierung auch unter Einbezug der Segmentierungsbeschreibung definieren. Dabei setzt man voraus, daß die Segmentierungsbeschreibung auf genügend Objekte zutrifft. Hierzu definieren MICHALSKI und STEPP zunächst die Differenz zwischen zwei Segmentbeschreibungen, $Te, Te' \in Te^{KNF}(A)$ als Anzahl der Klauseln, durch die sich die beiden Segmentbeschreibungen unterscheiden:[169]

$$Differenz(Te, Te') := Klmax + Klmax' - 2\cdot|\{Kl_1, ..., Kl_{Klmax}\} \cap \{Kl'_1, ..., Kl'_{Klmax'}\}|;$$

mit $Te=(Kl_1\wedge...\wedge Kl_{Klmax})$;

$Te'=(Kl'_1\wedge...\wedge Kl'_{Klmax'})$.

Auch hier ist – wie schon im Abschnitt zuvor – kritisch anzumerken, daß die bloße Anzahl der Klauseln wenig aussagekräftig ist, da sich die Wertebereiche einzelner Klauseln überlappen können, wie z.B. `Altere∈[10;20]` und `Altere∈[15;25]`. Die Differenz wird nun um diesen Aspekt erweitert:

**Definition 2-70: Differenz zweier Segmentbeschreibungen**

Es gelten dieselben Voraussetzungen wie in Definition 2-69. Dann ist die *Differenz zweier Segmentbeschreibungen*, $Te, Te' \in Te^{KNF}(A)$, wie folgt definiert:

$$\begin{aligned} Differenz: \quad & Te^{KNF}(A)\times Te^{KNF}(A) \rightarrow \{r \mid r\in\mathbf{R},\ 0\le r\le 1\} \\ & (Te, Te') \rightarrow Differenz(Te, Te') := \\ & \frac{Klmax + Klmax' - 2\cdot \sum_{a\in A^{gemeinsam}} Überlappung\left(Kl^a, Kl'^a\right)}{Klmax + Klmax'}; \end{aligned}$$

mit:

---

[168] Vgl. ESTER/SANDER (2000), S. 45.

[169] Vgl. MICHALSKI/STEPP (1983a), S. 345.

$Te = (Kl^{a_1}, \ldots, Kl^{a_{Klmax}})$;

$Te' = (Kl'^{a'_1}, \ldots, Kl'^{a'_{Klmax'}})$;

$A^{gemeinsam} = \{a \mid a \in \{a_1, \ldots, a_{Klmax}\}, a \in \{a'_1, \ldots, a'_{Klmax'}\}\}$;

$$\begin{aligned} \ddot{U}berlappung: \; Kl^{KNF}(A^{gemeinsam}) \times Kl^{KNF}(A^{gemeinsam}) &\rightarrow \{r \mid r \in \mathbf{R},\ 0 \leq r \leq 1\} \\ (Kl^a, Kl'^a) &\rightarrow \ddot{U}berlappung(Kl^a, Kl'^a) \\ &:= \frac{|WM \cap WM'|}{|WM \cup WM'|}; \end{aligned}$$

$Kl^a = (a = WM)$;

$Kl'^a = (a = WM')$. ◊

*Beispielsweise würde die Differenz der Terme* `Alter∈[10;20]` *und* `Alter∈[15;25]` *wie folgt bestimmt:*

*Differenz(Alter∈[10;20], Alter∈[15;25])*
*= (1 + 1 – 2·|[15;20]|/|[10;25]|) / (1 + 1)*
*= (1 + 1 – 2·5/15) / (1 + 1)*
*= (4/3) / 2*
*= 2/3.*

Ist die Differenz zwischen zwei Termen operationalisiert, läßt sich damit die Heterogenität einer Menge von Segmentbeschreibungen als Summe der Differenzen über alle Paare von Segmentbeschreibungen definieren:[170]

**Definition 2-71: Heterogenität einer Segmentierung**

Es gelten dieselben Voraussetzungen wie in Definition 2-67. Dann ist die *Heterogenität einer Segmentierung*, $M_{O^T} \in L$, wie folgt definiert.

$$\begin{aligned} Heterogenität^M: \; L &\rightarrow \mathbf{R} \\ M_{O^T} &\rightarrow Heterogenität(M_{O^T}) := \sum_{\substack{(Pr \rightarrow Ko),\\ (Pr' \rightarrow Ko') \in M_{O^T}}} Differenz(Pr, Pr'). \end{aligned}$$ ◊

Die Verwendung der Summe anstelle der durchschnittlichen Differenz führt dazu, daß die Heterogenität umso größer wird, je mehr Segmente es gibt. Dies ist plausibel, da eine sehr heterogene Objektmenge in viele Segmente zerlegt werden muß, wenn man eine Aufteilung in jeweils homogene Segmente erreichen möchte.

[170] Vgl. MICHALSKI/STEPP (1983a), S. 345.

#### *2.2.4.10 Die Bewertung der Nützlichkeit*

Problematisch stellen sich Versuche dar, die Nützlichkeit einer Aussage im Hinblick auf ihre praktische Verwendbarkeit zu bewerten. Die **Nützlichkeit** gliedert sich in mehrere Teilaspekte, von denen die folgenden eine gewisse Anwendungsunabhängigkeit aufweisen:[171]

- ⇨ Nutzenaspekte der erklärenden Attribute:
  - → Beeinflußbarkeit der erklärenden Attribute;
  - → Frühzeitigkeit der Bekanntheit der Werte der erklärenden Attribute;
  - → Bestimmungskosten für die Werte der erklärenden Attribute;
- ⇨ Nutzenaspekte der zu erklärenden Attribute:
  - → Bedeutsamkeit der zugewiesenen Klasse;
  - → Endgültigkeit der Aussagen (Zwischen- oder Endergebnisse);
- ⇨ Nutzenaspekte der beschriebenen Objektmengen:
  - → Aktualität der Objekteigenschaften;
  - → Erlöse bzw. Kosten, die mit einem Objekt verbunden sind.

MÜLLER, HAUSDORF und SCHNEEBERGER liefern nur für wenige Nützlichkeitsaspekte operationale Berechnungsvorschriften. Obwohl die genannten Kriterien plausibel sind, eignen sie sich nicht unbedingt zur Verwendung im Data Mining. Vor allem ist zu beachten, daß jede zusätzliche Interessantheitsfacette eine zusätzliche Zielsetzung für das zu lösende Optimierungsproblem darstellt. Gemäß Definition 2-7 bzw. Definition 2-5 wird zur Optimierung ein skalarer Interessantheitsgrad benötigt, d.h. Mehrfachzielsetzungen müssen aufgelöst werden.[172] Dies ist immer mit dem Problem verbunden, adäquate Gewichtungen für die Einzelziele zu finden. Allerdings zeigt die folgende Diskussion der einzelnen Nützlichkeitsaspekte, daß diese nicht unbedingt als eigene Zielsetzungen in das zu lösende Optimierungsproblem eingehen müssen:

---

[171] Vgl. MÜLLER/HAUSDORF/SCHNEEBERGER (1998), S. 257. Die Autoren beziehen sich mit ihren Facetten auf den medizinischen Bereich.

[172] Vgl. zur Auflösung von Mehrfachzielsetzungen: DINKELBACH (1982), S. 153 ff.

- ⇨ Die **Beeinflußbarkeit** der erklärenden Attribute kann bei der Vorauswahl der Attribute berücksichtigt werden. Unbeinflußbare Attribute werden nicht in die Trainingsmenge übernommen.
- ⇨ Auch die **Frühzeitigkeit** der Bekanntheit der Werte der erklärenden Attribute kann bei der Vorauswahl der Attribute berücksichtigt werden.
- ⇨ **Bestimmungskosten** für die Werte der erklärenden Attribute können ohnehin nur dann berücksichtigt werden, wenn sie sehr genau bekannt sind. Ansonsten würden sie die Modellbewertungen verzerren.
- ⇨ Die **Aktualität** der Objekteigenschaften in der Trainingsmenge unterscheidet sich nur dann, wenn die Datenbasis über längere Zeit zusammengetragen wurde. In diesem Fall sollte die Aktualität aber nicht als Interessantheitsfacette Berücksichtigung finden, sondern bei der Bestimmung des Modelloutputs. So könnten beispielsweise zur Bestimmung eines Prognosewertes – ähnlich wie bei dem bekannten Verfahren der exponentiellen Glättung[173] – aktuellere Daten stärker gewichtet werden als ältere.[174]
- ⇨ Die **Erlöse bzw. Kosten**, die mit einem Objekt verbunden sind, stellen aus betriebswirtschaftlicher Sicht eine besonders relevante Interessantheitsfacette dar und werden daher in der weiteren Untersuchung berücksichtigt.[175] Selbiges gilt für die **Bedeutsamkeit** der zugewiesenen Klasse und die **Endgültigkeit** der Aussagen, wenn sich diese Werte monetär quantifizieren lassen.

## 2.3 Voraussetzungen für die Anwendung von Data-Mining-Verfahren

Nachdem nun die technischen Grundlagen des Data Mining eingeführt wurden, kann zu dessen betriebswirtschaftlichen Anwendungspotentialen übergegangen werden. Die Brücke zur praktischen Anwendung von Data-Mining-Verfahren bilden die aus den technischen Grundlagen resultierenden Eignungskriterien. Dabei konzentrieren sich die

---

[173] Vgl. zur exponentiellen Glättung beispielsweise HANSEMANN (1995), S. 273.

[174] Vgl. NAKHAEIZADEH/TAYLOR/KUNISCH (1997), S. 131.

[175] Erlöse bzw. Kosten werden in Abschnitt 3.3.2 bei Erklärungs- und Beschreibungsmodellen als „Erfolgsbeitrag“ und bei Entscheidungsmodellen innerhalb eines erfolgsorientierten „Nutzwertes“ berücksichtigt.

Ausführungen auf die Eignungskriterien, die sich aus der Datenbasis des Anwendungsbereiches ergeben. Hier macht der Einsatz von Data-Mining-Methoden nur Sinn, wenn folgende Voraussetzungen erfüllt sind:

- ⇨ In der Datenbasis existiert potentiell eine **Menge unbekannter Zusammenhänge**, so daß der Einsatz von Data-Mining-Verfahren gegenüber einfacheren strukturprüfenden Datenanalysemethoden gerechtfertigt ist.[176] Eine notwendige Voraussetzung dafür ist, daß die Datenbasis viele Variablen bzw. Variablen mit großen Domänen umfaßt.

- ⇨ Die Datenbasis darf **nicht zu viele widersprüchliche**[177] **Datensätze** enthalten, da sonst sehr schnell zufällige Zusammenhänge statt der beabsichtigten Kausalzusammenhänge generiert werden. Die Anforderung ist umso eher erfüllt, je mehr erklärende Variablen zur Verfügung stehen, denn dann ist es umso unwahrscheinlicher, daß sich die Datensätze durch kein einziges Merkmal unterscheiden lassen. Ein einfaches Herausfiltern widersprüchlicher Datensätze löst das Problem nicht, wenn nicht ausgeschlossen werden kann, daß die herausgefilterten Datensätze durch das erlernte Modell erfaßt werden – andernfalls würden für diese ohne Datengrundlage erfaßten Objekte falsche Aussagen getroffen.

- ⇨ **Möglichst viele Einflußfaktoren** stehen als Attribute in der Datenbank zur Verfügung.[178] Jeder unbeobachtete Einflußfaktor verstärkt, wie in Abschnitt 2.1.1 erläutert wurde, das Datenrauschen und damit das zuvor genannte Problem widersprüchlicher Datensätze. Welche Kategorien von Variablen benötigt werden, hängt von der jeweiligen Problemsituation ab.

  *Beispielsweise reichen für Prognosen des Kundenverhaltens soziodemographische Merkmale i.d.R. nicht aus.*[179]

---

[176] Vgl. FAYYAD/PIATETSKY-SHAPIRO/SMYTH (1996), S. 24 f. und PIATETSKY-SHAPIRO ET AL. (1996), S. 94.

[177] Die Problematik der Widersprüche wurde im Zusammenhang mit Definition bereits für Muster in konjunktiver Normalform konkretisiert. Will man von der Einschränkung auf die konjunktive Normalform absehen und einen Widerspruch allgemein definieren, so daß die Definition bspw. auch für Regressionsmodelle gilt, so kann man zwei Objekte genau dann als „***widersprüchlich***" bezeichnen, wenn sie sich durch ihre erklärenden Merkmale kaum differenzieren lassen und sich in ihren zu erklärenden Merkmalen stark unterscheiden (sofern sich erklärendende und zu erklärende Merkmale unterscheiden lassen).

[178] Vgl. PIATETSKY-SHAPIRO ET AL. (1996), S. 94.

[179] Vgl. FAYYAD/PIATETSKY-SHAPIRO/SMYTH (1996), S. 24 f. und PIATETSKY-SHAPIRO ET AL. (1996), S. 94.

- ⇨ Die benötigten Daten sind in **ausreichender Qualität** vorhanden.[180] Wünschenswert wären gut gepflegte Felder mit wenigen Falscheinträgen und fehlenden Werten, so daß man überhaupt die Chance hat, Regelmäßigkeiten in den Daten aufzuspüren und den gefundenen Mustern ein hohes Vertrauen entgegenbringen kann. Falscheinträge und fehlende Werte erhöhen genau wie Datenrauschen die Wahrscheinlichkeit, daß zufällige anstelle der beabsichtigten Kausalzusammenhänge gefunden werden. Diese Wahrscheinlichkeit kann nicht so ohne weiteres bestimmt werden, da man hierzu das zu verwendende Data-Mining-Verfahren analysieren müßte, was bei komplexeren Verfahren kaum möglich ist. Daher kann keine Aussage darüber gemacht werden, wieviele Falscheinträge man tolerieren kann.

- ⇨ Die Daten sollten **über längere Zeit aussagekräftig** sein.[181] Insbesondere sollten sie keine Strukturbrüche enthalten. Grundsätzlich ist die Wahrscheinlichkeit für das Vorliegen eines Strukturbruchs umso größer, je dynamischer das Umfeld und je älter die Datenbasis ist bzw. je größer der Zeitraum ist, über den die Daten zusammengetragen wurden.

- ⇨ Schließlich müssen die relevanten Beziehungen gestützt von einer **genügenden Anzahl von Datenobjekten** in der Datenbasis vorliegen, um ein statistisch haltbares Modell zu erhalten.[182] Diese Anforderung kann erst während der eigentlichen Data-Mining-Phase exakt überprüft werden, da potentiell jede durch das Verfahren erzeugte Aussage eine andere Objektmenge abdeckt und der Aussage damit ein andere Stichprobenumfang zugrunde liegt.[183] Teilweise können intellektuell auch schon im vorhinein bestimmte Bezugsobjekttypen von der Untersuchung ausgeschlossen werden.

  *Beispielsweise können, wenn ein Unternehmen nur 100 Kunden besitzt, kaum brauchbare Datenmuster über die Kunden induziert werden, da jedes Datenmuster nur eine Teilmenge der ohnehin kleinen Kundenmenge beschreibt – z.B. die Kunden im Bundesland „Bayern“ mit der Lage „Stadtzentrum“ und „Anzahl Mitarbeiter > 200“.*

---

[180] Vgl. FAYYAD/PIATETSKY-SHAPIRO/SMYTH (1996), S. 24 f. und PIATETSKY-SHAPIRO ET AL. (1996), S. 94.

[181] Vgl. MEFFERT (1991a), S. 245.

[182] Vgl. FAYYAD/PIATETSKY-SHAPIRO/SMYTH (1996), S. 24 f.

[183] Dies wird in Abschnitt 3.3.2.1 deutlich.

⇨ Die Datenbasis darf **nicht zu viele Einzelfälle**[184] **aufweisen**, die sich sowohl in ihren erklärenden als auch in ihren zu erklärenden Merkmalen von anderen Objekten unterscheiden. Bei dem Versuch, diese Einzelfälle zu generalisieren, würden sich Widersprüche ergeben, so daß man vor dem o.g. Problem steht. Da das Auftreten dieser Widersprüche von den im Data Mining durchgeführten Generalisierungen abhängt, kann es erst während des Data Mining erkannt werden. Das Problem der Einzelfälle tritt insbesondere dann auf, wenn *viele relevante Variablen* existieren.

*Wenn beispielsweise der Zusammenhang zwischen 100 Produktparametern und den erwarteten Produktkosten abgebildet werden soll und jeder der 100 Parameter hat einen nicht zu vernachlässigenden Einfluß auf die Produktkosten, so werden die relevanten Unterschiede vergangener Produkte so groß sein, daß sich kaum generelle Aussagen ableiten lassen. Die Generalisierung scheitert dann an der zuvor genannten Anforderung, daß pro Aussage zu wenig identische Produkte erfaßt werden.*

Das Problem der nicht generalisierbaren Einzelfälle kann im Data Mining nur gelöst werden, indem man eine Repräsentationsform wählt, die Einzelfälle explizit als solche ausweist und nicht versucht, sie durch eine generelle Aussage abzudecken. Eine solche Repräsentationsform stellen in die in Abschnitt 2.2.2.3.3 eingeführten Entscheidungslisten dar. Für entsprechende Anwendungsbereiche mit vielen Entscheidungsvariablen, wie z.B. die *konstruktionsbegleitende Kalkulation*, sind andere Verfahren, die ohne Generalisierung auskommen, geeigneter.[185]

⇨ Die Planungsobjekte müssen eine **einheitliche Struktur** aufweisen, damit sie überhaupt zu generellen Aussagen zusammengefaßt werden können.

*Beispielsweise besitzen Planungsobjekte, deren Merkmale über längere Zeit und in unregelmäßigen Abständen zusammengetragen werden, wie z.B. Projektbeschreibungen, zu keinem festen Zeitpunkt eine einheitliche Struktur.*
*Auch Freitexte besitzen keine maschinell verwendbare Struktur. Beim Mining von Freitexten, dem sog. „**Text Mining**“[186], wird zur Vorstrukturierung jeder Text in einen Vektor $(h_1, ..., h_d)'$ transformiert, wobei jedes $h_i$ die Häufigkeit einer intellektuell vorzugebenden Zeichenkette, $ZK_i$, in dem Text angibt. Zusätzlich können weitere strukturierende Maßnahmen angewendet werden, wie z.B. das Streichen von vorgegebenen irrelevanten Wörtern oder die Reduzierung der Wörter auf eine Grundform. Die ermittelten Vektoren $(h_1, ..., h_d)'$ bilden die Datensätze der Trainingsmenge, so daß herkömmliche Data-Mining-Verfahren angewendet werden können. Beim sog. „**Web Mining**“[187], dem Mining von HTML-*

---

[184] Als „***Einzelfall***“ bezeichnet man ein Objekt, wenn es sich in mindestens einem betrachteten Merkmal von den gefundenen Datenmustern unterscheidet, so daß es durch kein Datenmuster abgedeckt wird.

[185] Einen Ansatz für solche Verfahren stellt z.B. das in Abschnitt 2.1.4 behandelte fallbasierte Schließen dar.

[186] Vgl. zum Text Mining ESTER/SANDER (2000), S. 246 f.

[187] Vgl. zum Web Mining ESTER/SANDER (2000), S. 246 ff. und Abschnitt 3.2.1

*und XML-Dokumenten aus dem World-Wide-Web, können zusätzlich die HTML- und XML-Befehle als Strukturinformationen genutzt werden.*
*Entsprechend bezeichnet man das Mining von geographischen Daten als „**Spatial Mining**“[188], das Mining von zeitlichen Daten als „**Temporal Mining**“[189], das Mining von Bildern als „**Image Mining**“[190] und das Mining von Multimedia-Daten als „**Multimedial Mining**“[191].*

⇨ Wenn existierende Data-Mining-Verfahren, welche ohne umfangreiches Vorwissen in maschinell verarbeitbarer Form auskommen, verwendet werden sollen, dürfen die potentiell interessanten Beziehungen in den Daten **nicht durch stärkere Zusammenhänge „überlagert“** werden. Denn wenn man Data Mining als Optimierungsproblem betreibt, so wird die interessanteste Datenmuster-Menge gesucht. Dies ist i.d.R. die mit den stärksten Zusammenhängen. Gerade im innerbetrieblichen Bereich sind die stärksten Zusammenhänge aber vorgegeben und damit trivial.

*Beispielsweise hängt der Lagerbestand eines Zwischenproduktes von den externen Kundenaufträgen bezüglich der Endprodukte, von den Produktionskoeffizienten und Vorlaufverschiebungen, von der Bestellpolitik und ihren Parametern (z.B. Meldebestand) und von der Auftragsfreigabe ab. Durch diese Zusammenhänge werden schwächere Einflüsse auf den Lagerbestand, wie z.B. der Disponent in der Beschaffung, überlagert, so daß diese nicht gefunden werden.*
*Oder die Durchlaufzeiten in der Produktion hängen ab von der Reihenfolgeplanung, der Auftragsfreigabe, dem Kapazitätsabgleich, den verwendeten Arbeitsplänen, der Losgrößenplanung und der Kundenauftragsdisposition. Auch hier werden schwächere Einflüsse auf die Durchlaufzeiten, wie z.B. der verantwortliche Entscheidungsträger in der Auftragsfreigabe, durch triviale Zusammenhänge überlagert.*

Um schwächere Einflüssen aufspüren zu können, müßte man zunächst alle trivialen Einflüsse intellektuell eleminieren. Daß dies nicht immer funktioniert, verdeutlicht das folgende Beispiel:

*Im Lagerbestand-Beispiel dürften nur identische Produkte und Stücklisten bei identischer Auftragslage, gleichbleibender Bestellpolitik und konstanter Auftragsfreigabe in die Analyse einbezogen werden, um den Einfluß des Disponenten zu ermitteln.*
*Im Durchlaufzeit-Beispiel dürften nur identische Fertigungsaufträge, die nach denselben Arbeitsplänen in derselben Maschinenbelegung bei vergleichbarer Kapazitätsauslastung gefertigt wurden, in die Analyse einbezogen werden, um den Einfluß des Entscheidungsträgers in der Auftragsfreigabe zu ermitteln.*

Die intellektuelle Eleminierung trivialer Zusammenhänge führt i.d.R. dazu, daß zu wenige Planungsobjekte übrig bleiben und somit das o.g. Eignungskriterium verletzt

---

[188] Vgl. zum Spatial Mining ESTER/SANDER (2000), S. 234 ff.

[189] Vgl. zum Temporal Mining ESTER/SANDER (2000), S. 223 ff.

[190] Vgl. zum Image Mining HAN/KAMBER (2001), S. 412 ff.

[191] Vgl. zum Multimedial Mining HAN/KAMBER (2001), S. 412 ff.

wird. Die Frage, die sich der Analytiker stellen muß, ist vor allem, ob solche schwächeren Einflüsse überhaupt von Interesse sind. Denn unter der Voraussetzung, daß die entsprechenden Größen mit vertretbarem Aufwand beeinflußt werden können (also Entscheidungsvariable darstellen), bestehen die größeren Rationalisierungspotentiale bei den starken Einflüssen, da hier entsprechend starke Wirkungen erzielt werden können. Nur wenn eine Einflußnahme unmöglich oder zu teuer ist oder wenn die mit den starken Wirkungen verbundenen Rationalisierungspotentiale bereits ausgeschöpft wurden, kann das Data Mining möglicherweise neue Potentiale offenlegen.

Würde man nun die Verwendung umfangreichen Vorwissens in maschinell verarbeitbarer Form zulassen, so könnte das Wissen dazu verwendet werden, triviale oder bereits bekannte Zusammenhänge, welche von einem Data-Mining-Verfahren generiert werden, abzuwerten. Eine solche Abwertung durch ein Neuheitsmaß wurde bereits in Abschnitt 2.2.4.5 angesprochen. Eine Komponente der Neuheit ist dabei die Unbekanntheit bezüglich bereits vorhandenen Wissens, welches in Abschnitt 3.3.2.3 neu entwickelt wird. Dort wird auch angesprochen, wie der Aufwand zur Wissensbereitstellung in Grenzen gehalten werden kann.

# 3 Data Mining als Instrument zur Entscheidungsunterstützung

Dieses Kapitel zeigt auf, wie das Data Mining genutzt werden kann, um das Management bei der Planung und Kontrolle von Entscheidungen zu unterstützen. Nach einer Einführung in betriebswirtschaftliche Entscheidungsprozesse und verschiedene Modelltypen zur Unterstützung von Entscheidungsprozessen in Abschnitt 3.1 werden in Abschnitt 3.2 existierende Anwendungen des Data Mining betrachtet und in die Phasen der Entscheidungsfindung eingeordnet. Dies ist notwendig, um in Abschnitt 3.3 ein allgemeines Problemlösungsschema für Data-Mining-Anwendungen induzieren zu können. Damit wird das erste in Abschnitt 1.3 aufgestellte Ziel (*„Lösungsschema konzipieren“*) verfolgt.

## 3.1 Modelle zur Unterstützung betriebswirtschaftlicher Entscheidungsprozesse

Entscheidungsprozesse stellen wenig standardisierte Führungsprozesse dar, welche die Phasen der Zielbildung, Situationsanalyse, Planung i.e.S., Realisierung und Kontrolle durchlaufen (vgl. Abbildung 3-1).

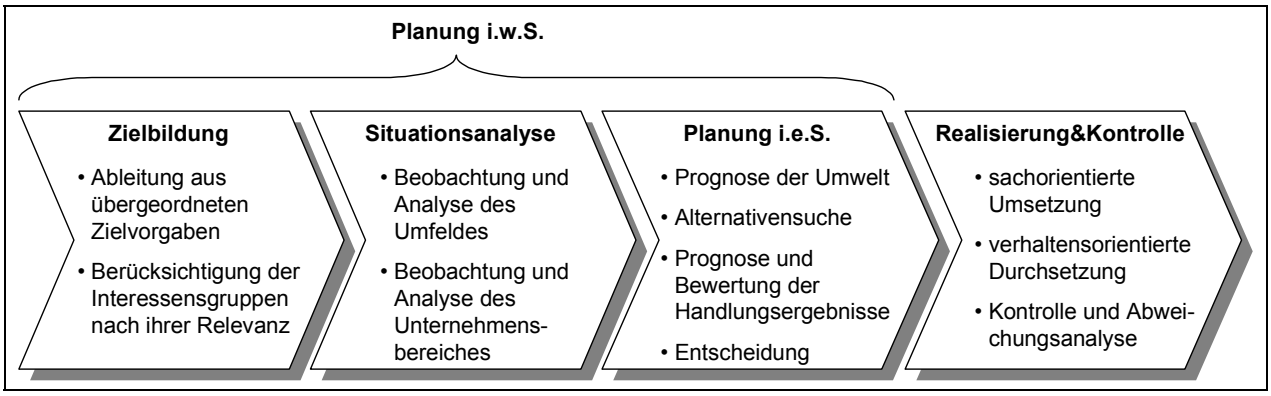



**Abbildung 3-1: Phasen von Entscheidungsprozessen**[192]

⇨ Die **Zielbildung** erfolgt unter Berücksichtigung übergeordneter Zielvorgaben. Dabei werden die von dem Verantwortungsbereich des jeweiligen Managers betroffenen Interessengruppen nach ihrer Relevanz berücksichtigt.[193]

[192] Die Darstellung ist idealisiert. Tatsächlich werden die Managementphasen nicht streng sequentiell durchlaufen.

- ⇨ Die **Situationsanalyse** dient der frühzeitigen Erkennung von Chancen und Risiken, die sich aus dem Umfeld des Verantwortungsbereiches ergeben, sowie der Erkennung von Potentialen und Gefahren innerhalb des Verantwortungsbereiches.[194] Sie umfaßt die permanente Beobachtung und Analyse des Verantwortungsbereiches und seines Umfeldes.
- ⇨ Die **Planungsphase** umfaßt die Prognose der Umwelt, die kreative Suche nach sinnvollen Handlungsalternativen, die Prognose der in bestimmten Situationen bei Durchführung bestimmter Handlungen erzielbaren Ergebnisse, deren Bewertung unter Berücksichtigung ihrer Wahrscheinlichkeiten sowie die Entscheidung für eine bestimmte Handlungsalternative.[195]
- ⇨ In der **Realisierungs- und Kontrollphase** unterscheidet man die sachorientierte Umsetzung von Entscheidungen und ihre verhaltensorientierte Durchsetzung bei den Betroffenen.[196] Zur Realisierung zählen die Ausstattung von Organisationseinheiten mit angemessenen Weisungsbefugnissen und Kommunikationsmöglichkeiten, deren Besetzung mit geeignetem Personal und die Vorgabe konkreter Sollwerte und Termine für die nachgelagerte Hierarchiestufe.[197] Bereits während der Realisierung wird die Zielerreichung kontrolliert – werden die geplanten Ergebnisse verfehlt, so sind in einer Abweichungsanalyse mögliche Ursachen herauszuarbeiten und eventuell Neuplanungen einzuleiten.

Entscheidungsprozesse können durch die Bildung und Anwendung von Modellen unterstützt werden. Ein ***Modell*** stellt eine Sammlung von Informationen über ein System dar, die zu einem bestimmten Zweck zusammengetragen wurden.[198] Informationen stellen letztendlich den wesentlichen Produktionsfaktor für die Entscheidungsfindung dar.[199] Trotzdem kann der Einsatzzweck eines betriebswirtschaftlichen Modells nicht nur in der Entscheidung für eine bestimmte Systemkonfiguration bestehen, sondern

---

[193] Vgl. WELGE/AL-LAHAM (1992), S. 51 ff.

[194] Vgl. WELGE/AL-LAHAM (1992), S. 83 ff.

[195] Vgl. GLUCHOWSKI/GABRIEL/CHAMONI (1997), S. 16 f.

[196] Vgl. WELGE/AL-LAHAM (1992), S. 387 ff.

[197] Vgl. GLUCHOWSKI/GABRIEL/CHAMONI (1997), S. 16 f.

[198] Vgl. GORDON (1969), S. 5.

[199] Vgl. GLUCHOWSKI/GABRIEL/CHAMONI (1997), S. 21.

auch in der Prognose zukünftiger Systemzustände, in der Erklärung von Systemzuständen oder in der Beschreibung von Systemzuständen.[200] Entsprechend sollen hier nach dem Modellzweck Beschreibungs-, Erklärungs-, Prognose- und Entscheidungsmodelle unterschieden werden.

***Beschreibungsmodelle*** sind Konstruktionen einzelner oder mehrerer, durch Klassifikationsschemata oder Definitionsgleichungen miteinander verknüpfter, vergangener und/oder zukünftiger betriebswirtschaftlicher Tatbestände.[201] Ihre *Aufgabe* liegt in der geordneten Erfassung und Darstellung *singulärer* betriebswirtschaftlicher Sachverhalte.[202]

***Erklärungsmodelle*** sind Konstruktionen mehrerer miteinander verknüpfter betriebswirtschaftlicher Tatbestände, die sich in „erklärende Größen“ und „zu erklärende Größen“ unterscheiden lassen.[203] Sie bilden die fundamentalen Ursache-Wirkungszusammenhänge zwischen dem Input und dem Output eines realen Systems ab. Die *Erklärungsaufgabe* besteht darin, ausgehend von gegebenen zu erklärenden Größen

- ⇨ die betriebswirtschaftlichen Sachverhalte, die als erklärende Größen (Ursachen) in Frage kommen und
- ⇨ die Art und Weise, wie die erklärenden und die zu erklärenden Größen miteinander verknüpft sind,

zu beschreiben.[204]

***Prognosemodelle*** sind Konstruktionen mehrerer miteinander verknüpfter betriebswirtschaftlicher Tatbestände, die sich in „abhängige Größen“ und „unabhängige Größen“ unterscheiden lassen; zumindest die abhängigen Größen beziehen sich dabei auf die Zukunft.[205] Die *Aufgabe eines Prognosemodells* besteht darin, aus gegebenen Werten

---

[200] Vgl. RIEPER (1992), S. 87. Die Systemzustände können dabei theoretisch, wie im folgenden deutlich wird, alle Zustandskomponenten eines Entscheidungsfeldes – also Handlungsalternativen, Umweltzustände oder Handlungsergebnisse – darstellen.

[201] Vgl. RIEPER (1992), S. 88 f. und SCHAFFT (1992), S. 5 f.

[202] Vgl. RIEPER (1992), S. 89.

[203] Vgl. RIEPER (1992), S. 89.

[204] Vgl. RIEPER (1992), S. 90.

[205] Vgl. RIEPER (1992), S. 89, S. 91.

der unabhängigen Größen und gegebenen Wirkungszusammenhängen die Werte der abhängigen Größen vorherzusagen.

***Entscheidungsmodelle*** sind formale Darstellungen von Entscheidungsproblemen.[206] Die *Aufgabe eines Entscheidungsmodells* besteht darin, in einer gegebenen Entscheidungssituation aus einer durch vorgegebene Restriktionen eingeschränkten Menge von Handlungsalternativen diejenige auszuwählen, die ein bestimmtes Zielkriterium optimiert.[207] Als „***Entscheidungssituation***" sei dabei die zum Zeitpunkt der Entscheidung vorliegende Umweltsituation bezeichnet, welche die relevante Rahmenbedingung für das Entscheidungsproblem darstellt.

*Wenn beispielsweise für einen Kreditantrag zu entscheiden ist, ob dieser angenommen oder abgelehnt werden soll, so beschreiben die Merkmale des Kreditantrags und des Kreditstellers die vorliegende Entscheidungssituation.*

Nach den zur Verfügung stehenden Handlungsalternativen unterscheidet man:[208]

⇨ Einzelentscheidungen,

⇨ Alternativentscheidungen und

⇨ Programmentscheidungen.

Bei einer **Einzelentscheidung** steht eine Alternative zur Auswahl, die entweder angenommen oder abgelehnt werden kann, d.h. die Alternativenmenge, $H$, umfaßt:

⇨ eine „Positiv-Entscheidung", $h^+$ (z.B. Kredit vergeben, Kunde kontaktieren oder Kunde akquirieren);

⇨ eine „Negativ-Entscheidung", $h^-$ (z.B. Kredit nicht vergeben, Kunde nicht kontaktieren oder Kunde nicht akquirieren).

**Alternativentscheidungen** sind dadurch charakterisiert, daß aus einer Menge von Handlungsalternativen genau eine auszuwählen ist. Bei einer **Programmentscheidung** soll eine Teilmenge von Alternativen aus einer Gesamtmenge ausgewählt werden. Charakteristisch für Programmentscheidungen ist das Vorliegen von Interdependenzen

---

[206] Vgl. DINKELBACH (1982), S. 29.

[207] Vgl. DÜSING (1997), S. 115 ff.

[208] Vgl. LACKES/MACK (2000), S. 57 ff.

zwischen den einzelnen Alternativen, wie sie z.B. durch die gemeinsame Nutzung einer Ressource entstehen können.

Abbildung 3-2 beschreibt das sog. „*Grundmodell der Entscheidungstheorie*“[209]. In der linken Randspalte sind die Alternativen $h_1, \ldots, h_{hmax}$, eingetragen, welche dem Entscheidungsträger zur Verfügung stehen. In der Kopfzeile sind die möglichen Umweltsituationen, welche die Entscheidung beeinflussen können, $u_1, \ldots, u_{umax}$, abgebildet. Die Wahrscheinlichkeit für das Eintreten der $i$-ten Umweltsituation wird mit $P(u_i)$ bezeichnet. Im Data Mining kann die Verteilung dieser Eintrittswahrscheinlichkeiten, $P(u_1), \ldots, P(u_{umax})$, durch die Trainingsdaten approximiert werden. Dieses Szenario, bei dem die $P(u_i)$ bekannt ist, wird in der Entscheidungstheorie auch als „***Entscheidung unter Risiko***“[210] bezeichnet.

In Abhängigkeit von der Entscheidung, $h_j$, und der Umweltsituation, $u_i$, wird das Handlungsergebnis $e_{j,i}$ erwartet. Den Zusammenhang $e_{j,i} = f^W(h_j,u_i)$ modelliert eine sog. „*Wirkungsfunktion*“, $f^W$. Da das Handlungsergebnis auch eine qualitative Größe, wie z.B. „Versicherungsvertrag abgeschlossen“, sein kann, muß es durch eine reelle Zahl bzw. durch einen Vektor von reellen Zahlen bewertet werden, z.B. durch die abgezinsten erwarteten Einzahlungsüberschüsse aus der Vertragslaufzeit. Diese Bewertung stellt der Zielbeitrag $z_{j,i}$ dar. Den Zusammenhang $z_{j,i} = f^Z(h_j,u_i)$ modelliert eine sog. „*Zielerreichungsfunktion*“, $f^Z$. Aus rechentechnischen Gründen könnte man die Wirkungsfunktion auch weglassen. Doch gehen zahlreiche entscheidungstheoretische Abhandlungen zunächst von Handlungsergebnissen aus, die nicht zwingend reelle Zahlen sein müssen und bewerten diese anschließend durch reellwertige Größen.[211] Daher soll auch in dieser Untersuchung auf diese Weise verfahren werden, so daß eine Bewertung der Handlungsergebnisse erforderlich wird.

Sind die Zielbeiträge mehrdimensional, so müssen sie, um eine optimale Entscheidung treffen zu können, durch Gewichtung der Einzelziele in einen eindimensionalen Wert transformiert werden. Außerdem kann, da der Eintritt einer Umweltsituation, $u$, ein Zufallsereignis darstellt, ein- und dieselbe Handlungsalternative zu verschiedenen Zielbeiträgen führen. Denn aufgrund des funktionalen Zusammenhangs $z = f^Z(h,u)$ ist

[209] Vgl. SCHNEEWEISS (1966), S. 125 ff.

[210] DINKELBACH (1982), S. 40

[211] Vgl. DINKELBACH (1982), S. 5 f.

trägen führen. Denn aufgrund des funktionalen Zusammenhangs $z = f^Z(h,u)$ ist auch die Realisierung eines bestimmten Zielbeitrags, $z$, zu einer Handlungsalternative, $h$, mit einem gewissen Risiko behaftet. Die Präferenzstruktur des Entscheidungsträgers bezüglich des Risikos und der Gewichtung der Einzelziele werden durch die sog. „*Zielfunktion*“ berücksichtigt.[212] Damit hat die Zielfunktion die Aufgabe, einer Handlungsalternative, $h_j$, einen skalaren Nutzwert, $n_j$, zuzuordnen, der es erlaubt, die verschiedenen Alternativen bezüglich ihrer Vorziehenswürdigkeit zu ordnen.

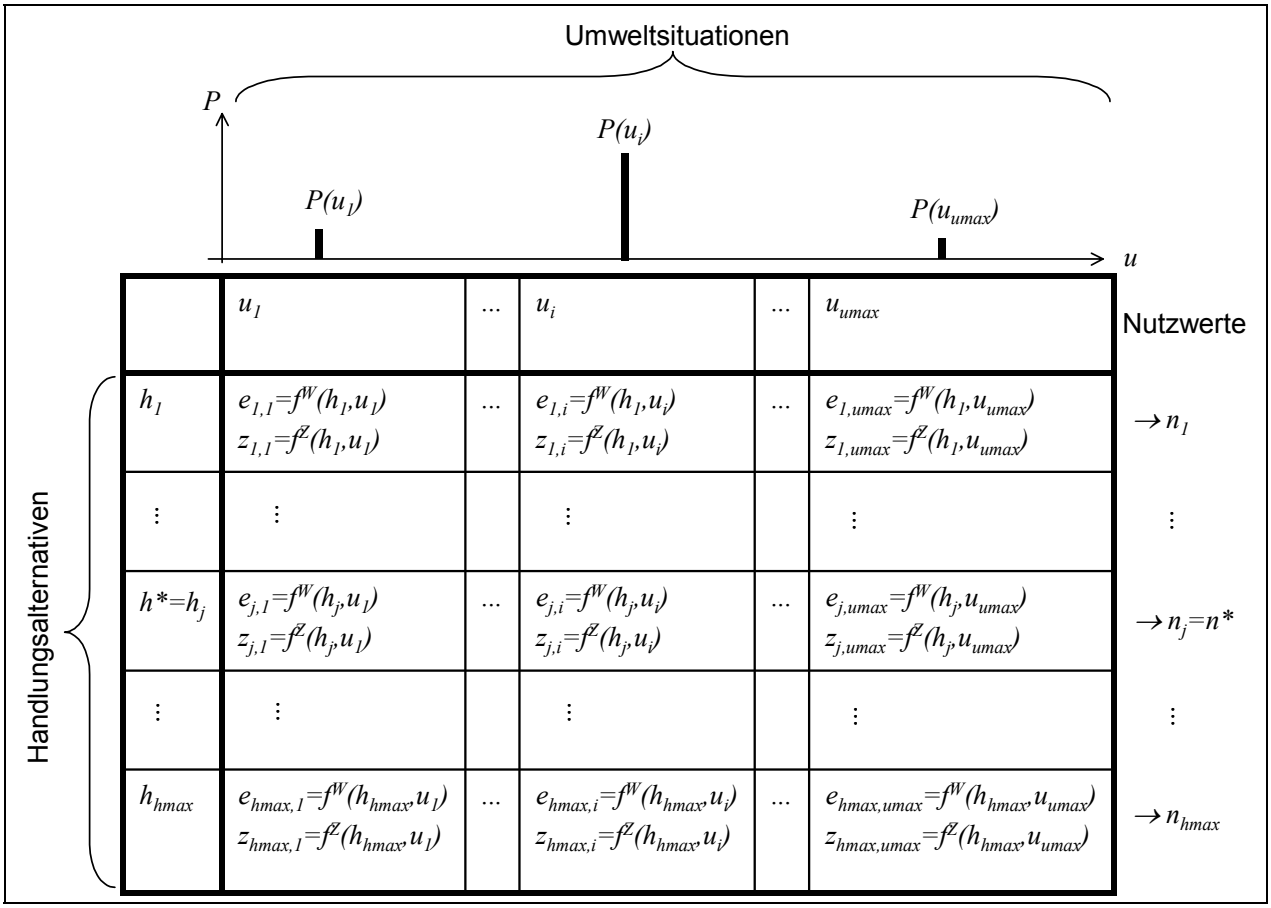



**Abbildung 3-2: Grundmodell der Entscheidungstheorie[213]**

Formal lassen sich die genannten Funktionen wie folgt definieren.[214]

**Definition 3-1: Wirkungs-, Zielerreichungs- und Zielfunktion**

Gegeben seien die Mengen der Umweltsituationen[215], $U$, der Handlungsalternativen, $H$, der Handlungsergebnisse, $E$, der Zielbeiträge, $Z$, und der Nutzwerte, $N$. Die Elemente

[212] Vgl. DINKELBACH (1982), S. 75.

[213] In Anlehnung an: DINKELBACH (1982), S. 6.

[214] Vgl. RIEPER (1992), S. 47 ff.

aus $U$, $H$, $E$ und $Z$ können Vektoren darstellen. Für die Menge der Nutzwerte, $N$, muß gelten, daß sie vollständig geordnet werden kann.

Dann lassen sich folgende funktionale Zusammenhänge zwischen diesen Komponenten definieren:

| | | | |
|---|---|---|---|
| *Wirkungsfunktion:* | $f^W$: | $H \times U$ | $\rightarrow E;$ |
| | | $(h, u)$ | $\rightarrow f^W(h, u);$ |
| *Zielerreichungsfunktion:* | $f^Z$: | $H \times U$ | $\rightarrow Z;$ |
| | | $(h, u)$ | $\rightarrow f^Z(h, u);$ |
| *Zielfunktion:* | $zf$: | $H$ | $\rightarrow N;$ |
| | | $h$ | $\rightarrow zf(h).$ ◊ |

Das Tupel $(U, H, f^W)$ wird auch als „***Entscheidungsfeld***“ bezeichnet.[216]

Unter bestimmten, noch zu erarbeitenden Voraussetzungen können derartige Entscheidungsmodelle durch ein Data-Mining-Verfahren automatisch generiert werden. Der häufigere Fall ist aber der, daß im Data Mining Prognosemodelle generiert werden, die eine oder mehrere Komponenten eines Entscheidungsmodells, d.h. Umweltsituationen, Handlungsergebnisse oder Zielbeiträge, vorhersagen und damit einen engen Entscheidungsbezug aufweisen. Die entsprechenden Vorgehensweisen werden noch herauszuarbeiten sein. Ist der Entscheidungsbezug weniger direkt, da das Entscheidungsproblem selbst noch zu unstrukturiert ist und z.B. die Handlungsalternativen noch nicht konkretisiert sind, so können möglicherweise Beschreibungs- oder Erklärungsmodelle durch Data-Mining-Verfahren produziert werden, welche die Konkretisierung eines Entscheidungsmodells unterstützen. Um die zur Modellgenerierung notwendigen Vorgehensweisen herausarbeiten zu können, seien im folgenden einige existerende Anwendungen des Data Mining betrachtet.

## 3.2 Betriebswirtschaftliche Anwendungen des Data Mining

Zur Zeit dominieren Data-Mining-Anwendungen im Marketing.[217] Aus sektoraler Sicht sind die dienstleistungs- und informationsintensiven Branchen *Versicherungen,*

[215] Demgegenüber wird im folgenden die zum Entscheidungszeitpunkt herrschende Umweltsituation als „***Entscheidungssituation***“ bezeichnet.

[216] Vgl. DÜSING (1997), S. 116.

*Banken, Telekommunikation* und *Handel* führend.[218] Insbesondere dort, wo große Datenmengen effizient erfaßt werden, wie z.B. in internetbasierten Märkten oder im Einzelhandel durch den Einsatz integrierter Warenwirtschaftssysteme am Point of Sale, sind hohe Erfolge realisierbar.[219] Der Einsatz des Data Mining erfolgt fast immer in Bereichen, in denen bereits *kleine Verbesserungen von zentraler Bedeutung für den Geschäftserfolg* sind.[220]

*Beispielsweise können kleine Verbesserungen der Prognosegenauigkeit von Kreditrisiken große Auswirkungen auf die Kreditvergabe und damit auf das Betriebsergebnis einer Bank haben.*

In Abschnitt 2.3 wurde argumentiert, daß aufgrund der Überlagerung durch stärkere, aber schon bekannte (vorgegebene) Einflüsse mit stärkeren Rationalisierungspotentialen die innerbetrieblichen Bereiche weniger geeignet für das Data Mining sind. So beziehen sich die meisten betriebswirtschaftlichen Anwendungen des Data Mining auf *unternehmensexterne Daten*, wie z.B. Kundendaten, oder auf Relationen zwischen externen und internen Daten. Unternehmensinterne Bereiche, die potentiell unbekannte Zusammenhänge aufweisen, sind eher technischer Natur (z.B. Fehlererkennung).

Die folgenden Abschnitte geben einen Überblick über die häufigsten betriebswirtschaftlichen Anwendungen des Data Mining. Dabei werden Unstimmigkeiten und Fehler einiger der betrachteten Anwendungen offengelegt und darauf verwiesen, wie sich diese vermeiden lassen. Die Anwendungen werden nach ihrer Verwendung im Entscheidungsprozeß strukturiert (vgl. Abbildung 3-3), so daß anschließend allgemeine Nutzenpotentiale des Data Mining in der Entscheidungsunterstützung herausgearbeitet werden können.

---

[217] Vgl. GROB (1999), S. 14.

[218] Vgl. KÜPPERS (1999), S. 142 ff. und FRAWLEY/PIATETSKY-SHAPIRO/MATHEUS (1991), S. 17 f.

[219] Vgl. GROB (1999), S. 18 ff.

[220] Vgl. zu dieser Einschätzung KÜPPERS (1999), S. 148 f.

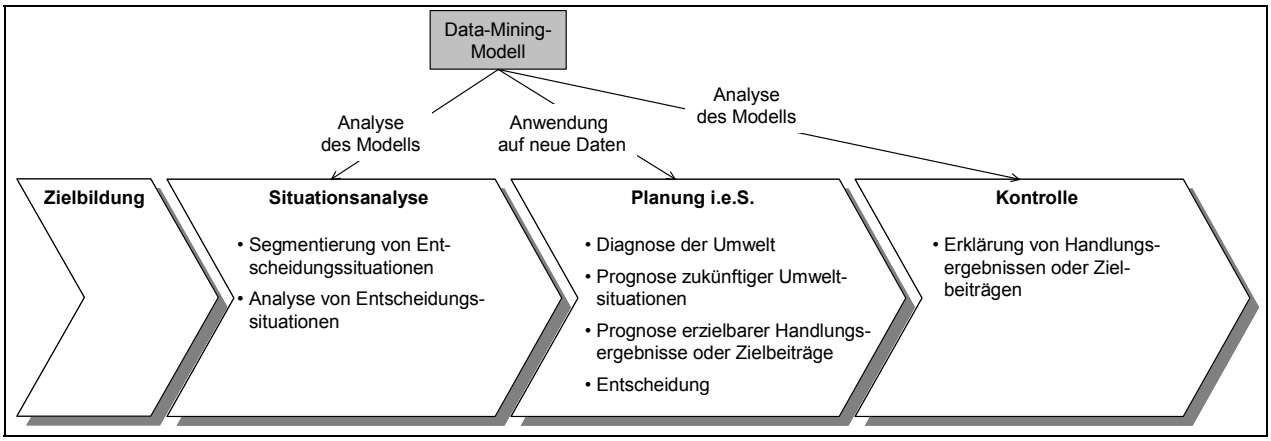



**Abbildung 3-3: Einordnung von Data-Mining-Anwendungen in die Phasen betrieblicher Entscheidungsprozesse**[221]

Abbildung 3-3 läßt erkennen, daß in der Situationsanalyse- und der Kontrollphase solche Anwendungen eingeordnet sind, die einen relativ geringen Entscheidungsbezug haben. Es liegt zwar eine Problemstellung zugrunde, die einer Entscheidung bedarf, doch häufig ist diese Problemstellung noch so unkonkret (z.B. da die Alternativen noch nicht klar strukturiert sind), daß das Entscheidungsproblem erst einmal strukturiert werden muß. Hier kann u.U. ein Beschreibungs- oder Erklärungsmodell erstellt und analysiert werden. Die Erkenntnisse aus der Analyse des Modells können dann in die Entscheidungsfindung eingehen.

In der Planungsphase sind Anwendungen eingeordnet, die ein Data-Mining-Modell auf neue Daten anwenden. Die neuen Daten charakterisieren dabei u.a. konkrete Entscheidungssituationen, womit, wie bisher, vorliegende Umweltsituationen bezeichnet seien, die als Problemsituationen wahrgenommen werden und daher einer Entscheidung bedürfen. Der Entscheidungsbezug ist hier wesentlich enger, da das Data-Mining-Modell entweder konkrete Parameter eines Entscheidungsmodells (Umweltsituationen, Handlungsergebnisse oder Zielbeiträge) liefert oder sogar direkt eine Entscheidung trifft.

[221] Die Erläuterungen zu den einzelnen in der Abbildung aufgeführten Anwendungsbereichen erfolgen in den nachfolgenden Abschnitten.

### 3.2.1 Anwendungen zur Segmentierung von Entscheidungssituationen

Dieser Abschnitt behandelt Anwendungen, welche Segmente ähnlicher Entscheidungssituationen identifizieren und durch typische Merkmale beschreiben. Die Segmentbeschreibungen können innerhalb des Managementprozesses einen Handlungsbedarf offenlegen. Idealerweise sollten die Segmente so generiert werden, daß jedem Segment eine eigene Handlung zugeordnet werden kann. Dabei geben die Segmentbeschreibungen Hinweise auf mögliche Entscheidungen. Die Entscheidung selbst ist keine modellendogene Größe, da die Datenbasis weder Handlungsalternativen noch -ergebnisse umfaßt. Da keine Größe erklärt oder vorhergesagt wird, handelt es sich hier um reine Beschreibungsmodelle.

Für die Segmentierung von Entscheidungssituationen eröffnet sich der Erstellung von Fehlerprofilen in technischen Systemen ein großes Anwendungsfeld.[222] Eher betriebswirtschaftlich ausgerichtet sind in diesem Zusammenhang folgende große Anwendungsbereiche:

- ⇨ die Kunden- bzw. Marktsegmentierung;
- ⇨ die Generierung typischer Warenkorbprofile und
- ⇨ die Segmentierung von Webseitenanforderungen.

Abbildung 3-4 ordnet diese Anwendungsbereiche in die Phasen betrieblicher Entscheidungsprozesse ein.

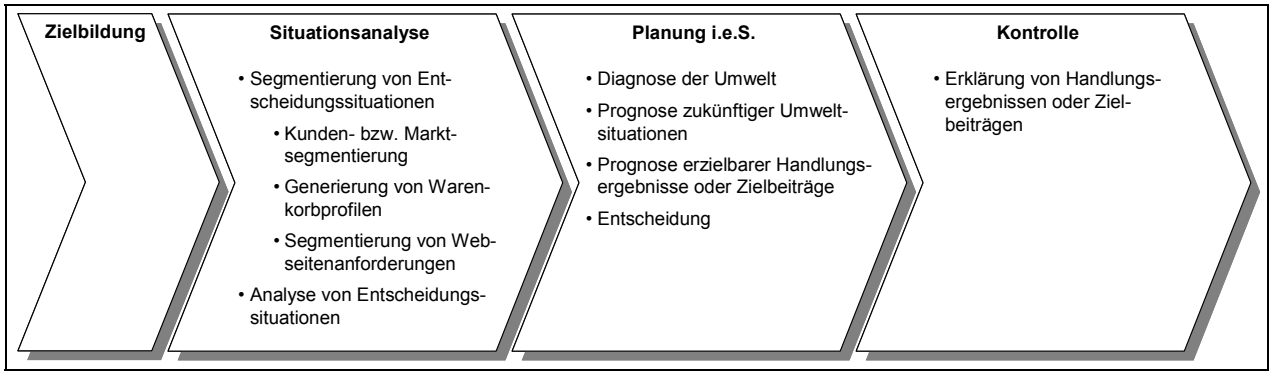



**Abbildung 3-4: Anwendungen zur Segmentierung von Entscheidungssituationen**

[222] Vgl. z.B. KÜPPERS (1999), S. 139 f.

Die ***Marktsegmentierung*** ist eine Teilfunktion der Situationsanalyse, die sich mit der Aufteilung des Gesamtmarktes in homogene Käufergruppen beschäftigt.[223] Das Hauptziel der Marktsegmentierung besteht darin, einen hohen Grad von Identität zwischen der angebotenen Marktleistung und den homogenen Käufern eines Segmentes zu erzielen.[224] Um dieses Ziel zu erreichen, wird für jedes relevante Marktsegment ein eigenes Marketingprogramm geplant. Daher sollten die zur Segmentierung genutzten Kriterien nicht nur eine gewisse Kaufverhaltensrelevanz aufweisen, sondern auch für den differenzierten Einsatz von Marketinginstrumenten geeignet sein.[225]

Die ***Kundensegmentierung*** bezieht sich in Abgrenzung zur Marktsegmentierung nicht auf den Gesamtmarkt, sondern auf die aktuellen Kunden eines Unternehmens, verfolgt aber dasselbe Ziel und verwendet dieselben Methoden, so daß Markt- und Kundensegmentierung hier gemeinsam betrachtet werden können. Durch die Fokussierung auf die eigenen Kunden ist im Bereich der Kundensegmentierung i.d.R. eine größere und verläßlichere Datenbasis gegeben als dies bei Marktsegmentierungen aufgrund von Umfragedaten der Fall ist. Damit sind die entsprechenden datenbasierten Eignungskriterien für den Einsatz von Data-Mining-Verfahren bei der Kundensegmentierung eher erfüllt.

*Beispielsweise führten HIPPNER und SCHMITZ eine Segmentierung von Privatkunden eines Kreditinstitutes durch, um dem Institut eine informationelle Grundlage für die differenzierte Kundenansprache zu schaffen.[226] POLONI und NELKE segmentieren die Homebanking-Nutzer einer Bank und definieren alle Kunden, die eine gewisse Zugehörigkeit zu den Segmenten unterschreiten, als Nichtnutzer.[227] Die gewonnenen Kundenprofile sollen zur gezielten Ansprache von Kunden für existierende und neue Produkte genutzt werden.[228] Und WAGNER, REISINGER und RUSS segmentieren die Charterkunden von Lauda Air nach den gewünschten Serviceleistungen, um diese Erkenntnisse bei der Gestaltung der Kommunikationspolitik zu nutzen.[229] Diese verfolgt das Ziel, ein spezielles Image der Charterflugleistungen aufzubauen, das sich eindeutig von den Linienflügen abhebt.[230]*

---

[223] Vgl. MEFFERT (1991a), S. 243 und MEFFERT (1991b), S. 40.

[224] Vgl. MEFFERT (1991a), S. 243.

[225] Vgl. zu diesen und weiteren Anforderungen an Segmentierungskriterien: FRETER (1983), S. 43 f.

[226] Vgl. HIPPNER/SCHMITZ (2001), S. 608 ff.

[227] Vgl. POLONI/NELKE (2001), S. 646 f.

[228] Vgl. POLONI/NELKE (2001), S. 643.

[229] Vgl. WAGNER/REISINGER/RUSS (2001), S. 883 ff.

[230] Vgl. WAGNER/REISINGER/RUSS (2001), S. 877.

Ein regelmäßig auftretendes Problem sowohl bei der Kunden- als auch bei der Martsegmentierung besteht darin, daß einige Kriterien eine höhere Kaufverhaltensrelevanz und andere eine höhere Aussagekraft bezüglich der Erreichbarkeit der Zielgruppen über bestimmte Marketinginstrumente aufweisen. Hier stellt sich die Frage, welche Kriterien zur Segmentierung herangezogen werden sollen.

*LÖBLER und PETERSOHN beantworteten diese Frage für die Marktsegmentierung im Automobilhandel wie folgt:[231] Sie befragten potentielle Autokäufer und bilden Segmente mit Merkmalen, die die Einstellungen der Personen gegenüber der vertriebenen Marke charakterisieren, wie z.B. die Einschätzung der Zuverlässigkeit, der Sicherheit, der Servicequalität und der Ausstattung. Von solchen Imagemerkmalen nimmt man an, daß sie eine gewisse Kaufverhaltensrelevanz aufweisen. Nachdem mehrere Segmente gebildet wurden, die sich in der Einstellung zu der Automarke unterscheiden, wurden diejenigen Segmente mit einer positiven Einstellung daraufhin untersucht, ob sie durch unterschiedliche Marketinginstrumente erreichbar sind. Hierzu wurden Merkmale wie Geschlecht, Alter, Beruf und Einkommen sowie die bevorzugte Informationsquelle beim Autokauf (Zeitungen, Autozeitschriften, Fernsehen, Händler, ADAC, Internet u.a.) hinzugezogen, die zuvor nicht zur Bildung der Segmente dienten, sonden die Erreichbarkeit der interessanten Segmente sicherstellen soll.*

Die Segmentierung aufgrund kaufverhaltensrelevanter Merkmale kann dazu führen, daß Segmente gebildet werden, die durch die zur Verfügung stehenden Marketinginstrumente nicht erreicht werden können. Möglicherweise können aber neue Instrumente entwickelt werden. Auf jeden Fall ist das alternative Vorgehen, erreichbarkeitsrelevante Merkmale in die Segmentierung aufzunehmen, schlechter zu beurteilen, da darunter die anvisierte Kaufverhaltenswirkung leidet. Das Problem (bedürfnis-differenzierte Marktbearbeitung) muß also zuerst analysiert werden (d.h., es müssen bedürfnisorientierte Marktsegmente gebildet werden), bevor über die Problemlösung (Gestaltung des Marketingsinstrumentariums) nachgedacht wird.

Viele Unternehmen sind bestrebt, ihre Aktivitäten auf die ertragreichen Segmente zu fokussieren. Dies führt zu dem weit verbreiteten Fehler, Erfolgsgrößen mit in die Segmentierung einzubeziehen. Als Segmentierungskriterien sollten aber ausschließlich solche Merkmale verwendet werden, die mit bestimmten Bedürfnissen der Marktteilnehmer korrespondieren. Erfolgsgrößen sollten erst nach erfolgter Segmentierung zur Charakterisierung der Segmente herangezogen werden.

*Beispielsweise werden für die Segmentierung von Bankkunden als besonders relevant erachtet: das Alter, das Nettoeinkommen, das bei der Bank angelegte Geldvermögen, die Höhe der bei der Bank in Anspruch*

[231] Vgl. zu diesem Beispiel: LÖBLER/PETERSOHN (2001), S. 623 ff.

*genommenen Kredite und der Jahresdeckungsbeitrag des Kunden.[232] Letzterer kann sicher nicht dazu dienen, Kunden mit ähnlichen Bedürfnissen zu gruppieren.*

Ein weiterer verbreiteter Fehler bei der Marktsegmentierung besteht darin, auch dann ein reines Beschreibungsmodell zu generieren, wenn eine geeignete zu erklärende Variable zur Verfügung steht.

*Beispielsweise identifiziert die LVM-Versicherung bei den Kunden der Sparte Kraftfahrzeug-Haftpflicht 25 Segmente und beschreibt diese durch Merkmale, die das Schadensrisiko determinieren, wie z.B. der Regionalstatistik des HUK-Verbandes, technische Merkmale, Geschlecht, Garagennutzung und Fahrleistung. Auf dieser Situationsanalyse aufbauend können Maßnahmen geplant werden, wie z.B. eine feine Tarifabstufung, das Versagen von Rabatten, das Aussprechen von Kündigungen oder Aktionen zur Verhütung von Schäden. Soweit wäre die Vorgehensweise korrekt, doch nach den Ausführungen von KÜPPERS wird zur Segmentbildung auch die Schadenshäufigkeit herangezogen, obwohl dies gerade das zu erklärende Phänomen charakterisiert.[233]*

Eine Variante der Kundensegmentierung stellt die **Generierung typischer Warenkorbprofile** dar. Die Generierung typischer Warenkorbprofile findet aufgrund der kostengünstig erhebbaren, großen Datenmengen vor allem in der Kassenbonanalyse Verwendung, ist aber nicht auf diese beschränkt. In anderen Branchen treten an die Stelle der Kassenbons Versicherungspolicen, Schadensmeldungen, Zahlungsbelege, Bestellungen, Reklamationen oder Rechnungen.[234]

*Ein mögliches Warenkorbprofil wäre beispielsweise:*

```
´Italienisches Abendessen´:
    Anzahl Einkäufe signifikant hoch bei der Warengruppenkombination:
        Spaghetti, Schinken, Basilikum, Knoblauch, Rotwein.
```

*Man erkennt bei solchen Warenkorbanalysen u.a., daß bestimmte Artikel häufig zusammen gekauft werden, z.B. Nudeln, Fleischsorten, Gewürze und Rotwein für ein italienisches Abendessen. Anstatt diese Artikel jeweils nur bei den Teigwaren, beim Fleisch, bei den Gewürzen und bei den Alkoholika zu plazieren, sollten sie zusammen plaziert werden. Dies kann dazu führen, daß der Kunde, da er durch die gemeinsam angebotenen Waren zusätzliche Kaufimpulse aufnimmt, anstelle von 25 DM nun 32 DM für das Abendessen ausgibt.[235]*

Während sog. „*Bedarfsverbünde*“, wie das erwähnte italienische Abendessen oder ein Mikrowellenherd und das passende Mikrowellengeschirr oder eine Uhr und die passende Batterie, aus komplementären Ge- bzw. Verbrauchsgewohnheiten entstehen und damit relativ offensichtlich sind, gibt es noch weitere, weniger triviale Verbundarten. So

---

[232] Vgl. ZIMMERMANN (1995), S. 144 f.

[233] Vgl. KÜPPERS (1999), S. 144 f.

[234] Vgl. STÄDTLER/FISCHER (1998), S. 340.

[235] Vgl. LACKES/MACK/TILLMANNS (1998), S. 253.

entsteht ein *Nachfrageverbund* allein durch besondere Wünsche und Einstellungen der Kunden (z.B. Windeln und Bier), ein *Auswahlverbund* entsteht durch Bedarfserweiterungseffekte (z.B. mehrere Sachbücher zum selben Thema), und ein *Akquisitionsverbund* entsteht durch die gemeinsame Absatzförderung mehrerer Artikel (z.B. in einem Prospekt gemeinsam angebotene Auslaufmodelle von Elektroartikeln).[236]

In der Konsumgüterbranche sind die Ergebnisse der Kassenbonanalyse insbesondere für den Handel, die Industrie und Branchenverbände interessant. Der Handel verwendet sie zum gemeinsamen Bewerben der assoziierten Artikel, zur Sortimentsplanung und zur Ladengestaltung; die Industrie verwendet sie zur Kontrolle, ob ihre Produkte durch den Handel optimal präsentiert und beworben werden; die Branchenverbände verwenden sie zur Versorgung ihrer Mitglieder mit Marktinformationen.[237]

Die letzte betrachtete Anwendung von Beschreibungsmodellen, die **Segmentierung von Webseitenanforderungen**, ist eine Aufgabe innerhalb des sog. „***Web Mining***“. Das Web Mining unterscheidet sich von dem üblichen Data Mining im wesentlichen in der Struktur der Datenbasis, welche sich auch auf die verwendeten Verfahren auswirken kann. Je nach Ausprägung des Web Mining umfaßt die ursprüngliche Datenbasis:

- ⇨ den HTML-Code der Webseite („***Web Content Mining***“),
- ⇨ die Linkstruktur einer Web-Präsenz („***Web Structure Mining***“) oder
- ⇨ die Zugriffe der Nutzer auf eine Web-Präsenz („***Web Usage Mining***“).[238]

Während die beiden erstgenannten Aufgaben des Web Mining eher technischen[239] Anwendungen dienen, kann das Web Usage Mining betriebswirtschaftliche Entscheidungen unterstützen. Werden die Zugriffe auf eine Web-Präsenz ausschließlich auf der Grundlage von Protokolldateien der Web-Server durchgeführt, spricht man auch vom

---

[236] Vgl. zu den genannten Verbundarten: MICHELS (2001), S. 934.

[237] Vgl. STÄDTLER/FISCHER (1998), S. 340 f.

[238] Vgl. KOSALA/BLOCKEEL (2000), S. 3 f.

[239] Das ***Web Content Mining*** segmentiert HTML-Seiten oder faßt sie inhaltlich zusammen und wird beispielsweise in Suchmaschinen eingesetzt (vgl. KOSALA/BLOCKEEL (2000), S. 4 ff.). Es umfaßt darüber hinaus Klassifikationstechniken zur maschinellen Typisierung von Web-Präsenzen, wie sie z.B. zum Zwecke des Aufbaus eines Suchdienstes für betriebliche E-Commerce-Angebote erforderlich ist (vgl. KURBEL/SZULIM/TEUTEBERG (2000), S. 223). Das ***Web Structure Mining*** segmentiert Web-Präsenzen nach ihrer Linkstruktur oder klassifiziert die einzelnen Seiten aufgrund der Linkstruktur in Einstiegs-, Verteiler- oder Inhaltsseiten, was beispielsweise innerhalb von Suchmaschinen als Indiz für die Relevanz einer Seite herangezogen wird (vgl. KOSALA/BLOCKEEL (2000), S. 7 ff.).

„***Web Log Mining***“[240] als Teilaufgabe des Web Usage Mining. Jenachdem, ob dabei die zeitliche Reihenfolge der Zugriffe auf die Webseiten berücksichtigt wird oder nicht, unterscheidet man die Pfad- von der Assoziationsanalyse. Die ***Assoziationsanalyse*** liefert typische Sammlungen von Seiten, die signifikant häufig in beliebiger Reihenfolge innerhalb einer Sitzung aufgerufen werden.[241] Diese Ergebnisse können zur Verlinkung der Web-Präsenz genutzt werden, um ein Cross Selling oder ein Bundling zu ermöglichen.[242]

*Es sei das Beispiel eines Internet-Shops betrachtet, der grundsätzlich hierarchisch strukturiert ist, d.h. man klickt top down von Abteilungen, Unterabteilungen, Warengruppen und –untergruppen bis zum gewünschten Artikel. Neben der hierarchischen Linkstruktur können zusätzliche Links zwischen Seiten mit komplementären Artikeln oder Warengruppen eingefügt werden, um die Navigation zu erleichtern. Das Einfügen von Links ist zwar nur mit geringen Kosten verbunden, so daß man auf die Idee kommen könnte, sehr viele Querverweise in die Web-Präsenz einzupflanzen – dies würde ab einer gewissen Anzahl von Querverweisen die Navigation aber eher erschweren als erleichtern.*

Die ***Pfadanalyse*** liefert typische Sequenzen von Seiten, die signifikant häufig in derselben Reihenfolge innerhalb einer Sitzung aufgerufen werden.[243] Die Ergebnisse der Pfadanalyse geben Aufschlüsse über das Informationssuchverhalten der Konsumenten und können zur Gestaltung der Web-Präsenz genutzt werden.[244]

*Die Erkenntnisse über die typischen Pfade können beispielsweise dazu genutzt werden, unnötig lange Pfade, welche häufig auftreten und zu wichtigen Seiten führen, zu verkürzen, indem die Links, welche zu einer Abkürzung führen würden, deutlicher sichtbar gemacht werden.*

Das Web Usage Mining weist gewisse Vorteile gegenüber der normalen Warenkorbanalyse auf. Erstens müssen die Produkte nicht unbedingt gekauft werden, da bereits der Aufruf von Informationsseiten über die Produkte genügt, um daraus Rückschlüsse für Marketingaktionen zu ziehen. Zweitens können Verbundbeziehungen nicht nur zwischen Artikeln, sondern zwischen beliebigen Web-Ressourcen, wie z.B. Informationsseiten, Bestellfunktionen, Suchanfragen oder Dateien mit Produktdemos, ermittelt werden.

---

[240] BENSBERG/WEIß (1999), S. 426.

[241] Vgl. GROB/BENSBERG (1999), S. 20.

[242] Vgl. BENSBERG/WEIß (1999), S. 430.

[243] Vgl. GROB/BENSBERG (1999), S. 20.

[244] Vgl. BENSBERG/WEIß (1999), S. 430.

Sowohl bei der Untersuchung typischer Warenkorbprofile als auch bei der Analyse assoziierter Webseiten unterliegen die ermittelten Zusammenhänge den Auswirkungen vergangener Entscheidungen, wie z.B. den vorhandenen Warenplazierungen im Supermarkt oder der vorhandenen Verlinkung der Webseiten. Dies hat zur Folge, daß die ermittelten Assoziationen nicht ungesehen zur Entscheidung über neue Plazierungen bzw. Verlinkungen verwendet werden dürfen. Nur solche Assoziationen, die nicht auf bereits bestehende Verbundentscheidungen zurückzuführen sind, dürfen bei neuen Entscheidungen eine Rolle spielen. Zwar stört es nicht weiter, wenn auch bereits bestehene Waren- oder Webseitenverbünde aufgedeckt werden – problematisch wird es dann, wenn diese so stark und so zahlreich sind, daß andere, schwächere, aber vielleicht unbekannte Zusammenhänge verborgen bleiben. Dieses Problem wurde in Abschnitt 2.3 bereits unter dem Stichwort „Überlagerung durch stärkere Zusammenhänge“ diskutiert.

### 3.2.2 Anwendungen zur Analyse von Umweltsituationen

Die Analyse von Umweltsituationen geht über die bloße Beschreibung der Umwelt, wie sie im letzten Abschnitt behandelt wurde, hinaus. Es wird nach Ursachen für bestimmte beobachtbare Umweltzustände gesucht. In der Situationsanalyse werden diese Ursachen im Umfeld des Verantwortungsbereiches des Managers gesucht – dagegen werden die Auswirkungen eigener Handlungen in der Kontrollphase untersucht. Zur Analyse der Umweltsituationen werden deren Eigenschaften in erklärende und zu erklärende Merkmale aufgeteilt. Die erklärenden Merkmale beschreiben dann die gesuchte Problemursache und die zu erklärenden Merkmale die wahrgenommene Problemsituation. Die Ursache-Wirkungszusammenhänge werden durch ein Erklärungsmodell abgebildet. Aus der Kenntnis der Problemursachen und der Ursache-Wirkungszusammenhänge können sich für den Manager sowohl ein Handlungsbedarf als auch Ansätze für die Maßnahmenplanung ergeben.

Große Anwendungsbereiche des Data Mining ergeben sich in diesem Zusammenhang bei der Früherkennung und Ursachenanalyse von Krankheiten[245] und von Fehlern in

[245] Vgl. WIEDERHOLD ET AL. (1986), S. 81 f.

technischen Systemen[246]. Diese werden hier aufgrund ihres geringen betriebswirtschaftlichen Bezugs ausgegrenzt. Die Früherkennung von Betrugsfällen im Telekommunikations-[247] oder Kreditkartengeschäft[248] sind weitere häufig referenzierte Anwendungen von Erklärungs- und sogar Beschreibungsmodellen. Diese Anwendungen weisen allerdings einen so engen Bezug zu Entscheidungen über die Prüfung der Verdachtsmomente auf, daß sich die Generierung von Entscheidungsmodellen anbietet. Dies wird in Abschnitt 3.2.6 anhand der Prüfung von Steuerbetrugsfällen gezeigt. An dieser Stelle werden die folgenden betriebswirtschaftlichen Anwendungsbereiche betrachtet:

- ⇨ die Erklärung von Kundenbeziehungen;
- ⇨ die Erklärung von Kundeneinstellungen;
- ⇨ die (gerichtete) Kassenbonanalyse und
- ⇨ die Erklärung der Varianten- oder Markenwahl von Kunden.

Abbildung 3-5 ordnet diese Anwendungsbereiche in die Phasen betrieblicher Entscheidungsprozesse ein.

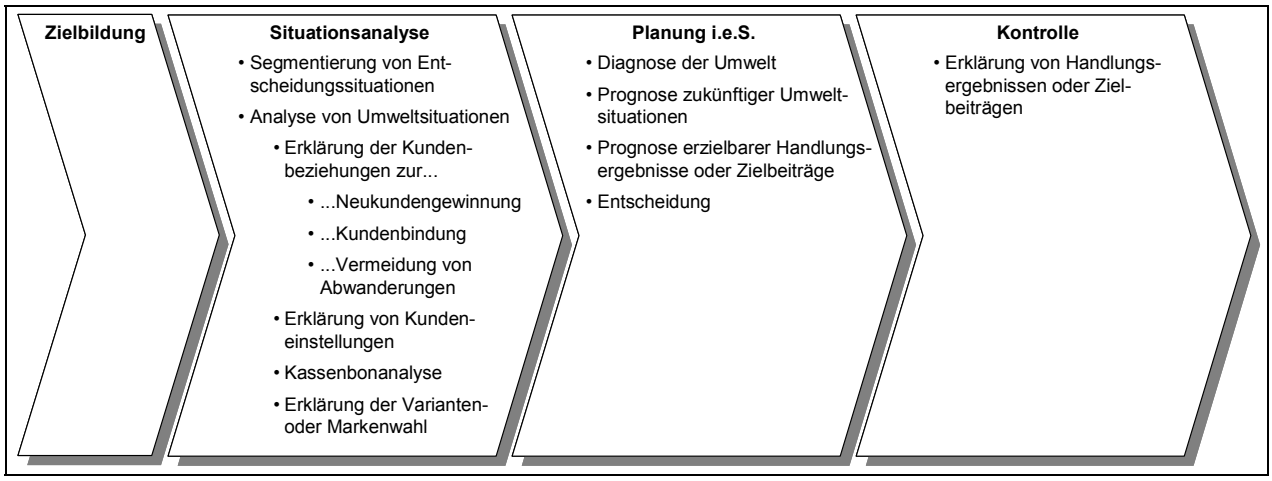



**Abbildung 3-5: Anwendungen zur Analyse von Entscheidungssituationen**

Die **Erklärung von Kundenbeziehungen** erfolgt gewöhnlich mit dem Ziel, ein Verständnis für das Kundenverhalten zu entwickeln, auf dessen Grundlage Maßnahmen

[246] Vgl. KÜPPERS (1999), S. 139 f.

[247] Vgl. HAN/KAMBER (2001), S. 457.

[248] Vgl. ESTER/SANDER (2000), S. 8.

zur gezielten Beeinflussung des Kundenverhaltens konzipiert werden können. Kundenbeziehungen können nach dem erreichten Lebenszyklusstadium differenziert werden in:

- ⇨ erste Interessenbekundungen (Interessent);
- ⇨ Eintritt in eine Kundenbeziehung (Neukunde);
- ⇨ Reifung zum Stammkunden (Stammkunde);
- ⇨ Abwanderung (Ex-Kunde).

Hier können Data-Mining-Modelle erstellt werden, die den Eintritt in ein neues Lebenszyklusstadium erklären.

*Als Beispiel zur Erklärung der Neukundengewinnung sei die Ursachenanalyse erfolgreicher Gewinnung von Sammelbestellern im Versandhandel betrachtet.[249] Sammelbesteller bündeln die Kaufwünsche ihrer meist in der Nähe wohnenden Bekannten zu größeren Bestellungen und nehmen dabei auch beratende Funktionen ein. Damit sind Sammelbesteller und ihre regionale Verteilung für Versandhandelshäuser von größter wirtschaftlicher Bedeutung. Dementsprechend viel wird in die Akquisition dieser Verkaufshelfer investiert. Günstig wäre es, wenn die Sammelbesteller entsprechend der Kaufkraft bzw. Kaufbereitschaft der einzelnen Regionen verteilt wären. Wie die Investitionen in Akquisitionsmaßnahmen regional dosiert werden sollen, ist eine kritische Frage, die regelmäßig untersucht wird. Das Data Mining kann hier unterstützen, indem ein Modell generiert wird, das die Anzahl der neu gewonnenen Sammelbesteller durch regionale Merkmale, wie z.B. Umsätze, Haushaltsgröße, Einwohnerzahl und Filialen pro Region erklärt.*

Eine **Erklärung von Kundeneinstellungen** wird vorgenommen, wenn die direkte Erklärung des Kaufverhaltens von Kunden nicht möglich ist. Insbesondere für neu einzuführende Produkte liegen keine Daten zu vergangenen Kaufentscheidungen vor, so daß Einstellungsvariablen als Indikatoren für spätere Kaufentscheidungen definiert und erklärt werden. Das Data Mining kann hier eingesetzt werden, um Konsumententypologien zu erstellen, die als aussagekräftig bezüglich späterer Kaufentscheidungen gelten.[250]

*Beispielsweise wurde im Auftrag des Verlagshauses Gruner+Jahr eine Befragung von 5518 Personen zu 1119 Variablen insbesondere des Lebens- und Konsumstils durchgeführt.[251] Ein Teil der zusammengetragenen Daten wurde in einer gesonderten Studie analysiert, um daraus den Konsumententyp „Billigkäufer" anhand soziodemographischer Merkmale zu erklären.[252]*

---

[249] Vgl. zu diesem Beispiel BISSANTZ/BRAUN (1998), S. 16 ff.

[250] Vgl. STECKING (2000), S. 88.

[251] Vgl. STECKING (2000), S. 88 f.

[252] Vgl. zu diesem Beispiel STECKING (2000), S. 90 ff.

*Die zu erklärende Variable stellt die Summe aus 14 binären Variablen, $x_1$, ..., $x_{14}$, dar. Die $i$-te Variable, $x_i$, steht dabei für: „Produktgruppe $i$ kaufe ich dort, wo es am billigsten ist" und kann die Ausprägungen 0 für „nein" und 1 für „ja" annehmen. Diese Frage wurde zu folgenden 14 Produktgruppen gestellt:*

- ⇨ *Bier („Kaufen Sie Bier dort wo es am billigsten ist?"),*
- ⇨ *Wein und Sekt („Kaufen Sie Wein und Sekt dort wo es am billigsten ist?"),*
- ⇨ *Spirituosen (usw.),*
- ⇨ *Alkoholfreie Getränke,*
- ⇨ *Käse,*
- ⇨ *Kosmetik, Körper- und Haarpflegemittel,*
- ⇨ *Mode, Kleidung und Schuhe,*
- ⇨ *Uhren und Schmuck,*
- ⇨ *Fotokameras und Zubehör,*
- ⇨ *Elektrische Haushaltsgeräte,*
- ⇨ *Möbel und Einrichtungsgegenstände,*
- ⇨ *Hifi-, Video- und Fernsehgeräte,*
- ⇨ *Autos,*
- ⇨ *Reise- und Hotelangebote.*

*Die aggregierte Variable „Billigkäufer" nimmt damit ganzzahlige Werte zwischen 0 („kaufe kein Produkt dort, wo es am billigsten ist") und 14 („kaufe alle Produkte dort, wo sie am billigsten sind") an. Als erklärende Variablen wurden folgende 14 soziodemographische Merkmale betrachtet:*

- ⇨ *Geschlecht (weiblich, männlich),*
- ⇨ *Alter (in Jahren: 18-70),*
- ⇨ *Familienstand (verheiratet, ledig, verwitwet, geschieden),*
- ⇨ *Mit Partner zusammenlebend (nein, ja),*
- ⇨ *Haushalt besorgend (gar nicht, auch, hauptsächlich),*
- ⇨ *Haushaltsvorstand (nein, ja),*
- ⇨ *Schulabschluß (Volksschule, Weiterführende ohne Abitur, Abitur, Studium),*
- ⇨ *Berufstätigkeit (berufstätig, nicht berufstätig, in Ausbildung, Rentner),*
- ⇨ *Anzahl der Einkommensbezieher im Haushalt (1 Person, 2 Personen, 3 und mehr),*
- ⇨ *Nettoeinkommen in DM (12 Einkommensklassen),*
- ⇨ *Haushalts-Nettoeinkommen in DM (11 Einkommensklassen),*
- ⇨ *Anzahl der Personen im Haushalt (1 Person, 2 Personen, 3 Personen, 4 und mehr),*
- ⇨ *Kinder unter 14 Jahre im Haushalt (nein, ja),*
- ⇨ *Ortsgröße (6 Klassen).*

*Zusammenhänge zwischen diesen erklärenden Merkmalen und der Variable „Billigkäufer" können dann analysiert und zur Vermarktung von Billigprodukten genutzt werden, z.B. durch Verteilung von Prospekten mit Sonderangeboten an Empfänger mit den für relevant befundenen soziodemographischen Merkmalsausprägungen.*

Ein weiterer Anwendungsbereich, bei dem es um die Erklärung des Kundenverhaltens geht, ist die **(gerichtete) Kassenbonanalyse**. Sie stellt eine weitere Aufgabe dar, welche sich der bereits in Abschnitt 3.2.1 behandelten Warenkorbanalysen zuordnen läßt. Sie soll zu der dort vorgestellten Generierung typischer Warenkorbprofile dadurch abgegrenzt werden, daß hier erstens weitere, dem Kassenbon zuordnungsfähigen Merkmale einbezogen und zweitens *gerichtete* Abhängigkeiten zwischen Artikelkäufen

untersucht werden. Während auf die einem Kassenbon zuordnungsfähigen Merkmale weiter unten eingegangen wird, bezeichnen *gerichtete* Abhängigkeiten Zusammenhänge wie z.B.:

Der Kauf von Artikel $A$ erklärt den Kauf von Artikel $B$.

Diese Information kann zur Planung von gerichteten artikelbezogenen Maßnahmen verwendet werden.

*Ein Beispiel für solche „gerichteten artikelbezogenen Maßnahmen" ist durch die Eliminierung von Artikeln aus dem Sortiment gegeben. Artikel $A$ habe einen negativen Deckungsbeitrag. Die Entscheidung, ob Artikel $A$ aus dem Sortiment eliminiert werden kann, hängt aber auch davon ab, ob durch eine Eliminierung der Absatz anderer, mit $A$ assoziierter Artikel gefährdet wäre. Wenn die o.g. Abhängigkeit zwischen $A$ und $B$ sehr stark ist und $B$ einen hohen Deckungsbeitrag erwirtschaftet, dann sollte $A$ nicht eliminiert werden.*

Des weiteren können sich unter Berücksichtigung der gerichteten Abhängigkeiten Kernartikel herauskristallisieren, die so stark mit anderen Artikeln assoziiert sind, daß es bei der Nachdisposition genügt, die Bestellmenge dieser Kernartikel zu disponieren und die Bedarfsmengen und -zeitpunkte aller assoziierten Artikel davon abzuleiten.[253]

Gerichtete Kassenbonanalysen müssen nicht unbedingt nur auf die Artikel abstellen, sondern können alle dem Kassenbon zuordnungsfähigen Merkmale umfassen. In Abhängigkeit von den verwendeten Merkmalen und dem Aufwand, der zu ihrer Erhebung betrieben werden muß, sollen hier verschiedene Stufen der Kassenbonanalyse unterschieden werden:

1. Die erste Stufe, die für das Data Mining relevant ist, besteht aus der o.g. Suche nach gerichteten Assoziationen zwischen Artikeln oder Warengruppen.
2. Die zweite Stufe der Kassenbonanalyse bezieht Merkmale des Kassenbons mit ein, wie z.B. Jahr, Monat, Wochentag, Uhrzeit, Kasse, Zahlungsart, Bonsumme, Anzahl Bonpositionen.

   *Beispielsweise können sich schon unter Hinzunahme der Uhrzeit interessante weitergehende Aussagen herauskristallisieren:*

   *WENN der Einkaufszeitpunkt nach 16 Uhr liegt*
   *DANN enthält der Kassenbon häufig alkoholfreie Getränke, Bier, Spirituosen.*

   *Offensichtlich kaufen viele Kunden nach Feierabend Getränke ein. Hieraus können zeit- und warengruppenbezogene Marketingmaßnahmen abgeleitet werden, wie z.B. Lautsprecherdurchsagen, die nach 16 Uhr auf Sonderangebote in der Getränkeabteilung hinweisen.*

[253] Vgl. MICHELS (1995), S. 38.

3. Die dritte Stufe der Kassenbonanalyse bezieht absatzfördernde Maßnahmen mit ein, wie z.B. ob eine Ware im Rahmen einer Sonderaktion präsentiert wurde. Damit können beispielsweise Spill-over-Effekte von Marketingmaßnahmen auf andere Warengruppen analysiert und bei zukünftigen Planungen berücksichtigt werden.[254]

4. Die vierte Stufe der Kassenbonanalyse bezieht Merkmale des Verkäufers mit ein, wie z.B. Geschlecht, besuchte Schulungen oder Qualifikationen. Hieraus lassen sich verkäuferbezogene Maßnahmen ableiten, wie z.B. der Besuch von Fortbildungen.

5. Die fünfte Stufe der Kassenbonanalyse bezieht regionale Merkmale mit ein, wie z.B. die Nähe bestimmter Konkurrenten. Dies kann die Sortimentsplanung beeinflussen.

   *Wird beispielsweise herausgefunden, das die Nähe der nächsten ALDI-Filiale einen starken negativen Einfluß auf die Anzahl der verkauften Konserven ausübt, so sollten in den Verkaufsstellen, die sich in der Nähe von ALDI-Filialen befinden, mit dem ALDI-Angebot konkurrierende Konserven aus dem Sortiment genommen werden – wenn dieser Entscheidung keine weiteren Verbundeffekte entgegenstehen.*[255]

6. Die sechste Stufe der Kassenbonanalyse bezieht den Kunden mit ein. Zahlt der Kunde mit einer Kreditkarte o.ä., so kann er identifiziert und früheren Einkäufen zugeordnet werden. Daher soll auf dieser Stufe die Kaufhistorie von Kunden betrachtet werden. Die resultierenden Datenmuster werden auch als „***Kaufsequenzen***“ bezeichnet.

   *Beispielsweise folgt auf den Kauf eines Mikrowellengerätes der Kauf von Mikrowellengeschirr, Fertiggerichten oder speziellen Kochbüchern.*

   Kaufsequenzen erklären oder prognostizieren einen zukünftigen Kundenbedarf. Wenn Kaufsequenzen als Erklärungsmodell dienen sollen, dann können die extrahierten Merkmale Anhaltspunkte zur Konzeption von Marketingmaßnahmen liefern, wie z.B. Sonderangebote für die Artikel, die weitere Bedarfe nach sich ziehen. Prognosemodelle werden in laufenden Geschäftsprozessen dazu eingesetzt, für aktuelle Artikelkäufe zukünftige Artikelbedarfe zu antizipieren. Die Anwendung derartiger Prognosemodelle ist Gegenstand von Abschnitt 3.2.4.

7. Die siebte Stufe der Kassenbonanalyse bezieht zusätzliche Stammdaten des Kunden mit ein. Diese sind dem Unternehmen beispielsweise dann bekannt, wenn es

---

[254] Vgl. MICHELS (2001), S. 947.

[255] Vgl. MICHELS (2001), S. 948.

eigene Kundenkarten ausgibt, die den Kunden gewisse Vorteile einräumen, wie z.B. einer Ratenzahlung, der Mitnahme von Auswahlartikeln, Rabatte oder der Erstattung von Parkgebühren.[256] Als Gegenleistung für diese Vorteile geben die Kunden persönliche Merkmale preis, wie z.B. ihren Beruf, ihr Alter, ihr Wohngebiet, ihre Haushaltsgröße oder den Träger der Kaufentscheidungen in ihrem Haushalt. Aus Datenmustern mit diesen Merkmalen lassen sich entsprechend Entscheidungen ableiten, die sich an den soziodemographischen und psychographischen Merkmalen orientieren, wie z.B. die regionale Differenzierung von Werbeprospekten.

Die gerichteten Abhängigkeiten zwischen Artikeln oder sonstigen zuordnungsfähigen Merkmalen eines Kassenbons werden auch als „***Assoziationsregeln***“ bezeichnet. Die Generierung von Assoziationsregeln ist in der Literatur sehr weit verbreitet. Dabei wird allerdings regelmäßig derselbe Fehler gemacht, der darin besteht, daß kein Optimierungs-, sondern ein Suchproblem gemäß Definition 2-6 bearbeitet wird. Die Problemstellung bei einem Suchproblem besteht nicht darin, die beste Regelmenge zu finden, sondern darin, alle zulässigen Assoziationsregeln der Form

WENN $A$ DANN $B$ (kurz: $A \rightarrow B$),

aufzuspüren, wobei die Zulässigkeit durch Mindestanforderungen an den *Support*, $|O^T[A \wedge B]|/|O^T|$, die *Konfidenz,* $|O^T[A \wedge B]|/|O^T[A]|$, und seltener den *Interest-Wert*, $|O^T[A \wedge B]|/(|O^T[A]| \cdot |O^T[B]|)$, geprüft wird.[257] Hier stellen $A$ und $B$ Mengen von Artikeln dar, d.h. eine Regel *Saft→Cola* wird gelesen als: „Wenn Saft eingekauft wird, dann wird gleichzeitig auch Cola eingekauft.“ Wie bereits in Abschnitt 2.1.3 diskutiert wurde, kann bei einem solchen Suchproblem nicht bewertet werden, wieviele Objekte durch keine Regel erfaßt werden. Weiter kann nicht berücksichtigt werden, inwieweit mehrere Regeln dieselben Objekte abdecken und möglicherweise widersprüchliche Aussagen repräsentieren.

Ein weiterer, weit verbreiteter Fehler bei der Generierung von Assoziationsregeln ist in der Verwendung der *Konfidenz* anstelle des *Interest-Wertes* zu sehen.[258] Es kommt

---

[256] Vgl. RÜTER (1994), S. 20 ff.

[257] Vgl. zu den beiden erstgenannten Interessantheitsfacetten Definition 2-55 („Allgemeingültigkeit“ oder „Support“) und Definition 2-52 („Sicherheit“ oder „Konfidenz“). Vgl. zum Interest-Wert: BRIN/MOTWANI/SILVERSTEIN (1997), S. 269).

[258] Vgl. BRIN/MOTWANI/SILVERSTEIN (1997), S. 269.

nicht darauf an, daß $A$ und $B$ häufig zusammen gekauft werden (relativ zu $A$ alleine), sondern darauf, daß $A$ und $B$ häufiger zusammen gekauft werden als man dies bei Unabhängigkeit von $A$ erwarten würde. Denn nur dann dürfen sich Entscheidungen, wie z.B. die erwähnte Elimination, an dieser Kaufverbundbeziehung orientieren.

Der Grund für diese beiden Fehler liegt wohl in der historisch bedingten Technikorientierung des Data Mining, da der Grundalgorithmus zur Generierung aller zulässiger Assoziationsregeln sehr elegant und schnell ist[259] und diese Kriterien bei der Entwicklung von Data-Mining-Verfahren oft im Vordergrund standen. Dies kann aus betriebswirtschaftlicher Sicht kein akzeptables Argument darstellen.

Der letzte hier zu behandelnde Anwendungsbereich von Erklärungsmodellen in der Situationsnalyse betrifft die **Erklärung der Varianten- oder Markenwahl** von Kunden. Sowohl die Varianten- als auch die Markenwahl sind Aspekte des beobachtbaren Kaufverhaltens. Sie unterscheiden sich darin, daß im ersten Fall die Entscheidung des Kunden für eine von mehreren Produktvarianten eines Herstellers und im zweiten Fall die Wahl zwischen den Marken verschiedener Hersteller erklärt werden soll. Falls die Einflußfaktoren auf die Varianten- oder Markenwahl ermittelt werden können, kann das Marketing an diesen Faktoren ansetzen, um seine Instrumente daran auszurichten.

*Beispielsweise behandeln DECKER und TEMME u.a. die Frage, wie die Produktprospekte eines Automobilherstellers neu zu gestalten sind.[260] Bei der Neugestaltung der Prospekte möchte der Hersteller sich daran orientieren, wie wichtig die einzelnen Fahrzeugeigenschaften aus Kundensicht sind, d.h. in welchem Maße sie geeignet sind, den Kauf einer bestimmten Fahrzeugvariante zu erklären. Als Datenbasis dienen daher Umfragedaten zu psychographischen Variablen, wie z.B. der subjektiven Bedeutung der Innenraumgestaltung, des Fahrkomforts oder der Motorleistung für die Kaufentscheidung.*

### 3.2.3 Anwendungen zur Diagnose von Entscheidungssituationen

Eine Entscheidungssituation stellt, wie bisher, eine Umweltsituation dar, die als Problem wahrgenommen wird und daher einer Lösung, d.h. einer Entscheidung und deren Umsetzung, bedarf. Bei der Diagnose geht es um das Schließen von einer konkret beobachteten Entscheidungs- oder Problemsituationen auf die vermutete Ursache. Im Gegensatz zu den Anwendungen aus den Abschnitten zuvor, in denen das generierte

[259] Vgl. AGRAWAL ET AL. (1996), S. 310 ff.

[260] Vgl. DECKER/TEMME (2001), S. 671 ff.

Data-Mining-Modell „nur“ analysiert wurde, wird es hier auf neue Daten angewendet, welche die Problemsituation beschreiben. Wenn die jeweilige Problemursache erst einmal diagnostiziert ist, dann kann modellextern über Maßnahmen zur Behebung des Problems entschieden werden.

Abbildung 3-6 ordnet die Anwendungen zur Diagnose von Entscheidungssituationen in die Phasen betrieblicher Entscheidungsprozesse ein. Man erkennt, daß die Diagnose der Planungsphase zugeordnet ist. Diese Zuordnung wurde eingangs des Abschnittes 3.2 vorgenommen, da ein Diagnosemodell auf neue Daten, die eine aktuelle Entscheidungssituation beschreiben, angewendet wird und damit einen engen Entscheidungsbezug besitzt.

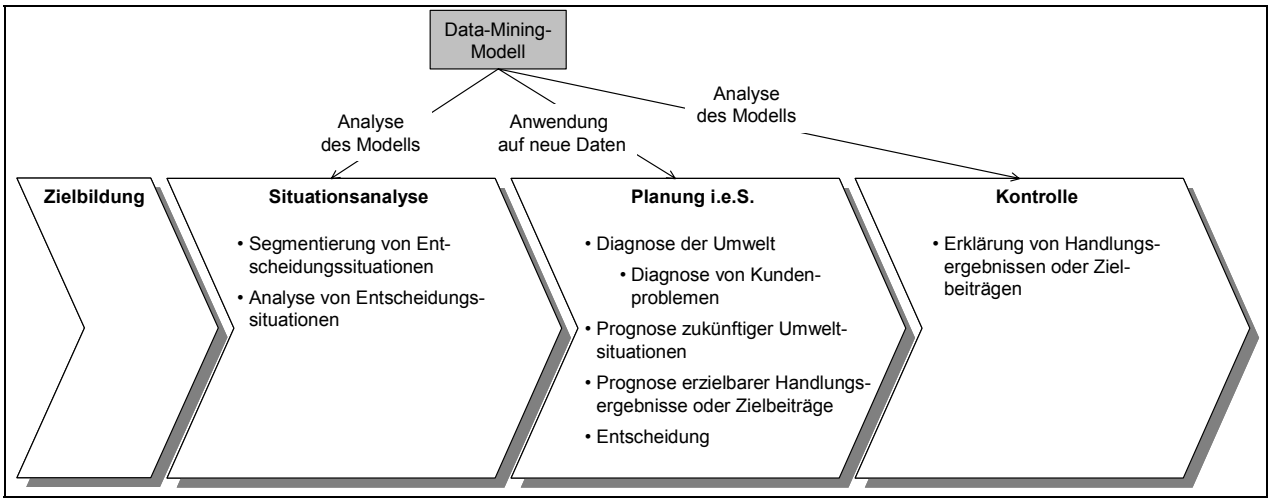



**Abbildung 3-6: Anwendungen zur Diagnose der Umwelt**

Große Anwendungsbereiche sind die Diagnose von Krankheiten[261] und die Diagnose von Produktmängeln und technischen Fehlern[262]. Stärker betriebswirtschaftlich orientiert ist die Diagnose von Kundenproblemen. Ist die Ursache für das Kundenproblem erst einmal bekannt, so hat der Kundenservice anschließend zu entscheiden, wie es gelöst werden soll.

*So können beispielsweise Kundenprobleme durch ein Call Center entgegengenommen und die wahrscheinlichste Problemursache durch Anwendung eines Diagnosemodells festgestellt werden. Das Diagnoseergebnis kann zur Planung des technischen Außendienstes verwendet werden.*

[261] Vgl. MCLEISH ET AL. (1991), S. 480 ff.

[262] Vgl. HÄTÖNEN ET AL. (1996), S. 116 f. und KÜPPERS (1999), S. 136 f.

Bei der Diagnose wird von Beobachtungen auf deren Ursache geschlossen, auch wenn dieser Schluß nach der strengen Logik nicht zulässig ist, da es mehrere mögliche Ursachen für dieselbe Beobachtung geben kann.[263] In Regelform läßt sich eine Diagnose wie folgt darstellen:

```
WENN beobachtete Problemsituation = ... DANN Ursache = ...
```

Diagnoseprobleme können nach ihrer Struktur zu den Prognoseproblemen gezählt werden, welche im nächsten Abschnitt behandelt werden. Die beiden Problemtypen unterscheiden sich nur darin, daß bei der Prognose zukünftige Sachverhalte vorhergesagt werden, während bei der Diagnose Sachverhalte festgestellt werden, die zwar auch gegenwärtig unbekannt, die aber bereits in der Vergangenheit aufgetreten sind und das aktuelle Problem verursacht haben.

### 3.2.4 Anwendungen zur Prognose zukünftiger Umweltsituationen

Die Prognose zukünftiger Umweltsituationen durch ein Prognosemodell dient dazu, für ein gegebenes Planungsobjekt die *erwartete* Umweltentwicklung vorherzusagen. Dies kann nicht mit der Modellierung der zukünftigen Umweltzustände beim Aufbau eines Entscheidungsmodells gleichgesetzt werden, da für eine Entscheidung unter Risiko die Kenntnis *aller möglichen* Umweltentwicklungen und deren Wahrscheinlichkeitsverteilung vorausgesetzt wird.

Data-Mining-Prognosemodelle werden häufig in folgenden Anwendungsbereichen eingesetzt:

- ⇨ Prognose von Finanzmarktwerten;
- ⇨ Antizipation zukünftiger Kundenbedarfe;
- ⇨ Prognose der Liquidität von Kreditnehmern;
- ⇨ Prognose der Dauer von Kundenbeziehungen.

Abbildung 3-7 ordnet diese Anwendungsbereiche in die Phasen betrieblicher Entscheidungsprozesse ein.

---

[263] Vgl. MEYER-FUJARA/PUPPE/WACHSMUTH (1993), S. 745.

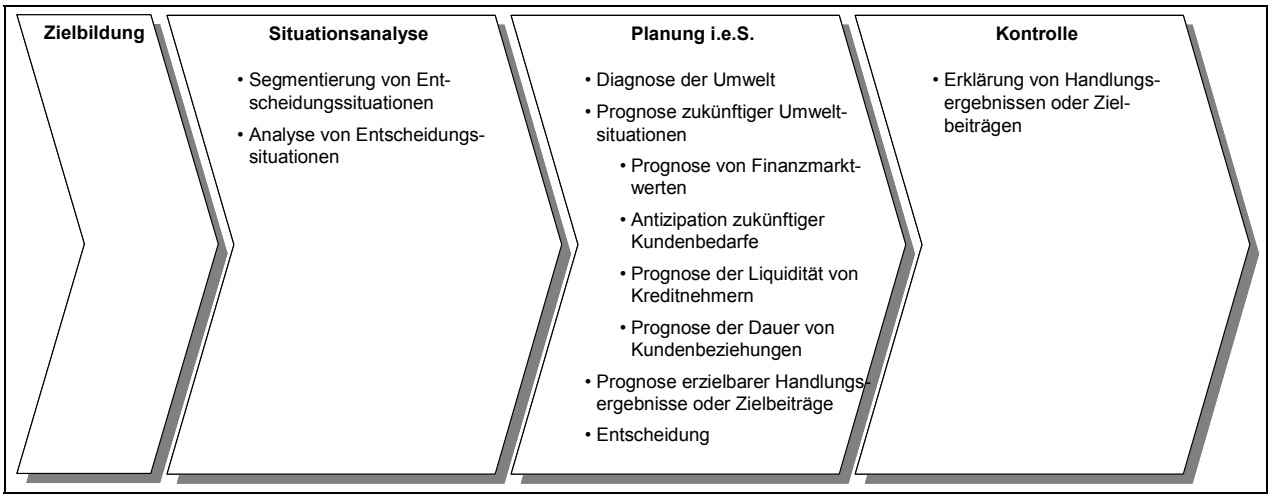



**Abbildung 3-7: Anwendungen zur Prognose zukünftiger Umweltsituationen**

Zu den Anwendungen in der **Prognose von Finanzmarktwerten** zählen die Aktienkursprognose[264] die Zinssatzprognose[265] und die Wechselkursprognose[266]. Betrachtet werden soll hier die häufigste Anwendung, die Aktienkursprognose. Man unterscheidet traditionell die Fundamentalanalyse und die technische Aktienanalyse.[267] Bei der ***Fundamentalanalyse*** werden Prognosemodelle mit unternehmensindividuellen, branchenspezifischen und gesamtmarktbezogenen Variablen aufgestellt. Dagegen werden bei der ***technischen Aktienanalyse*** Prognosemodelle aufgestellt, die ihre Vorhersagen primär aus der Zeitreihe selbst sowie aus dem Aktienmarkt zugehörigen Größen trifft. Das Data Mining bietet hier durch seine Ausrichtung auf Problemstellungen mit sehr vielen Einflußgrößen eine Möglichkeit, die Variablen aus der Fundamentalanalyse und der technischen Aktienanalyse zu kombinieren.

Die Prognose von Aktienkursen wird fast immer zur *Unterstützung von Kaufen-, Halten- oder Verkaufen-Entscheidungen* verwendet. Unter gewissen Voraussetzungen[268] kann hier ein Entscheidungs- anstelle eines Prognosemodells aufgestellt werden. Wenn diese Voraussetzungen nicht gegeben sind, so muß man sich mit der Generierung eines

[264] Einen Überblick über Anwendungen speziell neuronaler Netze in der Aktienkursprognose gibt BAUN (1994), S. 194 ff.

[265] Vgl. PODDIG (1994), S. 254 ff.

[266] Vgl. z.B. RAUSCHER (1998), S. 333 ff.

[267] Vgl. zu dieser Unterscheidung BAUN (1994), S. 142 ff.

[268] Abschnitt 3.2.6 geht auf die Voraussetzungen zur Unterstützung von Anlageentscheidungen durch Data-Mining-Verfahren ein.

Prognosemodells begnügen. Die meisten Prognoseverfahren und -modelle liefern allerdings, wie gesagt, nur die erwartete Prognose und nicht die vollständige Wahrscheinlichkeitsverteilung der möglichen Kursentwicklungen, die zur Auswahl der optimalen Alternative notwendig wäre.

Die **Antizipation zukünftiger Kundenbedarfe** ist von den in Abschnitt 3.2.2 angesprochenen Kaufsequenz-Mustern abzugrenzen, bei denen es nur auf die Reihenfolge der Artikelkäufe ankam. Im Gegensatz dazu werden hier aus vergangenen Artikelkäufen eines konkreten Kunden dessen zukünftige Bedarfe möglichst zeitpunktgenau vorhergesagt. Bei vielen Direktmarketingmaßnahmen, wie z.B. dem Angebot für eine Probefahrt mit einer neuen Automobil-Variante, hat der Zeitpunkt der Ansprache einen großen Einfluß auf die Wirksamkeit der Aktion.[269] Darüber hinaus unterstützt die Prognose von Kundenbedarfen die Bestimmung des Angebotsinhaltes, d.h., die Frage, *welche Produkte* angeboten werden sollen.

Die zeitpunktgenaue Vorhersage der Kundenbedarfe dient nicht der Neukonzeption von Maßnahmen, wie dies bei den Erklärungsmodellen der Fall war. Vielmehr werden die antizipierten Bedarfe unmittelbar in einer bereits existierenden Maßnahme verwendet, z.B. zur Selektion von Zielkunden für eine zeitpunktgerechte Kundenansprache.[270] Damit ist auch hier ein enger Entscheidungsbezug gegeben, so daß die Bedarfsprognosen nur dann gerechtfertigt sind, wenn die Generierung eines Entscheidungsmodells an einem unzureichenden Verfahren oder an unzureichenden Trainingsdaten scheitert. Wann genau ein Entscheidungsmodell durch Data-Mining-Verfahren erstellt werden kann, wird noch zu dikutieren sein.

Die Anwendungen zur **Prognose der Liquidität von Kreditnehmern** sind auch unter den Bezeichnungen „Bonitätsanalyse“, „Bonitätsbewertung“, „Kreditwürdigkeitsanalyse“ oder „Kreditwürdigkeitsprüfung“ bekannt. In den meisten Ausprägungen geht es um die Prognose einer Kennzahl, die zur Unterstützung von Entscheidungen über Kreditvergaben herangezogen wird. Diese Kennzahl charakterisiert das Rückzahlungspotential des Kunden, also eine Umweltgröße. Damit kann zumindest das kundenindividuelle Bonitätsrisiko eingegrenzt werden. Andere Risiken der Kreditvergabe, wie z.B. das

[269] Vgl. BAUSCH (1991), S. 86.

[270] Die Selektion von Zielkunden wird in den Abschnitten 3.2.5 und 3.2.6 noch einmal aufgegriffen.

Zinsänderungsrisiko, das Geldwertrisiko, das Währungsrisiko (bei Fremdwährungskrediten) oder das Besicherungsrisiko (bei der Stellung von Kreditsicherheiten), müssen eigens analysiert werden.[271]

I.d.R. wird mit der Prognose der Liquidität oder Kreditwürdigkeit unmittelbar eine Kreditentscheidung verbunden, die jedoch modellextern getroffen wird. Die extrene Entscheidungsfindung ist dann sinnvoll, wenn tatsächlich noch externe Überlegungen wie die Aussicht auf Folgeaufträge oder die Berücksichtigung des Zinsänderungsrisikos in die Entscheidungsfindung einfließen. Falls externe Überlegungen aber ohnehin ignoriert werden und jeder Antragsteller, für den das Prognosemodell eine hohe Bonitätskennzahl vorhersagt, einen Kredit bekommt, dann sollte u.U. eher ein Entscheidungs- als ein Prognosemodell generiert werden. Denn in einem Entscheidungsmodell können Risiko- und Zielvorstellungen der Entscheidungsträger integriert werden. Das entsprechende Vorgehen im Data Mining wird noch herauszuarbeiten sein.

Bei Anwendungen zur **Prognose der Dauer von Kundenbeziehungen** wird der erwartete Zeitpunkt abgeschätzt, zu dem ein Kunde seine Geschäftsbeziehungen zum Unternehmen kündigt und „abwandert“. Eigene Handlungen zur Verlängerung der Kundenverweildauer bleiben dabei i.d.R. unberücksichtigt.

Im Gegensatz zu der in Abschnitt 3.2.2 angesprochenen Erklärung von Kundenabwanderungen werden die hier behandelten Anwendungen der Planungsphase zugeordnet, da hier nicht die permanente Beobachtung der Umwelt, wie sie in der Situationsanalyse stattfindet, einen Handlungsbedarf offenlegt. Vielmehr wurde hier bereits ein Handlungsbedarf registriert, und nun ist für einen konkreten Kunden eine Entscheidung zu treffen, für die die Kenntnis der Kundenverweildauer von Bedeutung ist. Die Verweildauer soll also durch ein Prognosemodell vorhergesagt werden und dann extern eine übergeordnete Entscheidung unterstützen.

*Beispielsweise könnte eine Versicherungsgesellschaft vor dem Entscheidungsproblem stehen, die Akquisitionsbudgets kundenorientiert verteilen zu wollen, d.h. in die Akquisition von Neukunden, von denen man sich eine lange Vertragslaufzeit erwartet, soll entsprechend mehr investiert werden als in Personen, die potentiell weniger aussichtsreich sind. Wenn es anhand der Eigenschaften von zu besuchenden Interessenten gelänge, bereits bei der Akquisitionsplanung für jeden Kunden dessen voraussichtliche Vertragslaufzeit zu prognostizieren, so könnte der Akquisitionsaufwand an die durch die Kunden erwarteten Einnahmen angepaßt werden.*

---

[271] Vgl. SCHMIDT-VON RHEIN/REHKUGLER (1994), S. 493.

Wie das Beispiel zeigt, dient auch die Prognose der Kundenverweildauer unmittelbar einer Entscheidung. Falls keine modellexternen Überlegungen in die Entscheidung einfließen, empfiehlt sich auch hier die Generierung eines entsprechenden Entscheidungsmodells, das nicht etwa die Kundenverweilzeit prognostiziert, sondern direkt die Entscheidung (im Beispiel: das anzusetzende Aquisitionsbudget) trifft.

### 3.2.5 Anwendungen zur Prognose erzielbarer Handlungsergebnisse oder Zielbeiträge

Die Prognose erzielbarer Handlungsergebnisse oder Zielbeiträge verfolgt, wie die zuvor diskutierte Prognose von Umweltsituationen, bestimmte Parameter eines übergeordneten betriebswirtschaftlichen Entscheidungsmodells zu ermitteln. Im folgenden werden Anwendungen diskutiert, sich zumeist damit begnügen, in einer gegebenen Entscheidungssituation für eine bestimmte Handlungsalternative das erwartete Handlungsergebnis bzw. den erwarteten Zielbeitrag vorherzusagen. Die übergeordnete Entscheidung würde jedoch besser unterstützt, wenn das Prognosemodell in einer Entscheidungssituation für eine bestimmte Handlungsalternative *alle möglichen* Handlungsergebnisse bzw. Zielbeiträge und deren Wahrscheinlichkeitsverteilung ausgeben würde. Die Unterschiede in der Entscheidungsfindung werden noc herauszuarbeiten sein.[272]

In diesem Zusammenhang sind Data-Mining-Anwendungen zur **Prognose von Kundenreaktionen auf eine Direktmarketingaktion** weit verbreitet. Abbildung 3-8 ordnet diese Anwendungen in die Phasen betrieblicher Entscheidungsprozesse ein.

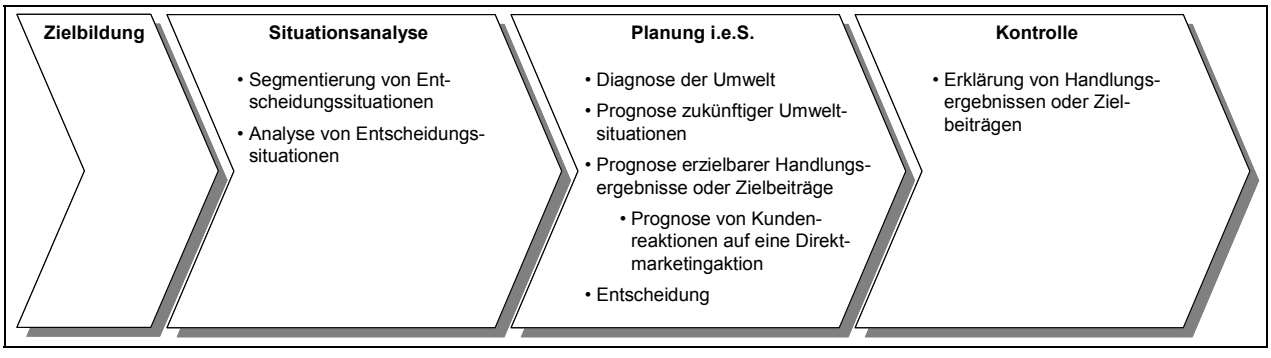



**Abbildung 3-8: Anwendungen zur Prognose von Handlungsergebnissen oder Zielbeiträgen**

[272] Dies erfolgt in Abschnitt 3.3.3.

Im *Direktmarketing* tritt das Unternehmen über geeignete Medien, wie z.B. Briefpost, Telefon oder E-Mail, in direkten Kontakt zum Kunden, um bei diesem eine überprüfbare Reaktion auszulösen.[273] Geht man davon aus, daß für eine Direktmarketingaktion, z.B. für ein Produktangebot, die wesentlichen Aktionsparameter, wie die Produktauswahl, der Zielmarkt, das Werbemedium, das Werbemittel, der Preis, die Zahlungskonditionen sowie die Lieferbedingungen, geklärt sind, stellt sich abschließend die Frage, welche Kunden man in die Direktmarketingaktion mit einbezieht. Die Einsparung überflüssiger Kundenkontakte reduziert die Aktionskosten und die Gefahr der Verärgerung von Kunden.[274]

Prognosemodelle sollen bezüglich eines gegebenen Aktionstyps für jeden Kunden – vorausgesetzt, er wird für die Aktion ausgewählt – dessen Reaktion auf die Direktmarketingaktion vorhersagen. Die Kenntnis der Kundenreaktion wird fast immer zur *Selektion von Zielkunden für eine Marketingmaßnahme* verwendet, so daß ein direkter Entscheidungsbezug vorliegt. Ein Kunde wird modellextern dann für eine Aktion ausgewählt, wenn er eine gewisse Bestell- bzw. Kaufwahrscheinlichkeit überschreitet.[275] Wenn aber modellextern ohnehin keine weiteren Überlegungen in die Entscheidungsfindung einfließen, dann kann auch gleich die Entscheidung modellintern und damit automatisch erfolgen. Entsprechende Anwendungen werden in dem nächsten Abschnitt diskutiert.

Die Reaktion auf eine Direktmarketingaktion kann nur dann vorhergesagt werden, wenn sie für eine Teilmenge von Kunden bekannt ist, anhand derer ein Prognosemodell erlernt werden kann. Solche Reaktionsdaten erhält man durch Pre-Tests, in denen einer Testauswahl von Kunden ein Sonderangebot o.ä. unterbreitet und deren Reaktionen erfaßt werden.

*Beispielsweise soll eine Direktmarketingaktion gestartet werden, in denen den Kunden eine CD-Serie zum Kauf angeboten wird.[276] Um die benötigten Reaktionsdaten zusammenzutragen, wird innerhalb eines Pre-Test eine Teilmenge von Kunden selektiert, denen die CD-Serie vorab angeboten wird. Nach einer fest*

---

[273] Vgl. BAUSCH (1991), S. 86.

[274] Vgl. PIATETSKY-SHAPIRO/MASAND (1999), S. 186.

[275] Der Schwellwert für die Kaufwahrscheinlichkeit kann exakt bestimmt werden, wie in Abschnitt 3.3.2.1 gezeigt wird.

[276] Vgl. BAUSCH (1991), S. 87 ff.

*definierten „Wartezeit" werden alle Kunden, die bislang nicht mit einem Kauf reagiert haben, als „Nichtkäufer" vermerkt.*

Wenn der Aufwand für solche Pre-Tests gescheut wird, versucht man, die Entscheidung ohne die Reaktionsdaten zu treffen. Zu diesem Zweck setzt man die Selektionsentscheidung für einen Kunden mit der Affinität des Kunden zu dem angebotenen Produkt gleich. Die Affinität kann dabei durch ein Modell vorhergesagt werden, welches anhand der gegebenen Kundendaten erlernt wird.

*Beispielsweise prognostizieren ARNDT, GERSTEN und WIRTH die Markenaffinität im Automobilsektor – sie klassifizieren Kunden anhand von zugeordneten Lifestyle-Daten in Mercedes-Besitzer und -Nichtbesitzer und setzten die Klasse mit der Auswahl bzw. Nichtauswahl neuer Kunden für eine Direktmarketingaktion gleich.[277] TIETZ, POSCHARSKY, ERICHSON und MÜLLER prognostizieren die Affinität von Bankkunden zu einem speziellen Sparvertrag – auch sie setzten die Nutzung bzw. Nichtnutzung des Sparvertrags durch aktuelle Kunden mit der Auswahl bzw. Nichtauswahl neuer Kunden für eine Direktmarketingaktion gleich.[278]*

Problematisch an der Gleichsetzung der Produktaffinität und der Selektionsentscheidung ist, daß auf diese Weise aus den Trainingsdaten keine Kaufwahrscheinlichkeit (sondern nur die Wahrscheinlichkeit für eine gewisse Produktaffinität) approximiert werden kann. Die Kaufwahrscheinlichkeit ist aber eine relevante Größe, da sie determiniert, ob ein Kunde selektiert werden soll oder nicht.

Auch wenn die benötigten Reaktionsdaten vorliegen, können bestimmte Probleme auftreten. So sind bei der reinen Prognose der binären Größe „*Kunde reagiert / reagiert nicht*" die beiden Ausprägungen i.d.R. sehr ungleich verteilt. Reagieren beispielsweise auf ein Produktangebot *1%* der kontaktierten Kunden, so liegt ein triviales Prognosemodell, das für alle Kunden „*reagiert nicht*" vorhersagt, in *99%* der Fälle richtig.[279] Dieses Problem läßt sich relativ einfach umgehen, indem man das Modell anhand der Reduzierung des Prognosefehlers gegenüber der trivialen Prognose bewertet.[280] Gravierender ist das folgende, oben bereits angedeutete Problem: Die Entscheidung, ob ein Kunde für eine Aktion selektiert werden soll oder nicht, beeinflußt unmittelbar den ökonomischen Erfolg der Aktion. Damit sollte hier eher ein Entscheidungs- als ein Prognosemodell aufgestellt werden.

---

[277] Vgl. ARNDT/GERSTEN/WIRTH (2001), S. 594 ff.

[278] Vgl. TIETZ ET AL. (2001), S. 769 ff.

[279] Vgl. LING/LI (1998), S. 74.

[280] Vgl. zur Reduzierung des Prognosefehlers: HILBERT (1998), S. 68 ff.

In der Literatur wurde der direkte Entscheidungsbezug zwar teilweise erkannt – dies zeigen die folgenden Anwendungsbeispiele – aber trotztem wird regelmäßig kein Entscheidungsmodell aufgestellt, sondern andere Lösungsversuche eingeschlagen.

*Beispielsweise stellen ITTNER, SIEBER und TRAUTZSCH ein Modell zur Klassifikation von Kunden als Reagierer bzw. Nichtreagierer auf das Direktangebotes eines Buches auf.[281] Für die Reagierer ergeben sich mehrere Merkmalsprofile, die sie von den Nichtreagierern abgrenzen. Die Profile beschreiben jeweils ein Kundensegment, für das sich Kontaktkosten angeben lassen. Da das Buch einen festen Preis hat, lassen sich für die Segmente auch Deckungsbeiträge ermitteln. Ein Segment wird entweder vollständig kontaktiert oder von der Aktion ausgeschlossen.*

*MASAND und PIATETSKY-SHAPIRO versuchen in einer ähnlichen Anwendung, aus Kunden- und Transaktionseigenschaften die Inanspruchnahme eines angebotenen Telekommunikationsdienstes vorherzusagen.[282] Nach der Modellgenerierung und der Anwendung auf die unbekannten Kunden wird für jeden Kunden eine durch das Prognosemodell ermittelte „Kauf"-Wahrscheinlichkeit mit dem Mittelwert der früheren Telekommunikationsausgaben des Kunden multipliziert, um den erwarteten Erlös und (nach Abzug der Kontaktkosten) den Deckungsbeitrag pro Kunde zu ermitteln. Ein zweiter Ansatz der Autoren besteht darin, direkt die erreichbaren Deckungsbeiträge zu prognostizieren.[283] Allerdings wird der über alle Kunden erzielte Deckungsbeitrag erst nach Anwendung des Modells berechnet.*

Die Anwendungen lassen sich so zusammenfassen, daß jeweils versucht wird, ein Prognosemodell zu generieren und bei der Anwendung des Modells den erzielbaren Deckungsbeitrag zu ermitteln. Sind die Deckungsbeiträge bekannt, so können für die Marketingaktion, falls kein Engpaß in Form eines Budgets besteht, alle Kunden mit positivem Deckungsbeitrag selektiert werden. Ist das für die Aktion zur Verfügung stehende Kapital beschränkt, so ist der relative Deckungsbeitrag das relevante Selektionskriterium.[284] Der relative Deckungsbeitrag ermittelt sich als Quotient aus absolutem Deckungsbeitrag und den Kontaktkosten pro selektiertem Kunden. Da dieser Kostenwert i.d.R. über alle Planungsobjekte konstant ist, genügt es, die Kunden nach ihren absoluten Deckungsbeiträgen zu ordnen und, beim höchsten Deckungsbeitrag angefangen, solange Kunden zu selektieren, bis das Budget ausgeschöpft ist.

Alle in den Anwendungen beschriebenen Möglichkeiten – die Prognose von Kundenreaktionen, Reaktionswahrscheinlichkeiten oder Deckungsbeiträgen für Kundensegmente oder einzelne Kunden – führen nicht zur deckungsbeitragsoptimalen Kundenselektion, da der Gesamtdeckungsbeitrag erst ermittelt wird, nachdem das Modell bereits

---

[281] Vgl. ITTNER/SIEBER/TRAUTZSCH (2001), S. 715 ff.

[282] Vgl. MASAND/PIATETSKY-SHAPIRO (1996), S. 198.

[283] Vgl. MASAND/PIATETSKY-SHAPIRO (1996), S. 199.

[284] Vgl. KILGER (1993), S. 839 f.

generiert wurde. Würde dagegen bereits während der Modellgenerierung eine Erfolgsgröße optimiert, so würde ein Modell generiert, das besonders erfolgswirksame Zusammenhänge besonders gut approximiert. Dies wird bei der Generierung von Entscheidungsmodellen durch die unterschiedliche Gewichtung verschiedener Arten von Modellfehlern berücksichtigt. So ist es i.d.R. günstiger, einen Nichtkäufer fälschlicherweise für einen Käufer zu halten und ihm ein Produktangebot zu unterbreiten, als einen Käufer fälschlicherweise für einen Nichtkäufer zu halten und ihm kein Produktangebot zukommen zu lassen. Eine nachträgliche Erfolgsbewertung von Prognosemodellen ist nur zu rechtfertigen, wenn kein Verfahren zur Generierung von Entscheidungsmodellen zur Verfügung steht oder wenn modellexogene Faktoren in die Entscheidungsfindung einfließen sollen.

### 3.2.6 Anwendungen zur Entscheidung für eine optimale Handlungsalternative

Data-Mining-Entscheidungsmodelle unterstützen den Entscheidungsträger in der Planungsphase. Sie liefern unter bestimmten Voraussetzungen optimale Entscheidungen für konkrete Planungsobjekte (Entscheidungssituationen). Zu diesen Voraussetzungen zählen:

- ⇨ die allgemeinen Einsatzvoraussetzungen des Data Mining;[285]
- ⇨ das Vorhandensein eines geeigneten Verfahrens zur Generierung von Data-Mining-Entscheidungsmodellen;
- ⇨ die während der Problemstrukturierung getroffenen Annahmen, wie z.B. das Absehen von bestimmten Variablen oder Interdependenzen. Im Data Mining betrifft dies regelmäßig Einflüsse, die nicht aus den Trainingsdaten hervorgehen, wie z.B. das Absehen von Pressemeldungen bei der Aktienanlage oder die Berücksichtigung von Erfahrungen aus vergangenen Kundenkontakten bei der Marketingaktionsplanung.

Die beiden letzten Abschnitte haben bereits gezeigt, daß viele Anwendungen einen engen Entscheidungsbezug aufweisen. Trotzdem werden diese Problemstellungen in den betrachteten Literaturquellen nicht durch Entscheidungs-, sondern durch Prognosemodelle angegangen, die bestimmte Parameter von Entscheidungsmodellen vorhersagen

[285] Die allgemeinen Einsatzvoraussetzungen des Data Mining wurden in Abschnitt 2.3 behandelt.

konnten. Die eigentliche Entscheidung wurde dann aber immer modellextern getroffen, was von Vorteil ist, wenn modellexterne Überlegungen in die Entscheidungsfindung einfließen sollen. Für einfach strukturierte Entscheidungsprobleme, die keine externe Einflußnahme erfordern, kommt eine automatisierte Entscheidungsfindung aber durchaus in Frage.

Die Potentiale der automatischen Entscheidungsfindung liegen in den folgenden Anwendungsfeldern:

⇨ Entscheidung über die Kundenselektion für Direktmarketingaktionen;

⇨ Entscheidung über die Vergabe von Krediten;

⇨ Entscheidung über die Prüfung von Verdachtsmomenten;

⇨ Entscheidung über den Handel mit Wertpapieren.

Abbildung 3-9 ordnet diese Anwendungsbereiche in die Phasen betrieblicher Entscheidungsprozesse ein.

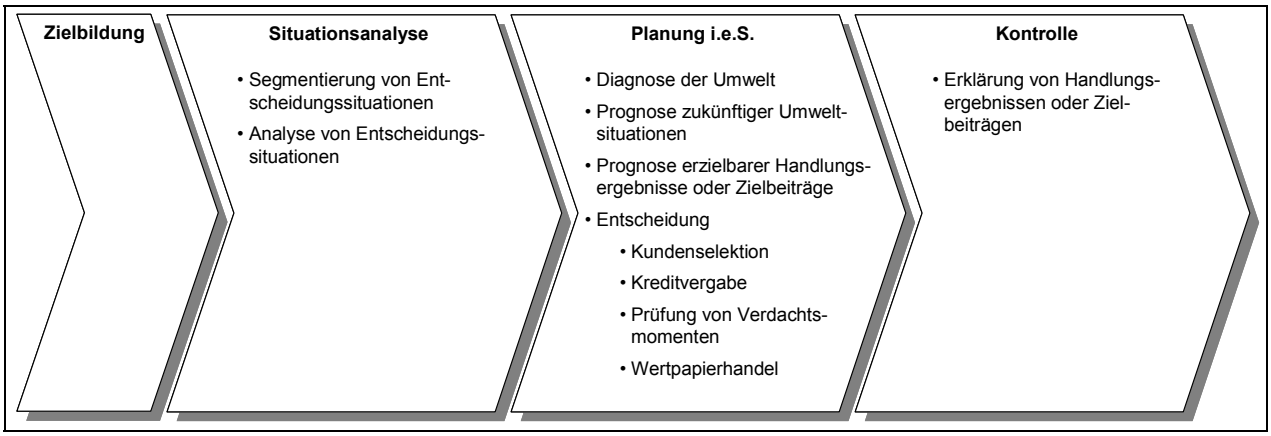



**Abbildung 3-9: Anwendungen zur Entscheidung für eine optimale Alternative**

Zunächst sollen Anwendungspotentiale des Data Mining zur **Entscheidung über die Kundenselektion für Direktmarketingaktionen** betrachtet werden. Dieses Feld wurde im Abschnitt zuvor bereits diskutiert, so daß die folgenden Ausführungen sich auf die Besonderheiten beschränken, die sich bei der Betrachtung der Kundenselektion als Entscheidungsproblem ergeben.

Eine Betrachtung der Kundenselektion als Entscheidungsproblem bedeutet, daß das produzierte Modell direkt die optimale Alternative ausgibt – und nicht etwa ein Handlungsergebnis, wie dies im Abschnitt zuvor der Fall war. Eine solche Zielsetzung führt

bei der Verwendung klassischer Data-Mining-Verfahren zu Problemen. Klassische Verfahren wie etwa neuronale Netze, Entscheidungsbaum- oder Entscheidungslistenverfahren produzieren Modelle, welche ausschließlich Variablen umfassen, die in den Trainingsdaten zur Verfügung stehen. Sie sind nicht in der Lage, eine neue Variable, wie die optimale Entscheidung, selbst zu erzeugen. D.h. die optimale Entscheidung muß in der Trainingsmenge bereits vorliegen, so daß die Trainingsmenge wie folgt aufgebaut ist:

$O^T \subseteq ES \times H.$

Dabei bezeichnet $ES$ die Menge aller möglichen Entscheidungssituationen, womit hier die vorliegenden Kundenmerkmale gemeint sind. Und $H$ bezeichnet die Menge der möglichen Handlungsalternativen, hier: $H = \{Kunde\ selektieren,\ Kunde\ nicht\ selektieren\}$. Daraus kann dann mit klassischen Verfahren ein funktionales Modell, $M$, abgeleitet werden, das zu einem gegebenen Kunden, $es \in ES$, die optimale Entscheidung, $h \in H$, ausgibt:

$$M: \quad ES \rightarrow H = \{Kunde\ selektieren,\ Kunde\ nicht\ selektieren\}$$
$$es \rightarrow M(es).$$

Eine derartige Modellierung, wie sie klassische Verfahren vornehmen, ist aus folgenden Gründen problematisch:

1. Die Präferenzvorstellungen der Träger vergangener Entscheidungen sind in den Trainingsdaten implizit enthalten. Damit werden sie auch durch das Data-Mining-Verfahren erlernt und fließen in das entstehende Data-Mining-Modell ein. Diese Fixierung im Modell kann dann gefährlich werden, wenn sich die Präferenzvorstellungen ändern.[286]
2. Der Entscheidungsträger hat keine Möglichkeiten mehr, eine andere Alternative zu wählen, da er deren Auswirkungen dem Modell nicht entnehmen kann.
3. Bei der Generierung des Modells kann keine ökonomische Zielfunktion optimiert werden. I.d.R. wird die Modellkorrektheit gemäß Definition 2-54 maximiert.[287] Dadurch wird nicht etwa auf die gute Empfehlung besonders erfolgskritischer

[286] Vgl. LACKES/MACK (2000), S. 62.

[287] Das damit verbundene Problem wurde bereits in Abschnitt 3.2.6 diskutiert.

Entscheidungen abgestellt, sondern auf eine im Durchschnitt über alle Entscheidungssituationen gute Empfehlung.

4. Das mit der Wahl einer Alternative verbundene Risiko bezüglich der zu erwartenden Zielbeiträge kann nicht in die Entscheidungsfindung einfließen.

Insbesondere der letzte Punkt wiegt so schwer, daß in der Literatur zumeist wie im Abschnitt zuvor besprochen vorgegangen wird und die Entscheidung modellextern erfolgt. Trotzdem stellt die Selektion von Zielkunden ein Entscheidungsproblem dar, das so einfach strukturiert ist, daß die Entscheidung auch modellendogen getroffen werden könnte – und zwar ohne die Problempunkte drei und vier. Dies setzt voraus, daß kein klassisches Verfahren verwendet wird, sondern ein neu zu entwickelndes, welches in der Lage wäre, die optimale Entscheidung als Variable in das Modell zu integrieren, ohne daß sie Bestandteil der Trainingsdaten ist.

Die Konzeption des Modelltyps und des Verfahrens wird an anderer Stelle nachgeholt. Da an dieser Stelle die Anwendungsmöglichkeiten von Data-Mining-Verfahren betrachtet werden, wird im folgenden das Problem der Zielkundenselektion als Entscheidungsproblem dargestellt.

Für jeden aktuellen oder potentiellen Kunden in der Adreßdatenbank ist zu entscheiden, ob ihm ein Angebot zugesendet werden soll. In Tabelle 3-1 ist die entsprechende Entscheidungsmatrix dargestellt. Falls dem Kunden ein Angebot zugesendet wird und der Kunde ein Kaufinteresse hat, erfolgt als Handlungsergebnis eine Bestellung. Dieses Handlungsergebnis kann durch einen positiven Deckungsbeitrag bewertet werden, der sich aus den direkt der Aktion zurechenbaren Erlösen und Kosten ermitteln läßt. Falls dem Kunden ein Angebot zugesendet wird und der Kunde kein Kaufinteresse hat, bestellt er nicht, und es fallen nur Kosten an. Außerdem ist der Kunde möglicherweise verärgert, was durch zusätzliche „Kosten“ bewertet werden kann. Falls kein Angebot versendet wird, erfolgt auch keine Bestellung. Wenn allerdings der Kunde ein Kaufinteresse hätte, entstehen Opportunitätskosten. Bei der Modellgenerierung wird dies dadurch berücksichtigt, daß ein Modell, welches alle Deckungsbeitragspotentiale ausschöpft (und außerdem keinen Nichtbesteller selektiert), am besten bewertet wird.

| | | zukünftige Umweltsituation | |
|---|---|---|---|
| | | **Kunde hat Interesse** | **Kunde hat kein Interesse** |
| **Handlungs-alternative** | **Angebot zusenden** | Kunde bestellt (positiver Deckungsbeitrag) | Kunde bestellt nicht (negativer Deckungsbeitrag) |
| | **Angebot nicht zusenden** | Kunde kann nicht bestellen (Deckungsbeitrag = 0) | Kunde kann nicht bestellen (Deckungsbeitrag = 0) |

**Tabelle 3-1: Handlungsergebnisse (und Bewertungsansätze) für die Versendung von Katalogen im Versandhandel**

Diese Entscheidungsmatrix zeigt, daß es sich bei der Zielkundenselektion um ein einfach strukturiertes Entscheidungsproblem handelt, bei dem externe Überlegungen keine Rolle spielen, so daß die Entscheidung für oder gegen die Selektion eines Kunden durchaus direkt durch ein automatisch generiertes Modell getroffen werden könnte. Die Konzeptionen des entsprechenden Modelltyps und der dazugehörigen Zielsetzungen für Data-Mining-Verfahren erfolgen in Abschnitt 3.3.2.1.

Ähnlich wie die Kundenselektionsentscheidungen eine Weiterentwicklung der Kundenreaktionsprognosen darstellen, kann man die **Entscheidung über die Vergabe von Krediten** als Weiterentwicklung der in Abschnitt 3.2.4 angerissenen Liquiditätsprognosen betrachten. Für jeden Antragsteller ist zu entscheiden, ob der beantragte Kredit bewilligt wird oder nicht. Falls der Kredit vergeben wird und der Kunde zu den Rückzahlungsterminen liquide sein wird, ist als Handlungsergebnis eine fristgerechte Rückzahlung zu erwarten, und ein positiver Betrag, wie etwa der Kapitalwert der Aus- und Einzahlungsströme, kann zur Bewertung dieses Handlungsergebnisses angesetzt werden. Falls der Kunde illiquide wird und seinen Kredit nicht fristgerecht zurückzahlt, muß ein negativer Betrag angesetzt werden, wenn man von dem Ausfall eines Großteils der Einzahlungen ausgehen kann. Falls kein Kredit vergeben wird, muß auch kein Kredit zurückgezahlt werden. Opportunitätskosten müssen, wie gesagt, nicht angesetzt werden. Die Entscheidungsmatrix läßt sich damit ähnlich aufbauen wie zuvor Tabelle 3-1.

Über die Höhe der anzusetzenden Beträge bestehen unterschiedliche Ansichten.[288] In der Praxis wird häufig lediglich eine möglichst hohe Modellkorrektheit nach Definition 2-54 ohne unterschiedliche monetäre Gewichtung der verschiedenen Fehlerarten

[288] Vgl. SCHMIDT-VON RHEIN/REHKUGLER (1994), S. 496.

angestrebt. Damit verzichtet man auf eine ökonomische Bewertung des Modells. In der Literatur sind zwar vereinzelt auch Anwendungen zu finden, die eine ökonomische Zielgröße optimieren – diese Optimierung findet aber regelmäßig erst nach Abschluß des Lernvorgangs statt.[289] Dann aber existiert das Modell schon, so daß es nicht möglich ist, ein Modell zu generieren, das besonders erfolgswirksame Zusammenhänge besonders gut abbildet.

Ein weiteres Entscheidungsfeld, das sich als *(2×2)*-Matrix modellieren läßt, tut sich in der **Entscheidung über die Prüfung von Verdachtsmomenten** auf. Falls ein Verdacht auf einen Betrugsversuch im Telekommunikations- oder Versicherungsbereich, im elektronischen Zahlungsverkehr[290] oder bei der Auswertung von Steuererklärungen aufkommt, so muß entschieden werden, ob dieser Verdachtsmoment geprüft werden soll. Die Prüfung ist mit Kosten verbunden; falls ein Betrugsversuch aufgedeckt wird, fallen Erlöse in Höhe des unterschlagenen Betrags (evtl. zuzüglich einer Strafe) an.[291]

Auch die **Entscheidung über den Handel mit Wertpapieren** läßt sich in einer – leicht veränderten – Entscheidungsmatrix abbilden (vgl. Tabelle 3-2). Die Prognose von Finanzmarktwerten wurde bereits in Abschnitt 3.2.4 angesprochen, so daß hier nur noch auf die Besonderheiten der Betrachtung als Entscheidungsproblem eingegangen werden muß.

Die Darstellung als Entscheidungsmatrix ist unter folgenden Voraussetzungen möglich:

- ⇨ Es wird ein einzelnes Wertpapier – ohne Berücksichtigung von Alternativanlagen oder Risikostreuungsüberlegungen – betrachtet.
- ⇨ Es gibt ein festes Budget für die Anlage in das betrachtete Wertpapier.
- ⇨ Der Prognosehorizont beträgt jeweils eine Periode.
- ⇨ Gewinne bzw. Verluste werden in der nächsten Periode unmittelbar realisiert.
- ⇨ Es fallen keine Transaktionskosten an.

---

[289] Beispielsweise führen HIPPNER und RUPP eine Kreditwürdigkeitsprüfung im Versandhandel durch. (Vgl. HIPPNER/RUPP (2001), S. 690 ff.) Für einen Kunden, der eine Bestellung vornimmt, soll geprüft werden, ob ihm ein Rechnungs- oder Ratenkauf angeboten werden kann. Eine ähnliche Anwendung zur Bonitätsprüfung im Konsumgütervertrieb über Außendienstmitarbeiter beschreiben BONNE und ARMINGER. (Vgl. BONNE/ARMINGER (2001), S. 653 ff. und BONNE (2000), S. 138 ff.)

[290] Vgl. KÜPPERS (1999), S. 124.

[291] Vgl. BONCHI ET AL. (1999), S. 177.

Damit läßt sich die Entscheidungsmatrix wie folgt konstruieren.[292] Falls das Wertpapier gekauft wird und sein Kurs steigt, kann aufgrund des unterstellten Verkaufs in der nächsten Periode eine Gewinnmitnahme verbucht werden. Falls das Wertpapier gekauft wird und sein Kurs fällt, muß ein Verlust verbucht werden. Falls das Wertpapier (eventuell durch einen Leerverkauf) verkauft wird, treten bei einem Kursabfall Gewinne und bei einem Kursabfall Verluste ein. Bei gleichbleibenden Kursen können weder Gewinne noch Verluste eintreten. Der Gewinn bzw. Verlust stellt sich jeweils in Höhe der Kursänderung ein.

| | | **Umweltsituation** | | |
|---|---|---|---|---|
| | | **Kurs steigt** | **Kurs fällt** | **Kurs unverändert** |
| **Handlungs-alternative** | **Kauf** | Gewinn | Verlust | weder Gewinn noch Verlust |
| | **Verkauf** | Verlust | Gewinn | weder Gewinn noch Verlust |

**Tabelle 3-2: Bewertungsansätze für den Handel mit Wertpapieren**

Trotz der etwas unterschiedlichen Bewertungsansätze können, wie Abschnitt 3.3.2.1 zeigt, alle hier vorgestellten Entscheidungsprobleme durch denselben Modelltyp abgebildet werden. Der Modelltyp des (Data-Mining-) Entscheidungsmodells wird zusammen mit der entsprechenden Data-Mining-Zielsetzung in Abschnitt 3.3.2.1 völlig neu entwickelt

### 3.2.7 Anwendungen zur Kontrolle von Handlungsergebnissen oder Zielbeiträgen

Ex-post versucht man im Entscheidungsprozeß, Ursachen für das Eintreten bereits realisierter Handlungsergebnisse oder Zielbeiträge aufzuspüren. Für unerwünschte Handlungsergebnisse oder Zielbeiträge, wie z.B. das Verfehlen einer Planvorgabe für eine Zielgröße, kann neben Umweltfaktoren die Qualität einer Entscheidung oder die Qualität ihrer Umsetzung verantwortlich gemacht werden. Die Zusammenhänge zwischen eigenen Handlungen und den realisierten Ergebnissen oder Zielbeiträgen werden als Erklärungsmodelle abgebildet.

[292] Vgl. PODDIG/DICHTL/PETERSMEIER (2000), S. 418 ff.

In der Kontrollphase wird das Data Mining sehr häufig zur Erfolgsanalyse von Maschinenkonfigurationen[293] oder von Therapien[294] eingesetzt. Aufgrund des geringen betriebswirtschaftlichen Bezugs werden hier statt dessen folgende Anwendungsbereiche betrachtet:

⇨ Werbeerfolgskontrolle;

⇨ Kennzahlanalysen im Ergebniscontrolling;

⇨ Kontrolle der Nutzung von Web-Ressourcen;

⇨ Kontrolle der Kreditwürdigkeitsprüfung.

Abbildung 3-10 ordnet diese Anwendungsbereiche in die Phasen betrieblicher Entscheidungsprozesse ein.

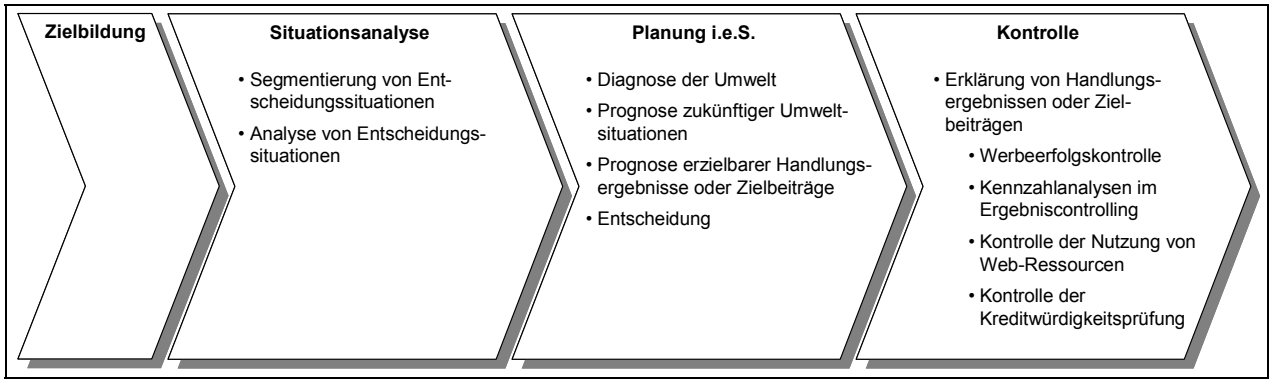



**Abbildung 3-10: Anwendungen zur Kontrolle erzielter Handlungsergebnisse oder Zielbeiträge**

Die **Werbeerfolgskontrolle** dient der Kontrolle der Erreichung geplanter Werbewirkungen. Es wird empfohlen, die Werbeplanung und -kontrolle nach der Wirkungskette zwischen dem Kontakt mit der Werbung und der Kaufentscheidung zu strukturieren – z.B. in Kontakt-, Bekanntheits-, Einstellungs- und Kaufverhaltenswirkungen.[295] Für jede Stufe dieser Wirkungskette lassen sich Kennzahlen definieren, die als zu erklärende Größe für das Data Mining in Frage kommen.

*Beispielsweise kämen mit*
*$a$: Anzahl erreichter Personen,*
*$b$: Anzahl Zielpersonen mit einer bestimmten Aufmerksamkeitsstufe,*

[293] Vgl. DECKER/FOCARDI (1995), S. 25.

[294] Vgl. KÜPPERS (1999), S. 147.

[295] Vgl. JANßEN (1999), S. 147 ff.

*c: Anzahl Zielpersonen mit aktiver Markenbekanntheit,*
*d: Anzahl Zielpersonen mit einer bestimmten Markeneinstellung,*
*e: Anzahl der Zielpersonen, die die betrachtete Marke kaufen*
*folgende zu erklärende Größen in Frage:*[296]

- ⇨ *als Kontaktkennzahl: b/a,*
- ⇨ *als Bekanntheitskennzahl: c/b,*
- ⇨ *als Einstellungskennzahl: d/c und*
- ⇨ *als Kaufverhaltenskennzahl: e/d.*

Als erklärende Größen kommen alle beobachtbaren und erfaßbaren Kundenmerkmale in Betracht, die die zu kontaktierenden Zielkunden charakterisieren. Da die zu erklärenden Kennzahlen Aggregationen über alle durch die erklärenden Merkmale beschriebenden Kunden darstellen, müßten die Kennzahlen durch das Data-Mining-Verfahren selbst berechnet werden. Damit kommen klassische Verfahren, wie etwa neuronale Lernverfahren oder Entscheidungsbaumalgorithmen, hierzu nicht in Frage. Die generierten Wirkungszusammenhänge zwischen erklärenden und zu erklärenden Größen können der Planung weiterer Werbemaßnahmen dienen.[297]

**Kennzahlanalysen im Ergebniscontrolling** dienen der Ergründung von Ursachen für bestimmte Ergebniskennzahlen. Das Ergebnis eines Geschäftsbereiches kann absolut (z.B. in Umsätzen oder Deckungsbeiträgen) oder relativ (z.B. in Umsätzen pro Vertriebsmitarbeiter) gemessen werden. Eine höhere Aussagekraft haben vergleichende Kennzahlen. Ein Kennzahlvergleich kann über die Zeit, zwischen Ist- und Sollwerten, zwischen verschiedenen Produktgruppen, Vertretern oder Regionen stattfinden (z.B. Umsatz- oder Deckungsbeitragwachstum in verschiedenen Perioden, Abweichung zwischen Plan- und Istkosten, Rabattgewährung aller Hauptvertreter als Histogramm).

Dieser Mehrdimensionalität des Betriebsergebnisses trägt das Umsatz- und Deckungsbeitragscontrolling Rechnung, indem es das Ergebnis differenziert nach mehreren Bezugsobjekten, wie z.B. Produkten, Kunden, Regionen und Vertriebskanälen, ausweist.[298] Stehen ebenso differenzierte Planergebnisrechnungen zur Verfügung, so lassen sich Plan-Ist-Abweichungen auf ihre Hauptverursacher in den einzelnen Dimensionen zurückführen. Der Verursacher kann sowohl eine eigene Handlung (z.B. der

[296] Die Kennzahlen wurden in Anlehnung an JANßEN aufgestellt (vgl. JANßEN (1999), S. 168).

[297] Vgl. JANßEN (1999), S. 165.

[298] Vgl. BISSANTZ (1996), S. 35.

Einsatz eines bestimmten Vertreters) als auch eine Umweltgröße sein (z.B. eine bestimmte Kundengruppe).

Das klassische Vorgehen des Controllers ist bisher durch überwiegend manuelle Tätigkeiten gekennzeichnet. I.d.R. fängt der Controller mit einer hoch verdichteten Kennzahl an (z.B. mit einer hohen negativen Erlösabweichung einer Sparte gegenüber der Vorperiode oder dem Planwert) und spaltet sie nach und nach auf, um ihre Hauptverursacher aufzuspüren. Der Analyseweg kann sich dabei über Regionen, Bezirke, Teilbezirke, Sparten, Artikelgruppen, Kundengruppen und Vetreter erstrecken. Diese Top-Down-Analyse wird durch *OLAP-Werkzeuge* unterstützt. Deren Unterstützungspotential liegt vor allem in einer grafischen Darstellung der aktuellen Analysesicht und schnellen Wechseln zwischen den Sichten.[299] Die Analysewege müssen im OLAP-Konzept intellektuell durch den Controller beschritten werden. Dabei orientiert sich der Controller an Heuristiken, wie z.B. der Streuung der Kennzahlausprägungen über eine Analysedimension. Diese Top-Down-Analyse kann durch eine rechnergestützte Abbildung der Heuristiken weitgehend automatisiert werden.[300]

Das Data Mining kann hier in einer anderen Art und Weise unterstützen, die auch als „Bottom-Up-Ansatz“[301] bezeichnet werden kann. Damit ist die Charakeristik von Data-Mining-Verfahren gemeint, aus unverdichteten Daten generelle Aussagen zu induzieren, welche Aggregate großer Datenmengen darstellen. Damit ist der Vorteil gegenüber dem Top-Down-Ansatz verbunden, daß die Analyse nicht auf fest vorgegebene Ergebniskennzahlen und Analysehierarchien festgelegt ist.[302]

Die **Kontrolle der Nutzung von Web-Ressourcen** ist ein weiteres Anwendungsgebiet des in Abschnitt 3.2.1 angesprochenen *Web Usage Mining*. Im Gegensatz zu den dort dargestellten Anwendungen werden hier Anwendungen von Erklärungsmodellen behandelt. Das zu erklärende Phänomen ist hier die Nutzung einer bestimmten Web-Ressource, wie z.B. einer Bestellfunktion oder einer Versorgung der Nutzer mit Echzeit-Informationen (Börsenkurse, Nachrichten o.ä.). Die erklärenden Größen können – je

[299] Vgl. JAHNKE/GROFFMANN/KRUPPA (1996), S. 321.

[300] Vgl. HAGEDORN (1996), S. 99 ff.

[301] Vgl. BISSANTZ (1996), S. 40.

[302] Vgl. BISSANTZ (1996), S. 40.

nach Ausnutzung der Reihenfolgeinformationen aus den Protokolldateien des Web-Servers – zeitlich geordnete oder reihenfolgeunabhängige Seitenaufrufe sein. Außerdem können sie – je nach Identifikationsmöglichkeit einzelner Nutzer – sitzungsübergreifend sein oder sich auf jeweils eine Sitzung beziehen.[303] Eine entsprechende Analyse bei sitzungsübergreifender Identifikation der Nutzer wird als „***Sequenzanalyse***" bezeichnet.[304] Dabei stehen zusätzlich zu den sitzungsübergreifenden „Click"-Historien u.U. Stammdaten der Nutzer als erklärende Größen zur Verfügung.

Wenn erst einmal die Ursachen für die Nutzung bestimmter Web-Ressourcen bekannt sind, kann versucht werden, Einfluß auf diese Ursachen zu nehmen. Insbesondere kann sich das Marketingcontrolling, wenn die Nutzer der Ressourcen genauer charakterisiert sind, bemühen, diese Nutzer individuell anzusprechen.

*Beispielsweise kann, falls einzelne Kunden sich durch Benutzername und Paßwort identifizieren und die Reihenfolge der Seitenaufrufe analysiert wird, erklärt werden, aufgrund welcher Seitenaufrufe ein Benutzer eine Bestellung vornimmt. Das Unternehmen kann aufgrund der Kundenidentifikation durch ein kundenindividuelles Echtzeit-Marketing eine Reduktion der Kontakthäufigkeit anstreben, z.B. durch das dynamische Generieren von Sonderangeboten.*[305]

Innerhalb einer Sitzung kann man bei der bereits angesprochenen *Pfadanalyse* versuchen, die Erreichung bestimmter Zielseiten über bestimmte Pfade zu erklären, um Ansatzpunkte zu deren Maximierung zu ermitteln.

*Beispielsweise interessiert die Anzahl der Nutzer, die von einer Einstiegsseite aus durch eine Gruppe von Pfaden eine bestimmte Web-Ressource, z.B. eine Bestellfunktion, gefunden haben. Dieses Maß wird auch als „**Konvertierungseffizienz**"*[306] *bezeichnet. Die Betrachtung einer bestimmten Gruppe von Pfaden läßt eine Unterscheidung zu, ob die Zielseite über besonders lange oder kurze Pfade erreicht wurde. Beispielsweise könnten zur Maximierung der Konvertierungseffizienz für kurze Pfade Links angepaßt, bestimmte Seiten entfernt oder attraktiver gestaltet werden. SPILIOPOULOU und BERENDT führen dies im Rahmen einer Non-Profit-Site (www.schulweb.de) vor.*[307] *Auf eine kommerzielle Site ausgerichtet ist die Studie von WEINGÄRTNER zur Erklärung der Online-Softwarebestellung.*[308] *Hier konnte u.a. herausgefunden werden, daß die zur Bestellung erforderlichen Seitenaufrufe und Registrierungsvorgänge nicht den Vorstellungen der Nutzer entsprachen. Dies führte – zu einem Zeitpunkt, an dem die Nutzer sich eigentlich*

---

[303] Vgl. zur Identifizierung von Nutzern und Sitzungen: SPILIOPOULOU (2001), S. 494 ff.

[304] Vgl. BENSBERG/WEIß (1999), S. 430.

[305] Vgl. zu diesem Beispiel: GROB/BENSBERG (1999), S. 20.

[306] SPILIOPOULOU (2001), S. 505

[307] Vgl. SPILIOPOULOU/BERENDT (2001), S. 856 ff.

[308] Vgl. WEINGÄRTNER (2001), S. 890 ff.

*bereits für einen Kauf entschieden hatten –zu Fehlbedienungen der Registrierungsfunktion und zu Abbrüchen der Bestellvorgänge.*[309]

Bei der **Kontrolle der Kreditwürdigkeitsprüfung** geht es um die Frage, ob bewilligte Kreditanträge zu recht bewilligt wurden. So sei für die Entscheidung über die Kreditvergabe aus Abschnitt 3.2.6 die Situation gegeben, daß die Datenbasis nur die Merkmale derjenigen Antragsteller umfaßt, deren Kreditantrag bewilligt wurde. Die Outputgröße gebe an, ob der Kredit fristgerecht zurückgezahlt wurde. Dann kann nicht zwischen kreditwürdigen und nicht kreditwürdigen Kunden diskriminiert werden, da die Merkmale der abgelehnten Anträge fehlen. Dennoch kann u.U. ein Erklärungsmodell generiert werden, daß für die bewilligten Anträge mit der Charakterisierung $Pr$ anzeigt, ob eine signifikant erhöhte Wahrscheinlichkeit einer Fristüberschreitung besteht, d.h. ob gilt:

$$P(\text{Fristüberschreitung} \mid \text{Kredit bewilligt} \wedge Pr) > P(\text{Fristüberschreitung} \mid \text{Kredit bewilligt}).$$

Falls dem so ist, muß das Kreditinstitut die durch $Pr$ charakterisierten Anträge besser prüfen. D.h. das Erklärungsmodell wird nicht etwa auf neue Kreditanträge angewendet, sondern dient der Verbesserung des Geschäftsprozesses „Antragsprüfung“.

### 3.2.8 Zusammenfassende Betrachtung der Data-Mining-Anwendungen

Die Betrachtung der Data-Mining-Anwendungen hat gezeigt, daß auch im Data Mining die Modelltypen des Beschreibungs-, Erklärungs-, Prognose- und Entscheidungsmodells unterschieden werden können. Ein Spezialfall war die Diagnose, bei der die Ausgabegröße vergangene Umweltsituationen, wie z.B. Krankheiten, Kundenprobleme oder Produktprobleme, darstellt. In jedem Fall sind die Umweltsituationen zum Zeitpunkt der Prognose unbekannt, so daß auch Diagnose- zu den Prognoseproblemen gezählt werden können. Die Unterscheidung in Beschreibungs-, Erklärungs-, Prognose- und Entscheidungsmodelle ist in der Data-Mining-Literatur nur ansatzweise vorhanden, wie ein Rückblick auf die Abgrenzung von Data-Mining-Aufgaben in Abschnitt 2.1.3 offenbart.

Wie beispielsweise die Anwendungen zur Marktsegmentierung und zur Erklärung von Kundeneinstellungen gezeigt haben, können Data-Mining-Beschreibungs- und –Erklärungsmodelle als Informationslieferanten für strategische Entscheidungen

[309] Vgl. WEINGÄRTNER (2001), S. 902.

dienen.[310] Sie unterstützen einmalige Entscheidungen, die von anderer Struktur sind als die Trainingsdaten. Die Entscheidungsunterstützung besteht darin, daß die Modelle einen Handlungsbedarf offenlegen und damit einen Planungsprozeß in Gang setzen. Damit ist die Phase der Situationsanalyse abgeschlossen, doch die Informationen aus den Beschreibungs- und Erklärungsmodellen können bei der weiteren Problemstrukturierung die Konkretisierung eines betriebswirtschaftlichen Entscheidungsmodells unterstützen. **Data-Mining-Beschreibungsmodelle** charakterisieren Segmente von Entscheidungssituationen zu dem Zweck, betriebswirtschaftlich orientierte Handlungen auf die spezifischen Eigenschaften des Segmentes zuzuschneiden. Damit unterstützen sie in erster Linie die Konkretisierung der Handlungsmöglichkeiten des Entscheidungsträgers, denn idealerweise geht mit jedem gebildeten Segment eine eigene Handlung einher (im Beispiel der Marktsegmentierung: eine eigene Produktkonfiguration). **Data-Mining-Erklärungsmodelle** dienen der Erklärung von Entscheidungssituationen in der Situationsanalyse und der Erklärung von Handlungsergebnissen oder Zielbeiträgen in der Kontrollphase. Die Zusammenhänge zwischen den analysierten Größen können bei der Konstruktion eines betriebswirtschaftlichen Entscheidungsmodells den Aufbau von Wirkungs- oder Zielerreichungsfunktionen unterstützen.

Zur direkten Entscheidungsunterstützung können Data-Mining-Prognose- und -Entscheidungsmodelle im Zusammenhang mit bereits erkannten und strukturierten Standard-Entscheidungen eingesetzt werden. Die Standard-Entscheidungen zeichnen sich dadurch aus, daß sie häufig unter vergleichbaren Rahmenbedingungen wiederholt werden. Zu beachten ist lediglich, daß das Modell, je nach Dynamik des Anwendungsbereiches, von Zeit zu Zeit aktualisiert werden muß. Im Gegensatz zu Beschreibungs- und Erklärungsmodellen werden diese Modelle auf neue Daten angewendet. Diese „neuen Daten“ charakterisieren Situationen, in denen eine Entscheidung zu treffen ist, also z.B. einen Kreditantrag, der angenommen oder abgelehnt werden soll. Die Entscheidungssituationen, auf die die Modelle angewendet werden, besitzen dieselbe Struktur wie die Trainingsmenge. Dabei können die Entscheidungen modellendogen durch ein **Entscheidungsmodell** getroffen oder durch die **Prognose** von Kenngrößen (Umweltsituationen, Handlungsergebnissen oder Zielbeiträgen) unterstützt werden.

---

[310] Vgl. zur Unterstützung strategischer Entscheidungen auch: DETERMANN/REY (1999), S. 146.

Solche hochstrukturierten und regelmäßig auftretenden Probleme sind eindeutig dem *operativen* Management zuzuordnen. Falls identisch strukturierte Entscheidungssituationen sehr häufig innerhalb eines Geschäftsprozesses auftreten, kann die entsprechende Data-Mining-Applikation als *Instrument zur Geschäftsprozeßunterstützung* eingestuft werden.

*Beispielsweise könnte innerhalb des Geschäftsprozesses „Kreditantragsbearbeitung" einer Bank ein Modell zur Kreditwürdigkeitsbewertung zum Einsatz kommen, das den Entscheidungsträger bei der Kreditvergabe unterstützt. Das Prognosemodell wird aus Kreditantragsdaten erlernt, die für jeden Kreditantrag identisch strukturiert sind. Angewendet wird das Modell auf neue Kreditanträge, die eine zu den Trainingsdaten analoge Struktur aufweisen.*

Data-Mining-Erklärungs- und -Beschreibungsmodelle können die Ablauforganisation eines Unternehmens dann beeinflussen, wenn sie einen Handlungsbedarf zur Umgestaltung von Geschäftsprozeßstrukturen offenlegen.

*Beispielsweise könnte ein Erklärungsmodell, das die ursächlichen Merkmale für eine geringe Rückzahlungsmoral von Kreditkunden offenlegt, Anlaß für das Kreditinstitut sein, innerhalb ihrer Kreditantragsbearbeitung diejenigen Anträge, die diese Merkmale aufweisen, genauer zu überprüfen. Wie schon das Prognosemodell wird auch dieses Erklärungsmodell aus Kreditantragsdaten erlernt, die für jeden Kreditantrag identisch strukturiert sind. Angewendet wird das Modell aber – anders als das Prognosemodell – bei der einmaligen Entscheidung, wie die Kreditantragsbearbeitung umstrukturiert werden kann.*

## 3.3 Konzeption eines Problemlösungsschemas für Data-Mining-Anwendungen

Im folgenden sollen zur Erfüllung des ersten aufgestellten Ziels (*„Lösungsschema konzipieren")* Lösungswege für gegebene Probemsituationen aufgezeigt werden. Zu dem Problemlösungsschema zählen folgende Schritte:

1. **Bestimmung der Problemklasse**: Zunächst wird die Problemsituation anhand der Trainingsmenge, der Planungsphase und der Möglichkeiten zur Entwicklung eines geeigneten Lösungsverfahrens erfaßt und in eine von 11 definierten Problemklassen eingeordnet. Dabei benötigt jede der 11 Problemklassen zur Lösung entweder ein Data-Mining-Entscheidungs-, -Prognose-, -Erklärungs- oder -Beschreibungsmodell.
2. **Konzeption eines Data-Mining-Ansatzes**: In Abhängigkeit von dem benötigten Data-Mining-Modelltyp ist ein Verfahren bereitzustellen, daß entsprechende Modelle

generieren kann. Die Anforderungen, die die Modelltypen an das Verfahren stellen, unterscheiden sich im wesentlichen durch die Interessantheitsbewertung.

3. **Anwendung des Verfahrens zur Entscheidungsunterstützung**: Nach der Bereitstellung eines Verfahrens erfolgt die Modellgenerierung. Wenn die erzeugten Modelle für verwendbar befunden werden, können sie zur Entscheidungsunterstützung eingesetzt werden. Diese erfolgt in Abhängigkeit von der definierten Problemklasse auf unterschiedliche Weise.

Diese drei Schritte werden in den folgenden drei Abschnitten eingehend behandelt.

### 3.3.1 Bestimmung der Problemklasse

Zunächst sollen die bei der Betrachtung der existierenden Data-Mining-Anwendungen getroffenen Aussagen zusammengefaßt werden, die sich auf die *Bestimmung der Problemklasse* beziehen.

Geht man davon aus, daß die in Abschnitt 2.3 herausgearbeiteten Anwendungsvoraussetzungen für den Einsatz von Data-Minig-Verfahren erfüllt sind, so enthält die Trainingsmenge eine ausreichend große Menge identisch strukturierter Planungsobjekte. Die Trainingsmenge, $O^T$, ist wie folgt aufgebaut:

$$O^T \subseteq (dom(SA) \times ES \times \bullet).$$

Dabei stehen, wie bisher, $SA$ für die Schlüsselattribute, $ES$ für die Menge der möglichen Entscheidungssituationen und $\bullet$ für noch zu konkretisierende Zusatzinformationen. Eine Entscheidungssituation, $es \in ES$, charakterisiert, wie bisher, eine als Problem wahrgenommene Merkmalswertekombination, die einer Lösung (Entscheidung) bedarf, z.B. die Merkmalsausprägungen eines Kreditantrags, für den eine Kreditentscheidung zu treffen ist. Die Zusatzinformationen, $\bullet$, welche bei der Bereitstellung der Trainingsmenge berücksichtigt werden müssen, sind nun wie folgt zu konkretisieren.

Zunächst ist danach zu differenzieren, ob eine einmalige Entscheidung oder eine hochstrukturierte Standardentscheidung zu treffen ist. Einmalige Entscheidungen wie die Konfiguration neuer Produkte oder die Umgestaltung von Geschäftsprozessen werden, wie gesehen, durch Beschreibungs- und Erklärungsmodelle unterstützt, die intellektuell interpretiert werden. Standardentscheidungen wie die Zielkundenselektion oder die

Kreditvergabe werden durch Prognose- und Entscheidungsmodelle unterstützt, die auf neue Daten angewendet werden. Dabei stellen die „neuen Daten“ konkrete Entscheidungssituationen dar, wie etwa die Beschreibung eines Kundenproblems, das diagnostiziert werden muß und für das der technische Kundendienst nun eine Lösung finden muß.

Das erste relevante Differenzierungskriterium bei der Unterstützung von Standardentscheidungen ist die Phase, in der sich der Entscheidungsprozeß befindet. Während in der Situationsanalyse Anhaltspunkte für einen Handlungsbedarf gesammelt werden, die in der Planungsphase zum Aufbau des Entscheidungsmodells beitragen können, beschäftigt sich die Kontrollphase mit der Analyse vergangener Entscheidungen und deren Umsetzung. Die Betrachtung der Data-Mining-Anwendungen hat gezeigt, daß die Situationsanalyse durch Data-Mining-Beschreibungs- und -Erklärungsmodelle unterstützt werden kann und die Kontrollphase ausschließlich durch Erklärungsmodelle.

Wie die betrachteten Anwendungen ergeben haben, stellen die beschriebenen Entscheidungssituationen manchmal *Indikatoren* für andere Merkmale dar, die nicht beobachtet werden können, wie z.B. für die Markenaffinität von Kunden oder für das Schadensrisiko von Versicherungsnehmern. Solche unbeobachteten Merkmale sind häufig gerade die entscheidungsrelevanten Merkmale. Könnten sie beobachtet werden, so würde man eher versuchen, diese Merkmale durch andere erfaßte Merkmale *zu erklären*. Beschreibungsmodelle werden eher dann gebildet, wenn solche Erklärungsversuche an dem Fehlen einer zu erklärenden Größe scheitern. Daher ist in der Situationsanalyse zu unterscheiden, ob in der Datenbasis erklärende und zu erklärende Größen differenziert werden können. Ohne geeignete zu erklärende Größen kommt nur noch die Generierung von Modellen zur **Segmentierung von Entscheidungssituationen** in Frage. Die Trainingsmenge ist dann wie folgt aufgebaut:

$O^T \subseteq (dom(SA) \times ES)$,

d.h. sie enthält neben den Attributen $SA$, die nur zu Identifikation der Datenobjekte dienen, ausschließlich Entscheidungssituationen aus $ES$, wie etwa ein Bündel assoziierter Waren oder eine Kombination von Kundenmerkmalen, die eine marktsegmentierungsrelevante Gruppe beschreiben.

Können dagegen erklärende und zu erklärende Größen unterschieden werden, z.B. Kundenmerkmale als erklärende Größen und die Einstellung zu einer Marke als zu erklärende Größe, dann kann der Zusammenhang zwischen diesen Größen durch ein Erklärungsmodell abgebildet werden. Die Trainingsmenge, $O^T$, ist dann wie folgt aufgebaut:

$O^T \subseteq (dom(SA) \times ES \times U)$.

Dabei bezeichnet $U$ die Menge der möglichen Umweltsituationen. $U$ bildet die Domäne für die zu erklärende Variable, die hier mit $u$ bezeichnet sei. Somit kann das Data Mining auch zur **Erklärung von Umweltsituationen** eingesetzt werden.

*In dem Beispiel der Erklärung der Einstellung zu einer Marke wäre $ES$ die Domäne der Kundenmerkmale und $U$={positive Einstellung, gleichgültige Einstellung, negative Einstellung} die Domäne der Einstellung gegenüber der Marke.*

In der Kontrollphase muß, wie gesagt, nicht mehr zwischen Erklärungs- und Beschreibungsmodellen unterschieden werden. Hier kommen Modelle zur **Erklärung von Handlungsergebnissen oder Zielbeiträgen** zur Anwendung, wie z.B. Modelle zur Erklärung einer Werbewirkungskennzahl. Diese modellieren dann *Wirkungs- oder Zielerreichungsfunktionen*, jenachdem, ob die Trainingsmenge neben den Entscheidungssituationen Handlungsergebnisse oder Zielbeiträge umfaßt:

$O^T \subseteq (dom(SA) \times ES \times E)$ bzw. $O^T \subseteq (dom(SA) \times ES \times Z)$.

Dabei bezeichnen $E$ die Menge der Handlungsergebnisse und $Z$ die Menge der Zielbeiträge.

Damit sind alle bei der Unterstützung einmaliger Entscheidungen zu differenzierenden Fälle abgehandelt. Abbildung 3-11 stellt die Bestimmung der Problemklasse bei einmaligen Entscheidungen im Überblick dar.

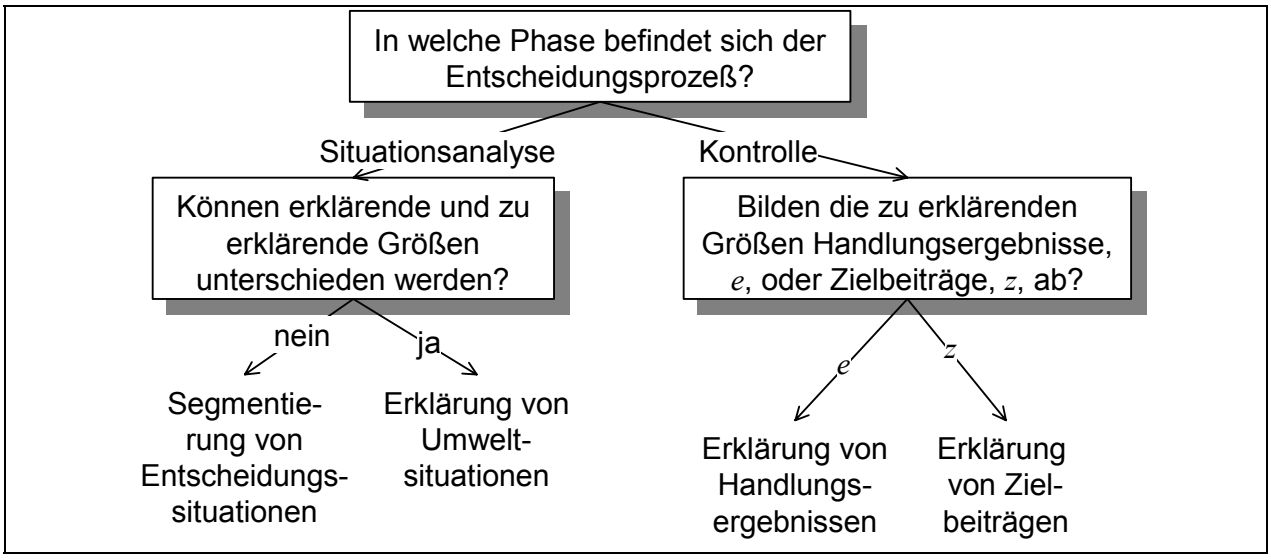



**Abbildung 3-11: Bestimmung der Problemklasse bei der Unterstützung einmaliger Entscheidungen**

Bleibt noch zu klären, wie die Problemklasse für das Data Mining bestimmt werden kann, wenn eine häufig zu wiederholende Standardentscheidung innerhalb eines existierenden Geschäftsprozesses, wie etwa die Kreditvergabe oder die Zielkundenselektion, zu treffen ist.

Handelt es sich bei der gegebenen Problemstellung um ein Standardproblem, dann kommt die Generierung von Prognose- oder Entscheidungsmodellen in Betracht. Beide Modelltypen lösen bereits erkannte und strukturierte Probleme in konkreten Entscheidungssituationen. Sie werden auf neue Situationen angewendet, die dieselbe Struktur wie die Trainingsmenge aufweisen. Allein die zu erklärende Variable fehlt bei den neuen Daten. Die Struktur der bereitzustellenden Trainingsmenge wird in Abhängigkeit von der Problemklasse wie folgt bestimmt.

Zunächst ist zu unterscheiden, ob ausschließlich klassische Data-Mining-Verfahren zur Verfügung stehen oder ob auch individuelle Lösungen entwickelt werden können. Klassische Verfahren, wie etwa neuronale Lernverfahren, Assoziationsregel- oder Entscheidungsbaumalgorithmen, zeichnen sich dadurch aus, daß sie Modelle produzieren, die ausschließlich Variablen enthalten, welche auch als Attribute in der Trainingsmenge vorhanden sind. D.h. falls z.B. Daten über Entscheidungs- und Umweltsituationen bereitgestellt werden können ($O^T \subseteq (dom(SA) \times ES \times U)$), kann zu einer neuen Entscheidungssituation auch nur eine Umweltsituation bzw. eine Verteilung von Umweltsituationen

vorhergesagt werden (und nicht etwa Zielbeiträge oder gar optimale Entscheidungen ausgegeben werden).

Andere Möglichkeiten ergeben sich, wenn man die Grenzen klassischer Verfahren aufhebt und die Entwicklung eines Verfahrens in Betracht zieht, das die zur Entscheidungsfindung fehlenden Komponenten, wie Zielbeiträge oder Nutzwerte, selbst berechnet. Dann könnte u.U. auch aus einer Datenbasis mit der o.g. Struktur ($O^T \subseteq (dom(SA) \times ES \times U)$) ein Entscheidungsmodell generiert werden.

*Beispielsweise sei ein Entscheidungsproblem betrachtet, bei dem eine Menge von Kunden zusammenzustellen ist, denen eine CD-Serie mit klassischer Musik angeboten werden soll. Die Datenbasis enthalte Kundenmerkmale und die Variable* $Interesse\ an\ klassischer\ Musik$ *mit der Domäne* $U = \{1,2,3,4,5\}$ *(*$5$ *= maximales,* $1$ *= minimales Interesse), wie sie etwa in einer Umfrage erhoben worden sein könnten. Dann könnte bspw. ein Entscheidungsbaumverfahren ein Modell generieren, das für Kunden, welche nicht an der Umfrage teilgenommen haben, ausgibt, wie hoch ihr Interesse an klassischer Musik ist. Extern würde man dann denjenigen Kunden ein Angebot unterbreiten, die mit einer bestimmten Mindestwahrscheinlichkeit ein bestimmtes Mindestinteresse an klassischer Musik besitzen. Der Entscheidungsbaum würde durch das Verfahren so generiert, daß seine Korrektheit (z.B. gemäß Definition 2-54) optimiert würde. Klassische Verfahren berücksichtigen nicht, daß etwa Kunden mit einem Interesse von* $6$ *erfolgswirksamer sind als Kunden mit einem Interesse von* $1$*. Gerade die erfolgswirksamen Zusammenhänge gilt es, besonders gut zu erlernen. Hierzu müßte ein Verfahren entwickelt werden, das in der Lage wäre, anhand der erklärenden Kundenmerkmale direkt die Entscheidung,* $Kunde\ selektieren \in \{ja, nein\}$ *zu treffen und die Zusammenhänge zwischen den erklärenden Merkmalen und der Entscheidung ökonomisch zu bewerten. Da in der Trainingsmenge nur Entscheidungssituationen (Kundenmerkmale) und Umweltsituationen (Interesse an klassicher Musik) vorliegen, müßte das Verfahren die Handlungsalternativen (*$Kunde\ selektieren \in \{ja, nein\}$*) kennen. Ebenso müßte es die Zielerreichungsfunktion kennen, welche aus der Handlungsalternative und dem Musikinteresse die Zielbeiträge berechnet, die quantifizieren, wie hoch ein bestimmtes Musikinteresse bei Selektion bzw. Nichtselektion des Kunden zu bewerten ist. Außerdem müssen die Risikopräferenzen des Entscheidungsträgers in dem Verfahren abgebildet sein, damit die Verteilung der Zielbeiträge bei Wahl einer bestimmten Alternative bewertet werden kann.*

Das Beispiel hat einen möglichen Fall verdeutlich, in dem die in der Datenbasis fehlenden Komponenten eines Entscheidungsmodells extern hinzugefügt werden. Insgesamt sind folgende mögliche Strukturen der Trainingsmenge zu unterscheiden:

⇨ Fall 1: Falls – wie in dem Beispiel – in der Datenbasis Entscheidungs- und Umweltsituationen abgebildet sind, hat die Trainingsmenge folgende Struktur:

$O^T \subseteq (dom(SA) \times ES \times U)$.

In diesem Fall müssen die Handlungsalternativen, $H$, die Zielerreichungsfunktion, $f^Z$, und die Ziel- und Risikopräferenzen in das Data-Mining-Verfahren integriert werden, um eine modellendogene Entscheidungsfindung zu ermöglichen.

⇨ Fall 2: Falls die Datenbasis Entscheidungssituationen und Handlungsergebnisse umfaßt, so sind die Handlungsergebnisse nur verwertbar, wenn man weiß, zu welcher Handlung sie gehören.

*Man betrachte als Beispiel die Selektion von Kunden für eine Direktmarketingaktion. Die für das Data Mining benötigte Datenbasis erhält man i.d.R. durch einen Pre-Test, in dem einer kleinen Stichprobe von Kunden ein Angebot zugesendet wird. D.h. es werden nur Daten für den Fall zusammengetragen, daß die Handlung $h^+$=„Angebot zusenden“ gewählt wird. Das Ergebnis für die alternative Handlung, $h^-$= „Angebot nicht zusenden“, ist ohnehin bekannt, da in diesem Fall keine Bestellung erfolgen kann. Die möglichen Ergebnisse als Reaktion auf die Direktmarketingaktion $h^+$ seien mit $E^+$ bezeichnet: $E^+$={Kunde bestellt, Kunde bestellt nicht}.*

Mit $E^+$ als Menge der Handlungsergebnisse, die sich bei Wahl der positiven Handlung, $h^+$, ergeben können, besitzt die Trainingsmenge folgende Struktur:

$O^T \subseteq (dom(SA) \times ES \times E^+$.

In diesem Fall müssen die alternative Handlung, $h^-$, die Zielerreichungsfunktion, $f^Z$, und die Ziel- und Risikopräferenzen in das Data-Mining-Verfahren integriert werden, um eine modellendogene Entscheidungsfindung zu ermöglichen. Der Umwelteinfluß ist implizit durch die Streuung der Handlungsergebnisse gegeben.

⇨ Fall 3: Ein ähnlicher Fall ist dann gegeben, wenn die Datenbasis statt der Handlungsergebnisse $E^+$ die Zielbeiträge $Z^+$ umfaßt, die sich bei Wahl der Handlung $h^+$ ergeben. Damit besitzt die Trainingsmenge die folgende Struktur:

$O^T \subseteq (dom(SA) \times ES \times Z^+$.

In diesem Fall müssen die alternative Handlung, $h^-$, ein eindeutiger Zielbeitrag, $z^-$, und die Ziel- und Risikopräferenzen in das Data-Mining-Verfahren integriert werden, um eine modellendogene Entscheidungsfindung zu ermöglichen.

*In dem Beispiel würde $Z^+$ die möglichen Werte der Bestellungen der Kunden in Folge der Direktmarketingaktion $h^+$={Kunde selektieren} bezeichnen. Der Zielbeitrag der negativen Entscheidung, $h^-$={Kunde nicht selektieren}, beträgt stets $z^-=0$.*

⇨ Fall 4: Ein Sonderfall ist dann gegeben, wenn die Datenbasis neben den Entscheidungssituationen unmittelbar die optimalen Handlungsalternativen, $H^*$, enthalten:

$O^T \subseteq (dom(SA) \times ES \times H^*)$.

Das Wertesystem des Entscheidungsträgers ist dann in solchen Entscheidungsempfehlungen bereits implizit enthalten.[311]

*Man betrachte als Beispiel Trainingsdaten mit Kreditanträgen und den zugehörigen Kreditentscheidungen:*
*ES: Merkmale der Kreditanträge;*
$H^*$*={Kredit vergeben, Kredit nicht vergeben}.*

⇨ Fall 5: Falls die Datenbasis Entscheidungssituationen, Handlungsalternativen und -ergebnisse umfaßt, besitzt die Trainingsmenge folgende Struktur:

$$O^T \subseteq (dom(SA) \times ES \times H \times E).$$

*Man betrachte als Beispiel Trainingsdaten mit potentiellen Versicherungskunden, den zur Akquisition der Kunden betriebenen Programme und den realisierten Laufzeiten der Versicherungsverträge:*
*ES: Merkmale von potentiellen Neukunden im Versicherungsgeschäft;*
*H: durch die Versicherung betriebene Akquisitionsprogramme;*
*E: Laufzeiten der Versicherungsverträge.*

In diesem Fall müssen die Zielerreichungsfunktion, $f^Z$, und die Ziel- und Risikopräferenzen in das Data-Mining-Verfahren integriert werden, um eine modellendogene Entscheidungsfindung zu ermöglichen.

⇨ Fall 6: Falls die Datenbasis Entscheidungssituationen, Handlungsalternativen und Zielbeiträge umfaßt, besitzt die Trainingsmenge folgende Struktur:

$$O^T \subseteq (dom(SA) \times ES \times H \times Z).$$

*Man betrachte als Beispiel wieder Trainingsdaten mit potentiellen Versicherungskunden, den zur Akquisition der Kunden betriebenen Programme und den Werten der Vertragslaufzeiten zum Planungszeitpunkt:*
*ES: Merkmale von potentiellen Neukunden im Versicherungsgeschäft;*
*H: durch die Versicherung betriebene Akquisitionsprogramme;*
*Z: abgezinste Versicherungsbeiträge über die Laufzeit der Verträge, abzüglich der anfänglichen Akquisitionskosten.*

Wird eine Software speziell zur Generierung von Data-Mining-Entscheidungsmodellen entwickelt, so ist eine echte **Entscheidung** durch das Modell möglich, wenn die fehlenden Entscheidungsmodellkomponenten durch das Data-Mining-Verfahren ergänzt werden. Können aber durch das Verfahren die Handlungsalternativen, $H$, die Zielerreichungsfunktion, $f^Z$, oder die Ziel- und Risikopräferenzen nicht ergänzt werden, so kann

[311] Vgl. DÜSING (1996), S. 143.Die Nachteile einer solchen Form der Entscheidungsfindung wurden bereits in Abschnitt 3.2.6 angesprochen.

jeweils nur die in der Trainingsmenge vorhandene Outputgröße prognostiziert werden. Konkret sind dann noch folgende Prognosen möglich:

- ⇨ im 1. Fall: **Prognose von Umweltsituationen**, $u \in U$, in der aktuellen Entscheidungssituation, $es \in ES$;
- ⇨ im 2. Fall: **Prognose von Handlungsergebnissen**, $e^+ \in E^+$, für die in der aktuellen Entscheidungssituation, $es \in ES$, implizit unterstellte Positiv-Handlung, $h^+ \in H$;
- ⇨ im 3. Fall: **Prognose von Zielbeiträgen**, $z^+ \in Z^+$, für die in der aktuellen Entscheidungssituation, $es \in ES$, implizit unterstellte Positiv-Handlung, $h^+ \in H$;
- ⇨ im 4. Fall: **„Prognose" der optimalen Entscheidung**, $h^* \in H$, in der aktuellen Entscheidungssituation, $es \in ES$;
- ⇨ im 5. Fall: **Prognose von Handlungsergebnissen**, $e \in E$, für die explizit vorliegende Handlung, $h \in H$, in der aktuellen Entscheidungssituation, $es \in ES$;
- ⇨ im 6. Fall: **Prognose von Zielbeiträgen**, $z \in Z$, für die explizit vorliegende Handlung, $h \in H$, in der aktuellen Entscheidungssituation, $es \in ES$.

Die fehlenden Entscheidungsmodellkomponenten werden in diesen Fällen nicht durch ein automatisches Verfahren, sondern intellektuell hinzugefügt. Abschnitt 3.3.3.3 beschreibt, wie die sechs möglichen Prognosen der modellexternen Entscheidungsfindung dienen können.

Abbildung 3-12 zeigt die Bestimmung der Problemklasse bei der Unterstützung häufig wiederholter Standardentscheidungen im Überblick.

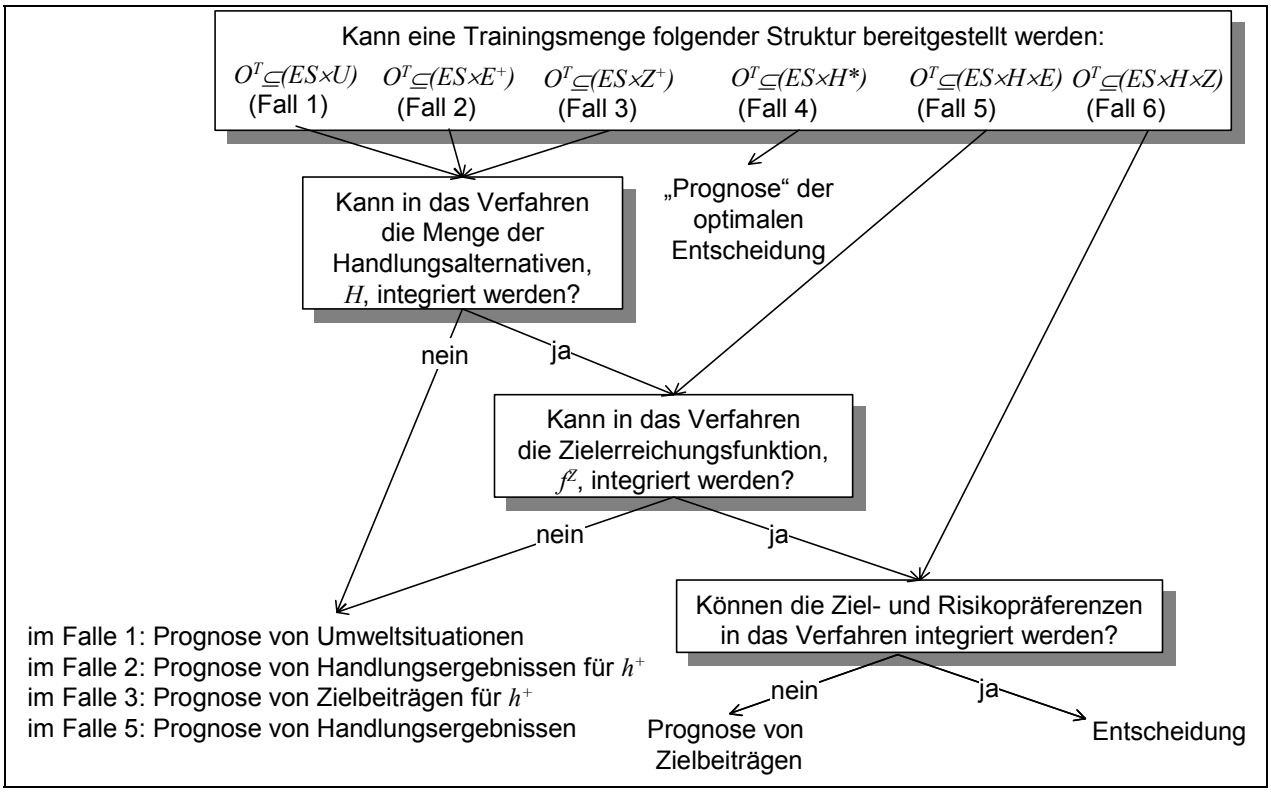



**Abbildung 3-12: Bestimmung der Problemklasse bei der Unterstützung von Standardentscheidungen**

### 3.3.2 Konzeption eines geeigneten Data-Mining-Ansatzes

Der zweite Schritt innerhalb des Problemlösungsschemas für Data-Mining-Anwendungen besteht in der Konzeption eines Data-Mining-Verfahrens, das geeignet ist, ein Modell des benötigten Typs zu generieren. Die Darstellung der Konzeption beschränkt sich dabei auf die wesentlichen Verfahrenskomponenten: den Aufbau und die Bewertung von Modellen des jeweiligen Typs. Die folgenden vier Abschnitte führen dies für die vier differenzierten Modelltypen vor. Anschließend faßt Abschnitt 3.3.2.5 die konzipierten Interessantheitsbewertungen zusammen und stellt sie gegenüber.

#### *3.3.2.1 Konzeption zur Generierung von Entscheidungsmodellen*

Im folgenden wird der in der Data-Mining-Literatur bisher nicht bekannte Modelltyp des „*Data-Mining-Entscheidungsmodells*" eingeführt.

Im allgemeinen ist es notwendig, für verschiedene Entscheidungssituationen verschiedene Entscheidungsmodelle aufzustellen. Im Data Mining werden aber ähnlich strukturierte Entscheidungssituationen betrachtet, für die dasselbe Entscheidungsmodell

verwendet werden kann. Damit die Entscheidungsfindung vollständig modellendogen stattfinden kann, muß das generierte Modell für eine neue Entscheidungssituation entweder

⇨ direkt die optimale Handlungsalternative oder

⇨ zu allen möglichen Alternativen deren Nutzwerte

liefern. Im folgenden wird von einer direkten Empfehlung der optimalen Alternative ausgegangen. Die zur Generierung eines solchen Modells notwendige Trainingsmenge, $O^T$, kann, wie der letzte Abschnitt gezeigt hat, verschiedene Komponenten eines Entscheidungsmodells enthalten und daher unterschiedlich aufgebaut sein. Im letzten Abschnitt wurden dabei sechs Fälle behandelt, die für den Aufbau von Entscheidungsmodellen relevant sind. Hier soll von unterschiedlichen Strukturen der Trainingsmenge abstrahiert werden. Wichtig ist hier nur, daß die Trainingsmenge in jedem Fall die Entscheidungssituationen umfaßt, in denen eine Entscheidung zu treffen ist. D.h. die Trainingsmenge besitzt folgende Struktur:

$O^T \subseteq (dom(SA) \times ES \times \bullet)$,

wobei $SA$ die Menge der Schlüsselattribute bezeichnet, $ES$ die Menge möglicher Entscheidungssituationen und $\bullet$ die Zusatzinformationen gemäß der Ausführungen aus dem Abschnitt zuvor.[312]

Ein Entscheidungsmodell läßt sich dann wie folgt definieren:

**Definition 3-2: Data Mining-Entscheidungsmodell**

Gegeben sei mit $H$ eine Menge unabhängiger Handlungsalternativen, mit $O=(dom(SA) \times ES \times \bullet)$ eine Menge von Planungsobjekten und mit $O^T \subseteq O$ eine Trainingsmenge. Ein *Data-Mining-Entscheidungsmodell*, $M^{Ent}$, ist ein funktionales Data-Mining-Modell gemäß Definition 2-22, das geeignet ist, in jeder möglichen Entscheidungssituation eine optimale Handlungsalternative, $h^* \in H$, zu empfehlen:

[312] Für die Zusatzinformationen gilt gemäß Abschnitt 3.3.1: $\bullet \in \{U, E^+, Z^+, H^*, U \times E, Z \times E\}$ mit $U$: Menge der Umweltsituationen, $E^+$: Menge der Handlungsergebnisse bei Wahl der positiven Handlung, $Z^+$: Menge der Zielbeiträge bei Wahl der positiven Handlung, $H^*$: Menge der für optimal befundenen Handlungsalternativen, $E$: Menge der Handlungsergebnisse und $Z$: Menge der Zielbeiträge.

$$M^{Ent}: \; O \rightarrow H;$$
$$o \rightarrow M^{Ent}(o) = h^*. \qquad \Diamond$$

Die *Aufgabe eines Data-Mining-Entscheidungsmodells* besteht in der Empfehlung von Entscheidungen für neue Inputdaten, $o \in O^{neu}$. Dabei gilt:

$O^{neu} \subseteq O,\ O^T \subseteq O,\ O^{neu} \cap O^T = \varnothing.$

Unterstützt werden können dabei nur *Standard-Entscheidungen* – das sind Entscheidungen, die bei stabilen Rahmenbedingungen in ähnlicher Weise für eine große Menge von Planungsobjekten getroffen werden. I.d.R. ist jede einzelne Entscheidung nicht von großer Bedeutung für den Unternehmenserfolg, sondern erst die Gesamtheit aller Entscheidungen.

Die Regeln in Data-Mining-Entscheidungsmodellen, $(Pr \rightarrow Ko) \in M_{O^T}$, werden hier als „***Entscheidungsregeln***" bezeichnet und besitzen folgende Form:[313]

WENN die Entscheidungssituation $Pr$ vorliegt,
DANN ist die Entscheidung $Ko$ optimal.

Dabei charakterisiert der Term $Pr$ die Entscheidungssituation, die dafür verantwortlich ist, daß einem Planungsobjekt, $o \in O[Pr]$, eine genau definierte, optimale Entscheidung, $h^*$, aus der Alternativenmenge, $H$, zugeordnet wird.

SCHNEEWEIß schlägt vor, ein Entscheidungsmodell als *black box* zu betrachten und nur anhand des **Nutzwertes**[314] *der für optimal befundenen Entscheidung* zu bewerten. Der Nutzwert könnte – den Ausführungen aus Abschnitt 3.2.6 folgend – eine Erfolgsgröße darstellen, wie z.B. der Gewinn oder der Deckungsbeitrag pro Planungsobjekt. Der Nutzwert wird in Abhängigkeit von dem Risiko, Fehlentscheidungen zu treffen, vermindert – z.B. wenn ein Kredit vergeben wird, obwohl der Antragsteller den Kredit nicht zurückzahlen kann.

Nach SCHNEEWEIß kann auf eine empirische Überprüfung von einzelnen Datenmustern verzichtet werden.[315] Allerdings muß sichergestellt sein, daß das Modell auf neue

---

[313] Vgl. DÜSING (1996), S. 143.

[314] Zur besseren Orientierung sind die für die weitere Untersuchung relevanten Interessantheitsfacetten fett gedruckt.

[315] Vgl. SCHNEEWEIß (1984), S. 483.

Entscheidungssituationen verallgemeinerbar ist. Im folgenden wird ein Ansatz entwickelt, der die **Allgemeingültigkeit** eines Datenmusters zwar nicht als eigenes Gütemaß berechnet, sie aber in das Entscheidungsmodell integriert und letztendlich dessen Nutzwert beeinflußt. Dieser Ansatz soll zunächst anhand dreier Beispiele eingeführt und anschließend verallgemeinert werden.

*Man betrachte eine Direktmarketingaktion, für die Kunden selektiert werden, denen eine CD-Serie angeboten wird. Per Data Mining soll ein Zusammenhang zwischen Kundenmerkmalen und der Entscheidungsvariable mit den Handlungsalternativen* $h^+$*=Kunde selektieren und* $h^-$*=Kunde nicht selektieren hergestellt werden. Die Trainingsmenge umfaßt für die in einem Pre-Test kontaktierten Kunden,* $\forall o \in O^T$*, die Handlungsergebnisse „Käufer der CD-Serie = ja“ bzw. „Käufer der CD-Serie = nein“ oder formal:*

⇨ $Käufer^+(o)=1$ *(ja) bzw.*

⇨ $Käufer^+(o)=0$ *(nein).*

*Das hochgestellte +-Symbol soll dabei andeuten, daß nur die Handlungsergebnisse für die positive Handlung,* $h^+$*, vorliegen, da alle Kunden in* $O^T$ *kontaktiert wurden. Es handelt sich hier demnach um Fall 2 aus Abbildung 3-12: Gegeben sind die Entscheidungssituation,* $es$*, und das Ergebnis,* $e$*, für die implizit unterstellte Handlung,* $h^+$*:* $(es,e^+)$*.*

*Aus einer anderen Menge von Kunden,* $O^{neu}$*, sind die zu kontaktierenden Kunden so auszuwählen, daß sie möglichst hohe Kaufwahrscheinlichkeiten aufweisen. Dazu sollen aus den Trainingsdaten Entscheidungsregeln der folgenden Form erzeugt werden:*

`WENN Kundenmerkmale ... DANN Entscheidung = Kunde selektieren bzw. nicht selektieren.`

*Um die durch eine Entscheidung realisierbaren Handlungsergebnisse in einen Nutzwert zu transformieren, müssen die Handlungsergebnisse bewertet werden. Durch die Kontaktierung eines Kunden fallen direkte Kosten in Höhe von* $dk$ *an. Falls der Kunde die CD-Serie bestellt, wird ein Erlös in Höhe von* $er$ *erzielt. Direkte Kosten und Erlöse fallen aber nur dann an, wenn ein Kunde selektiert wird. Ein Kunde,* $o \in O^T$*, wird genau dann selektiert, wenn das Modell,* $M^{Ent}$*, für ihn eine positive Selektionsentscheidung empfielt:* $M^{Ent}(o) = h^+ = Kunde\ selektieren.$

*Eine Möglichkeit, den Nutzwert des gesamten Modells,* $Modellnutzwert(M^{Ent})$*, zu bestimmen, besteht nun darin, den in der Trainingsmenge erzielten Deckungsbeitrag über alle Kunden zu berechnen, die zu selektieren lohnt:*

$$Modellnutzwert(M^{Ent}) = \sum_{\substack{\forall o \in O^T | \\ M^{Ent}(o)=h^+}} -dk + er \cdot Käufer^+(o)\ .$$

*Welche Kunden zu selektieren lohnt, besagen die Konklusionen der Entscheidungsregeln, die als Input in die Akkumulation gemäß Definition 2-21 eingehen. Zu klären bleibt noch, wie die Konklusionen der Entscheidungsregeln zustandekommen, wenn die Entscheidungen selbst nicht als Attributwerte in der Trainingsmenge vorliegen. Der Konklusion einer Entscheidungsregel,* `WENN Pr DANN Ko`*, soll genau dann eine positive Selektionsentscheidung zugewiesen werden, wenn durch eine Selektion ein positiver Deckungsbeitrag erzielt würde. Dies wird für das Segment* $O^T[Pr]$ *genau dann erwartet, falls gilt:*

$$\frac{|O^T[(Käufer^+ = 1) \wedge Pr]|}{|O^T[Pr]|} > \frac{dk}{er} \quad \Leftrightarrow \quad |O^T[(Käufer^+ = 1) \wedge Pr]| \cdot er > |O^T[Pr]| \cdot dk\ .$$

[316]

*D.h., die aus den Trainingsdaten abgeschätzte Bestellwahrscheinlichkeit muß über dem Verhältnis aus direkten Kosten und Erlösen liegen.*

---

[316] Vgl. BAUSCH (1991), S. 87 ff.

*Als weiteres Beispiel für die Anwendung von Data-Mining-Entscheidungsmodellen sei die Akquisition von Neukunden durch ein Versicherungsunternehmen betrachtet. Für potentielle Neukunden soll entschieden werden, ob ein umfangreiches ($h_1$), ein weniger umfangreiches ($h_2$) oder gar kein Akquisitionsprogramm gestartet werden soll ($h_3$).*

*Neben den erklärenden Merkmalen enthält die Trainingsmenge die für einen Kunden, $o$, realisierte Vertragslaufzeit in Jahren, $Laufzeit(o)$, und den für diesen Kunden betriebenen Akquisitionsaufwand, $h(o) \in \{h_1,h_2,h_3\}$. Es handelt sich hier demnach um Fall 5 aus Abbildung 3-12: Gegeben sind die Entscheidungssituation, $es$, die explizit vorliegende Handlung, $h$, und das Handlungsergebnis, $e$: $(es,h,e)$.*

*Eine Bewertung könnte wie im obigen Beispiel erfolgen – nur daß jetzt drei verschiedene Alternativen zu berücksichtigen sind. Für die $i$-te Alternative ($i = 1, 2, 3$) würde $kv_i$ dann die einmaligen Akquisitionskosten in DM pro Kunde darstellen, und $er_i$ würde die Erlöse (Versicherungsbeiträge) pro Jahr und Kunde bezeichnen. In diesem Beispiel gilt: $er_1=er_2=er_3$ und $kv_1>kv_2>kv_3=0$. Damit ergäbe sich folgender Nutzwert:*

$$Modellnutzwert(M^{Ent}) = \sum_{\forall o \in O^T} -dk_i + er_i \cdot Laufzeit(o) \quad \text{mit} \quad i : M^{Ent}(o) = h_i, i \in \{1,2,3\}$$

*Eine Entscheidungsregel, $\texttt{WENN } Pr \texttt{ DANN } Ko$, erhält genau dann die Konklusion $Ko := (\texttt{Entscheidung} = h_i)$, wenn mit dem Erwartungswert $\mu_i := E\left[Laufzeit \mid Pr \wedge (h = h_i)\right]$ gilt: $\mu_i \cdot er_i - dk_i = \max_{j=1,2,3} \mu_j \cdot er_j - dk_j$.*

*Man wird sich also bei einem Kunden mit den Merkmalen $Pr$ für die Akquisitionsmaßnahme $h_i$ entscheiden, für die der erwartete Zielbeitrag, $\mu_i \cdot er_i - dk_i$, maximal ist. Dabei kann für die Kundenmerkmale $Pr$ nur die beste der in der Trainingsmenge zu $Pr$ vorhandenen Alternative empfohlen werden. Der Erwartungswert kann wie folgt aus den Trainingsdaten geschätzt werden:*

$$\hat{\mu}_i := \frac{1}{|O^T[Pr \wedge (h = h_i)]|} \sum_{o \in O^T[Pr \wedge (h = h_i)]} Laufzeit(o).$$

*Als drittes Beispiel betrachte man den Handel mit Aktien gemäß Tabelle 3-2.* Mit *$h_1 = Kaufen$, $h_2 = Verkaufen$, $er_1=+1$, $er_2=-1$, $dk_1=dk_2=0$ und $\Delta K(o)$=Kursänderung für Datensatz $o$ ergeben sich:*[317]

$$Modellnutzwert(M^{Ent}) = \sum_{\forall o \in O^T} -dk_i + er_i \cdot \Delta K(o) \quad \text{mit} \quad i : M^{Ent}(o) = h_i, i \in \{1,2\}$$

*und zur Bestimmung der Entscheidungsregeln:*

$$\hat{\mu}_i := \frac{1}{|O^T[Pr \wedge (h = h_i)]|} \sum_{o \in O^T[Pr \wedge (h = h_i)]} \Delta K(o).$$

Aus den Beispielen können folgende Aspekte verallgemeinert werden:

Gegeben sei eine diskrete Alternativenmenge, $H = \{h_1, ..., h_{hmax}\}$. Bei einer Entscheidung für die Alternative $h_i$ fallen Zielbeiträge in Höhe von $-dk_i + er_i \cdot e(o)$ an. Dabei stellt $dk_i$ für die $i$-te Alternative die direkten Kosten pro Planungsobjekt dar, die unabhängig davon anfallen, welches Handlungsergebnis eintritt. $e(o)$ steht für das realisierte Handlungsergebnis des Planungsobjektes $o$. Bei dichotomen Handlungsergebnissen umfaßt der Wertebereich $dom(e)=\{0,1\}$ die Werte $0$ für das gewünschte (z.B. Kunde kauft die

[317] Das Modell berücksichtigt nicht, wie stark die Kursprognosen von den tatsächlichen Kursen abweichen. Ein entsprechendes Bewertungskonzept würde nur dann Sinn machen, wenn diese Kursdifferenz-Informationen für Allokationsentscheidungen zwischen konkurrierenden Anlagen genutzt würden (vgl. PODDIG (1999), S. 468). Dies wurde aber in Abschnitt 3.2.6 ausgeschlossen.

CD-Serie) und *1* für das nicht gewünschte Handlungsergebnis (z.B. Kunde kauft die CD-Serie nicht). Und $er_i$ quantifiziert die Erlöse, die sich bei Eintreten des gewünschten Handlungsergebnisses erzielen lassen. Bei mehrwertigen Handlungsergebnissen quantifiziert $er_i$ die Erlöse, die sich pro Einheit des Handlungsergebnisses (z.B. pro Jahr der Vertragslaufzeit) erzielen lassen.

Zu erstellen seien Entscheidungsregeln der Form:

`WENN Entscheidungssituation` *Pr* `vorliegt, DANN ist die Entscheidung` *Ko* `optimal.`

$\mu_i$ bezeichne den unbekannten Erwartungswert des Handlungsergebnisses, das sich in einer durch *Pr* charakterisierten Entscheidungssituation bei Durchführung der Handlung $h_i$ erzielen läßt:

$$\mu_i := E\left[e \mid Pr \wedge (h = h_i)\right].$$

Da das erwartete Handlungsergebnis, $\mu_i$, unbekannt ist, muß es aus der Trainingsmenge geschätzt werden. Ein erwartungstreuer Schätzer, $\hat{\mu}_i$, ist der Stichprobenmittelwert der Handlungsergebnisse, $\bar{e}_i$:

$$\hat{\mu}_i := \bar{e}_i = \frac{1}{|O^T[Pr \wedge (h = h_i)]|} \sum_{\forall o \in O^T [Pr \wedge (h = h_i)]} e(o).$$

Eine Entscheidungsregel, *(Pr→Ko)*, erhält die Konklusion

*Ko* `:= (Entscheidung =` $h_i$`)`

mit dem *i*, für das gilt:

$$\bar{e}_i \cdot er_i - dk_i = max\left\{\bar{e}_j \cdot er_j - dk_j \quad \middle| \quad j = 1,...,hmax\right\}.$$

Damit kann der Nutzwert eines Data-Mining-Entscheidungsmodells wie folgt definiert werden:

$$Modellnutzwert(M^{Ent}) = \sum_{\forall o \in O^T} -dk_i + er_i \cdot e(o) \quad \text{mit} \quad i : M^{Ent}(o) = h_i;\ i = 0,...,hmax;$$

$h_0$ sei dabei eine Dummy-Alternative mit $er_0 = dk_0 = 0$, die ausgegeben wird, falls das Modell für bestimmte Planungsobjekte keine Entscheidung aussprechen kann.

Die beschriebene Vorgehensweise funktioniert auch dann, wenn die Trainingsmenge Zielbeiträge, *z(o)*, statt der Handlungsergebnisse, *e(o)*, enthält (Fall 6), denn es gilt: $z(o) = -dk_i + er_i \cdot e(o)$. Die Parameter $dk_i$ und $er_i$ werden dann nicht benötigt.

Komplexe Interdependenzen, wie sie für Programmentscheidungen typisch sind, werden hier nicht berücksichtigt, so daß sich das hier eingeführte Bewertungskonzept nur zur Unterstützung von Alternativ- und Einzelentscheidungen eignet.

Damit steht ein erster Bewertungsansatz für das Data Mining zur Verfügung. Dieser Ansatz basiert auf dem ***Erwartungswertmodell***[318] der Entscheidungstheorie. Er ist dann gerechtfertigt, wenn die Entscheidungssituationen mit den durch $Pr$ beschriebenen Eigenschaften häufig auftreten und die zugrundeliegenden Objektmengen $O^T[Pr \wedge (h=h_i)]$ „groß genug" sind, so daß sich für die $i$-te Handlung im Mittel näherungsweise das erwartete Handlungsergebnis, $\mu_i$, einstellt. Bei kleineren Objektmengen ist das Risiko, daß sich das erwartete Ergebnis nicht einstellt, explizit zu berücksichtigen. Die Entscheidungstheorie schlägt hier u.a. das ***Erwartungswert-Varianz-Modell***[319] vor. Nach diesem Modell würde eine Entscheidungsregel, $(Pr \rightarrow Ko)$, die Konklusion

$Ko$ := (Entscheidung = $h_i$)

mit dem $i$ erhalten, für das gilt:

$$\mu_i^z - \pi \cdot \sigma_i^2 = max\left\{\mu_j^z - \pi \cdot \sigma_j^2 \quad \middle| \quad j = 1,...,hmax\right\}$$

mit $\pi \geq 0$.

$\pi$ kann dabei als „***Risikoaversität***" bezeichnet werden. $\mu_i^z$ ist der Erwartungswert der Zielbeiträge, die sich bei der Wahl der Alternative $h_i$ einstellen. Er kann durch den Mittelwert der Stichprobe, $\bar{z}_i$, abgeschätzt werden. Und $\sigma_i^2$ ist die Varianz der Zielbeiträge in der Grundgesamtheit, die sich bei der Wahl der Alternative $h_i$ einstellen. Auch sie ist unbekannt, kann aber wie folgt aus der Stichprobe $O^T[Pr \wedge (h=h_i)]$ geschätzt werden:[320]

$$\hat{\sigma}_i^2 = \frac{1}{|O^T[Pr \wedge (h=h_i)]| - 1} \sum_{\forall o \in O^T[Pr \wedge (h=h_i)]} (z(o) - \bar{z}_i)^2 \ .$$

Die Varianz fällt umso geringer aus, je mehr Datenobjekte der Trainingsmenge von der Entscheidungsregel erfaßt werden, d.h. je größer $|O^T[Pr \wedge (h=h_i)]|$ wird. Eine hohe Anwendbarkeit einer Regel trägt also zur Reduktion des Entscheidungsrisikos bei.

---

[318] Vgl. zu dem Erwartungswertmodell DINKELBACH (1982), S. 78 ff.

[319] Vgl. zu dem Erwartungswert-Varianz-Modell DINKELBACH (1982), S. 84 ff.

[320] Vgl. BAMBERG/BAUR (1998), S. 165.

Das Erwartungswert-Varianz-Modell hat jedoch einen Nachteil, der seine Einsatzmöglichkeiten im Data Mining einschränkt: Die Risikoaversität $\pi$ ist sehr schwer sinnvoll einzustellen. Ähnliches gilt für den Parameter $t_0$ des entscheidungstheoretischen ***Aspirationsmodells***[321]. Nach diesem Modell wird diejenige Alternative für optimal befunden, die die Wahrscheinlichkeit für das Erreichen oder Übertreffen eines gewünschten Zielfunktionswertes $t_0$ maximiert. Im Data Mining kann die Erreichbarkeit bestimmter Zielfunktionswerte erst nach vielen Durchläufen abgeschätzt werden.

Daher erscheint im Data Mining das ***Fraktilmodell***[322] praktikabler. Danach gibt man eine Wahrscheinlichkeit $1\text{-}\alpha$ vor und befindet die Alternative für optimal, die eine Untergrenze für den Zielfunktionswert maximiert, der mit dieser Wahrscheinlichkeit erreichbar ist. Sei $ug_i^{Pr}(1-\alpha)$ eine Untergrenze für die erwarteten Zielbeiträge einer Regel mit der Prämisse $Pr$, dann ermittelt man die optimale Alternative, $h_i$, wie folgt:

$$ug_i^{Pr}(1-\alpha) = max\left\{ug_j^{Pr}(1-\alpha) \quad \middle| \quad j = 1,...,hmax\right\}.$$

Die Untergrenze $ug_i^{Pr}(1-\alpha)$ läßt sich durch Rückgriff auf die Stichprobentheorie statistisch fundieren. Das Stichprobenmittel $\bar{z}_i$ stellt die Realisierung einer Zufallsvariable $\overline{Z}_i$ dar. Für einen Stichprobenumfang von über $30$ gilt für beliebig verteilte Grundgesamtheiten näherungsweise:[323]

$$P\left(\mu_i^z < \overline{Z}_i - \frac{c(1-\alpha)}{\sqrt{\left|O^T[Pr \wedge (h = h_i)]\right|}} \cdot \sigma\right) = \alpha\,.$$

Der Wert $c(1-\alpha)$ wird aus der Tafel der Standardnormalverteilung bzw. der Student-$t$-Verteilung abgelesen.[324] Wie im Fraktilmodell beabsichtigt, kann eine Untergrenze für den Erwartungswert der Zielbeiträge, $\mu_i^z$, angegeben werden, der mit der Wahrscheinlichkeit $1-\alpha$ nicht unterschritten wird:

$$ug_i^{Pr}(1-\alpha) = \bar{z}_i - \frac{c(1-\alpha)}{\sqrt{\left|O^T[Pr \wedge (h = h_i)]\right|}} \cdot \sigma\,.$$

---

[321] Vgl. zu dem Aspirationsmodell DINKELBACH (1982), S. 91 ff.

[322] Vgl. zum Fraktilmodell DINKELBACH (1982), S. 88 ff.

[323] Vgl. BAMBERG/BAUR (1998), S. 162, S. 166.

[324] Die Standardnormalverteilung entspricht für Stichprobenumfänge größer $30$ näherungsweise der Student-$t$-Verteilung.

Man erkennt die Ähnlichkeit zum Erwartungswert-Varianz-Modell – mit dem Unterschied, daß anstelle der Varianz die Standardabweichung, $\sigma$, verwendet wird und daß der Koeffizient $c(1-\alpha)/\sqrt{|O^T[Pr \wedge (h = h_i)]|}$ nicht willkürlich gewählt und konstant ist, sondern statistisch fundiert wurde und von dem Stichprobenumfang abhängt.

Nach diesen Überlegungen berechnet sich die Lösungsgüte wie folgt:

**Definition 3-3: Nutzwert eines Data-Mining-Entscheidungsmodells**

$O^T$ sei eine Trainingsmenge, $M^{Ent}$ ein Data-Mining-Entscheidungsmodell, $Z$ die Menge der Zielbeiträge, und $H = \{h_0, h_1, ..., h_{hmax}\}$ eine diskrete Alternativenmenge mit $h_0$ als Dummy-Alternative. Das Entscheidungsmodell $M^{Ent}$ sei auf der Trainingsmenge, $O^T$, erlernt worden:

$M^{Ent}(o) = Akkumulation(M_{O^T}, o)$.

Dann ergibt sich der Nutzwert des Entscheidungsmodells als Summe der Nutzwert-Untergrenzen über alle Planungsobjekte:

$$Modellnutzwert_\alpha(M^{Ent}) = \sum_{\forall o \in O^T} ug_i^{Pr}(1-\alpha);$$

$o \in O^T[Pr];$

$i : M^{Ent}(o) = h_i; i = 0,...,hmax.$

◊

### *3.3.2.2 Konzeption zur Generierung von Prognosemodellen*

Überträgt man die in Abschnitt 3.1 vorgestellte Beschreibung eines allgemeinen Prognosemodells auf die im Data Mining generierten Datenmuster, so kann man ein Data-Mining-Prognosemodell wie folgt definieren:

**Definition 3-4: Data-Mining-Prognosemodell**

Gegeben sei mit $dom(D)$ der Wertebereich einer zu prognostizierenden Größe $D$, $D \in A$, mit $O$ die Grundgesamtheit aller Planungsobjekte und mit $O^T \subseteq O$ eine Trainingsmenge. Dann ist ein *Data-Mining-Prognosemodell*, $M^{Pro}$, ein funktionales Data-Mining-Modell gemäß Definition 2-22, das geeignet ist, in einer gegebenen Prognosesituation, $o \in O$, eine möglichst präzise und korrekte Vorhersage zukünftiger Outputs, $M^{Pro}(o) \in dom(D)$, zu treffen:

$$M^{Pro}: \; O \rightarrow dom(D);$$
$$o \rightarrow M^{Pro}(o). \qquad \Diamond$$

Die *Aufgabe eines Data-Mining-Prognosemodells* besteht in seiner Anwendung auf neue Inputdaten, $O^{neu} \subseteq O$, für die es Prognosen liefert. Dabei gilt: $O^{neu} \cap O^T = \varnothing$.

Wie die Anwendungen in Abschnitt 3.2 gezeigt haben, sind die für das praktische Data Mining relevantesten Fälle die, daß ein Prognosemodell entweder Umweltsituationen, Handlungsergebnisse oder Zielbeiträge vorhersagt.

Im Falle der *Prognose von Handlungsergebnissen oder Zielbeiträgen* bildet das Modell eine Wirkungs- oder Zielerreichungsfunktion gemäß Definition 3-1 ab. Diese wird durch eine Menge von Datenmustern, $M_{O^T} = \{(Pr_1 \rightarrow Ko_1), ..., (Pr_M \rightarrow Ko_M)\}$, der Form

WENN Entscheidungssituation $Pr_i$ gegeben ist,
DANN prognostiziere Handlungsergebnis bzw. Zielbeitrag $Ko_i$

sowie durch eine Akkumulationsfunktion gemäß Definition 2-21 modelliert. Jedes Planungsobjekt in einer durch $Pr_i$ beschriebene Entscheidungssituation erhält als Prognose das durch $Ko_i$ beschriebene Handlungsergebnis.

Im Falle der *Prognose von Umweltsituationen* haben die Datenmuster die folgende Form:

WENN Entscheidungssituation $Pr_i$ gegeben ist,
DANN prognostiziere Umweltsituation $Ko_i$.

Jedes Planungsobjekt in einer durch $Pr_i$ beschriebenen Entscheidungssituation erhält als Prognose die durch $Ko_i$ beschriebene Umweltsituation. Auch hier wird die endgültige Prognose durch die Akkumulation von Einzelprognosen ermittelt.

Für Prognosemodelle gilt, wie auch schon für Entscheidungsmodelle, daß die generierte Datenmuster-Menge, $M_{O^T}$, selbst nicht von Interesse ist, sondern nur – im Sinne einer *black box* – die Güte der Prognosen, $M^{Pro}(o)$, für neue Inputdaten, $o \in O^{neu}$. Diese ist allerdings unbekannt, so daß an ihrer Stelle Treffsicherheitsmaße, wie z.B. die **Modellkorrektheit** aus Definition 2-54, verwendet werden. Treffsicherheitsmaße messen den Erfolg von Prognosemodellen in der Vergangenheit, können aber aufgrund des in

Abschnitt 2.1.1 diskutierten Induktionsproblems nie den Erfolg von Prognosemodellen in zukünftigen Anwendungen vorhersagen.[325]

In praktischen Anwendungen werden Prognosemodelle fast immer allein anhand ihrer Modellkorrektheit bewertet. Die Reliabilität, also die Korrektheit bezüglich neuer Daten, wird dabei i.d.R. erst nach der Erstellung des Modells *geprüft*. Dabei können nach den Ausführungen aus Abschnitt 2.2.4.2 durch Gewährleistung einer gewissen **Allgemeingültigkeit**[326] der an der Prognose beteiligten Datenmuster durchaus Modelle *erzeugt* werden, die mit hoher Wahrscheinlichkeit reliabel sind (anstelle die Reliabilität nur im Nachhinein zu prüfen).

Schließlich sind die ausgegebenen Prognosen anhand ihrer **Präzision**[327] zu beurteilen. Je präziser die Aussage für eine gegebene Entscheidungssituation ausfällt, umso größer ist die Planungssicherheit für den Entscheidungsträger. Entscheidungen aufgrund von unpräzisen Prognosen werden hier in Anlehnung an die Entscheidungstheorie als „***Entscheidungen unter Unsicherheit***“[328] bezeichnet. Entscheidungen unter Unsicherheit zeichnen sich durch bekannte Wertebereiche und unbekannte Wahrscheinlichkeiten der zugrundeliegenden Zufallsvariablen aus.

*Beispielsweise ist die Prognose „$Aktienkurs \in [50;90]$“ nicht sehr präzise, wenn der aktuelle Kurs $70$ Euro betrifft. Selbst wenn die Wahrscheinlichkeit $P(Aktienkurs \in [50;90]) = 0{,}99$ mit einer geringen Irrtumswahrscheinlichkeit vorhergesagt werden kann, so ist die Wahrscheinlichkeitsverteilung innerhalb des Intervalls $[50;90]$ unbekannt. Deren Kenntnis wäre aber relevant für eine Kaufen-/Verkaufen-Entscheidung.*

#### *3.3.2.3 Konzeption zur Generierung von Erklärungsmodellen*

Überträgt man den Begriff des Erklärungsmodells aus Abschnitt 3.1 auf die im Data Mining generierten Datenmuster, so kann man ein Data-Mining-Erklärungsmodell wie folgt definieren:

---

[325] Vgl. TIETZEL (1985), S. 104.

[326] Die Allgemeingültigkeit wurde in Definition 2-55 operational definiert.

[327] Die Präzision wurde in Definition 2-62 operational definiert.

[328] DINKELBACH (1982), S. 40

**Definition 3-5: Data-Mining-Erklärungsmodell**

Gegeben sei mit $A$ eine Menge von Attributen, mit $D \subset A$ eine Menge von Output- oder Wirkungsvariablen und mit $O^T$ eine Trainingsmenge. Ein *Data-Mining-Erklärungsmodell*, $M^{Erk} = M_{O^T}$, ist eine Menge von Datenmustern, welche die fundamentalen Ursache-Wirkungszusammenhänge in folgender Form abbildet:

$M^{Erk} = \{ s_{O_1}, \ldots, s_{O_M} \}$ mit

$s_{O_i} = (Pr_i \rightarrow Ko_i);$

$O_i = O^T[Pr_i];$

$Ko_i = (D \in D_i);$

$D_i \subset dom(D);$

$i = 1, \ldots, M.$

Ein Datenmuster, $Pr_i \rightarrow Ko_i$, kann wie folgt gelesen werden:

$Pr_i$ `erklärt` $Ko_i$. ◊

Dabei charakterisiert der Term $Pr_i$ die Ursache, die die durch $Ko_i$ charakterisierte Wirkung hervorruft. Die Konklusionen umfassen Teilmengen, $D_i$, aus dem Wertebereich der Outputvariablen, $dom(D)$.

Erklärungsmodelle sind nach SCHNEEWEIß anhand der *Verifikationsgüte* ihrer einzelnen Datenmuster zu beurteilen.[329] Die Verifikationsgüte wird hier in die Komponenten „**Allgemeingültigkeit**“ und „**Stärke des Ursache-Wirkungszusammenhangs**“ zerlegt. Als Bewertungsansatz für eine Regel kämen die Interessantheitsmaße aus Definition 2-55 und Definition 2-65 in Betracht.

Weiterhin sind vor allem diejenigen Aussagen interessant, die solche Phänomene erklären, welche für den betriebswirtschaftlichen Erfolg des Verantwortungsträgers *besonders relevant* sind – z.B. die Einkäufe einer besonders deckungsbeitragsstarken Warengruppe $X$. Wenn ein solches Phänomen dann auch noch *vollständig erklärt* werden kann – also z.B. alle Einkäufe der Warengruppe $X$ – dann liefert das Modell einen besonders hohen Erfolgsbeitrag. Dieser Aspekt der Interessantheit soll als „***Erfolgsrelevanz***“ bezeichnet und wie folgt definiert werden:

[329] Vgl. SCHNEEWEIß (1984), S. 483.

**Definition 3-6: Erfolgsrelevanz eines Modells**[330]

$L$ sei die Menge aller möglichen Modelle gemäß Definition 2-15 und $\boldsymbol{R}$ die Menge der reellen Zahlen. $ER(o) \in \{r | r \in \boldsymbol{R}, 0 \leq r \leq 1\}$ quantifiziere für jedes Planungsobjekt, $o \in O^T$, dessen relative Erfolgsrelevanz. Dann ist die *Erfolgsrelevanz eines Modells* aus $L$ als Summe der Erfolgsrelevanzen derjenigen Objekte, die durch mindestens eine Regel des Modells abgedeckt werden, definiert:

$$Erfolgsrelevanz^M: \quad L \rightarrow \{r \mid r \in \boldsymbol{R},\ 0 \leq r \leq 1\};$$

$$M_{O^T} \rightarrow Erfolgsrelevanz^M(M_{O^T});$$

$$\text{mit} \quad Erfolgsrelevanz^M(M_{O^T}) := \sum_{\substack{o \in O^T [Pr \wedge Ko], \\ (Pr \rightarrow Ko) \in M_{O^T}}} ER(o);$$

$$\sum_{o \in O^T} ER(o) = 1. \qquad \Diamond$$

Demnach ist die Erfolgsrelevanz eines Modells umso höher, je mehr Objekte durch mindestens eine Regel des Modells erfaßt werden und je höher die Erfolgsbeiträge der erfaßten Objekte sind. Die Summe der objektbezogenen Erfolgsbeiträge soll $1$ ergeben, damit die Erfolgsrelevanz des Modells auf das Intervall von $[0;1]$ normiert ist.

Neben diesem *mengen- oder wertmäßigen Anteil des Phänomens*, der erklärt werden kann, bleibt zu untersuchen, *wie gut das Phänomen erklärt werden kann*. Dieser Aspekt der Interessantheit soll in Anlehnung an das auf eine einzelne Regel bezogene Zusammenhangsmaß aus Definition 2-65 (S. 95), $Zusammenhang^g$, wie folgt definiert werden:

**Definition 3-7: Stärke der Zusammenhänge eines Erklärungsmodells**

Es gelten dieselben Voraussetzungen wie in der Definition zuvor. Dann ist die *Stärke der Zusammenhänge eines Erklärungsmodells* aus $L$ wie folgt definiert:

$$Zusammenhang^{M,g}: \quad L \rightarrow \{r \mid r \in \boldsymbol{R},\ 0 \leq r \leq 1\};$$

$$M^{Erk} \rightarrow Zusammenhang^{M,g}(M^{Erk});$$

[330] Diese Definition wird für Erklärungs- und Beschreibungsmodelle verwendet, so daß hier $M_{O^T}$ anstelle von $M^{Erk}$ bzw. $M^{Bes}$ geschrieben wird.

mit $Zusammenhang^{M,g}(M^{Erk}) := \frac{1}{M}\sum_{j=1}^{M} Zusammenhang^{g}(Pr_j \rightarrow Ko_j)$. ◊

Beide Aspekte werden unter der Interessantheitsfacette „***Erklärungsgüte***“ zusammengefaßt:

**Definition 3-8: Erklärungsgüte eines Erklärungsmodells**

Es gelten dieselben Voraussetzungen wie in den beiden Definitionen zuvor. Dann ist die *Erklärungsgüte* eines Erklärungsmodells, $M^{Erk} = \{(Pr_1 \rightarrow Ko_1), ..., (Pr_M \rightarrow Ko_M)\}$, wie folgt definiert:

*Erklärungsgüte*: $L \rightarrow \{r \mid r \in \mathbf{R}, 0 \leq r \leq 1\}$;

$M^{Erk} \rightarrow Erklärungsgüte(M^{Erk})$;

mit $Erklärungsgüte(M^{Erk}) :=$
$Erfolgsrelevanz^{M}(M^{Erk}) \cdot w^{E} + Zusammenhang^{M}(M^{Erk}) \cdot (1-w^{E})$;
$w^{E} \in \{r \mid r \in \mathbf{R}, 0 \leq r \leq 1\}$. ◊

Neben der oben auf das ganze Modell bezogenen Erfolgsrelevanz macht es auch Sinn, die Erfolgsrelevanz für eine einzelne Regel zu definieren, um irrelevante Regeln von der Analyse auszuschließen:

**Definition 3-9: Erfolgsrelevanz einer Regel**

$A$ sei die Menge der beobachteten Attribute, $D$ sei eine Menge von zu erklärenden, $C$ eine Menge von erklärenden Attributen mit $D,C \subset A$, $C \cap D = \emptyset$, $C,D \neq \emptyset$. $DM^{KNF}(C,D)$ sei die Menge der möglichen Datenmuster in KNF bzgl. $C$ und $D$ und $\mathbf{R}$ die Menge der reellen Zahlen. $ER(o) \in \{r \mid r \in [0;1] \cap \mathbf{R}\}$ quantifiziere für jedes Planungsobjekt, $o \in O^T$, dessen relative Erfolgsrelevanz. Dann ist die *Erfolgsrelevanz einer Regel* aus $L$ als Summe der Erfolgsrelevanzen derjenigen Objekte, die durch die Regel abgedeckt werden, definiert:

*Erfolgsrelevanz*: $DM^{KNF}(C,D) \rightarrow \{r \mid r \in \mathbf{R}, 0 \leq r \leq 1\}$;

$Pr \rightarrow Ko \rightarrow Erfolgsrelevanz(Pr \rightarrow Ko)$;

mit $Erfolgsrelevanz(M^{Erk}) := \sum_{o \in O^T[Pr \wedge Ko]} ER(o)$;

$\sum_{o \in O^T} ER(o) = 1$. ◊

*Man betrachte das in Tabelle 3-3 dargestellte Beispiel.*

| Nr | Artikelgruppe | Kundengruppe | Region | Umsatzgruppe | Umsatz | ER(o) |
|---|---|---|---|---|---|---|
| 1 | Tourenräder | Fachhandel | Süd | niedrig | 5.000 | 0,01 |
| 5 | Tourenräder | Fachhandel | Süd | hoch | 100.000 | 0,20 |
| 2 | Cityräder | Discounter | West | hoch | 150.000 | 0,30 |
| 8 | Cityräder | Discounter | West | hoch | 110.000 | 0,22 |
| 3 | Mountainbikes | Fachhandel | Nord | mittel | 50.000 | 0,10 |
| 4 | Tourenräder | Fachhandel | Nord | niedrig | 5.000 | 0,01 |
| 6 | Mountainbikes | Kaufhaus | Nord | mittel | 40.000 | 0,08 |
| 7 | Mountainbikes | Discounter | West | mittel | 40.000 | 0,08 |
| | | | | **Summe:** | 500.000 | 1,00 |

***Tabelle 3-3: Beispiel zur Berechnung der Erklärungsgüte***

*Beispielsweise ermittelt man für die Regel:*
*(Pr→Ko) =* `WENN Artikelgruppe = Cityräder UND Kundengruppe = Discounter UND Region = West DANN Umsatzgruppe = hoch`
*den Zusammenhang und den regelbezogenen Erfolgsbeitrag wie folgt:*

$|O^T[Ko]| = 3.$

$|O^T[Pr \wedge Ko]| = 2.$

$|O^T| = 8.$

$|O^T[Pr]| = 2.$

$Zusammenhang^g(Pr \rightarrow Ko) = 1 - (3/8)/(2/2) = 5/8 = 0{,}625.$[331]

$ER(\text{Objekt Nr. 2}) + ER(\text{Objekt Nr. 8}) = 0{,}30 + 0{,}22 = 0{,}52.$

*Nach diesem Rechenschema ergeben sich bei Vernachlässigung der ersten Regel (mit der widersprüchlichen Konklusion) folgende Regeln (mit den entsprechenden Werten für den Zusammenhang und die regelbezogenen Erfolgsbeitrag in Klammern):*

`(Pr→Ko) = WENN Artikelgruppe = Cityräder UND Kundengruppe = Discounter UND Region = West DANN Umsatzgruppe = hoch` *(0,625; 0,52).*

`WENN Artikelgruppe = Mountainbikes UND Kundengruppe = Fachhandel UND Region = Nord DANN Umsatzgruppe = mittel` *(0,625; 0,1).*

`WENN Artikelgruppe = Tourenräder UND Kundengruppe = Fachhandel UND Region = Nord DANN Umsatzgruppe = niedrig` *(0,75; 0,01).*

`WENN Artikelgruppe = Mountainbikes UND Kundengruppe = Kaufhaus UND Region = Nord DANN Umsatzgruppe = mittel` *(0,625; 0,08).*

`WENN Artikelgruppe = Mountainbikes UND Kundengruppe = Discounter UND Region = West DANN Umsatzgruppe = mittel` *(0,625; 0,08).*

*Damit ergibt sich für die Regelmenge ein Zusammenhang von 1/5·(0,625 + 0,625 + 0,75 + 0,625 + 0,625) = 0,65, eine Erfolgsrelevanz von 0,52 + 0,1 + 0,01 + 0,08 + 0,08 = 0,79 und mit $w^E=0{,}5$ eine Erklärungsgüte von 0,65·0,5 + 0,79·0,5 = 0,72.*

[331] Vgl. zur Formel für den Zusammenhang: Definition 2-65 (S. 95).

Darüber hinaus muß ein Erklärungsmodell für den Benutzer verständlich und interpretierbar sein, damit er daraus einen möglichen Handlungsbedarf ableiten und einen Planungsprozeß in Gang setzen kann. Die **Verständlichkeit** eines Modells kann – wie an Definition 2-57 bzw. Definition 2-58 gesehen – auf syntaktischer Ebene problemlos definiert werden.[332] Der Einbezug der Semantik scheitert jedoch an dem unzureichenden Kontextwissen, das einem Data-Mining-Verfahren unter vertretbarem Aufwand zur Verfügung gestellt werden kann.

Der angestrebte Erkenntniszuwachs tritt nicht ein, wenn die Regeln bereits bekannte Zusammenhänge abbilden. Daher ist für Erklärungsmodelle *Neuheit* zu fordern. Aufgrund der in Abschnitt 2.2.4.5 geführten Diskussion werden hier besonders die **Redundanzfreiheit** der Datenmuster innerhalb eines Erklärungsmodells und die **Unbekanntheit** der Datenmuster für den Benutzer als relevante Facetten der Neuheit erachtet. Während erstere – wie Abschnitt 5.2.1 zeigt – durch die Definition des Modelltyps gewährleistet werden kann, ist nun noch die Unbekanntheit einer Aussage für den Benutzer formal zu definieren. Um die Unbekanntheit berechnen zu können, muß das Verfahren „wissen“, was dem Benutzer bereits bekannt ist. Die Auswertung dieser Facette erfordert damit ein Zugriff auf Vorwissen aus dem Benutzermodell. Daher stellt dieses Interessantheitsmaß – im Gegensatz zu den übrigen Facetten der Interessantheit – eine subjektive Größe dar. Es ist von dem Kenntnisstand des Benutzers und nicht von den Daten oder der Syntax des Modells abhängig. Dieses Maß läßt sich künstlich „objektivieren“, indem man sich bei der Erfassung des Vorwissens auf einen allgemein als bekannt angenommenen Wissensstand beschränkt. Dieser allgemein als bekannt angenommene Wissensstand umfaßt vor allem triviale Aussagen, wie z.B.:

*WENN Beruf = Student UND Alter ∈ [18;28] DANN Einkommen = niedrig.*

Dabei soll die Unbekanntheit nicht – wie dies beispielsweise GEBHARDT[333] praktiziert – in Abhängigkeit von der Anzahl der durch zwei Regeln gemeinsam erfaßten Objekte definiert werden, da sonst für alle Regeln des Benutzermodells die in der aktuellen Datenbasis erfaßten Objekte ermittelt und gespeichert werden müßten. Dies würde zum

---

[332] Der Begriff wurde hier von „Einfachheit“ in „Verständlichkeit“ geändert, da sich „Verständlichkeit“ nur auf einen der in Abschnitt 2.2.4.3 diskutierten Aspekte der Einfachheit bezieht (auf die erste Version von „Occam’s razor“).

[333] GEBHARDT bezeichnet sein Redundanzmaß als „Affinität“ zwischen zwei Objektmengen. Vgl. GEBHARDT (1994), S. 13.

einen die Laufzeit erhöhen. Zum anderen zielt der Begriff der Unbekanntheit eher auf die logische Aussage als auf die abgedeckten Objekte – schließlich kann eine Aussage auch dann unbekannt sein, wenn sie ähnliche Objektmengen abdeckt wie bereits bekannte Aussagen.

Im folgenden wird ein Unbekanntheitsmaß entwickelt, das sich an dem Konzept des Informationsgehaltes aus Abschnitt 2.2.4.4 orientiert. Danach ist eine Regel umso unbekannter, je allgemeingültiger ihre Prämisse und je präziser ihre Konklusion im Vergleich zu einer bereits bekannten Regel aus dem Benutzermodell ist. Damit ergibt sich folgende Definition:

**Definition 3-10: Unbekanntheit einer Regel im Vergleich zu einer gegebenen Regel**

Es gelten dieselben Voraussetzungen wie in der Definition zuvor. Die *Unbekanntheit einer neuen Regel, $(Pr \rightarrow Ko)$, im Vergleich zu einer gegebenen Regel, $(Pr^{BM} \rightarrow Ko^{BM})$,* setzt sich wie folgt aus dem Allgemeinheits- und Präzisionsgewinn der neuen Regel zusammen:

$$\textit{Unbekanntheit}: \quad DM^{KNF}(C,D) \times DM^{KNF}(C,D) \rightarrow \{r \mid r \in \mathbf{R},\ 0 \le r \le 1\};$$

$$((Pr \rightarrow Ko), (Pr^{BM} \rightarrow Ko^{BM})) \rightarrow Unbekanntheit((Pr \rightarrow Ko), (Pr^{BM} \rightarrow Ko^{BM}));$$

mit $Unbekanntheit((Pr \rightarrow Ko), (Pr^{BM} \rightarrow Ko^{BM}))$

$$:= \begin{cases} 1 & \text{falls } a^{Ko} \neq a^{BM,Ko}; \\ \frac{1}{2}(ag + pg) & \text{sonst;} \end{cases}$$

$ag := Allgemeinheitsgewinn(Pr, Pr^{BM});$

$pg := Präzisionsgewinn(Ko, Ko^{BM});$

wobei $Ko = (a^{Ko} \in WM)$ und $Ko^{BM} = (a^{BM,Ko} \in WM^{BM})$ Konklusionsklauseln in KNF darstellen. Falls es sich bei dem Attribut um ein kardinales oder ordinales Feld handelt, stellt die Wertemenge, wie gehabt, ein Intervall dar:

$WM = [ug;og)$ bzw. $WM^{BM} = [ug^{BM};og^{BM})$.

Die Klauseln der Prämissen werden wie folgt geordnet:

$Pr = (Pr_1 \wedge \ldots \wedge Pr_{Prmax});$
$Pr^{BM} = (Pr_1^{BM} \wedge \ldots \wedge Pr_{Prmax}^{BM});$
$Pr_i = (a_i \in WM_i);$
$Pr_i^{BM} = (a_i \in WM_i^{BM});$
$i = 1,\ldots, Prmax.$

D.h., das Attribut der $i$-ten Klausel, $a_i$, ist in beiden Regeln identisch. Falls ein Attribut in der neuen Regel oder in der Regel des Benutzermodells nicht auftaucht, wird eine künstliche Klausel eingeführt mit:

$WM_i^{BM} = dom(a_i)$ bzw. $WM_i = dom(a_i)$.

Sei $\#Kl(Pr)$ die Anzahl der Klauseln in der Prämisse $Pr$ (ohne die eventuell künstlich eingeführten Klauseln). Dann setzt sich der Allgemeinheitsgewinn wie folgt aus dem Allgemeinheitsgewinn der einzelnen Klauseln zusammen:

$$\begin{aligned}
&\textit{Allgemeinheitsgewinn:} \quad Te^{KNF}(A) \times Te^{KNF}(A) \rightarrow \{r \mid r \in \mathbf{R},\ 0 \le r \le 1\};\\
&\qquad (Pr, Pr^{BM}) \rightarrow Allgemeinheitsgewinn(Pr, Pr^{BM});\\
&\qquad \text{mit } Allgemeinheitsgewinn(Pr, Pr^{BM}) :=\\
&\qquad \frac{1}{\#Kl(Pr)} \sum_{i=1}^{Prmax} Allgemeinheitsgewinn^{Klausel}(Pr_i, Pr_i^{BM});\\
&\qquad \#Kl(Pr) := \sum_{i=1}^{Prmax} \begin{cases} 1 & \text{falls } Allgemeinheitsgewinn^{Klausel}(Pr_i, Pr_i^{BM}) > 0; \\ 0 & \text{sonst.} \end{cases}
\end{aligned}$$

Der Allgemeinheitsgewinn einer neuen Klausel gegenüber der entsprechenden Klausel aus dem Benutzermodell wird schließlich wie folgt ermittelt:

$$\begin{aligned}
&Allgemeinheitsgewinn^{Klausel}: \quad Kl^{KNF}(A) \times Kl^{KNF}(A) \rightarrow \{r \mid r \in \mathbf{R},\ 0 \le r \le 1\};\\
&\qquad (Kl, Kl^{BM}) \rightarrow Allgemeinheitsgewinn^{Klausel}(Kl, Kl^{BM});
\end{aligned}$$

mit $Allgemeinheitsgewinn^{Klausel}(Kl, Kl^{BM}) :=$

$$\begin{cases}
max\left\{0, \dfrac{|WM - WM^{BM}|}{|WM|}\right\} & \text{falls } a \text{ nominal;}\\
\dfrac{max\{0, ug^{BM} - ug\} + max\{0, og - og^{BM}\}}{max\{og - ug, og^{BM} - ug^{BM}\}} & \text{falls } a \text{ kardinal;}\\
\dfrac{max\{0, R(ug^{BM}) - R(ug)\} + max\{0, R(og) - R(og^{BM})\}}{max\{R(og) - R(ug), R(og^{BM}) - R(ug^{BM})\}} & \text{falls } a \text{ ordinal.}
\end{cases}$$

Der Präzisionsgewinn wird analog zum Allgemeinheitsgewinn mit umgekehrten Vorzeichen ermittelt.

*Es liege beispielsweise die folgende Regel im Benutzermodell vor:*

```
WENN Alter ∈ [18;25) UND studiert = ja
DANN Einkommen  ∈ [0;2000).
```

*Die Redundanz zu einer neuen Regel,*

*WENN Alter ∈ [20;28) UND Geschlecht = männlich*
*DANN Einkommen ∈ [0;3000),*

*wird dann durch Vergleich der folgenden Klauseln ermittelt:*

*Alter ∈ [20;28)* **und** *Alter ∈ [18;25);*
*studiert ∈ {ja, nein}* **und** *studiert = ja;*
*Geschlecht = männlich* **und** *Geschlecht ∈ {weiblich, männlich};*
*Einkommen ∈ [0;3000)* **und** *Einkommen ∈ [0;2000).*

*Der Vergleich der Klauseln ergibt folgende Allgemeinheits- und Präzisionsgewinne:*

$Allgemeinheitsgewinn^{Klausel}((Alter \in [20;28)), (Alter \in [18;25))) = (28–25)/(28–20) = 3/8;$
$Allgemeinheitsgewinn^{Klausel}((studiert \in \{ja, nein\}), (studiert \in \{ja\})) = 1/2;$
$Allgemeinheitsgewinn^{Klausel}((Geschlecht \in \{männlich\}), (Geschlecht \in \{weiblich, männlich\})) = 0;$
$Allgemeinheitsgewinn((Alter \in [20;28)) \wedge (studiert \in \{ja, nein\}) \wedge (Geschlecht \in \{männlich\}),$
$(Alter \in [18;25)) \wedge (studiert \in \{ja\}) \wedge (Geschlecht \in \{weiblich, männlich\})) = 1/2 \cdot (3/8+1/2+0) = 7/16;$
$Präzisionsgewinn((Einkommen \in [0;3000)), (Einkommen \in [0;2000))) = 0.$

*Damit ergibt sich schließlich folgende Unbekanntheit:*

$Unbekanntheit(((Alter \in [20;28)) \wedge (studiert \in \{ja, nein\}) \wedge (Geschlecht \in \{männlich\}) \rightarrow (Einkommen \in [0;3000))), ((Alter \in [18;25)) \wedge (studiert \in \{ja\}) \wedge (Geschlecht \in \{weiblich, männlich\})) \rightarrow (Einkommen \in [0;2000))) = 1/2 \cdot (7/16+0) = 7/32 \approx 0{,}219.$

*Für den umgekehrten Fall, daß die zweitgenannte Regel im Benutzermodell vorliegt und die erstgenannte neu generiert wurde, ergibt sich eine Unbekanntheit von* $1/2 \cdot (3/8+1/3) = 17/48 \approx 0{,}354$. *Dieser Wert liegt über dem im ersten Fall, da die erste Regel eine wesentlich präzisere Konklusion und nur eine geringfügig speziellere Prämisse aufweist als die zweite Regel.*

Damit läßt sich nun die Unbekanntheit einer Regel als minimale Unbekanntheit zu den Regeln des Benutzermodells definieren:

**Definition 3-11: Unbekanntheit einer Regel**

Es gelten dieselben Voraussetzungen wie in Definition 3-10. Weiterhin sei ein Benutzermodell gegeben, das Vorwissen in Regelform enthält:

$B = \{(Pr_1^{BM} \rightarrow Ko_1^{BM}), \ldots, (Pr_{Bmax}^{BM} \rightarrow Ko_{Bmax}^{BM})\};$
$(Pr_i^{BM} \rightarrow Ko_i^{BM}) \in D^{KNF}(C,D);$
$i = 1, \ldots, Bmax.$

Die Unbekanntheit einer Regel, $(Pr \rightarrow Ko) \in DM^{KNF}(C,D)$, sei nun wie folgt definiert:

$Unbekanntheit_B$: $DM^{KNF}(C,D) \rightarrow \{r \mid r \in \mathbf{R}, 0 \le r \le 1\};$

$Pr \rightarrow Ko \rightarrow Unbekanntheit_B(Pr \rightarrow Ko);$

mit $Unbekanntheit_B(Pr \rightarrow Ko) :=$
$\min_{i=1,\ldots,Bmax} Unbekanntheit((Pr \rightarrow Ko), (Pr_i^{BM} \rightarrow Ko_i^{BM}))$. ◊

Da die Unbekanntheit von dem Benutzermodell $B$ abhängt und $B$ zwischen mehreren Ausführungen der Data-Mining-Phase ändern kann, kann die Interessantheit von Erklärungsmodellen nicht über mehrere Data-Mining-Läufe hinweg verglichen werden.

#### *3.3.2.4 Konzeption zur Generierung von Beschreibungsmodellen*

Überträgt man den Begriff des Beschreibungsmodells aus Abschnitt 3.1 auf die im Data Mining generierten Datenmuster, so kann man ein Data-Mining-Beschreibungsmodell wie folgt definieren:

**Definition 3-12: Data-Mining-Beschreibungsmodell**

Gegeben sei mit $A$ eine Menge von Attributen und mit $O^T$ eine Trainingsmenge. Ein *Data-Mining-Beschreibungsmodell*, $M^{Bes} = M_{O^T}$, ist eine Menge von Datenmustern, die in der Trainingsmenge betriebswirtschaftlich relevante Segmente identifiziert und beschreibt. Formal:

$M^{Bes} = \{ s_{O_1}, \ldots, s_{O_M} \}$ mit

$s_{O_i} = ( Pr_i \rightarrow Ko_i );$

$Ko_i = (\text{Segment } i );$

$i = 1, \ldots, M.$

Ein Datenmuster, $Pr_i \rightarrow Ko_i$, kann wie folgt als Regel gelesen werden:

WENN Planungsobjekte die Merkmale $Pr_i$ aufweisen,
DANN (und nur dann) bezeichne sie als „Segment $i$".

Dabei stellt „Segment $i$" die Bezeichnung, $Pr_i$ die Beschreibung und $O_i = O^T[Pr_i]$ die Menge der Planungsobjekte des $i$-ten Segmentes dar. ◊

Beschreibungsmodellen kommt im Data Mining die Aufgabe zu, betriebswirtschaftliche Planungsobjekte geordnet zu erfassen und darzustellen. Zur Generierung eines Beschreibungsmodells gehört die *Identifikation und Beschreibung der relevanten Objektklassen*. Anhand einer derartigen Beschreibung kann für jedes Objekt entschieden werden, in welche Klasse(n) es eingeordnet wird. Außerdem kann anhand der Segmentbeschreibung die Planung von segmentspezifischen Maßnahmen vorgenommen werden.

Der Segmentierung von Entscheidungssituationen liegen nach MEFFERT folgende Zielsetzungen zugrunde:[334]

⇨ **Homogenität**[335] **innerhalb eines Segmentes**: Die Planungsobjekte innerhalb eines Segmentes sollen möglichst ähnliche entscheidungsrelevante Eigenschaften

[334] Vgl. MEFFERT (1991a), S. 243.

[335] Ein mögliches Maß für die Homogenität wurde in Definition 2-69 bereits vorgestellt.

aufweisen, so daß für alle Objekte eines Segmentes dieselbe Entscheidung getroffen werden kann. Je homogener das Segment ist, desto wirksamer sind die aus den Entscheidungen resultierenden Maßnahmen.

- ⇨ **Heterogenität**[336] **zwischen verschiedenen Segmenten**: Objekte aus verschiedenen Segmenten sollten – falls mit den verschiedenen Segmenten sich gegenseitig ausschließende Entscheidungen assoziiert sind – möglichst unterschiedliche entscheidungsrelevante Eigenschaften aufweisen, da nur dann klar getrennt werden kann, welchem Planungsobjekt welche Entscheidung zuzuordnen ist.
- ⇨ **Erfolgsrelevanz**[337]: Jedes Segment sollte eine gewisse Mindestgröße aufweisen, und die darin enthaltenen Planungsobjekte sollten einen gewissen Erfolgsbeitrag leisten, da die Planung und Kontrolle von kleinen bzw. irrelevanten Segmenten unwirtschaftlich ist.
- ⇨ **Anzahl der Segmente**: Wenn für jedes Segment eine eigene Entscheidung getroffen und realisiert werden soll, so steigen mit der Anzahl der Segmente die Kosten der Planung und Realisierung. Daher ist die Anzahl der Segmente gering zu halten.
- ⇨ **Operationalität einer Segmentbeschreibung**: Ein Segmentierungsmodell muß operational sein, damit der Benutzer daraus Handlungsbedarfe erkennen und Maßnahmen ableiten kann. Die Operationalität kann kaum durch ein automatisches Verfahren bewertet werden, so daß hier auf eine Teilkomponente der Operationalität, die *Verständlichkeit*[338] *einer Segmentbeschreibung*, abgestellt wird.

Diese in der Literatur genannten Zielvorstellungen reichen nicht aus, um Segmente zu bilden, deren Bearbeitung *Synergieeffekte* freisetzt.

*Werden beispielsweise Warenkorbprofile gebildet, um die in einem Warenkorb enthaltenen Artikel gemeinsam zu plazieren, so soll die gemeinsame Plazierung zusätzliche Kaufanreize auslösen. Der Synergieeffekt besteht dann darin, daß die Kaufhäufigkeit für den gemeinsam plazierten Warenkorb die Kaufhäufigkeit bei unabhängiger Plazierung der einzelnen Artikel signifikant übersteigt.*

Um diese Synergieeffekte zu berücksichtigen, wird ein Maß für die *Stärke der symmetrischen Zusammenhänge* zwischen den assoziierten Planungsobjekten benötigt.[339] Das

---

[336] Die Heterogenität einer Segmentierung wurde bereits in Definition 2-71 operational bestimmt.

[337] Die Erfolgsrelevanz eines Segmentes ist dieselbe wie in Definition 3-9 für eine erklärende Regel.

[338] Vgl. das zur Verständlichkeit von Erklärungsmodellen im Abschnitt zuvor Gesagte.

[339] Ein enstprechendes Interessantheitsmaß für ein Datenmuster liefert Definition 2-64.

Data-Mining-Verfahren soll solche Merkmale zur Segmentierung auswählen, die eine starke gegenseitige Abhängigkeit aufweisen. Im Vergleich zu der Stärke des Ursache-Wirkungszusammenhangs bei Erklärungsmodellen geht es hier um ungerichtete (symmetrische, gegenseitige) Abhängigkeiten, da bei Beschreibungsmodellen keine erklärenden und zu erklärenden Größen unterschieden werden. Ähnlich wie bei den Erklärungsmodellen wird hier der symmetrische Zusammenhang eines Datenmusters aus Definition 2-64 auf eine Datenmuster-Menge erweitert:

**Definition 3-13: Stärke der Zusammenhänge eines Beschreibungsmodells**

$L$ sei die Menge aller möglichen Beschreibungsmodelle gemäß Definition 3-12 und $\boldsymbol{R}$ die Menge der reellen Zahlen. Dann ist die *Stärke der Zusammenhänge eines Erklärungsmodells* aus $L$ wie folgt definiert:

$$Zusammenhang^{M,s}: \; L \rightarrow \{r \mid r \in \boldsymbol{R},\ 0 \le r \le 1\};$$

$$M^{Bes} \rightarrow Zusammenhang^{M,s}(M^{Bes});$$

$$\text{mit } Zusammenhang^{M,s}(M^{Bes}) := \frac{1}{M} \sum_{j=1}^{M} Zusammenhang^{s}(Pr_j \rightarrow Ko_j). \quad \Diamond$$

Analog zur Erklärungsgüte sollte hier durch die ***Beschreibungsgüte*** gemessen werden, *wie gut* und *wie vollständig* die Menge der Planungsobjekte beschrieben werden kann. Die Beschreibungsgüte setzt sich entsprechend aus den Facetten der „***Stärke der Zusammenhänge eines Beschreibungsmodells***“ und der „***Erfolgsrelevanz***“ der insgesamt beschriebenen Objektmenge zusammen:

**Definition 3-14: Beschreibungsgüte eines Beschreibungsmodells**

Es gelten dieselben Voraussetzungen wie in der Definition zuvor sowie in Definition 3-6. Dann ist die *Beschreibungsgüte* eines Beschreibungsmodells, $M^{Bes} = \{(Pr_1 \rightarrow Ko_1), ..., (Pr_M \rightarrow Ko_M)\}$, wie folgt definiert:

$$Beschreibungsgüte: \; L \rightarrow \{r \mid r \in \boldsymbol{R},\ 0 \le r \le 1\};$$

$$M^{Bes} \rightarrow Beschreibungsgüte(M^{Bes});$$

$$\text{mit } Beschreibungsgüte(M^{Bes}) :=$$
$$Erfolgsrelevanz^{M}(M^{Bes}) \cdot w^{E} + Zusammenhang^{M,s}(M^{Bes}) \cdot (1-w^{E});$$
$$w^{E} \in \{r \mid r \in \boldsymbol{R},\ 0 \le r \le 1\}. \quad \Diamond$$

Hinzu kommen die bereits im Zusammenhang mit Erklärungsmodellen erwähnten Facetten der **Unbekanntheit** und der **Redundanzfreiheit**. Die Genauigkeit der Segmentierungsbeschreibung nach Definition 2-68 wird in diesem Bewertungskonzept nicht berücksichtigt, da dieses Maß nur für endliche Domänen geeignet ist und darüber hinaus die Häufigkeitsverteilung der zur Segmentierung verwendeten Merkmale ignoriert.

#### *3.3.2.5 Zusammenfassende Betrachtung der Konzeptionen*

In den vorangegangenen Abschnitten wurden relevante Interessantheitsfacetten für die Bewertung der eingeführten Modelltypen hergeleitet. Abbildung 3-13 faßt diese Facetten noch einmal zusammen.

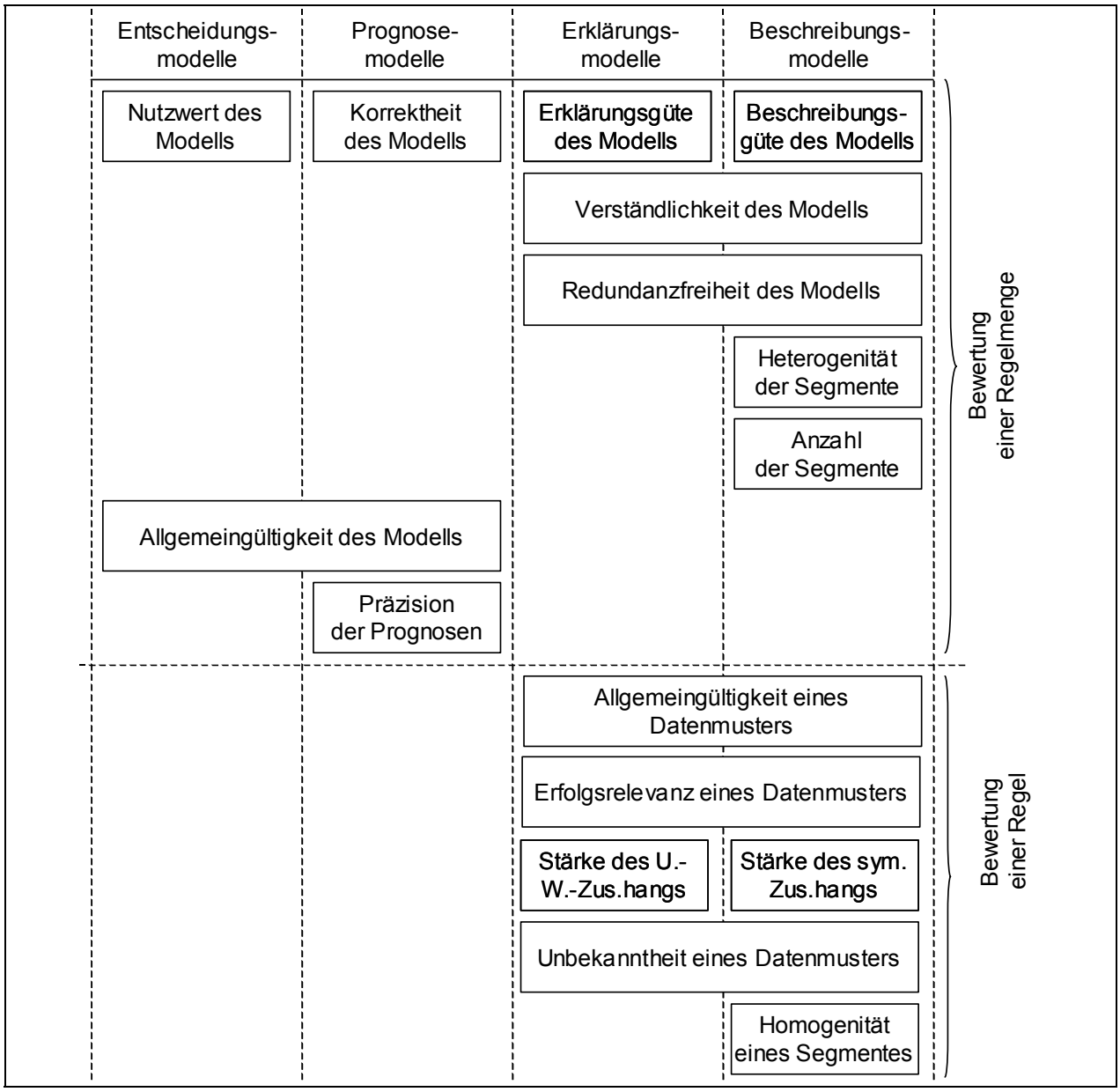



**Abbildung 3-13: Interessantheitskonzeption für das Data-Mining-Verfahren**

Die meisten Facetten wurden bereits hinreichend besprochen und müssen hier nicht noch einmal erläutert werden. Bemerkenswert sind folgende vergleichende Aussagen der Abbildung:

- ⇨ Die Interessantheitsfacetten in der ersten Zeile der Abbildung – d.h. der **Nutzwert**, die **Korrektheit**, die **Erklärungsgüte** und die **Beschreibungsgüte** – machen jeweils den Kern der Modellbewertung aus und grenzen sie voneinander ab.
- ⇨ Die **Verständlichkeit** eines Modells ist nur bei den Modelltypen, die intellektuell interpretiert werden – also bei Erklärungs- und Beschreibungsmodellen –, relevant. Selbiges gilt für die **Redundanzfreiheit** des Gesamtmodells und alle Facetten, die ein einzelnes Datenmuster bewerten.
- ⇨ Die **Allgemeingültigkeit** von Regeln garantiert die generelle Verwendbarkeit der Merkmale eines Data-Mining-Erklärungsmodells. Prognose- und Entscheidungsmodelle sind funktionale Data-Mining-Modelle gemäß Definition 2-22, deren Outputs durch Anwendung mehrerer Regeln zustande kommen können, so daß hier anstelle der Allgemeingültigkeit einzelner Regeln die **Allgemeingültigkeit**[340] **des gesamten Modells** relevant ist.
- ⇨ Aus demselben Grund wie bei der Allgemeingültigkeit muß auch die **Präzision der Prognosen**[341] für das gesamte Modell bestimmt werden. Dagegen müssen die Outputs von Entscheidungsmodellen nicht bezüglich ihrer Präzision beurteilt werden, da in der vorgestellten Konzeption diskrete Alternativenmengen vorgegeben werden – gibt man die Alternativen präzise vor, so sind auch die Outputs des Modells präzise.

---

[340] Diese könnte als gewichtetes Mittel der an der Prognose bzw. Entscheidung beteiligten Regeln definiert werden – wobei das Gewicht dasselbe sein müßte, mit dem die jeweilige Regel die Prognose bzw. Entscheidung beeinflußt. Formal definiert werden kann dieses Maß nur für eine konkrete Akkumulationsfunktion gemäß Definition 2-21, da diese festlegt, in welcher Weise der Modelloutput erzeugt wird.

[341] Ein entsprechendes Präzisionsmaß ließe sich beispielsweise in Anlehnung an Definition 2-62 formalisieren. Wenn die Ausgabe des Prognosemodells als Term in KNF dargestellt würde, so könnte man die Präzision als Gegenwahrscheinlichkeit zur A-priori-Wahrscheinlichkeit für die Erfülltheit der Prognose gemäß Definition 2-61 berechnen.

### 3.3.3 Anwendung des Data-Mining-Modells zur Entscheidungsunterstützung

Der dritte Schritt innerhalb des Problemlösungsschemas für Data-Mining-Anwendungen besteht in der Anwendung des konzipierten Data-Mining-Verfahrens. Dies stellt den Data-Mining-Schritt innerhalb des KDD-Prozesses dar. Im Rahmen der betriebswirtschaftlichen Entscheidungsfindung ist aber eher die Anwendung der im Data Mining generierten Modelle von Interesse. Die Anwendung der Modelle zur Entscheidungsunterstützung wird im folgenden für die 11 in Abschnitt 3.3.1 eingeführten Problemklassen vorgestellt. Abschnitt 3.3.3.1 diskutiert die Anwendung von Beschreibungsmodellen, Abschnitt 3.3.3.2 die Anwendung von Erklärungsmodellen, Abschnitt 3.3.3.3 die Anwendung von Prognosemodellen und Abschnitt 3.3.3.4 die Anwendung von Entscheidungsmodellen.

#### *3.3.3.1 Anwendung von Data-Mining-Beschreibungsmodellen*

Data-Mining-Beschreibungsmodelle charakterisieren, wie gesehen, Segmente typischer Entscheidungssituationen in folgender Form:

Segment $l$ := Beschreibung der Entscheidungssituation $l$;

$\vdots$

Segment $M$ := Beschreibung der Entscheidungssituation $M$.

Betrachtet man die durch ein Beschreibungsmodell charakterisierten Objektmengen als Aufteilung der Gesamtmenge aller Planungsobjekte, so sind die entstehenden Teilmengen kleiner und bezüglich bestimmter Kriterien homogener und damit leichter zu planen und zu kontrollieren als die gesamte Objektmenge. Insbesondere liegt der Nutzen der Aufteilung in einer *Erhöhung der Erfolgswirksamkeit von Maßnahmen*, die sich nun nicht auf die heterogene Gesamtmenge, sondern auf homogenere Teilmengen beziehen.

*Beispielsweise kann die Gesamtmenge der Kunden derart segmentiert werden, daß sich in jedem Segment Kunden mit ähnlichem Verhalten befinden. Insbesondere die Reaktion auf eine gegebene Marketingmaßnahme soll bei den Kunden eines Segmentes identisch sein. Auf der Grundlage einer derartigen Segmentbildung mit entsprechender Charakterisierung der darin enthaltenen Kunden können dann Marketingmaßnahmen sehr differenziert auf das jeweilige Segment zugeschnitten werden (z.B. Mailings an Kunden mit Interesse an Artikel X).*

Betrachtet man die charakterisierten Objektmengen umgekehrt als Zusammenfassung von Einzelobjekten, so liegt der Nutzen einer derartigen Zusammenfassung in der

Komplexitätsreduktion. In betriebswirtschaftlichen Anwendungen ist damit i.d.R. eine *Kostensenkung* verbunden.

*In dieser Sichtweise kann man Kunden so zu Kundengruppen zusammenfassen, daß nicht für jeden Kunden eine eigene Marketingmaßnahme konzipiert und umgesetzt werden muß, sondern nur eine Maßnahme pro Kundengruppe.*

In jedem Fall geschieht die Beschreibung der Objektmengen zu dem Zweck, betriebswirtschaftlich orientierte *Handlungen auf die spezifischen Eigenschaften einer Objektmenge zuzuschneiden*. D.h., letztendlich ist auch hier eine betriebswirtschaftliche Gestaltungsaufgabe und damit ein Entscheidungsproblem zu lösen. Dies wirft das Problem auf, daß sich bereits die Vorauswahl der Segmentierungsmerkmale an der zu unterstützenden Entscheidung orientieren muß, da sonst Segmente mit unbrauchbaren Merkmalen gebildet werden. Der Entscheider wird also nicht dadurch unterstützt, daß das Data Mining ihm die relevanten Segmentierungsmerkmale liefert, sondern dadurch, daß für ein noch nicht genau strukturiertes Planungsproblem die zu betrachtenden Entscheidungssituationen gruppiert und beschrieben werden. Durch die Gruppierung wird das gesamte Planungsproblem in mehrere Unterprobleme aufgeteilt, für die anhand der Beschreibungen Handlungsalternativen konkretisiert werden können.

*Beispielsweise könnte das Fernsehverhalten der Deutschen segmentiert werden. Als Variablen könnten z.B. die durchschnittliche Sehdauer von Genres wie Actionfilmen, Serien, Nachrichten, Sport usw. zur Verfügung gestellt werden. Damit würden sich möglicherweise Segmente ergeben, die verschiedene Typen von Zuschauern kennzeichnen. Auf der Grundlage der Zuschauertypen könnte ein Unternehmen seine Werbeschaltungen auf die zwischen den entsprechenden Fernsehangeboten liegenden Werbeblöcke konzentrieren. Wenn ein Segment mehrere Genres mit überdurchschnittlicher Sehdauer erkennen läßt, besteht für die Werbetreibenden eine erhöhte Chance, dieselben Zuschauer mehrfach zu kontaktieren. Der Nutzen der Segmentierung liegt demnach darin, die große Anzahl von Genre-Kombinationen, die als Alternativen für Werbekampagnen zur Verfügung stünden, auf eine überschaubare Anzahl einzugrenzen.*

#### *3.3.3.2 Anwendung von Data-Mining-Erklärungsmodellen*

Im Rahmen der Anwendung von Data-Mining-Erklärungsmodellen werden gemäß Abbildung 3-11 drei Problemklassen unterschieden.

Der Fall, daß Erklärungsmodelle **Abhängigkeiten zwischen Entscheidungssituationen** charakterisieren, läßt sich wie folgt in Regelform schreiben:

$$\texttt{Entscheidungssituation } Pr \texttt{ erklärt Entscheidungssituation } Ko.$$

Bei der Erklärung von Entscheidungssituationen geht es in betriebswirtschaftlichen Anwendungen fast immer um die Analyse des Verhaltens oder der Einstellung von Kunden. Derartige Erklärungsmodelle erhöhen den Kenntnisstand des Marketingplaners über die Kunden und können neue Potentiale offenlegen, die Kunden anhand der ermittelten Merkmale anzusprechen, um eine gewünschte Reaktion hervorzurufen. Das erklärte Phänomen muß dabei in engem Zusammenhang mit der erwünschten Kundenreaktion stehen.

*Falls beispielsweise die Variable „Billigkäufer" erklärt wurde, so sollen die Kunden zum Kauf von Billigprodukten animiert werden.*

In der Kontrollphase kommen – je nach Datenbasis – Modelle zur **Erklärung von Handlungsergebnissen oder Zielbeiträgen** zur Anwendung. Diese modellieren dann *Wirkungs- oder Zielerreichungsfunktionen* in der Form

$$\texttt{Entscheidungssituation} \wedge \texttt{Alternative } Pr \texttt{ erklärt Handlungsergebnis } Ko$$

bzw.

$$\texttt{Entscheidungssituation} \wedge \texttt{Alternative } Pr \texttt{ erklärt Zielbeitrag } Ko.$$

Die Ursachen für das Zustandekommen der Handlungsergebnisse bzw. Zielbeiträge können im Verantwortungsbereich der eigenen Handlungsplanung oder -realisierung oder in der tatsächlich eingetretenen Umweltentwicklung liegen. Aufgrund der Allgemeingültigkeit der Aussagen können nicht nur die Verursacher unerwünschter Ergebnisse zur Verantwortung gezogen werden. Vielmehr können die erklärenden Merkmale die Grundlage für zukünftige Planungen bilden, so daß die Kontrollphase fließend in die weitere Planung übergeht.

*Um beispielsweise einen Vertreter für die von ihm vergebenen, ungerechtfertigt hohen Rabatte zur Verantwortung zu ziehen, genügt die Identifizierung des betreffenden Vertreters. Data-Mining-Erklärungsmodelle leisten über die Identifizierung hinaus eine Beschreibung solcher Sachverhalte. Die extrahierten Merkmale von Vertretern, die hohe Rabatte vergeben, können dann unabhängig von den konkreten Vertretern – beispielsweise bei der zukünftigen Personalauswahl – berücksichtigt werden.*

#### *3.3.3.3 Anwendung von Data-Mining-Prognosemodellen*

Bei der Anwendung von Data-Mining-Prognosemodellen werden nach der Zusammensetzung der Trainingstabelle gemäß Abbildung 3-12 sechs Fälle unterschieden.

**Im Falle 1** kann ein Modell, $ES \rightarrow U$, zur Prognose einer zukünftigen Umweltsituation, $u^* \in U$, auf der Grundlage der aktuell vorliegenden Entscheidungssituation, $es \in ES$,

induziert werden. In Abbildung 3-14 soll das obere Rechteck andeuten, daß die Umweltsituation $u^*$ *modellendogen* bestimmt wird. Das darunterliegende Rechteck soll andeuten, welche Handlungsergebnisse bzw. Zielbeiträge, ausgehend von der Kenntnis von $u^*$, im folgenden *modellexogen* betrachtet werden. Dann kann in einem manuellen Entscheidungsprozeß eine *modellexogene* Zielerreichungsfunktion über alle Handlungsalternativen maximiert werden, um die für die vorhergesagte Umweltentwicklung, $u^*$, günstigste Entscheidung zu treffen:

$\max_{h \in H} f^Z(h,u^*)$.[342]

Das kleine Rechteck im linken Teil der Abbildung deutet an, daß die optimale Handlung, $h^*$, *modellexogen* bestimmt wird.

Im Hinblick auf eine vollständige Entscheidungsunterstützung sind folgende Aspekte kritisch zu sehen:

- ⇨ Es wird i.d.R. nur eine bestimmte Umweltsituation, $u^*$, und nicht die gesamte Wahrscheinlichkeitsverteilung prognostiziert, so daß die Risikopräferenzen des Entscheidungsträgers nicht in das Modell einfließen können. Dies ist dann tolerierbar, wenn die Wahrscheinlichkeit $P(u^*)$ sehr hoch ist, was jedoch a-priori nicht bekannt ist, sondern sich erst bei der Anwendung des Prognosemodells zeigt. Somit bietet sich diese Anwendung vor allem dann an, wenn nur zwei Umweltsituationen unterschieden werden. In diesem Falle ergibt sich die Wahrscheinlichkeit für die alternative Umweltsituation, $\neg u^*$, als $P(\neg u^*) = 1 - P(u^*)$.
- ⇨ Bei der Generierung des Prognosemodells wird keine ökonomische Zielfunktion optimiert, sondern die Modellkorrektheit gemäß Definition 2-54 maximiert.[343] Dadurch wird nicht etwa auf die gute Empfehlung besonders erfolgskritischer Entscheidungen abgestellt, sondern auf eine im Durchschnitt über alle Entscheidungssituationen gute Empfehlung.

---

[342] In dieser Notation wurde von eindimensionalen Zielbeiträgen ausgegangen. Mehrfachzielsetzungen müßten noch in eine eindimensionale Zielfunktion transformiert werden, um sie optimieren zu können.

[343] Das damit verbundene Problem wurde bereits in Abschnitt 3.2.6 diskutiert.

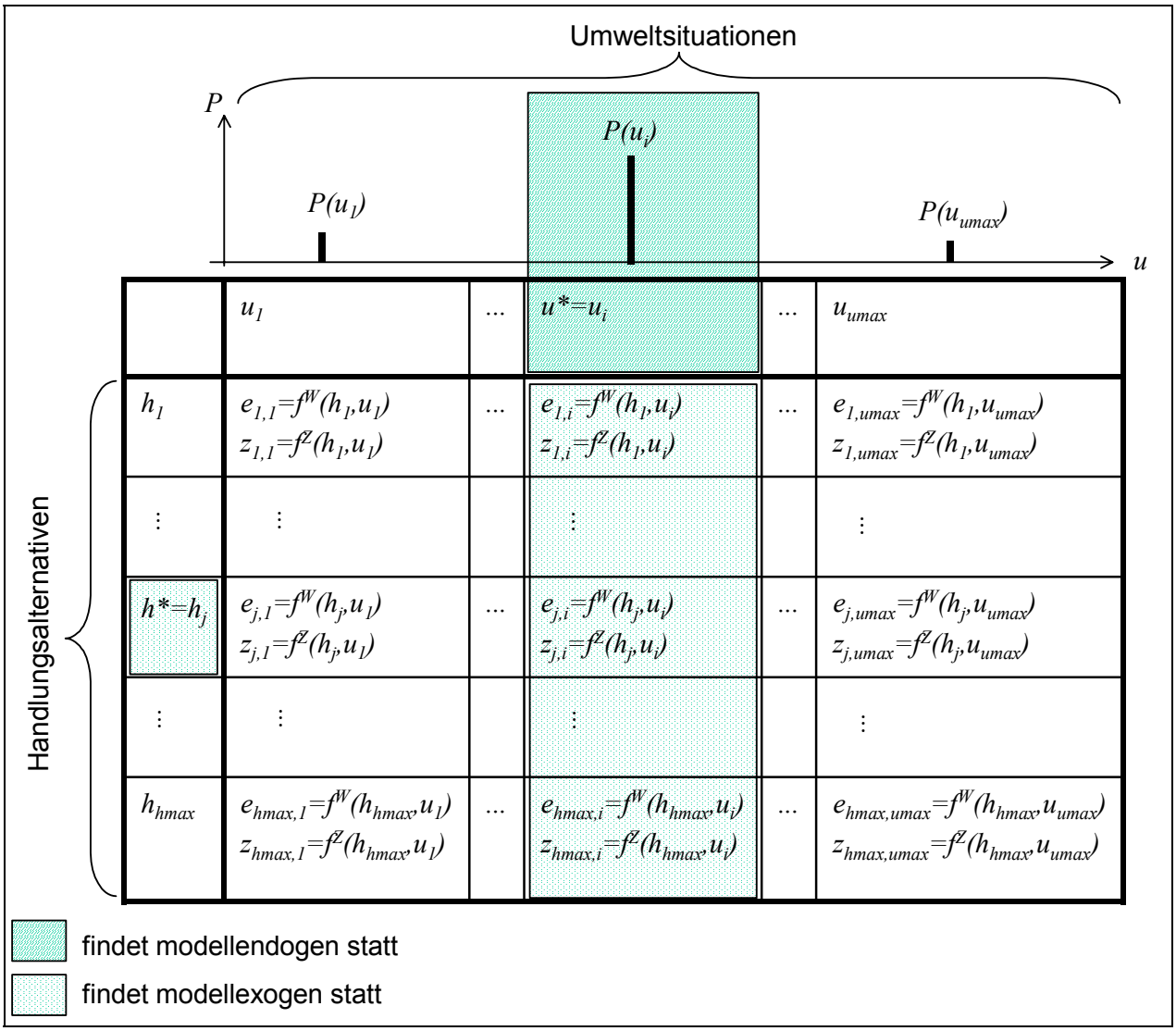



**Abbildung 3-14: Entscheidungsfindung bei Verwendung von Data-Mining-Modellen zur Prognose einer Umweltsituation (Fall 1)**

Um den zweitgenannten Nachteil zu umgehen, müßte, wie im nächsten Abschnitt diskutiert wird, ein Data-Mining-Entscheidungsmodell generiert werden. Und um den erstgenannten Nachteil zu umgehen, müßte – was kaum ein Verfahren leistet – die gesamte Wahrscheinlichkeitsverteilung der zukünftigen Umweltsituationen prognostiziert werden. Die Entscheidungsfindung für diesen Fall stellt Abbildung 3-15 dar. Während die Umweltentwicklung *modellendogen* bestimmt wird, muß die Bewertung aller *(u,h)*-Kombintionen *modellexogen* erfolgen – ebenso wie die anschließende Ermittlung der Nutzwerte und der Auswahl der Alternative, $h^*$, mit dem höchsten Nutzwert, $n^*$.

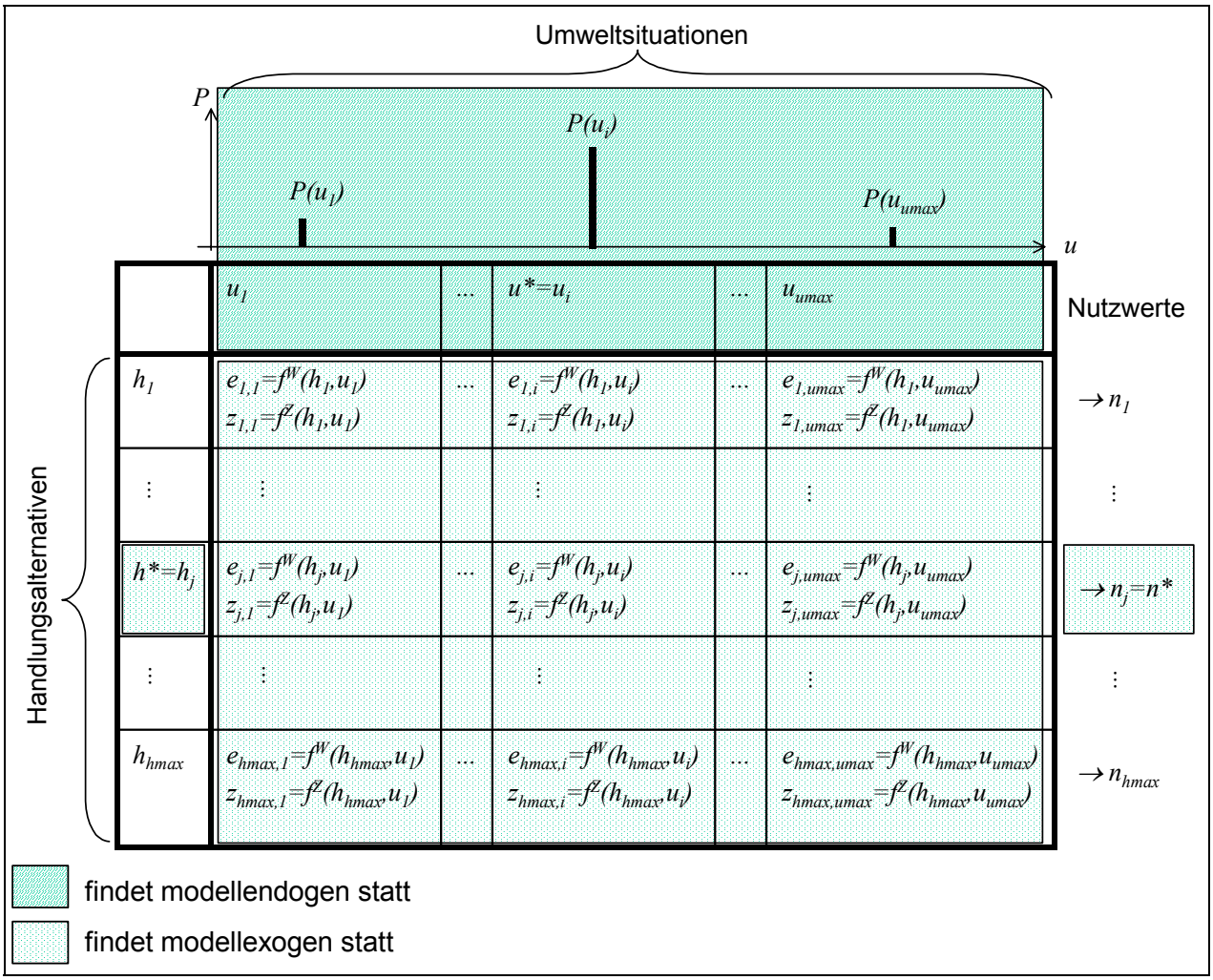



**Abbildung 3-15: Entscheidungsfindung bei Verwendung von Data-Mining-Modellen zur Prognose der Verteilung der Umweltsituationen (Fall 1)**

**Im Falle 2 bzw. 3** kann ein Modell zur Prognose der erzielbaren Handlungsergebnisse, $e_i^+$, bzw. Zielbeiträge, $z_i^+$, induziert werden, welche bei impliziter Unterstellung einer Handlung, $h^+$, erreicht werden können, wenn eine implizit prognostizierte Umweltsituation, $u^*=u_i$, eintritt. Im mittleren Teil der Abbildung 3-16 soll das Rechteck andeuten, daß $e_i^+$ bzw. $z_i^+$ *modellendogen* bestimmt wird. Da $h^+$ implizit unterstellt wurde, läßt sich anhand der Prognose von $e_i^+$ bzw. $z_i^+$ aus dem Modell heraus nicht entscheiden, ob die anderen Alternativen aus $H-\{h^+\}$ besser oder schlechter abschneiden als $h^+$. Im Falle 2 müssen die Handlungsergebnisse exogen in Zielbeiträge transformiert werden können. Wie die Anwendungen in Abschnitt 3.2.6 zeigen, ist die Alternativenmenge häufig zweiwertig mit $H=\{h^+,h^-\}$ und in der Trainingsmenge wird $h^+$ implizit vorausgesetzt, da dort bspw. nur Kunden enthalten sind, denen ein Direktmarketingangebot zugensendet wurde. Der Zielbeitrag der Negativ-Handlung, $h^-$, beträgt Null für alle Umweltsituationen,

da bei Unterlassen der Handlung (z.B. kein Angebot zusenden) weder eine positive noch eine negative Wirkung erzielt wird: $\forall u \in U: z^- = f^Z(h^-,u) = 0$. Damit ist, wenn $z_i^+ > 0$ ($z_i^+ < 0$) vorhergesagt wird, eindeutig, daß $h^+$ ($h^-$) gewählt werden sollte. Diese *modell-exogene* Ermittlung der Entscheidung $h^*$ wird im linken Teil der Abbildung 3-16 durch das kleine Rechteck angedeutet.

Im Hinblick auf eine vollständige Entscheidungsunterstützung sind folgende Aspekte kritisch zu sehen:

- ⇨ Falls die Zielbeiträge der modellexogenen Alternativen aus $H - \{h^+\}$ nicht bekannt sind, kann die optimale Handlungsalternative nicht bestimmt werden.
- ⇨ Wie im Falle 1 wird auch hier keine betriebswirtschaftlich orientierte Zielfunktion optimiert.
- ⇨ Da implizit genau eine bestimmte Umweltsituation, $u^*$, prognostiziert wird, gelten wieder die im Falle 1 genannten Einschränkungen (keine Berücksichtigung der Risikopräferenzen).

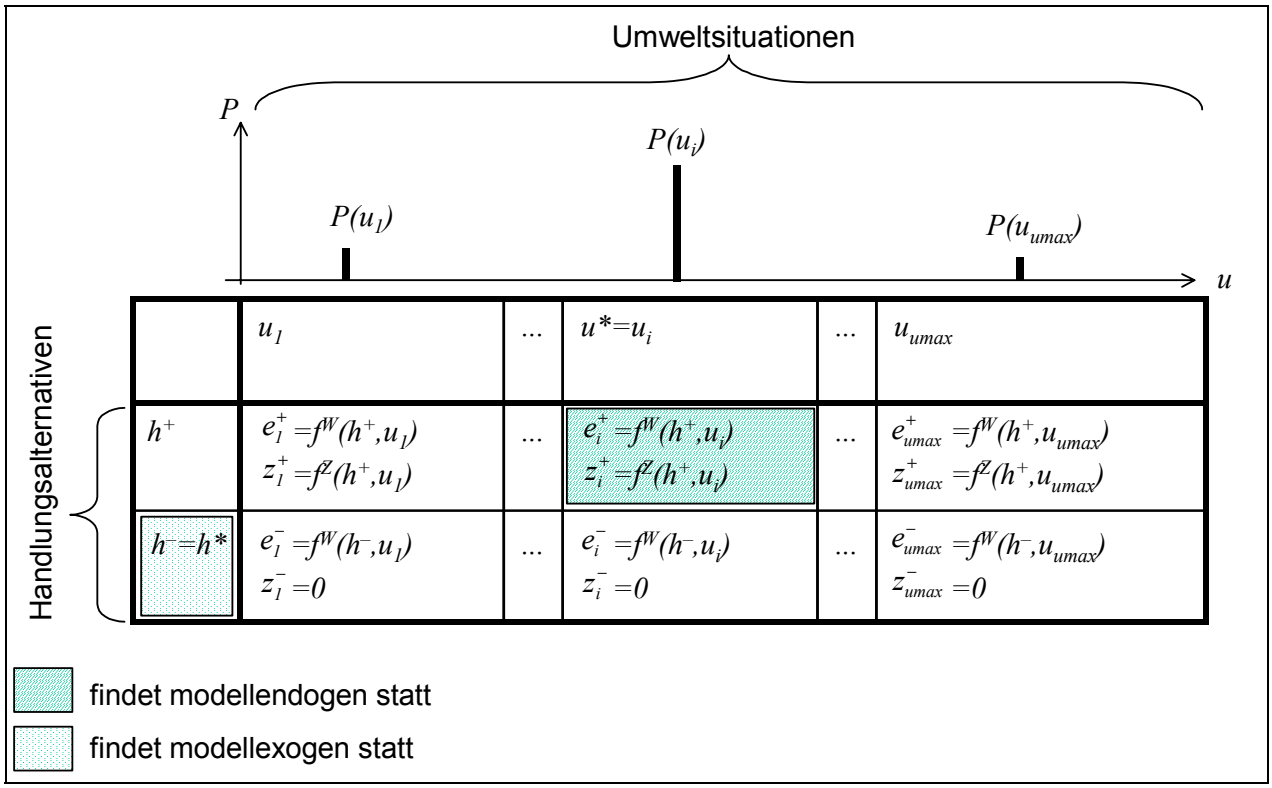



**Abbildung 3-16: Entscheidungsfindung bei Verwendung von Modellen zur Prognose eines Handlungsergebnisses oder Zielbeitrags (Fall 2/3)**

Um auch in diesem Fall den letztgenannten Nachteil zu umgehen, müßte wieder die gesamte Wahrscheinlichkeitsverteilung der erzielbaren Handlungsergebnisse bzw. Zielbeiträge prognostiziert werden. Die Entscheidungsfindung für diesen Fall stellt Abbildung 3-17 dar. Während die gesuchte Verteilung *modellendogen* bestimmt wird, erfolgt die Auswahl der Alternative, $h^*$, mit dem höchsten Nutzwert, $n^*$, *modellexogen*. Auch diese Entscheidungsfindung ist nur dann möglich, wenn gilt: $H=\{h^+,h^-\}$, $n^-$ bekannt, wobei $n^-$ den Nutzwert der Negativ-Handlung, $h^-$, darstellt.

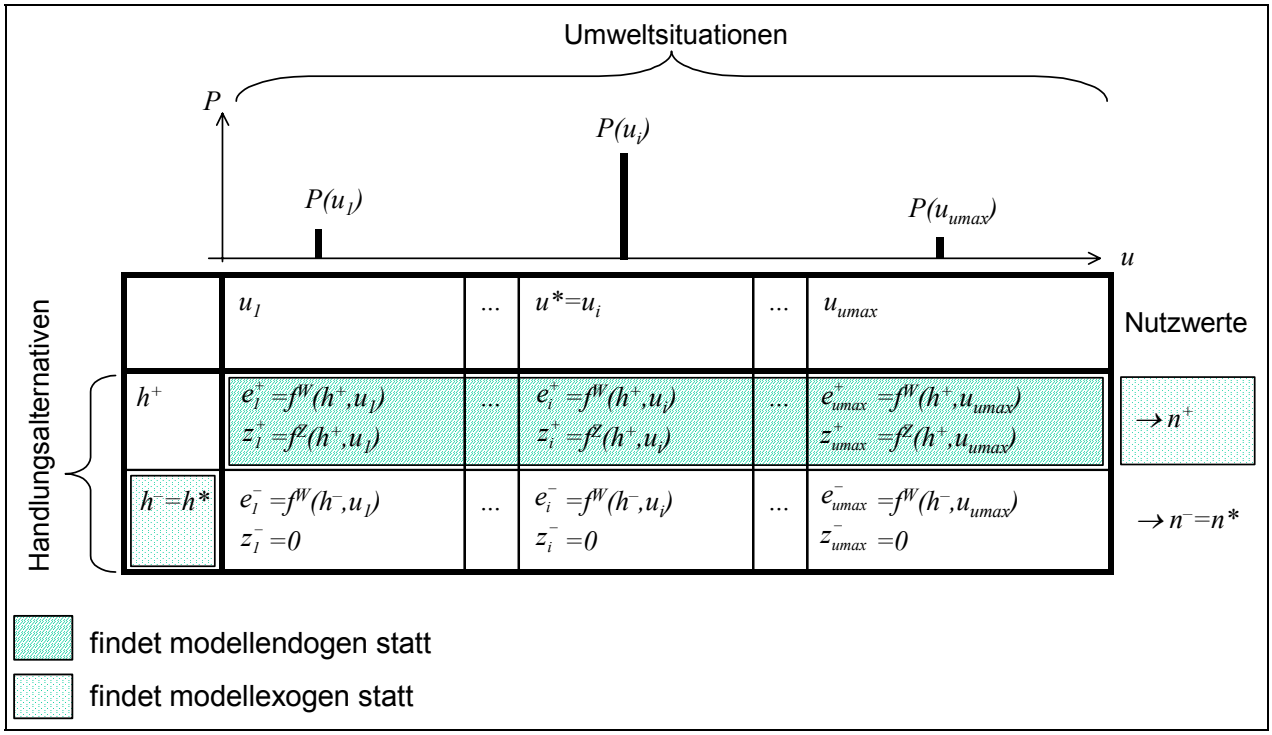



**Abbildung 3-17: Entscheidungsfindung bei Verwendung von Modellen zur Prognose der Ergebnis- oder Zielbeitragsverteilung (Fall 2/3)**

**Im Falle 4** wird ein Prognosemodell aufgestellt, das direkt für eine vorliegende Entscheidungssituation die Entscheidung ausgibt, die nach Maßstäben der Vergangenheit optimal war. Diese vollautomatische Entscheidungsfindung ist aus folgenden Gründen kritisch zu beurteilen:

1. Die Präferenzvorstellungen der Träger vergangener Entscheidungen sind in den Trainingsdaten implizit enthalten. Damit werden sie auch durch das Data-Mining-Verfahren erlernt und fließen in das entstehende Data-Mining-Modell ein. Diese

Fixierung im Modell kann dann gefährlich werden, wenn sich die Präferenzvorstellungen ändern[344] oder neue Handlungsalternativen möglich werden.

2. Der Entscheidungsträger hat keine Möglichkeiten mehr, eine andere Alternative zu wählen, da er deren Auswirkungen dem Modell nicht entnehmen kann.
3. Wie schon im Falle 1 wird auch hier während des Lernvorgangs keine betriebswirtschaftlich orientierte Zielfunktion optimiert.
4. Die mit dem Modelloutput verbundene Ungewißheit kann nicht in die Entscheidungsfindung einfließen.

**Im Falle 5 bzw. 6** kann, wenn kein Verfahren zur Generierung von Entscheidungsmodellen zur Verfügung gestellt werden kann, ein Modell zur Prognose der erzielbaren Handlungsergebnisse, $e_{j,i}$, bzw. Zielbeiträge, $z_{j,i}$, induziert werden. In Abbildung 3-18 sollen die gestreift ausgefüllten Rechtecke andeuten, daß $e_{j,i}$ bzw. $z_{j,i}$ *modellendogen* bestimmt wird. Diese Prognose muß für alle Alternativen, $h_j$ $(j = 1, ..., hmax)$, durchgeführt werden. Im Falle 5 müssen die Handlungsergebnisse exogen in Zielbeiträge transformiert werden können. Die *modellexogene* Ermittlung der Entscheidung $h^*$ wird in der Abbildung durch das kleine, gepunktet ausgefüllte Rechteck angedeutet.

Im Hinblick auf eine vollständige Entscheidungsunterstützung sind folgende Aspekte kritisch zu sehen:

⇨ Wenn das Modell aufgrund mangelnder Trainingsdaten nicht für alle Alternativen, $h_j$ $(j = 1, ..., hmax)$, Prognosen liefern kann, läßt sich nicht entscheiden, ob diese Alternativen besser oder schlechter abschneiden als die, auf die das Modell angewendet werden konnte.

⇨ Wie im Falle 1 wird auch hier keine betriebswirtschaftlich orientierte Zielfunktion optimiert.

⇨ Da implizit eine unbekannte Umweltsituation prognostiziert wird, gelten wieder die im Falle 1 genannten Einschränkungen (keine Berücksichtigung der Risikopräferenzen).

---

[344] Vgl. LACKES/MACK (2000), S. 62.

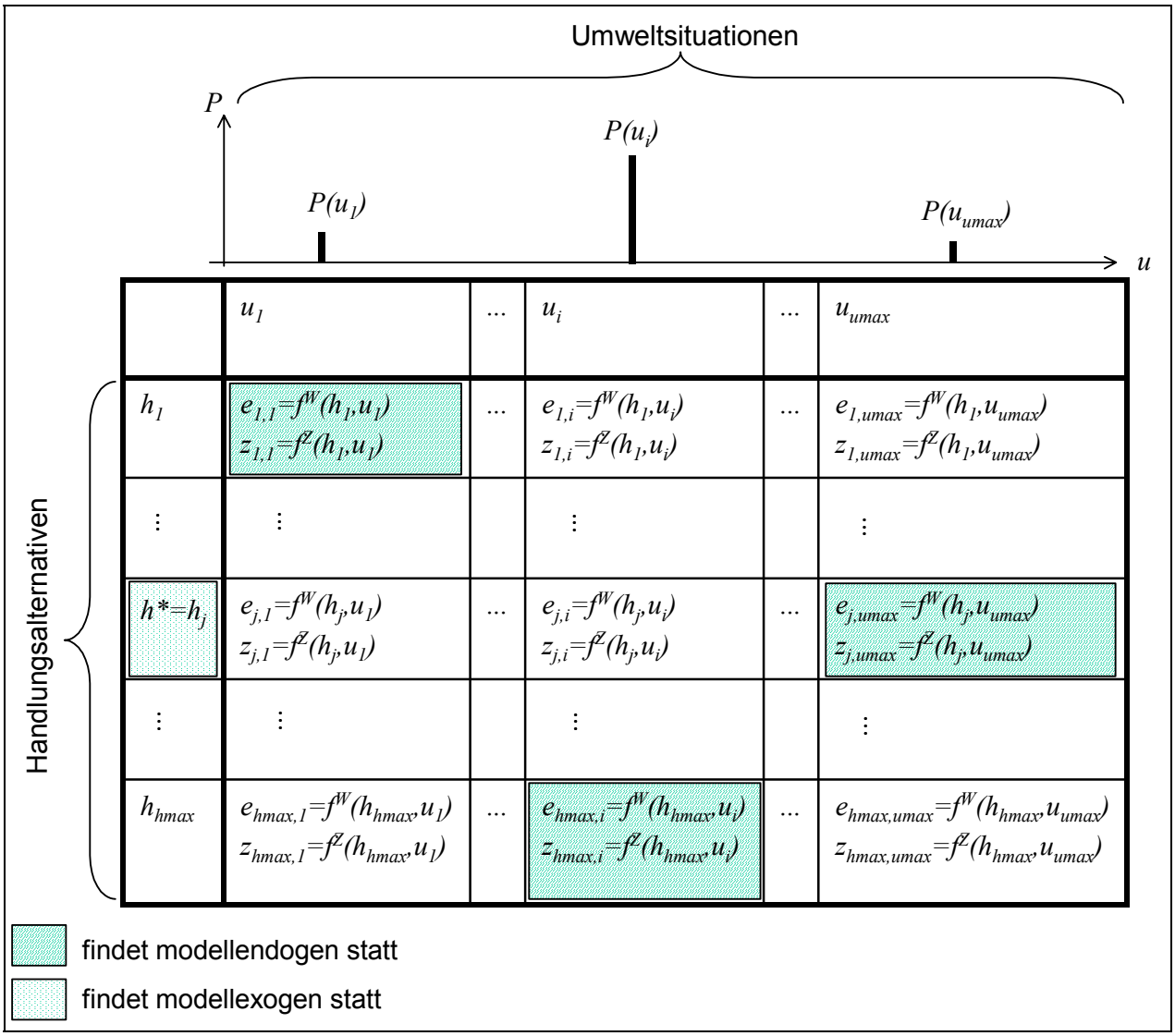



**Abbildung 3-18: Entscheidungsfindung bei Verwendung von Modellen zur Prognose eines Handlungsergebnisses oder Zielbeitrags (Fall 5/6)**

Um auch in diesem Fall den letztgenannten Nachteil zu umgehen, müßte wieder die gesamte Wahrscheinlichkeitsverteilung der erzielbaren Handlungsergebnisse bzw. Zielbeiträge prognostiziert werden. Die Entscheidungsfindung für diesen Fall stellt Abbildung 3-19 dar. Während die gesuchte Verteilung *modellendogen* bestimmt wird, erfolgt die Auswahl der Alternative, $h^*$, mit dem höchsten Nutzwert, $n^*$, *modellexogen*.

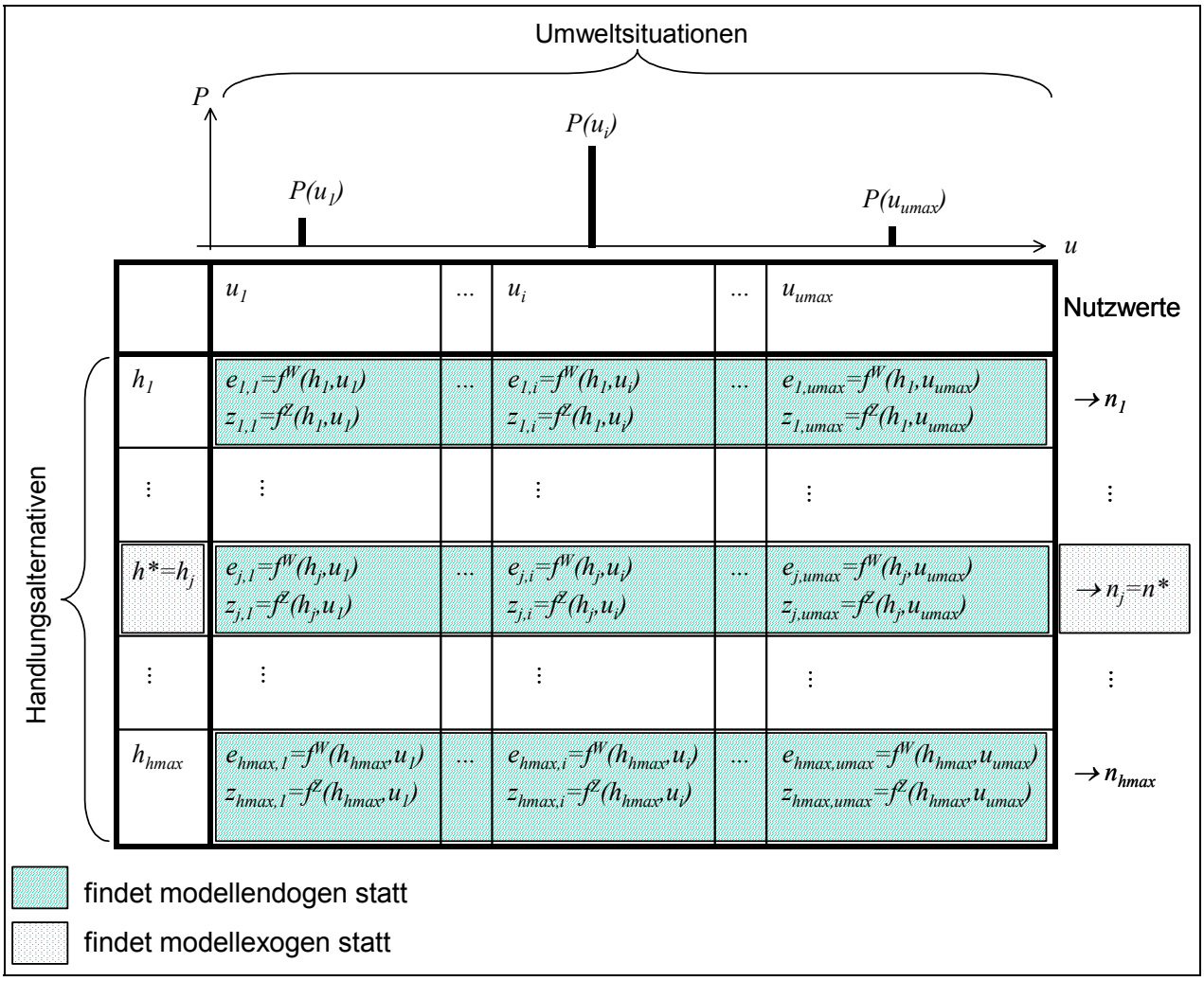



**Abbildung 3-19: Entscheidungsfindung bei Verwendung von Modellen zur Prognose der Ergebnis- oder Zielbeitragsverteilung (Fall 5/6)**

#### *3.3.3.4 Anwendung von Data-Mining-Entscheidungsmodellen*

Wurde die aus Abbildung 3-12 bekannte Problemklasse *„Entscheidung“* identifiziert und besteht die Möglichkeit, ein Data-Mining-Verfahren zur Generierung von Entscheidungsmodellen zu entwickeln, so kann die Entscheidungsfindung autonom stattfinden, und die bei der Anwendung von Prognosemodellen auftretenden Nachteile können umgangen werden.

Als Output liefert ein Data-Mining-Entscheidungsmodell direkt eine aus mehreren vorgebenen Handlungsempfehlungen, so daß zur Entscheidungsfindung keine manuellen Aktivitäten mehr notwendig sind. Diese vollautomatische Entscheidungsfindung ist nicht unkritisch zu sehen: Der Entscheidungsträger hat keine Möglichkeiten mehr, eine andere Alternative zu wählen, da er deren Auswirkungen dem Modell nicht entnehmen kann.

Im Gegensatz zu der zuvor im Falle 4 behandelten „Prognose“ einer Entscheidungsempfehlung fällt das Problem der fixierten Präferenzvorstellungen hier weniger ins Gewicht. Denn die Präferenzvorstellungen der Entscheidungsträger sind hier nicht in den Trainingsdaten implizit enthalten, sondern explizit im Data-Mining-Verfahren. Sie können parametrisiert werden, so daß ein schnelles Relearning möglich ist, wenn sich die Präferenzvorstellungen oder Handlungsalternativen ändern.

Da die Anwendung von Data-Mining-Entscheidungsmodellen direkt eine Entscheidung liefert und keine exogene Einflußnahme vorsieht, erübrigt sich an dieser Stelle jede weitere Diskussion. Es sei lediglich noch auf Abschnitt 6.2 verwiesen, der die Anwendung von Data-Mining-Entscheidungsmodellen auf eine reale Problemstellung behandelt.

# 4 Anforderungen an Data-Mining-Verfahren und ihre Erfüllung durch existierende Verfahren

Aus den betriebswirtschaftlichen Überlegungen des vorangegangenen Kapitels und aus den technischen Aussagen des Kapitels 2 lassen sich Anforderungen an Data-Mining-Verfahren ableiten, die erfüllt werden müssen, um das in Abschnitt 1.3 aufgestellte Ziel 2 (*„Lösungsverfahren entwickeln“*) zu erreichen. Die folgenden vier Abschnitte definieren diese Anforderungen – nach den Komponenten von Data-Mining-Verfahren gegliedert. Danach gibt Abschnitt 4.5 einen Überblick über die Erfüllung der aufgestellten Anforderungen durch existierende Data-Mining-Verfahren. Abschließend zieht Abschnitt 4.6 aus den vorangegangenen Ausführungen einige Schlußfolgerungen für die Entwicklung eines neuen Data-Mining-Verfahrens.

## 4.1 Anforderungen an den Modelltyp

Zunächst werden einige wichtige und weit verbreitete Anforderungen an den Typ von Data-Mining-Modellen behandelt, die bereits in Abschnitt 2.2.2.1 motiviert wurden und daher keiner weiteren Erläuterung bedürfen:

- ⇨ Die formale Sprache sollte **für den Benutzer leicht verständlich** sein, um die Interpretation der Modelle zu erleichtern.[345]
- ⇨ Es sollen **Variablen unterschiedlicher Skalenniveaus** aufgenommen werden können, damit keine Skalentransformationen vorgenommen werden müssen.
- ⇨ Jeder denkbare Zusammenhang im Realsystem sollte approximierbar sein. Etwas schwächer formuliert sollte die verwendete Sprache den Suchraum nicht so einengen, daß bestimmte Zusammenhänge ausgegrenzt werden.[346] Insbesondere soll der Modelltyp so mächtig sein, daß er auch potentiell „unerwartete“ Zusammenhänge umfaßt.[347] Auf der anderen Seite sollte der Modelltyp auch nicht so viele Variablen enthalten, daß der Suchraum zu groß wird. Diese konfliktären Zielsetzungen

[345] Vgl. FAYYAD/PIATETSKY-SHAPIRO/SMYTH (1996), S. 16.

[346] Vgl. FAYYAD/PIATETSKY-SHAPIRO/SMYTH (1996), S. 16.

[347] Vgl. DOMINGOS (1998), S. 41.

lassen sich zu der Anforderung zusammenfassen, daß eine **Sprache von angemessener Komplexität** benötigt wird.

⇨ Um eine kontrollierte Suche zu ermögliche, sollte der Modelltyp **leicht in eine Suchstrategie integrierbar** sein. Dies ist dann der Fall, wenn sich wenige sinnvolle Modifikationsmöglichkeiten einer Modellinstanz anbieten, z.B. ausschließlich durch Generalisierung- und Spezialisierungsoperatoren, wie sie in Definition 39 bis Definition 44 vorgestellt wurden.[348]

Neben diesen bereits in Abschnitt 2.2.2.1 motivierten Anforderungen sollte für jede Modellinstanz in der verwendeten Sprache **genau eine Möglichkeit existieren, diese Modellinstanz zu repräsentieren**. Könnte ein- und dieselbe logische Aussage durch mehrere sprachliche Konstrukte abgebildet werden, so würde der Suchraum unnötig aufgebläht und die Suche eventuell verzerrt.

*Dies passiert leicht bei zu mächtigen Sprachen, die beispielsweise folgende logisch identische Aussagen zulassen:*

```
WENN Beruf = Beamter ODER Beruf = Selbständiger
DANN  Kreditlimit ≥ 100.000.
```

```
WENN Beruf ∈ {Beamter, Selbständiger}
DANN  Kreditlimit ∈ [100.000; ∞).
```

Weiterhin gilt es zu berücksichtigen, daß für Analysen auch sog. „multirelationale Muster" oder „$1{:}N$-Beziehungen“ zwischen verschiedenen Objekttypen der Realsphäre relevant sein können.

*Beispielsweise kauft ein Kunde $N$ Produkte, ein Produkt wird $N$ mal nachgefragt, ein Unternehmen kontaktiert $N$ Zielkunden mit einer Marketingmaßnahme, und ein Versicherungsnehmer meldet $N$ Schadensfälle.*

Um solche $1{:}N$-Beziehungen durch herkömmliche Modelltypen, wie etwa Entscheidungsbäume oder neuronale Netze, repräsentieren zu können, baut man die Trainingsmenge so auf, daß jede Zeile ein Objekt des zu untersuchenden Typs repräsentiert (z.B. je einen Kassenbon) und jede Spalte die relevanten Objekte, mit denen die zu untersuchenden Objekte in Beziehung stehen können (z.B. jeden für die Untersuchung relevanten Artikel).

*Beispielsweise würde eine Warenkorbanalyse eine Trainingstabelle erfordern, die wie Tabelle 4-1 strukturiert ist.*

---

[348] Selbst bei diesen einfachen Operatoren mußte bereits zwischen nominalen und nichtnominalen Klauseln unterschieden werden.

| ***Warenkorb*** | | | | | | |
|---|---|---|---|---|---|---|
| ***Bon_id*** | ***enthält Milch*** | ***enthält Cola*** | ***enthält Chips*** | ***enthält Fanta*** | ***enthält Bier*** | ***enthält Kaffee*** |
| *1* | *falsch* | *Wahr* | *falsch* | *wahr* | *wahr* | *wahr* |
| *2* | *falsch* | *Falsch* | *wahr* | *wahr* | *falsch* | *falsch* |
| *...* | *...* | *...* | *...* | *...* | *...* | *...* |

***Tabelle 4-1: Trainingstabelle für eine Warenkorbanalyse***

Mit einem derartigen Aufbau der Trainingstabelle hat man die Interpretation der $1{:}N$-Beziehung zwischen einem Kassenbon und jedem der $N$ Artikel festgelegt. So kann diese Beziehung nur als „Kassenbon enthält mindestens einmal Artikel X“ gelesen werden. Denkbar sind – wenn man die Trainingsmenge anders aufbauen würde – auch andere Interpretationen. Insgesamt erscheinen folgende Interpretationen einer Beziehung zwischen einem Objekt aus einer Objektmenge, $O^1$, und mehreren Objekten aus einer anderen Objektmenge, $O^N$, sinvoll:

- ⇨ **„Mindestens 1“-Interpretation:** Ein Objekt aus $O^1$ bezieht sich auf *mindestens eines* der Objekte aus $O^N$ (z.B. „Kassenbon enthält Bier“).
- ⇨ **„Alle“-Interpretation:** Ein Objekt aus $O^1$ bezieht sich auf *alle* Objekte aus $O^N$ (z.B. „Kassenbon enthält ausschließlich Elektroartikel“).
- ⇨ **„Überwiegend“-Interpretation:** Ein Objekt aus $O^1$ bezieht sich auf *den überwiegenden Anteil* der Objekte aus $O^N$ (z.B. „Kassenbon enthält überwiegend Niedrigpreisartikel“).
- ⇨ **„Genau 1“-Interpretation:** Ein Objekt aus $O^1$ bezieht sich auf *genau eines* der Objekte aus $O^N$ (z.B. „Kassenbon enthält genau einmal Bier“).
- ⇨ **„Nicht“-Interpretation:** Ein Objekt aus $O^1$ ist *nicht* mit Objekten aus $O^N$ *assoziiert* (z.B. „Kunde hat noch nie reklamiert“).[349]

Um Beziehungen der genannten Typen automatisch generieren zu können, muß die Trainingsmenge jeweils unterschiedlich aufgebaut werden, da unterschiedliche Variablen generiert werden müssen (z.B. „enthält mindestens einmal Elektroartikel“ versus

[349] Diese Beziehungsinterpretation macht nur dann Sinn, wenn die Integritätsregeln des Datenmodells nicht verlangen, daß mit einem Objekt aus $O^1$ mindestens ein Objekt aus $O^N$ assoziiert ist, wie dies für die Datenobjekttypen „*Kassenbon*“ und „*Artikel*“ der Fall wäre.

„enthält überwiegend Elektroartikel“). Da offensichtlich aus einem in der Datenbank vorliegenden Attribut, wie z.B. „Artikelbezeichnung“, je nach Beziehungsinterpretation unterschiedliche Analysevariablen, wie z.B. „enthält mindestens einen Elektroartikel“ und „enthält überwiegend Elektroartikel“, erzeugt werden können, soll die Beziehungsinterpretation in der formalen Repräsentationssprache von dem originären Attribut getrennt werden. D.h., das Attribut aus der Datenbank und die Beziehungsinterpretation werden in dem Modelltyp eigens codiert, wie z.B.:

*WENN für ein Objekt vom Typ „Kunde“ gilt:*
*es referenziert ausschließlich:*
$Reklamation.Datum < 31.12.2001$,
*UND es referenziert mindestens 1:*
$Bestellung.Datum > 31.12.2001$,
*DANN gilt für dieses Objekt auch:*
$Kunde.Zufriedenheit = hoch$.

*Dieses Datenmuster besagt, daß ein Kunde als hoch zufrieden eingestuft wird, wenn seine letzte Reklamation vor dem 31.12.2001 stattfand und er danach noch mindestens eine Bestellung aufgegeben hat.*

Als Anforderung an den Modelltyp kann somit festgehalten werden, daß er die ***1:N*-Beziehungen zwischen relevanten Objekttypen in einer geeigneten Sprache repräsentieren** können muß.

## 4.2 Anforderungen an das Suchverfahren

Die folgenden Anforderungen an das Suchverfahren wurden bereits in Abschnitt 2.2.3 motiviert und können dementsprechend kurz abgehandelt werden:

- ⇨ Suchverfahren sollten **kurzfristige Verschlechterungen der Lösungen zulassen**, um den Einflußbereich lokaler Optima zu verlassen und in neue Bereiche des Suchraums vorzustoßen.
- ⇨ Suchverfahren sollten **Zyklen in der Lösungssuche vermeiden**.
- ⇨ Suchverfahren sollten Möglichkeiten zur **Intensivierung und Diversifizierung** der Suche bereitstellen.

Hinzuzufügen sind noch die folgenden Anforderungen:

- ⇨ Die **Laufzeit** des Data-Mining-Verfahrens sollte nicht zu hoch sein, um die Akzeptanz des Verfahrens bei den Benutzern nicht zu gefährden. Die wichtigsten Einflußfaktoren auf die Laufzeit sind: die Effizienz des Datenzugriffs, die Größe des

Suchraums und die Navigation durch den Suchraum. Die Forderung nach einer geringen Laufzeit schließt insbesondere Data-Mining-Verfahren aus, die den Suchraum vollständig durchsuchen.

Zwar wird die Datenbasis i.d.R. über einen längeren Zeitraum zusammengetragen, so daß im Vergleich dazu das Data Mining einen relativ geringen Zeitbedarf einnimmt. Auf der anderen Seite erkennt man an Abbildung 2-2, daß die Data-Mining-Phase innerhalb des KDD-Prozesses u.U. sehr häufig durchlaufen werden muß. So kann sich der KDD-Prozeß über mehrere Arbeitstage oder -wochen hinziehen, bis verwertbare Ergebnisse vorliegen.

⇨ Der Algorithmendesigner sollte **„intelligente" Suchoperatoren** in das Suchverfahren implementieren. Erfahrungen aus der Entwicklung von Data-Mining-Verfahren zeigen, daß die bloße Übernahme von allgemeinverwendbaren Meta-Strategien, wie z.B. Tabu Search oder genetischen Algorithmen, aufgrund der komplexen Suchräume, mit denen man im Data Mining konfrontiert wird, nicht ausreicht. Suchoperatoren sollen hier als „intelligent" bezeichnet werden, wenn sie speziell auf die Generierung eines Data-Mining-Modells zugeschnitten sind und die Auswirkungen von Zügen a-priori abschätzen können.

*Beispielsweise könnte ein intelligenter Suchoperator für eine Regel, deren Prämisse eine geringe Allgemeinheit besitzt, eine Generalisierung durchführen und für eine Regel, deren Korrektheit gering ist, eine Spezialisierung.*

⇨ Aus Laufzeitgründen wird für das Data Mining eine **enge Kopplung zwischen Suchverfahren und Datenzugriff** gefordert.[350] Vor allem sollte das Suchverfahren möglichst selten Zugriffe auf externe Speicher anfordern.

⇨ Viele lokale Suchstrategien funktionieren bei großen Lösungsänderungen nicht und verlangen daher kleine Lösungsänderungen.[351] Dies liegt zumeist an den verwendeten Akzeptanzkriterien, die auf kleine Lösungs- und Zielfunktionsänderungen zugeschnitten sind. Mit großen Änderungen der Lösung gehen i.d.R. auch große Änderungen der Zielfunktionswerte einher. Große Lösungsänderungen sind zum einen erforderlich, um den Einflußbereich lokaler Optima zu verlassen. Zum anderen kann

---

[350] Vgl. IMIELINSKI/MANNILA (1996), S. 59.

[351] Vgl. beispielsweise zu Simulated Annealing, Threshold Accepting, Great Deluge und Record-to-Record Travel: DUECK (1993), S. 86 f.

der Lösungsraum mit großen Schritten schneller durchforstet werden als mit kleinen Schritten. Letztendlich sollte das **Suchverfahren mit großen Änderungen der Zielfunktionswerte umgehen können**.

## 4.3 Anforderungen an die Bewertung der Modelle

Die wichtigste Anforderung an die Bewertung von Data-Mining-Modellen besteht darin, daß die Bewertung an eine betriebswirtschaftlich fundierte Aufgabenstellung angepaßt sein muß. Dies bedeutet zum einen, daß in der Bewertungsfunktion der **Entscheidungsbezug** – also der Grad der Entscheidungsunterstützung – des bewerteten Modells zum Ausdruck kommen muß. Zum anderen bedeutet die Anpassung an eine betriebswirtschaftlich fundierte Aufgabenstellung, daß nicht einfach eine Sammlung von vielen Interessantheitsfacetten in das Verfahren implementiert werden darf, zwischen denen der Benutzer dann zu wählen hat oder eine große Anzahl von Zielgewichtungen einstellen muß. Vielmehr muß eine **Klassifizierung betriebswirtschaftlicher Aufgabenstellungen** implementiert werden, denen die jeweils sinnvollen Bewertungen zugeordnet sind. Der Benutzer muß dann nur noch den Aufgabentyp wählen, anstatt erst noch die benötigten Interessantheitsfacetten zusammenzustellen.

Eine grobe Klassifizierung der relevanten Aufgabenstellungen ist in dieser Arbeit bereits erfolgt. So wurden in Kapitel 3 Data-Mining-Entscheidungs-, -Prognose-, -Erklärungs- und -Beschreibungsmodelle differenziert. Und auch die jeweils relevanten Interessantheitsfacetten wurden dort bereits herausgearbeitet und in Abbildung 3-5 zusammengefaßt. Als Anforderung an die Bewertung von Data-Mining-Modellen kann damit festgehalten werden, daß die **Bewertung der Modelltypen gemäß Abbildung 3-5** zu erfolgen hat. Dazu gehört insbesondere die bereits in Abschnitt 2.1.3 erkannte Notwendigkeit, **das Modell als Ganzes zu bewerten**. Eine ebenso wichtige Forderung, die bereits diskutiert wurde, ist die **Übertragbarkeit des Modells auf neue Daten**.

Weiterhin müßte ein geschlossenes Bewertungskonzept alle dargestellten Interessantheitsfacetten nicht nur verbal umschreiben, sondern **operational definieren** und für verschiedene Problemstellungen empfehlen, welche als **zu optimierende Größe** gewählt werden soll und **welche Interessantheitsfacetten zu satisfizieren sind**.

Eine weitere Anforderung an das Bewertungskonzept ergibt sich dadurch, daß $1{:}N$-Beziehungen zwischen einem Bezugsobjekt und den Objekten eines anderen Typs zugelassen werden. Dadurch kann jedes Bezugsobjekt durch mehrere Regeln erfaßt werden.

*Beispielsweise werden bei Warenkorbanalysen regelmäßig Bezugsobjekte, also Warenkörbe, durch mehrere Regeln erfaßt, wie z.B. ein Warenkorb mit Bier und Cola durch die folgenden Regeln:*

```
WENN der Warenkorb Cola enthält DANN ...
WENN der Warenkorb Bier enthält DANN ...
```

Eine solche Mehrfacherfassung derselben Bezugsobjekte darf nicht additiv in die Bewertung einfließen, da sonst ein Modell umso besser abschneiden würde, je mehr Regeln es umfaßt. Attribute, die in einer $N{:}1$-Beziehung zu den Bezugsobjekten stehen, seien als „***Mehrfach-Attribute***“ bezeichnet, und Attribute, die in einer $1{:}1$-Beziehung zu den Bezugsobjekten stehen, als „***Einfach-Attribute***“. Damit kann man die Anforderung formulieren, daß das Bewertungskonzept **für Mehrfach- und Einfach-Attribute gleichermaßen geeignet** sein sollte.

## 4.4 Anforderungen an den Datenzugriff

Die Motivation des Data Mining besteht nach Abschnitt 1.1 darin, neuartige Datenmuster aufzuspüren. Bereits in Abschnitt 2.2.4.5 wurde darauf hingewiesen, daß eine notwendige Voraussetzung für die Erreichung dieses Ziels darin besteht, daß man einen sehr großen Raum von Aussagentypen definiert, der potentiell viele unbekannte Aussageninstanzen enthält. Eine Möglichkeit, einen großen Suchraum zu definieren, besteht darin, eine große Anzahl von Variablen vorzusehen, zwischen denen nach Zusammenhängen gesucht werden soll. Dies kann ein Data-Mining-Verfahren dadurch realisieren, daß es selbständig und wiederholt Attribute aus der Datenbank auswählt, anstatt die Attributauswahl dem Benutzer zu überlassen.

Da im Data Mining statistisch haltbare Aussagen ermittelt werden sollen, muß zu den Ausprägungen aller interessierenden Variablen eine genügend große Anzahl von Objekten vorliegen. Zu einer großen Anzahl von Objekten und einer großen Anzahl von Variablen mit großen Domänen läßt sich jedoch i.d.R. keine Trainingsmenge aufstellen, da diese viel zu groß für den Hauptspeicher werden würde. Auch die externe Speicherung stellt keine praktikable Lösung dar, da externe Speicher zu langsam sind, als daß

das Data-Mining-Verfahren permanent auf sie zugreifen könnte. Die Anzahl der Variablen würde die Breite der Trainingstabelle bestimmen und die Anzahl der Objekte die Länge der Trainingstabelle. Ein weiterer Umstand, der die Trainingstabelle sehr groß werden läßt, besteht darin, daß in einer solchen Trainingstabelle die Trainingsdaten hochgradig redundant abgelegt würden.

*Diese Redundanz läßt sich bereits an einem kleinen Auszug aus einer Trainingstabelle verdeutlichen:*
*In Tabelle 4-2 enthalten die ersten vier Spalten Merkmale des Objekttyps „Kassenbon“ und die übrigen Spalten Merkmale des Objekttyps „Bonposition“. Eine Bonposition läßt sich jeweils genau einem Kassenbon zuordnen. Umgekehrt kann ein Kassenbon durchaus mehrere Bonpositionen umfassen. Ein solcher Beziehungstyp zwischen den Objekttypen „Kassenbon“ und „Bonposition“ wird auch als „$1{:}N$-Beziehungstyp“ bezeichnet. Die Redundanz besteht nun darin, daß die Informationen, daß der Kassenbon mit der Bon_id 1 an einem Montag um 16:25 Uhr an Kasse 3 ausgestellt wurde und der Kassenbon mit der Bon_id 2 an einem Mittwoch um 13:12 Uhr an Kasse 1, nicht nur einmal abgespeichert wird, sondern einmal pro Bonposition.*

| ***Bon_id*** | ***Wochentag*** | ***Uhrzeit*** | ***Kasse*** | ***Bonpos_id*** | ***Artikel_id*** | ***Menge*** |
|---|---|---|---|---|---|---|
| *1* | *Montag* | *16:25* | *3* | *1* | *A008* | *1* |
| *1* | *Montag* | *16:25* | *3* | *2* | *A010* | *1* |
| *1* | *Montag* | *16:25* | *3* | *3* | *A120* | *1* |
| *2* | *Mittwoch* | *13:12* | *1* | *1* | *A220* | *2* |
| *2* | *Mittwoch* | *13:12* | *1* | *2* | *A008* | *1* |
| *2* | *Mittwoch* | *13:12* | *1* | *3* | *A001* | *1* |
| *2* | *Mittwoch* | *13:12* | *1* | *4* | *A800* | *1* |

***Tabelle 4-2: Redundante Datenhaltung in der Trainingsmenge***

Neben der Datenredundanz gibt es im Zusammenhang mit $1{:}N$-Beziehungen noch einen anderen Faktor, der eine Trainingsmenge unnötig aufbläht: die Speicherung von Eigenschaften, welche die Objekte *nicht* aufweisen. Der Grund, warum solche nicht vorhandenen Eigenschaften überhaupt gespeichert werden, liegt darin, daß die meisten Data-Mining-Verfahren verlangen, daß die Trainingsmenge aus genau einer Tabelle besteht, in der jede Zeile ein eigenes Objekt repräsentiert. Die Objekte des Typs mit der Kardinalität $N$ werden dann nicht zeilenweise in einer eigenen Tabelle gespeichert, sondern spaltenweise in derselben Tabelle wie die Bezugsobjekte. Folgendes Beispiel soll dies verdeutlichen.

*Gegeben sei eine Datenbank mit den Tabellen „Kassenbon“, „Bonposition“ und „Artikel“ und den in den folgenden Tabellen abgebildeten Datenobjekten.*

| Kassenbon | |
|---|---|
| **Bon_id** | **Wochentag** |
| 1 | Montag |
| 2 | Samstag |

| Bonposition | | |
|---|---|---|
| **Bon_id** | **Bonpos_id** | **Artikel_id** |
| 1 | 1 | 3 |
| 1 | 2 | 5 |
| 1 | 3 | 4 |
| 1 | 4 | 1 |
| 2 | 1 | 2 |
| 2 | 2 | 3 |

| Artikel | |
|---|---|
| **Artikel_id** | **Bezeichnung** |
| 1 | Cola |
| 2 | Chips |
| 3 | Fanta |
| 4 | Bier |
| 5 | Kaffee |

***Tabelle 4-3: Eine Beispiel-Datenbank***

*Es sollen nun Aussagen über typische Warenkörbe getroffen werden, z.B.:*

`WENN ein Warenkorb an einem Montag eingekauft wurde und Cola enthält, DANN enthält er auch Bier.`

*Einem Warenkorb entspricht in der Datenbank ein Kassenbon. Damit stellt die Tabelle „Kassenbon" die sog. „**Bezugsrelation**" dar, d.h. über ihre Objekte sollen Aussagen induziert werden. Zur Konstruktion einer geeigneten Trainingstabelle geht man wie folgt vor:*

*Die neue Relation „Warenkorb" erhält denselben Primärschlüssel wie die Bezugsrelation „Kassenbon", also die „Bon_id". Alle Attribute, die in einer 1:1-Beziehung zur „Bon_id" stehen, können ohne Transformation in die Relation „Warenkorb" übernommen werden. Dies trifft im Beispiel auf das Attribut „Wochentag" zu. 1:N-Beziehungen zur „Bon_id" werden als „`enthält mindestens 1`" interpretiert. Dies trifft auf die Beziehungen zu den Attributen „Bonpos_id", „Artikel_id" und „Bezeichnung" zu. Von diesen Attributen wird die Bezeichnung der Artikel mit jedem möglichen Wert und der Interpretation „`enthält mindestens 1`" als eigenes Attribut in die Trainingstabelle „Warenkorb" eingefügt. Das Ergebnis ist in Tabelle 4-4 aufgeführt.*

| Warenkorb | | | | | | |
|---|---|---|---|---|---|---|
| **Bon_id** | **Wochentag** | **enthält Cola** | **enthält Chips** | **enthält Fanta** | **enthält Bier** | **enthält Kaffee** |
| 1 | Montag | wahr | falsch | wahr | wahr | wahr |
| 2 | Samstag | falsch | wahr | wahr | falsch | falsch |

***Tabelle 4-4: Eine Trainingstabelle mit Eigenschaften, welche die Warenkörbe nicht aufweisen***

Die dargestellte Transformation hat für das Data-Mining-Verfahren den Vorteil, daß die Ergebnistabelle genau eine Zeile pro Bezugsobjekt besitzt, so daß nur noch nach Abhängigkeiten zwischen den Attributen einer Zeile (sog. „***inter-field patterns***"[352]) und

[352] FRAWLEY/PIATETSKY-SHAPIRO/MATHEUS (1991), S. 12

nicht nach Abhängigkeiten zwischen verschieden Zeilen (sog. „***inter-record patterns***“[353]) gesucht werden muß. Der Nachteil an dieser Art der Repräsentation liegt darin, daß die Tabelle eventuell so groß wird, daß sie nicht mehr handhabbar ist.

*Als Konsequenz würde man im Beispiel die Anzahl der Artikel einschränken oder nur noch nach Zusammenhängen zwischen Warengruppen suchen. Dann könnten keine Assoziationen zwischen den weggelassenen Artikeln oder zwischen Artikeln und Warengruppen extrahiert werden. Diese im voraus durchzuführende Einschränkung von potentiell relevanten Größen und potentiell extrahierbaren Aussagen will man im Data Mining jedoch minimieren.*

Zusammengefaßt führen sowohl eine redundante Datenhaltung als auch eine Speicherung nicht vorhandener Eigenschaften bei einer großen Anzahl von Variablen und Objekten zu einer extrem großen und unhandlichen Trainingstabelle. Um dies zu vermeiden, ohne den Benutzer zu zwingen, sich auf eine kleine Anzahl von Variablen und Aussagen zu beschränken und damit zu riskieren, daß die generierten Aussagen bereits bekannt sind, werden die folgenden Anforderungen an den Datenzugriff gestellt:

- ⇨ Die **Trainingsmenge** soll nicht manuell vordefiniert, sondern durch das Data-Mining-Verfahren **dynamisch aus mehreren Tabellen zusammengestellt** werden.
- ⇨ Dies hat zur Konsequenz, daß das Data-Mining-Verfahren **direkt auf die Datenbank zugreifen** muß.
- ⇨ Der **Datenzugriff sollte effizient sein**, da er – gerade bei einem permanenten Zugriff auf die Datenbank – einen entscheidenden Einflußfaktor auf die Laufzeit und damit auf die Akzeptanz des Data-Mining-Verfahrens darstellt.
- ⇨ Der **Datenzugriff** sollte **entsprechend der Interpretation von *1:N*-Beziehungen** als „`Genau 1`“, „`Mindestens 1`“, „`Alle`“, „`Überwiegend`“ oder „`Nicht`“ erfolgen. Diese Anforderung betrifft die zum Zugriff auf die Datenbank benötigten SELECT-Befehle.

Der Zugriff auf die Unternehmensdaten kann nicht direkt auf der operativen Datenbank erfolgen, die die Transaktionen des Tagesgeschäftes abbildet, da die in schneller Folge abgesendeten Lesezugriffe des Data-Mining-Verfahrens die Datenbank zu lange blockieren würden. Aus diesem Grunde ist ein analysezweckbezogener Auszug aus den

[353] FRAWLEY/PIATETSKY-SHAPIRO/MATHEUS (1991), S. 12

operativen Daten erforderlich, wie er im Zusammenhang mit sog. *Data-Warehouse-Architekturen*[354] diskutiert wird.

Wichtige Größen für die Aufstellung betriebswirtschaftlicher Modelle liegen aufgrund der redundanzarmen Speicherung nicht in relationalen Datenbeständen vor, sondern müssen aus den Daten berechnet werden, wie z.B. die Kassenbonsumme, die Laufzeit von Versicherungsverträgen oder die Anzahl Reklamationen eines Kunden. Als Berechnungen kommen vor allem Aggregationen wie COUNT(*) für diskrete Attribute und SUM bzw. MEAN für stetige Attribute in Frage. Hier liegen zwei Ansatzpunkte nahe, solche Berechnungsfelder mit in die Analyse einzubeziehen:

- ⇨ Man könnte das Data-Mining-Verfahren so konzipieren, daß es selbständig Aggregationen über $1{:}N$-Beziehungen bildet, wie z.B. die Summe der einzelnen Kassenbonpositionen. Diese würde aber eine enorme Vergrößerung des Suchraums bewirken, da potentiell jedes Feld im Hinblick auf jeden Objekttyp, mit dem es über eine $1{:}N$-Beziehung verbunden ist, aggregiert werden kann.
- ⇨ Eine einfachere Lösung bestünde darin, manuell vor dem eigentlichen Data Mining die relevanten Berechnungsfelder in der Datenbank als virtuelle Felder zu definieren. Damit stehen sie dem Data-Mining-Verfahren genauso zur Verfügung, als wären die Feldwerte physisch in der Datenbank gespeichert. Hierunter würde allerdings die Performance leiden, da potentiell bei jedem Lesezugriff auf das Feld Neuberechnungen stattfinden müssen. Dieses Problem läßt sich durch die redundante Speicherung der Berechnungsfelder lösen, wie sie in dem erwähnten Data-Warehouse-Konzept realisiert werden.

Obwohl die zweitgenannte, manuelle Vorgehensweise der ursprünglichen Idee des Data Mining widerspricht, möglichst autonom herauszufinden, welche Felder relevant sind, wird sie hier weiter verfolgt, um den Suchraum nicht noch weiter zu vergrößern. Damit ergeben sich keine weiteren Anforderungen an das Data-Mining-Verfahren.

---

[354] Vgl. TILLMANNS (2000), S. 672.

## 4.5 Betrachtung existierender Verfahrenskomponenten bezüglich der Anforderungen

Im folgenden sollen die vorhandenen Forschungsergebnisse im Data Mining den aufgeführten Anforderungen gegenübergestellt werden. Damit soll beurteilt werden, welche Forschungsergebnisse Einzug in das zu entwickelnde Data-Mining-Verfahren erhalten können und welche Teilkomponenten neu konzipiert werden müssen. Tabelle 4-5 faßt die definierten Anforderungen an Data-Mining-Verfahren zusammen.

| Modelltyp | Suchverfahren | Bewertung | Datenzugriff |
|---|---|---|---|
| ⇨ Verständlichkeit für den Benutzer<br>⇨ Variablen unterschiedlicher Skalenniveaus<br>⇨ Sprache von angemessener Komplexität<br>⇨ Leichte Integrierbarkeit in eine Suchstrategie<br>⇨ Eindeutige Repräsentation von Modellen<br>⇨ Repräsentation von *1:N*-Beziehungen mit „`Genau 1`“-, „`Mindestens 1`“-, „`Alle`“-, „`Überwiegend`“- und „`Nicht`“-Interpretation | ⇨ Zulassen kurzfristiger Verschlechterungen<br>⇨ Vermeidung von Zyklen<br>⇨ Intensivierung der Suche möglich<br>⇨ Diversifizierung der Suche möglich<br>⇨ Geringe Laufzeit<br>⇨ „Intelligente“ Suchoperatoren<br>⇨ Enge Kopplung mit dem Datenzugriff<br>⇨ Umgang mit grossen Änderungen der Zielfunktionswerte | ⇨ Entscheidungsbezug<br>⇨ Klassifizierung betriebswirtschaftlicher Aufgabenstellungen<br>⇨ Bewertung der Modelltypen gemäß Abbildung 3-5, insbesondere:<br>⇨ Bewertung eines Modells als Ganzes<br>⇨ Gewährleistung der Übertragbarkeit auf neue Daten<br>⇨ Operationale Definitionen<br>⇨ Deklaration von Optimierungs- und Satisfizierungszielen<br>⇨ Bewertung für Mehrfach- und Einfach-Attribute gleichermaßen geeignetet | ⇨ Dynamische Zusammenstellung der Trainingsmenge aus mehreren Tabellen<br>⇨ Direkter Zugriff auf die Datenbank<br>⇨ Effizienter Datenzugriff<br>⇨ Datenzugriff entsprechend der Interpretation von *1:N*-Beziehungen als „`Genau 1`“, „`Mindestens 1`“, „`Alle`“, „`Überwiegend`“ oder „`Nicht`“ |

**Tabelle 4-5: Überblick über die Anforderungen an Data-Mining-Verfahren**

Abschnitt 4.5.1 stellt die existierenden Modelltypen den definierten Anforderungen gegenüber. Abschnitt 4.5.2 geht analog für die existierenden Suchverfahren vor, Abschnitt 4.5.3 für existierende Interessantheitskonzepte und Abschnitt 4.5.4 für existierende Datenzugriffskomponenten.

### 4.5.1 Betrachtung existierender Modelltypen und darauf aufbauender Data-Mining-Verfahren

Das maschinelle Lernen, das den Fortschritt der Data-Mining-Verfahren am stärksten vorangetrieben hat, fokussiert die Entwicklung von Heuristiken, die speziell auf die

Konstruktion von Modellen eines ganz bestimmten Typs ausgerichtet sind. Es hat eine Vielzahl von Repräsentationsformen für Modellwissen hervorgebracht, die im Rahmen der automatischen Wissensakquisition eingesetzt werden können.[355]

Außer der Prädikatenlogik eignet sich keine andere verbreitete Repräsentationsform unverändert zur Repräsentation von $1{:}N$-Beziehungen. Verfahren auf Basis der Prädikatenlogik wurden in großer Anzahl in dem Forschungsbereich der „*induktiven logischen Programmierung*“ entwickelt. Kaum eines dieser Verfahren bietet jedoch einen direkten Zugriff auf die Datenbank, wie er oben aus Gründen der Speichereffizienz gefordert wurde.[356] Gegen die Verwendung der Prädikatenlogik spricht, daß ihre Ausdrucksfähigkeit weit über das hier geforderte Niveau hinausgeht, d.h. sie ist unnötig komplex, schwer verständlich und der Suchraum unnötig groß.

Ein Ansatz, der bei dem Data-Mining-Verfahren RDT/DB[357] verfolgt wird, um das Problem der Suchraumgröße zu lösen, besteht darin, dem Benutzer die Aufgabe und die Bürde zu übertragen, den Suchraum einzuschränken. Er muß zu diesem Zweck Schablonen für die zu erzeugenden Regeln vorgeben, die dann durch das Data-Mining-Verfahren instanziiert werden. Durch die explizite Vorgabe von Meta-Wissen in Form von Regelschablonen kann – gerade bei einer so mächtigen Sprache wie der Prädikatenlogik – der Benutzer leicht überfordert werden: Wird die Sprache zu stark eingeschränkt, so grenzt man von vornherein interessante Muster aus dem Suchraum aus. Bleibt die Sprache zu mächtig, überfordert man das Suchverfahren.

Andere Autoren schlagen daher vor, den Benutzer die Ergebnisse der Data-Mining-Phase analysieren zu lassen, eine veränderte Einschränkung des Modelltyps vorzunehmen und diese Schleife erneut zu durchlaufen.[358] Dies wiederum widerspricht den eingangs gesetzten Zielen, möglichst autonome Verfahren zu entwickeln, um die Gesamtzeit der Datenanalyse in Grenzen zu halten. UTGOFF hat sich mit Versuchen

---

[355] Vgl. zu Beispielen und Literaturreferenzen von Wissensrepräsentationsformen Abschnitt 2.2.2.

[356] Vgl. KNOBBE/SIEBES/VAN DER WALLEN (1999), S. 380.

[357] Auf das Data-Mining-Verfahren RDT/DB wird hier mehrfach referenziert, da es dasjenige aus der Literatur bekannte Verfahren ist, das – gemessen an den definierten Anforderungen an die Repräsentation multirelationaler Datenmuster – am besten zu beurteilen ist. Vgl. zu RDT/DB: MORIK/BROCKHAUSEN (1997), S. 292 ff.

[358] Vgl. SHEN/LENG (1996), S. 901.

beschäftigt, die Anpassung des Suchraums automatisch durchführen zu lassen.[359] Er verfolgt das Ziel, den geeigneten Modelltyp automatisch zu erlernen, um ihn für spätere Data-Mining-Läufe zu verwenden.[360] Dieses „Meta"-Lernen empfielt sich allerdings nur für sehr ähnliche Data-Mining-Aufgaben, so daß es hier nicht weiter verfolgt wird.

Durch die Verwendung komplexer Sprachen werden die benötigten Suchalgorithmen sehr anspruchsvoll. Dieses Problem wird bei MORIK und BROCKHAUSEN nicht gelöst – so umfaßt RDT/DB ein mehrphasiges Suchverfahren, das kein Backtracking zwischen den Phasen zuläßt.[361] Beispielsweise werden für kardinale Attribute zuerst die zu verwendenden Intervallgrenzen gelernt und in einer nachgelagerten Phase die Klauseln, die diese Intervallgrenzen verwenden. Erst in dieser Phase stellt sich heraus, ob die erlernten Intervallgrenzen gut gewählt wurden – ein Zurück gibt es nicht mehr. Außerdem führt RDT/DB bis auf einen einfachen Pruning-Mechanismus[362] eine vollständige Exploration des Suchraums durch.[363] Dies muß bei größeren Problemen zu unakzeptablen Laufzeiten führen.

Noch schwerer wiegt der Nachteil, daß in der referenzierten Literatur im wesentlichen einzelne Regeln, aber nicht eine Regelmenge als Ganzes bewertet wird. Die damit verbundenen Probleme wurden bereits in Abschnitt 2.1.3 diskutiert.

Einen ähnlichen Lösungsansatz für multirelationale Datenmuster unter Verwendung prädikatenlogischer Ausdrücke präsentiert WROBEL für sein Data-Mining-Verfahren „MIDOS". Er beschränkt allerdings die Lernaufgabe auf das Auffinden der $k$ besten Beschreibungen von Teilmengen einer vorgegebenen Objektmenge – wobei $k$ durch den Benutzer vorgegeben werden muß.[364] Ansonsten fällt die Kritik ähnlich aus wie oben für das System RDT/DB: Die erforderlichen Benutzervorgaben sind sehr komplex, die

[359] Vgl. UTGOFF (1986), S. 113 ff.

[360] Vgl. Abschnitt 2.2.3.4 unter dem Stichwort „Erlernen einer vielversprechenden Teilmenge des Suchraums".

[361] Vgl. MORIK/BROCKHAUSEN (1997), S. 292 ff.

[362] Der Pruning-Mechanismus umgeht Bereiche des Suchraums, von denen bekannt ist, daß die Lösungen die Nebenbedingungen einer gewissen Allgemeingültigkeit nicht erfüllen oder daß sie nur noch speziellere als die bereits gefundenen Aussagen enthalten.

[363] Vgl. MORIK/BROCKHAUSEN (1997), S. 292.

[364] Vgl. WROBEL (1997), S. 79, S. 82.

Regelmenge wird nicht als Ganzes bewertet, und das Suchverfahren ist für größere Problemstellungen zu wenig „intelligent“.

Viele weitere Forschungsarbeiten aus dem Gebiet der induktiven logischen Programmierung – wie z.B. die Verfahren ICL[365], TILDE[366], CLAUDIEN[367], S-CART[368], WARMR[369], RIBL2[370], RDBC[371], FORC[372] oder FFOIL[373] – weisen dieselben Nachteile auf, so daß die Vermutung naheliegt, daß aus diesem Gebiet weder ein geeignetes Suchverfahren mit einem direkten Datenzugriff noch ein ökonomisches Bewertungskonzept übernommen werden kann.

Die konjunktive Normalform stellt nach den Ausführungen aus Abschnitt 2.2.2.2 eine geeignete Ausgangsbasis für die Konzeption einer *verständlichen Sprache von angemessener Komplexität* dar, die ohne Codierung *Variablen unterschiedlicher Skalenniveaus* aufweisen kann und *leicht in eine Suchstrategie integrierbar* ist. Die Regeldarstellung in konjunktiver Normalform kann nicht unverändert aus Abschnitt 2.2.2.2 übernommen werden, da in Abschnitt 4.1 einige Anforderungen an die *Repräsentation von 1:N-Beziehungen* gestellt wurden, die diese Wissensrepräsentation nicht erfüllt.

*Beispielsweise geht folgende 1:N-Beziehung über die Beschreibungsfähigkeit von Regeln in KNF hinaus:*

```
WENN für ein Objekt vom Typ „Kunde“ gilt:
     es referenziert kein Objekt vom Typ:
          Reklamation,
UND  es referenziert mindestens 1:
          Bestellung.Datum  > 1.1.2000,
DANN gilt für dieses Objekt auch:
     Kunde.Zufriedenheit = hoch.
```

Zum einen muß der Objekttyp, auf den sich die Klauseln beziehen (im Beispiel: „*Kunde*“), angegeben werden. Zum anderen müssen die Beziehungsinterpretationen für die einzelnen Klauseln in der Beschreibung enthalten sein. Diejenigen Beziehungen mit

---

[365] Vgl. DE RAEDT ET AL. (2001), S. 114 f.

[366] Vgl. DE RAEDT ET AL. (2001), S. 116 f.

[367] Vgl. DE RAEDT ET AL. (2001), S. 117 f.

[368] Vgl. KRAMER/WIDMER (2001), S. 145 ff.

[369] Vgl. DEHASPE/TOIVONEN (2001), S. 201 ff.

[370] Vgl. KIRSTEN/WROBEL/HORVÁTH (2001), S. 220 f.

[371] Vgl. KIRSTEN/WROBEL/HORVÁTH (2001), S. 221 ff.

[372] Vgl. KIRSTEN/WROBEL/HORVÁTH (2001), S. 223 ff.

[373] Vgl. QUINLAN (2001), S. 294 ff.

„Nicht“-Interpretation (im Beispiel: „es referenziert kein Objekt vom Typ *Reklamation*“) erfordern keine Angabe von Attributen und Wertemengen.

Zusammengefaßt muß die Regeldarstellung in konjunktiver Normalform um den Bezugsobjekttyp und um die Interpretation des Beziehungstyps erweitert werden. Und es müssen Klauseln zugelassen werden, die nur aus der Angabe der Interpretation „keines“ und eines Objekttyps bestehen.

Eine Erweiterung der konjunktiven Normalform, die diesen Anforderungen schon recht nahekommt, wird von KNOBBE, SIEBES und VAN DER WALLEN vorgestellt.[374] Bemerkenswert ist dabei die Repräsentation in Form von erweiterten Entscheidungsbäumen, welche eindeutig in SELECT-Anweisungen der Standard-Datenbankschnittstelle SQL übersetzt werden können. Die zur Wissensrepräsentation verwendete Sprache bietet jedoch keine Beziehungen mit „Überwiegend“- und „Alle“-Interpretation. Außerdem sind Entscheidungsbäume in ihrer Approximationsfähigkeit dadurch eingeschränkt, daß sie, wie in Abschnitt 2.2.2.3.2 diskutiert wurde, eine bestimmte Strukturierung der Regelmenge verlangen. Des weiteren sind die zur Generierung von Entscheidungsbäumen eingesetzten Suchverfahren einfache Heuristiken, deren Nachteile im nächsten Abschnitt angesprochen werden.

Eine Möglichkeit, die Komplexität des Modelltyps in Grenzen zu halten, bieten sog. „***attributorientierte Verfahren***“[375]. Diese zeichnen sich dadurch aus, daß eine Lösung einzig aus einer Kombination von erklärenden und/oder zu erklärenden Attributen, $D \subset A$, $C \subseteq A$, $C \cap D = \emptyset$, $C \neq \emptyset$, besteht, aus der eine Regelmenge eindeutig abgeleitet werden kann. Ein Beispiel für eine solche Ableitung stellt die in Definition 2-36 eingeführte Induktion einer Regelmenge nach dem Rough-Set-Ansatz dar.

Änderungen der Attributmenge $C$ oder $D$ wirken sich auf jede einzelne Regel der Regelmenge aus. Da jede Lösung sich nur aus einer Kombination von Attributen zusammensetzt, ist der Suchraum um ein vielfaches kleiner, als wenn jede Regel aus der Regelmenge einzeln variiert werden müßte. Inwieweit auf diese Weise tatsächlich keine relevanten Aussagen ausgegrenzt werden, wird später diskutiert.[376]

---

[374] Vgl. KNOBBE/SIEBES/VAN DER WALLEN (1999), S. 381 f.

[375] Vgl. HAN/FU (1996), S. 401 ff.

[376] Die kritische Diskussion des Modelltyps erfolgt in Abschnitt 5.2.2.

Noch zu klären bleibt, nach welcher Suchstrategie die Attributmengen $C$ und $D$ aus $A$ zusammengestellt werden soll.

### 4.5.2 Betrachtung existierender Suchstrategien

Im maschinellen Lernen wurde eine große Anzahl von Suchstrategien entwickelt. Die meisten davon sind Algorithmen, die – ohne ein Backtracking zu erlauben – auf ein lokales Optimum zusteuern und dieses als beste gefundene Lösung ausgeben. Viele dieser Suchverfahren sind sehr eng mit einer bestimmten Form der Wissensrepräsentation verknüpft und nicht auf andere Repräsentationsformen übertragbar. Neben diesen speziellen existieren auch allgemeinverwendbare Suchverfahren. Die bekanntesten sind:

- ⇨ genetische Algorithmen,
- ⇨ Evolutionsstrategien,
- ⇨ Simulated Annealing und
- ⇨ Tabu Search.

Data-Mining-Verfahren, die auf **genetischen Algorithmen** basieren, erfüllen die meisten aufgestellten Anforderungen an Suchverfahren schon aufgrund ihrer globalen Suche.[377] So *lassen sie Verschlechterungen zu*, haben *keine Probleme mit Zyklen* oder mit *großen Änderungen der Zielfunktionswerte*, *intensivieren die Suche* durch die bevorzugte Selektion guter Lösungen und *diversifizieren die Suche* durch die Mutation, wenn sie so konzipiert ist, daß sie der Implementierung von neuem genetischen Material in die Population dient.

Bereits in Abschnitt 2.2.3.4 wurde Kritik an genetischen Algorithmen geäußert. Über die dort behandelten Kritikpunkte hinaus sind genetische Algorithmen bezüglich der Möglichkeit, *„intelligente" Suchoperatoren* zu implementieren, kritisch zu beurteilen. Die Initialisierung der ersten Population sowie die Suchschritte erfordern im Data Mining aufgrund der Größe des Lösungsraums häufig Spezialwissen aus einem bestimmten

[377] Vgl. beispielsweise ISHIBUCHI/MURATA (1997), S. 261 f. zu genetischen Algorithmen, die eine Regelmenge als ein Individuum modellieren.

Anwendungsbereich.[378] Noch gravierender ist das Problem der *Kopplung mit dem Datenzugriff*. Die Häufigkeit des Zugriffs auf langsame externe Speicher sollte begrenzt werden. Dies ließe sich bei lokalen Suchstrategien relativ einfach verwirklichen, indem die jeweils untersuchte lokale Umgebung der aktuellen Lösung im Hauptspeicher gehalten wird. Bei genetischen Algorithmen ist dies so nicht möglich, da sie bei jeder Populationsbewertung potentiell auf die gesamte Datenbank zugreifen können. Darüber hinaus sind genetische Algorithmen gerade im Zusammenhang mit der attributorientierten Suche bezüglich ihrer *Laufzeit* kritisch zu beurteilen. Als Operatoren für die attributorientierte Suche bieten sich insbesondere das Streichen und Hinzufügen einzelner Attribute an. Vor allem das Streichen von Attributen kann sehr laufzeiteffizient konzipiert werden, wie in Abschnitt 5.4 gezeigt wird. Dadurch bietet es sich an, wenn eine Lösung mit den Variablen $s=\{a_1, ..., a_n\}$ existiert, durch die schnellen Generalisierungsoperationen auch alle Teil-Lösungen, $\forall s' \in Pot(s)$, zu erzeugen und zu bewerten. Dies ist in genetischen Algorithmen nicht vorgesehen.

Aus diesen Gründen soll hier auf die Entwicklung eines genetischen Algorithmus' verzichtet werden. Allein die in Abschnitt 2.2.3.4 diskutierte Grundidee des Erlernens guter Lösungsbestandteile soll für die anstehende Konzeption eines Suchverfahrens übernommen werden.

Die Entwicklung von **Evolutionsstrategien** setzt voraus, daß Schrittweiten für die Lösungsänderungen definiert werden können.[379] Die im Data Mining relevanten Lösungsänderungen sind überwiegend diskret, wie z.B. die Auswahl der hinzuzufügenden oder zu streichenden Attribute. Oben wurde bereits darauf hingewiesen, daß gerade das Streichen von Attributen sehr effizient realisiert werden kann. Da sich bei solchen Suchoperationen kaum sinnvolle Schrittweiten definieren lassen, kommt die eigentliche Stärke der Evolutionsstrategien (geeignete Schrittweiten zu erlernen und diese zur Lösung des Optimierungsproblems zu nutzen) nicht zum tragen.[380] Somit ist bereits eine verfahrensspezifische Voraussetzung nicht erfüllt, so daß sich die Diskussion der oben definierten Anforderungen erübrigt.

---

[378] Vgl. BÄCK/SCHÜTZ (2001), S. 413.

[379] Vgl. BÄCK/HAMMEL/SCHWEFEL (1997), S. 5.

[380] Vgl. auch Ausführungen aus Abschnitt 2.2.3.4 zum Erlernen von guten Schrittweiten.

Das **Simulated Annealing** scheidet bezüglich der aufgestellten Anforderungen schlecht ab. Zwar *läßt es kurzfristige Verschlechterungen zu* – dies geschieht nach einer Folge von „Temperatur"-Parametern, welche durch den Benutzer gesetzt werden müssen.[381] Aber die Bestimmung einer geeigneten Folge von „Temperaturen" für eine Problemklasse, welche unabhängig von der Startlösung zum Optimum führt, kann sich sehr aufwendig gestalten.[382] Außerdem bietet das Simulated Annealing keine besonderen Mechanismen zur *Vermeidung von Zyklen*, zur *Intensivierung und Diversifizierung der Suche*. *Große Verschlechterungen der Zielfunktionswerte* werden i.d.R. – je nach Abkühlungsparametern – verboten. Zur *Laufzeit* kann wenig Allgemeingültiges ausgesagt werden. Sie liegt – je nach Abbruchkriterium – i.d.R. deutlich unter der für eine vollständige Exploration benötigten Laufzeit. Zwar bietet das Simulated Annealing aufgrund seiner einfachen Kontrollstruktur die Möglichkeit, *„intelligente" Suchoperatoren* zu integrieren und diese *eng an den Datenzugriff zu koppeln* – insgesamt überwiegen jedoch die Nachteile, so daß das Simulated Annealing hier nicht als Ansatz für die Entwicklung eines Suchverfahrens in Frage kommt.

Die Grundidee des **Tabu Search** bestand nach Abschnitt 2.2.3.4 darin, das wiederholte Besuchen von Lösungen durch eine Tabu-Liste explizit zu verbieten oder dessen Wahrscheinlichkeit gering zu halten, um *Zyklen in der Problemlösung* zu vermeiden. Ein zur Tabu-Liste analoger Mechanismus kann für die folgende Entwicklung eines Suchverfahrens hilfreich sein. *Intensivierungs- und Diversifizierungszüge* ließen sich durch Statistiken über den Erfolg von Lösungsbestandteilen realisieren.[383] Solche Statistiken können aber prinzipiell in jedes Suchverfahren integriert werden, so daß sie nicht dem Tabu Search zuzurechnen sind. Im Hinblick auf die übrigen aufgestellten Anforderungen bietet das Tabu Search keine besonderen Mechanismen.

Speziell im Zusammenhang mit Rough-Set-orientierten Regelmengen wurden bislang zumeist Greedy-Algorithmen entwickelt.[384] Auf o.g. Metastrategien wird nur selten zurückgegriffen. Eine Ausnahme stellt der genetische Algorithmus von HASHEMI ET AL.

---

[381] Die Funktion der „Temperatur"-Parameter im Simulated Annealing wurde in Abschnitt 2.2.3.5 beschrieben.

[382] Vgl. WOODRUFF (1994), S. 835.

[383] Vgl. DOMSCHKE/KLEIN/SCHOLL (1996), S. 609.

[384] Vgl. beispielsweise TSUMOTO (1997), S. 62 f.

dar.[385] Dieser codiert jedoch jede Regel als eigenes Individuum und verstößt damit gegen die Anforderung, Regelmengen als Ganzes zu bewerten. Außerdem verwendet er obere Näherungen – diese ist, wie in Kapitel 2 erläutert, nur dann nichtleer, wenn eine spezielle Teilmenge von Objekten (wie z.B. die Menge der Kunden mit hohem Umsatz) beschrieben werden soll. Auf dieses Spezialproblem soll das zu entwickelnde Verfahren aber nicht beschränkt werden, so daß es ohne eine obere Näherung auskommen muß.

### 4.5.3 Betrachtung existierender Interessantheitskonzepte

Existierende Ansätze zur Bewertung der Interessantheit wurden bereits in Abschnitt 2.2.4 vorgestellt. In der dort angegebenen Literatur werden nur einzelne Interessantheitsfacetten vorgeschlagen, ohne ein geschlossenes Bewertungskonzept zu bieten, welches die aufgestellten Anforderungen erfüllt. Insbesondere weisen viele Bewertungskonzepte die Unzulänglichkeit auf, daß nicht alle produzierten Aussagen als Gesamtheit, sondern *jede Aussage für sich genommen bewertet wird*.[386] Dadurch werden extrem redundante oder in sich widersprüchliche Aussagenmengen erzeugt. Darüber hinaus sind die generierten Modelle häufig nicht reliabel, da sie innerhalb des Lernprozesses *nicht nach ihrer Allgemeingültigkeit evaluiert werden*.[387]

Gerade bei der Bewertung der Interessantheit existiert ein großes Verbesserungspotential, zumal hier auch die Möglichkeit besteht, betriebswirtschaftliche Überlegungen mit in die Entwicklung eines Data-Mining-Verfahrens einzubeziehen.

⇨ Erstens können unter bestimmten Voraussetzungen die Interessantheitsfacetten selbst Kosten- und Erfolgsfunktionen darstellen, so daß die Modellbewertung in betriebswirtschaftlichen Größen vorgenommen werden kann. Speziell bei der Generierung von Data-Mining-Entscheidungsmodellen kann der betriebswirtschaftliche Nutzwert über alle Entscheidungssituationen maximiert werden. Damit wäre ein direkter *Entscheidungsbezug* gegeben, den klassische Klassifikations- und Prognoseverfahren, wie z.B. neuronale Lernverfahren, Entscheidungsbaumverfahren oder

---

[385] Vgl. HASHEMI ET AL. (1997), S. 164 ff.

[386] Vgl. z.B. KRABS (1994), S. 42 ff., MORIK/BROCKHAUSEN (1997), S. 292 und WROBEL (1997), S. 82.

[387] Vgl. z.B. KRABS (1994), S. 42 ff.

diskriminanzanalytische Verfahren, nicht bieten. Die klassischen Verfahren sagen lediglich einen einzelnen Wert voraus, der i.d.R. die erwartete Umweltsituation widerspiegelt.[388] Um eine Entscheidung fällen zu können, muß jedoch nach den Ausführungen von Abschnitt 3.3 die gesamte Verteilung der erzielbaren Handlungsergebnisse ermittelt und ökonomisch bewertet werden können.

Auch bei den Modelltypen mit geringerem Entscheidungsbezug, den Beschreibungs- und Erklärungsmodellen, kann über die in den Abschnitten 3.3.2.4 und 3.3.2.3 eingeführte Erfolgsrelevanz eine ökonomisch fundierte Bewertung in das Gütemaß integriert werden.

⇨ Zweitens kann nach Abschnitt 3.3 ein direkter *Bezug zwischen der dort vorgenommenen Klassifizierung betriebswirtschaftlicher Aufgabenstellungen und den anzuwendenden Bewertungsvorschriften* hergestellt werden. Gerade dies leistet kein einziger Literaturbeitrag in hinreichendem Maße. Selbst ein Forschungsprojekt, dem explizit das Ziel zugrunde lag, ein Instrumentarium zu entwickeln, anhand dessen die Einsatzpotentiale von Data Mining im betrieblichen Umfeld beurteilt werden können,[389] vernachlässigt den überaus wichtigen Bezug zwischen der betriebswirtschaftlichen Problemstellung und den anzuwendenden Bewertungsvorschriften und liefert diesbezüglich kein operationales Instrument. Andere betriebswirtschaftlich orientierte Forschungsbeiträge konzentrieren sich jeweils auf einen speziellen Problemtyp und stellen damit auch keine Verbindung zwischen verschiedenen Problemtypen und Interessantheitsmaßen bereit.[390]

### 4.5.4 Betrachtung existierender Datenzugriffskomponenten

Die Datenzugriffskomponente ist in den meisten Data-Mining-Verfahren ausgesprochen rudimentär realisiert. So wird die gesamte Datenbasis häufig in den Hauptspeicher geladen, was eine Anwendung auf große Datenbestände ausschließt. Das Ziehen von Stichproben löst das Problem nur teilweise, da hier nur die Anzahl der Datensätze, aber nicht die Anzahl der Variablen eingeschränkt wird.

---

[388] Dieser Fall wurde in Abschnitt 3.3.3.3 als „Fall 1“ bezeichnet.

[389] Vgl. KÜPPERS (1999), S. 14.

[390] Vgl. BISSANTZ (1996), S. 91 ff. und SÄUBERLICH (2000), S. 139 ff.

Eine Gruppe von Verfahren, welche *direkt auf große Datenbanken zugreifen, die Trainingsmenge dynamisch aus mehreren Tabellen zusammenstellen* und in annehmbaren Laufzeiten Datenmuster mit $1{:}N$-Beziehungen-Beziehungen extrahieren können, werden unter dem Oberbegriff „***Assoziationsregelverfahren***“ zusammengefaßt. Diese Verfahren zeichnen sich dadurch aus, daß sie Abhängigkeiten zwischen Objekten eines Typs (z.B. zwischen $N$ Artikeln) extrahieren, welche zusammen ein Objekt eines anderen Typs (z.B. einen Warenkorb) bilden.[391] Aufgrund ihrer einfachen Wissensrepräsentation und Interessantheitsbewertung konnten leistungsfähige Suchverfahren mit *effizienten Datenzugriffen* entwickelt werden, so daß die Verarbeitung großer Datenmengen möglich geworden war. Einige wenige Lösungsansätze existieren bereits, die das Problem des Datenzugriffs in Kombination mit der Repräsentation komplexerer relationaler Wissensstrukturen angehen.[392] Sie bieten jedoch, wie bereits in den Abschnitten zuvor herausgestellt wurde, weder ein leistungsfähiges Suchverfahren, das in der Lage ist, große Suchräume effektiv und effizient zu durchforsten, noch eine umfassende, betriebswirtschaftlich orientierte Bewertung der Interessantheit.

Wurde oben bereits die Entwicklung eines neuen Suchverfahrens nahegelegt, so folgt daraus, daß auch die Datenzugriffskomponente neu konzipiert werden muß, denn der Datenzugriff sollte eng mit dem Suchverfahren gekoppelt sein, um gute Laufzeiten zu erzielen, da die Laufzeit zum größten Teil durch die Datenzugriffe auf externe Speicher determiniert wird.[393]

## 4.6 Schlußfolgerungen für die Entwicklung eines neuen Data-Mining-Verfahrens

In den ersten vier Abschnitten dieses Kapitels wurden Anforderungen an die Komponenten von Data-Mining-Verfahrens definiert. Die daruf folgende Betrachtung existierender Verfahrenskomponenten hat gezeigt, daß in einigen Bereichen bereits verwendbare Lösungen existieren. Interessant erscheinen vor allem:

---

[391] Vgl. AGRAWAL ET AL. (1996), S. 308.

[392] Vgl. WROBEL (1997), S. 78 ff., MORIK/BROCKHAUSEN (1997), S. 287 ff. und KNOBBE/SIEBES/VAN DER WALLEN (1999), S. 378 ff.

[393] Vgl. IMIELINSKI/MANNILA (1996), S. 59 f.

- ⇨ *zur Wissensrepräsentation*: eine Mischform zwischen der Prädikatenlogik und der konjunktiven Normalform;
- ⇨ *zur Steuerung der Suche*: ein attributorientiertes Verfahren in Kombination mit einigen Grundideen verschiedener Meta-Heuristiken (Tabu-Mechanismus, Erlernen guter Lösungsbestandteile, Intensivierungs- und Diversifizierungsmechanismen);

Der *Datenzugriff* wird aufgrund der engen Kopplung mit dem Suchverfahren neu zu entwickeln sein. Auch die *Bewertungskomponente* kann nicht aus der Literatur übernommen werden, da dort kaum problemtyporientierte ökonomischen Bewertungsansätze existieren.

Bei der Aufstellung der Anforderungen wurde keine Rücksicht darauf genommen, ob sie im Rahmen dieser Untersuchung auch erfüllt werden können. Aufgrund des hohen Entwicklungsaufwandes soll hier mit den Data-Mining-Entscheidungsmodellen nur einer der vier in Kapitel 3 definierten Modelltypen realisiert werden. Auf die Entscheidungsmodelle fiel die Wahl aus folgenden Gründen:

- ⇨ **Entscheidungsbezug**: Data-Mining-Entscheidungsmodelle weisen den engsten Entscheidungsbezug auf. In die ökonomische Bewertung können unmittelbar die Konsequenzen der empfohlenen Entscheidung einfließen. Damit besitzen Entscheidungsmodelle den größten Wert für den Entscheidungsträger.
- ⇨ **Innovation**: Es gibt bereits viele und weit entwickelte Segmentierungs- und Prognoseverfahren. Im Rahmen des Data Mining sind hier im Hinblick auf die Segmentierungsverfahren (aufgrund der integrierten Clusteridentifikation und -beschreibung) vor allem das konzeptionelle Clustern und im Hinblick auf die Prognoseverfahren (aufgrund der Lernfähigkeit beliebiger stetiger Funktionen) neuronale Netze zu nennen. Für erklärende Aufgaben werden vor allem Entscheidungsbaum- und Assoziationsregelverfahren eingesetzt. Beide Verfahrenstypen weisen Schwächen im Hinblick auf eine ökonomische Bewertung und eine intelligente Suche auf. Somit bestünde in diesen Bereichen ein Verbesserungspotential. Am wenigsten erforscht und wirklich *innovativ* ist aber die Generierung von Entscheidungsmodellen.

# 5 Entwicklung eines Data-Mining-Verfahrens zur Generierung von Entscheidungsmodellen

Dieses Kapitel verfolgt das in Abschnitt 1.3 aufgestellte Ziel 2 *(„Lösungsverfahren entwickeln“).* In Abschnitt 4.6 wurde begründet, warum die Entwicklung eines Data-Mining-Verfahrens zur Generierung von Entscheidungsmodellen bezüglich deren Innovationsgrades und Entscheidungsbezugs anderen Modelltypen vorzuziehen ist.

Die Konzeption des Verfahrens wird in den folgenden vier Abschnitten – getrennt nach den eingeführten Komponenten von Data-Mining-Verfahren – vorgestellt und kritisch diskutiert.

## 5.1 Der Datenzugriff

Als erste der vier Verfahrenskomponenten wird der Zugriff auf die relationale Datenbank konzipiert (vgl. Abschnitt 5.1.1). Diese Konzeption wird in Abschnitt 5.1.2 einer kritischen Diskussion unterzogen.

### 5.1.1 Konzeption des Datenzugriffs

In Abschnitt 4.4 wurde die Anforderung aufgestellt, daß die Trainingsmenge aus der relationalen Datenbank dynamisch zu erzeugen sei. Auch wenn hier relationale Datenbanken mit beliebig vielen Relationen betrachtet werden, bezieht sich eine Data-Mining-Analyse auf genau einen Objekttyp, zu dem Aussagen generiert werden sollen. Dieser wird im folgenden als „***Bezugsobjekttyp***“ und die Objekte als „***Bezugsobjekte***“ bezeichnet. Im Hinblick auf die relationale Datenbank wird der Bezugsobjekttyp auch „***Bezugsrelation***“ genannt. Zur Notation des Datenzugriffs dienen neben den in Anhang A dargestellten Grundlagen von Datenbanken auch die folgende Begriffe:

**Definition 5-1: Jointabellen, Joinpfad, Mehrfachjoin**

Gegeben sei eine Menge von Relationen, $RM = \{R_1, ..., R_{Rmax}\}$, und eine ausgezeichnete Bezugsrelation, $BR \in RM$. $R_i.A$ sei die Menge der Attribute der $i$-ten Relation, $R_i$. $R_i.SA$ sei eine vom Benutzer definierte Attributmenge mit $R_i.SA \subseteq R_i.A$, welche die Datenobjekte

der $i$-ten Relation, $R_i$ ($i = 1, ..., Rmax$), eindeutig identifiziert.[394] Der Benutzer gibt zu jeder relevanten Relation, $R \in RM$, einen Pfad an, über den $R$ mit der Bezugsrelation, $BR$, verknüpft werden soll:

$Joinpfad_{BR}(R) := (R_{i(1)}.SA = R_{i(2)}.SA) \wedge (R_{i(2)}.SA = R_{i(3)}.SA) \wedge ... \wedge (R_{i(f-1)}.SA = R_{i(f)}.SA)$

mit $f, i(j) \in \{1,...,Rmax\};$

$R_{i(j)}.SA \subseteq R_{i(j)}.A$

$j = 1, ..., f;$

$BR, R \in \{R_{i(1)}, ..., R_{i(f)}\} \subseteq RM.$

Ein Joinpfad kann, wie im nächsten Abschnitt gezeigt wird, in der WHERE-Klausel eines SELECT-Befehls angewendet werden, um die Relationen $R_{i(1)}, ..., R_{i(f)}$ über Join-Operationen zu verknüpfen. Weiter wird definiert:

$Jointabellen_{BR}(R) := \{R_{i(1)}, ..., R_{i(f)}\};$

$Mehrfachjoin_{BR}(R) := R_{i(1)}[R_{i(1)}.SA=R_{i(2)}.SA]R_{i(2)}...R_{i(f-1)}[R_{i(f-1)}.SA=R_{i(f)}.SA]R_{i(f)}.$ ◊

Die Jointabellen stellen eine Menge von Relationen dar, der Joinpfad ist ein sprachlicher Ausdruck und der Mehrfachjoin eine Relation als Ergebnis mehrerer Join-Operationen.

Diese Begriffe werden an folgendem Beispiel verdeutlicht:

*Gegeben sei eine Relationenmenge mit den Relationen Kassenbon, Bonposition und Artikel und den in den folgenden Tabellen abgebildeten Datenobjekten.*

[394] Eine eindeutig identifizierende Attributmenge ist der Primärschlüssel (vgl. Definition 7-3) einer Relation.

| Kassenbon | |
|---|---|
| **Bon_id** | **Wochentag** |
| 1 | Montag |
| 2 | Samstag |

| Bonposition | | |
|---|---|---|
| **Bon_id** | **Bonpos_id** | **Artikel_id** |
| 1 | 1 | 3 |
| 1 | 2 | 5 |
| 1 | 3 | 4 |
| 1 | 4 | 1 |
| 2 | 1 | 2 |
| 2 | 2 | 3 |

| Artikel | |
|---|---|
| **Artikel_id** | **Bezeichnung** |
| 1 | Cola |
| 2 | Chips |
| 3 | Fanta |
| 4 | Bier |
| 5 | Kaffee |

***Tabelle 5-1: Beispiel-Datenbank mit der Bezugsrelation „Kassenbon“***

*Es sollen nun Aussagen über typische Warenkörbe getroffen werden, z.B.:*

`WENN ein Warenkorb an einem Montag eingekauft wurde und Cola enthält, DANN enthält er auch Bier.`

*Einem Warenkorb entspricht in der Relationenmenge ein Kassenbon. Damit stellt die Relation Kassenbon die sog. „Bezugsrelation“, BR, dar. Die oben definierten Begriffe stellen sich in diesem Beispiel wie folgt dar:*

$Jointabellen_{Kassenbon}(Artikel) = \{Artikel, Bonposition, Kassenbon\};$

$Jointabellen_{Kassenbon}(Bonposition) = \{Bonposition, Kassenbon\};$

$Jointabellen_{Kassenbon}(Kassenbon) = \{Kassenbon\};$

$Joinpfad_{Kassenbon}(Artikel) = (Artikel.Artikel\_id = Bonposition.Artikel\_id \wedge Bonposition.Bon\_id = Kassenbon.Bon\_id);$

$Joinpfad_{Kassenbon}(Bonposition) = (Bonposition.Bon\_id = Kassenbon.Bon\_id);$

$Joinpfad_{Kassenbon}(Kassenbon) = ().$

*Die folgenden Tabellen zeigen die Ergebnisse der Mehrfachjoins.*[395]

| $Mehrfachjoin_{Kassenbon}(Artikel)$ | | | | |
|---|---|---|---|---|
| **Kassenbon. Bon_id** | **Bonposition. Bonpos_id** | **Artikel. Artikel_id** | **Artikel. Bezeichnung** | **Kassenbon. Wochentag** |
| 1 | 1 | 3 | Fanta | Montag |
| 1 | 2 | 5 | Kaffee | Montag |
| 1 | 3 | 4 | Bier | Montag |
| 1 | 4 | 1 | Cola | Montag |
| 2 | 1 | 2 | Chips | Samstag |
| 2 | 2 | 3 | Fanta | Samstag |

***Tabelle 5-2: Mehrfachjoin für die Relation „Artikel“***

[395] Bei den folgenden Darstellung von Joins wird von zwei identischen Spalten, wie z.B. von *Bonposition.Artikel_id* und *Artikel.Artikel_id*, jeweils eine unterdrückt.

| *Mehrfachjoin$_{Kassenbon}$(Bonposition)* | | | |
|---|---|---|---|
| ***Kassenbon.Bon_id*** | ***Bonposition.Bonpos_id*** | ***Artikel.Artikel_id*** | ***Kassenbon.Wochentag*** |
| *1* | *1* | *3* | *Montag* |
| *1* | *2* | *5* | *Montag* |
| *1* | *3* | *4* | *Montag* |
| *1* | *4* | *1* | *Montag* |
| *2* | *1* | *2* | *Samstag* |
| *2* | *2* | *3* | *Samstag* |

***Tabelle 5-3: Mehrfachjoin für die Relation „Bonposition“***

| *Mehrfachjoin$_{Kassenbon}$(Kassenbon)* | |
|---|---|
| ***Kassenbon.Bon_id*** | ***Kassenbon.Wochentag*** |
| *1* | *Montag* |
| *2* | *Samstag* |

***Tabelle 5-4: Mehrfachjoin für die Relation „Kassenbon“***

Die explizite Angabe der Joinpfade bzw. –tabellen ist deswegen erforderlich, weil es nach dem Datenschema mehrere Möglichkeiten geben kann, eine beliebige Relation mit der Bezugsrelation zu verknüpfen.

*Im Beispiel könnte man, wenn man das Datenschema gemäß Abbildung 5-1 erweitern würde, folgende alternative Jointabellen definieren:*

*Jointabellen$_{Kassenbon}$(Artikel) := (Artikel, Bonposition, Kassenbon);*

*Jointabellen$_{Kassenbon}$(Artikel) := (Artikel, Bonposition, Zuordnung, Kassenbon).*

*Die erstgenannten Jointabellen liefern, wenn der entsprechende Mehrfach-Join ausgeführt wird, alle Kassenbons mit allen Artikeln, die gekauft wurden (vgl. Tabelle 5-2). Die zweitgenannten Jointabellen liefern, wenn der entsprechende Mehrfach-Join ausgeführt wird, diejenigen Kassenbons, die Artikel aufführen, welche mindestens einmal reklamiert wurden, zusammen mit den reklamierten Artikeln.*

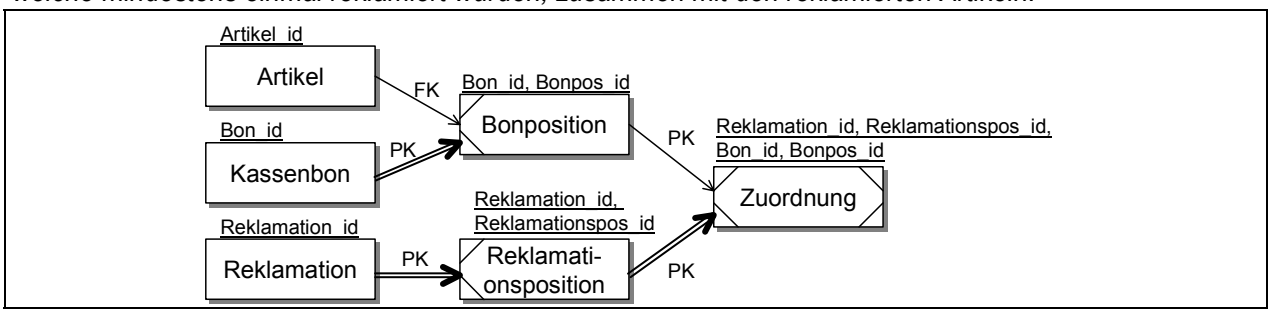



***Abbildung 5-1: Erweitertes Datenschema[396] für das Beispiel aus Tabelle 5-1***

Die verschiedenen Verknüpfungsmöglichkeiten unterscheiden sich in ihrer semantischen Interpretation. Daher kann nur der Benutzer die Jointabellen sinnvoll und der

[396] Das Datenschema ist hier als strukturiertes Entity-Relationship-Diagramm dargestellt, dessen Notation in Anhang C erläutert wird.

Problemstellung angemessen definieren. Allerdings kann sich die Definition der Jointabellen für alle potentiell interessanten Relationen als recht aufwendig erweisen, so daß eine Data-Mining-Software eine Default-Definition für die Jointabellen vorsehen sollte, falls der Benutzer keine eigene Definition vornimmt. Diese Default-Definition könnte beispielsweise die (erste gefundene) kürzeste Verbindung zwischen der angegebenen Relation, $R$, und der Bezugsrelation, $BR$, darstellen.

Für die angestrebte dynamische Erzeugung der Trainingsmenge wird noch der folgende Begriff definiert:

**Definition 5-2: Datenbasis**

Es gelten dieselben Voraussetzungen wie in Definition 77. Weiterhin sei mit $A_{BR}$ die Menge der Attribute bezeichnet, für die eine Verbindung zur Bezugsrelation, $BR$, definiert wurde:

$$A_{BR} := \bigcup_{\substack{i \in \{1,\ldots,Rmax| \\ \exists Joinpfad_{BR}(R_i), R_i \in RM\}}} R_i.A \quad .$$

Gegeben sei weiterhin eine aktuelle Auswahl von Attributen, $A=\{a_1, \ldots, a_{amax}\} \subseteq A_{BR}$. Die Relation, die das Attribut $a_i$ enthält, sei mit $R_{a_i}$ bezeichnet, ihre identifizierenden Attribute mit $R_{a_i}.SA \;\; (i = 1,\ldots,amax)$. Die identifizierenden Attribute der Bezugsrelation, $BR$, werden als $BR.SA$ bezeichnet. Weiterhin gelte folgende Kurzschreibweise für Mehrfachjoin und anschließende Projektion[397] auf die relevanten Attribute:

$$MJ^{a_i} := Mehrfachjoin_{BR}(R_{a_i})[\{a_i\} \cup R_{a_i}.SA].$$

Damit definiert man die ***Datenbasis***, $DB(BR,A)$, wie folgt:

$$DB(BR,A) := MJ_{a_1}[MJ_{a_1}.BR.SA = MJ_{a_2}.BR.SA]MJ_{a_2} \ldots$$
$$\ldots MJ_{a_{amax-1}}[MJ_{a_{amax-1}}.BR.SA = MJ_{a_{amax}}.BR.SA]MJ_{a_{amax}}.$$

D.h., die Konstruktion der Datenbasis erfolgt als Join über die Mehrfachjoins der Bezugsrelation, $BR$, mit den Relationen, in denen die ausgewählten Attribute aus $A$ vorkommen.[398] ◊

---

[397] Die Projektion wurde in Definition 2-50 eingeführt.

[398] Bei Definition 5-2 ist zu beachten, daß es sich bislang um eine rein formale Notation auf konzeptioneller Ebene handelt. Die dynamische Erzeugung der Trainingsmenge wird keineswegs als sequenzielle Ausführung von Joins und Projektionen implementiert, da diese Vorgehensweise bezüglich Laufzeit und Speicherbedarf ineffizient wäre. Eine schnellere und speichersparendere Alternative bietet die Entwicklung eigener Datenbankoperatoren.

*Als Fortführung des obigen Beispiels sollen die Attribute $A = \{Artikel.Bezeichnung, Reklamation.Wochentag\}$ bezüglich der Bezugsrelation $BR = Kassenbon$ in die aktuelle Trainingsmenge aufgenommen werden. Die Mehrfach-Joins $MJ_{Artikel.Bezeichnung}$ und $MJ_{Reklamation.Wochentag}$ werden in Tabelle 5-5 dargestellt.*[399]

| $Mehrfachjoin_{Kassenbon}(Artikel)$ [{Kassenbon.Bon_id, Artikel.Bezeichnung}] | |
|---|---|
| ***Kassenbon.Bon_id*** | ***Artikel.Bezeichnung*** |
| *1* | *Fanta* |
| *1* | *Kaffee* |
| *1* | *Bier* |
| *1* | *Cola* |
| *2* | *Chips* |
| *2* | *Fanta* |

| $Mehrfachjoin_{Kassenbon}(Reklamation)$ [{Kassenbon.Bon_id, Reklamation.Wochentag}] | |
|---|---|
| ***Kassenbon.Bon_id*** | ***Reklamation.Wochentag*** |
| *1* | *Dienstag* |

***Tabelle 5-5: Mehrfachjoin für die Relationstypen „Kassenbon“ und „Reklamation“ nach Projektion auf die relevanten Attribute***

*Die Datenbasis ergibt sich wie folgt:*

$DB(Kassenbon, \{Artikel.Bezeichnung, Reklamation.Wochentag\})$

$:= MJ_{Artikel.Bezeichnung}\ [MJ_{Artikel.Bezeichnung}.Kassenbon.Bon\_id = MJ_{Reklamation.Wochentag}.Kassenbon.Bon\_id]\ MJ_{Reklamation.Wochentag}$.

*Das Ergebnis ist in Tabelle 5-6 dargestellt.*

| ***DB(Kassenbon, {Artikel.Bezeichnung, Reklamation.Wochentag})*** | | |
|---|---|---|
| ***Kassenbon.Bon_id*** | ***Artikel.Bezeichnung*** | ***Reklamation.Wochentag*** |
| *1* | *Fanta* | *Dienstag* |
| *1* | *Kaffee* | *Dienstag* |
| *1* | *Bier* | *Dienstag* |
| *1* | *Cola* | *Dienstag* |

***Tabelle 5-6: Datenbasis mit einem Bezugsobjekt (Kassenbon)***

Bei Zugriffen auf die Datenbasis (wie z.B. zur Berechnung der Sicherheit oder der Allgemeingültigkeit) muß im folgenden berücksichtigt werden, daß die Datenbasis nicht länger genau aus einer Zeile pro Bezugsobjekt besteht. So erhält man die Anzahl der Objekte, die einen Term, $Te$, erfüllen, nicht – wie bisher – durch eine einfache Selektion, $|O^T[Te]|$, sondern durch Selektion und anschließende Projektion auf die identifizierenden Attribute der Bezugsrelation:

$|DB(BR,A)[Te][BR.SA]|$.

[399] Die zugrundeliegenden Tabellen „*Reklamation*“, „*Reklamationsposition*“ und „*Zuordnung*“ sind hier nicht von Interesse und wurden daher nicht abgebildet.

### 5.1.2 Kritische Diskussion des Datenzugriffs

Die Zusammenstellung der Datenbasis, $DB(BR,A)$, aus einer relationalen Datenbank erfolgt durch SQL-Befehle. Es gibt mehrere Möglichkeiten, dies zu realisieren. Prinzipiell könnte die Datenbasis direkt durch den Datenbank-Server zusammengestellt und das Ergebnis, $DB(BR,A)$, in das Data-Mining-Verfahren eingelesen werden. Allerdings erwies sich diese Möglichkeit als zu langsam. Die verwendeten Datenbank-Server (MS-Access und Borland Interbase) verarbeiteten die entsprechenden SQL-Befehle offenbar zu ungeschickt. Daher wurde ein anderer Lösungsweg eingeschlagen: Jedes benötigte Attribut wird zusammen mit den identifizierenden Attributen in einer eigenen Abfrage eingelesen. Der Join über die eingelesenen identifizierenden Attributwerte erfolgt dann im Rahmen des implementierten Data-Mining-Verfahrens. Dieser Weg ist zwar wesentlich aufwendiger zu implementieren, aber dafür – zumindest mit den verwendeten Datenbankservern – viel schneller.

Auf die Darstellung der Algorithmen zum Einlesen der SQL-Abfrageergebnisse soll hier verzichtet werden. Stattdessen werden nur die abgesendeten SQL-Befehle betrachtet. Zum Einlesen eines nominalen Attributes, $R.a$, aus der Relation $R$ werden folgende Datenbankbefehle ausgeführt:[400]

```
CREATE VIEW "chti1" AS
   SELECT DISTINCT   R.a, COUNT(*)
   FROM              Jointabellen^BR(R)
   WHERE             Joinpfad^BR(R)
   GROUP BY          R.a
   HAVING            COUNT(*) ≥ (min_Allgemeingültigkeit · |BR|)

SELECT DISTINCT      R.a, BR.SA
FROM                 Jointabellen^BR(R), „chti1", „Stichprobe"
WHERE                Joinpfad^BR(R) AND
                     („Stichprobe".SA = BR.SA) AND
                     (BR.SA = „chti1".SA)
ORDER BY             R.a ASC, BR.SA ASC
```

[400] Dabei wird angenommen, daß zuvor eine Tabelle “Stichprobe“ erzeugt wurde, die eine zufällig zusammengestellte Menge von identifizierenden Schlüsseln aus der Bezugsrelation enthält.

*Für das Beispiel aus dem Abschnitt zuvor ergeben sich folgende Datenbankbefehle als spezielle Instanzen des allgemeinen Schemas:*

```
CREATE VIEW "chti1" AS
    SELECT DISTINCT     "Artikel"."Bezeichnung", COUNT(*)
    FROM                "Kassenbon", "Artikel", "Bonposition"
    WHERE               ("Artikel"."Artikel_id" = "Bonposition"."Artikel_id") AND
                        ("Kassenbon"."Bon_id" = "Bonposition"."Bon_id")
    GROUP BY            "Artikel"." Bezeichnung "
    HAVING              COUNT(*) ≥ 30

SELECT DISTINCT         "Artikel"." Bezeichnung ", "Kassenbon"."Bon_id"
FROM                    "Kassenbon", "Artikel", "Bonposition", chti1, "Stichprobe"
WHERE                   ("Artikel"."Artikel_id" = "Bonposition"."Artikel_id") AND
                        ("Kassenbon"."Bon_id" = "Bonposition"."Bon_id") AND
                        (chti1."Artikel"." Bezeichnung " = "Artikel"." Bezeichnung ")
                        AND
                        ("Stichprobe"."Bon_id" = "Kassenbon"."Bon_id")
ORDER BY                "Artikel"." Bezeichnung " ASC, "Kassenbon"."Bon_id" ASC
```

Bei Kardinal- und Ordinalattributen, $R.a$, der Relation $R$, wird für jedes Intervall, $[cp_{a,i}; cp_{a,i+1})$ mit $i = 1, ..., cpmax(a)-2$, eine eigene Datenbankabfrage durchgeführt:

```
SELECT DISTINCT     BR.SA
FROM                Jointabellen^BR(R), „Stichprobe"
WHERE               Joinpfad^BR(R) AND
                    („Stichprobe".SA = BR.SA) AND
                    (R.a ≥ cp_{a,i}) AND
                    (R.a < cp_{a,i+1})
ORDER BY            BR.SA ASC
```

*Für das Beispiel ergibt sich – wenn man die Tabelle „Kassenbon" um das kardinale Attribut „Uhrzeit" erweitert – u.a. folgende Datenbankabfrage:*

```
SELECT DISTINCT     "Kassenbon"."Bon_id"
FROM                "Kassenbon", "Stichprobe"
WHERE               ("Stichprobe"."Bon_id" = "Kassenbon"."Bon_id") AND
                    ("Kassenbon"."Uhrzeit" >= 10) AND
                    ("Kassenbon"."Uhrzeit" < 12)
ORDER BY            "Kassenbon"."Bon_id" ASC
```

Leider dauert bereits das Absenden solch einfacher SQL-Befehle und das Verarbeiten durch den Datenbankserver sehr lange. Die Ausführungszeit für einen einzelnen SELECT-Befehl liegt für eine Testdatenbank[401] im Mittel bei *1,8* Sekunden[402]. Die Zeit

[401] Die als Bezugsrelation gewählte Tabelle der Datenbank umfaßt *1215* Datensätze. Als Stichprobenumfang wurde *50%* (gut *600* Datensätze) gewählt. Die größte der *15* Tabellen umfaßte knap *20.000* Datensätze – die meisten Tabellen waren jedoch wesentlich kleiner.

für das Absenden des CREATE-VIEW-Befehls ist vernachlässigbar. Da jede Lösung *2* bis *Prmax_max+1* Variablen umfaßt, dauert die reine SQL-Verarbeitung pro Lösung *2·1,8 = 3,6* bis (für *Prmax_max=5*) *6·1,8 = 10,8* Sekunden. Falls in einem Suchprozeß *10.000* Lösungen zu testen sind, so muß das Data-Mining-Verfahren – selbst bei dieser kleinen Datenbank – zwischen *10* und *30* Stunden auf den Datenbankserver warten. In dieser Zeit hat das Verfahren noch keinen einzigen Befehl abgearbeitet. Diese Laufzeit ist für praktische Zwecke ungeeignet.

Nun kann die Laufzeit über spezielle Zwischenspeicher verringert werden. Das Verfahren wurde so konzipiert, daß die am häufigsten benötigten Felder (das Konklusionsfeld und die gut bewerteten Felder) im Hauptspeicher zwischengespeichert werden, solange dieser ausreicht. Mit einem derartig modifizierten Verfahren dauert das Einlesen eines Feldes im Mittel (wobei der Durchschnitt über das Einlesen aus der Datenbank und aus dem Hauptspeicher gebildet wurde) *1,340* Sekunden. Pro Lösung sind damit *2·1,340 = 2,680* bis *6·1,340 = 8,040* Sekunden anzusetzen. Bei *10.000* Lösungen ergibt sich eine Laufzeit von *7,44* bis *22,33* Stunden für das Einlesen der Datensätze – wobei nun bereits die Laufzeit für das Einlesen der SQL-Abfrageergebnisse in den Hauptspeicher mit berücksichtigt wurde – insgesamt konnte die Laufzeit also deutlich verringert werden. Trotzdem ist sie – wenn man sich noch einmal vor Augen führt, daß die Testdatenbank sehr klein ist – zu hoch, um das Verfahren praktisch verwenden zu können.

Die Konsequenz aus der hohen Wartezeit auf den Datanbankserver kann also nur lauten, daß ein direkter Zugriff auf eine relationale Datenbank mit herkömmlichen Datenbankservern nicht realisierbar ist. Eine direkte Kopplung des Suchverfahrens mit der Datenbank muß also noch näher an der Datenbanktechnik aufsetzen. Bereits die Funktionalität des Servers muß mit den Datenzugriffen des Suchverfahrens abgestimmt werden. Hierzu müßten die eingangs gezogenen Grenzen, die Realisierung auf PC-Basis vorzunehmen, aufgehoben und Client-Server-Systeme betrachtet werden. Ein Ansatzpunkt zur Beschleunigung des Datenzugriffs besteht darin, einen eigenen Server einzurichten, der sowohl das Datenbanksystem umfaßt als auch Funktionen zur

[402] Diese Zeitangabe gilt für das Einlesen nominaler Attribute. Metrische Attribute werden schneller eingelesen. Dafür muß dann jedes Intervall einzeln eingelesen werden, so daß für alle Attribute hier dieselbe Zeit veranschlagt wird.

Ausführung von sog „*KDD-Queries*“[403] bereitstellt. Dabei handelt es sich um Abfragen ähnlich zu SELECT-Befehlen in SQL, deren Ergebnisse nicht Mengen von Datenobjekten, sondern Mengen von Datenmustern (oder Vorformen davon) darstellen. Diese können dann über ein lokales Netz zu den Clients übertragen werden, so daß die Netzlast wesentlich geringer ist als bei der Übertragung der sehr viel größeren Mengen von Datenobjekten. Die Suchstrategie kann ebenfalls auf dem Server implementiert sein. Lediglich die Dialogkomponte zur Kommunikation mit dem Benutzer sollte auf den Clients installiert sein. Derartige „*Data-Mining-Server*“ können beispielsweise das Verwalten von speziellen Datenstrukturen, die Optimierung der Datenzugriffe mehrerer KDD-Queries, die Speicherung aggregierter Informationen oder das Ziehen von Stichproben übernehmen.[404] Derartige Lösungsansätze können aufgrund mangelnder Ressourcen jedoch nicht weiter verfolgt werden. Daher muß das Fazit lauten:

| |
|---|
| Die Anforderung, einen direkten Zugriff auf die relationale Datenbank zu realisieren, kann nicht erfüllt werden. |

## 5.2 Der Modelltyp

Der folgende Abschnitt 5.2.1 stellt die Konzeption des Modelltyps vor, die dem Data-Mining-Verfahren zugrunde liegt. Diese wird in Abschnitt 5.2.2 den definierten Anforderungen gegenübergestellt.

### 5.2.1 Konzeption des Modelltyps

Da der direkte Datenzugriff auf eine relationale Datenbank verworfen wurde, entfällt auch die Anpassung der konjunktiven Normalform auf multirelationale Datenmuster. Der Modelltyp ist ohne diese Anpassung identisch zu den in Abschnitt 2.2 vorgestellten Regelmengen in KNF. Dort wurde auch der Lösungsraum definiert. Demgegenüber umfaßt der Suchraum alle durch das Suchverfahren erreichbare Lösungen. Der Suchraum soll hier definiert und diskutiert werden, womit der Konzeption des Suchverfahrens vorgegriffen wird.

---

[403] Vgl. IMIELINSKI/MANNILA (1996), S. 60 ff.

[404] Vgl. HOLSHEIMER ET AL. (1995), S. 151 ff.

In Abschnitt 4.5.2 wurden attributorientierte Verfahren mit der Bemerkung erwähnt, daß sie die Anforderung nach einem (relativ zum Lösungsraum) nicht zu großen Suchraum erfüllen. Während der Lösungsraum nahezu beliebige Regelmengen umfaßt, wird der Suchraum im folgenden auf der Grundlage des Rough-Set-Ansatzes[405] sehr viel restriktiver konzeptioniert. Dabei werden geordnete Wertemengen in Intervalle aufgeteilt. Für kardinale und ordinale Domänen eines Attributes $a$, $dom(a)$, werden sog. „***Cutting Points***" definiert – dies sind Werte aus $dom(a)$, die zur Aufteilung der Wertemenge verwendet werden. $cpmax^V(a)$ bezeichne die Anzahl der für das Attribut $a$ definierten Cutting Points. Vorgegeben sei jeweils eine Menge von Cutting Points, $CP^V(a) \subseteq dom(a)$ mit $cpmax^V(a) \geq 3$ ($\forall a \in A$, $a$ geordnet). Jede Regelmenge mit den erklärenden Attributen, $C$, und dem zu erklärenden Attribut, $D=\{a^D\}$, basiert dann auf einer aktuellen Auswahl von Cutting Points, $\forall a^C \in C$: $CP(a^C)$, $CP(a^D)$. Für ein geordnetes Attribut, $a \in A$, ist $CP(a)$ wie folgt definiert:

$$CP(a) = \{cp_{a,1},...,cp_{a,cpmax(a)}\} \text{ mit } \min_{w \in dom(a)} w = cp_{a,1} < ... < cp_{a,cpmax(a)} = \max_{w \in dom(a)} w.$$

Dann kann für $cpmax(a) \geq 3$ die Domäne $dom(a)$ in die folgenden $cpmax(a)-1$ disjunkten Intervalle aufgeteilt werden:

$$I_1 := [cp_{a,1}; cp_{a,2});$$
$$I_2 := [cp_{a,2}; cp_{a,3});$$
$$\vdots$$
$$I_{cpmax(a)-2} := [cp_{a,cpmax(a)-2}; cp_{a,cpmax(a)-1})$$
$$I_{cpmax(a)-1} := [cp_{a,cpmax(a)-1}; cp_{a,cpmax(a)}].$$

Bis auf das letzte Intervall sind alle Intervalle rechts offen, d.h. die obere Intervallgrenze gehört nicht mehr zum Intervall. Die Intervalle können, wie das folgende Beispiel zeigt, zur Diskretisierung der Datenbasis genutzt werden.

*Gegeben sei die Datenbasis im linken Teil von Tabelle 5-7. Die Menge erklärender Attribute sei $C = \{Artikel.Bezeichnung, Kassenbon.Uhrzeit\}$, und die Menge zu erklärender Attribute sei $D = \{Artikel.Bezeichnung\}$.*

[405] Der Rough-Set-Ansatz wurde in Abschnitt 2.2.2.3.4 eingeführt.

| *Artikel. Bezeichnung* | *Kassenbon. Uhrzeit* |
|---|---|
| *Fanta* | *16:25* |
| *Kaffee* | *16:25* |
| *Bier* | *16:25* |
| *Cola* | *16:25* |
| *Chips* | *13:40* |
| *Fanta* | *13:40* |
| *Chips* | *13:12* |
| *Fanta* | *13:12* |

| *Artikel. Bezeichnung* | *Kassenbon. Uhrzeit* |
|---|---|
| *Fanta* | *[14:00; 18:00]* |
| *Kaffee* | *[14:00; 18:00]* |
| *Bier* | *[14:00; 18:00]* |
| *Cola* | *[14:00; 18:00]* |
| *Kaffee* | *[13:00; 14:00)* |
| *Fanta* | *[13:00; 14:00)* |
| *Chips* | *[13:00; 14:00)* |
| *Fanta* | *[13:00; 14:00)* |

***Tabelle 5-7: Diskretisierung der Uhrzeit***

*Zur Diskretisierung der Uhrzeit seien folgende Cutting Points in der aktuellen Lösung: $CP(Uhrzeit) = \{9:00, 13:00, 14:00, 18:00\}$. Die Diskretisierung der Uhrzeit führt zum rechten Teil der Tabelle 5-7.*

Die Vorgabe von Cutting Points bzw. Intervallen setzt voraus, daß entsprechendes Expertenwissen zur Verfügung steht, um einen stetigen Wertebereich sinnvoll in diskrete Intervalle zu transformieren. Ist diese Voraussetzung nicht gegeben, so wäre es wünschenswert, wenn das Data-Mining-Verfahren die Intervalle selbst erzeugen würde. Dies kann auch unter Verwendung des Rough-Set-Ansatzes geschehen, indem man die Intervallgrenzen eines kardinalen bzw. ordinalen Attributes in boole'sche Attribute transformiert. Folgendes Beispiel soll dies verdeutlichen:

*Eine Trainingsmenge umfasse die Attribute $A = \{Geschlecht, Einkommen\}$. Für das Attribut „Einkommen" treten in der Trainingsmenge die Werte $\{0, 615, 1235, 2500, 3200, 4000, 5230, 8900\}$ auf. Dann kann die ursprüngliche Trainingsmenge in eine Trainingsmenge der folgenden Form transformiert werden:*

| **Geschlecht** | **E≥615** | **E≥1235** | **E≥2500** | **E≥3200** | **E≥4000** | **E≥5230** | **E≥8900** |
|---|---|---|---|---|---|---|---|
| *w* | *falsch* | *falsch* | *falsch* | *falsch* | *falsch* | *falsch* | *falsch* |
| *w* | *wahr* | *falsch* | *falsch* | *falsch* | *falsch* | *falsch* | *falsch* |
| *w* | *wahr* | *wahr* | *falsch* | *falsch* | *falsch* | *falsch* | *falsch* |
| *m* | *wahr* | *wahr* | *wahr* | *falsch* | *falsch* | *falsch* | *falsch* |
| *m* | *wahr* | *wahr* | *wahr* | *wahr* | *falsch* | *falsch* | *falsch* |
| *m* | *wahr* | *wahr* | *wahr* | *wahr* | *falsch* | *falsch* | *falsch* |
| *m* | *wahr* | *wahr* | *wahr* | *wahr* | *wahr* | *falsch* | *falsch* |
| *w* | *wahr* | *wahr* | *wahr* | *wahr* | *wahr* | *wahr* | *falsch* |
| *m* | *wahr* | *wahr* | *wahr* | *wahr* | *wahr* | *wahr* | *wahr* |

***Tabelle 5-8: Transformierte Trainingsdaten***

*Dabei steht „E" für „Einkommen", „m" für „männlich" und „w" für „weiblich". Zwei Datenobjekte sind ununterscheidbar im Sinne von Definition 2-29.*

Die so transformierte Tabelle kann wie bisher modifiziert werden, indem Spalten gestrichen (oder hinzugefügt) werden. Bei kardinalen Attributen entspricht das Streichen einer Spalte dem Streichen einer Intervallgrenze, so daß zwei Intervalle zusammenfallen und ein generelleres Intervall entsteht, das mehr Datenobjekte abdeckt.

*Streicht man aus der Tabelle 5-8 die Intervallgrenze „3200“, so sind nun drei Datenobjekte ununterscheidbar.*

| Geschlecht | E≥615 | E≥1235 | E≥2500 | E≥4000 | E≥5230 | E≥8900 |
|---|---|---|---|---|---|---|
| *w* | *falsch* | *falsch* | *falsch* | *falsch* | *falsch* | *falsch* |
| *w* | *wahr* | *falsch* | *falsch* | *falsch* | *falsch* | *falsch* |
| *w* | *wahr* | *wahr* | *falsch* | *falsch* | *falsch* | *falsch* |
| *m* | *wahr* | *wahr* | *wahr* | *falsch* | *falsch* | *falsch* |
| *m* | *wahr* | *wahr* | *wahr* | *falsch* | *falsch* | *falsch* |
| *m* | *wahr* | *wahr* | *wahr* | *falsch* | *falsch* | *falsch* |
| *m* | *wahr* | *wahr* | *wahr* | *wahr* | *falsch* | *falsch* |
| *w* | *wahr* | *wahr* | *wahr* | *wahr* | *wahr* | *falsch* |
| *m* | *wahr* | *wahr* | *wahr* | *wahr* | *wahr* | *wahr* |

***Tabelle 5-9: Transformierte Trainingsdaten ohne die Intervallgrenze „3200“***

Die Anzahl der Intervallgrenzen für ein Attribut kann sehr groß werden. Der Suchraum vergrößert sich, wie im nächsten Abschnitt diskutiert wird, exponentiell mit der Anzahl der Variablen bzw. Cutting Points. Diese Vergrößerung des Suchraums ist durch die Genauigkeit des Verfahrens nicht zu rechtfertigen, denn eigene Experimente mit diesem Ansatz haben gezeigt, daß viele Lösungen, die sich nur durch einige Cutting Points unterschieden, nur marginal unterschiedlich bewertet wurden. Aus diesem Grunde wird ein anderer Ansatz vorgezogen, der auf der intellektuellen Vorgabe von Wertehierarchien basiert:

Ähnlich wie in Abbildung 2-8 eine Konzepthierarchie für ein nominales Attribut aufgestellt wurde, können solche Hierarchien auch für kardinale und ordinale Attribute definiert werden (vgl. Abbildung 5-2). Dies hat zur Konsequenz, daß eine Rough-Set-Lösung nicht nur die verwendeten Variablen speichern muß, sondern auch die Hierarchieebene.

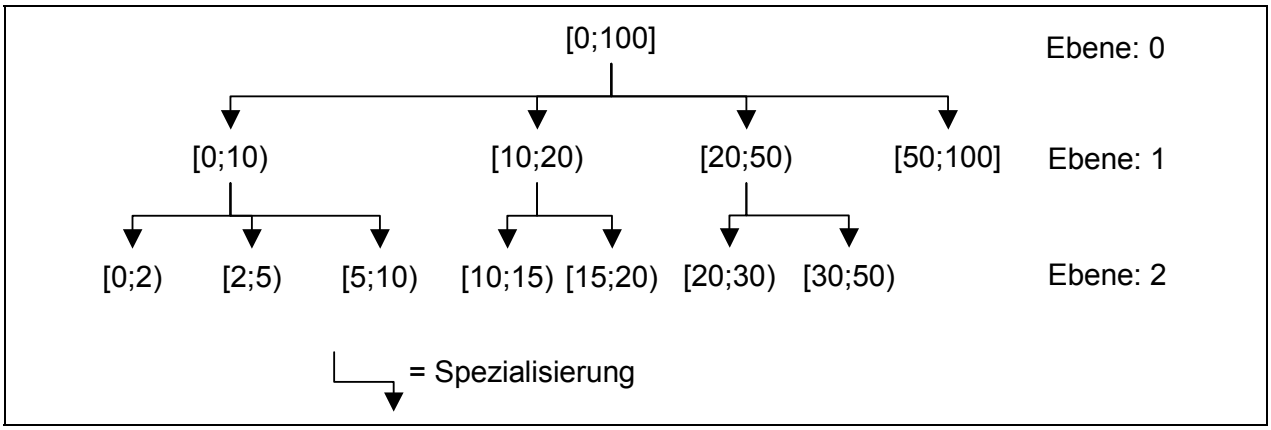



**Abbildung 5-2: Wertehierarchie für das kardinale Attribut „Alter“**

Unter Verwendung von Wertehierarchien kann eine Lösung als Punkt im Suchraum nun wie folgt definiert werden:

**Definition 5-3: Lösung im Lösungsbereich (Rough-Set-Lösung)**
Gegeben sei eine Menge von erklärenden Attributen, $C = \{ a_1^C, ..., a_{cmax}^C \}$. Weiterhin sei $himax_i$ die höchste erlaubte Hierarchieebene des Attributes $a_i^C$. Dann ist eine „*Lösung im Lösungsbereich*“ oder *„Rough-Set-Lösung“, s,* wie folgt definiert:

$s := (s[1], ..., s[cmax])$;
$s[i] \in \{0, ..., himax_i\}$: aktuelle Hierarchieebene von $a_i^C$;

$i = 1, ..., cmax$. ◊

*Beispielsweise stellt die Lösung (020110) mit der Attributmenge C = {Einkommen, Alter, Geschlecht, Beruf, Ausbildung, Sozialer Stand} und den maximalen Hierarchieebenen 3, 3, 1, 1, 1 und 1 die folgende Variablenkombination dar:*
*Alter (2. Hierarchieebene), Beruf, Ausbildung.*
*Die Variablen, die sich auf der höchsten Generalisierungsstufe (Hierarchieebene 0) befinden, können ignoriert werden, da sie die gesamte Domäne umfassen.*

Damit ergibt sich folgende Suchraum-Definition:

**Definition 5-4: Suchraum**
Es gelten dieselben Voraussetzungen wie in Definition 5-3. Weiterhin sei eine maximal erlaubte Anzahl von Klauseln in der Prämisse, $Prmax\_max$, gegeben. Dann umfaßt der *Suchraum* die Menge aller zulässigen Rough-Set-Lösungen:

$S(C, Prmax\_max) := \{ s \mid$ (*s* ist eine Rough-Set-Lösung gemäß Definition 5-3;
$\left| \{ s[i] \mid i = 1,...,cmax; \quad s[i] > 0 \} \right| \leq Prmax\_max \}$. ◊

Ein so definierter Suchraum ist sehr viel kleiner als der Lösungsraum aus Definition 2-15, da hier nur noch Attribut-Kombinationen variiert werden und nicht jede Regel einzeln. Wie aus der Attributkombination eine Regelmenge induziert wird, wurde bereits im Zusammenhang mit Rough Sets in Abschnitt 2.2.2.3.4 beschrieben.

Da auf die eine oder andere der beiden vorgeführten Möglichkeiten kardinale und ordinale Attributwerte diskretisiert werden können, wird im folgenden nicht mehr explizit auf das Skalenniveau der Attribute eingegangen – es kann unterstellt werden, daß alle Attribute diskret sind.

### 5.2.2 Kritische Diskussion des Modelltyps

Im folgenden wird der konzipierte Modelltyp einer kritischen Diskussion unterzogen, wobei auf die in Abschnitt 4.1 gestellten Anforderungen Bezug genommen wird.

Die allgemeinen Vor- und Nachteile von Regelmengen in konjunktiver Normalform waren bereits Gegenstand von Abschnitt 2.2.2.2. Da der hier vorgestellte Modelltyp an die konjunktive Normalform angelehnt ist, gilt auch für ihn, daß er

- ⇨ leicht verständlich ist,
- ⇨ Variablen unterschiedlicher Skalenniveaus repräsentieren kann,
- ⇨ von angemessener Komplexität ist und
- ⇨ leicht in ein Suchverfahren integriert werden kann.

Die **leichte Verständlichkeit** von Regeln in KNF wird durch die vorgenommene Modifikation eher verbessert, da es sich dabei um eine Einschränkung der Sprachkomplexität handelt. Rough-Set-Regelmengen können übersichtlich in Tabellenform dargestellt werden.

Der Modelltyp kann, wie gezeigt, durch intellektuelle Vorgabe von Wertehierarchien oder Cutting Points, **Variablen unterschiedlicher Skalenniveaus** repräsentieren. Dadurch, daß die Regeln kardinale und ordinale Werte zu Intervallen zusammenfassen, muß zwar durch die Repräsentation ein Informationsverlust hingenommen werden – die Intervallbildung nutzt jedoch die Ordnung der Wertebereiche aus.

*Die Ordnung spielt insofern eine Rolle, als daß nur Klauseln mit zusammenhängenden Intervallen gebildet werden, wie z.B.* `Alter∈[20;30)`*, nicht aber Klauseln mit Einzelwerten, wie z.B.* `Alter∈{20, 21, 27, 29}`*.*

Im Gegensatz zur Prädikatenlogik erlaubt die Konzeption des Modelltyps **keine Repräsentation von multirelationalen Beziehungen**. Dies hätte auch keinen Sinn, da bereits der Datenzugriff auf relationale Datenbanken an der Laufzeit der Datenbankabfragen gescheitert ist.

Ein Vorteil des attributorientierten Ansatzes gegenüber beliebigen Regelmengen in KNF ist darin zu sehen, daß jedes Modell aus dem Suchraum **eindeutig repräsentiert** ist, d.h. es können keine zwei logisch äquivalente Modelle mit unterschiedlicher Repräsentation generiert werden. Da, wie noch definiert wird, eine Lösung als Menge von Attributen repräsentiert werden kann, spielt die Reihenfolge der Attribute keine Rolle. So kann nicht etwa dieselbe Lösung durch unterschiedliche Permutationen derselben Attribute repräsentiert werden.

Die **leichte Integrierbarkeit in ein Suchverfahren** beruht darauf, daß als Züge prinzipiell nur das Hinzufügen und das Streichen von Attributen in Frage kommen. Das Auf- oder Absteigen in einer Wertehierarchie kann, wie in Abschnitt 5.4.1.1 verdeutlicht wird, ähnlich behandelt werden wie das Streichen bzw. Hinzufügen eines Attributes, dessen Domäne die Werte der jeweilgen Hierarchiebene umfaßt.

Weiterhin muß beurteilt werden, ob die zur Wissensrepräsentation verwendete **Sprache von angemessener Komplexität** ist. Während die Komplexität[406] einer Modellinstanz in einer definierten Sprache recht einfach berechnet werden kann, muß zur Berechnung der Komplexität der Sprache selbst auf die Größe des durch die Sprache beschriebenen Suchraums abgestellt werden. Um beurteilen zu können, ob die Komplexität des definierten Modelltyps „angemessen“ ist, muß untersucht werden, inwiefern eine Begrenzung der Suchraumsgröße zu lasten der Fähigkeit zur Approximation beliebiger Zusammenhänge geht.

Dadurch, daß bei dem hier verfolgten attributorientierten Ansatz eine Lösung nur aus den variablen Komponenten *s[1], ..., s[cmax]* besteht (mit *cmax*: Anzahl der erklärenden Attribute in einer Lösung), ist, wie gefordert wurde, der **Suchraum relativ zum Lösungsraum nicht zu groß**. Eine weitere Einschränkung der Suchraumgröße erfolgt durch die Vorgabe der maximal erlaubten Anzahl Klauseln in der Prämisse, *Prmax_max*.

---

[406] Vgl. die Berechnung der Einfachheit eines Datenmusters und einer Datenmuster-Menge in Abschnitt 2.2.4.3.

Derartige Einschränkungen sind in der Literatur weit verbreitet. Sie wurden mit der in Abschnitt 2.2.4.3 angesprochenen Fehlinterpretation von „Occam's razor" begründet. Als zulässiges Argument kommt neben der Verringerung der Laufzeit aber nur das Streben nach möglichst verständlichen Modellen in Frage. Ob durch diesen Ansatz betriebswirtschaftlich relevante Aussagen aus dem Suchraum ausgegrenzt werden, kann a-priori nicht festgestellt werden. A-priori verneint werden kann diese Frage für folgende Restriktion des Suchraums: In dem hier verfolgten Ansatz wird nicht mehr (wie noch in Definition 8) zugelassen, daß mehrere Nominalwerte in einer Klausel kombiniert werden, wie z.B. `Artikel ∈ {Chips, Bier, Cola}`. Vielmehr ist hier *pro Nominalklausel genau ein Nominalwert erlaubt*. Diese Einschränkung ist sinnvoll, da es bei nominalen Attributen keine inhaltlich begründeten Hinweise darauf gibt, welche Attributwerte sinnvollerweise kombiniert werden sollten.

*Beispielsweise gibt es in dem mehrfach bemühten Warenkorb-Beispiel keine sinnvolle Methode, ohne Vorwissen Artikel in einer Klausel zu kombinieren. Wenn eine hierarchische Ordnung über den Artikeln existiert, wie z.B. die Zusammenfassung von Artikeln zu Warengruppen, so kann statt des Attributes* `Artikel.Bezeichnung` *gleich das Attribut* `Artikel.Warengruppe` *in der Klausel verwendet werden.*

Die Größe des Suchraums beträgt bei $|C|$ erklärenden Variablen mit den Hierarchieebenen $0$ (d.h. Variable gehört nicht zu der Lösung) und $1$ (d.h. Variable gehört zu der Lösung):

$$|S(C, Prmax\_max)| = \sum_{k=1}^{Prmax\_max} \binom{|C|}{k}.$$

Abbildung 5-3 verdeutlicht, wie die Anzahl der Lösungen mit $k$ Variablen,

$$\binom{|C|}{k},$$

von der Anzahl der vorgegebenen Variablen, $|C|$, abhängt. Man beachte die logarithmische Skalierung der Ordinate.

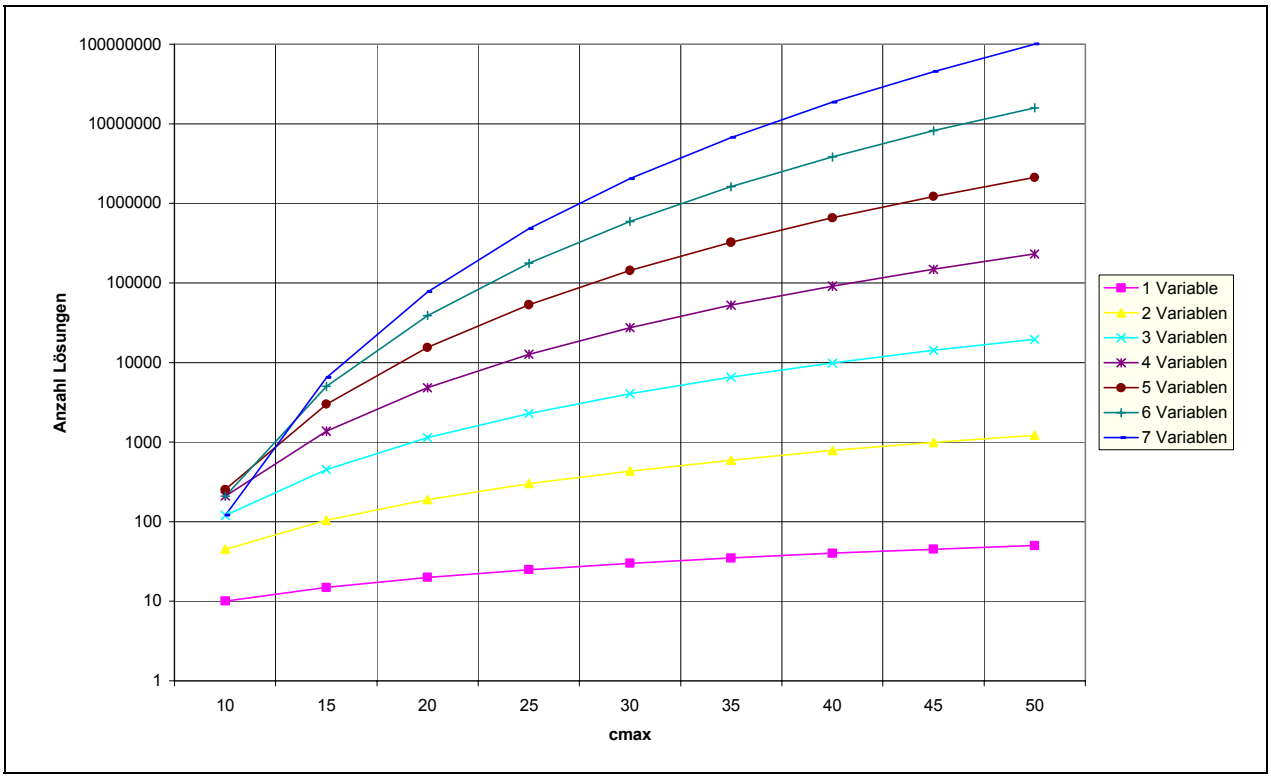



**Abbildung 5-3: Anzahl Lösungen im Suchraum (Ordinate logarithmisch skaliert!)**

Der Suchraum kann also immer noch sehr groß werden, aber eine weitere Einschränkung ist nicht mehr möglich. Die Darstellung verdeutlicht, warum im Abschnitt zuvor der Ansatz, Cutting Points geordneter Attribute wie selbständige Variablen zu betrachten und in die Variablenmenge $C$ aufzunehmen, verworfen werden mußte.

Nachdem nun die Größe des Suchraums analysiert wurde, muß zur Beurteilung, ob die Komplexität des Modelltyps angemessen ist, dessen **Fähigkeit zur Approximation** des unbekannten wahren Zusammenhangs diskutiert werden. Generell approximieren Regeln in KNF nahezu beliebige Zusammenhänge, obwohl sie nur aus achsenparallelen Schnitten durch den Merkmalsraum bestehen. Durch die Festlegung derselben Variablen und Cutting Points für die gesamte Regelmenge ist die Approximation bei attributorientierten Verfahren relativ grob, so daß dieser Modelltyp nicht für Anwendungen in Frage kommt, in denen präzise Outputs geliefert werden müssen.

*Abbildung 5-4 zeigt Beobachtungsdaten in einem zweidimensionalen Merkmalsraum. Die gestrichelten Linien deuten die achsenparallele Schnitte durch den Merkmalsraum an, die durch die vorgegebenen Intervalle entstehen. Jedes Rechteck entspricht einer potentiellen Regel der Form:*

```
WENN Einkommen ∈ [ug_Einkommen; og_Einkommen)
DANN Laufzeit ∈ [ug_Laufzeit; og_Laufzeit).
```

*Nun sind in dem konzipierten Verfahren gewisse Nebenbedingung an die Bewertung von Aussagen vorgesehen. Im Beispiel erfüllen Regeln, die weniger als fünf Beobachtungsdaten abdecken, nicht die*

*Nebenbedingung der minimalen Allgemeingültigkeit. In der Abbildung 5-4 sind die entsprechenden Beobachtungen durch Kreise mit weißer Füllung symbolisiert. Man erkennt, daß fast die Hälfte der Beobachtungen nicht durch zulässige Regeln erfaßt wird.*

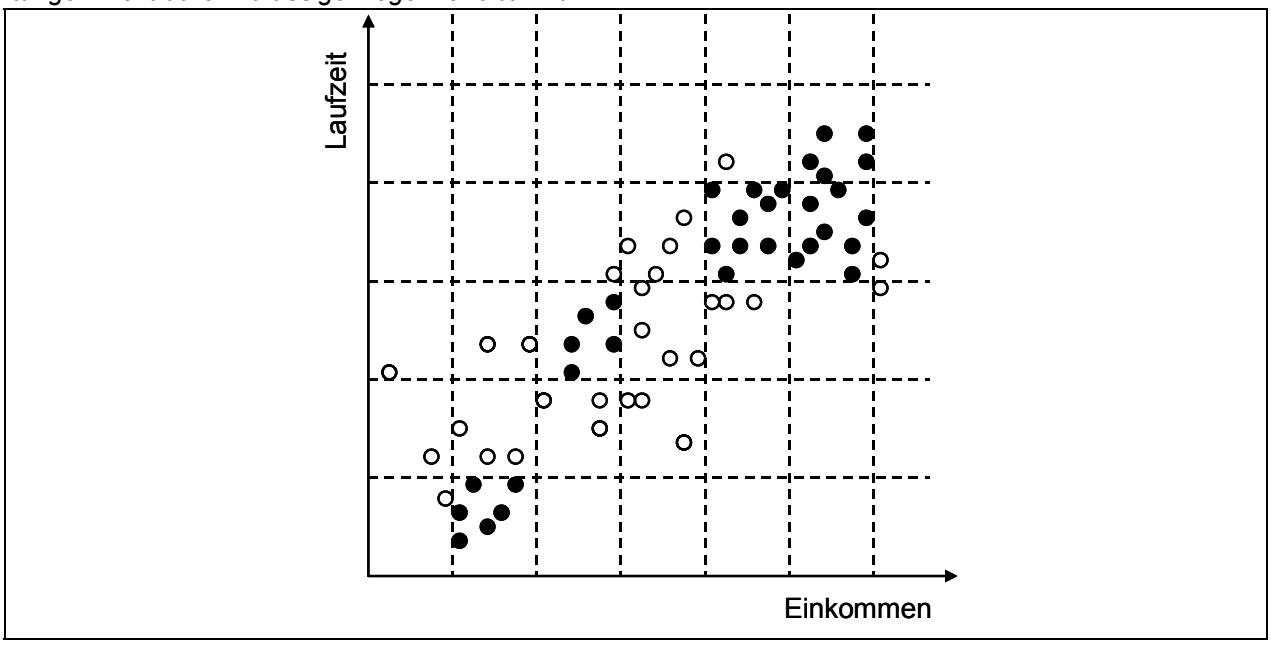



***Abbildung 5-4: Eingeschränkte Approximationsfähigkeit attributorientiert induzierter Regelmengen***

Wenn Regeln die geforderten Nebenbedingungen verletzen, so kann der durch sie abgebildete Zusammenhang nicht approximiert werden. Eine Möglichkeit, insbesondere die Nebenbedingung einer minimal zu erzielenden Allgemeingültigkeit „aufzuweichen“ und die Allgemeingültigkeit der Aussagen zu erhöhen, besteht in der Verwendung eines Fuzzy-Regelmodells. Dies wirft dann allerdings neue Probleme auf, wie z.B. eine erhöhte Laufzeitkomplexität oder die Konzeption einer geeigneten Defuzzyfizierung[407]. Genauere Modelltypen, die in den Klauseln einer Regel lineare Funktionen[408] oder konvexe Hüllen[409] zulassen, sind schwerer interpretierbar und führen zu wesentlich größeren Suchräumen und aufwendigeren Suchverfahren.

Eine weitere Möglichkeit, die Allgemeingültigkeit von Rough-Set-Regeln zu erhöhen, besteht darin, die Klauseln der Regelmenge daraufhin zu überprüfen, ob sie gestrichen werden können, ohne daß sich die Modellgüte signifikant verschlechtert.[410] Ließe man das Streichen einzelner Klauseln zu, so wäre der Modelltyp allgemeiner und

[407] Vgl. zur Defuzzyfizierung ZIMMERMANN (1993), S. 99 ff.

[408] Vgl. QUINLAN (1992), S. 343.

[409] Vgl. NEWLANDS/WEBB (1999), S. 307.

[410] Vgl. SHAN ET AL. (1995), S. 267 f.

ausdrucksstärker als ein Entscheidungsbaum, während ohne diese Erweiterung der Entscheidungsbaum ausdrucksstärker ist.

*Man betrachte als Beispiel das attributorientierte Modell in Tabelle 5-10 und den Entscheidungsbaum in Abbildung 5-5. Beide Modelle bilden denselben Sachverhalt ab, wobei jeweils von diskreten Domänen ausgegangen wird, wie sie bspw. in der Marktforschung üblich sind.*
*Würde man nun in dem attribtorientierten Modell verlangen, daß alle Zellen durch einen diskreten Attributwert gefüllt sein müssen, so könnte der Sachverhalt nicht abgebildet werden. Der Entscheidungsbaum würde den mächtigeren Modelltyp darstellen.*
*Läßt man dagegen, wie in diesem Beispiel zu sehen, leere Zellen zu, so stellt der Rough-Set-Modelltyp die ausdrucksstärkere Variante dar. Würde man in Tabelle 5-10 bspw. noch einen beliebigen Wert aus der Spalte „Nutzung des WWW" streichen, so könnte der entsprechende Sachverhalt nicht mehr als Entscheidungsbaum modelliert werden, da an den Kanten eines Entscheidungsbaum jeweils genau ein diskreter Attributwert stehen muß.*

| Nutzung des WWW | Altersgruppe | Nutzung Paging | Geschlecht | Telefontyp | Anzahl Objekte |
|---|---|---|---|---|---|
| Nein | 44 und jünger | Ja | | Schnurlosgerät | 10 |
| Nein | 44 und jünger | Nein | | Einfaches Gerät | 90 |
| Nein | 45 und älter | | Männlich | Komfortgerät | 20 |
| Nein | 45 und älter | | Weiblich | Kombigerät | 50 |
| Ja | 44 und jünger | | | Komfortgerät | 30 |
| Ja | 45 und älter | | | Einfaches Gerät | 100 |

***Tabelle 5-10: Beispiel für ein Rough-Set-Modell mit einzelnen gestrichenen Klauseln***

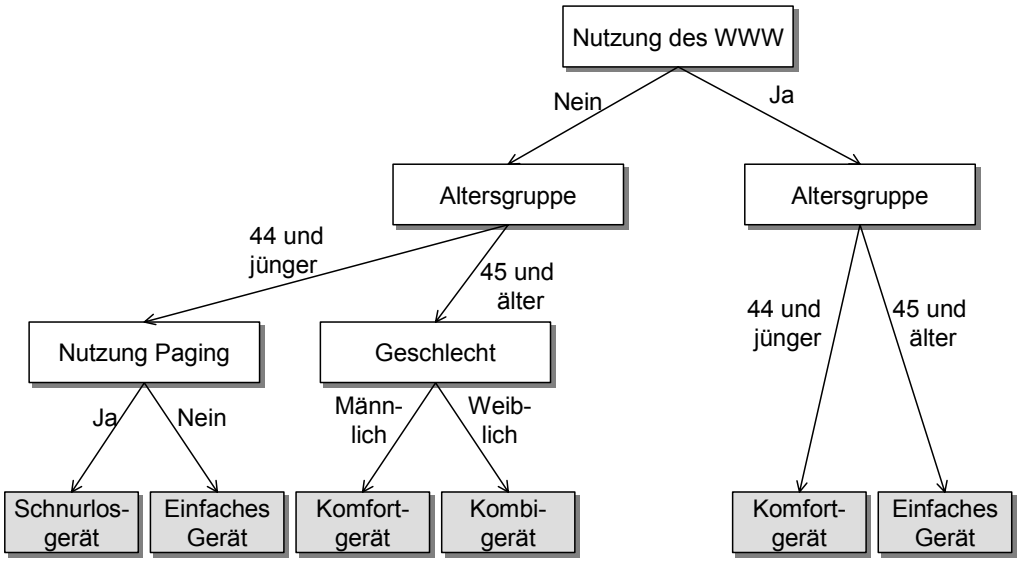



***Abbildung 5-5: Beispiel für einen Entscheidungsbaum***

Das Streichen von Klauseln aus einzelnen Regeln erhöht, wie das Beispiel zeigt, die Ausdrucksfähigkeit des Modellpyts. Auf der anderen Seite sind mit dem Streichen einzelner Klauseln folgende Probleme verbunden:

⇨ Die Modifikationen an den einzelnen Regeln würden kleine Lösungsänderungen darstellen. Trotz der geringeren Auswirkungen auf die Zielfunktionwerte müßte für diese kleinen Lösungsänderungen derselbe Zeitaufwand betrieben werden wie für das Streichen eines gesamten Attributes, da die gesamte Trainingsmenge durchlaufen werden muß, um zu prüfen, welche Objekte nach dem Streichen zusätzlich durch die generalisierte Regel erfaßt würden. Das Verhältnis von Zeitaufwand und Zusatznutzen ist demnach kritisch zu beurteilen.

⇨ Da durch das Streichen von Klauseln einzelne Regeln geändert würden, kann es dazu kommen, daß mehrere Regeln dieselben Objekte abdecken. Daher müßten Strategien zum Umgang mit Redundanzen und Widersprüchen integriert werden. Eigene Forschungserfahrungen haben gezeigt, daß dieses Problem schwer zu lösen ist. Daher geht bspw. das hier entwickelte Gütemaß zur Bewertung der Lösungen von einem eindeutigen Output, $M^{Ent}(o)$, aus.[411]

⇨ Der Suchraum vergrößert sich um ein Vielfaches, da potentiell für jede Rough-Set-Lösung mit $cmax$ erklärenden Variablen und $M$ Zeilen durch die Erweiterung $\left(\sum_{k=0}^{cmax-1}\binom{cmax}{k}\right)^{M}$ Varianten mit $0$ bis $cmax-1$ gestrichenen Klauseln pro Zeile hinzukommen. Möglicherweise müssen nicht alle diese Varianten durchgetestet werden, wenn sich eine intelligente Heuristik für dieses eigenständige Optimierungsproblem entwickeln läßt. Möglicherweise muß auch dieses Optimierungsproblem nicht für jede Rough-Set-Lösung gelöst werden, sondern nur für die besten gefundenen Rough-Set-Lösungen. Trotzdem würde, wie eigene Forschungserfahrungen gezeigt haben, eine entsprechende Heuristik für das zusätzliche Optimierungsproblem je nach Rahmenbedingungen einige Minuten bis zu einer Stunde in Anspruch nehmen. D.h. selbst wenn man diesen zusätzlichen Aufwand nur für ca. hundert gute Lösungen durchführte, käme man auf einen zusätzlichen Zeitbedarf von einigen Stunden bis Tagen, was nicht akzeptabel erscheint. Es bleibt offenbar nur die Hintereinanderausführung eines attributorientierten Verfahrens und eines auf der besten gefundenen Lösung aufsetzenden Zusatzverfahrens zum Streichen einzelner Klauseln. Damit würde das gesamte Optimierungsproblem in zwei sukzessiv zu

[411] Vgl. zur Definition des Gütemaßes Definition 3-3.

lösende Optimierungssprobleme aufgeteilt, deren Lösung sich von dem Optimum des Gesamtproblems unterscheiden kann.

Aufgrund der genannten Problempunkte wird auf die Integration einer Heuristik zum Streichen einzelner Klauseln verzichtet und die damit verbundene geringere Ausdrucksfähigkeit in Kauf genommen.

Die bisherige Diskussion zusammenfassend kann bzgl. des konzipierten Modelltyps kritisiert werden:

⇨ die eingeschränkte Approximationsfähigkeit;

⇨ die unberücksichtigte Repräsentation multirelationaler Beziehungen.

Die Stärken der Konzeption liegen vor allem in:

⇨ der Begrenzung des Suchraums,

⇨ der leichten Verständlichkeit für den Benutzer und

⇨ der leichten Integrierbarkeit in ein Suchverfahren.

## 5.3 Die ökonomische Bewertung

In diesem Abschnitt wird der im weiteren zu verwendende ökonomische Bewertungsansatz für Data-Mining-Entscheidungsmodelle bestimmt. Abschnitt 5.3.1 beschreibt die Konzeption dieses Bewertungsansatzes, und der darauffolgende Abschnitt 5.3.2 unterzieht diese Konzeption einer kritischen Betrachtung.

### 5.3.1 Konzeption der ökonomischen Bewertung

Zur weiteren Verwendung wird der in Abschnitt 3.3.2.1 eingeführte Ansatz zur ökonomischen Bewertung von Data-Mining-Entscheidungsmodellen übernommen. Damit kann das zu lösende Optimierungsproblem wie folgt formuliert werden:

**Definition 5-5: Optimierungsproblem**

Es gelten dieselben Voraussetzungen wie in den Definitionen des Abschnitts 3.3.2.1 sowie Definition 5-4. Das Optimierungsproblem lautet:

$max.\ Modellnutzwert_{\alpha}(M^{Ent})$

unter den Nebenbedingungen:

$$\alpha \in \{r \mid r \in \mathbf{R},\ 0<r<1\};$$
$$M^{Ent} \in S(C,\ Prmax\_max);$$
$$Prmax\_max \in \{1,\ ...,\ |C|\}. \qquad \Diamond$$

Demnach ist der Nutzwert eines Entscheidungsmodells zu maximieren, der nach Definition 3-3 (S. 172) als Summe der für die einzelnen Planungsobjekte zu erwartenden, vorsichtig geschätzten Zielbeiträge definiert ist.

### 5.3.2 Kritische Diskussion der ökonomischen Bewertung

Im folgenden wird der konzipierte Bewertungsansatz einer kritischen Diskussion unterzogen, wobei auf die in Abschnitt 4.3 gestellten Anforderungen Bezug genommen wird.

Die vorgestellte Konzeption orientiert sich an der in Kapitel 3 vorgenommenen **Klassifizierung betriebswirtschaftlicher Aufgabenstellungen**. Der Analytiker muß also die Bewertung nicht mehr selbst definieren, sondern nur die vorliegende Problemklasse auswählen. Aus Aufwandsgründen wird hier nur die Generierung von Entscheidungsmodellen betrachtet. Doch in Kapitel 3 wurden auch Ansätze zur Bewertung der übrigen Modelltypen vorgeschlagen.

Durch die Fokussierung auf Entscheidungsmodelle liegt ein unmittelbarer **Entscheidungsbezug** vor, denn die Bewertung orientiert sich sowohl an den erwarteten Zielbeiträgen, welche für jedes Planungsobjekt aus direkten Kosten und Erlösen ermittelt werden, als auch an dem Entscheidungsrisiko.

Weiterhin **orientiert sich das Bewertungskonzept an Abbildung 3-13** (S. 186). Und das **Modell wird als Ganzes** bewertet, so daß auch nicht erfaßte Planungsobjekte (und widersprüchliche[412]) Regeln in die Bewertung eingehen. Und die **Übertragbarkeit auf neue Datensätze** wird statistisch fundiert. Allerdings ist die Übertragbarkeit nicht automatisch sichergestellt, da die Stichprobentheorie davon ausgeht, daß zur Aufstellung der Hypothesen anderes Datenmaterial gedient hat als zu deren Überprüfung.[413]

---

[412] Dieser Aspekt ist hier irrelevant, da widersprüchliche Regeln schon durch die zuvor vorgenommene Modellkonzeption ausgeschlossen wurden.

[413] Vgl. BAMBERG/BAUR (1998), S. 179.

Dieser Fall ist beim maschinellen Lernen grundsätzlich zunächst nicht gegeben, da die Hypothesen anhand der Trainingsmenge aufgestellt und auch überprüft werden. Daher muß das fertig erlernte Modell auf einer unabhängigen Datenmenge überprüft werden, die streng genommen nur ein einziges Mal verwendet werden darf.[414]

Bezugnehmend auf die nächste Anforderung an das Bewertungskonzept kann konstatiert werden, daß das Bewertungsmaß **operational definiert** wurde. Der Modellnutzwert stellt die **zu optimierende Zielgröße** dar, in die auch die Übertragbarkeit auf neue Datensätze integriert wurde. Eine **zu satisfizierende Größe** ist in dieser Konzeption nicht unbedingt erforderlich.

Die **Berücksichtigung von Mehrfach-Attributen** entfiel dagegen bei der Konzeption des Interessantheitsmaßes, da multirelationale Datenmuster bereits bei der Konzeption des Datenzugriffs verworfen wurden.

Bzgl. des konzipierten Bewertungsansatzes kann kritisiert werden:

⇨ die unberücksichtigte Bewertung multirelationaler Beziehungen.

Die größten Stärken der Konzeption liegen in:

⇨ ihrer Orientierung an der Klassifizierung betriebswirtschaftlicher Aufgabenstellungen;

⇨ der Berücksichtigung von erwarteten Kosten und Erlösen sowie Risikoerwägungen bei der Entscheidungsfindung und

⇨ der statistischen Fundierung (einer vorsichtigen Schätzung der erwarteten Deckungsbeiträge).

## 5.4 Das Suchverfahren

Die folgenden beiden Abschnitte stellen die Konzeption des Suchverfahrens und die dazugehörige kritische Diskussion dar.

[414] Vgl. zu der Problematik der Bereitstellung einer unabhängigen Testdatenmenge Abschnitt 2.2.4.2.

### 5.4.1 Konzeption des Suchverfahrens

Der folgende Algorithmus gestattet einen groben Überblick über das Suchverfahren:

```
   Input: -
   Output: -
   Globale Variablen: -
1  s^alt := Initialisierung
   repeat
2      (s^neu, ig^neu) := Lokale Optimierung(s^alt)
3      if ig^neu ≤ 0 then s^neu := Diversifizierung(s^alt)
4                    else s^neu := Intensivierung(s^neu)
5      s^alt := s^neu
6  until Abbruchkriterium
```

Auf die Unterprogramme `Initialisierung`, `Lokale Optimierung`, `Diversifizierung`, `Intensivierung` und `Abbruchkriterium` wird im folgenden einzeln eingegangen.

Die **Initialisierung** (Zeile 1) erfolgt derart, daß $Prmax\_max$ Variablen zufällig-gleichverteilt ausgewählt und in die Startlösung, $s^{alt}$, übernommen werden. Die Güte der Initialisierung hat keinen besonderen Einfluß auf die Güte des weiteren Suchprozesses, so daß hier eine zufällige Variablenwahl ausreicht. Außerdem stehen zu Beginn keine Informationen über die Güte von Lösungen oder Lösungsbestandteilen zur Verfügung.

Die **lokale Optimierung** (Zeile 2) geht von einer Startlösung, $s^{alt}$, aus und generalisiert diese Schritt für Schritt, wobei alle möglichen Kombinationen von Lösungen (welche, bis auf $s^{alt}$ selbst, allesamt genereller als $s^{alt}$ sind) getestet werden. Abbildung 5-6 zeigt die Struktur des Teils-Suchraums ab $s^{alt}$. Die beste in diesem Prozeß gefundene Lösung stellt ein lokales Optimum dar, welches in $s^{neu}$ gespeichert wird. $ig^{neu}$ stellt den Interessantheitsgrad von $s^{neu}$ dar. Falls kein lokales Optimum gefunden werden kann, welches noch nicht bewertet wurde, resultiert $ig^{neu} = -1$. Falls ein lokales Optimum gefunden wird, welches keine anwendbare Regel enthält, resultiert $ig^{neu} = 0$. In beiden Fällen ist das lokale Optimum unbrauchbar, so daß in Zeile 3 eine Diversifizierung durchgeführt werden muß.[415]

[415] Das Akzeptanzkriterium, $ig^{neu} > 0$, wird im Verlauf der Testphase verbessert.

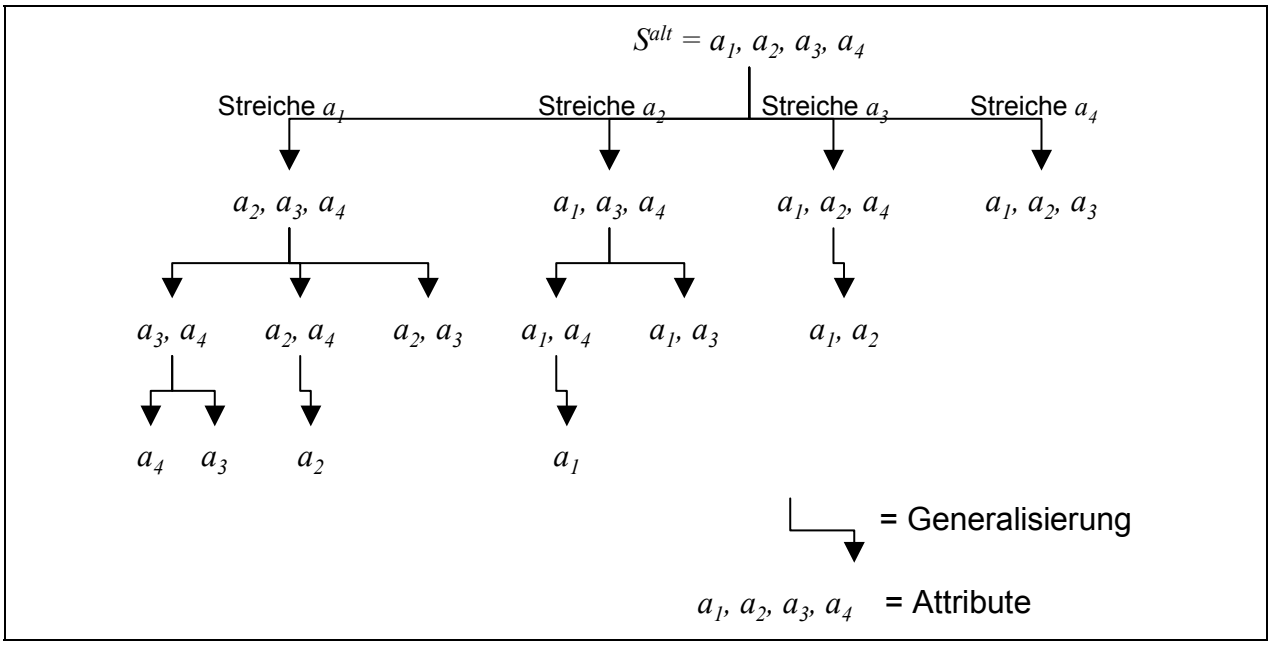



**Abbildung 5-6: Struktur des Teil-Suchraums ab $s^{alt}$**

Eine **Diversifizierung** (Zeile 3) liefert eine neue Lösung, $s^{neu}$, die von der alten Lösung, $s^{alt}$, möglichst weit entfernt liegt und bisher noch nicht besucht wurde. Möglichst weit entfernt bedeutet, daß keine Variable, die in $s^{alt}$ vorkommt, Bestandteil von $s^{neu}$ sein sollte. Dies ist notwendig, da die Diversifizierung eingesetzt wird, wenn von $s^{alt}$ aus durch lokale Optimierung nur eine unbrauchbare Lösung produziert werden konnte.

Die **Intensivierung** (Zeile 4) wird eingesetzt, wenn die Input-Lösung (das lokale Optimum, $s^{neu}$) brauchbar ist. Sie liefert eine veränderte Output-Lösung, $s^{neu}$, die aus denselben Variablen wie die Input-Lösung besteht. Zusätzlich werden soviele Variablen hinzugefügt, bis die Output-Lösung $Prmax\_max$ Variablen umfaßt. Das Hinzufügen der zusätzlichen Variablen geschieht derart, daß eine Hitliste der besten $X$ Lösungen geführt wird und solche Variablen, die zu diesen besonders guten Lösungen geführt haben, bevorzugt zur Intensivierung herangezogen werden.

In dem ungünstigen Fall, daß der Input bereits $Prmax\_max$ Variablen umfaßt, wird im nächsten Schleifendurchlauf bei der lokalen Optimierung festgestellt, daß die Lösung bereits bekannt ist. Dann resultiert $ig^{neu} = -1$, so daß im nächsten Schritt – wie beschrieben – eine Diversifizierung eingeleitet wird.

In Zeile 5 wird die neue Lösung, $s^{neu}$, für den nächsten Schleifendurchlauf zur aktuellen Startlösung, $s^{alt}$, erklärt.

Als **Abbruchkriterium** (Zeile 6) wird eine vorgegebene Anzahl von Iterationen verwendet.

Die in Abbildung 5-6 dargestellte Struktur des Suchraums legt die Anwendung eines Branch&Bound-Verfahrens[416] nahe. Die Idee von Branch&Bound-Verfahren besteht darin, in einem baumartig strukturierten Suchraum jeweils rekursiv in die Teilbäume zu verzweigen („branch“) und einen Teilbaum abzuschneiden („bound“), sobald man erkennt, daß er das Optimum nicht enthalten kann. Hierzu ist es erforderlich, für einen Teilbaum eine Obergrenze für die noch erreichbare Lösungsgüte zu ermitteln. Diese wird dann jeweils mit der besten bereits realisierten Lösungsbewertung verglichen. Ist die Obergrenze kleiner oder gleich der realisierten Bewertung, kann der Teilbaum abgeschnitten werden.

Als Obergrenze für den zuvor konzipierten Modellnutzwert liegt es zunächst nahe, die ursprüngliche Definition des Modellnutzwertes (ohne Berücksichtigung des Risikos) zu verwenden:

$$Obergrenze(M^{Ent}) = \sum_{\forall o \in O^T} dk_i + er_i \cdot e(o) \quad \text{mit} \quad i : M^{Ent}(o) = h_i ; i = 1,...,hmax.$$

Wie bereits in Abschnitt 3.3.2.1 bezeichne $hmax$ die Anzahl der Handlungsalternativen, $M^{Ent}(o)$ die Empfehlung des Entscheidungsmodells $M^{Ent}$ in der Entscheidungssituation $o$, $O^T$ die Trainingsmenge, $dk_i$ bzw. $er_i$ die direkten Kosten bzw. Erlöse bei Durchführung der Handlungsalternative $h_i$ und $e(o)$ das in der Entscheidungssituation $o$ realisierte Handlungsergebnis.

Damit wird so getan, als ob die Stichprobe unendlich groß wäre, so daß das Risiko vollständig eliminiert wäre. Tatsächlich steigt der Stichprobenumfang bei einer Generalisierungsoperation an, so daß sich die Modellnutzwert an die Obergrenze annähert. Allerdings kann durch eine solche Generalisierungsoperation die optimale Alternative, $M^{Ent}(o)=h_i$, wechseln, so daß u.U. andere Kosten und Erlöse, $dk_i$ und $er_i$, anzusetzen wären. Dadurch kann die „Obergrenze“ auch überschritten werden, so daß sie als Obergrenze nicht in Frage kommt. Setzt man dagegen jeweils die höchstmöglichen Zielbeiträge an, dann wird – wie eigene Experimente gezeigt haben – die Obergrenze so hoch, daß ein Bound so gut wie nie stattfindet.

---

[416] Vgl. zu Branch&Bound-Verfahren : JOEREßEN/SEBASTIAN (1998), S. 163 ff.

Aus diesen Gründen kommt der Branch&Bound-Ansatz hier nicht in Frage.

#### *5.4.1.1 Die lokale Optimierung*

Die lokale Optimierung verfolgt das Ziel, ausgehend von einer Startlösung, $s^{alt}$, ein lokales Optimum, $s^{neu}$, (einschließlich seiner Bewertung, $ig^{neu}$) zu erzeugen. Dabei werden alle Lösungen generiert und bewertet, die durch Streichen von Attributen aus $s^{alt}$ erzeugt werden können (vgl. Abbildung 5-6). Ebenso werden alle Lösungen analysiert, deren Variablen sich auf einer höheren Generalisierungsebene befinden als in $s^{alt}$ (wie z.B. die „Region“ auf „Bundesland“- anstelle der „Stadt“-Ebene).

Die lokale Optimierung folgt dem nachstehenden Schema:

**Input:** $s^{alt}$ (Lösung)
**Output:** *result* (Lösung, Bewertung)
**Globale Variablen:** -

```
1   if Lösung bereits bekannt(s^alt)
2        then result := (,', –1)
         else begin
3                ig^lokal-best := –1
4                Generalisierungsprozeß(s^alt, 0)
5                result := (s^lokal-best, ig^lokal-best)
             end
```

Wenn die Input-Lösung, $s^{alt}$, bereits erzeugt und getestet wurde (Zeile 1), so wird dies dem aufrufenden Programm durch eine undefinierte Output-Lösung mit einem Interessantheitsgrad von $-1$ zu erkennen gegeben (Zeile 2).

Andernfalls wird das lokale Optimum in einem Generalisierungsprozeß gesucht, der mit der Startlösung $s^{alt}$, der Bewertung $ig^{lokal\text{-}best}$ := $-1$ (Zeile 3) und dem Index der zu generalisierenden Variable, $0$, initialisiert wird. Der Generalisierungsprozeß wird in Zeile 4 aufgerufen. Er liefert das lokale Optimum in $s^{lokal\text{-}best}$ und dessen Bewertung in $ig^{lokal\text{-}best}$. Beide Variablen werden an das aufrufende Programm zurückgeliefert (Zeile 5).

Der **Generalisierungsprozeß** wird mit den beiden Input-Parametern $s^{alt}$ und *ab_Variable* aufgerufen. Er liefert das lokale Optimum in $s^{lokal\text{-}best}$ und dessen Bewertung in $ig^{lokal\text{-}best}$. Beide Variablen sind aus Sicht des Generalisierungsprozesses global.

Der Generalisierungsprozeß läuft nach folgendem Schema ab:

```
   Input: s^alt (Lösung), ab_Variable (Ordinalzahl)
   Output: -
   Globale Variablen: s^lokal-best (Lösung), ig^lokal-best (Bewertung)
1  g := ab_Variable
2  while Generalisierung möglich(s^alt, g) do
   Begin
3      s^neu := Generalisierung durchführen(s^alt, g)
4      ig^neu := ig(s^neu)
5      if ig^neu > ig^lokal-best then
       Begin
6          ig^lokal-best := ig^neu
7          s^lokal-best := s^neu
       End
8      Generalisierungsprozeß(s^neu, g)
9      g := g+1
   End
```

Zunächst wird der Input-Parameter *ab_Variable* in der lokalen Variable $g$ festgehalten (Zeile 1). Dieser gibt jeweils die als nächstes zu generalisierende Variable an. Solange in der Lösung $s^{alt}$ ab der $g$-ten Variable noch Generalisierungen möglich sind (Zeile 2), wird folgender Schleifenrumpf durchlaufen:

Die $g$-te Generalisierung wird – wie weiter unten detailliert geschildert wird – durchgeführt und das Ergebnis in $s^{neu}$ festgehalten (Zeile 3). Anschließend wird der Interessantheitsgrad von $s^{neu}$ berechnet und in $ig^{neu}$ gespeichert (Zeile 4). Falls es sich bei $s^{neu}$ um das (vorläufige) lokale Optimum handelt (Zeile 5), wird es in $s^{lokal\text{-}best}$ und seine Bewertung in $ig^{lokal\text{-}best}$ zwischengespeichert (Zeilen 6-7).

Der Generalisierungsprozeß wird nun rekursiv weitergeführt – und zwar mit der bereits generalisierten Lösung, $s^{neu}$, und der $g$-ten Generalisierungsmöglichkeit als Input-Parameter (Zeile 8).

Nach der Abarbeitung der untergeordneten Generalisierungen wird der Generalisierungszähler, $g$, um eins erhöht (Zeile 9), so daß beim nächsten Schleifendurchlauf die nächste Generalisierungsmöglichkeit getestet wird.

Das Unterprogramm „**Generalisierung durchführen**“ stellt den Kern des gesamten Suchverfahrens dar und wird aus diesem Grunde hier detaillierter dargestellt.

*Man betrachte die Beispiel-Lösung in Tabelle 5-11. Die Spalten 2 bis 4 enthalten die Werte der erklärenden Attribute. Die Zeilen lassen sich unmittelbar als Regelprämissen lesen, z.B. die erste Zeile als:*

`WENN Artikelgruppe = Tourenräder UND Kundengruppe = Fachhandel UND Region = Süd.`

*Die erste Spalte repräsentiert die Anzahl der Datensätze, die durch die jeweilige Prämisse erfaßt werden. Beispielsweise gilt für* $Pr = ((Artikelgruppe=Tourenräder) \wedge (Kundengruppe = Fachhandel) \wedge (Region = Süd))$: $|O[Pr]|=150$.

*Die letzten drei Spalten fassen die Anzahl der Datensätze zusammen, die durch eine Regel abgedeckt werden, welche als Konklusion „Umsatz = niedrig", „Umsatz = mittel" bzw. „Umsatz = hoch" besitzt. Beispielsweise gilt für* $Pr = ((Artikelgruppe=Tourenräder) \wedge (Kundengruppe = Fachhandel) \wedge (Region = Süd))$: $|O[Pr \wedge (Umsatz=niedrig)]|=120$. *Um – wie unten gezeigt wird – eine schnelle Ausführung der Generalisierungsprozedur zu erlauben, wird nicht nur eine Spalte für die Konklusionen geführt, sondern eine pro möglicher Ausprägung des Konklusionsattributes.*

| $\|O[Pr]\|$ | Artikel-gruppe | Kunden-gruppe | Region | $\|O[Pr \wedge (Umsatz = niedrig)]\|$ | $\|O[Pr \wedge (Umsatz = mittel)]\|$ | $\|O[Pr \wedge (Umsatz = hoch)]\|$ |
|---|---|---|---|---|---|---|
| 150 | Tourenräder | Fachhandel | Süd | 120 | 20 | 10 |
| 480 | Cityräder | Discounter | West | 400 | 30 | 50 |
| 305 | Mountainbikes | Fachhandel | Nord | 200 | 100 | 5 |
| 250 | Tourenräder | Fachhandel | Nord | 60 | 150 | 40 |
| 140 | Tourenräder | Fachhandel | Süd | 20 | 20 | 100 |
| 95 | Mountainbikes | Kaufhaus | Nord | 10 | 5 | 80 |
| 500 | Mountainbikes | Discounter | West | 450 | 50 | 0 |
| 345 | Cityräder | Discounter | West | 250 | 5 | 90 |

***Tabelle 5-11: Struktur einer Lösung***

*Zu generalisieren sei nun die Spalte „Region". Angenommen, die Generalisierungshierarchie zum Attribut „Region" sehe als Zusammenfassung der Werte „Nord", „Süd" und „West" den Wert „Deutschland gesamt" vor. Dann würden nach der Generalisierung alle Datensätze denselben Wert „Deutschland gesamt" aufweisen. Damit unterscheiden sich die Datensätze in dieser Spalte nicht mehr, so daß sie ohne Informationsverlust gelöscht werden kann. Falls die Region-Spalte noch weitere Attributwerte aufweisen würde, welche nicht zu „Deutschland gesamt" zusammengefaßt werden können (z.B. „Schweiz", „Österreich"), so würde die Region-Spalte erhalten bleiben und nur die deutschen Teilregionen durch „Deutschland gesamt" ersetzt. Das Ergebnis zeigt Tabelle 5-12.*

| $\|O[Pr]\|$ | Artikel-gruppe | Kunden-gruppe | $\|O[Pr \wedge (Umsatz = niedrig)]\|$ | $\|O[Pr \wedge (Umsatz = mittel)]\|$ | $\|O[Pr \wedge (Umsatz = hoch)]\|$ |
|---|---|---|---|---|---|
| 825 | Cityräder | Discounter | 650 | 35 | 140 |
| 500 | Mountainbikes | Discounter | 450 | 50 | 0 |
| 305 | Mountainbikes | Fachhandel | 200 | 100 | 5 |
| 95 | Mountainbikes | Kaufhaus | 10 | 5 | 80 |
| 540 | Tourenräder | Fachhandel | 200 | 190 | 150 |

***Tabelle 5-12: Lösung nach Löschen der Spalte „Region"***

Im folgenden wird der Algorithmus dargestellt, der die im vorstehenden Beispiel gezeigte Generalisierungsfunktion realisiert. Er generalisiert die *zu_generalisierende_Spalte* der Lösung *Quelle* und speichert das Ergebnis als Lösung *result*.

```
   Input: Quelle (Lösung), zu_generalisierende_Spalte (Ordinalzahl)
   Output: result (Lösung)
   Globale Variablen: -
1  for Zeile := 0 to Quelle.Anzahl_Zeilen-1 do
   Begin
2    result.Zeile := result.Binaerbaum.Zeile_suchen(Zeile, zu_generalisierende Spalte)
3    if result.Zeile = nil {Zeile nicht gefunden} then
4      result.Binaerbaum.Sortiere_ein(Zeile, zu_generalisierende_Spalte)
     else
     begin {für vorhandene Zeile: Zähler-Spalten addieren}
5      result.Zeile[0] :=result.Zeile[0] + Quelle.Zelle[0,Zeile]
6      ksQ := Quelle.erste_abhaengige_Spalte
7      for ksZ := result.erste_abhaengige_Spalte
       to result. erste_abhaengige_Spalte + result.Anzahl_Konklusionsspalten-1 do
       begin
8        result.Zeile[ksZ] := result.Zeile[ksZ] + Quelle.Zelle[ksQ, Zeile]
9        ksQ := ksQ+1
       End
     End
   End
```

Folgender Schleifenrumpf wird über alle Zeilen der Quelltabelle durchlaufen (Zeile 1):

In Zeile 2 wird die aktuelle *Zeile* der Queltabelle (ohne die *zu_generalisierende Spalte*) in der Tabelle *result* gesucht. Hierzu wird ein ausgeglichener Binärbaum verwendet, der eine Laufzeitkomplexität von $O(\log Quelle.Anzahl\_Zeilen)$ aufweist. Im schlechtesten Fall ist dies $O(\log N)$, wobei $N$ die Anzahl der Zeilen in der Trainingsmenge darstellt. Im Regelfall ist *Quelle.Anzahl_Zeilen* aber wesentlich kleiner als $N$, da die Quelltabelle – wie das Beispiel zeigt – schon stark generalisiert ist.

Falls die *Zeile* (ohne die *zu_generalisierende Spalte*) in der Tabelle *result* noch nicht existiert, ist *result.Zeile = nil*, und sie wird (ohne die *zu_generalisierende Spalte*) neu angelegt (Zeilen 3-4). Letzterer erfolgt wieder über den Binärbaum, so daß das oben zur Laufzeitkomplexität Gesagte analog gilt.

Falls die *Zeile* (ohne die *zu_generalisierende Spalte*) in der Tabelle *result* bereits existiert, kann die *i*-te Spalte dieser Zeile über *result.Zeile[i]* angesprochen werden. In Zeile 5 wird in der gefundenen Zeile der Ergebnistabelle die Spalte mit dem Index *0* – also der Zähler $|O[Pr]|$ – um den Zählerstand der Quelltabelle erhöht.

In Zeile 6 wird die erste abhängige Spalte der Quelltabelle – im Beispiel der Zähler $|O[Pr \wedge (Umsatz=niedrig)]|$ – in der Variablen *ksQ* (Konklusionsspalte der Quelltabelle) vermerkt. Der Schleifenkopf in Zeile 7 zählt (über den Zähler *ksZ*, Konklusionsspalte der Zieltabelle) die abhängigen Spalten der Ergebnistabelle durch. Dasselbe wird in Zeile 9 für *ksQ* realisiert. In Zeile 8 wird in der gefundenen Zeile der Ergebnistabelle der aktuelle Konklusionszähler um den entsprechenden Zählerstand der Quelltabelle erhöht.

#### *5.4.1.2 Die Diversifizierung*

Die Diversifizierung verfolgt das Ziel, eine neue Lösung, $s^{neu}$, zu erzeugen, die von der alten Lösung, $s^{alt}$, möglichst weit entfernt liegt und bisher noch nicht besucht wurde.

Die Diversifizierung läuft wie folgt ab:

```
   Input: s^alt (Lösung)
   Output: result (Lösung)
   Globale Variablen: -
   repeat
1      result := zufällig-diversifizierte Lösung erzeugen(s^alt)
2      if Lösung bereits bekannt(result)
3           then s^alt := Tausche eine Variable aus(s^alt)
4  until not Lösung bereits bekannt(result)
```

Zunächst wird eine vorläufige Lösung, *result*, mit *Prmax_max* Variablen erzeugt, von denen keine in der Input-Lösung, $s^{alt}$, vorkommen darf (Zeile 1). Die Auswahl der Variablen erfolgt zufällig-gleichverteilt.

Falls die vorläufige Lösung, *result*, bereits erzeugt und getestet wurde (Zeile 2), werden die ursprünglichen Anforderungen an die Diversifizierung aufgeweicht, indem die Input-Lösung, $s^{alt}$, verändert wird. Die Veränderung besteht darin, daß eine Variable in $s^{alt}$ ausgetauscht wird (Zeile 3).

Der Prozeß wiederholt sich solange, bis eine unbekannte Output-Lösung, *result*, produziert wird (Zeile 4). Theoretisch – falls alle Lösungen mit *Prmax_max* Variablen bereits

getestet wurden – kann der Prozeß in eine Endlosschleife laufen. Dieser Fall wurde durch eine maximale Anzahl durchzuführender Diversifizierungsversuche abgefangen. Da bei realistischen Problemstellungen dieser Fall nie eintritt, wurde er nicht mit in den o.g. Algorithmus aufgenommen.

#### *5.4.1.3 Die Intensivierung*

Die Intensivierung wird mit einem lokalen Optimum, $s$, als Input-Parameter aufgerufen und liefert eine zu $s$ ähnliche Output-Lösung, *result*, die aber spezieller als $s$ ist. Entstehen soll eine möglichst gute Output-Lösung, die als Ausgangspunkt für eine erneute lokale Optimierung dient. Damit stellt die erneute lokale Optimierung eine Intensivierung der Suche in der Nähe des letzten lokalen Optimum, $s$, dar.

Die Intensivierung erfolgt gemäß dem nachstehenden Schema:

```
   Input: s (Lösung)
   Output: result (Lösung)
   Globale Variablen: -
1  result := s
2  while (Anzahl Variablen in result) < Prmax_max do
   Begin
3      s^TOP := Wähle Lösung zufällig-fitneßproportional aus TOP
4      i := Wähle zufällig-gleichverteilt einen Variablenindex
        aus {1,...,cmax} mit s^TOP[i] > 0
5      result[i] := himax_i
   End
```

Zunächst wird die Input-Lösung, $s$, zur Output-Lösung, *result*, erklärt (Zeile 1). Dies ist für den Fall relevant, daß die nachfolgende Schleifeneingangsbedingung sofort falsch ist. In dieser Bedingung wird getestet, ob der Output-Lösung, *result*, noch Variablen hinzugefügt werden dürfen, ohne daß die vorgegebene Anzahl erklärender Variablen, *Prmax_max*, überschritten wird (Zeile 2). Solange dies der Fall ist, wird folgender Schleifenkörper ausgeführt:

Eine Lösung, $s^{TOP}$, wird aus der Liste der besten Lösungen, *TOP*, ausgewählt (Zeile 3):

$s^{TOP}$ := $s_i^{TOP}$ mit $TOP = \{s_1^{TOP}, s_2^{TOP}, ..., s_X^{TOP}\}$.

Die Auswahl der $i$-ten Lösung, $s_i^{TOP}$, erfolgt zufällig – aber nicht etwa gleichverteilt, sondern mit einer Wahrscheinlichkeit, $P(s_i^{TOP})$, die proportional zur relativen Güte (Fitneß) der $i$-ten Lösung in der TOP-Liste ist:

$$P(s_i^{TOP}) \sim ig(s_i^{TOP})/\Sigma_{s \in TOP}\, ig(s).$$

*Beispielsweise enthalte die TOP-Liste folgende Lösungen (mit ihren Bewertungen in Klammern), wobei eine 1 an der i-ten Position für „Lösung enthält Attribut $a_i$" und eine 0 für „Lösung enthält Attribut* $a_i$ *nicht" steht:*

$s_1^{TOP}$ = 0000000101 (20,38)
$s_2^{TOP}$ = 0000000111 (18,15)
$s_3^{TOP}$ = 0000110001 (10,01)

*In dieser Situation würde die Lösung $s_1^{TOP}$ mit einer Wahrscheinlichkeit von $P(s_1^{TOP}) = 20{,}38/(20{,}38+18{,}15+10{,}01) \approx 0{,}42$ ausgewählt. Weiter ergeben sich $P(s_2^{TOP}) = 18{,}15/(20{,}38+18{,}15+10{,}01) \approx 0{,}37$ und $P(s_3^{TOP}) = 10{,}01/(20{,}38+18{,}15+10{,}01) \approx 0{,}21$.*

Nach der Auswahl der Lösung $s^{TOP}$ wird daraus zufällig-gleichverteilt ein Attribut, $a^C{}_i$, bestimmt (Zeile 4), das der Output-Lösung, *result*, hinzugefügt wird (Zeile 5). Dabei wird die höchste Hierarchieebene, $himax_i$, festgesetzt, von der aus noch alle Generalisierungsstufen möglich sind.

Die fitneßproportionale Lösungsselektion führt dazu, daß relativ häufiger Variablen aus guten Lösungen – welches potentiell die brauchbareren Variablen darstellen – gewählt und in die neue Lösung, *result*, implantiert werden. Sowohl die fitneßproportionale Selektion als auch das Rekombinieren alter Lösungen zu einer neuen sind Ansätze, die den bereits diskutierten genetischen Algorithmen entnommen wurden. Ein wesentlicher Unterschied besteht darin, daß hier nicht aus einer großen Population „beliebiger" Lösungen gewählt wird (die annahmegemäß aus Lösungen mit guten Lösungsbestandteilen (sog. „Schemata") bestehen), sondern aus einer sehr viel kleineren Liste der besten bisher produzierten Lösungen.

### 5.4.2 Kritische Diskussion des Suchverfahrens

Das entwickelte Suchverfahren kann anhand der in Abschnitt 2.2.3 vorgestellten generellen Kontrollstruktur von Suchverfahren bezüglich seiner Initialisierung, der Auswahl zu testender Suchoperationen, des Akzeptanzkriteriums, der Auswahl der akzeptierten Lösungen und des Abbruchkriteriums diskutiert werden. Dabei wird das Suchverfahren an den im Abschnitt 4.2 aufgestellten Anforderungen gemessen:

⇨ **Initialisierung**: Die Initialisierungsfunktion erzeugt eine zufällige Lösung, die $Prmax\_max$ Variablen umfaßt (also so speziell wie möglich ist). Die Anfangslösung übt kaum einen Einfluß auf den weiteren Suchprozeß aus.

⇨ **Auswahl zu testender Suchoperationen**: Da die attributorientierte Generalisierung von Lösungen – wie weiter unten diskutiert wird – sehr effizient ist, wurde sie als Kernfunktion der Suchstrategie konzipiert. Im Rahmen der lokalen Optimierung sind daher nur Generalisierungszüge zulässig. Als Kandidat für den nächsten Zug kommt jede mögliche Generalisierung in Frage. Da sich die Zahl der Variablen pro Lösung, $Prmax\_max$, in engen Grenzen hält, stellt das Durchtesten jeder Generalisierungsmöglichkeit kein Problem dar. Die **lokalen Generalisierungsschritte** sorgen für eine vollständige Exploration eines Teilbereiches des gesamten Suchraums, der durch die Startlösung eingegrenzt wird. Durch die vollständige Exploration generiert das Verfahren zu jeder Attributkombination, $X$, auch alle Teilmengen, $Y \in Pot(X)$, so daß es nicht vorkommen kann, daß die ausgegebene Variablenkombination schlechter ist als eine ihrer Teilmengen.

Zu den lokalen Generalisierungsschritten kommen die genannten Intensivierungs- und Diversifizierungsmethoden hinzu, die jeweils nur einen einzigen Kandidaten produzieren. **Intensivierungsmethoden** sorgen für eine intensive Suche in vielversprechenden Bereichen des Suchraums. **Diversifizierungsmethoden** sorgen dafür, daß „Sackgassen“ im Suchraum verlassen werden. „**Intelligent**“ im Sinne von Abschnitt 4.2 ist die Entscheidung zwischen diesen beiden Suchoperatoren: Eine Intensivierung wird nur dann durchgeführt, wenn ihre Startlösung „brauchbar“[417] erscheint. Beide Methoden berücksichtigen die Suchhistorie: Die Diversifizierung orientiert sich an dem **Tabu-Gedächtnis** und ermittelt eine unbekannte Lösung (mit möglichst wenig zu der Input-Lösung identischen Variablen). Und die Intensivierung orientiert sich an guten Lösungen aus der **TOP-Liste** und greift daraus einzelne Bestandteile heraus. Beide Methoden werden im Laufe der Testphase noch verbessert. Leider bietet der gewählte Ansatz keine Möglichkeit, gute Lösungsbestandteile (Variablen) **kennzahlgesteuert** – also gezielt – auszuwählen. Folgende Optionen,

[417] Bisher gilt eine Startlösung als brauchbar, wenn sie besser als mit $0$ bewertet wurde. Dieses einfache Akzeptanzkriterium wird in Abschnitt 6.1.3.8 noch verbessert.

eine Kennzahl $KZ(a)$ für die Güte des Attributes $a$ zu definieren, wurden implementiert und getestet:

→ Das in Abschnitt 5.3.1 konzipierte Gütemaß bewertet eine Lösung und kommt als Kennzahl für eine einzelne Variable, $a$, nicht in Frage, da Lösungen, die aus genau einer Variablen bestehen, zumeist mit $0$ bewertet werden, d.h. es gilt für fast alle $a \in A: KZ(a)=0$.

→ Die durchschnittliche Güte einer Lösung mit der Variable $a$ kommt als Kennzahl $KZ(a)$ nicht in Frage, da sie in keinem Zusammenhang zu der Güte des Optimums steht. Die Variablen, die in der optimalen Lösung vorkommen, weisen also nicht zwangsläufig einen höheren Kennzahlwert auf.

→ Selbiges gilt für die relative Anzahl oder das durchschnittliche Ausmaß an Verschlechterungen, die durch das Streichen der Variable $a$ bewirkt wurde.

⇨ **Akzeptanz der generierten Lösungen**: Akzeptiert wird jede generierte Lösung – unabhängig davon, ob sie eine Verbesserung oder eine *Verschlechterung* darstellt. Einzige Bedingung ist, daß sie zuvor noch nicht generiert wurde, so daß *die Suche nicht in einen Zyklus geraten kann*. Dieses einfache Akzeptanzkriterium kann dazu führen, daß unnötig viele schlechte Lösungen getestet werden. Aus diesen Gründen wird im Zuge der Testphase ein verbessertes Akzeptanzkriterium entwickelt. Der Vorteil dieses Akzeptanzkriteriums liegt darin, daß es *unempfindlich gegenüber großen Änderungen der Zielfunktionswerte* ist.

⇨ **Auswahl der neuen Lösung**: Die Auswahl der neuen Lösung aus den akzeptierten Lösungen erfolgt in der lokalen Optimierungsphase derart, daß nur die beste Lösung, das lokale Optimum, ausgewählt wird. Im Rahmen der Intensivierung und der Diversifizierung wird nur eine einzige Lösung generiert, akzeptiert und ausgewählt.

⇨ **Abbruchkriterium**: Kritisch zu beurteilen ist das einfache Abbruchkriterium. Wenn stets nach einer vorgegebenen Anzahl Iterationen abgebrochen wird, dann kann es sowohl vorkommen, daß die Suche weitergeführt wird, obwohl sich in der Nutzenentwicklung schon lange nichts mehr tut – oder daß die Suche abgebrochen wird, obwohl die Nutzenentwicklung noch in vollem Gange ist. Daher wird das Abbruchkriterium im Verlauf der Anwendungen verbessert.

Damit das Verfahren auch in der Praxis akzeptiert wird, wurde viel Wert auf eine geringe Laufzeit gelegt. Während die tatsächlich realisierbaren Laufzeiten in der Testphase ermittelt werden, kann folgendes zur Laufzeit*komplexität* gesagt werden:

- ⇨ Die Funktion `Lösung bereits bekannt(`$s^{alt}$`)` wird über eine Hashstruktur in $O(1)$ realisiert. Bei der Verwendung von reinen Hashstrukturen besteht regelmäßig das Problem eines hohen Speicherbedarfs. Hier beträgt der Speicherbedarf für die Hashstruktur $2^{amax}$ Bits, wobei $amax$ die Anzahl der Attribute darstellt. In der Testphase werden Problemstellungen mit maximal $30$ Variablen bearbeitet, so daß hierfür ein Speicherbedarf von $2^{30}$ Bits = $128$ MB entsteht. Dies ist mit heutigen PCs problemlos machbar. Für größere Problemstellungen, wie etwa die in Abschnitt 6.2, wird die Hashstruktur derart angepaßt, daß jede Hashadresse eine *Liste* von Lösungseinträgen aufnehmen kann. Dadurch erhöht sich die Laufzeitkomplexität, und die Laufzeiten verhalten sich proportional zur Länge der Liste.
- ⇨ Die Prozedur `Generalisierung durchführen(`$s$`)` wird über einen ausgeglichenen Binärbaum (einen sog. „Rot-Schwarz-Baum“) realisiert, der im schlechtesten Fall $N$ mal aufgerufen wird, wobei $N$ für die Anzahl der Datensätze steht. Somit beträg die Laufzeitkomplexität $O(N \cdot log\ N)$.

Durch die Fokussierung auf den Generalisierungsprozeß werden nur wenige Lösungen mit vielen Variablen getestet. Da in jeder Iteration genau eine Startlösung mit allen $Prmax\_max$ Variablen getestet wird, werden insgesamt nur $G$ Lösungen mit allen Variablen getestet (mit $G$: Anzahl durchzuführender Iterationen). Für $amax=30$ Attribute, $Prmax\_max=5$ Prämissenklauseln und $G=100$ Iterationen ergab sich in einem Beispiel-Lauf folgendes Bild:

```
Anzahl getesteter Lösungen mit 1 Variablen: 30 von 30 zulässigen Lösungen (100%);
Anzahl getesteter Lösungen mit 2 Variablen: 393 von 435 zulässigen Lösungen (90,34%);
Anzahl getesteter Lösungen mit 3 Variablen: 879 von 4060 zulässigen Lösungen (21,65%);
Anzahl getesteter Lösungen mit 4 Variablen: 495 von 27405 zulässigen Lösungen (1,81%);
Anzahl getesteter Lösungen mit 5 Variablen: 100 von 142506 zulässigen Lösungen (0,07%).
In der Summe wurden 1897 (1,09%) der 174.436 zulässigen Lösungen getestet.
```

Es hat den Anschein, als würde es – wenn das Optimum viele Variablen umfaßt – sehr unwahrscheinlich, daß es durch das Suchverfahren gefunden wird. Dies ist aber aus zwei Gründen falsch:

1. Durch eine entsprechend große Vorgabe des Parameters $Prmax\_max$ werden entsprechend mehr Lösungen mit vielen Variablen getestet. D.h. selbst wenn man nur an Lösungen mit ein bis fünf Variablen interessiert ist, kann es sinnvoll sein,

$Prmax\_max>5$ zu wählen. Es erscheint allerdings nicht sinnvoll, *generell* mehr große Lösungen zu testen als benötigt und dies schon im Verfahren zu fixieren. Schließlich verhält sich, wenn man die bereits besuchten Lösungen vernachlässigt, die Laufzeit pro Iteration proportional zu $\sum_{k=1}^{Prmax\_max} \binom{Prmax\_max}{k}$.

2. Die Auswahl der zu testenden Lösungen ist nicht zufällig. In den Intensivierungsschritten werden Bestandteile aus Lösungen, die zuvor positiv aufgefallen sind, zu größeren Lösungen zusammengesetzt.

Zusammengefaßt sind positiv zu beurteilen:

⇨ die relativ geringe Laufzeitkomplexität;

⇨ die Vermeidung von Zyklen durch effiziente Abfrage bereits besuchter Lösungen;

⇨ die Integration von lokalen Suchschritten, Diversifizierungs- und Intensivierungsfunktionen, einem Tabu-Gedächtnis und einer Liste der besten Lösungen;

⇨ Unempfindlichkeit gegenüber großen Änderungen der Zielfunktionswerte;[418]

⇨ das Zulassen kurzfristiger Verschlechterungen.

Negativ zu beurteilen sind:

⇨ das einfache Abbruchkriterium (dies wird in der Testphase verbessert);

⇨ das einfache Akzeptanzkriterium (dies wird in der Testphase verbessert);

⇨ die fehlende Möglichkeit, Attribute kennzahlgesteuert zu selektieren;

⇨ der hohe Speicherbedarf der Hashstruktur bei sehr großen Problemstellungen (oder alternativ: die schlechtere Laufzeitkomplexität).

[418] Dagegen sind Verfahren, die auf den in Tabelle 2-10 aufgeführten Akzeptanzkriterien basieren, auf kleine Lösungs- und Zielfunktionsänderungen ausgelegt. Große Verschlechterungen der Zielfunktionswerte werden dort nicht zugelassen. Dies würde für die großen Lösungsänderungen, mit denen das hier konzipierte Verfahren umgehen muß, bedeuten, daß sehr viele Lösungswege (Suchrichtungen) verschlossen blieben.

# 6 Anwendungen des Verfahrens

In diesem Kapitel wird das eingangs aufgestellte Ziel, das entwickelte Verfahren zu evaluieren, verfolgt. Das entwickelte Verfahren wird in Abschnitt 6.1 auf künstlich erzeugte Testdaten angewendet, um es analysieren und verbessern sowie die Parameter bestmöglich einstellen zu können. Die gewonnenen Erkenntnisse werden in Abschnitt 6.2 genutzt, wo das Verfahren auf Realdaten angewendet wird.

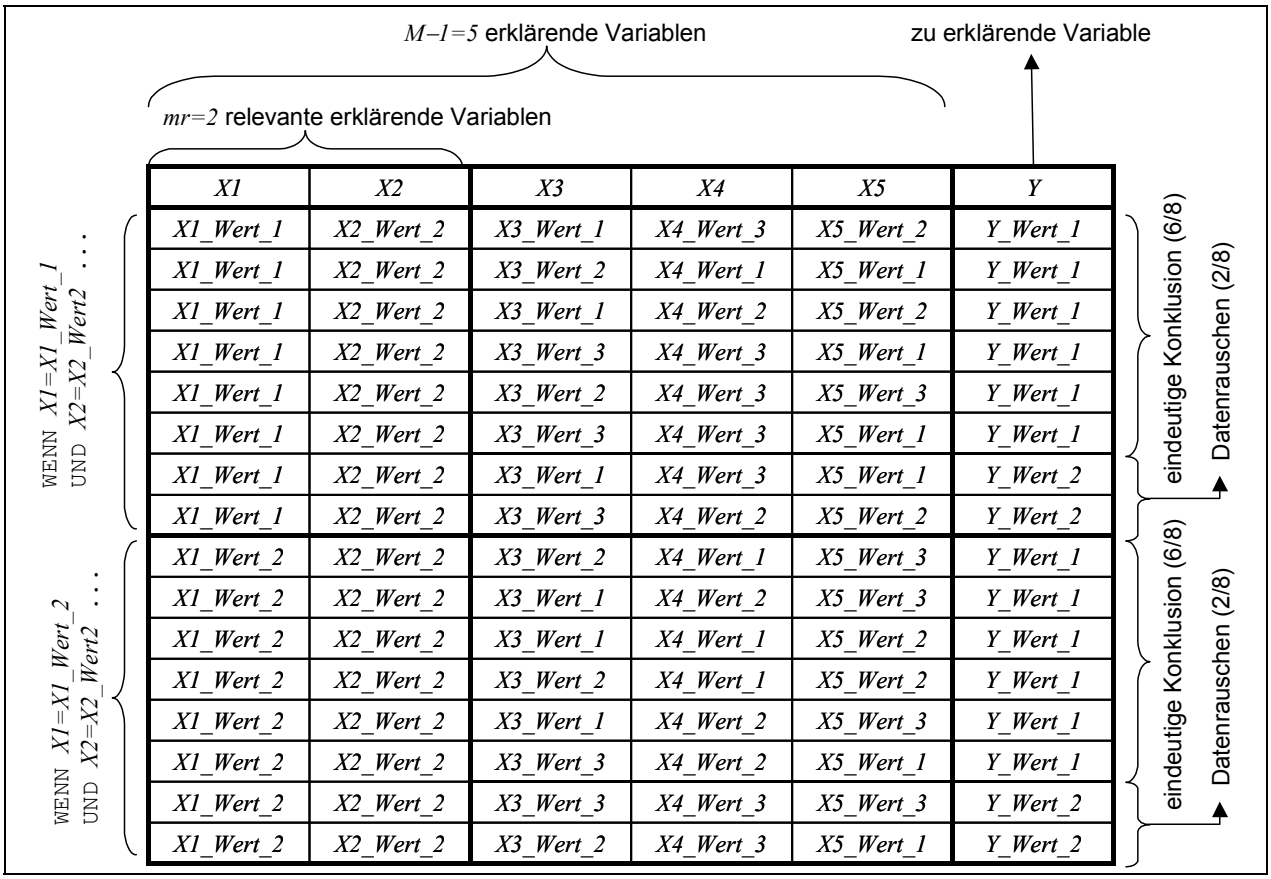
## 6.1 Anwendung des Verfahrens auf künstlich erzeugte Testdaten

### 6.1.1 Generierung der Testdaten

Das entwickelte Data-Mining-Verfahren soll anhand von künstlich erzeugten Daten getestet werden. Dies hat den Vorteil gegenüber der Verwendung von Realdaten, daß künstliche Daten sich sehr schnell erzeugen lassen, so daß das Verfahren wesentlich intensiver – anhand unterschiedlichster Datenbasen – getestet werden kann. Durch die künstliche Datengenerierung können beliebige Testszenarien erstellt werden.

Anstatt hier den Algorithmus zur Datengenerierung vorzustellen, soll im folgenden die Struktur der generierten Datenmengen diskutiert werden (vgl. Abbildung 6-1).

Die Datenbasis wird derart gebildet, daß die ersten $mr$ Variablen, $X1, ..., Xmr$, erklärungsrelevant sind, d.h. sie und nur sie sollten in den Prämissen der optimalen Regelmenge erscheinen. Die Verteilungen der übrigen erklärenden Variablen wurden so generiert, daß sie keinen Beitrag zur Erklärung der Variable $Y$ leisten. $Y$ ist die zu erklärende Variable. Sie erhält für einen bestimmten Anteil der Datensätze, die bezüglich der erklärungsrelevanten Variablen eine Prämisse bilden, denselben Wert – für die übrigen Datensätze erhält $Y$ einen anderen Wert, wodurch ein gewisses Maß an Datenrauschen entsteht.




| X1 | X2 | X3 | X4 | X5 | Y |
|---|---|---|---|---|---|
| X1_Wert_1 | X2_Wert_2 | X3_Wert_1 | X4_Wert_3 | X5_Wert_2 | Y_Wert_1 |
| X1_Wert_1 | X2_Wert_2 | X3_Wert_2 | X4_Wert_1 | X5_Wert_1 | Y_Wert_1 |
| X1_Wert_1 | X2_Wert_2 | X3_Wert_1 | X4_Wert_2 | X5_Wert_2 | Y_Wert_1 |
| X1_Wert_1 | X2_Wert_2 | X3_Wert_3 | X4_Wert_3 | X5_Wert_1 | Y_Wert_1 |
| X1_Wert_1 | X2_Wert_2 | X3_Wert_2 | X4_Wert_3 | X5_Wert_3 | Y_Wert_1 |
| X1_Wert_1 | X2_Wert_2 | X3_Wert_3 | X4_Wert_3 | X5_Wert_1 | Y_Wert_1 |
| X1_Wert_1 | X2_Wert_2 | X3_Wert_1 | X4_Wert_3 | X5_Wert_1 | Y_Wert_2 |
| X1_Wert_1 | X2_Wert_2 | X3_Wert_3 | X4_Wert_2 | X5_Wert_2 | Y_Wert_2 |
| X1_Wert_2 | X2_Wert_2 | X3_Wert_2 | X4_Wert_1 | X5_Wert_3 | Y_Wert_1 |
| X1_Wert_2 | X2_Wert_2 | X3_Wert_1 | X4_Wert_2 | X5_Wert_3 | Y_Wert_1 |
| X1_Wert_2 | X2_Wert_2 | X3_Wert_1 | X4_Wert_1 | X5_Wert_2 | Y_Wert_1 |
| X1_Wert_2 | X2_Wert_2 | X3_Wert_2 | X4_Wert_1 | X5_Wert_2 | Y_Wert_1 |
| X1_Wert_2 | X2_Wert_2 | X3_Wert_1 | X4_Wert_2 | X5_Wert_3 | Y_Wert_1 |
| X1_Wert_2 | X2_Wert_2 | X3_Wert_3 | X4_Wert_2 | X5_Wert_1 | Y_Wert_1 |
| X1_Wert_2 | X2_Wert_2 | X3_Wert_3 | X4_Wert_3 | X5_Wert_3 | Y_Wert_2 |
| X1_Wert_2 | X2_Wert_2 | X3_Wert_2 | X4_Wert_3 | X5_Wert_1 | Y_Wert_2 |

**Abbildung 6-1: Struktur der Trainingsdaten (Auszug)**

Um verschiedene Datenszenarien erzeugen zu können, wurde der Algorithmus zur Datengenerierung mit folgenden Parametern versehen:

- **Anzahl Datensätze** ($|O^T|$): Dieser Parameter beeinflußt die Laufzeit des Data-Mining-Verfahrens, spielt aber ansonsten keine Rolle.
- **Anzahl erklärender Variablen** ($amax–1$): Dieser Parameter beeinflußt die Größe des Suchraums und damit die Schwierigkeit der Problemstellung.
- **Anzahl relevanter erklärender Variablen** ($mr$): Dieser Parameter bestimmt die Anzahl der Variablen (beginnend mit $X1$), die erklärungsrelevant sind, d.h. die die Prämissen der optimalen Regelmenge bilden.
- **Sicherheit bzw. Datenrauschen** ($|O^T[Pr \wedge Ko]| / |O^T[Pr]|$): Die Anzahl der Datensätze, die die Prämissen erfüllen, sowie die Anzahl der Objekte, welche sowohl die Prämissen als auch die Konklusionen erfüllen, können separat eingestellt werden. Ihr Quotient ergibt die Interessantheitsfacette „Sicherheit“, deren Komplement das „Datenrauschen“. Je verrauschter die Daten sind, desto schwerer wird es, das

vermeintliche Optimum zu erkennen. Der Parameter $|O^T[Pr \wedge Ko]|$ beeinflußt außerdem die statistische Zuverlässigkeit der Nutzenabschätzungen. Dies wird weiter unten im Zusammenhang mit Abbildung 6-2 sowie Abbildung 6-3 noch diskutiert.

⇨ **Anzahl Werte pro erklärender Variable** ($y$): Dieser Parameter gibt die Größe der Domänen für die erklärenden Variablen an. Er wird so bestimmt, daß der kleinste ganzzahlige Wert gewählt wird, der

$$y \geq \sqrt[mr]{\frac{|O^T|}{|O^T[Pr]|}}$$

erfüllt. Dadurch bleibt gerade noch gewährleistet, daß $||O^T|/|O^T[Pr]||$ verschiedene Regeln mit $mr$ relevanten Variablen erzeugt werden können. $y$ wird minimiert, damit die Wahrscheinlichkeit möglichst klein ist, daß zufällig andere als die vorgesehenen Variablenkombinationen optimal werden.

⇨ **Anzahl Werte für die zu erklärende Variable**: Dieser Parameter gibt die Größe der Domänen für die zu erklärenden Variable, $Y$, an: $|dom(Y)|$.

⇨ **Konzentration auf das erste Handlungsergebnis** $Y\_Wert\_1$: Dieser Parameter gibt die Wahrscheinlichkeit an, mit der in dem Algorithmus zur Erzeugung der Trainingsdaten einen Wert aus der Wertemenge $dom(Y)$ auswählt. Dies führt allerdings nur bei einem Datenrauschen von $0$ zu einer relativen Häufigkeit von $95\%$ in der Trainingsmenge. Bei einem Datenrauschen von beispielsweise $10/30$ werden – falls $Y\_Wert\_1$ ausgewürfelt wurde – für jede Regel $20$ Datensätze mit $Y\_Wert\_1$ und $10$ Datensätze mit $Y\_Wert\_2$ generiert.

### 6.1.2 Aufstellung des Versuchsplans

Die Aufstellung des Versuchsplans dient dazu, die durchzuspielenden Parameterkonfigurationen soweit möglich im vorhinein festzulegen, um ein systematisches „Abtasten“ des gesamten Parameterraums zu gewährleisten. Im Einzelfall können dann bestimmte Parameterwertebereiche ausgelassen werden, wenn sich während der Tests ergibt, daß diese zu suboptimalen Ergebnissen führen. Zu unterscheiden sind die Parameter der Trainingsmengengenerierung von den Verfahrens- und Modellparametern.

Um die Anzahl der Versuche im Rahmen zu halten, werden für die verschiedenen Parameter zunächst die erlaubten Werte unabhängig voneinander festgesetzt. Das Ergebnis für die Parameter der Trainingsmengen-Generierung zeigt Tabelle 6-1.

| Parameter der Trainingsmenge | erlaubte Werte |
|---|---|
| Anzahl Datensätze $\|O^T\|$ | *1.000, 10.000, 100.000* |
| Anzahl erklärender Variablen (*amax-1*) | *5, 10, 30* |
| Anzahl relevanter erklärender Variablen (*mr*) | *2, 4, 6* |
| $\|O^T[Pr \wedge Ko]\| / \|O^T[Pr]\|$ | *20/20, 30/30, 25/30, 20/30, 30/50* |
| Anzahl Werte pro erklärender Variable ($y$) | $y \geq \sqrt[mr]{\frac{\|O^T\|}{\|O^T[Pr]\|}}$, $y$ ganzzahlig[419] |
| Anzahl Werte für die zu erklärende Variable | *2, 8* |
| Konzentration auf das erste Handlungsergebnis *(P(Y_Wert_1))* | *50%, 35%, 5%* |

**Tabelle 6-1: Erlaubte Parameterwerte für die Trainingsmenge**

Würde man versuchen, aus den erlaubten Parameterwerten alle erlaubten Trainingsmengen zu bilden, so müßten sich $3 \cdot 3 \cdot 3 \cdot 5 \cdot 2 \cdot 3 = 810$ verschiedene Trainingsmengen getestet werden. Da dies nicht durchführbar und auch nicht notwendig ist, werden daraus *12 verschiedene Trainingsmengen* ausgewählt, die unterschiedliche Anforderungen an das Data-Mining-Verfahren stellen. Die genaue Anzahl von 12 und die einzelnen Parameterwertekombinationen werden erst im Verlaufe der Versuchsdurchführungen ermittelt, um uninteressante Untersuchungen auszuschließen.

Steht die Trainingsmenge fest, so sind im zweiten Schritt die Verfahrens- und Modellparameter festzusetzen. Die Vorgehensweise ist analog zu der Festsetzung der Parameter für die Trainingsmenge. Erlaubt sind die in Tabelle 6-2 aufgeführten Parameterwerte. Den aufgeführten Konfidenzwahrscheinlichkeiten entsprechen für genügend große Stichproben approximativ die $1-\alpha$-Fraktile der Standardnormalverteilung, $c(99\%) = 2{,}326$, $c(95\%) = 1{,}645$ und $c(90\%) = 1{,}282$.[420]

[419] Die Anzahl der Werte pro erklärender Variable, $y$, wird (wie im Abschnitt zuvor diskutiert wurde) so festgesetzt, daß gerade genügend Werte verwendet werden, um die benötigte Anzahl an Regeln, $|O^T|/|O^T[Pr]|$, so zu generieren, daß sie durch *mr* Variablen unterschieden werden können.

[420] Vgl. Backhaus et al. (2000), S. 647.

| Verfahrens- und Modellparameter | erlaubte Werte |
|---|---|
| Anzahl Iterationen ($G$) | *20, 30, 50, 100, 200* |
| Länge der TOP-Liste ($X$) | *1, 3, 5, 10* |
| $min\_Anwendbarkeit \cdot \lvert O^T \rvert$ | *10, 20, 30* |
| *Prmax_max* | *5, 6, 7, 8, 9* |
| Konfidenzwahrscheinlichkeit ($1-\alpha$) | *99%, 95%, 90%* |
| Bewertung der Handlungsergebnisse[421] | Bei zwei möglichen Handlungsergebnissen:<br>bei Wahl der Alternative „Aktion durchführen“:<br>*30/–5* bzw. *100/–10* bzw. *30/–30* bzw. *30/–20*<br>bei Wahl der Alternative “keine Aktion“: *0/0*<br><br>Bei acht möglichen Handlungsergebnissen:<br>bei Wahl der Alternative „Aktion durchführen“:<br>*20/30/40/50/60/70/80/–10*<br>bei Wahl der Alternative “keine Aktion“: *0/0/0/0/0/0/0/0* |

**Tabelle 6-2: Erlaubte Werte für die Verfahrens- und Modellparameter**

Auch hier werden nicht alle $5 \cdot 4 \cdot 3 \cdot 5 \cdot 3 \cdot 5 = 4500$ Kombinationsmöglichkeiten durchgetestet, sondern pro Trainingsmenge nur einige wenige. Im folgenden wird kurz auf die einzelnen Verfahrens- und Modellparameter eingegangen:

- ⇨ **Anzahl durchzuführender Iterationen** *(G)*: Dieser Wert definiert das Abbruchkriterium für die Suche. Je mehr Iterationen durchgerechnet werden, desto mehr Lösungen werden generiert und getestet, desto länger ist die Laufzeit des Verfahrens, und desto wahrscheinlicher ist es, daß das globale Optimum gefunden wird.
- ⇨ **Länge der *TOP*-Liste** *(X)*: Dieser Wert bestimmt, wieviele der besten gefundenen Lösungen zwischengespeichert werden, um eingangs der Intensivierungsphasen aus diesen Lösungen Variablen auszuwählen, die in eine neue Startlösung implantiert werden. Je länger die TOP-Liste ist, desto kleiner ist die Wahrscheinlichkeit, daß eine gute Variablen*kombination* übernommen wird, und desto kleiner ist die Gefahr, daß immer wieder dieselbe Lösung mit denselben Variablen gewählt wird und die Suche sich in bestimmten Bereichen des Suchraums festfährt.
- ⇨ **Bewertung der Handlungsergebnisse**: Die Bewertung der Handlungsergebnisse sollten aus dem Entscheidungsmodell der zugrundeliegenden Anwendung heraus begründet sein und nicht in verschiedenen Variationen durchgespielt werden. Einzig

[421] Die Bewertung wird in der Reihenfolge der Werte angegeben, wie sie durch den Datengenerator numeriert werden: *Y_Wert_1/Y_Wert_2/...*

muß beachtet werden, ob es sich bei den gewählten Werten in Zusammenhang mit den zugrundeliegenden Daten um ein triviales Entscheidungsmodell handelt. Dies ist dann der Fall, wenn es *immer* ökonomisch sinnvoll ist, dieselbe Alternative zu wählen. Dieser Fall wird in der folgenden Testphase einmal aufgezeigt.

⇨ ***Prmax_max***: Dieser Wert legt die maximale Anzahl an Variablen bzw. Klauseln fest, die eine Regelprämisse umfassen darf. Je größer dieser Wert gewählt wird, desto größer wird der Suchraum. Da die lokale Suche stets bei einer Lösung mit $Prmax\_max$ Variablen beginnt und alle Nachfolger-Lösungen generiert, determiniert dieser Parameter auch die Anzahl der generierten Lösungen und damit die Laufzeit des Verfahrens. $Prmax\_max$ sollte immer genügend groß gesetzt werden, da pro Iteration nur eine einzige Lösung mit allen $Prmax\_max$ Variablen generiert wird. Bspw. ist es sehr unwahrscheinlich, in einem Suchraum mit $30$ Variablen das globale Optimum innerhalb von $100$ Iterationen zu finden, wenn es $5$ Variablen umfaßt und $Prmax\_max := 5$ gesetzt wird – denn in diesem Falle würden nur $100$ Lösungen mit $5$ Variablen generiert (von $5$ aus $30$ möglichen Lösungen mit $5$ Variablen).

⇨ **Minimale Anwendbarkeit**: Dieser Wert setzt die Anforderung fest, wieviele Datensätze durch die Prämisse einer Regel erfaßt werden müssen, damit die Regel zulässig ist.[422] Je größer die tatsächliche Anwendbarkeit, $|O[Pr]|/|O^T|$, ist, desto besser kann die Intervallschätzung des erwarteten Nutzens statistisch abgesichert werden.[423] Allerdings besteht bei zu großen Anforderungen an die Anwendbarkeit der Regeln die Gefahr, daß kaum noch zulässige Regeln gefunden werden. In der ökonomischen Bewertung in Abschnitt 5.3.1 wurde die Anwendbarkeit nicht explizit als Nebenbedingung in das Optimierungsproblem aufgenommen, da sie bereits in den Nutzwert einfließt; sie dient hier nur dazu, das implementierte Verfahren flexibler zu gestalten.

⇨ **Konfidenzwahrscheinlichkeit** $(1-\alpha)$: Dieser Wert gibt die Wahrscheinlichkeit vor, mit der der erwartete Nutzen einer Lösung mindestens so groß wie die aus der

---

[422] Die Anwendbarkeit wird hier absolut angegeben, $min\_Anwendbarkeit$ ist jedoch ein Wert zwischen $0$ und $1$, so daß er noch mit $|O^T|$ multipliziert werden muß.

[423] Man beachte, daß in Abschnitt 3.3.2.1 gilt: $|O[Pr]| = |O[Pr \wedge (h=h_i)]|$, da alle Datensätze aus $O[Pr]$ denselben Wert $h_i$ für die abhängige Variable $h$ zugewiesen bekommen..

Stichprobe errechnete Untergrenze ist. Der Einfluß dieses Wertes wird im folgenden diskutiert.

*Gegeben sei eine Regel, für die 50% der durch die Prämisse erfaßten Datensätze positiv regieren, was mit 30 bewertet wird; der Rest reagiert negativ, was mit –5 bewertet wird. Abbildung 6-2 zeigt für drei verschiedene Irrtumswahrscheinlichkeiten, wie sich die Untergrenze mit steigender Anwendbarkeit der Regel an den mittleren Zielbeitrag von 13 annähert. Zu beachten ist bei der Interpretation, daß die Normalverteilungsannahme erst ab ca. 30 Datensätzen gilt.*

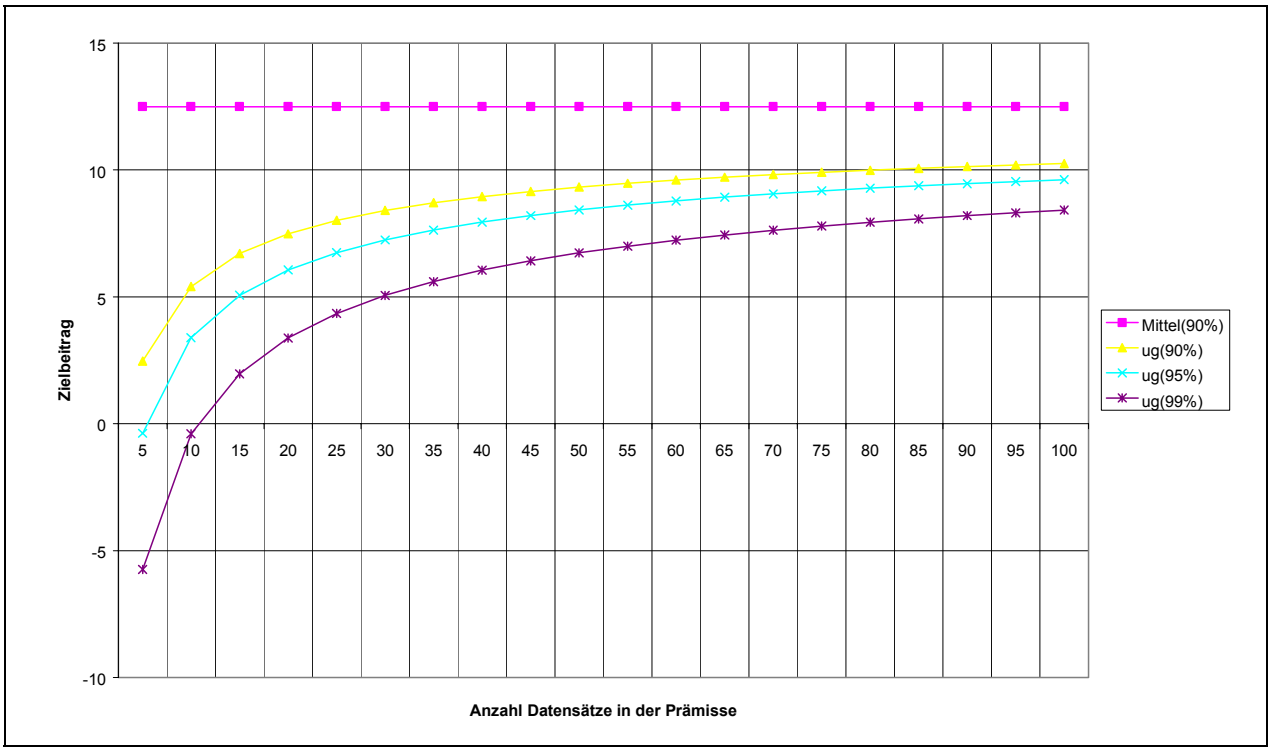



**Abbildung 6-2: Untergrenze des Zielbeitrags einer Handlungsempfehlung „Aktion durchführen“ bei einer Responsequote von *50%***

*Gegeben sei nun zum Vergleich eine weitere Regel, für die 20% der durch die Prämisse erfaßten Datensätze positiv regieren; der Rest reagiert negativ. Abbildung 6-3 zeigt für dieselben drei Irrtumswahrscheinlichkeiten wie oben, wie sich die Untergrenze mit steigender Anwendbarkeit der Regel an den mittleren Zielbeitrag von 2 annähert. Man erkennt, daß nur bei einer relativ hohen Irrtumswahrscheinlichkeit von 10% und bei einer relativ hohen Anwendbarkeit von über 80 Datensätzen ein positiver Zielbeitrag erzielt und somit die Aktion durchgeführt wird.*

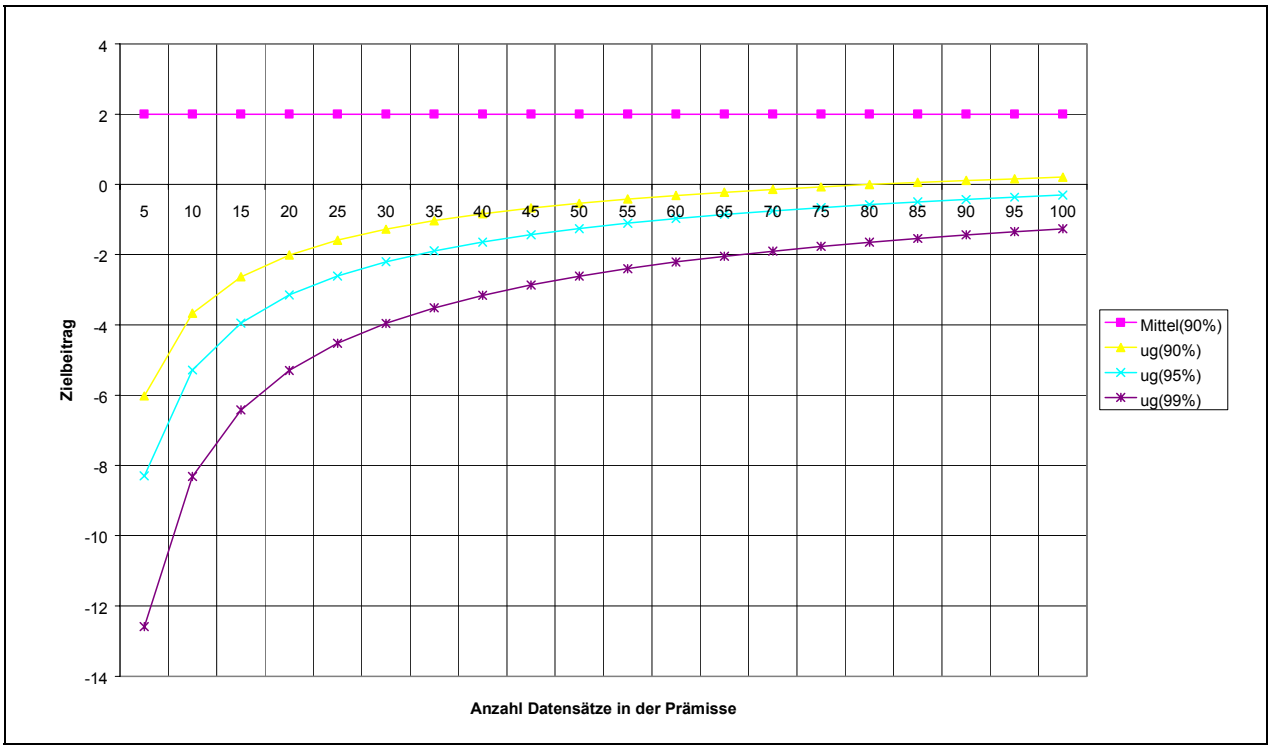



**Abbildung 6-3: Untergrenze des Zielbeitrags einer Handlungsempfehlung „Aktion durchführen“ bei einer Responsequote von *20%***

Für die Auswertung eines Verfahrensdurchlaufs von Interesse sind vor allem die folgenden **7 Ergebnisgrößen**:

⇨ die Laufzeit[424] des Verfahrens (ab Abschnitt 6.1.3.10);

⇨ die Nummer der Iteration, in der die beste Lösung gefunden wurde;

⇨ der Nutzwert der besten gefundenen Lösung;

⇨ die Anzahl der insgesamt getesteten Lösungen (im Vergleich zu der Anzahl der Lösungen im Suchraum);

⇨ die Anzahl der im Suchprozeß durchgeführten Diversifizierungen;

⇨ die Erreichung bzw. Nichterreichung des globalen Optimums.

Da das Verfahren zufallsabhängig arbeitet, genügt es nicht, pro Versuch (d.h. pro Parameterkombination) einen Testlauf durchzuführen. Daher werden *in jedem Versuch zehn Testläufe* durchgeführt und für jede Ergebnisgröße der minimale und der maximale Wert, das arithmetische Mittel und die Standardabweichung festgehalten.

[424] Als Testumgebung wurde ein PC mit einem Intel Celeron 466 MHz-Prozessor und 512 MB Hauptspeicher verwendet.

### 6.1.3 Durchführung und Interpretation der Versuche

Zur besseren Orientierung werden die benutzten Trainingsmengen bereits an dieser Stelle abgebildet, obwohl sie erst nach und nach (im Rahmen der Versuchsphase) benötigt werden. Tabelle 6-3 zeigt die Trainingsmengen im Überblick.

| Parameter der Trainingsmenge | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Anzahl Datensätze | *1.000* | | | | | | | | *10.000* | | *100.000* | *10.000* |
| Anzahl erklärender Variablen | *5* | *10* | *30* | | *10* | *30* | | | | | | |
| Anzahl relevanter erklärender Variablen | *2* | | *4* | | | *6* | | | | | | |
| $\lvert O^T[Pr \wedge Ko] \rvert / \lvert O^T[Pr] \rvert$ | *20/20* | | *30/30* | *25/30* | | *20/30* | | | | *30/50* | *25/30* | |
| Anzahl Werte pro erklärender Variable | *8* | | *3* | | | | | | *2* | *5* | *4* | *3* |
| Anzahl Werte für die zu erklärende Variable | *2* | | | | | | | | | | | *8* |
| Konzentration auf das positive Handlungsergebnis | *50%* | | | | | *35%* | | *5%* | | | | |

**Tabelle 6-3: Die zu testeten Trainingsmengen im Überblick**

Die folgenden Abschnitte beschreiben die Versuche mit den verschiedenen Trainingsmengen und die im Rahmen der Versuche für notwenig erachteten Modifikationen an dem Suchverfahren.

#### *6.1.3.1 Versuche mit Trainingsmenge 1*

Da die erste Trainingsmenge relativ geringe Anforderungen an das Data-Mining-Verfahren stellt, wurden die Verfahrens- und Modellparameter entsprechend anspruchslos gewählt. Wichtig ist hier nur die Einstellung $Prmax\_max:=5$.

Bereits kurz nach dem Start des Verfahrens lagen die Ergebnisse vor. Bereits im Bruchteil einer Sekunde konnte der gesamte Suchraum – der nur aus $\sum_{k=1}^{5}\binom{5}{k} = 31$ Lösungen besteht[425] – durchforstet und somit auch das globale Optimum gefunden werden. Daß der gesamte Suchraum betrachtet wurde, liegt daran, daß für die Startlösung des Suchprozesses laut Voreinstellung $Prmax\_max = 5$ zugelassen wurde. D.h. die Startlösung umfaßt alle fünf Variablen der Trainingsmenge, und nach Abschnitt 5.4.1.1

[425] Vgl. zur Berechnung der Suchraumgröße S. 239.

werden von der Startlösung ausgehend alle Teillösungen mit weniger Variablen betrachtet.

Aufgrund ihrer Anspruchslosigkeit wird diese Trainingsmenge nicht weiter betrachtet.

### *6.1.3.2 Versuche mit Trainingsmenge 2*

Im Vergleich zur ersten Trainingsmenge wird der Anspruch an das Suchverfahren erhöht, indem die Anzahl der erklärenden Variablen von fünf auf zehn gesteigert wird. Die Verfahrens- und Modellparameter bleiben unverändert. Tabelle 6-4 zeigt die Verfahrens- und Modellparameter sowie die Ergebnisgrößen zu diesem Versuch.

| **Verfahrens- und Modellparameter** | | **Versuch 1** |
|---|---|---|
| Anzahl Iterationen ($G$) | | *20* |
| Länge der TOP-Liste ($X$) | | *1* |
| $min\_Anwendbarkeit \cdot \lvert O^T \rvert$ | | *10* |
| $Prmax\_max$ | | *5* |
| Konfidenzwahrscheinlichkeit $(1-\alpha)$ | | *99%* |
| Bewertung der Handlungsergebnisse | | *30/–5*<br>*0/0* |
| **Ergebnisgrößen** | | |
| Iterationen bis zum Erreichen der besten Lösung | Minimum | *1* |
| | Maximum | *14* |
| | Mittel | ***4,5*** |
| | Standardabweichung | *3,77* |
| Nutzwert der besten Lösung | Minimum | *13394,06* |
| | Maximum | *13394,06* |
| | Mittel | *13394,06* |
| | Standardabweichung | *0* |
| Anzahl getesteter Lösungen | Minimum | *228* |
| | Maximum | *252* |
| | Mittel | *239,6* |
| | Standardabweichung | *7,47* |
| Anzahl Lösungen im Suchraum | | *637* |
| Erreichung des globalen Optimums | | ***100%*** |

**Tabelle 6-4: Versuch 1 zu Trainingsmenge 2**[426]

[426] Im Text referenzierte Ergebnisse werden hier und in den folgenden Tabellen fett markiert.

Man erkennt, daß das Optimum in allen zehn (*100%*) Einzeldurchläufen gefunden wurde – und zwar im Mittel bereits nach *4,5* Iterationen. So scheint auch diese Trainingsmenge noch zu anspruchslos zu sein, so daß auch bei dieser Datenbasis auf weitere Versuche verzichtet werden kann.

#### *6.1.3.3 Versuche mit Trainingsmenge 3*

Im Vergleich zur zweiten Trainingsmenge wird der Anspruch an das Suchverfahren weiter erhöht, indem die Anzahl der erklärenden Variablen von *10* auf *30* gesteigert wird. Die Verfahrens- und Modellparameter bleiben zunächst weiter unverändert. Tabelle 6-5 zeigt die Verfahrens- und Modellparameter sowie die Ergebnisgrößen zu diesem und den weiteren Versuchen.

Diese dritte Trainingsmenge führt nun dazu, daß nicht mehr in jedem Durchlauf innerhalb von *20* Iterationen das globale Optimum, „*X1, X2*“, gefunden wird, sondern nur noch in *60%* der Durchläufe (**Versuch 1**). In den übrigen *40%* der Läufe wurde entweder das lokale Optimum „*X1*“ oder das lokale Optimum „*X2*“ als beste *gefundene* Lösung ausgegeben. Daß nicht in allen Läufen das globale Optimum gefunden wurde liegt daran, daß der Suchraum nun *174.436* Lösungen umfaßt, von denen in *20* Generationen im Mittel nur knapp *500* (ca. *0,3%*) getestet werden können.

Die einfachste Möglichkeit, diese Trefferquote von *60%* zu erhöhen, besteht darin, mehr Iterationen rechnen zu lassen, so daß mehr Lösungen getestet werden. Zunächst sollen aber andere Wege beschritten werden, um den Einfluß der *TOP*-Länge, *X*, zu ergründen. So wird für den **zweiten Versuch** *X:=10* gesetzt, so daß bei Intensivierungen die Lösungen (aus denen die Startlösung für die nächste Iteration zusammengesetzt wird) aus einer TOP-*10*-Liste ausgewählt werden. Das Ergebnis ist jetzt sogar noch schlechter: Nur in *30%* der Durchläufe wurde das Optimum „*X1, X2*“ gefunden. Dies mag daran liegen, daß zehn Lösungen schon zuviel „schlechtes Material“ (also Variablen, die nicht zu absolut guten Lösungen führen) enthalten, aus denen die neue Startlösung dann zusammengesetzt wird.

Aufgrund dieser Vermutung wurde für den **dritten Versuch** *X:=5* gesetzt – und tatsächlich konnte in *80%* der Fälle das globale Optimum gefunden werden.

| **Verfahrens- und Modellparameter** | | **1** | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|---|---|
| Anzahl Iterationen ($G$) | | *20* | | | ***100*** | | |
| Länge der TOP-Liste ($X$) | | *1* | ***10*** | ***5*** | ***10*** | ***5*** | |
| $min\_Anwendbarkeit \cdot \lvert O^T \rvert$ | | *10* | | | | | |
| $Prmax\_max$ | | *5* | | | | | |
| Konfidenzwahrscheinlichkeit $(1-\alpha)$ | | *99%* | | | | | |
| Bewertung der Handlungsergebnisse | | *30/–5 0/0* | | | | | ***100/–10*** *0/0* |
| **Ergebnisgrößen** | | | | | | | |
| Iterationen bis zum Erreichen der besten Lösung | Minimum | *1* | *1* | *2* | ***2*** | ***2*** | ***6*** |
| | Maximum | *20* | *9* | *19* | ***66*** | ***70*** | ***39*** |
| | Mittel | *11,1* | *4,3* | *10,3* | *23,4* | *27,7* | *22,1* |
| | Standardabweichung | *6,11* | *3,1* | *5,71* | ***19,38*** | ***20,81*** | ***11,09*** |
| Nutzwert der besten Lösung | Minimum | *11.180* | *11.180* | *11.467* | *11.467* | *15.314* | ***51.394*** |
| | Maximum | *15.314* | *15.314* | *15.314* | *15.314* | *15.314* | ***51.394*** |
| | Mittel | *13.746* | *12.592* | *14.545* | *14.929* | *15.314* | ***51.394*** |
| | Standardabweichung | *1.921* | *1.783* | *1.538* | *1.154* | *0* | *0* |
| Anzahl getesteter Lösungen | Minimum | *490* | *474* | *475* | *1655* | *1519* | *1499* |
| | Maximum | *513* | *511* | *510* | *1754* | *1608* | *1596* |
| | Mittel | ***498*** | *494,1* | *496,4* | *1702,9* | *1557,7* | *1556,3* |
| | Standardabweichung | *7,14* | *12,9* | *10,56* | *28,88* | *27,85* | *25,85* |
| Anzahl Lösungen im Suchraum | | ***174.436*** | | | | | |
| Anzahl durchgeführter Diversifizierungen | Minimum | ***1*** | ***1*** | ***1*** | ***16*** | ***22*** | ***21*** |
| | Maximum | ***1*** | ***1*** | ***1*** | ***24*** | ***28*** | ***26*** |
| | Mittel | *1* | *1* | *1* | *20,3* | *24,4* | *24,1* |
| | Standardabweichung | *0* | *0* | *0* | *2,37* | *1,96* | *1,64* |
| Erreichung des globalen Optimums | | ***60%*** | ***30%*** | ***80%*** | ***90%*** | ***100%*** | ***100%*** |

**Tabelle 6-5: Versuche zu Trainingsmenge 3**

Man erkennt hier, daß der Parameter $X$ schwer a-priori zu bestimmen ist. Dies ist allerdings nicht so tragisch, da eine schlechte Wertewahl durch eine entsprechend höhere Iterationszahl ausgeglichen werden kann – allerdings zu Lasten der Laufzeit. Dies soll in einem **vierten Versuch** überprüft werden, indem der schlechteste $X$-Wert, nämlich zehn, in Kombination mit einer Iterationszahl von *100* festgesetzt wird. Mit dieser Parameterkonfiguration konnte in *90%* der Einzelläufe das globale Optimum gefunden werden.

Der Vollständigkeit halber werden im **fünften Versuch** *100* Iterationen mit *X=5* durchlaufen. Durch Kombination dieser beider Einstellungen, die sich als vorteilhaft erwiesen haben, konnte nun in *100%* der Einzelläufe das globale Optimum gefunden werden.

Um einmal zu testen, inwieweit die Bewertung der Handlungsergebnisse den Suchprozeß beeinflußt, wird ein **sechster Versuch** durchgeführt. Die Ergebnisse der Handlung „Aktion durchführen“ werden nun bei positiver Resonanz mit *100* und bei negativer Resonanz mit *–10* bewertet. Es stellte sich erwartungsgemäß heraus, daß sich nur die erzielten Nutzwerte geändert haben; der Suchprozeß blieb unbeeinflußt.

Da die im fünften und sechsten Versuch eingestellten Parameterkonfigurationen sicher zum Entdecken des globalen Optimums führen, kann die Versuchsreihe mit der dritten Trainingsmenge beendet werden. Ein letzter, **siebter Versuch** soll zuvor noch testen, welche Konsequenzen es hat, wenn gefordert wird, daß mindestens *30* Datensätze durch die Regeln abgedeckt werden sollen – wobei die Daten ja so generiert wurden, daß die einzelnen Äquivalenzklassen jeweils nur *20* Datensätze umfassen. Keine der vorgegebenen Regeln kann also gefunden werden. Das Verfahren liefert in diesem Versuch in allen Einzelläufen „*X1*“als beste Lösung.

Über die geschilderten Versuche hinweg kann folgendes zusammengefaßt werden:

- ⇨ Wann immer das Verfahren das globale Optimum nicht entdecken konnte, so wurde doch zumindest immer „*X1*“ oder „*X2*“ (die ja zusammen das globale Optimum „*X1, X2*“ bilden) als beste gefundene Lösung ausgegeben. Auch diese Ergebnisse sind verwertbar – wenn auch nicht optimal.
- ⇨ Es fällt auf, daß im vierten, fünften und sechsten Versuch die Suche einige Male diversifiziert werden mußte, wogegen in den ersten drei Versuchen nur die Anfangslösung diversifiziert wurde (was immer geschieht). Diversifiziert wird dann, wenn in einer lokalen Optimierungsphase keine brauchbare Lösung produziert werden kann. Dies kommt umso häufiger vor, je mehr Iterationen durchgeführt werden. Eine solche Abhängigkeit ist trivial, so daß die Anzahl durchgeführter Diversifizierungen in den folgenden Tabellen nicht mit aufgeführt wird.
- ⇨ Die Anzahl der Iterationen streut bis zur Erreichung des Optimums enorm. Dies ist bei den Versuchen vier bis sechs bereits innerhalb der zehn Testläufe eines Versuches zu beobachten und somit nicht durch die Variation der Parameterkonfiguration

zu erklären. Daher wird diese Ergebnisgröße in den folgenden Tabellen nicht mehr aufgeführt

#### *6.1.3.4 Versuche mit Trainingsmenge 4*

Die Anforderungen an das Suchverfahren sollen nun weiter gesteigert werden. Während die hohe Zahl von *30* unabhängigen Variablen beibehalten wird, wird die Anzahl der relevanten Variablen von zwei auf vier verdoppelt und $|O^T[Pr \wedge Ko]| = |O^T[Pr]| = 30$ gesetzt (also weiterhin ohne Datenrauschen).

Es wird mit den Verfahrens- und Modellparametern begonnen, mit denen die letzte Versuchsreihe abgeschlossen wurde (vgl. Tabelle 6-6). Allein die Anwendbarkeit der Regeln sollte nun mindestend *30* Datensätze umfassen. In diesem **ersten Versuch** konnte in *9* von *10* Testläufen die global-optimale Lösung „*X1, X2, X3, X4*“ gefunden werden. In einem Testlauf wurde die Lösung „*X2, X3, X4*“ ausgegeben, die nicht viel schlechter bewertet wird als das globale Optimum.

| **Verfahrens- und Modellparameter** | | **1** | **2** |
|---|---|---|---|
| Anzahl Iterationen ($G$) | | *100* | |
| Länge der TOP-Liste ($X$) | | *5* | |
| $min\_Anwendbarkeit \cdot \|O^T\|$ | | ***30*** | |
| *Prmax_max* | | *5* | ***6*** |
| Konfidenzwahrscheinlichkeit $(1-\alpha)$ | | *99%* | |
| Bewertung der Handlungsergebnisse | | *100/–10<br>0/0* | |
| **Ergebnisgrößen** | | | |
| Nutzwert der besten Lösung | Minimum | *39.995* | *39.995* |
| | Maximum | *45.238* | *45.238* |
| | Mittel | *44.713* | *44.713* |
| | Standardabweichung | *1572* | *1572* |
| Anzahl getesteter Lösungen | Minimum | *1370* | *3017* |
| | Maximum | *1468* | *3094* |
| | Mittel | *1405,2* | ***3064,1*** |
| | Standardabweichung | *26,93* | *24,48* |
| Anzahl Lösungen im Suchraum | | *174.436* | ***768.211*** |
| Erreichung des globalen Optimums | | ***90%*** | ***100%*** |

**Tabelle 6-6: Versuche zu Trainingsmenge 4**

Für einen **zweiten Versuch** wurde *Prmax_max:=6* gesetzt. Dadurch vergrößert sich der Suchraum auf *768.211* Lösungen, von denen durchschnittlich gut *3000* Lösungen getestet wurden. Da nun mehr als doppelt soviele Lösungen wie im ersten Versuch verarbeitet wurden, konnte in allen Fällen das Optimum gefunden werden. Da dieses Ergebnis nicht mehr zu verbessern ist, kann die Testreihe an dieser Stelle abgebrochen werden.

#### *6.1.3.5 Versuche mit Trainingsmenge 5*

Als nächste Steigerung der Anforderungen an das Data-Mining-Verfahren wurde eine Trainingsmenge mit Datenrauschen generiert. Jede Regel deckt *30* Datensätze ab, von denen aber nur *25* eine einheitliche Konklusion (*Y = Y_Wert_1* bzw. *Y = Y_Wert_2*) aufweisen. Die übrigen fünf Datensätze haben den jeweils anderen *Y*-Wert. Damit weist die Datenbasis ein Rauschen von $5/30 \approx 17\%$ auf.

Behält man die zuletzt gewählten Parameterwerte bei, ergeben sich im **ersten Versuch** zunächst überraschende Resultate: Nicht – wie beabsichtigt – die Lösung „*X1, X2, X3, X4*“ stellt das globale Optimum dar, sondern die Lösung „*X4*“– wobei alle Lösungen mit einer Variablen ungefähr gleichgut abschnitten. Der Grund für dieses Ergebnis liegt darin, daß die Datenbasis bisher so generiert wurde, daß *Y_Wert_1* und *Y_Wert_2* gleichverteilt vergeben wurden. Aufgrund der ungleichen Bewertung der Handlungsergebnisse (mit *100/–10* bei Durchführung einer Handlung, wie z.B. dem Versenden eines Angebotes an einen Kunden, und *0/0* bei Nichtdurchführung der Handlung) ist es nun *immer* ökonomisch sinnvoll, eine Handlung durchzuführen, da in allen Regeln genügend Datensätze vorliegen, um insgesamt einen positiven Nutzenbeitrag zu erwirtschaften. Dies gilt selbst dann, wenn *25* negative Handlungsergebnisse und nur *5* positive erzeugt werden: $5 \cdot 100 - 25 \cdot 10 = 250 > 0$. Also macht ein aufwendiges Data-Mining-Verfahren keinen Sinn, da das zugrundeliegende Entscheidungsproblem trivial ist.

Daher wird zu einem **zweiten Versuch** übergegangen, der nun eine geänderte Bewertung der Handlungsergebnisse durchspielt. Mit den ansonsten gleichgebliebenen Einstellungen konnte auf Anhieb in *100%* der Testläufe das globale Optimum „*X1, X2, X3, X4*“ gefunden werden.

Da diese Trainingsmenge dem Suchverfahren offensichtlich keine besonderen Schwierigkeiten bereitet, wird zur nächsten Trainingsmenge übergegangen.

### *6.1.3.6 Versuche mit Trainingsmenge 6*

Um die Schwierigkeit weiter zu erhöhen, wird in der sechsten Trainingsmenge ein Datenrauschen von *10/30* (ca. *33%*) erzeugt. Da sich auch hier bei Gleichverteilung der Handlungsergebnisse „*Y_Wert_1*“ und „*Y_Wert_2*“ stets eine Aktion rentieren würde (*10·30–20·10 = 100 > 0*), werden die Ergebnisse der Handlung „Aktion durchführen“ nun mit *30/–30* bewertet (*10·30–20·30 = –300 < 0*). Weiterhin wird die Trainingsmenge mit einer Konzentration von *35%* auf *Y_Wert_1* generiert.

Für diese Trainingsmenge können im **ersten Versuch** die zuletzt getesteten Verfahrens- und Modellparameter weiterverwendet werden. Wiederum konnte so auf Anhieb in *100%* der Testläufe das globale Optimum gefunden werden. Allerdings handelt es sich hier nicht um die Lösung „*X1, X2, X3, X4*“, sondern um „*X1, X3, X4*“. Dies war sogar die einzige Lösung mit positivem Interessantheitsgrad. Alle anderen Bewertungen waren gleich *0*, so daß entsprechend in jeder Iteration diversifiziert werden mußte. Dies war allerdings kein Problem des Verfahrens (das Optimum wurde ja immer gefunden), sondern ein Problem der Datenmenge, welches oben im Zusammenhang mit Abbildung 6-2 und Abbildung 6-3 bereits diskutiert wurde. Aus den Abbildungen geht auch hervor, daß eine Lösung des Problems darin bestehen kann, die Konfidenzwahrscheinlichkeit abzusenken.

So wird im **zweiten Versuch** die Konfidenzwahrscheinlichkeit auf *90%* abgesenkt. Das Problem bleibt aber noch bestehen – die Lösung „*X1, X2, X3, X4*“ hat nun einen positiven Nutzwert (insgesamt den zweithöchsten), optimal ist aber weiterhin „*X1, X3, X4*“.

Daher wird im **dritten Versuch** die Bewertung der Handlungsergebnisse bei Durchführung einer Aktion auf *30* im positiven und *–20* im negativen Falle angepaßt. Jetzt lautet das Optimum wie beabsichtigt „*X1, X2, X3, X4*“ und wird auch in *100%* der Testläufe gefunden.

Da diese Trainingsmenge dem Suchverfahren offensichtlich keine besonderen Schwierigkeiten bereitet, wird zur nächsten Trainingsmenge übergegangen.

### *6.1.3.7 Versuche mit Trainingsmenge 7*

Um die Schwierigkeit weiter zu erhöhen, wird in der siebten Trainingsmenge bei ansonsten gleichen Bedingungen die Anzahl der Variablen auf *30* erhöht.

Bei der Durchführung der Versuche traten wieder dieselben unerwünschten Effekte auf, wie im Abschnitt zuvor beschrieben wurde. Da das tatsächliche Optimum für verschiedene Bewertungsmöglichkeiten wieder nicht der beabsichtigten Lösung „*X1, X2, X3, X4*“ entspricht (und bei *30* Variablen auch kaum ausgemacht werden kann), wird diese Trainingsmenge sofort verworfen.

#### *6.1.3.8 Erste Modifikation des Verfahrens*

Durch Analyse der Protokolldateien von weniger erfolgreichen Suchläufen hat ein Problem offengelegt, welches das Suchverfahren betrifft und von der Datenbasis unabhängig ist. Die Analyse ergab zum einen, daß am Anfang der Suche über viele Iterationen jeweils Intensivierungen und keine Diversifizierung stattfinden. Zum anderen werden am Anfang kaum brauchbaren Variablen (*X1*, *X2*, *X3* oder *X4*) getestet. Somit stehen in der TOP-Liste auch keine guten Lösungen oder wenigstens Lösungen, die gute Variablen enthalten. Dadurch können die Intensivierungen auch nicht zu guten Startlösungen führen.

Gerade zu Beginn, wenn die TOP-Liste noch keine absolut guten Inhalte umfaßt, müssen mehr verschiedene Variablen durchgetestet werden. Hier sollte die Diversifizierung also eine stärkere Rolle spielen. Dies wird durch drei Änderungen im Algorithmus verwirklicht:

1. Die Diversifizierung wird in Zeile 1 (vgl. Abschnitt 5.4.1.2) so angepaßt, daß zunächst sichergestellt wird, daß alle Variablen einmal in eine Lösung eingepflanzt werden. Erst danach werden die Variablen zufällig ausgewürfelt.
2. Im Hauptprogramm wird in Zeile 3a (s.u.) veranlaßt, daß in den ersten $X$ Iterationen immer diversifiziert und nicht intensiviert wird. $X$ stellt wie bisher die gewünschte Länge der TOP-$X$-Liste dar. In Kombination mit Punkt 1 bewirkt dies, daß die TOP-$X$-Liste nach $X$ Iterationen möglichst verschiedene Variablen enthält.
3. Im Hauptprogramm wird anstelle von ($ig^{neu} > 0$) in Zeile 3a nun ein anderes Akzeptanzkriterium geprüft. Eine Lösung, $s^{neu}$, wird dann akzeptiert und intensiviert, wenn ihre Güte, $ig^{neu}$, nicht mehr als $c\%$ unter der der besten bisherigen Lösung, $Rekord_g$,

liegt. Das Akzeptanzkriterium[427] ähnelt nun dem des Record-To-Record-Travel[428], prüft aber die relative Abweichung zum Rekord. Dadurch kann der Benutzer die Konstante $c$ besser vorgeben als beim Record-To-Record-Travel, da er die absolute Höhe der Rekorde a-priori nicht kennt.

**Input: -**
**Output: -**
**Globale Variablen: -**

1 $s^{alt}$ := Initialisierung
**repeat**
2 $(s^{neu}, ig^{neu})$ := Lokale Optimierung($s^{alt}$)
3a **if** $(ig^{neu} \leq (Rekord_g \cdot (1\text{-}c/100)))$ **or** $(|TOP| \leq X)$
3b **then** $s^{neu}$ := Diversifizierung($s^{alt}$)
4 **else** $s^{neu}$ := Intensivierung($s^{neu}$)
5 $s^{alt}$ := $s^{neu}$
6 **until** Abbruchkriterium

Der neue Toleranzparameter, $c$, wurde auf $30\%$ eingestellt, d.h. solange das jeweilige lokale Optimum den bisherigen Rekord um nicht mehr als $30\%$ unterschreitet, wird intensiviert. Zunächst wird das geänderte Data-Mining-Verfahren noch einmal auf die Trainingsmenge 4[429] angewendet, um zu prüfen, ob die Verfahrensänderungen tatsächlich eine Verbesserung herbeigeführt haben. Obwohl die Parameterwerte mit denen des ersten Versuches aus Tabelle 6-6 identisch sind, konnte mit dem geänderten Verfahren statt in $90\%$ nun in $100\%$ der Fälle das Optimum gefunden werden.

#### *6.1.3.9 Versuche mit Trainingsmenge 8 und zweite Modifikation des Verfahrens*

Um das Optimum wieder eindeutig auf die Lösung „*X1, X2, X3, X4*“ einzustellen, wird eine neue Trainingsmenge erzeugt, die in $5\%$ der Datensätze *Y_Wert_1* enthält. Dies

---

[427] Man könnte eventuell auch auf die Idee kommen, das neue Akzeptanzkriterium auch im Rahmen des Generalisierungsprozesses zu prüfen. So könnte in Abbildung 5-6 ein Teilbaum ignoriert werden, dessen Startlösung (Wurzel des Teilbaums) das Akzeptanzkriterium verfehlt. Allerdings funktioniert dies auf den verwendeten Daten nicht, da bereits die obersten Lösungen (mit *Prmax_max* Variablen) zumeist mit $0$ bewertet werden und somit das Akzeptanzkriterium schon nicht erfüllen. Die weiter unten liegenden Lösungen können aber dennoch sehr gut sein. Daher wird dieser Ansatz hier nicht weiter verfolgt.

[428] Vgl. Abschnitt 2.2.3.5.

[429] Diese Trainingsmenge wurde gewählt, da sie von denjenigen Datenmengen, für die das globale Optimum sicher bekannt ist, bislang die höchsten Anforderungen an das Suchverfahren stellt.

stellt nun eher eine realistische Verteilung dar, wenn man bedenkt, daß bspw. beim Direktmarketing je nach Marketingaktion i.d.R. zwischen *1* und *5* Prozent der kontaktierten Kunden reagiert. Da der erwünschte Wert *Y_Wert_1* so selten vorkommt, wird er nun entsprechend hoch bewertet (mit *100* gegenüber *–10* für das unerwünschte Handlungsergebnis). Tabelle 6-7 beschreibt den Versuchsverlauf.

| **Verfahrens- und Modellparameter** | | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|---|
| Anzahl Iterationen ($G$) | | *100* | | | | |
| Länge der TOP-Liste ($X$) | | *5* | | | ***3*** | |
| $min\_Anwendbarkeit \cdot \lvert O^T \rvert$ | | *30* | | | | ***10*** |
| *Prmax_max* | | *5* | | | | |
| Konfidenzwahrscheinlichkeit $(1-\alpha)$ | | *90%* | | | | |
| Toleranz bei der Unterschreitung des Rekordes, $c$ | | ***30%*** | ***100%*** | ***60%*** | | |
| Bewertung der Handlungsergebnisse | | *100/–10*<br>*0/0* | | | | |
| **Ergebnisgrößen** | | | | | | |
| Nutzwert der besten Lösung | Minimum | *464,34* | *654,13* | *915,45* | *915,45* | *1469,89* |
| | Maximum | *2153,59* | *2153,59* | *2153,59* | *2153,59* | *2153,59* |
| | Mittel | *1737,04* | *1788,29* | *1769,22* | *1801,49* | *1880,11* |
| | Standardabweichung | *646,87* | *573,14* | *502,34* | *454,55* | *334,94* |
| Anzahl getesteter Lösungen | Minimum | *1881* | *1188* | *1864* | *1825* | *1753* |
| | Maximum | *1928* | *1270* | *1912* | *1928* | *1927* |
| | Mittel | *1906,6* | *1220,8* | *1889,6* | *1870,7* | *1854,9* |
| | Standardabweichung | *17,38* | *27,71* | *14,08* | *32,8* | *42,18* |
| Anzahl Lösungen im Suchraum | | *174.436* | | | | |
| Erreichung des globalen Optimums | | ***70%*** | ***70%*** | ***60%*** | ***60%*** | ***60%*** |

**Tabelle 6-7: Versuche zu Trainingsmenge 8**

Die **ersten drei Versuche** unterscheiden sich nur durch ihre Toleranzgrenze. Das Ergebnis ist jedesmal ähnlich: In *6* bzw. *7* von *10* Testläufen konnte das Optimum gefunden werden. Und wenn das Optimum nicht gefunden wurde, so doch zumindest brauchbare Lösungen, die drei der vier Variablen *X1, X2, X3* und *X4* umfassen. Erzielt wurden diese ähnlichen Ergebnisse jedoch auf etwas anderen Weise: So wurde beim ersten Versuch *95* bis *100* mal diversifiziert, beim zweiten *17* bis *24* mal und beim dritten *89* bis *98* mal. Gerade wenn eine strenge Toleranzschwelle, $c$, gewählt wird, kann das

Problem auftreten, daß – wenn einmal eine gute Lösung vorliegt – nicht mehr intensiviert, sondern nur noch diversifiziert wird.

Außerdem konnte beobachtet werden, daß sehr häufig unbrauchbare Variablen aus der *TOP*-Liste gewählt wurden. Dies wurde in einem **vierten Versuch** durch Verringerung der *TOP*-Listengröße auf drei Einträge zu verbessern versucht. Das Ergebnis ist nicht besser als vorher. Diesmal wurde die Suche *83* bis *100* mal diversifiziert.

Um eine weitere mögliche Ursache für dieses Problem auszuschließen, wurde in einem **fünften Versuch** die minimale Anwendbarkeit auf zehn Datensätze abgesenkt. Auch durch diese Maßnahme konnte das Ergebnis nicht verbessert werden. Diesmal wurde die Suche *73* bis *99* mal diversifiziert.

Da die Diversifizierung häufig benötigt wird und damit eine wichtige Rolle spielt, sollte sie verbessert werden. So könnte sich die Diversifizierung (wie die Intensivierung) auch der TOP-Liste bedienen, um Variablen in die nächste Startlösung einzupflanzen. Damit die Suche noch genügend diversifiziert wird, müßte dann mindestens einer der aus der TOP-Liste selektierten Einträge gelöscht werden. Hierzu wird die Methode `zufällig-diversifizierte Lösung erzeugen(`$s^{alt}$`)` in Zeile 1 der Diversifizierung (vgl. Abschnitt 5.4.1.2) wie folgt definiert:

**Input:** $s^{alt}$ (Lösung)

**Output:** *result* (Lösung)

**Globale Variablen:** *Variablen* (Menge aller noch nicht gewählter Variablenindices)

```
1   result = (result[1], ..., result[cmax]) := (0, 0, 0, ..., 0)
2   vorbestimmte_Anzahl := wähle zufällig-gleichverteilt eine Zahl
    aus {1, ..., Prmax_max-1}
3   while (Anzahl Variablen in result < vorbestimmte_Anzahl)
    and (|TOP| > 0) do begin
4       s := wähle zufällig-gleichverteilt eine Lösung aus TOP
5       i := wähle zufällig-gleichverteilt einen Variablenindex
        aus {1, ..., cmax} mit s[i]>0
6       result[i] := himax_i
    end
7   Entferne Lösung s aus TOP
8   while (Anzahl Variablen in result < Prmax_max) do begin
9       if Variablen ≠ ∅ then begin
10          i := wähle zufällig-gleichverteilt einen Index aus Variablen
```

```
11          Variablen := Variablen – {i}
        end else
12          i := wähle zufällig-gleichverteilt einen Variablenindex
13      result[i] := himax_i
    end
```

Nachdem eine leere Startlösung, *result*, erzeugt wurde (Zeile 1), wird eine Anzahl zwischen *1* und *Prmax_max–1* ausgewürfelt (Zeile 2). Diese Anzahl bestimmt, wieviele Lösungen zufällig-gleichverteilt aus der *TOP*-Liste ausgewählt werden, um daraus je einen Variablenindex, *i*, zu wählen und die höchstmögliche Hierarchieebene, $himax_i$, in die neue Lösung, *result*, einzupflanzen (Zeilen 3-6). Anschließend wird in Zeile 7 der zuletzt gewählte *TOP*-Eintrag freigegeben.

Mindestens eine weitere Variable wird nun auf die folgende Weise der Lösung *result* hinzugefügt: Falls es noch Variablen gibt, die bisher in keiner Lösung vorkamen, stehen deren Indices in der Menge *Variablen* (Zeile 9). In diesem Fall wird einer dieser Variablenindices selektiert (Zeile 10) und aus *Variablen* entfernt (Zeile 11). Andernfalls wird ein Variablenindex zufällig bestimmt (Zeile 12). Die auf die erste oder zweite Weise bestimmte Variable, $a^C_i$, wird anschließend auf höchster Hierarchieebene, $himax_i$, der neuen Lösung, *result*, hinzugefügt (Zeile 13).

Tabelle 6-8 zeigt den Versuchsverlauf für die geänderte Diversifizierung.

Der **erste Versuch** nach der Verfahrensmodifikation wird wieder mit unveränderten Parameterwerten durchgeführt. Das Optimum konnte diesmal statt in sechs nun in sieben von zehn Testläufen gefunden werden. Daß die Ergebnisverbesserung so gering ausfiel, mag darin liegen, daß die Suche immerhin noch *67* bis *95* mal diversifiziert wurde. Dies ist zu viel, da die Diversifizierung nicht das Ziel verfolgt, direkt eine Verbesserung der Lösungsqualität zu erzielen.

Um die Anzahl durchzuführender Diversifizierungen gering zu halten, wird in einem **zweiten Versuch** die Toleranz auf *90%* des Rekordes erhöht. Offensichtlich konnte das Optimum immerhin in *80%* der Testläufe gefunden werden, da nur *24* bis *70* Diversifizierungen durchgeführt wurden. In dem schlechtesten Testlauf (mit einem Nutzwert-Ergebnis von *654,13*) wurde nur *24* mal diversifiziert und dementsprechend häufig intensiviert. Allerdings wurden bei den Intensivierungen aus *TOP* fast nur schlechte Variablen ausgewählt, obwohl auch gute Variablen in der *TOP*-Liste enthalten waren. Offenbar

ist die *Intensivierung unzureichend*. Die Probleme und Lösungsmöglichkeiten sollen nun detaillierter diskutiert werden:

| **Verfahrens- und Modellparameter** | | **1** | **2** | **3** |
|---|---|---|---|---|
| Anzahl Iterationen ($G$) | | *100* | | |
| Länge der TOP-Liste ($X$) | | *3* | | ***1*** |
| $min\_Anwendbarkeit \cdot \lvert O^T \rvert$ | | *10* | | |
| $Prmax\_max$ | | *5* | | |
| Konfidenzwahrscheinlichkeit $(1-\alpha)$ | | *90%* | | |
| Toleranz bei der Unterschreitung des Rekordes, $c$ | | *60%* | ***90%*** | |
| Bewertung der Handlungsergebnisse | | *100/–10<br>0/0* | | |
| **Ergebnisgrößen** | | | | |
| Nutzwert der besten Lösung | Minimum | *523,04* | ***654,13*** | *2153,59* |
| | Maximum | *2153,59* | *2153,59* | *2153,59* |
| | Mittel | *1772,22* | *1879,83* | *2153,59* |
| | Standardabweichung | *626,11* | *550,63* | *0* |
| Anzahl getesteter Lösungen | Minimum | *1638* | *1216* | *1400* |
| | Maximum | *1848* | *1615* | *1606* |
| | Mittel | *1717,5* | *1413,7* | *1535* |
| | Standardabweichung | *65,85* | *115,6* | *72,96* |
| Anzahl Lösungen im Suchraum | | *174.436* | | |
| Anzahl durchgeführter Diversifizierungen | Minimum | ***67*** | ***24*** | *47* |
| | Maximum | ***95*** | ***70*** | *76* |
| | Mittel | *81,8* | *45,4* | *65,6* |
| | Standardabweichung | *9,43* | *14,52* | *9,7* |
| Erreichung des globalen Optimums | | ***70%*** | ***80%*** | ***100%*** |

**Tabelle 6-8: Versuche zu Trainingsmenge 8 mit geänderter Diversifizierung**

Zum einen führt in Zeile 3 der Intensivierung (vgl. Abschnitt 5.4.1.3) eine erfolgsproportionale Auswahlwahrscheinlichkeit nur auf lange Sicht zu erfolgreicheren Lösungsselektionen – *100* Iterationen sind hier zu kurz. Zum anderen kann auch bei Selektion einer erfolgreichen Lösung eine schlechte Variable aus dieser Lösung gewählt werden, wenn diese Wahl zufällig-gleichverteilt erfolgt.

Zumindest der erste Punkt kann durch eine reine Parametervariation getestet werden. So wird im **dritten Versuch** $X:=1$ gesetzt, so daß immer die beste bislang erzielte Lösung als Basis für die Intensivierung selektiert wird. Und tatsächlich stellt sich der

erhoffte Erfolg ein: Das Optimum konnte in allen Testläufen gefunden werden, und zwar bereits nach *8* bis *45* Iterationen..

Die gewonnenen Erkenntnisse sollen nun auch für die Verfahrenskonzeption verwendet werden. So kann die Intensivierung in zweifacher Hinsicht modifiziert werden:

- ⇨ Da der dritte Versuch aus Tabelle 6-8 mit der Einstellung $X=1$ so erfolgreich war, wird bei der Intensivierung nun stets deterministisch der beste Eintrag aus $TOP$ selektiert, um daraus Variablen für eine Intensivierung auszuwählen. Damit wird derselbe Effekt erzielt wie durch die Parametereinstellung $X=1$. Der Parameter $X$ beeinflußt dann nur noch die Diversifizierung (s.o.).
- ⇨ Objektiv unbrauchbare Lösungen mit einer Bewertung kleiner gleich *0* erhalten keinen Einzug in die $TOP$-Liste.

Der erste Verbesserungsvorschlag schlägt sich in der Intensivierungsmethode derart nieder, daß Zeile 3 wie folgt zu ersetzen und vor die Schleife zu platzieren ist:

**Input:** $s$ (Lösung)
**Output:** *result* (Lösung)
**Globale Variablen:** -

1 *result* := $s$

3[430] $s^{TOP} := s',\quad ig(s') = \max_{s'' \in TOP} ig(s'')$

2 **while** (Anzahl Variablen in *result*) < *Prmax_max* **do**
**begin**

4 $i$ := Wähle zufällig-gleichverteilt einen Variablenindex aus $\{1, ..., cmax\}$ mit $s^{TOP}[i]>0$

5 $result[i] := himax_i$

**end**

Nun ist die Wahrscheinlichkeit, eine gute Variablen*kombination* zu übernehmen, höher als in der alten Intensivierungsprozedur, da die neue Intensivierung jede in die neue Lösung einzupflanzende Variable deterministisch aus derselben Lösung, $s^{TOP}$, ermittelt.

Eine Verbesserung der Variablenselektion aus $s^{TOP}$ ist nicht möglich, wie bereits bei der kritischen Diskussion des Suchverfahrens in Abschnitt 5.4.2 verdeutlicht wurde, da hierzu keine sinnvollen Kennzahlen gebildet werden können.

---

[430] Hier wird die ursprüngliche Zeilennumerierung aus Abschnitt 5.4.1.3 beibehalten.

#### *6.1.3.10 Versuche mit Trainingsmenge 9*

Um die Anforderungen an das Data-Mining-Verfahren weiter zu erhöhen, wurde Trainingsmenge Nr. 9 generiert, die nun sechs statt vier erklärungsrelevante Variablen umfaßt. Außerdem wurde die Anzahl der Datensätze auf $N=10.000$ erhöht. Den Versuchsverlauf zeigt Tabelle 6-9. Im folgenden sollen nun auch die benötigten Laufzeiten betrachtet werden.

Im **ersten Versuch** wird mit den bislang bewährten Einstellungen begonnen. Eine Ausnahme bildet der Parameter *Prmax_max*, der mindestens so groß sein muß wie das vorgegebene Optimum, damit es überhaupt gefunden werden kann. Damit das Verfahren eine reelle Chance hat, das Optimum aufzuspüren, wird *Prmax_max:=7* gesetzt. Die Ergebnisse sind auf Anhieb recht gut: In neun von zehn Testläufen konnte das Optimum gefunden werden. Einmal wurde die kaum schlechtere Lösung „*X2, X3, X4, X5, X6*“ ausgegeben.

Im **zweiten Versuch** wird dieses Ergebnis durch eine Verdoppelung der Anzahl durchzuführender Iterationen noch zu verbessern versucht. Tatsächlich stellt sich in *100%* der Testläufe das erwünschte Optimum ein – allerdings wurde es spätestens nach *80* Iterationen gefunden, so daß die Erhöhung auf *200* Iterationen hier überflüssig war.

Im **dritten Versuch** wird getestet, ob dasselbe Ergebnis auch durch Vorgabe von *Prmax_max=8* erreicht werden kann. Die Konfidenzwahrscheinlichkeit wurde auf *95%* eingestellt, was aber nur einen Einfluß auf die Höhe der Bewertungen hat. Aufgrund der höheren Anzahl an Lösungen, die pro Generation getestet wurden, konnte das globale Optimum im Mittel bereits nach *12,6* und maximal nach *43* Iterationen gefunden werden.

Um die Laufzeitentwicklung abschätzen zu können, wurde noch ein **vierter Versuch** mit *50* Iterationen und *Prmax_max=9* durchgeführt. Der extrem große Teil-Suchraum (mit ein bis neun Variablen aus der jeweiligen Startlösung), der in jeder Iteration vollständig exploriert wird, führt zu Laufzeiten zwischen ca. *24* und *28* Minuten. Diese extreme Einstellung ist bei den hier aufgestellten Versuchen mit bis zu sechs relevanten Variablen nicht erforderlich – das Optimum kann, wie gesehen, auch anders (und schneller) gefunden werden. In dieser Testreihe wurde das Optimum nach maximal *26* Iterationen (gut *14* Minuten) gefunden.

| **Verfahrens- und Modellparameter** | | **1** | **2** | **3** | **4** |
|---|---|---|---|---|---|
| Anzahl Iterationen ($G$) | | *100* | ***200*** | *100* | ***50*** |
| Länge der TOP-Liste ($X$) | | *3* | | | |
| $min\_Anwendbarkeit \cdot \lvert O^T \rvert$ | | *20* | | | |
| *Prmax_max* | | *7* | | ***8*** | ***9*** |
| Konfidenzwahrscheinlichkeit $(1-\alpha)$ | | *90%* | | *95%* | |
| Toleranz bei der Unterschreitung des Rekordes, $c$ | | *90%* | | | |
| Bewertung der Handlungsergebnisse | | *100/–10*<br>*0/0* | | | |
| **Ergebnisgrößen** | | | | | |
| Laufzeit des Verfahrens (in Sekunden) | Minimum | *209,37* | *448,37* | *764,04* | ***1464,59*** |
| | Maximum | *251,48* | *475,28* | *918,2* | ***1671,23*** |
| | Mittel | *228,16* | *463,44* | *862,55* | *1576,95* |
| | Standardabweichung | *12,99* | *9,65* | *42,66* | *70,65* |
| Iterationen bis zum Erreichen der besten Lösung | Minimum | *7* | *7* | *7* | *8* |
| | Maximum | *90* | ***80*** | ***43*** | ***26*** |
| | Mittel | *31,9* | *27,2* | ***12,6*** | *15,3* |
| | Standardabweichung | *27,48* | *23,16* | *10,4* | *6,12* |
| Nutzwert der besten Lösung | Minimum | *15.205,48* | *22.162,2* | *17.604,08* | *17.604,08* |
| | Maximum | *22.162,2* | *22.162,2* | *17.604,08* | *17.604,08* |
| | Mittel | *21.466,53* | *22.162,2* | *17.604,08* | *17.604,08* |
| | Standardabweichung | *2.087,01* | *0* | *0* | *0* |
| Anzahl getesteter Lösungen | Minimum | *6.546* | *12.994* | *13.658* | *15239* |
| | Maximum | *8.284* | *13.950* | *16.891* | *18830* |
| | Mittel | *7.283* | *13.519,2* | *15.615,6* | *17092,3* |
| | Standardabweichung | *512,4* | *317,06* | *927,07* | *1134,45* |
| Anzahl Lösungen im Suchraum | | *2.804.011* | | *8.656.936* | *22.964.086* |
| Anzahl durchgeführter Diversifizierungen | Minimum | *53* | *144* | *54* | *17* |
| | Maximum | *86* | *163* | *79* | *32* |
| | Mittel | *67* | *156* | *67,5* | *25,1* |
| | Standardabweichung | *10,14* | *5,78* | *7,24* | *4,95* |
| Erreichung des globalen Optimums | | ***90%*** | ***100%*** | ***100%*** | ***100%*** |

**Tabelle 6-9: Versuche zu Trainingsmenge 9**

### *6.1.3.11 Versuche mit Trainingsmenge 10*

Eine weitere Erhöhung der Anforderungen besteht darin, die Anzahl der Werte der erklärenden Variablen auf fünf zu erhöhen. Dadurch entstehen wesentlich mehr

Äquivalenzklassen, und die Laufzeit steigt. Bei der Betrachtung der Ergebnisse in Tabelle 6-10 sollte daher besonderes Augenmerk auf die realisierten Laufzeiten gelegt werden.

| **Verfahrens- und Modellparameter** | | **1** | **2** |
|---|---|---|---|
| Anzahl Iterationen ($G$) | | *100* | ***50*** |
| Länge der TOP-Liste ($X$) | | *3* | |
| $min\_Anwendbarkeit \cdot \lvert O^T \rvert$ | | *30* | ***20*** |
| $Prmax\_max$ | | *7* | ***8*** |
| Konfidenzwahrscheinlichkeit $(1-\alpha)$ | | *90%* | |
| Toleranz bei der Unterschreitung des Rekordes, $c$ | | *90%* | |
| Bewertung der Handlungsergebnisse | | *100/–10*<br>*0/0* | |
| **Ergebnisgrößen** | | | |
| Laufzeit des Verfahrens (in Sekunden) | Minimum | *1481,28* | *2629,39* |
| | Maximum | *1723,92* | *2897,77* |
| | Mittel | ***1611,58*** | ***2780,67*** |
| | Standardabweichung | *69,35* | *100,92* |
| Iterationen bis zum Erreichen der besten Lösung | Minimum | *5* | *5* |
| | Maximum | ***32*** | ***16*** |
| | Mittel | ***18,2*** | *7* |
| | Standardabweichung | *9,65* | *3,1* |
| Nutzwert der besten Lösung | Minimum | *47.851,27* | *47.851,27* |
| | Maximum | *47.851,27* | *47.851,27* |
| | Mittel | *47.851,27* | *47.851,27* |
| | Standardabweichung | *0* | *0* |
| Anzahl getesteter Lösungen | Minimum | *6061* | *6968* |
| | Maximum | *7030* | *8043* |
| | Mittel | *6560,9* | *7579,9* |
| | Standardabweichung | *298,79* | *339,43* |
| Anzahl Lösungen im Suchraum | | *2.804.011* | *8.656.936* |
| Anzahl durchgeführter Diversifizierungen | Minimum | *29* | *13* |
| | Maximum | *48* | *21* |
| | Mittel | *38,6* | *16,4* |
| | Standardabweichung | *5,75* | *2,65* |
| Erreichung des globalen Optimums | | ***100%*** | ***100%*** |

**Tabelle 6-10: Versuche zu Trainingsmenge 10**

Im **ersten Versuch** wird wieder mit den bislang bewährten Einstellungen begonnen. Auf Anhieb konnte in allen Testläufen das Optimum gefunden werden, und zwar im Mittel nach *18,2* Iterationen (knapp *5* Minuten), maximal nach *32* Iterationen (ca. *8,5* Minuten).

Ein **zweiter Versuch** soll dieses Ergebnis unter etwas anderen Rahmenbedingungen bestätigen. Neben einer Änderung der minimalen Allgemeingültigkeit auf *20* wurde die Anzahl der Iterationen auf *50* halbiert, und *Prmax_max* wurde auf *8* Prämissenklauseln erhöht. Das Optimum wurde wieder in allen Testläufen gefunden, und zwar jetzt bereits nach durchschnittlich *7* Iterationen (ca. *6,5* Minuten), maximal nach *16* Iterationen (knapp *15* Minuten).

#### *6.1.3.12 Versuche mit Trainingsmenge 11*

Eine weitere Erhöhung der Anforderungen besteht darin, die Anzahl der Datensätze auf *100.000* zu erhöhen. Die dazugehörigen Versuche sind in Tabelle 6-11 dargestellt.

Im **ersten Versuch** konnte in nur *50%* der Testläufe das Optimum „*X1, X2, X3, X4, X5, X6*“ gefunden werden. Die Parameter *G* und *Prmax_max* waren offensichtlich zu klein gewählt; durchschnittlich *9* Minuten Rechenzeit genügten nicht, um das Optimum immer aufzuspüren.

Daher soll ein **zweiter Versuch** mit *50* Iterationen und *Prmax_max=8* den erwünschten Erfolg bringen. Tatsächlich stellte sich in allen Testläufen das Optimum ein, wenn auch in einem Testlauf erst in der *49.* Iteration. Die Laufzeit betrug im Mittel ca. *51* Minuten.

| Verfahrens- und Modellparameter | | 1 | 2 |
|---|---|---|---|
| Anzahl Iterationen ($G$) | | *30* | ***50*** |
| Länge der TOP-Liste ($X$) | | *3* | |
| $min\_Anwendbarkeit \cdot \lvert O^T \rvert$ | | *30* | |
| $Prmax\_max$ | | *7* | ***8*** |
| Konfidenzwahrscheinlichkeit $(1-\alpha)$ | | *90%* | |
| Toleranz bei der Unterschreitung des Rekordes, $c$ | | *90%* | |
| Bewertung der Handlungsergebnisse | | *100/–10<br>0/0* | |
| **Ergebnisgrößen** | | | |
| Laufzeit des Verfahrens (in Sekunden) | Minimum | *519,47* | *2881,12* |
| | Maximum | *609,66* | *3183,82* |
| | Mittel | ***555,74*** | ***3059,01*** |
| | Standardabweichung | *26,3* | *158,3* |
| Iterationen bis zum Erreichen der besten Lösung | Minimum | *3* | *6* |
| | Maximum | *27* | ***49*** |
| | Mittel | *16* | *17,8* |
| | Standardabweichung | *7,75* | *13,83* |
| Nutzwert der besten Lösung | Minimum | *574,54* | *72.803,63* |
| | Maximum | *72.803,64* | *72.803,63* |
| | Mittel | *38.116,32* | *72.803,63* |
| | Standardabweichung | *34.771,3* | *0* |
| Anzahl getesteter Lösungen | Minimum | *1947* | *6905* |
| | Maximum | *2620* | *8711* |
| | Mittel | *2152,5* | *8080,3* |
| | Standardabweichung | *216,59* | *527,33* |
| Anzahl Lösungen im Suchraum | | *2.804.011* | *8.656.936* |
| Anzahl durchgeführter Diversifizierungen | Minimum | *7* | *14* |
| | Maximum | *22* | *34* |
| | Mittel | *11,8* | *28,1* |
| | Standardabweichung | *4,94* | *5,72* |
| Erreichung des globalen Optimums | | ***50%*** | ***100%*** |

**Tabelle 6-11: Versuche zu Trainingsmenge 11**

#### *6.1.3.13 Versuche mit Trainingsmenge 12*

Abschließend soll ein neuer Aspekt in die Testszenarien eingehen: Die Anzahl der Werte der zu erklärenden Variablen, $|dom(Y)|$, wird auf acht festgesetzt. Die Anzahl der

Datensätze wird wieder auf *10.000* zurückgesetzt. Die dazugehörigen Versuche sind in Tabelle 6-12 dargestellt.

| **Verfahrens- und Modellparameter** | | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
|---|---|---|---|---|---|---|---|---|
| Anzahl Iterationen ($G$) | | *50* | ***100*** | | | | | |
| Länge der TOP-Liste ($X$) | | *3* | | *10* | *1* | *3* | | |
| $min\_Anwendbarkeit \cdot \lvert O^T \rvert$ | | *20* | | | | | | |
| *Prmax_max* | | *7* | ***8*** | | | | | |
| Konfidenzwahrscheinlichkeit $(1-\alpha)$ | | *90%* | | | | | | |
| Toleranz bei der Unterschreitung des Rekordes, $c$ | | *90%* | | | | ***10%*** | ***100%*** | *50%* |
| Bewertung der Handlungsergebnisse | | *20/30/40/50/60/70/80/–10 0/0/0/0/0/0/0/0* | | | | | | |
| **Ergebnisgrößen** | | | | | | | | |
| Laufzeit des Verfahrens (in Sekunden) | Minimum | *180,76* | *1291,45* | *1240,49* | *1380,97* | *1619,83* | *998,67* | *1600,11* |
| | Maximum | *208,71* | *1459,28* | *1417,47* | *1577* | *1655,45* | *1120,79* | *1648,77* |
| | Mittel | *197,13* | *1384,71* | *1310,03* | *1495,62* | *1632,82* | *1069,84* | *1628,78* |
| | Standardabw. | *8,29* | *45,71* | *61,2* | *51,55* | *10,73* | *33,82* | *16,11* |
| Iterationen bis zum Erreichen der besten Lösung | Minimum | *7* | *5* | *13* | *7* | *5* | *7* | *6* |
| | Maximum | *48* | ***28*** | ***55*** | ***93*** | *21* | *77* | *53* |
| | Mittel | *19,6* | ***14,3*** | ***23,1*** | ***24,9*** | *11* | *17,7* | *17,1* |
| | Standardabw. | *15,26* | ***8,12*** | ***12,57*** | ***24,39*** | *4,92* | *20,32* | *14,33* |
| Nutzwert der besten Lösung | Minimum | *4.197,48* | *6128,94* | *6128,94* | *6128,94* | *4197,48* | *6128,94* | *6128,94* |
| | Maximum | *6.128,94* | *6128,94* | *6128,94* | *6128,94* | *6128,94* | *6128,94* | *6128,94* |
| | Mittel | *5.742,65* | *6128,94* | *6128,94* | *6128,94* | *5742,65* | *6128,94* | *6128,94* |
| | Standardabw. | *772,59* | *0* | *0* | *0* | *772,59* | *0* | *0* |
| Anzahl getesteter Lösungen | Minimum | *3.384* | *13.791* | *12.846* | *15136* | *18143* | *10036* | *18007* |
| | Maximum | *4.378* | *15.973* | *15.561* | *17705* | *18651* | *11077* | *18732* |
| | Mittel | *3.911,3* | *14.970,3* | *13.914,7* | *16643,7* | *18368,1* | *10615,2* | *18398,8* |
| | Standardabw. | *282,94* | *679,77* | *807,2* | *712,77* | *171,58* | *362,44* | *206,08* |
| Anzahl Lösungen im Suchraum | | *2.804.011* | *8.656.936* | | | | | |
| Anzahl durchgeführter Diversifizierungen | Minimum | *23* | *55* | *49* | *62* | ***95*** | *17* | ***93*** |
| | Maximum | *41* | *76* | *72* | *84* | ***99*** | *27* | ***98*** |
| | Mittel | *31,2* | *63,4* | *57,9* | *73,8* | *97* | *22,3* | *95,9* |
| | Standardabw. | *5,29* | *6,55* | *5,99* | *7,76* | *1,61* | *2,45* | *1,92* |
| Erreichung des globalen Optimums | | ***80%*** | ***100%*** | ***100%*** | ***100%*** | ***80%*** | ***100%*** | ***100%*** |

**Tabelle 6-12: Versuche zu Trainingsmenge 12**

Im **ersten Versuch** konnte in acht von zehn Testläufen das Optimum „*X1, X2, X3, X4, X5, X6*“ gefunden werden. Zweimal wurde als beste Lösung „*X2, X3, X4, X5, X6*“ ausgegeben.

Im **zweiten Versuch** wurde die Anzahl der Iterationen auf *100* und *Prmax_max* auf acht erhöht. Das Optimum konnte nun in allen Testläufen gefunden werden, und zwar maximal nach *28* Iterationen.

Im **dritten bzw. vierten Versuch** wurde, um das Verfahren auf seine Sensitivität bzgl. der *TOP*-Länge zu analysieren, *X:=10* bzw. *X:=1* gesetzt. In beiden Versuchen konnte das Optimum in allen Testläufen gefunden werden, jedoch im Mittel erst ca. *9-10* Iterationen später als mit der Einstellung *X=3*. Auch die Streuung der Iterationszahl bis zum Auffinden des Optimums hat von *8,12* auf *12,57* bzw. *24,39* zugenommen, und im schlechtesten Testlauf hätten *50* Iterationen nicht mehr ausgereicht, um das Optimum aufzuspüren.

Im **fünften bzw. sechsten Versuch** wurde, um das Verfahren auf seine Sensitivität bzgl. der Toleranzgrenze für die Unterschreitung des Rekordes zu analysieren, *c:=10%* bzw. *c:=100%* gesetzt. Letzteres entspricht dem Verfahren vor der Modifikation des Akzeptanzkriteriums. Mit *c=10%* konnte in acht von zehn Testläufen das Optimum gefunden werden. Zweimal wurde „*X2, X3, X4, X5, X6*“ als beste Lösung ausgegeben. In diesen beiden Testläufen war nach sieben bzw. zehn Iterationen keine Verbesserung mehr möglich. Überhaupt wurde in allen Testläufen in *95* bis *99* von *100* Iterationen diversifiziert. Eine Toleranz von *c=10%* ist also viel zu niedrig, um eine produktive Suche zu ermöglichen.

Auch eine Toleranz von *c=50%* im **siebten Versuch** führt in allen Testläufen zum Optimum. Die Anzahl Diversifizierungen war ähnlich hoch wie im fünften Versuch mit *c=10%*.

### 6.1.4 Zusammenfassung der Testphase

Nach der Auswertung der Testläufe können folgende Aussagen festgehalten werden:

- ⇨ **Realisierbare Laufzeiten**: In Abhängigkeit von der Anzahl der Datensätze, *N*, der Anzahl Iterationen, *G*, und der maximalen Anzahl an Prämissenklauseln, *Prmax_max*, können folgende Aussagen über die Laufzeit getroffen werden:

→ Legt man $N=10.000$ Datensätze zugrunde, so liegt das Laufzeitmittel je nach Parametereinstellungen zwischen $4$ und $56$ Sekunden pro Iteration. Die $56$ Sekunden waren in Versuch 2 der Trainingsmenge 10 mit $Prmax\_max=8$ zustandegekommen. Berücksichtigt man, daß in Versuch 1 bereits $Prmax\_max=7$ genügte, um das Optimum in allen Testläufen aufzuspüren, so liegt das *Laufzeitmittel zwischen $4$ und $17$ Sekunden pro Iteration*.

→ Die Laufzeit verhält sich proportional zu $N \cdot log\ N$ (wobei für $N$ eigentlich die Anzahl der tatsächlich gebildeten Äquivalenzklassen angesetzt werden müßte, aber die ist a-priori unbekannt).

→ Die Laufzeit verhält sich proportional zu $G$.

→ Legt man wieder $N=10.000$ Datensätze zugrunde, so liegt das Laufzeitmittel je nach Daten und Parametereinstellungen *zwischen $5$ und $37$ Sekunden pro $100$ getesteter Lösungen*.

→ Der Zusammenhang zwischen der Anzahl zu testender Lösungen und $Prmax\_max$ geht aus der Struktur der zu untersuchenden Teil-Suchräume hervor (vgl. Abbildung 5-6). Pro Iteration werden $\binom{Prmax\_max}{i}$ Lösungen mit $i$ Variablen getestet, insgesamt also $\sum_{i=1}^{Prmax\_max}\binom{Prmax\_max}{i}$ Lösungen pro Iteration.

Für $Prmax\_max=7$ umfaßt der Teil-Suchraum $127$ Lösungen, für $Prmax\_max=8$ umfaßt er $255$ Lösungen. Allerdings müssen bereits besuchte Lösungen nicht erneut getestet werden, so daß diese Zahlen Obergrenzen darstellen. Obwohl sich die Zahl der Lösungen verdoppelt, verdreikommafünffacht sich die Laufzeit.[431] Dies liegt daran, daß nicht nur mehr, sondern auch größere Lösungen generiert werden: Sowohl die Anzahl der Variablen (Breite der Tabellen) erhöht sich, als auch die Anzahl der gebildeten Äquivalenzklassen (Höhe der Tabellen).

⇨ **Auffindbarkeit des Optimums**: Das Optimum konnte letztendlich in allen Testumgebungen in $100\%$ der Fälle gefunden werden. Hierzu reichten relativ kurze Iterationszahlen aus. Wurde $Prmax\_max$ genügend groß gewählt, so konnte das Optimum

[431] Diese Beobachtung läßt sich beispielsweise an Tabelle 6-10 oder Tabelle 6-12 überprüfen, wenn man die Änderung der Iterationszahl, $G$, herausrechnet.

in allen Fällen in den ersten $100$, fast immer sogar in den ersten $50$ Iterationen gefunden werden. Positiv ist, daß allein durch eine Anpassung der Iterationszahl (und damit der Laufzeit) immer das Optimum gefunden werden konnte. Einzige Voraussetzung dafür war, daß $Prmax\_max$ um mindestens eins größer als die Anzahl Variablen im Optimum gewählt wurde.

⇨ **Abbruchkriterium**: Um zum einen überflüssige Berechnungen zu vermeiden und zum anderen relativ sicher zu sein, das globale Optimum zu finden, sollte die Anzahl Iterationen vorgegeben werden, nach denen das Verfahren abbricht, wenn während dieser Iterationen keine Verbesserung des Rekordes stattfindet. Aufgrund der vorstehenden Ausführungen zur Auffindbarkeit des Optimums reicht es (mit einem großen Puffer) aus, $50$ Iterationen nach der letzten Verbesserung abzubrechen. Da die Laufzeit besser kontrollierbar ist, wenn ein fester Wert für die Anzahl Iterationen vorgegeben wird, wird für die Anwendung auf die Realdaten in Abschnitt 6.2 noch an dem alten Abbruchkriterium festgehalten.

⇨ **Gütemaß-Parameter**: Die Höhe der Lösungsgüte hängt von drei Parametern ab: Erstens von der *Bewertung der Handlungsergebnisse*, zweitens von der vorgegebenen *Irrtumswahrscheinlichkeit*, $\alpha$, und drittens von der geforderten *minimalen Anwendbarkeit* der Regeln. Keiner der Faktoren hat einen besonderen Einfluß auf den Suchprozeß, sondern nur auf die Höhe der Lösungsqualitäten. Dabei ist zu beachten, daß die gewählten Einstellungen nicht zu einem trivialen Entscheidungsproblem führen dürfen, in dem stets dieselbe Alternative optimal ist und alle Lösungen nahezu gleichgut bewertet werden. Hinzu kommt bei der minimalen Anwendbarkeit, daß eine zu strenge Anforderung dazu führen kann, daß überhaupt keine zuläsigen Regeln gebildet werden können.

⇨ **Akzeptanzkriterium**: Das Kriterium, nach dem die lokalen Optima aus den Generalisierungsprozessen geprüft und darauf aufsetzend eine Intensivierung oder eine Diversifizierung eingeleitet wird, konnte in der Testphase verbessert werden. Allerdings sollte mit bspw. $90\%$ eine hohe Toleranz gewählt werden, damit nicht zu häufig diversifiziert wird. (Das ursprüngliche Akzeptanzkriterium entsprach einer Toleranz von $100\%$.)

⇨ **Länge der TOP-Liste**: Die Sensitivitätsanalysen der Versuche 2 bis 4 auf Trainingsmenge 12 haben gezeigt, daß der Suchprozeß gegenüber diesem Parameter

relativ unempfindlich ist, da $X$ nach der Verfahrensänderung nur noch die Diversifizierung beeinflußt. Eine schlechte Wertewahl, wie z.B. in den Versuchen 3 und 4 auf Trainingsmenge 12, kann durch eine größere Anzahl an Iterationen oder Prämissenklauseln aufgefangen werden.

Darüber hinaus konnten im Laufe der Testphase auch die Intensivierung und die Diversifizierung verbessert werden.

## 6.2 Anwendung des Verfahrens auf Realdaten

### 6.2.1 Beschreibung der Problemstellung

Die den folgenden Ausführungen zugrundeliegende Problemstellung wurde einem Data-Mining-Wettbewerb entnommen, dessen Ergebnisse zum Teil veröffentlicht wurden.[432] Für den Wettbewerb hatte ein Versicherungsunternehmen seine Kundendaten zur Verfügung gestellt. Es handelt sich hierbei um eine Trainingsmenge, $O^T$, und eine Evaluierungsmenge, $O^E$. Erstere umfaßt $5822$ Datensätze, letztere $4000$. Beide Datenmengen wurden bereits vorverarbeitet und enthalten keine fehlenden Werte. Die Evaluierungsmenge umfaßt dieselben Attribute wie die Lernmenge – nur das Zielattribut fehlte den Wettbewerbsteilnehmern. Nachträglich wurde es publiziert, so daß es hier zur Evaluierung des erlernten Modells verwendet werden kann. Es handelt sich dabei um das $86$. Attribut, *caravan*, das angibt, ob der entsprechende Kunde eine Caravan-Versicherung besitzt (Wert „*1*“) oder nicht (Wert „*0*“). Die ersten $43$ Attribute sind soziodemographische Merkmale, die nicht den einzelnen Kunden beschreiben, sondern die Region, in der der Kunde wohnt. Die Attribute $44$ bis $65$ beschreiben die Beitragshöhen zu verschiedenen Versicherungsprodukten, wie z.B. „Feuer“, „Boot“, „Motorrad“ oder „Kfz“. Die Attribute $66$ bis $85$ bilden die Anzahl der Policen ab, die der Kunde von einem dieser Produkte besitzt.

Sowohl die Trainings- als auch die Evaluierungsmenge enthält Käufer und Nichtkäufer. Die hier betrachtete Aufgabe besteht darin, anhand der Trainingsmenge ein Modell zu erlernen, das Käufer und Nichtkäufer bestmöglich trennt. Das Modell soll für die $4000$

[432] Vgl. zu der Problemstellung und den Ergebnissen des Wettbewerbs VAN DER PUTTEN (2000) ET AL., S. 1 ff.

Evaluierungsdatensätze klassifizieren, welche Kunden potentielle Käufer einer Caravan-Versicherung sind. Die *800* Kunden mit der höchsten Kaufwahrscheinlichkeit sollen dann selektiert und mit einer Direktmarketingmaßnahme kontaktiert werden. Unter den *4000* Datensätzen befinden sich *238* Käufer. Das Modellziel besteht demnach darin, die *800* Kunden so zu selektieren, daß sich darunter möglichst viele von den *238* Käufern befinden (vgl. Abbildung 6-4). Bei zufälliger Selektion von *800* Kunden würde man darunter *238/4000*800 = 47,6* Käufer erwarten. Nur ein Modell, das signifikant mehr korrekte Käufer identifiziert, liefert einen brauchbaren Erkenntniszuwachs gegenüber der zufälligen Selektion.

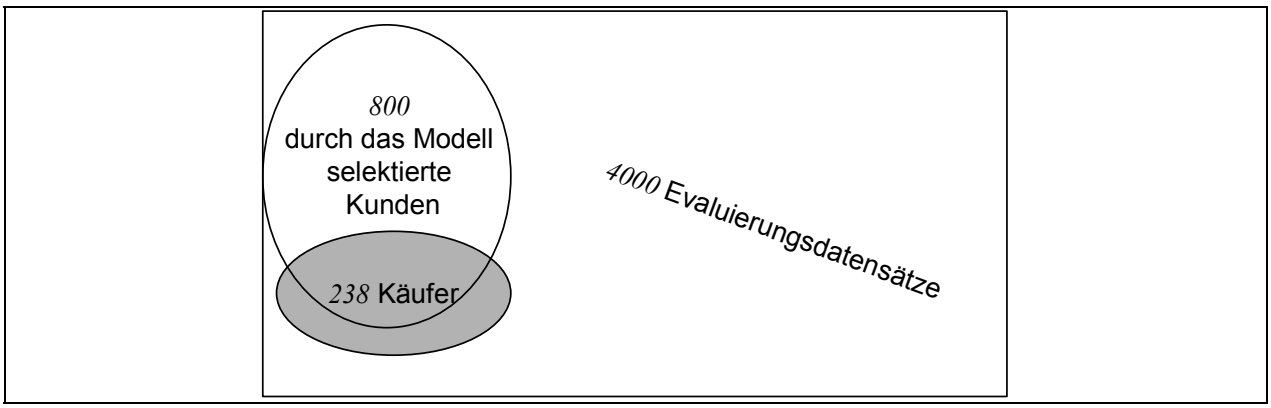



**Abbildung 6-4: Modellgestützte Selektion von *800* Kunden**

### 6.2.2 Modifikation des Gütemaßes zur Anpassung an die Problemstellung

Eigentlich ist das in dieser Arbeit konzipierte Gütemaß bestens zur Bearbeitung der im Abschnitt zuvor geschilderten Problemstellung geeignet. Leider wurde die wahre Problemstellung für die Zwecke des Wettbewerbs approximiert und vereinfacht.[433] So ergibt sich die Anzahl der zu kontaktierenden Kunden nicht auf der Grundlage der erwarteten Kosten und Erlöse – stattdessen ist, wie im Abschnitt zuvor gesagt, eine vorgegebene Anzahl von *800* der insgesamt *4000* Kunden zu selektieren.

Entsprechend kann das Gütemaß nun nicht mehr auf den Kosten und Erlösen aufbauen, da diese Parameter nicht bekannt sind. Es wird daher so modifiziert, daß es die erwartete Anzahl von Käufern maximiert. Als erste Nebenbedingung wird für jede Regel

[433] Vgl. VAN DER PUTTEN ET AL. (2000), S. 2.

eine gewisse Anwendungsrelevanz[434] gefordert. Da nur $800$ von $4000$ Kunden selektiert werden dürfen, wird dieses Verhältnis in einer zweiten Nebenbedingung auf die Trainingsmenge übertragen, welche $5822$ Kunden umfaßt. Diese Übertragung ist zulässig, da man im Data Mining generell unterstellt, daß die Verteilungen der betrachteten Variablen während des Trainings und bei der Anwendung des Modells identisch sind. Das geänderte Optimierungsproblem lautet:

$$max \sum_{(Pr \to Ko) \in M_{O^T}} \left| O^T \left[ Pr \wedge (caravan = 1) \right] \right|;$$

$$\frac{|O^T[Pr]|}{|O^T|} \geq min\_Anwendungsrelevanz; .$$

$$\sum_{(Pr \to Ko) \in M_{O^T}} \left| O^T \left[ Pr \right] \right| \leq \frac{800}{4000} \cdot 5822.$$

Wenn das Modell $M_{O^T}$ erst einmal fertig erlernt wird, kann es für jeden Kunden der Evaluierungsmenge, der durch die Prämisse einer Regel, $(Pr \to Ko) \in M_{O^T}$, erfaßt wird, dessen Kaufwahrscheinlichkeit durch $|O^T[Pr \wedge (caravan=1)]|/|O^T[Pr]|$ punktschätzen. Anschließend werden die Kunden nach absteigenden Kaufwahrscheinlichkeiten sortiert und die ersten $800$ Kunden selektiert.

### 6.2.3 Durchführung der Versuche

Zunächst steht die Konkurrenz zu den Teilnehmern des Wettbewerbs im Vordergrund. Daher wird im nächsten Abschnitt versucht, ein Modell zu erlernen, welches besser als die im Wettbewerb erzielten Ergebnisse bewertet wird. Dieses Ziel wird auch erreicht, aber es stellt sich dabei heraus, daß die erzielbaren Ergebnisse aufgrund der Wettbewerbsbedingungen zu stark vom Zufall abhängen, um einen fairen Vergleich zu ermöglichen. Daher wird im darauffolgenden Abschnitt 6.2.3.2 von den Wettbewerbsbedingungen abgerückt und mit einer neuen Versuchsserie das ursprüngliche Optimierungsproblem aus Abschnitt 5.3 bearbeitet.

---

[434] Die minimale Anwendungsrelevanz als Nebenbedingung erscheint an dieser Stelle sinnvoller als eine minimale Allgemeingültigkeit, da die Allgemeingültigkeit bereits durch das zu maximierende Krietrium angestrebt wird. Die Anwendungsrelevanz wurde in Definition 2-59 eingeführt.

### *6.2.3.1 Versuchsserie 1*

Das Verfahren wurde in einer ersten Versuchsserie auf eine Trainingsmenge mit den Feldern *44* bis *86* gestartet. Die soziodemographischen Merkmale *1* bis *43* wurden von vornherein von der Analyse ausgeschlossen, da diese nicht einzelne Kunden charakterisieren, sondern die Regionen, aus denen die Kunden stammen. Es wird angenommen, daß diese Merkmale keine genügend differenzierte Prognose der Kaufwahrscheinlichkeit erlauben.[435] Tabelle 6-13 beschreibt den Aufbau der Datenbasis. Die Bedeutung der Attribute kann Anhang D entnommen werden. Bei den Attributen *44* bis *64*, welche die Beitragshöhen zu den Versicherungen repräsentieren, beginnt die Attributkurzbezeichnung jeweils mit „*c_*“ für „contribution“ (engl.: Beitrag). Fehlt das Präfix „*c_*“, so bezeichnet das Attribut die Anzahl der Policen, die ein Kunde von dem jeweiligen Versicherungstyp abgeschlossen hat.

| Symbole für die Attribute | $a_1, a_2, \ldots, a_{43}$ | $a_{44}, a_{45}, \ldots, a_{64}$ | $a_{65}, a_{66}, \ldots, a_{86}$ |
|---|---|---|---|
| Bedeutung der Attribute | Soziodemographische Merkmale der Region | Versicherungsbeiträge | Anzahl Policen zur jeweiligen Versicherung |
| Beispiele | Durchschnittsalter, -kaufkraft, -haushaltsgröße in der Region | Beitrag[436] zur Kfz-, Surfbrett- oder Feuerversicherung | Anzahl_Policen zur Kfz-, Surfbrett- oder Feuerversicherung |
| Benennung der Attribute | keine (Diese Attribute werden von der Analyse ausgeschlossen.) | *c_car, c_surfboard, c_fire* | *car, surfboard, fire* |

**Tabelle 6-13: Aufbau der Datenbasis**

Bereits der erste Probelauf ergab folgende Variablenkombination, welche in der Evaluierungsmenge *122* Käufer korrekt klassifizierte:

*{c_third_party, c_car, c_surfboard, third_party, moped}*.

Dieses Ergebnis ist sogar besser als das des Wettbewerbssiegers, der *121* Käufer korrekt klassifizierte. Allerdings hat dies nichts zu bedeuten, wie folgende Überlegungen zeigen:

---

[435] Ähnlich argumentiert auch der Gewinner des Wettbewerbs (vgl. ELKAN (2000) o.S.).

[436] Falls ein Kunde mehrere Versicherungen desselben Typs besitzt (z.B. mehrere Kfz-Versicherungen), so gebene diese Attribute den höchsten Beitrag dieser Versicherungen an.

Ein Rough-Set-Modell besteht aus disjunkten Regeln, die (bei Vernachlässigung der Nebenbedingungen[437]) die gesamte Datenmenge abdecken, wobei jedes Objekt von genau einer Regel erfaßt wird (vgl. Abbildung 6-5).

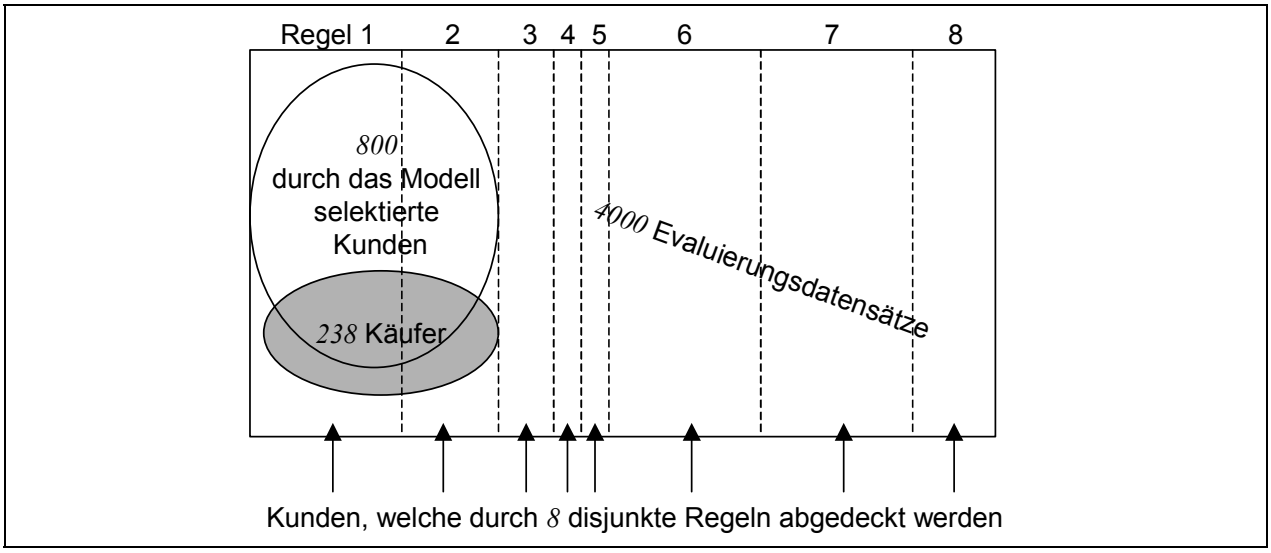



**Abbildung 6-5: Selektion von 800 Kunden durch disjunkte Regeln**

Laut Abschnitt 6.2.2 wird für alle Kunden der Evaluierungsmenge, $O^E$, die durch die Prämisse einer Regel, $(Pr_i \rightarrow Ko_i) \in M_{O^T}$, erfaßt werden, deren Kaufwahrscheinlichkeit durch die Konfidenz der Regel, $Konfidenz(Pr_i \rightarrow Ko_i) = |O^T[Pr_i \wedge (caravan86=1)]| / |O^T[Pr_i]|$, punktgeschätzt. Ordnet man $M$ disjunkte Regeln nach absteigender Konfidenz, so daß gilt:

$$Konfidenz(Pr_1 \rightarrow Ko_1) \geq \ldots \geq Konfidenz(Pr_M \rightarrow Ko_M),$$

so werden die besten $n$ Regeln zur Selektion verwendet – in Abbildung 6-5 sind dies gerade die Regeln 1 und 2. Allgemein wird $n$ so bestimmt, daß gilt:

$$\sum_{i=1}^{n-1} \left| O^T[Pr_i] \right| < 800 \leq \sum_{i=1}^{n} \left| O^T[Pr_i] \right|, \quad 1 \leq n \leq M.$$

Da mit den ersten $n$ Regeln nicht genau $800$, sondern u.U. weit mehr Kunden abgedeckt werden, muß die Menge der abgedeckten Kunden auf $800$ begrenzt werden. Das Problem bei der Verwendung von Regeln liegt darin, daß für alle Kunden in $O^E[Pr_i]$ *dieselbe* Kaufwahrscheinlichkeit geschätzt wird. Man hat damit kein sinnvolles Kriterium, die Menge der zu selektierenden Kunden auf $800$ zu begrenzen. Daher ist es willkürlich,

[437] Nach den o.g. Ausführungen wird eine minimale Anwendungsrelevanz für jede Regel verlangt.

welche der durch die $n$ besten Regeln abgedeckten Kunden letztendlich selektiert werden.

*Das Beispiel aus Tabelle 6-14 zeigt, wie willkürlich es ist, gerade 800 Kunden zu selektieren, denn die 1427 durch die dritte Regel erfaßten Kunden weisen alle dieselbe Kaufwahrscheinlichkeit von 11,44% auf.*

| Angewendete Regel | Kaufwahrscheinlichkeit (= Konfidenz der Regel) | Anzahl Kunden mit derselben Kaufwahrscheinlichkeit | Kumulierte Anzahl abgedeckter Kunden | Anzahl Caravanversicherungskäufer unter den Kunden | Kumulierte Anzahl Caravanversicherungskäufer |
|---|---|---|---|---|---|
| Regel 1 | *15,69* | *36* | *36* | *10* | *10* |
| Regel 2 | *14,81* | *16* | *52* | *2* | *12* |
| Regel 3 | *11,44* | *1427* | *1479* | *140* | *152* |
| ... | ... | ... | ... | ... | ... |

***Tabelle 6-14: Ergebnisse der Anwendung der ersten drei Regeln***

*Würde man 52 Kunden selektieren, so wären darunter 12 Käufer, und würde man 1479 Kunden selektieren, so wären darunter 152 Käufer. Bei mehr als 52 oder weniger als 1479 Selektionen hängt die Anzahl Käufer davon ab, welche Kunden man willkürlich auswählt. In dem oben beschriebenen Testlauf waren es insgesamt zufällig 122 Käufer.*

Da die Begrenzung der zu selektierenden Kunden willkürlich ist, muß der Sinn des gesamten Wettbewerbs in Frage gestellt werden (oder alle regelorientierten Verfahren müßten von dem Wettbewerb ausgeschlossen werden).

Aufgrund dieser ernüchternden Erkenntnis wird im nächsten Versuch wieder die ursprüngliche, ökonomisch ohnehin sinnvollere Interessantheitsbewertung aus Abschnitt 5.3 vorgenommen. Die Ergebnisse sind nicht mit den (Zufalls-) Ergebnissen des Wettbewerbs vergleichbar, da hier eine bestimmte Kosten- und Erlössituation unterstellt wird. Aus diesem Grund wird auch nicht näher auf die im Wettbewerb erzielten Ergebnisse eingegangen.

#### *6.2.3.2 Versuchsserie 2*

Das Verfahren wurde in einer zweiten Versuchsserie mit dem konzipierten Modellnutzwert[438] als Gütemaß auf eine Trainingsmenge mit den Feldern *44* bis *86* gestartet. Tabelle 6-15 zeigt die dazugehörigen Parametereinstellungen und Ergebnisse.

[438] Vgl.zum Modellnutzwert Definition 3-3, S. 172.

| Verfahrens- und Modellparameter | | 1 | 2 | 3 |
|---|---|---|---|---|
| Anzahl Iterationen ($G$) | | ***200*** | | ***300*** |
| Länge der TOP-Liste ($X$) | | *3* | | *3* |
| $min\_Anwendbarkeit \cdot \lvert O^T \rvert$ | | *20* | | *30* |
| *Prmax_max* | | *10* | | ***11*** |
| Konfidenzwahrscheinlichkeit $(1-\alpha)$ | | *90%* | | *90%* |
| Toleranz bei der Unterschreitung des Rekordes, $c$ | | *90%* | | *90%* |
| Bewertung der Handlungsergebnisse | | *100/–10 0/0* | ***200****/–10 0/0* | ***300****/–10 0/0* |
| **Ergebnisgrößen** | | | | |
| Laufzeit des Verfahrens (in Sekunden) | Minimum | *809,83* | *553,65* | *3493,28* |
| | Maximum | *1257,63* | *878,48* | *7017,01* |
| | Mittel | *939,11* | *707,44* | *5549,02* |
| | Standardabweichung | *163,63* | *127,50* | *1092,37* |
| Iterationen bis zum Erreichen der besten Lösung | Minimum | *58* | *109* | ***3*** |
| | Maximum | *199* | *198* | ***13*** |
| | Mittel | *131* | *151,2* | *8,2* |
| | Standardabweichung | *48,82* | *38,64* | *3,31* |
| Nutzwert der besten Lösung | Minimum | ***6916,78*** | ***27556,44*** | *52103,65* |
| | Maximum | ***6978,21*** | ***27556,44*** | *52103,65* |
| | Mittel | ***6959,78*** | ***27556,44*** | *52103,65* |
| | Standardabweichung | ***22,8*** | ***0*** | *0* |
| Anzahl getesteter Lösungen | Minimum | *81752* | *78906* | *279.218* |
| | Maximum | *94364* | *102477* | *316.981* |
| | Mittel | *88580,2* | *94266,6* | *299.881,2* |
| | Standardabweichung | *4452,06* | *8696,29* | *12.372,29* |
| Anzahl Lösungen im Suchraum | | *2.068.564.063* | | *6.349.125.439* |
| Anzahl durchgeführter Diversifizierungen | Minimum | *39* | *31* | *12* |
| | Maximum | *52* | *39* | *35* |
| | Mittel | *46* | *34,4* | *22,8* |
| | Standardabweichung | *4,43* | *2,87* | *8,89* |

**Tabelle 6-15: Versuchsserie zu den Caravanversicherungsdaten**

Im **ersten Versuch** wurden fünf Testläufe mit identischen Parametereinstellungen durchgeführt. Die Wahl der Einstellungen leitet sich aus den Testergebnissen des Abschnittes 6.1 ab. Aufgrund der großen Anzahl an Variablen wurde die Iterationszahl auf

*200* gesetzt. Die fünf Ergebnislösungen der fünf Testläufe sind in Tabelle 6-16 widergegeben.

| Nummer der Lösung | *c_third_party* | *c_car* | *c_van* | *c_lorry* | *c_trailer* | *c_private_accidents* | *c_diyability* | *c_fire* | *c_surfboard* | *third_party* | *van* | *lorry* | *trailer* | *private_accidents* | *disability* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | $a_{47}$ | $a_{48}$ | | | | | $a_{59}$ | $a_{60}$ | $a_{67}$ | | $a_{71}$ | $a_{72}$ | | $a_{79}$ |
| 2 | $a_{46}$ | $a_{47}$ | $a_{48}$ | $a_{50}$ | | $a_{56}$ | $a_{58}$ | $a_{59}$ | $a_{60}$ | | | | | | $a_{79}$ |
| 3 | $a_{46}$ | $a_{47}$ | | | | | | $a_{59}$ | $a_{60}$ | | $a_{69}$ | | $a_{72}$ | $a_{77}$ | $a_{79}$ |
| 4 | $a_{46}$ | $a_{47}$ | $a_{48}$ | $a_{50}$ | $a_{51}$ | | $a_{58}$ | $a_{59}$ | $a_{60}$ | | | | | $a_{77}$ | |
| 5 | | $a_{47}$ | | | | | $a_{58}$ | $a_{59}$ | $a_{60}$ | $a_{67}$ | $a_{69}$ | $a_{71}$ | $a_{72}$ | $a_{77}$ | |

**Tabelle 6-16: Attribute der fünf Lösungen aus dem ersten Versuch**

Mindestens eine der fünf Ergebnislösungen wurde erst in der *199*. Iteration gefunden, was vermuten läßt, daß der Lösungsraum noch bessere Lösungen umfaßt. Allerdings wurden die fünf Lösungen fast identisch gut bewertet, was wiederum die Vermutung nahelegt, daß die potentiell besseren Lösungen auch nicht viel besser als die erzielten Lösungen sind. Daher wird hier auf eine weitergehende Exploration des Suchraums verzichtet.

In allen fünf Lösungen erscheinen die Attribute *c_car*, *c_fire* und *c_surfboard*, wobei sich das erst- und das letztgenannte Attribut unmittelbar inhaltlicht begründen lassen, da ein Caravan durch ein Kraftfahrzeug gezogen wird und sowohl ein Caravan als auch ein Surfbrett kausal mit Urlaubsreisen im Zusammenhang stehen. Zur eingehenderen Untersuchung wird im weiteren beispielhaft die fünfte erzielte Lösung betrachtet. Sie wird in Abschnitt 6.2.4 unter der Bezeichnung „**Rough-Set-Modell 1**“ diskutiert.

Im **zweiten Versuch** wurden weitere fünf Testläufe mit einer Bewertung von *200* für das positive Handlungsergebnis „Kunde kauft Caravan-Versicherung““ durchgeführt. Die fünf Ergebnislösungen aus diesen fünf Testläufen zeigt Tabelle 6-17.

| Nummer der Lösung | *c_third_party* | *c_car* | *c_lorry* | *c_aggricultura* | *c_surfboard* | *lorry* | *aggricultural* | *surfboard* |
|---|---|---|---|---|---|---|---|---|
| 1 | $a_{46}$ | $a_{47}$ | | | $a_{60}$ | $a_{71}$ | $a_{74}$ | $a_{81}$ |
| 2 | $a_{46}$ | $a_{47}$ | | $a_{53}$ | | | $a_{74}$ | $a_{81}$ |
| 3 | $a_{46}$ | $a_{47}$ | $a_{50}$ | $a_{53}$ | | | $a_{74}$ | $a_{81}$ |
| 4 | $a_{46}$ | $a_{47}$ | | | $a_{60}$ | $a_{71}$ | $a_{74}$ | $a_{81}$ |
| 5 | $a_{46}$ | $a_{47}$ | $a_{50}$ | | $a_{60}$ | | $a_{74}$ | |

**Tabelle 6-17: Attribute der fünf Lösungen aus dem zweiten Versuch**

Zu den diskriminierungsrelevanten Phänomenen zählt der Besitz bzw. Nichtbesitz einer Pkw-, Lkw-, Landmaschinen und Surfbrett-Versicherung sowie einer Versicherungspolice bei einem anderen Versicherungsunternehmen.

Alle Ergebnisse erzielten exakt denselben Nutzwert, was vermuten läßt, daß es sich hier tatsächlich um die globalen Optima handelt. Beim Vergleich der fünf Modelle wird die inhaltliche Austauschbarkeit zwischen *c_surfboard* und *surfboard* bzw. *c_lorry* und *lorry* bzw. *c_aggricultural* und *aggricultural* offensichtlich. Entscheidender als die Fragen, wieviele Versicherungspolicen ein Kunde besitzt oder wie hoch seine Versicherungsbeiträge sind, scheint also die Frage zu sein, ob ein Kunde überhaupt eine Versicherung des entsprechenden Typs besitzt (bzw. ob er überhaupt einen Beitrag hierfür leistet). Lediglich das Attribut *c_car* scheint im Vergleich zu *car* eine eigene Bedeutung zu besitzen, da ein hoher Kfz-Versicherungsbeitrag für eine besondere Neigung zu Caravan-Versicherungskäufen steht (große Autos, Diesel, Vielfahrer).[439]

Beispielhaft wird in Abschnitt 6.2.4 die fünfte Lösung betrachtet. Sie wird dort unter der Bezeichnung „**Rough-Set-Modell 2**“ diskutiert.

Als **dritter Versuch** wird eine völlig neue Parameterkonstellation getestet. Um mit hoher Wahrscheinlichkeit das Optimum zu finden, wurde die Anzahl durchzuführender Iterationen auf *300* und *Prmax_max* auf *11* erhöht. Das positive Handlungsergebnis

[439] Diese Beobachtung wird im folgenden noch deutlicher.

wurde mit *300* bewertet. Es stellte sich heraus, daß die Einstellungen unnötig hoch gewählt wurden, denn bereits nach *3* bis *13* Iterationen wurde die jeweils beste Lösungen gefunden. Die gefundenen Lösungen sind in Tabelle 6-18 zusammengestellt.

| Nummer der Lösung | *c_car* | *c_lorry* | *c_aggricultural* | *c_surfboard* | *lorry* | *aggricultural* | *surfboard* |
|---|---|---|---|---|---|---|---|
| 1 | $a_{47}$ | | $a_{53}$ | | $a_{71}$ | | $a_{81}$ |
| 2 | $a_{47}$ | $a_{50}$ | $a_{53}$ | | | | $a_{81}$ |
| 3 | $a_{47}$ | $a_{50}$ | | | | $a_{74}$ | $a_{81}$ |
| 4 | $a_{47}$ | | $a_{53}$ | $a_{60}$ | $a_{71}$ | | |
| 5 | $a_{47}$ | $a_{50}$ | $a_{53}$ | | | | $a_{81}$ |

**Tabelle 6-18: Attribute der fünf Lösungen aus dem dritten Versuch**

Diese Lösungen unterscheiden sich nicht in ihrer Güte, so daß die Vermutung naheliegt, daß es sich hier tatsächlich um die globalen Optima handelt. Dies gilt umso mehr, wenn man wie oben unterstellt, daß die jeweils unterschiedlichen Attribute austauschbar sind. Auffällig ist die geringe Anzahl von nur vier Variablen in jeder Lösung. Dies liegt vor allem an dem günstigen Verhältnis zwischen den Bewertungen des positiven gegenüber dem negativen Handlungsergebnis. Es ist hier nicht mehr so wichtig wie zuvor, daß Käufer und Nichtkäufer gut getrennt werden, so daß wenige gute Variablen zur Diskriminierung ausreichen. Zu den diskriminierungsrelevanten Phänomenen zählt nun der Besitz bzw. Nichtbesitz einer Pkw-, Lkw-, Landmaschinen und Surfbrett-Versicherung.

Beispielhaft wird wieder die fünfte Lösung herausgegriffen. Sie wird in Abschnitt 6.2.4 unter der Bezeichnung „**Rough-Set-Modell 3**“ diskutiert.

Dieselben Versuche wurden noch einmal durchgeführt, wobei zuvor für jedes Attribut eine Wertehierarchie definiert wurde. Die Wertehierarchie umfaßt neben der Ebene mit der Wurzel jeweils zwei weitere Hierarchieebenen. Auf der unteren Ebene befinden sich die Ausprägungen aus der Datenbasis. Auf der oberen Ebene wurden jeweils die Werte größer Null zu einem Intervall zusammengefaßt, da oben vermutet wurde, daß

inhaltlich relevant nur die Tatsache sei, ob eine Kunde eine bestimmte Versicherungpolice besitze oder nicht – nicht aber die Anzahl der Policen oder die Höhe der Beiträge. Die Ergebnisse unterschieden sich nicht nennenswert von den in Tabelle 6-15 dargestellten Ergebnissen, so daß diese hier nicht weiter diskutiert werden.[440]

### 6.2.4 Interpretation der Ergebnisse und Vergleich mit den Ergebnissen eines Entscheidungsbaumverfahrens

An dieser Stelle sollen die erzielten Ergebnisse interpretiert und mit denen eines Entscheidungsbaumverfahrens verglichen werden, da letzteres dasjenige überwachte Lernverfahren darstellt, welches die größte Verbreitung in Theorie und Praxis gefunden hat. Bei dem Vergleich gilt folgendes zu berücksichtigen:

- ⇨ Entscheidungsbaumverfahren können den realen Zusammenhang genauer approximieren, da nicht jedes Attribut in jeder Regel vorkommen muß.
- ⇨ Entscheidungsbaumverfahren sind i.d.R. greedy-Algorithmen, d.h., sie bauen den Baum von der Wurzel aus auf, indem sie ihn Knoten für Knoten aufspalten, ohne ein Backtracking zu bereits aufgespalteten Knoten zu erlauben. Dadurch sind sie sehr schnell, können aber in Anwendungsbereichen mit vielen Variablen leicht in lokalen Optima enden.
- ⇨ Klassische Entscheidungsbaumverfahren optimieren kein betriebswirtschaftliches Gütemaß. Es wurde hier zwar gezielt ein Entscheidungsbaumverfahren gewählt, das unterschiedliche Fehlklassifikationskosten für falsch erkannte Käufer und Nichtkäufer berücksichtigen kann, aber erstens finden diese Kosten erst bei der Anwendung des Entscheidungsbaumes und nicht während des Lernens Berücksichtigung, und zweitens kennt das Gütemaß keine Sicherheitspuffer, falls die Entscheidungsempfehlung nur auf wenigen Datensätzen beruht. Man hat hier nur die Möglichkeit, sehr umständlich die Baumkomplexität einzuschränken und zu „hoffen“, daß sich dadurch Regeln ergeben, welche viele Datensätze abdecken. Drittens kann das gewählte Verfahren nur zwei mögliche Handlungsalternativen (Selektion, keine

[440] Einzig die Laufzeiten waren wesentlich größer, da in jeder Iteration wesentlich mehr Lösungen generiert und getestet werden. Denn die Einführung zusätzlicher Hierarchieebenen hat ähnliche Auswirkungen auf den Suchraum wie die Einführung zusätzlicher Variablen.

Selektion) und zwei mögliche Handlungsergebnisse (kauft, kauft nicht) unterscheiden, was hier zwar ausreicht, aber für andere Entscheidungsprobleme möglicherweise nicht.

Als Entscheidungsbaumsoftware wurde der Discoverer2000 der Firma „prudsys“ verwendet, da er eine leichte Handhabung und vielfältige Funktionen zur Analyse von Entscheidungsbäumen bietet – u.a. auch die hier benötigte (wenn auch nachträgliche) Bewertung des Entscheidungsbaums mit ökonomischen Größen.

Als Methode wurde der von prudsys entwickelte Entscheidungsbaumalgorithmus für achsenparallele Trennfunktionen eingesetzt. Es wurden mehrere Versuche mit unterschiedlichen Parameterkonfigurationen durchgeführt, die zwar zu unterschiedlich großen Bäumen führten, wobei sich aber die relevanten Regeln, welche Kunden als Caravanpolice-Käufer einstuften, kaum unterschieden. Folgende Parametereinstellungen führten zu dem im weiteren diskutierten Modell:

- ⇨ Baumkomplexität: *20* (relativ gering);[441]
- ⇨ Minimalanteil des Nachfolgeknotens: *5%*;[442]
- ⇨ Indexgrenzen: *1% – 10.000%*.[443]

Nach einer Laufzeit von nicht einmal zehn Sekunden wurde die folgende Regelmenge ausgegeben (Dabei stellt der Score-Wert die Konfidenz der jeweiligen Regel dar.):

*(c_car > 5.5)* and *(c_fire > 2.5)* and *(c_fire <= 5.5)* and *(car > 1.5)* => *Score = 0.220779*

*(c_car > 5.5)* and *(c_fire > 2.5)* and *(c_fire <= 5.5)* and *(car <= 1.5)* => *Score = 0.154004*

*(c_car > 5.5)* and *(c_fire <= 2.5)* and *(c_fire <= 1.5)* => *Score = 0.07*

*(c_car <= 5.5)* and *(c_moped <= 1)* and *(c_fire > 3.5)* => *Score = 0.0397644*

*(c_car <= 5.5)* and *(c_moped <= 1)* and *(c_fire <= 3.5)* and *(c_fire <= 0.5)* and *(c_car <= 2)* and *(c_motorcycle <= 2)* => *Score = 0.0386572*

*(c_car > 5.5)* and *(c_fire > 2.5)* and *(c_fire > 5.5)* => *Score = 0.0357143*

*(c_car > 5.5)* and *(c_fire <= 2.5)* and *(c_fire > 1.5)* => Score = *0.0331126*

---

[441] Es werden nur noch Segmente zerlegt, die mehr als eine bestimmte Anzahl von Objekten aufweisen. Genauere Angaben über die Anzahl der Objekte macht das Benutzerhandbuch nicht. Vgl. S. 148.

[442] D.h. ein untergeordneter Knoten muß mindestens *5%* der Objekte aus dem übergeordneten Knoten enthalten.

[443] D.h. der Anteil der Käufer bzw. Nichtkäufer in einem Knoten wird nicht begrenzt.

(*c_car* <= *5.5*) and (*c_moped* <= *1*) and (*c_fire* <= *3.5*) and (*c_fire* > *0.5*) and (*c_fire* > *2.5*) => *Score = 0.029304*

(*c_car* <= *5.5*) and (*c_moped* <= *1*) and (*c_fire* <= *3.5*) and (*c_fire* <= *0.5*) and (*c_car* > *2*) => *Score = 0.00857143*

(*c_car* <= *5.5*) and (*c_moped* <= *1*) and (*c_fire* <= *3.5*) and (*c_fire* > *0.5*) and (*c_fire* <= *2.5*) => *Score = 0.00392157*

(*c_car* <= *5.5*) and (*c_moped* > *1*) => *Score = 0*

(*c_car* <= *5.5*) and (*c_moped* <= *1*) and (*c_fire* <= *3.5*) and (*c_fire* <= *0.5*) and (*c_car* <= *2*) and (*c_motorcycle* > *2*) => *Score = 0*

Da der Anteil der Käufer in der Evaluierungsmenge 23*8/4000 = 5,95%* beträgt, kann man davon ausgehen, daß – wenn überhaupt – nur die ersten drei Regeln einen Kausalzusammenhang abbilden, der den Kauf einer Caravan-Versicherung in der Grundgesamtheit zu erklären vermag. So läßt sich für die kritischste der drei Regeln – nämlich die dritte Regel mit einem Score von *0,07* – folgender Hypothesentest durchführen:

Getestet wird die Nullhypothese:

$H_0$: Die erwartete Anzahl der Käufer mit *(c_car>5.5)* and *(c_fire<=1.5)* entspricht der erwarteten Anzahl aller Käufer.

Die Gegenhypothese lautet:

$H_1$: Die erwartete Anzahl der Käufer mit *(c_car>5.5)* and *(c_fire<=1.5)* ist größer als die erwarteten Anzahl aller Käufer.

$H_0$ ist gegen $H_1$ zu verwerfen, wenn gilt: [444]

$$\frac{\bar{x}-\bar{y}}{\sqrt{\frac{\left(e^x+e^y\right)\cdot\left(n^x+n^y-e^x-e^y\right)}{\left(n^x+n^y\right)\cdot n^x\cdot n^y}}} \geq c(1-\alpha);$$

$$5 \leq e^x \leq n^x - 5;$$

$$5 \leq e^y \leq n^y - 5.$$

Mit:

$\alpha = 1\%$,

$n^x = |O^T[(c\_car>5.5)$ and $(c\_fire<=1.5)]| = 1000$,

$n^y = |O^E| = 4000$,

$e^x = |O^T[(c\_car>5.5)$ and $(c\_fire<=1.5)$ and $(caravan=1)]| = 70$,

$e^y = |O^E[caravan=1]| = 238$,

[444] Vgl. BAMBERG/BAUR (1998), S. 193 f.

$\bar{x} = 0{,}07$ und
$\bar{y} = 0{,}0595$
ergibt sich:
$1{,}24 \geq 0{,}84;\ 5 \leq 70 \leq 995,\ 5 \leq 238 \leq 3995,$

so daß $H_0$ verworfen und $H_1$ akzeptiert wird. Dasselbe Resultat ergibt sich, wenn man Trainings- und Evaluierungsdaten austauscht, wodurch zum Aufstellen der Hypothesen anderes Datenmaterial als zum Testen verwendet wird. Die Prämisse der dritten Regel besitzt daher mit mindestens $99$-prozentiger Wahrscheinlichkeit einen gewissen Einfluß auf den Kauf einer Caravan-Versicherung.

Die ersten drei Regeln sind nicht nur statistisch haltbar, sondern lassen sich auch am besten interpretieren: Erstens stellt der Besitz eines Automobils ($(c\_car > 0)$ oder auch ($car > 0$)) die Voraussetzung für die Beförderung eines Caravans dar. Zweitens sind besonders hohe Beiträge zur Kfz-Versicherung ($c\_car = 6$) Anzeichen für große Autos, Diesel und Vielfahrer, die sicher in engem Zusammenhang mit Caravan-Besitzern stehen. Drittens kann bei den ersten beiden Regeln ein hoher Beitrag zur Feuerversicherung ($c\_fire > 2.5$) ein Indikator für das Bestreben sein, „rundum gut versichert“ sein zu wollen.

Um einen Vergleich zwischen dem Entscheidungsbaum und den generierten Rough-Set-Modellen durchzuführen, sind folgende Fragen zu beantworten:

1. Anhand welcher Kriterien soll der Vergleich durchgeführt werden?
2. Welches Modell ist nach den gewählten Vergleichskriterien das bessere?
3. Falls der Entscheidungsbaum besser ist: Warum konnte das Rough-Set-Verfahren kein besseres Modell liefern? Und falls das Rough-Set-Modell besser ist: Warum konnte das Entscheidungsbaumverfahren kein besseres Modell liefern?

Diese Fragen sollen in den folgenden drei Abschnitten beantwortet werden.

#### *6.2.4.1 Entwicklung der Vergleichskriterien*

In diesem Abschnitt wird die Frage beantwortet, anhand welcher Kriterien die verschiedenen Modelle verglichen werden sollen. Ein mögliches Vergleichskriterium ist der mit einer bestimmten Konfidenzwahrscheinlichkeit zu erwartende Deckungsbeitrag des Modells. Dieses Kriterium wurde in Definition 3-3 für die Trainingsmenge formal

definiert und soll hier als „*Nutzwert auf der Trainingsmenge*“ bezeichnet werden. Es kann durch Multiplikation mit $4000/5822$ auf die Evaluierungsmenge umgerechnet werden und heißt dann: „*Erwarteter Nutzwert ohne Kenntnis der neuen Verteilung*“, da die Verteilungen in der Evaluierungsmenge nicht bekannt sind, sondern einfach die alten Verteilungen aus der Trainingsmenge übertragen wurden. Wendet man das Modell auf die Evaluierungsdaten an, so läßt sich der „*Erwartete Nutzwert mit Kenntnis der neuen Verteilung*“ berechnen:

**Definition 6-1: Erwarteter Nutzwert mit Kenntnis der neuen Verteilung**

Es gelten dieselben Voraussetzungen wie in Definition 3-3. Weiter sei $O^E$ eine Evaluierungsmenge. Dann ist der *erwartete Nutzwert mit Kenntnis der neuen Verteilung* wie folgt definiert:

$$Modellnutzwert_{\alpha}^{E}(M^{Ent}) = \sum_{\forall o \in O^E} ug_i^{Pr,E}(1-\alpha);$$

$$o \in O^E[Pr];$$

$$i : M^{Ent}(o) = h_i; i = 0,...,hmax.$$

$$ug_i^{Pr,E}(1-\alpha) = \bar{z}_i - \frac{c(1-\alpha)}{\sqrt{\left|O^E[Pr \wedge (h = h_i)]\right|}} \cdot \sigma . \qquad \Diamond$$

Dieser Nutzwert ist eine vorsichtige Schätzung des Deckungsbeitrags, der zu erwarten ist, wenn man die durch das Modell empfohlenen Entscheidungen umsetzt. Berechnet man die *Abweichung der Nutzwerte mit und ohne Kenntnis der Verteilung*, so erhält man ein Maß für den Unterschied zwischen den Verteilungen der Trainings- und der Evaluierungsmenge, welche durch das erlernte Modell erfaßt werden. Dabei wird nur der Unterschied bezüglich der *erklärenden* Variablen des Entscheidungsmodells quantifiziert, da sich Definition 6-1 von Definition 3-3 nur durch die Anwendung der Regelprämissen auf die jeweilige Datenmenge unterscheidet: $o \in O^E[Pr]$ bzw. $o \in O^T[Pr]$. Die Entscheidung, $(h=h_i)$, ist im Modell $M^{Ent}$ verankert, welches in den beiden Definitionen identisch ist.

Durch Anwendung der zu vergleichenden Modelle auf die Evaluierungsmenge kann auch der *tatsächlich realisierte Deckungsbeitrag* gemessen werden. Damit läßt sich auch die *Abweichung* dieses Deckungsbeitrags *von dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung* messen, um zu beurteilen, ob die vorsichtige Schätzung

auf Basis der Trainingsdaten (wie beabsichtigt) unter den tatsächlich realisierten Deckungsbeiträgen liegt.

Dabei muß beachtet werden, daß die Evaluierungsmenge – genau wie die Trainingsmenge – lediglich eine Stichprobe ist, so daß der darauf realisierbare Deckungsbeitrag eine Zufallsvariable darstellt. Damit stellt auch der auf der Evaluierungsmenge realisierte Deckungsbeitrag oder dessen Abweichung von der Schätzung kein verläßliches Vergleichskriterium dar.

Besser geeignet wäre hier eine Untergrenze für den erwarteten Deckungsbeitrag, der mit einer vorgegebenen Konfidenzwahrscheinlichkeit in der Grundgesamtheit vorliegt.[445] Die o.g. Nutzwerte können hierzu nicht verwendet werden, da diese anhand der Trainingsmenge berechnet wurden, die auch schon zur Generierung der Hypothesen herangezogen wurde.[446] Daher ist der Nutzwert aus Definition 6-1 auf die Evaluierungsmenge, $O^E$, zu übertragen:

**Definition 6-2: Nutzwert auf der Evaluierungsmenge**

Gegeben sei ein auf der Evaluierungsmenge erlerntes Entscheidungsmodell:

$M^{Ent}(o) = Akkumulation(M_{O^E}, o)$.

Ansonsten gilt Definition 6-1 analog. ◊

Dieses Kriterium wird hier als „*Nutzwert auf der Evaluierungsmenge*“ bezeichnet. Es stellt eine vorsichtige Schätzung der erwarteten Deckungsbeiträge dar, wenn das Modell auf $4000$ neue Kunden angewendet werden würde.

Die Entscheidung über die Vorziehenswürdigkeit eines Modells soll aufgrund der diskutierten Vorbehalte bezüglich der anderen Kriterien anhand des Nutzwertes auf der Evaluierungsmenge getroffen werden. Interessant ist auch die Frage, inwieweit der für die Evaluierungsmenge zu erwartende Nutzwert den tatsächlichen Nutzwert auf der Evaluierungsmenge approximieren kann. Schließlich steht letzterer während der Lernphase nicht zur Verfügung, und ein Lernverfahren ist auf eine gute Approximation des auf der Grundgesamtheit erzielbaren Nutzwertes angewiesen. Daher wird die *Abweichung* des

---

[445] Realisiert werden kann der Erwartungswert allerdings nur bei unendlich häufiger Anwendung des Modells.

[446] Vgl. hierzu die kritischen Ausführungen aus Abschnitt 5.3.2.

Nutzwertes auf der Evaluierungsmenge *von dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung* gemessen.

Schließlich kann noch die *Abweichung* des Nutzwertes auf der Evaluierungsmenge *von dem erwarteten Nutzwert mit Kenntnis der neuen Verteilung* gemessen werden, um den Unterschied zwischen Trainings- und Evaluierungsdaten zu quantifizieren. Dieser Unterschied bezieht sich im Gegensatz zur o.g. Abweichung (der Nutzwerte mit und ohne Kenntnis der Verteilung) auf die bedingte Verteilung der Caravanversicherungskäufer in Abhängigkeit von den erklärenden Variablen des Modells. Die entsprechenden Abweichungskennzahlen wurden auch für alle Modelle und Entscheidungsprobleme berechnet – dabei stellte sich jedoch heraus, daß sie sich jeweils nur marginal unterschieden. Aufgrund des mangelnden Erkenniszuwachses wird daher im folgenden auf die Darstellung dieser letzten Kennzahl verzichtet.

#### *6.2.4.2 Interpretation und Vergleich der Modelle*

Zunächst sei die Evaluierung des Entscheidungsbaumes für die drei Bewertungsszenarien betrachtet (vgl. Tabelle 6-19). Wird der Abschluß einer Caravan-Versicherung mit *100* bewertet, so liefern zwei Regeln einen positiven Nutzenbeitrag, bei *200* und *300* jeweils drei Regeln. Die übrigen Regeln mit negativem Nutzenbeitrag sind nur aufgeführt, um sie mit Tabelle 6-20 zu vergleichen. Sie würden nicht zur Anwendung kommen, da man sich bei Nichtselektion der entsprechenden Kunden besser stellt.

Betrachtet man das erste Entscheidungsproblem (mit der Bewertung *100/–10*), so erwirtschaften die beiden ersten Regeln auf der Trainingsmenge einen Nutzwert von *6.645,55*. Rechnet man diesen Nutzwert auf die *4000* Evaluierungsdatensätze um, so ergibt sich ohne Kenntnis der neuen Verteilung ein erwarteter Nutzwert von *4.565,82*. Mit Kenntnis der Verteilung beträgt der erwartete Nutzwert *4213,27*, was *7,72%* unter dem zuvor genannten Nutzwert liegt. Dies spricht dafür, daß die Trainings- und die Evaluierungsmenge in der Verteilung der in den Regelprämissen auftretenden Variablen ein wenig voneinander abweichen.

| c_car | c_fire | car | c_moped | c_motor-cycle | caravan = [0 ;0] | caravan = [1 ;1] | Nutzen (*100/–10*) | Nutzen (*200/–10*) | Nutzen (*300/–10*) |
|---|---|---|---|---|---|---|---|---|---|
| *> 5,5* | *(2,5;5,5]* | *> 1,5* | | | *120* | *34* | *1.474,14* | *4.218,28* | *6.954,41* |
| *> 5,5* | *(2,5;5,5]* | *≤ 1,5* | | | *824* | *150* | *5.171,41* | *18.727,25* | *32.283,08* |
| *> 5,5* | *≤ 1,5* | | | | *930* | *70* | *-3.437,81* | *2.527,81* | *8.493,43* |
| *≤ 5.5* | *> 3.5* | | *≤ 1* | | *652* | *27* | *-4538,04* | *-2490,81* | *-443,58* |
| *≤ 2* | *≤ 3.5* | | *≤ 1* | *≤ 2* | *945* | *38* | *-6502,34* | *-3477,19* | *-452,04* |
| *> 5.5* | *> 5.5* | | | | *81* | *3* | *-749,85* | *-667,90* | *-585,95* |
| *> 5.5* | *(1,5;2,5]* | | | | *146* | *5* | *-1270,07* | *-1051,94* | *-833,82* |
| | | | | | | Nutzwert auf der Trainingsmenge (Summe der positiven Nutzenbeiträge): | *6.645,55* | *25.469,34* | *47.730,92* |
| | | | | | | Erwarteter Nutzwert ohne Kenntnis der neuen Verteilung: (Nutzwert auf der Trainingsmenge mal *4000/5822*): | *4.565,82* | *17.498,69* | *32.793,49* |
| | | | | | | Erwarteter Nutzwert mit Kenntnis der neuen Verteilung (vorsichtige Schätzung der erwarteten Deckungsbeiträge): | *4.213,27* | *16.519,55* | *31.262,19* |
| | | | | | | Abweichung der Nutzwerte mit und ohne Kenntnis der Verteilung: | *-7,72%* | *-5,59%* | *-4,67%* |
| | | | | | | Anzahl der fälschlicherweise selektierten Kunden: | *672* | *1294* | *1294* |
| | | | | | | Anzahl der korrekt selektierten Kunden (Caravan-Käufer): | *114* | *150* | *150* |
| | | | | | | **Tatsächlich realisierte Deckungsbeiträge:** | ***4.680*** | ***17.060*** | ***32.060*** |
| | | | | | | **Abweichung von dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung:** | ***2,50%*** | ***-2,51%*** | ***-2,24%*** |
| | | | | | | **Nutzwert auf der Evaluierungsmenge:** | ***2.899,61*** | ***12.090,57*** | ***24.724,17*** |
| | | | | | | **Abweichung von dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung:** | ***-36,49%*** | ***-30,91%*** | ***-24,61%*** |

**Tabelle 6-19: Evaluierung des Entscheidungsbaums (nach dem hier entwickelten Nutzenkriterium)**

Auf der Evaluierungsmenge konnten *114* Käufer korrekt identifiziert werden; *672* Kunden wurden fälschlicherweise für Käufer gehalten und kontaktiert. Daraus ergibt sich ein Deckungsbeitrag von $114 \cdot 100 - 672 \cdot 10 = 4.680$. Dies liegt *2,50%* über dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung. Da letzterer eine sehr vorsichtige Schätzung darstellt, verwundert es ein wenig, daß der tatsächlich realisierte Deckungsbeitrag nur so geringfügig über dem Schätzwert liegt. In den beiden anderen Bewertungsszenarien liegt er gar *2,51%* bzw. *2,24% unter* dem Schätzwert, welcher eigentlich die Funktion einer Untergrenze für die zu erwartenden Deckungsbeiträge einnimmt. Die Schätzung beruht auf der Annahme, daß die Verteilung der Evaluierungsdaten der der Trainingsdaten entspricht. Diese Annahme ist hier offensichtlich verletzt. Auch der Nutzwert

auf der Evaluierungsmenge liegt *36,49%* bzw. *30,91%* bzw. *24,61%* unter dem auf der Grundlage der Trainingsdaten erwarteten Nutzwert. All dies spricht dafür, daß die Caravanversicherungskäufer (in Abhängigkeit von den erklärenden Variablen des Modells) in der Evaluierungsmenge anders verteilt sind als in der Trainingsmenge. Diese Beobachtung wird im folgenden noch mehrfach bestärkt.

Das verwendete Softwaretool benutzt nicht das hier entwickelte Nutzenkriterium zur Klassifikation der Caravanversicherungskäufer, sondern allein den erwarteten Deckungsbeitrag. Von diesem Erwartungswert wird kein Sicherheitspuffer in Abhängigkeit von der Anwendbarkeit der Regeln abgezogen, so daß er größer als der entsprechende Nutzwert ist. Dies hat zur Folge, daß nach dem Deckungsbeitragskriterium u.U. mehr Regeln zur Klassifikation von Käufern angewendet werden als nach dem Nutzwertkriterium – mindestens aber genausoviele. Tendenziell werden also nach dem Deckungsbeitragskriterium mehr Kunden selektiert und kontaktiert als nach dem Nutzenkriterium. Dieser Unterschied erfordert eine getrennte Evaluierung (vgl. Tabelle 6-20).

Der Wegfall des Sicherheitspuffers führt (unter ansonsten gleichen Rahmenbedingungen) dazu, daß jetzt wesentlich größere negative Abweichungen von dem erwarteten Deckungsbeitrag auftreten. Die realisierten Deckungsbeiträge sind in den ersten beiden Entscheidungsproblemen dieselben wie zuvor; bei einer Bewertung von *300/–10* liegen sie mehr als *8000* Einheiten über den Deckungsbeiträgen, die bei Verwendung des Nutzenkriteriums realisiert worden wären. Diese Verbesserung wird – wie ein Vergleich mit Tabelle 6-19 zeigt – durch die Anwendung von vier zusätzlichen Regeln ermöglicht, welche auf der Trainingsmenge einen negativen Nutzen-, aber einen positiven Deckungsbeitrag erwirtschaften. Gerade die letzten beiden Regeln sind mit *3* bzw. *5* Käufern und einem Deckungsbeitrag von *90* bzw. *40* äußerst fragwürdig und erzielen wohl nur zufällig auf den Evaluierungsdaten einen positiven Deckungsbeitrag. Einen Kausalzusammenhang zum Kauf einer Caravan-Versicherung bilden sie ohnehin nicht ab, wie oben bereits festgestellt wurde. Solche Zufälligkeiten werden durch das Nutzenkriterium (mit einer bestimmten Irrtumswahrscheinlichkeit) ausgeschlossen.

| c_car | c_fire | car | c_moped | c_motor-cycle | caravan = [0;0] | caravan = [1;1] | DB (*100/–10*) | DB (*200/–10*) | DB (*300/–10*) |
|---|---|---|---|---|---|---|---|---|---|
| *> 5,5* | *(2,5;5,5]* | *> 1,5* | | | *120* | *34* | *2200* | *5600* | *9000* |
| *> 5,5* | *(2,5;5,5]* | *≤ 1,5* | | | *824* | *150* | *6760* | *21760* | *36760* |
| *> 5,5* | *≤ 1,5* | | | | *930* | *70* | *-2300* | *4700* | *11700* |
| *≤ 5.5* | *> 3.5* | | *≤ 1* | | *652* | *27* | *-3820* | *-1120* | *1580* |
| *≤ 2* | *≤ 3.5* | | *≤ 1* | *≤ 2* | *945* | *38* | *-5650* | *-1850* | *1950* |
| *> 5.5* | *> 5.5* | | | | *81* | *3* | *-510* | *-210* | *90* |
| *> 5.5* | *(1,5;2,5]* | | | | *146* | *5* | *-960* | *-460* | *40* |
| Deckungsbeitrag (DB) auf der Trainingsmenge (Summe der positiven Deckungsbeiträge): | | | | | | | *8.960* | *32.060* | *61.120* |
| Erwarteter Deckungsbeitrag ohne Kenntnis der neuen Verteilung: (Deckungsbeitrag auf der Trainingsmenge mal *4000/5822*): | | | | | | | *6.155,96* | *22.026,79* | *41.992,44* |
| Erwarteter Deckungsbeitrag mit Kenntnis der neuen Verteilung (vorsichtige Schätzung der erwarteten Deckungsbeiträge): | | | | | | | *6.108,91* | *21.900,52* | *41.893,95* |
| Abweichung der Deckungsbeiträge mit und ohne Kenntnis der Verteilung: | | | | | | | *-0,76%* | *-0,57%* | *-0,23%* |
| Anzahl der fälschlicherweise selektierten Kunden: | | | | | | | *672* | *1294* | *2615* |
| Anzahl der korrekt selektierten Kunden (Caravan-Käufer): | | | | | | | *114* | *150* | *221* |
| **Tatsächlich realisierte Deckungsbeiträge:** | | | | | | | ***4.680*** | ***17.060*** | ***40.150*** |
| **Abweichung von dem erwarteten Deckungsbeitrag ohne Kenntnis der neuen Verteilung:** | | | | | | | ***-23,98%*** | ***-22,55%*** | ***-4,39%*** |
| **Nutzwert auf der Evaluierungsmenge:** | | | | | | | ***2.899,61*** | ***12.090,57*** | ***26.851,38*** |

**Tabelle 6-20: Evaluierung des Entscheidungsbaums (nach dem Deckungsbeitragskriterium des Softwaretools)**

Dem Entscheidungsbaum werden nun die Rough-Set-Modelle 1, 2 und 3 gegenübergestellt. Dei wird das einfache Deckungsbeitragskriterium durch den Modellnutzwert aus Definition 3-3 ersetzt. Zur Erinnering wird seine formale Definition hier noch einmal wiederholt:

$$Modellnutzwert_{\alpha}(M^{Ent}) = \sum_{\forall o \in O^T} ug_i^{Pr}(1-\alpha);$$

$$o \in O^T\,[Pr];$$

$$i : M^{Ent}(o) = h_i; i = 0,...,hmax.$$

Dabei stellt, wie gehabt, $O^T$ die Trainingsmenge, $Pr$ eine Prämisse, $\alpha$ eine vorzugebende Irrtumswahrscheinlichkeit, $M^{Ent}(o)=h_i$ die $i$-teHandlungsempfehlung aus einer

gegebenen Alternativenmenge für ein Datenobjekt $o$ und $ug_i^{Pr}(1-\alpha)$ die Untergrenze[447] für den Deckungsbeitrag dar, welcher zu erwarten ist, wenn für ein Datenobjekt, $o \in O^T[Pr]$, die Handlung $h_i$ durchgeführt wird. Der Modellnutzwert ergibt sich aus der Summe der Untergrenzen für die Deckungsbeiträge über alle Datenobjekte, die bei der jeweils optimalen Entscheidung (Selektion bzw. Nichtselektion) zu erwarten sind.

Zunächst wird anhand von Tabelle 6-21 die Evaluierung des Rough-Set-Modells 1 betrachtet. Der Kauf einer Caravan-Versicherung wird hier mit *100* bewertet, so daß das Modell dem Entscheidungsbaum im ersten Bewertungsszenario gegenübrgestellt werden kann. Das Rough-Set-Modell 1 umfaßt zwei Regeln, die für die Handlungsempfehlung „Caravan-Angebot unterbreiten“ einen positiven Nutzwert ergeben:

WENN *c_car = 6* UND *c_disability = 0* UND *c_fire = 4* UND *c_surfboard = 0* UND *third_party = 0* UND *van = 0* UND *lorry = 0* UND *trailer = 0* UND *private_accidents = 0* DANN Caravan-Angebot unterbreiten.

WENN *c_car = 6* UND *c_disability = 0* UND *c_fire = 3* UND *c_surfboard = 0* UND *third_party = 0* UND *van = 0* UND *lorry = 0* UND *trailer = 0* UND *private_accidents = 0* DANN Caravan-Angebot unterbreiten.

Beschränkt man die inhaltliche Interpretation dieser Regeln zunächst auf die Klauseln mit Werten größer *0*, also „*c_car = 6*“ in Kombination mit „*c_fire = 4*“ sowie „*c_car = 6*“ in Kombination mit „*c_fire = 3*“, so sind wieder die oben bereits erkannten Einflußfaktoren "Vielfahrer/Diesel/große Autos" und "rundum gut versichert" feststellbar.

Die übrigen Regeln, die von der Unterbreitung eines Angebotes abraten, sind hier nicht von Interesse. Eine Ausnahme bildet die dritte in der folgenden Tabelle 6-21 abgebildete Regel:

WENN *c_car = 6* UND *c_disability = 0* UND *c_fire = 5* UND *c_surfboard = 0* UND *third_party = 0* UND *van = 0* UND *lorry = 0* UND *trailer = 0* UND *private_accidents = 0* DANN Caravan-Angebot unterbreiten.

Diese Regel ergibt zwar einen negativen Nutzenbeitrag auf der Trainingsmenge, erzielt aber auf der Evaluierungsmenge einen positiven Nutzwert. Da die Evaluierungsmenge nicht bereits zum Erlernen des Modells diente, ist dieses Zahlenmaterial verläßlicher, so daß die dritte Regel im Nachhinein auch akzeptiert wird.

[447] Diese Untergrenze wurde in Abschnitt 3.3.2.1 statistisch motiviert.

| c_car | c_disabi-lity | c_fire | c_surf-board | third_ party | van | lorry | trailer | private_ accident | caravan = [0;0] | caravan = [1;1] | Nutzen (*100/–10*) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *6* | *0* | *4* | *0* | *0* | *0* | *0* | *0* | *0* | *524* | *124* | *5737,63* |
| *6* | *0* | *3* | *0* | *0* | *0* | *0* | *0* | *0* | *294* | *51* | *1220,09* |
| *6* | *0* | *5* | *0* | *0* | *0* | *0* | *0* | *0* | *44* | *6* | *-164,04* |
| | | | | | | | | | | Nutzwert auf der Trainingsmenge (Summe der positiven Nutzenbeiträge): | *6957,72* |
| | | | | | | | | | | Erwarteter Nutzwert ohne Kenntnis der neuen Verteilung: (Nutzwert auf der Trainingsmenge mal *4000/5822*): | *4780,30* |
| | | | | | | | | | | Erwarteter Nutzwert mit Kenntnis der neuen Verteilung (vorsichtige Schätzung der erwarteten Deckungsbeiträge): | *4858,08* |
| | | | | | | | | | | Abweichung der Nutzwerte mit und ohne Kenntnis der Verteilung: | *1,63%* |
| | | | | | | | | | | Anzahl der fälschlicherweise selektierten Kunden: | *589* |
| | | | | | | | | | | Anzahl der korrekt selektierten Kunden (Caravan-Käufer): | *101* |
| | | | | | | | | | | **Tatsächlich realisierte Deckungsbeiträge:** | ***4210*** |
| | | | | | | | | | | **Abweichung von dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung:** | ***-11,93%*** |
| | | | | | | | | | | **Nutzwert auf der Evaluierungsmenge:** | ***2487,12*** |
| | | | | | | | | | | **Abweichung von dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung:** | ***-47,97%*** |

**Tabelle 6-21: Evaluierung des Rough-Set-Modells 1**

Die beiden ersten Regeln führen auf der Trainingsmenge zu einem Nutzwert von *6957,72*. Rechnet man diesen Nutzwert auf die *4000* Evaluierungsdatensätze um, so ergibt sich ohne Kenntnis der Verteilung ein erwarteter Nutzwert von *4780,30*. Mit Kenntnis der Verteilung beträgt der erwartete Nutzwert *4858,08*, was nur *1,63%* von dem zuvor genannten Nutzwert abweicht. Dies spricht dafür, daß die Trainings- und die Evaluierungsmenge in der Verteilung der in den Regelprämissen auftretenden Variablen nur gering voneinander abweichen. Anders verhält es sich mit der bedingten Verteilung der Variable *caravan*, denn trotz der vorsichtigen Schätzung der erwarteten Deckungsbeiträge kann nur ein Deckungsbeitrag von $101 \cdot 100 - 589 \cdot 10 = 4210$ tatsächlich realisiert werden. Dieser Deckungsbeitrag liegt trotz der vorsichtigen Schätzung um *11,93%* unter der Schätzung.

Bei der Berechnung ist Nutzwertes auf der Evaluierungsmenge wurde auch die dritte Regel berücksichtigt, so daß sich insgesamt ein Nutzwert von *2487,12* ergibt. Obwohl nun eine zusätzliche Regel positive Nutzenbeiträge liefert, ist der Nutzwert auf der Evaluierungsmenge weit geringer als der erwartete Nutzwert mit Kenntnis der Verteilung

der erklärenden Variablen. Dies verdeutlicht die oben bereits erkannte stark von der Trainingsmenge abweichende bedingte Verteilung der Caravan-Käufer (in Abhängigkeit von den erklärenden Variablen des Modells).

Vergleicht man die tatsächlich realisierten Deckungsbeiträge und den Nutzwert auf der Evaluierungsmenge mit den entsprechenden Kennzahlen des Entscheidungsbaums, so schneidet der Entscheidungsbaum bezüglich beider Kriterien *besser* ab. Dieses Resultat überrascht und wird daher in Abschnitt 6.2.4.3 ausführlich diskutiert.

Es stellt sich nun die Frage, ob die in den Regelprämissen enthaltenen Klauseln der Form „*Versicherung* $x = 0$“ (d.h. der Kunde besitzt keine Versicherung $x$) überhaupt für den Kauf einer Caravan-Versicherung verantwortlich sind. Schließlich besitzen nur die Klauseln mit den Attributen *c_car* und *c_fire* von $0$ verschiedene Werte. Dafür, daß auch die Klauseln mit $0$-Werten den Kauf einer Caravan-Versicherung beeinflussen, spricht, daß das entwickelte Verfahren zu jeder Attributkombination, $X$, auch alle Teilmengen, $Y \in Pot(X)$, generiert und bewertet. Das Modell *{c_car, c_fire}* ist also auf der Trainingsmenge schlechter bewertet worden als das Rough-Set-Modell 1. Es drängt sich die Frage auf, ob dies nur ein Phänomen der Trainingsmenge ist, oder ob es auch für die Evaluierungsmenge gilt, d.h. ob die tatsächlich realisierten Deckungsbeiträge und der Nutzwert auf der Evaluierungsmenge ebenfalls für das Rough-Set-Modell 1 sprechen. Hierzu sei die Evaluierung des Modells *{c_car, c_fire}* in Tabelle 6-22 betrachtet.

Vergleicht man die Evaluierung der *100/–10*-Bewertung mit der Evaluierung des Rough-Set-Modells 1, so erkennt man, daß nun tatsächlich – wie zuvor anhand des Data-Mining-Verfahrens begründet wurde – der Nutzwert auf der Trainingsmenge geringer ist als beim Rough-Set-Modell 1. Wie man an den Regeln mit positivem Nutzenbeitrag sieht, konnten im Rough-Set-Modell 1 durch die einschränkenden Klauseln der Form „*Versicherung* $x = 0$“ einige Nichtkäufer ausgeschlossen werden. Solche $0$-Klauseln können nicht pauschal durch ein automatisches Verfahren ignoriert werden, da sie z.T. inhaltlich begründet werden können, so z.B. „*c_third_party* $= 0$“. Es erscheint plausibel, daß der Kauf einer Caravan-Versicherung wahrscheinlicher ist, wenn der Kunde *keine* Privatversicherung bei einem anderen Versicherungsunternehmen besitzt, denn sonst würde er die Caravan-Versicherung u.U. ebenfalls dort abschließen. Möglicherweise können auch die Klauseln „*c_lorry* $= 0$“, „*c_surfboard* $= 0$“ und „*aggricultural* $= 0$“ inhaltlich

begründet werden. (Vielleicht sind Lastwagen- und Landmaschinen-Besitzer sowie Surfbrett-Urlauber nicht die typischen Caravan-Urlauber.)

| c_car | c_fire | caravan = [0;0] | caravan = [1;1] | Nutzen (*100/–10*) | Nutzen (*200/–10*) | Nutzen (*300/–10*) |
|---|---|---|---|---|---|---|
| *6* | *4* | *556* | *124* | *5420,05* | *16529,18* | *27638,31* |
| *6* | *3* | *307* | *52* | *1189,62* | *5534,73* | *9879,83* |
| *6* | *0* | *887* | *68* | *-3190,72* | *2590,45* | *8371,62* |
| *6* | *5* | *63* | *8* | *-205,72* | *252,71* | *711,15* |
| *5* | *4* | *78* | *7* | *-437,41* | *-62,33* | *312,75* |
| *0* | *4* | *428* | *20* | *-2896,42* | *-1456,81* | *-17,19* |
| | | | Nutzwert auf der Trainingsmenge (Summe der positiven Nutzenbeiträge): | *6609,67* | *24907,07* | *46913,66* |
| | | | Erwarteter Nutzwert ohne Kenntnis der neuen Verteilung: (Nutzwert auf der Trainingsmenge mal *4000/5822*): | *4541,17* | *17112,38* | *32231,99* |
| | | | Erwarteter Nutzwert mit Kenntnis der neuen Verteilung (vorsichtige Schätzung der erwarteten Deckungsbeiträge): | *4575,98* | *17195,78* | *32352,41* |
| | | | Abweichung der Nutzwerte mit und ohne Kenntnis der Verteilung: | *0,77%* | *0,49%* | *0,37%* |
| | | | Anzahl der fälschlicherweise selektierten Kunden: | *615* | *1262* | *1327* |
| | | | Anzahl der korrekt selektierten Kunden (Caravan-Käufer): | *104* | *148* | *152* |
| | | | **Tatsächlich realisierte Deckungsbeiträge:** | ***4250*** | ***16980*** | ***32330*** |
| | | | **Abweichung von dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung:** | ***-6,41%*** | ***-0,77%*** | ***0,30%*** |
| | | | **Nutzwert auf der Evaluierungsmenge:** | ***2407,97*** | ***11796,82*** | ***25381,01*** |
| | | | **Abweichung von dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung:** | ***-46,07%*** | ***-31,06%*** | ***-21,26%*** |

**Tabelle 6-22: Evaluierung des Rough-Set-Modells *{c_car, c_fire}* für *$1-\alpha=90\%$***

Auch bei Kenntnis der neuen Verteilung der Variablen *c_car* und *c_fire* in den Evaluierungsdaten schneidet das Rough-Set-Modell 1 besser ab, was für die Ähnlichkeit der Verteilung dieser Variablen zu der in der Trainingsmenge spricht. Diese Ähnlichkeit quantifiziert auch die geringe Abweichung der Nutzwerte mit und ohne Kenntnis der Verteilung (*0,77%*). Allerdings schneidet das Modell *{c_car, c_fire}* bezüglich der tatsächlich erzielten Deckungsbeiträge und bezüglich des Nutzwertes auf der Evaluierungsmenge geringfügig besser ab als das Rough-Set-Modell 1. Der Unterschied ist so gering, daß er durch den Zufallsfehler beim Aufteilen der Daten in Trainings- und Evaluierungsdaten erklärt werden kann.

Betrachtet wird nun die Evaluierung des Rough-Set-Modells 2 in Tabelle 6-23. Der Kauf einer Caravan-Versicherung wird hier mit *200* bewertet, so daß das Modell dem Entscheidungsbaum im zweiten Bewertungsszenario gegenübergestellt werden kann.

| c_thirdparty | c_car | c_lorry | c_surfboard | aggricultural | caravan = [0;0] | caravan = [1;1] | Nutzen (*200/–10*) |
|---|---|---|---|---|---|---|---|
| *0* | *6* | *0* | *0* | *0* | *1973* | *259* | *27996,44* |
| | | | | | | Nutzwert auf der Trainingsmenge (Summe der positiven Nutzenbeiträge): | *27996,44* |
| | | | | | | Erwarteter Nutzwert ohne Kenntnis der neuen Verteilung: (Nutzwert auf der Trainingsmenge mal *4000/5822*): | *18932,63* |
| | | | | | | Erwarteter Nutzwert mit Kenntnis der neuen Verteilung (vorsichtige Schätzung der erwarteten Deckungsbeiträge): | *19228,74* |
| | | | | | | Abweichung der Nutzwerte mit und ohne Kenntnis der Verteilung: | *1,56%* |
| | | | | | | Anzahl der fälschlicherweise selektierten Kunden: | *1376* |
| | | | | | | Anzahl der korrekt selektierten Kunden (Caravan-Käufer): | *157* |
| | | | | | | **Tatsächlich realisierte Deckungsbeiträge:** | ***17640*** |
| | | | | | | **Abweichung von dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung:** | ***-6,83%*** |
| | | | | | | **Nutzwert auf der Evaluierungsmenge:** | ***14444,08*** |
| | | | | | | **Abweichung von dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung:** | ***-23,71%*** |

**Tabelle 6-23: Evaluierung des Rough-Set-Modells 2**

Das Rough-Set-Modell 2 schneidet nun nicht nur auf den Trainings-, sondern auch auf den Evaluierungsdaten deutlich besser ab als das Rough-Set-Modell *{c_car, c_fire}*. Insbesondere der Nutzwert auf der Evaluierungsmenge liegt mit *14444,08* deutlich über dem Wert *11796,82*, der bei dem Modell *{c_car, c_fire}* erzielt wurde.

Vergleicht man die tatsächlich realisierten Deckungsbeiträge und den Nutzwert auf der Evaluierungsmenge mit den entsprechenden Kennzahlen des Entscheidungsbaums, so schneidet der Entscheidungsbaum bezüglich beider Kriterien *schlechter* ab.

Ähnliches gilt auch bei der nächsten Modellevaluierung. Um Wiederholungen zu vermeiden, werden die Evaluierungsdaten in Tabelle 6-24 kommentarlos widergegeben. Der Kauf einer Caravan-Versicherung wird hier mit *300* bewertet, so daß das Modell dem Entscheidungsbaum im dritten Bewertungsszenario gegenübrgestellt werden kann. Die Ergebnisse sind wiederum sowohl im Vergleich zu dem Entscheidungsbaum als auch im Vergleich zu dem Rough-Set-Modell *{c_car, c_fire}* wesentlich besser.

| c_car | c_lorry | c_aggricultural | surfboard | caravan = [0;0] | Caravan = [1;1] | Nutzen (*300/–10*) |
|---|---|---|---|---|---|---|
| *6* | *0* | *0* | *0* | *2044* | *262* | *52103,65* |
| | | | | | Nutzwert auf der Trainingsmenge (Summe der positiven Nutzenbeiträge): | *52103,65* |
| | | | | | Erwarteter Nutzwert ohne Kenntnis der neuen Verteilung: (Nutzwert auf der Trainingsmenge mal *4000/5822*): | *35797,77* |
| | | | | | Erwarteter Nutzwert mit Kenntnis der neuen Verteilung (vorsichtige Schätzung der erwarteten Deckungsbeiträge): | *35677,22* |
| | | | | | Abweichung der Nutzwerte mit und ohne Kenntnis der Verteilung: | *-0,34%* |
| | | | | | Anzahl der fälschlicherweise selektierten Kunden: | *1422* |
| | | | | | Anzahl der korrekt selektierten Kunden (Caravan-Käufer): | *157* |
| | | | | | **Tatsächlich realisierte Deckungsbeiträge:** | ***32880*** |
| | | | | | **Abweichung von dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung:** | ***-8,15%*** |
| | | | | | **Nutzwert auf der Evaluierungsmenge:** | ***28154,39*** |
| | | | | | **Abweichung von dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung:** | ***-21,35%*** |

**Tabelle 6-24: Evaluierung des Rough-Set-Modells 3**

### *6.2.4.3 Erklärung der Vergleichsergebnisse*

Die Evaluierung im Abschnitt zuvor hat ergeben, daß der Entscheidungsbaum in dem ersten Entscheidungsproblem, in dem der Kauf einer Caravan-Versicherung durch einen selektierten Kunden mit *100*, der nichtkauf mit *–10* bewertet wurde, besser abschneidet als das Rough-Set-Modell 1, in dem zweiten (*200/–10*) und dritten Entscheidungsproblem (*300/–10*) aber schlechter als die Rough-Set-Modelle 2 und 3.

Zu klären sind nun drei Fragen:

1. Warum hat im ersten Entscheidungsproblem das entwickelte Rough-Set-Verfahren die durch den Entscheidungsbaumalgorithmus generierte Variablenkombination *{c_car, c_fire, car}* nicht als beste Lösung ausgegeben?
2. Warum hat im zweiten und dritten Entscheidungsproblem der Entscheidungsbaumalgorithmus kein Modell mit den durch das Rough-Set-Verfahren vorgeschlagenen Variablenkombinationen *{c_third_party, c_car, c_lorry, c_surfboard, aggricultural}* und *{c_car, c_lorry, c_aggricultural, surfboard}* produziert?
3. Warum liegen die tatsächlich realisierten Deckungsbeiträge fast immer unter dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung, obwohl dieser eine *vorsichtige* Schätzung der erwarteten Deckungsbeiträge darstellt?

**Ad 1:** Konstruiert man das Modell *{c_car, c_fire, car}* nach dem hier verfolgten attributorientierten Ansatz mit einer minimalen Anwendbarkeit von *20* Kundendatensätzen und einer Konfidenzwahrscheinlichkeit von *90%,* so ergeben sich die in Tabelle 6-25 dargestellten Regeln. Man erkennt, daß die Regeln *31+10=41* Datensätze oder mehr abdecken. Dabei werden nur diejenigen Regeln aufgeführt, die in mindestens einer Parameterkonstellation zu einem positiven Nutzwert führen und somit überhaupt zur Entscheidungsunterstützung auf neue Datensätze angewendet werden.

| c_car | c_fire | car | caravan = [0;0] | caravan = [1;1] | Nutzen (*100/–10*) | Nutzen (*200/–10*) | Nutzen (*300/–10*) |
|---|---|---|---|---|---|---|---|
| *6* | *4* | *1* | *494* | *102* | *3963,35* | *12984,59* | *22005,82* |
| *6* | *4* | *2* | *60* | *22* | *1034,2* | *2719,84* | *4405,48* |
| *6* | *3* | *1* | *276* | *42* | *588,58* | *4014,55* | *7440,53* |
| *6* | *3* | *2* | *31* | *10* | *302,23* | *949,72* | *1597,21* |
| *6* | *5* | *1* | *50* | *6* | *-226,4* | *76,88* | *380,15* |
| *6* | *0* | *1* | *822* | *63* | *-2998,74* | *2320,59* | *7639,93* |
| *5* | *4* | *1* | *78* | *7* | *-437,41* | *-62,33* | *312,75* |
| | | | | Nutzwert auf der Trainingsmenge (Summe der positiven Nutzenbeiträge): | *5888,36* | *23066,17* | *43781,87* |
| | | | | Erwarteter Nutzwert ohne Kenntnis der neuen Verteilung: (Nutzwert auf der Trainingsmenge mal *4000/5822*): | *4045,59* | *15847,59* | *30080,30* |
| | | | | Erwarteter Nutzwert mit Kenntnis der neuen Verteilung (vorsichtige Schätzung der erwarteten Deckungsbeiträge): | *3944,29* | *15645,34* | *29787,48* |
| | | | | Abweichung der Nutzwerte mit und ohne Kenntnis der Verteilung: | *-2,50%* | *-1,28%* | *-0,97%* |
| | | | | Anzahl der fälschlicherweise selektierten Kunden: | *611* | *1208* | *1273* |
| | | | | Anzahl der korrekt selektierten Kunden (Caravan-Käufer): | *104* | *144* | *148* |
| | | | | **Tatsächlich realisierte Deckungsbeiträge:** | ***4290*** | ***16720*** | ***31670*** |
| | | | | **Abweichung von dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung:** | ***6,04%*** | ***5,51%*** | ***5,28%*** |
| | | | | **Nutzwert auf der Evaluierungsmenge:** | ***2079,59*** | ***11004,19*** | ***23462,23*** |
| | | | | **Abweichung von dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung:** | ***-48,60%*** | ***-30,56%*** | ***-22,00%*** |

**Tabelle 6-25: Evaluierung der Rough-Set-Lösung *{c_car, c_fire, car}* für *1–α=90%***

Wendet man bei einer Bewertung von *100/–10* die vier Regeln mit positivem Nutzwert auf die Evaluierungsdaten an, so werden *715* Kunden selektiert, von denen sich *104* als Caravan-Käufer herausstellen. Dies entspricht einem Deckungsbeitrag von *104·100 – (715-104)·10 = 10400 – 6110 = 4290*. Dies ist dem Rough-Set-Modell 1, welches einen Deckungsbeitrag von *4210* erwirtschaftete, zwar knapp überlegen. Bezüglich des

Nutzwertes auf der Evaluierungsmenge schneidet es allerdings wesentlich schlechter ab als das Rough-Set-Modell 1, was dem Rough-Set-Modell 1 einen höheren erwarteten Deckungsbeitrag in der Grundgesamtheit zuspricht. Entscheidend für die Beantwortung der Frage, warum Rough-Set-Modell 1 und nicht das Modell *{c_car, c_fire, car}* ausgegeben wurde, ist der höhere Nutzwert auf der Trainingsmenge (*6957,72* im Vergleich zu *5888,36*). Im Vergleich zu den ersten drei Regeln des Entscheidungsbaums, die ja dieselben Attribute umfassen, besteht der Unterschied hier darin, daß das Attribut *car* in jeder Regel des Rough-Set-Modells vorkommt. Das schlechtere Abschneiden des Rough-Set-Modells *{c_car, c_fire, car}* ist also in dem gewählten Modelltyp begründet, der reale Zusammenhänge nur grob approximieren kann. Es scheint sich dabei allerdings um ein zufälliges Phänomen zu handeln, denn ein Vergleich mit der Evaluierung des Rough-Set-Modells *{c_car, c_fire}* zeigt, daß der Nutzwert auf der Evaluierungsmenge ohne das Attribut *car* besser ist als mit diesem Attribut. Ohnehin wurde ja bereits festgestellt, daß es eine hohe inhaltliche Redundanz zu dem Attribut *c_car* aufweist.

**Ad 2:** Die Entscheidungsbaum-Software funktioniert derart, daß zunächst ein Entscheidungsbaum generiert wird, ohne erzielbare Kosten und Erlöse abzuschätzen. Erst bei der Anwendung des Baums auf die Evaluierungsdaten werden die ökonomischen Größen berücksichtigt. Wenn in einem Blatt das Verhältnis der Käufer zu den Nichtkäufen einen bestimmten Schwellwert, $d$, überschreitet, wird die Durchführung einer Marketingaktion empfohlen. Der Schwellwert $d$ muß umso höher gewählt werden, je ungünstiger das Verhältnis aus direkten Kosten, $dk$, und Erlösen, $er$, ist. Genauer muß nach den Ausführungen aus Abschnitt 3.3.2.1 gelten:

$$\frac{|O^T[(Käufer^+ = 1) \wedge Pr]|}{|O^T[Pr]|} > \frac{dk}{er} \quad \Leftrightarrow \quad |O^T[(Käufer^+ = 1) \wedge Pr]| \cdot er > |O^T[Pr]| \cdot dk \,.$$

Für die drei Entscheidungsprobleme sollte $d$ wie folgt bestimmt werden:

1. Mit $dk = 10$ und $er = 100+10 = 110$ ergibt sich: $d := dk/er = 10/110 = 0{,}09$.
2. Mit $dk = 10$ und $er = 200+10 = 210$ ergibt sich: $d := dk/er = 10/210 = 0{,}05$.
3. Mit $dk = 10$ und $er = 300+10 = 310$ ergibt sich: $d := dk/er = 10/310 = 0{,}03$.

So wurde beispielsweise im ersten Entscheidungsproblem der Tabelle 6-20 die dritte Regel des Entscheidungsbaums mit einem Käuferanteil von *0,07* nicht angewendet, da gilt $0,07 < d = 0,09$.

Erst nach abgeschlossener Lernphase kann in der verwendeten Software die eigentliche Bewertung der Handlungsergebnisse angegeben werden. D.h. in allen drei Entscheidungsproblemen wird derselbe Baum benutzt. Dies führt möglicherweise dazu, daß der Baum nicht optimal auf das Entscheidungsproblem abgestimmt ist und mag eventuell das schlechtere Abschneiden im Vergleich zu den Rough-Set-Modellen 2 und 3 erklären, denn diese wurden explizit für die jeweilige Bewertungslage erlernt.

Ein weiterer Grund für das schlechtere Abschneiden des Entscheidungsbaumverfahrens wurde bereits diskutiert. Entscheidungsbaumverfahren verwenden i.d.R. kein Backtracking, so daß sie besonders bei Problemstellungen mit vielen Attributen oder großen Domänen suboptimale Lösungen liefern.

**Ad 3:** Eine mögliche Erklärung für dieses ungewöhnliche Phänomen besteht darin, daß die bedingte Verteilung der Käufer bzw. Nichtkäufer in den Evaluierungsdaten nicht der in den Trainingsdaten entspricht. Dies hängt zum einen von der Bedingung ab, d.h. in Abhängigkeit welcher erklärender Variablen die Verteilung der Käufer und Nichtkäufer betrachtet wird. So hat die Evaluierung der Rough-Set-Lösung *{c_car, c_fire, car}* ein anderes Bild ergeben – hier war die Deckungsbeitragsschätzung vorsichtig genug.

In Abhängigkeit der in den Rough-Set-Modellen 1, 2 und 3 verwendeten Variablen lag der tatsächlich realisierte Deckungsbeitrag unter der vorsichtigen Schätzung des erwarteten Deckungsbeitrags, so daß bezüglich dieser Variablen die Trainings- und Evaluierungsmenge verschiedene *caravan*-Verteilungen aufweisen. Dies kann vorkommen, wenn der Kundendatenbestand sehr ungünstig in Trainings- und Evaluierungsdaten aufgeteilt wurde. Ob dem so ist, kann überprüft werden, indem beide Datenmengen zusammengefügt und zufällig wieder aufgetrennt werden. Dies wird hier derart durchgeführt, daß wieder *5822* Datensätze in der neuen Trainingsmenge enthalten sind und *4000* in der neuen Evaluierungsmenge. Die Anzahl der Käufer in der Evaluierungsmenge beträgt nun *237* statt *238*, so daß der Unterschied vernachlässigbar ist.

Tatsächlich stellt sich das vermutete Ergebnis ein (vgl. Tabelle 6-26 bis Tabelle 6-28): Für alle drei betrachteten Modelle sind die auf der Evaluierungsmenge erzielten

Ergebnisse besser als bei der ersten Evaluierung im Abschnitt zuvor. Außerdem liegen die tatsächlich realisierten Deckungsbeitrage – wie eigentlich auch bei der ersten Evaluierungsphase zu erwarten war – *über* der vorsichten Schätzung der erwarteten Deckungsbeiträge. Demgegenüber sind die auf der Trainingsmenge erzielten Ergebnisse schlechter als bei der ersten Evaluierung. Würde man erneut die Lernphase anstoßen, so würden wahrscheinlich *bessere Modelle gefunden werden*. Es wäre dann zu prüfen, ob diese Modelle den besten auffindbaren Entscheidungsbaum in allen Entscheidungsproblemen schlagen würden. Darauf wird hier verzichtet, denn erstens schnitt das Verfahren schon in den durchgeführten Versuchen *bei zwei von drei Problemstellungen* besser ab als das Entscheidungsbaumverfahren, und zweitens können die wichtigsten Schlußfolgerungen aus der Testphase bereits jetzt gezogen werden (vgl. den folgenden Abschnitt 6.2.5).

| c_car | c_disabi-lity | c_fire | c_surf-board | third_party | van | lorry | trailer | private_accident | caravan = [0;0] | caravan = [1;1] | Nutzen (*100/–10*) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *6* | *0* | *4* | *0* | *0* | *0* | *0* | *0* | *0* | *535* | *109* | *4208,08* |
| *6* | *0* | *3* | *0* | *0* | *0* | *0* | *0* | *0* | *315* | *45* | *465,11* |
| *6* | *0* | *5* | *0* | *0* | *0* | *0* | *0* | *0* | *54* | *11* | *133,70* |
| Nutzwert auf der Trainingsmenge (Summe der positiven Nutzenbeiträge): | | | | | | | | | | | *4806,88* |
| Erwarteter Nutzwert ohne Kenntnis der neuen Verteilung: (Nutzwert auf der Trainingsmenge mal *4000/5822*): | | | | | | | | | | | *3302,56* |
| Erwarteter Nutzwert mit Kenntnis der neuen Verteilung (vorsichtige Schätzung der erwarteten Deckungsbeiträge): | | | | | | | | | | | *3325,33* |
| Abweichung der Nutzwerte mit und ohne Kenntnis der Verteilung: | | | | | | | | | | | *0,69%* |
| Anzahl der fälschlicherweise selektierten Kunden: | | | | | | | | | | | *582* |
| Anzahl der korrekt selektierten Kunden (Caravan-Käufer): | | | | | | | | | | | *125* |
| **Tatsächlich realisierte Deckungsbeiträge:** | | | | | | | | | | | ***6680*** |
| **Abweichung von dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung:** | | | | | | | | | | | ***102,27%*** |
| **Nutzwert auf der Evaluierungsmenge:** | | | | | | | | | | | ***4677,17*** |
| **Abweichung von dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung:** | | | | | | | | | | | ***41,62%*** |

**Tabelle 6-26: Erneute Evaluierung des Rough-Set-Modells 1**

| c_thirdparty | c_car | c_lorry | c_surfboard | aggricultural | caravan = [0;0] | caravan = [1;1] | Nutzen (*200/–10*) |
|---|---|---|---|---|---|---|---|
| *0* | *6* | *0* | *0* | *0* | *2006* | *243* | *24576,48* |
| | | | | | | Nutzwert auf der Trainingsmenge (Summe der positiven Nutzenbeiträge): | *24576,48* |
| | | | | | | Erwarteter Nutzwert ohne Kenntnis der neuen Verteilung: (Nutzwert auf der Trainingsmenge mal *4000/5822*): | *16885,25* |
| | | | | | | Erwarteter Nutzwert mit Kenntnis der neuen Verteilung (vorsichtige Schätzung der erwarteten Deckungsbeiträge): | *16566,45* |
| | | | | | | Abweichung der Nutzwerte mit und ohne Kenntnis der Verteilung: | *-1,89%* |
| | | | | | | Anzahl der fälschlicherweise selektierten Kunden: | *1343* |
| | | | | | | Anzahl der korrekt selektierten Kunden (Caravan-Käufer): | *173* |
| | | | | | | **Tatsächlich realisierte Deckungsbeiträge:** | ***21170*** |
| | | | | | | **Abweichung von dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung:** | ***25,38%*** |
| | | | | | | **Nutzwert auf der Evaluierungsmenge:** | ***17837,13*** |
| | | | | | | **Abweichung von dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung:** | ***5,64%*** |

**Tabelle 6-27: Erneute Evaluierung des Rough-Set-Modells 2**

| c_car | c_lorry | c_aggricultural | surfboard | caravan = [0;0] | caravan = [1;1] | Nutzen (*300/–10*) |
|---|---|---|---|---|---|---|
| *6* | *0* | *0* | *0* | *2071* | *245* | *46907,61* |
| | | | | | Nutzwert auf der Trainingsmenge (Summe der positiven Nutzenbeiträge): | *46907,61* |
| | | | | | Erwarteter Nutzwert ohne Kenntnis der neuen Verteilung: (Nutzwert auf der Trainingsmenge mal *4000/5822*): | *32227,83* |
| | | | | | Erwarteter Nutzwert mit Kenntnis der neuen Verteilung (vorsichtige Schätzung der erwarteten Deckungsbeiträge): | *31778,09* |
| | | | | | Abweichung der Nutzwerte mit und ohne Kenntnis der Verteilung: | *-1,40%* |
| | | | | | Anzahl der fälschlicherweise selektierten Kunden: | *1395* |
| | | | | | Anzahl der korrekt selektierten Kunden (Caravan-Käufer): | *174* |
| | | | | | **Tatsächlich realisierte Deckungsbeiträge:** | ***38250*** |
| | | | | | **Abweichung von dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung:** | ***18,69%*** |
| | | | | | **Nutzwert auf der Evaluierungsmenge:** | ***33306,89*** |
| | | | | | **Abweichung von dem erwarteten Nutzwert ohne Kenntnis der neuen Verteilung:** | ***3,35%*** |

**Tabelle 6-28: Erneute Evaluierung des Rough-Set-Modells 3**

### 6.2.5 Zusammenfassung der Anwendungen auf Realdaten

Im Rahmen der Anwendungen auf die Caravanversicherungsdaten hat sich das entwickelte Nutzenkriterium bewährt. Nachdem die Trainings- und Evaluierungsdaten

zusammengeführt und zufällig wieder aufgespalten wurden, stellten die ermittelten Nutzwerte tatsächlich Untergrenzen für die erwarteten Deckungsbeiträge dar, welche nicht unterschritten wurden. Die Berücksichtigung eines Sicherheitspuffers bewirkt, daß ein Modell erlernt wird, welches auch bei kleineren Marketingaktionen oder leicht veränderten Stichprobenverteilungen eine gewisse Erfolgsgarantie verspricht. Es gewährleistet (unter den diskutierten Einschränkungen) die Übertragbarkeit der Ergebnisse auf die Grundgesamtheit. Der Sicherheitspuffer hat starke Auswirkungen auf die Modellgenerierung und ist daher von erheblicher Relevanz für das Data-Mining-Verfahren. Diese Relevanz resultiert aus der starken Abhängigkeit von der Stichprobengröße. Da bei regelorientierten Verfahren die Stichproben aus den durch die Prämissen, $Pr$, abgedeckten Objekten bestehen, besitzen sie nicht etwa den Umfang $|O^T|$, sondern nur $|O^T[Pr]|$.[448] Dadurch hat man es häufig mit relativ kleinen Stichproben zu tun, so daß die tatsächlich realisierten Deckungsbeiträge erheblich von dem Stichprobenmittel abweichen können. Eine vorsichtige Schätzung der realisierbaren Deckungsbeiträge ist hier also angebracht. Denn durch die veränderte Deckungsbeitragsschätzung ergibt sich auch ein anderes optimales Modell. Der genaue Einfluß der Stichprobengröße auf den Sicherheitspuffer wurde bereits in Abbildung 6-2 und Abbildung 6-3 (vgl. S. 267) verdeutlicht.

Wie die Rough-Set-Modelle 2 und 3 im Vergleich zu dem Entscheidungsbaum gezeigt haben, ist es wichtig, daß die direkten Kosten und Erlöse der möglichen Handlungen bereits beim Lernen in das Modell einfließen, damit das ökonomisch optimale Modell erlernt wird.

Das entwickelte Suchverfahren hat sich in den praktischen Anwendungen bewährt. Seine Laufzeiten liegen zwar über denen des Entscheidungsbaumverfahrens, aber der Suchraum wird wesentlich besser erforscht, so daß (trotz des ungenaueren Modelltyps) die Ergebnisse ähnlich gut oder besser als die des Entscheidungsbaumverfahrens waren.

Bei attributorientierten Modellen stellt sich generell die Frage, ob nicht auch in einigen Regeln eine kleinere Menge von Klauseln zur Erklärung des abhängigen Phänomens

[448] Da für alle Objekte, die durch eine Regelprämisse abgedeckt werden, dieselbe Handlung, $h_i$, empfohlen wird, gilt für den Stichprobenumfang: $|O^T[Pr]| = |O^T[Pr \wedge (h=h_i)]|$.

ausreichen würde. Diese Frage kann durch das entwickelte Verfahren nicht beantwortet werden. Vielmehr müßten neue Hypothesen mit kleineren Regeln gebildet und überprüft werden. Ein Verkleinern der Regeln wäre vor allem im Zusammenhang mit erklärenden Fragestellungen wünschenswert, zu deren Beantwortung eine intellektuelle Interpretation der Regeln erforderlich ist. Das hier entwickelte Verfahren unterstützt aber vor allem Standardentscheidungen im Unternehmen, so daß es sich an dem Nutzen messen lassen muß, den es in neuen Entscheidungssituationen erzielt. Ein Verkleinern der Regeln kann hier auch bei der Generierung von Prognose- oder Entscheidungsmodellen hilfreich sein, falls dadurch das Realsystem genauer approximiert werden kann.[449] Die Evaluierung des Rough-Set-Modells 1 im Vergleich zu dem diskutierten Entscheidungsbaum hat gezeigt, daß das Streichen von überflüssigen Klauseln in den Regelprämissen den Nutzwert des Modells erhöhen kann.

[449] Dieser Aspekt wurde bereits in Abschnitt 5.2.2 diskutiert.

# 7 Fazit und Ausblick

Die Ziele dieser Arbeit bestanden darin,

- ⇨ eine Systematik zur Lösung betriebswirtschaftlicher Problemstellungen zu konzipieren (**Ziel 1: „Lösungsschema konzipieren“**),
- ⇨ ein Data-Mining-Verfahren zur Lösung der Problemstellungen zu entwickeln (**Ziel 2: „Lösungsverfahren entwickeln“**) und
- ⇨ durch Anwendung des Verfahrens auf eine Datenbasis dessen praktische Eignung zu evaluieren (**Ziel 3: „Verfahren evaluieren“**).

Zur Erfüllung dieser Ziele wurden in Kapitel 2 zunächst die notwendigen technischen und praktischen Grundlagen des Data Mining eingeführt. Während andere Arbeiten zu den Grundlagen des Data Mining i.d.R. ausgewählte klassische Verfahren beschreiben und Problemklassen suchen, für die sich die Verfahren mehr oder weniger eignen, wurde in dieser Untersuchung ein anderer Weg verfolgt. So wurden Data-Mining-Verfahren in ihre Komponenten zerlegt und für jede Komponente Ansätze zu ihrer Gestaltung aufgezeigt, die je nach Problemstellung individuell zusammengesetzt und so optimal auf das Problem abgestimmt werden können. Für die Komponente des Suchverfahrens wurde ein allgemeingültiges Verfahrensschema aufgezeigt, in das sich bestehende Verfahren als konkrete Schemainstanzen einordnen lassen. Damit leistet die hier gewählte Darstellung Hilfestellung bei der Auswahl bestehender und der Entwicklung neuer Data-Mining-Verfahren.

Der betriebswirtschaftliche Bezug wurde in Kapitel 3 hergestellt. Andere Arbeiten zu Anwendungspotentialen des Data Mining konzentrieren sich auf ausgewählte Anwendungsbereiche, z.B. das Ergebniscontrolling oder die Kreditwürdigkeitsanalyse. In dieser Untersuchung stand nicht ein konkreter Anwendungsbereich im Vordergrund, sondern eine strukturierte Darstellung der Planungs- und Kontrollprozesse, in die sich die einzelnen Data-Mining-Anwendungen einordnen ließen. Diese generalisierte Betrachtung bot den Vorteil, daß sich daraus ein allgemeingültiges Problemlösungsschema ableiten ließ. Damit wurde in Kapitel 3 auch das erste Ziel (**„Lösungsschema entwickeln“**) erreicht. Das entwickelte Lösungsschema zeigt in Abhängigkeit von der Art und Weise der Entscheidung sowie von den zur Verfügung stehenden Trainingsdaten und

Data-Mining-Verfahren die möglichen Unterstützungspotentiale des Data Mining im Planungs- und Kontrollprozeß auf. Dazu zählt die Empfehlung eines geeignet aufgebauten Data-Mining-Modelltyps und eines geeigneten Bewertungskonzeptes. Dies konnte sowohl für Beschreibungs-, Erklärungs-, Prognose- und Entscheidungsmodelle geschehen. Ein vergleichbarer betriebswirtschaftliche Bezug wird von keiner anderen Untersuchung hergestellt.

Der vorgestellte Versuch, ein allgemeingültiges Problemlösungsschema zu entwickeln, muß wie folgt relativiert werden: Eigene und andere[450] Forschungen haben gezeigt, daß Data-Mining-Verfahren häufig nicht direkt auf die vorliegende Problemstellung anwendbar sind. Wenn die Problemstellung nicht durch den zugrundeliegenden Modelltyp abgebildet werden kann, muß der Modelltyp und die Interessentheitsbewertung (und unter Umständen auch das Suchverfahren mit dem Datenzgriff) modifiziert bzw. neu entwickelt werden. Daher macht es Sinn, individuelle Data-Mining-Applikationen zur direkten oder indirekten Unterstützung spezieller Standardentscheidungen, wie z.B. der Selektion von Zielkunden für Direktmarketingmaßnahmen, zu entwickeln. Diese Applikationen müßten dann speziell auf das relevante Entscheidungsfeld zugeschnittene Modelle und Bewertungskonzepte sowie ein geeignetes Lösungsverfahren bereitstellen. Als Grundlage hierzu können die in dieser Untersuchung erarbeiteten Konzeptionen dienen.

Um das zweite Ziel (**„Lösungsverfahren entwickeln“**) zu erreichen, wurden in Kapitel 4 Anforderungen an das zu enwickelnde Verfahren aufgestellt und untersucht, inwieweit bestehende Verfahren diese Anforderungen erfüllen. Aus Aufwandsgründen konnte mit den Entscheidungsmodellen nur einer der vier untersuchten Modelltypen berücksichtigt werden. Dieser Modelltyp ist innovativ und bietet den engsten Entscheidungsbezug.

Auf den Anforderungen und den bestehenden Ansätzen im Data Mining aufbauend konnte in Kapitel 5 eine an die konjunktive Normalform angelehnte Wissensrepräsentation und ein an das allgemeine Schema von Suchstrategien angelehntes Suchverfahren entwickelt und ein Bewertungskonzept aus dem zuvor entwickelten Problemlösungsschema übernommen werden. Der Zugriff auf externe Datenbanken mußte aus

[450] Vgl. SÄUBERLICH (2000), S. 194.

Laufzeitgründen verworfen werden. Andere Arbeiten, die die Entwicklung eines Data-Mining-Verfahrens dokumentieren, beschränken sich zumeist auf das Suchverfahren, das nur eine der vier Komponenten von Data-Mining-Verfahren darstellt. Die besonderen Neuigkeitsaspekte dieser Arbeit liegen in einem an betriebswirtschaftlichen Entscheidungen angelehnten Bewertungskonzept. Das konzipierte attributorientierte Suchverfahren durchforstet den durch den Modelltyp (allgemeiner Regelmengen in konjunktiven Normalform) aufgespannten Lösungsraum in großen Schritten, was relativ schnell zu guten Attributkombinationen führt, aber zu Lasten der erreichbaren Approximationsgüte geht.

Kapitel 6 verfolgt das dritte aufgestellte Ziel (**„Verfahren evaluieren“**) und demonstriert die praktische Eignung des entwickelten Verfahrens. Die Laufzeit des Verfahrens ist – obwohl auf ihre Begrenzung großer Wert gelegt wurde – schlechter als die eines zum Vergleich eingesetzten Entscheidungsbaumverfahrens. Dafür sind ermittelten Modelle trotz der erwarteten geringeren Approximationsgüte vergleichbar mit denen des Entscheidungsbaumverfahrens. In jedem Fall wird der zu erwartende ökonomische Erfolg des Modells vorsichtiger abgeschätzt; trotz dieser vorsichtigen Schätzung konnte das Modell in zwei von drei Entscheidungsproblemen den Entscheidungsbaum in dem erzielten Gesamtdeckungsbeitrag übertreffen.

Zukünftige Untersuchungen können an den in Kapitel 5 zu den vier Verfahrenskomponenten diskutierten Kritikpunkten ansetzen und das entwickelte Data-Mining-Verfahren in diesen Punkten verbessern. So kann beispielsweise die *erreichbare Güte der Approximation* verbessert werden, indem an das hier vorgestellte Verfahren ein Verfahren anschließt, welches die Regeln der generierten Regelmengen einzeln modifiziert und so die Regelmenge verbessert. Dieser Ansatz wurde bereits in Abschnitt 6.2.5 diskutiert.

Ein vielversprechender Ansatzpunkt zur Verbesserung der *Laufzeit* und damit eine neue Möglichkeit, den hier verworfenen Datenzugriff doch noch zu realisieren, ergibt sich, wenn man die eingangs gezogenen Grenzen, die Realisierung auf PC-Basis vorzunehmen, aufhebt, und Client-Server-Systeme betrachtet: Ein Ansatzpunkt zur Beschleunigung des Datenzugriffs mit speziellen "Data-Mining-Servern" wurde bereits in Abschnitt 5.1.2 diskutiert.

Neben diesen technischen Überlegungen sei auch ein Ausblick auf betriebswirtschaftliche Entwicklungen gestattet. Potentiale zur Unterstützung von Entscheidungsprozessen bietet die Entwicklung von geeigneten Verfahren für Data-Mining-Beschreibungs-, -Erklärungs- und -Prognosemodelle. Besonders innovativ wären zum einen Verfahren zur Generierung von Erklärungsmodellen, welche die in Abschnitt 3.3.2.3 konzipierten Bewertungsvorschriften nutzen, da diese Vorschriften im Gegensatz zu den üblichen Interessantheitsmaßen ökonomische und statistische Überlegungen sowie ein Unbekanntheitsmaß integrieren. Zum anderen bietet die Entwicklung eines Verfahrens Innovationspotentiale, welches Prognosemodelle generiert, die die gesamte Wahrscheinlichkeitsverteilung einer Zufallsgröße ausgäben. Ein solches Verfahren würde einen höheren Grad an Entscheidungsunterstützung bieten als die üblicherweise verwendeten „Punktprognosen“, die den Wert mit dem geringsten erwarteten Prognosefehler vorhersagen, da nach den Ausführungen aus Abschnitt 3.1 die gesamte Wahrscheinlichkeitsverteilung der für jede Alternative erreichbaren Zielbeiträge in die Entscheidungsfindung eingeht. Im Vergleich zu den in dieser Arbeit generierten Entscheidungsmodellen könnten solche Prognosemodelle dann in komplexeren Planungsprozessen eingesetzt werden, die externe Überlegungen mit in die Entscheidungsfindung einbeziehen.

Darüber hinaus können weitere Untersuchungen an das konzipierte Problemlösungsschema anknüpfen. Zunächst müßten alle in Abbildung 3-13 dargestellten Bewertungsansätze in eine Data-Mining-Verfahren integriert werden. Wenn die entsprechenden Data-Mining-Verfahren erst einmal zur Verfügung stünden, könnten die in Abbildung 3-12 dargestellten Anwendungsfälle 1 bis 10 auf ihre Praxisrelevanz überprüft werden, indem für jeden dieser Fälle eine entsprechende praktische Fallstudie projektiert wird. Für die Generierung von Entscheidungsmodellen könnten dann auch komplexere Nutzwert-Funktionen optimiert werden, wie sie beispielsweise bei der Berücksichtigung von unterschiedlichen Budgetinanspruchnahmen bei der Kundenselektion benötigt werden. Bei Verzicht auf die Neuentwicklung geeigneter Data-Mining-Verfahren könnten die Fallstudien der Anwendungsfälle 1 bis 10 alternativ mit bereits existierenden Verfahren bearbeitet werden, wobei dann aber die entsprechenden, in Abschnitt 4.5 diskutierten Einschränkungen in Kauf genommen werden müßten (von denen die

unzureichenden Bewertungskonzepte sicher die gravierendsten Einschränkungen darstellen).

Als Fazit kann festgehalten werden, daß nicht nur für die technische, sondern auch für die betriebswirtschaftlich ausgerichtete Data-Mining-Forschung noch Potentiale brachliegen. Um insbesondere die betriebswirtschaftlichen Potentiale ausschöpfen, ist es aber – im Gegensatz zu dem überwiegendenen Anteil der bertiebswirtschftlich orientierten Forschungsprojekte – notwendig, sich von den Restriktionen bestehender Data-Mining-Verfahren zu lösen und problemorientierte Modelltypen und Bewertungskonzepte zu entwickeln und in ein geeignetes Lösungsverfahren zu integrieren.

# Literaturverzeichnis

**Agrawal, Rakesh; Mannila, Heikki; Srikant, Ramakrishnan; Toivonen, Hannu; Verkamo, A. Inkeri:** Fast Discovery of Association Rules, in: Fayyad, Usama M.; Piatetsky-Shapiro, Gregory; Smyth, Padhraic; Uthurusamy, Ramasamy (Hrsg.): Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press: Menlo Park, California et al., 1996, S. 307-328.

**Arndt, Dirk; Gersten, Wendy; Wirth, Rüdiger:** Kundenprofile zur Prognose der Markenaffinität im Automobilsektor, in: Hippner, Hajo; Küsters, Ulrich; Meyer, Matthias; Wilde, Klaus (Hrsg.): Handbuch Data Mining im Marketing – Knowledge Discovery in Marketing Databases, Vieweg: Braunschweig et al., 2001, S. 591-606.

**Backhaus, Klaus; Weiber, Rolf:** Entwicklung einer Marketing-Konzeption mit SPSS/PC$^+$, Springer: Berlin et al., 1989.

**Backhaus, Klaus; Erichson, Bernd; Plinke, Wulff; Weiber, Rolf:** Multivariate Analysemethoden – Eine Anwendungsorientierte Einführung, 9. Aufl., Springer: Berlin et al., 2000.

**Bäck, Thomas; Schütz, Martin:** Evolution Strategies for Mixed-Integer Optimization of Optical Multilayer Systems, in: McDonnel, John R.; Reynolds, Robert G.; Fogel, David B. (Hrsg.): Evolutionary Programming IV – Proceedings of the Fourth Annual Conference on Evolutionary Programming, MIT Press: Cambridge, Massachusetts et al., 1995, S. 33-51.

**Bäck, Thomas; Schütz, Martin:** Evolutionäre Algorithmen im Data Mining, in: Hippner, Hajo; Küsters, Ulrich; Meyer, Matthias; Wilde, Klaus (Hrsg.): Handbuch Data Mining im Marketing – Knowledge Discovery in Marketing Databases, Vieweg: Braunschweig et al., 2001, S. 403-426.

**Bäck, Thomas; Hammel, Ulrich; Schwefel, Hans-Paul:** Evolutionary Computation: Comments on the History and Current State, in: IEEE Transactions on Evolutionary Computation, Vol. 1 (1997), No. 1, S. 1-15.

**Bamberg, Günter; Baur, Franz:** Statistik, 10. Aufl., Oldenbourg: München, 1998.

**Bausch, Thomas:** Gewinnoptimale Kundenselektion im Direkt-Marketing, in: Marketing ZFP, Jahrgang 13 (1991), H. 2, S. 86-96.

**Bensberg, Frank; Weiß, Thorsten:** Web Log Mining als Marktforschungsinstrument für das World Wide Web, in: Wirtschaftsinformatik, Jahrg. 41 (1999), H. 5, S. 426-432.

**Bissantz, Nicolas:** CLUSMIN – Ein Beitrag zur Analyse von Daten des Ergebniscontrollings mit Datenmustererkennung (Data Mining); Diss., Erlangen-Nürnberg, 1996.

**Bissantz, Nicolas; Braun, Gerhard:** Data Mining im Versandhandel, in: Datenbank Fokus, (1998), H. 2, S. 16-20.

**Baun, Susanne:** Neuronale Netze in der Aktienkursprognose, in: Rehkugler, Heinz; Zimmermann, Hans Georg (Hrsg.): Neuronale Netze in der Ökonomie – Grundlagen und finanzwirtschaftliche Anwendungen, Vahlen: München, 1994, S. 131-208.

**Bonne, Thorsten:** Kostenorientierte Klassifikationsanalyse – Theorie und betriebswirtschaftliche Anwendungen; Eul: Lohmar et al., 2000.

**Bonne, Thorsten; Arminger, Gerhard:** Der Einsatz automatischer Klassifikation zur Bonitätsprüfung im Direktvertrieb, in: Hippner, Hajo; Küsters, Ulrich; Meyer, Matthias; Wilde, Klaus (Hrsg.): Handbuch Data Mining im Marketing – Knowledge Discovery in Marketing Databases, Vieweg: Braunschweig et al., 2001, S. 653-670.

**Borgelt, Christian; Kruse, Rudolf; Lindner, Guido:** Lernen probabilistischer und possibilistischer Netze aus Daten - Theorie und Anwendungen, in: Zeitschrift Künstliche Intelligenz KI, (1998), H. 1, S. 11-17.

**Brachman, Ronald J.; Anand, Tej:** The Process of Knowledge Discovery in Databases - A Human-Centered Approach, in: Fayyad, Usama M.; Piatetsky-Shapiro, Gregory; Smyth, Padhraic; Uthurusamy, Ramasamy (Hrsg.): Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press: Menlo Park, California et al., 1996, S. 37-57.

**Brachman, Ronald, J.; Khabaza, Tom; Klösgen, Willi; Piatetsky-Shapiro, Gregory; Simoudis, Evangelos:** Mining Business Databases, in: Communications of the ACM, Vol. 39 (1996), No. 11, S. 42-48.

**Brin, Sergey; Motwani, Rajeev; Silverstein, Craig:** Beyond Market Baskets: Generalizing Association Rules to Correlations, in: SIGMOD Record, Vol. 26 (1997), No. 2, S. 265-276.

**Breitner, C.A.; Lockemann, P.C.; Schlösser, J.A.:** Die Rolle der Informationsverwaltung im KDD-Prozeß, in: Nakhaeizadeh, Gholamreza: Data Mining - Theoretische Aspekte und Anwendungen, Physica-Verlag: Heidelberg, 1998, S. 34-60.

**Buhmann, Joachim M.**: Knowledge Discovery - Wie können wir Muster und Strukturen in Daten zuverlässig erkennen?, in: Zeitschrift Künstliche Intelligenz KI, (1998), H. 1, S. 37.

**Cap, Clemens H.:** Wirtschaftliche Anwendungen - Die neuen Grand Challenges der Informatik?, in: Praxis der Wirtschaftsinformatik HMD, Jahrg. 35 (1998), H. 203 S. 50-57.

**Chapman, Pete et al.:** CRISP-DM 1.0 – Step-by-step data mining guide, in: Internet www.crisp-dm.org, Zugriff am 11.09.2000.

**Chaudhuri, Surajit; Madigan, David (Hrsg.):** KDD-99 - The Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego/Kalifornien/USA, ACM: New York, 1999.

**Chmielewicz, Klaus:** Forschungskonzeptionen der Wirtschaftswissenschaft, 2. Aufl., Poeschel: Stuttgart, 1979.

**Clark, Peter; Niblett, Tim:** The CN2 Induction Algorithm, in: Machine Learning, Vol. 3 (1989), No.4, S. 261-283.

**Decker, Karsten M.; Focardi, Sergio:** Technology Overview - A Report on Data Mining, Technical Report (CSCS TR-95-02), Swiss Scientific Computing Center: Zürich, 1995.

**Decker, Reinhold; Temme, Thorsten:** CHAID als Instrument der Werbemittelgestaltung und Zielgruppenbestimmung im Marketing, in: Hippner, Hajo; Küsters, Ulrich; Meyer, Matthias; Wilde, Klaus (Hrsg.): Handbuch Data Mining im Marketing – Knowledge Discovery in Marketing Databases, Vieweg: Braunschweig et al., 2001, S. 671-683.

**Dehaspe, Luc; Toivonen, Hannu:** Discovery of Relational Association Rules, in: Dzeroski, Saso; Lavrac, Nada (Hsrg.): Relational Data Mining, Springer: Berlin et al., 2001, S. 189-212.

**De Raedt, Luc; Blockeel, Hendrick; Dehaspe, Luc; Van Laer, Wim:** Three Companions for Data Mining in First Order Logic, in: Dzeroski, Saso; Lavrac, Nada (Hsrg.): Relational Data Mining, Springer: Berlin et al., 2001, S. 105-139.

**Determann, Lorenz; Rey, Michael:** Chancen und Grenzen des Data Mining im Controlling, in: Controlling, Jahrg. 8 (1999), H. 3, S. 143-147.

**Domingos, Pedro:** Occam's Two Razors - The Sharp and the Blunt, in: Agrawal, Rakesh; Stolorz, Paul; Piatetsky-Shapiro, Gregory: The Fourth International Conference on Knowledge Discovery & Data Mining, Proceedings, AAAI Press: Menlo Park, California, 1998, S. 37-42.

**Dueck, Gunter:** New Optimization Heuristics – The Great Deluge Algorithm and the Record-to-Record Travel, in: Journal of Computational Physics, Vol.104 (1993), S. 86-92.

**Dueck, Gunter; Scheuer, Tobias:** Threshold Accepting - A General Purpose Optimization Algorithm Appearing Superior to Simulated Annealing, in: Journal of Computational Physics, Vol. 90 (1990), S. 161-175.

**Düsing, Roland**: Betriebswirtschaftliche Anwendungsbereiche Konnektionistischer Systeme, Duisburger Betriebswirtschaftliche Schriften, Band 14, S+W Steuer- und Wirtschaftsverlag: Hamburg, 1997.

**Dzeroski, Saso; Lavrac, Nada (Hsrg.):** Relational Data Mining, Springer: Berlin et al., 2001.

**Elkan, Charles:** Coil Challenge 2000 Entry, in: Internet www.liacs.nl/~putten/library/cc2000/ report2.html; Zugriff am 28.01.2002.

**Engels, Robert:** Component-based user guidance in knowledge discovery and data mining, Infix: Sankt Augustin, 1999.

**Ester, Martin; Sander, Jörg:** Knowledge Discovery in Databases – Techniken und Anwendungen, Springer: Berlin et al., 2000.

**Fayyad, Usama M.; Piatetsky-Shapiro, Gregory; Smyth, Padhraic:** From Data Mining to Knowledge Discovery: An Overview, in: Fayyad, Usama M.; Piatetsky-Shapiro, Gregory; Smyth, Padhraic; Uthurusamy, Ramasamy (Hrsg.): Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press: Menlo Park, California et al., 1996, S. 1-34.

**Fayyad, Usama M.; Piatetsky-Shapiro, Gregory; Smyth, Padhraic; Uthurusamy, Ramasamy (Hrsg.):** Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press: Menlo Park, California et al., 1996.

**Fischer, Thomas:** Computergestützte Warenkorbanalyse - dargestellt auf der Grundlage von Scanningdaten des Lebensmitteleinzelhandels unter besonderer Berücksichtigung einer selbsterstellten Analysesoftware, Lang: Frankfurt a.M. et al., 1993.

**Frawley, W.J.; Piatetsky-Shapiro, G.; Matheus, C.J.:** Knowledge Discovery in Databases: An Overview, in: Frawley, W.J.; Piatetsky-Shapiro, G.; Matheus, C.J. (Hrsg.): Knowledge Discovery in Databases, AAAI Press/MIT Press: Menlo Park, California et al., 1991, S. 1-27.

**Freitas, Alex A.:** Understanding the Crucial Differences Between Classification and Discovery of Association Rules – A Position Paper, in: SIGKDD Explorations Vol. 2 (2000), No. 1, S. 65-69.

**Freter, Hermann:** Marktsegmentierung, Kohlhammer: Stuttgart, 1983.

**Friedman, Jerome H:** An Overview of Predictive Learning and Function Approximation, in: Cherkassky, Vladimir; Friedman, Jerome H.; Wechsler, Harry (Hrsg.): From Statistics to Neural Networks – Theory and Pattern Recognition Applications, Springer: Berlin et al. 1994, S. 1-61.

**Gebhardt, Friedrich:** Interessantheit als Kriterium für die Bewertung von Ergebnissen, in: Informatik – Forschung und Entwicklung, Jahrg. 9 (1994), H.1, S. 9-21.

**Glover, Fred:** Tabu Search – Part I, in: ORSA Journal on Computing, Vol. 1 (1989), No. 3, S. 190-206.

**Glover, Fred; Laguna, Manuel:** Tabu Search, Kluwer Academic Publishers: Dordrecht, 1997.

**Gluchowski, Peter; Gabriel, Roland; Chamoni, Peter:** Management Support Systeme - Computergestützte Informationssysteme für Führungskräfte und Entscheidungsträger, Springer: Berlin et al., 1997.

**Goldberg, David E.:** Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley: Reading, Massachusetts, 1989.

**Gordon, Geoffrey:** System Simulation, Prentice Hall, Inc.: Englewood Cliffs, New Jersey, 1969.

**Grob, Heinz Lothar; Bensberg, Frank:** Das Data-Mining-Konzept, Arbeitsbericht Nr. 8 der Reihe „Computergestütztes Controlling“, Institut für Wirtschaftsinformatik der Westfälischen Wilhelms-Universität Münster: Münster, 1999.

**Hätönen, K.; Klemettinen, M.; Mannila, H.; Ronkainen, P.; Toivonen, H.:** Knowledge Discovery from Telecommunication Network Alarm Databases, in: Su, S. (Hrsg.): Proceedings of the Twelfth International Conference on Data Engineering, IEEE Computer Society Press: Los Alamitos, Calif. 1996, S. 115-122.

**Hagedorn, Jürgen:** Die automatische Filterung von Controlling-Daten unter besonderer Berücksichtigung der Top-Down-Navigation (BETREX II), Diss., Erlangen-Nürnberg, 1996.

**Han, Jiawei; Fu, Yongjian:** Attribute-Oriented Induction in Data Mining, in: Fayyad, Usama M.; Piatetsky-Shapiro, Gregory; Smyth, Padhraic; Uthurusamy, Ramasamy (Hrsg.): Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press: Menlo Park, California et al., 1996, S. 399-421.

**Hashemi, Ray; Pearce, Bruce; Arani, Ramin; Hinson, Willam; Paule, Merle:** A Fusion of Rough Sets, Modifies Rough Sets, and Genetic Algorithms for Hybrid Diagnostic Systems, in: Lin, T.Y.; Cercone, N. (Hrsg.): Rough Sets and Data Mining: Analasys of Imprecise Data, Kluwer Academic Publishers: Boston et al., 1997, S. 149-175.

**Heistermann, Jochen:** Genetische Algorithmen – Theorie und Praxis evolutionärer Optimierung, Teubner: Stuttgart, 1994.

**Hilbert, Andreas:** Zur Theorie der Korrelationsmaße; Lohmar, Eul.: Köln, 1998.

**Hippner, Hajo; Meyer, Matthias; Wilde, Klaus D. (Hrsg.):** Computer Based Marketing – Das Handbuch zur Marketinginformatik, Vieweg: Braunschweig, 1998.

**Hippner, Hajo; Schmitz, Berit:** Data Mining in Kreditinstituten – Die Clusteranalyse zur zielgruppengerechten Kundenansprache, in: Hippner, Hajo; Küsters, Ulrich; Meyer, Matthias; Wilde, Klaus (Hrsg.): Handbuch Data Mining im Marketing – Knowledge Discovery in Marketing Databases, Vieweg: Braunschweig et al., 2001, S. 607-622.

**Hippner, Hajo; Rupp, Andreas:** Kreditwürdigkeitsprüfung im Versandhandel, in: Hippner, Hajo; Küsters, Ulrich; Meyer, Matthias; Wilde, Klaus (Hrsg.): Handbuch Data Mining im Marketing – Knowledge Discovery in Marketing Databases, Vieweg: Braunschweig et al., 2001, S. 685-706.

**Holsheimer, Marcel; Siebes, Arno:** Data Mining – The Search for Knowledge in Databases, Report CS-R9406, CWI: Amsterdam, 1994.

**Holsheimer, Marcel; Kersten, Martin; Mannila, Heikki; Toivonen, Hannu:** A Perspective on Databases and Data Mining, in: First International Conference on Knowledge Discovery and Data Mining (KDD'95), AAAI Press: Montreal, Canada, 1995, S. 150 – 155.

**Holte, Robert C.:** Very Simple Classification Rules Perform Well on Most Commonly Used Databases, in: Machine Learning, Vol. 2 (1993), S.63-91.

**Imielinski, Tomasz; Mannila, Heikki:** A Database Perspective on Knowledge Discovery, in: Communications of the ACM, Vol. 59 (1996), No. 11, S. 58-64.

**Ishibuchi, Hisao; Murata, Tadahiko:** Minimizing the Fuzzy Rule Base and Maximizing its Performance by a Multi-Objective Genetic Algorithm, in: Mantaras, R.L. (Hrsg.): Proceeding 6th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE '97), Barcelona, IEEE Press: Piscataway, NJ, 1997, S. 259-264.

**Ittner, Andreas; Sieber, Holm; Trautzsch, Sascha:** Nichtlineare Entscheidungsbäume zur Optimierung von Direktmailingaktionen, in: Hippner, Hajo; Küsters, Ulrich; Meyer, Matthias; Wilde, Klaus (Hrsg.): Handbuch Data Mining im Marketing – Knowledge Discovery in Marketing Databases, Vieweg: Braunschweig et al., 2001, S. 707-723.

**Jahnke, Bernd; Groffmann, Hans-Dieter; Kruppa, Stephan:** On-Line Analytical Processing (OLAP), in: Wirtschaftsinformatik, Jahrg. 38 (1996), H. 3, S. 321-324.

**Janßen, Volker:** Einsatz des Werbecontrolling – Aufbau, Steuerung und Simulation einer werblichen Erfolgskette, Wiesbaden: Gabler: Wiesbaden, 1999.

**Joereßen, Anton; Sebastian, Hans-Jürgen:** Problemlösung mit Modellen und Algorithmen, Teubner: Stuttgart et al., 1998.

**Kafka, Cornelia:** Konzeption und Umsetzung eines Leitfadens zum industriellen Einsatz von Data-Mining, Karlsruhe, Diss., 1999.

**Kearns, Michael J.; Vazirani, Umesh V.:** An Introduction to Computational Learning Theory, MIT Press: Cambridge Mass. et al. 1994.

**Keim, Daniel A.; Kriegel, Hans-Peter:** Visualization Techniques for Mining Large Databases: A Comparison, in: IEEE Transactions on Knowledge and Data Engineering, Vol. 8 (1996), No. 7, S. 923-938.

**Kilger, Wolfgang:** Flexible Plankostenrechnung und Deckungsbeitragsrechnung, 10. Aufl., Gabler: Wiesbaden, 1993.

**Kirkpatrick, S.; Gelatt, C.D.; Vecchi, M.P.:** Optimization by Simulated Annealing, in: Science, Vol. 220 (1983), No. 4598, S. 671-680.

**Kirsten, Mathias; Wrobel, Stefan; Horváth, Tamás:** Distance Based Approaches to Relational Learning and Clustering, in: Dzeroski, Saso; Lavrac, Nada (Hsrg.): Relational Data Mining, Springer: Berlin et al., 2001, S. 213-232.

**Knobbe, Arno J.; Siebes, Arno; van der Wallen, Daniel:** Multi-Relational Decision Tree Induction, in: Zytkow, Jan M.: Principles of data mining and knowledge discovery – third European conference, (PKDD '99), Prague, Czech Republik, Springer: Berlin et al., 1999.

**Kosala, Raymond; Blockeel, Hendrik:** Web Mining Research: A Survey, in: SIGKDD Explorations, ACM SIGKDD, Vol. 2 (2000), 1, S. 1-15.

**Krabs, Michael:** Das ROSA-Verfahren zur Modellierung dynamischer Systeme durch Regeln mit statistischer Relevanzbewertung, Forschungsberichte VDI, Reihe 8: Meß-, Steuerungs- und Regelungstechnik, Nr. 404, VDI-Verlag:Düsseldorf, 1994.

**Krahl, Daniela; Windheuser, Ulrich; Zick, Friedrich-Karl:** Data Mining: Einsatz in der Praxis, Addison-Wesley-Longman: Bonn et al. 1998.

**Kramer, Stefan; Widmer, Gerhard:** Inducing Classification and Regression Trees in First Order Logic, in: Dzeroski, Saso; Lavrac, Nada (Hsrg.): Relational Data Mining, Springer: Berlin et al., 2001, S. 140-159.

**Küppers, Bertram:** Data mining in der Praxis: ein Ansatz zur Nutzung der Potentiale von Data mining im betrieblichen Umfeld, Lang:Frankfurt a.M. et al., 1999.

**Kurbel, Karl; Szulim, Daniel; Teuteberg, Frank:** Künstliche Neuronale Netze zum Filtern und Klassifizieren betrieblicher E-Commerce-Angebote im World Wide Web – eine vergleichende Untersuchung, in: Wirtschaftsinformatik Jahrg. 42 (2000), H. 3, S. 222-232.

**Lackes, Richard; Mack, Dagmar:** Neuronale Netze in der Unternehmensplanung : Grundlagen, Entscheidungsunterstützung, Projektierung, Vahlen: München, 2000.

**Lackes, Richard; Mack, Dagmar; Tillmanns, Christoph:** Data Mining in der Marktforschung, in: Hippner, Hajo; Meyer, Matthias; Wilde, Klaus D. (Hrsg.): Computer Based Marketing – Das Handbuch zur Marketinginformatik, Vieweg: Braunschweig et al., 1998, S. 249-258.

**Lenz, Mario:** Fallbasiertes Schließen für die Produktionsplanung, Arbeitspapiere der Gesellschaft für Mathematik und Datenverarbeitung, Nr. 863, 1994.

**Lenz, Mario; Auriol, Eric; Manago, Michel:** Diagnosis and Decision Support, in: Lenz, Mario; Bartsch-Spörl, Brigitte; Burkhard, Hans-Dieter; Wess, Stefan (Hrsg.): Case-Based Reasoning Technology – From Foundations to Applications, Springer: Berlin et al., 1998.

**Ling, Charles, X.; Li, Chenghui:** Data Mining for Direct Marketin – Problems and Solutions, in: Agrawal, Rakesh; Stolorz, Paul; Piatetsky-Shapiro, Gregory: The Fourth International Conference on Knowledge Discovery & Data Mining, Proceedings, AAAI Press: Menlo Park, California, 1998, S. 73-79.

**Löbler, Helge; Petersohn, Helge:** Kundensegmentierung im Automobilhandel zur Verbesserung der Marktbearbeitung, in: Hippner, Hajo; Küsters, Ulrich; Meyer,

Matthias; Wilde, Klaus (Hrsg.): Handbuch Data Mining im Marketing – Knowledge Discovery in Marketing Databases, Vieweg: Braunschweig et al., 2001, S. 623-642.

**Mannila, Heikki:** Methods and Problems in Data Mining, in: Afrati, Foto N.: Database Theory - ICDT '97, Proceedings of the 6th International Conference, Delphi, Greece, Springer: Berlin, 1997, S. 41-55.

**Masand, Brij; Piatetsky-Shapiro, Gregory:** A Comparison of Approaches For Maximizing Business Payoff of Prediction Models, in: Simoudis, Evangelos et al. (Hrsg.): KDD-96, AAAI Press: Menlo Park, Calif., 1996, S. 195-201.

**McLeish, Mary; Yao, P.; Garg, M.; Stirtzinger, Tatiana:** Discovery of Medical Diagnostic Information: An Overview of Methods and Results, in: Piatetsky-Shapiro, Gregory; Frawley, William J. (Hrsg.): Knowledge Discovery in Databases, AAAI Press: Menlo Park, Calif., 1991, S. 477-490.

**Meffert, Heribert:** Marketing – Grundlagen der Absatzpolitik, 7. Aufl., Gabler: Wiesbaden, 1991.

**Meffert, Heribert:** Marktorientierte Unternehmensführung und Direct Marketing, in: Dallmer, Heinz (Hrsg.): Handbuch Direct Marketing, 6. Aufl., Gabler: Wiesbaden, 1991, S. 31-49.

**Meyer-Fujara, Josef; Puppe, Frank; Wachsmuth, Ipke:** Expertensysteme und Wissensmodellierung, in: Görtz, Günther (Hrsg.): Einführung in die künstliche Intelligenz, Addison-Wesley: Bonn, 1993, S. 714-766.

**Michalski, Ryszard:** Inferential Theory of Learning: Developing Foundations for Multistrategy Learning, in: Michalski, Ryszard S.; Tecuci, Gheorghe (Hrsg.): Machine Learning – A Multistrategy Approach - Volume IV, Morgan Kaufmann Publishers: San Francisco, 1994; S. 3-62.

**Michalski, Ryszard; Mozetic, Igor; Hong, Jiarong; Lavrac, Nada:** The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains, in: AAAI-86, Proceedings of the fifth national conference on artificial intelligence, Philadelphia, Kaufmann: Los Altos, Calif., 1986, S. 1041-1045.

**Michalski, Ryszard; Stepp, Robert:** Learning from Observation: Conceptual Clustering, in: Michalski, Ryszard; Carbonell, Jaime; Mitchell, Tom (Hrsg.): Machine Learning – An Artificial Intelligence Approach, Tioga Publishing Company: Palo Alto, 1983a; S. 331-363.

**Michalski, Ryszard; Stepp, Robert:** Automated Construction of Classifications: Conceptual Clustering Versus Numerical Taxonomy, in: IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 5 (1983b), No. 4, S. 396-410.

**Michels, Edmund:** Datenanalyse mit Data Mining: Kassenbons – die analysierbaren Stimmzetel der Konsumenten, in: Dynamik im Handel, (1995), H. 11, S. 37-43.

**Michels, Edmund:** Data Mining Analysen im Handel – konkrete Einsatzmöglichkeiten und Erfolgspotenziale, in: Hippner, Hajo; Küsters, Ulrich; Meyer, Matthias; Wilde, Klaus (Hrsg.): Handbuch Data Mining im Marketing – Knowledge Discovery in Marketing Databases, Vieweg: Braunschweig et al., 2001, S. 933-950.

**Morik, Katharina; Brockhausen, Peter:** A Multistrategy Approach to Relational Knowledge Discovery in Databases, in: Machine Learning, Vol. 27 (1997), No.3, S. 287-312.

**Müller, Michael; Hausdorf, Carsten; Schneeberger, Josef:** Zur Interessantheit bei der Entdeckung von Wissen in Datenbanken, in: Nakhaeizadeh, Gholamreza: Data Mining - Theoretische Aspekte und Anwendungen, Physica-Verlag: Heidelberg, 1998, S. 248-264.

**Nakhaeizadeh, Gholamreza; Taylor, C.C.; Kunisch, G.:** Dynamic Supervised Learning: Some Basic Issues and Application Aspects, in: Klar, Rüdiger; Opitz, Otto (Hrsg.): Classification and Knowledge Organization, Proceedings of the 20th Annual Conference of the Gesellschaft für Klassifikation e.V., Springer: Berlin et al., 1997.

**Nakhaeizadeh, Gholamreza:** Data Mining - Theoretische Aspekte und Anwendungen, Physica-Verlag: Heidelberg, 1998.

**Newlands, D.A.; Webb, G.I.:** Convex Hulls in Concept Induction, in: Zhong, Ning; Zhou, Lizhu (Hrsg.): Methodologie for Knowledge Discovery and Data Mining,

Third Pacific-Asia Conference, Proceedings, PAKDD-99, Springer: Berlin et al., 1999, S. 306-316.

**Padmanabhan, Balaji; Tuzhilin, Alexander:** A Belief-Driven Method for Discovering Unexpected Patterns, in: Agrawal, Rakesh; Stolorz, Paul; Piatetsky-Shapiro, Gregory: The Fourth International Conference on Knowledge Discovery & Data Mining, Proceedings, AAAI Press: Menlo Park, California, 1998, S. 94-100.

**Pawlak, Zdzislaw:** Reasoning about Data – A Rough Set Perspective, in: Polkowski, Lech; Skowron, Andrzej (Hrsg.): Rough Sets and Current Trends in Computing; First International Conference, Proceedings, RSCTC'98, Springer: Berlin et al., 1998.

**Piatetsky-Shapiro, Gregory; Brachman, Ron; Khabaza, Tom; Klösgen, Willi; Simoudis, Evangelos:** An Overview of Issues in Developing Industrial Data Mining and Knowledge Discovery Applications, in: Simoudis, Evangelos et al. (Hrsg.): KDD-96, AAAI Press: Menlo Park, Calif., 1996, S. 89-95.

**Piatetsky-Shapiro, Gregory; Masand, Brij:** Estimating Campaign Benefits and Modeling Lift, in: Chaudhuri, Surajit; Madigan, David (Hrsg.): KDD-99 - The Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego/Kalifornien/USA, ACM: New York, 1999, S. 185-193.

**Poddig, Thorsten:** Mittelfristige Zinsprognose mittels KNN und ökonometrischer Verfahren, in: Rehkugler, Heinz; Zimmermann, Hans Georg (Hrsg.): Neuronale Netze in der Ökonomie – Grundlagen und finanzwirtschaftliche Anwendungen, Vahlen: München, 1994, S. 209-290.

**Poddig, Thorsten:** Handbuch Kursprognose – Quantitative Methoden im Asset Management; Uhlenbruch: Bad Soden, 1999.

**Poddig, Thorsten; Dichtl, Hubert; Petersmeier, Kerstin:** Statistik, Ökonometrie, Optimierung – Methoden und ihre praktischen Anwendungen in Finanzanalyse und Portfoliomanagement; Uhlenbruch: Bad Soden, 2000.

**Poloni, Marco; Nelke, Martin:** Kundensegmentierung und Zielgruppendefinition im Database Marketing am Beispiel von Direktvertriebsprodukten, in: Hippner, Hajo; Küsters, Ulrich; Meyer, Matthias; Wilde, Klaus (Hrsg.): Handbuch Data Mining im

Marketing – Knowledge Discovery in Marketing Databases, Vieweg: Braunschweig et al., 2001, S. 643-650.

**Quinlan, J. Ross:** Induction of Decision Trees, in: Machine Learning, Vol. 1 (1986), S. 81-106.

**Quinlan, J. Ross:** Learning With Continuous Classes, in: Proceedings of the 5th Australien Joint Conference on Artificial Intelligence, World Scientific: Singapore, 1992, S. 343-348.

**Quinlan, J. Ross:** C4.5: Programms for Machine Learning, Morgan Kaufmann: San Mateo, Calif., 1993.

**Quinlan, J. Ross:** Relational Learning and Boosting, in: Dzeroski, Saso; Lavrac, Nada (Hsrg.): Relational Data Mining, Springer: Berlin et al., 2001, S. 292-306.

**Rauscher, Folke Axel:** Neuronale Kointegration – Ein Anwendungsbeispiel zur Wechselkursprognose, in: Nakhaeizadeh, Gholamreza: Data Mining - Theoretische Aspekte und Anwendungen, Physica-Verlag: Heidelberg, 1998, S. 328-340.

**Rechenberg, Ingo:** Evolutionsstrategie – Optimierung technischer Systeme nach Prinzipien der biologischen Evolution, Friedrich Frommann Verlag: Stuttgart, 1973.

**Reeves, Colin R.:** Modern Heuristic Techniques for Combinatorical Problems, Blackwell: Oxford, 1993.

**Rieper, Bernd:** Betriebswirtschaftliche Entscheidungsmodelle – Grundlagen, Verl. Neue Wirtschafts-Briefe: Herne, 1992.

**Rüter, Horst:** Handelskarten – der weite Weg zum Data Base Marketing, in: Dynamik im Handel, ohne Jahrg. (1994), H. 8, S. 20-22.

**Säuberlich, Frank:** KDD und Data Mining als Hiflsmittel zur Entscheidungsunterstützung, Lang: Frankfurt a.M, 2000.

**Schafft, Edgar:** Modellbildung und Modellbewertung dargestellt an einem Beispiel der Produktionsplanung bei Sortenfertigung, Hohenheim, Diss., 1992.

**Scheer, August-Wilhelm:** Wirtschaftsinformatik: Referenzmodelle für industrielle Geschäftsprozesse, Springer: Berlin et al., 1995.

**Schlageter, Gunter; Stucky, Wolffried:** Datenbanksysteme: Konzepte und Modelle, 2. Aufl., Teubner: Stuttgart, 1983.

**Schmidt von Rhein, Andreas; Rehkugler, Heinz:** KNN zur Kreditwürdigkeitsprüfung bei Privatkundenkrediten, in: Rehkugler, Heinz; Zimmermann, Hans Georg (Hrsg.): Neuronale Netze in der Ökonomie – Grundlagen und finanzwirtschaftliche Anwendungen, Vahlen: München, 1994, S. 491-545.

**Schnedlitz; Peter, Reutterer, Thomas; Joos, Walter:** Data-Mining und Sortimentsverbundanalyse im Einzelhandel, in: Hippner, Hajo; Küsters, Ulrich; Meyer, Matthias; Wilde, Klaus (Hrsg.): Handbuch Data Mining im Marketing – Knowledge Discovery in Marketing Databases, Vieweg: Braunschweig et al., 2001, S. 951-972.

**Schneeweiß, Christoph:** Elemente einer Theorie betriebswirtschaftlicher Modellbildung, in: Zeitschrift für Betriebswirtschaft ZfB, Jahrg. 54 (1984), H. 5, S. 480-504.

**Schneeweiss, H.:** Das Grundmodell der Entscheidungstheorie, in: Statistische Hefte ohne Jahrg. (1966) 7, S. 125-137.

**Schwefel, Hans-Paul:** Numerical Optimization of Computer Models, John Wiley & Sons: Chichester et al., 1981.

**Shan, Ning; Ziarko, Wojciech; Hamilton, Howard; Cercone, Nick:** Using Rough Sets as Tools for Knowledge Discovery, in: Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal, Canada, August, AAAI Press: Menlo Park, Calif., 1995, S. 263-268.

**Shen, Wei-Min; Leng, Bing:** A Metapattern-Based Automated Discovery Loop for Integrated Data Mining – Unsupervised Learning of Relational Patterns, in: IEEE Transactions on Knowledge and Data Engineering, Vol. 8 (1996), No. 6, S. 898-910.

**Sinz, Elmar J.:** Das Strukturierte Entity-Relationship-Modell (SER-Modell), in: Angewandte Informatik, Jahrg. 30 (1988), H. 5, S. 191-202.

**Spiliopoulou, Myra:** Web Usage Mining: Data Mining über die Nutzung des Web, in: Hippner, Hajo; Küsters, Ulrich; Meyer, Matthias; Wilde, Klaus (Hrsg.): Handbuch

Data Mining im Marketing – Knowledge Discovery in Marketing Databases, Vieweg: Braunschweig et al., 2001, S. 489-510.

**Spiliopoulou, Myra; Berendt, Bettina:** Kontrolle der Präsentation und Vermarktung von Gütern im WWW, in: Hippner, Hajo; Küsters, Ulrich; Meyer, Matthias; Wilde, Klaus (Hrsg.): Handbuch Data Mining im Marketing – Knowledge Discovery in Marketing Databases, Vieweg: Braunschweig et al., 2001, S. 855-873.

**Städler, Michael; Fischer, Joachim:** Warenkorb- und Bondatenanalyse im Computer Integrated Trading, in: Hippner, Hajo; Küsters, Ulrich; Meyer, Matthias; Wilde, Klaus (Hrsg.): Handbuch Data Mining im Marketing – Knowledge Discovery in Marketing Databases, Vieweg: Braunschweig et al., 2001, S. 339-348.

**Stecking, Ralf:** Marktsegmentierung mit Neuronalen Netzen, Dt. Univ.-Verl.: Wiesbaden, 2000.

**Stepp, Robert E.:** Conjunctive Conceptual Clustering: A Methodology and Experimentation, Univ. Microf. Internat: Ann Arbor, Mich.,1987; Urbana, Univ. of Illinois, Ph.D.Thesis, 1984.

**Tietz, Christiane; Poscharsky, Nikolaus; Erichson, Bernd; Müller, Holger:** Ein Vergleich führender Data-Mining-Methoden zur Cross-Selling-Optimierung von Finanzprodukten, in: Hippner, Hajo; Küsters, Ulrich; Meyer, Matthias; Wilde, Klaus (Hrsg.): Handbuch Data Mining im Marketing – Knowledge Discovery in Marketing Databases, Vieweg: Braunschweig et al., 2001, S. 767-785.

**Tietzel, Manfred:** Kriterien für die Qualität von Wirtschaftsprognosen, in: Sparkasse, Jahrg. 102 (1985), H. 3, S. 100-106.

**Tillmanns, Christoph:** Data Warehouse, in: Gabler Wirtschaftslexikon, Gabler: Wiesbaden, 2000, S. 671-673.

**Totok, Andreas:** Data Warehouse und OLAP als Basis für betriebliche Informationssysteme, Arbeitsbericht Nr. 97/03, Institut für Wirtschaftswissenschaften der Technischen Universität Braunschweig, Braunschweig, 1997.

**Tsumoto, Shusaku:** Extraction of Experts' Decision Process from Clinical Databases Using Rough Set Model, in: Komorowski, J.; Zytkow, J. (Hrsg.): Principles of

Data Mining and Knowledge Discovery, First European Symposium, PKDD'97, Proceedings, Berlin: Springer, 1997, S. 58-67.

**Vaessens, R.J.M.; Aarts, E.H.L.; Lenstra, J.K.:** A Local Search Template, in: Männer, Reinhard (Hrsg.): Parallel Problem Solving from Nature, 2, North-Holland: Amsterdam et al., 1992, S. 65-74.

**Voß, Stefan; Fiedler, Claudia; Greistorfer, Peter:** Meta-Heuristiken als moderne Lösungskonzepte für komplexe Optimierungsprobleme, in: Das Wirtschaftsstudium WISU, Jahrg. 29 (2000), H. 4, S. 552-566.

**Wagner, Udo; Reisinger, Heribert; Russ, Reinhold:** Der Einsatz von Methoden des Data Mining zur Unterstützung kommunikationspolitischer Aktivitäten der Lauda Air, in: Hippner, Hajo; Küsters, Ulrich; Meyer, Matthias; Wilde, Klaus (Hrsg.): Handbuch Data Mining im Marketing – Knowledge Discovery in Marketing Databases, Vieweg: Braunschweig et al., 2001, S. 875-888.

**Weingärtner, Stefan:** Web-Mining – Ein Erfahrungsbericht, in: Hippner, Hajo; Küsters, Ulrich; Meyer, Matthias; Wilde, Klaus (Hrsg.): Handbuch Data Mining im Marketing – Knowledge Discovery in Marketing Databases, Vieweg: Braunschweig et al., 2001, S. 889-903.

**Wiederhold, Gio C.M.; Walker, Michael G.; Blum, Robert L.; Downs, Stephen M.:** Acquisition of Knowledge from Data, in: ACM SIGART International Symposium on Methodologies for Intelligent Systems, Knoxville, Tennessee, Assoc. for Computing Machinery: New York, 1986, S. 74-84.

**Wittmann, Thomas:** Wissensentdeckung in Datenbanken mit adaptiven Regelsystemen – Entwicklung eines Methodenbaukastens auf Basis von Neuro-Fuzzy-Systemen; Lang: Frankfurt a. M., 2000.

**Woodruff, David L.:** Simulated Annealing and Tabu Search: Lessons from a Line Search, in: Computers Ops. Res.., Vol. 21 (1994), No. 8, S. 823-839.

**Wrobel, Stefan:** An Algorithm for multi-relational discovery of subgroups, in: Komorowski, J.; Zytkow, J. (Hrsg.): Principles of Data Mining and Knowledge Discovery: First European Symposium; Proceedings/PKDD'97, Springer: Berlin et al., 1997, S. 78-87.

**Yao, Y.Y.; Zhong, Ning:** An Analysis of Quantitative Measures Associated with Rules, in: Zhong, Ning; Zhou, Lizhu (Hrsg.): Methodologies for Knowledge Discovery and Data Mining, Third Pacific-Asia Conference, PAKDD-99, Proceedings, Beijing, China, Springer: Berlin et al., 1999, S. 479-488.

**Zimmermann, Hans-Jürgen:** Datenanalyse – Anwendung von DataEngine mit Fuzzy Technologien und Neuronalen Netzen, VDI Verlag: Düsseldorf, 1995.

**Zimmermann, Hans-Jürgen:** Fuzzy Technologien – Prinzipien, Werkzeuge, Potentiale, VDI Verlag: Düsseldorf, 1993.

# Anhang A: Begriffe aus der Datenbankforschung

Dieser Anhang definiert einige in Abschnitt 5.1 benötigte Begriffe aus der Datenbankforschung: [451]

**Definition 7-1: Relation**

Gegeben sei eine Attributmenge $R.A = \{R.a_1, ..., R.a_n\}$ mit den Wertebereichen (Domains) $dom(R.a_1), ..., dom(R.a_n)$. Dann ist eine Relation, $R$, als Teilmenge des kartesischen Produktes der Attributdomänen definiert:

$R \subseteq dom(R.a_1) \times ... \times dom(R.a_n)$. ◊

**Definition 7-2: Datensatz/Datenobjekt/Objekt:**

Ein $n$-Tupel, $r = (r_1, r_2, ..., r_n) \in R$ ($r_i \in dom(R.a_i)$ für $i = 1, 2, ..., n$), wird auch als *"Datensatz", "Datenobjekt" oder "Objekt" der Relation R* bezeichnet. ◊

**Definition 7-3: Primärschlüssel**

$R$ sei eine Relation und $PS$ eine Teilmenge von Attributen aus $R$: $PS \subseteq R.A$. $PS$ heißt genau dann „*Primärschlüssel* von $R$“, wenn gilt:

$\forall r_1, r_2 \in R: r_1.PS = r_2.PS \Rightarrow r_1 = r_2$. ◊

Primärschlüssel werden zur Kennzeichnung unterstrichen.

**Definition 7-4: Join/Verbund**

$R_1$ und $R_2$ seien Relationen mit den Attributmengen $R_1.A = \{R_1.a_1, ..., R_1.a_n\}$ und $R_2.A = \{R_2.a_1, ..., R_2.a_m\}$. Seien $B_1 \subseteq R_1.A$ und $B_2 \subseteq R_2.A$ Teilmengen aus diesen Attributmengen und gelte $B_1 = \{b_{1,1}, ..., b_{1,K}\}$ und $B_2 = \{b_{2,1}, ..., b_{2,K}\}$, dann ist der Join oder Verbund bezüglich $B_1$ und $B_2$ wie folgt definiert:

$R_1[B_1=B_2]R_2 := \{r_1, ..., r_n, q_1, ..., q_m \mid b_{1,k} = b_{2,k}; k = 1, ..., K; r_i \in dom(R_1.a_i); q_j \in dom(R_2.a_j); i = 1, ..., n; j = 1, ..., m\}$. ◊

---

[451] Vgl. zu den folgenden Definitionen: SCHLAGETER, GUNTER; STUCKY, WOLFFRIED (1983), S. 46, S. 81 ff. Die Definitionen wurden angepaßt, da für die Zwecke dieser Arbeit keine semantischen Integritätsbedingungen benötigt werden.

Die Begriffe "Selektion" und "Projektion" werden im Haupttext definiert (vgl. Definition 2-50 bzw. Definition 2-51).

# Anhang B: Ereignisgesteuerte Prozeßketten zur Darstelllung der Ablauflogik

Ereignisgesteuerte Prozeßketten (EPK) stellen ein Beschreibungsmittel für die Ablauflogik von Prozessen dar.[452] Sie beschreiben Funktionen, Ereignisse und deren Beziehungen durch die in Abbildung B-1 dargestellten Beschreibungselemente.

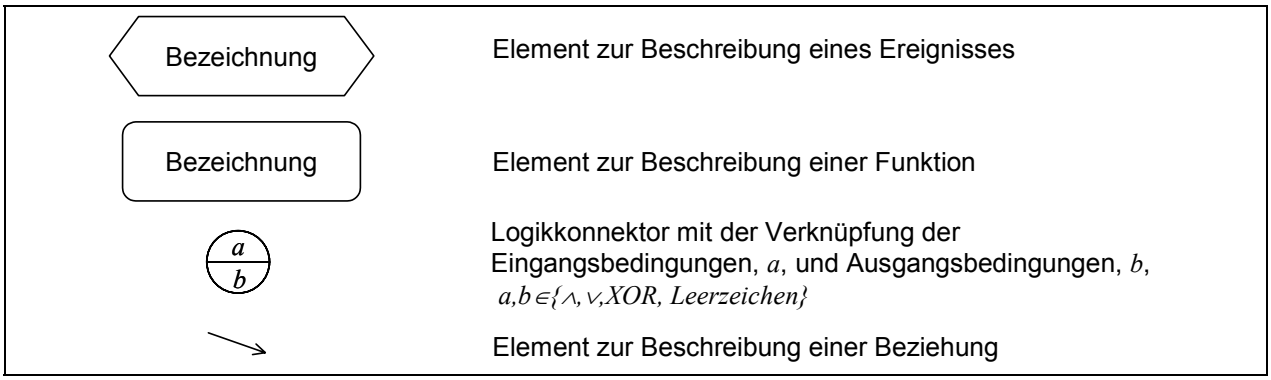



**Abbildung B-1: Beschreibungselemente einer EPK**

Bei der Modellierung sind folgende Regeln zu beachten:

- ⇨ Als Funktionen werden aktive, zeitverbrauchende Tätigkeiten modelliert.
- ⇨ Als Ereignisse werden zeitpunktbezogene Datenänderungen modelliert.
- ⇨ Eine Funktion muß in ihrem Vor- und Nachbereich – eventuell durch einen oder mehrere Logikkonnektoren getrennt – mit jeweils mindestens einem Ereignis verknüpft sein. Dadurch beginnt und endet jeder Prozeß mit mindestens einem Ereignis.
- ⇨ Logikkonnektoren verknüpfen im Ein- und Ausgangsbereich jeweils mehrere Ereignisse oder mehrere Funktionen durch $\wedge$ (und), $\vee$ (oder) oder $XOR$ (entweder oder). Einder der beiden Bereiche darf auch ein Leerzeichen enthalten, was bedeutet, daß diese Bereich mit genau einer Funktion oder einem Ereignis verbunden ist.
- ⇨ Ereignisse können keine Entscheidungen treffen, d.h. auf ein Ereignis darf kein Logikkonnektor mit einem $\vee$- oder $XOR$-Ausgang folgen.

[452] Vgl. zu den Ausführungen in diesem Anhang: SCHEER (1995), S. 49 ff.

II

# Anhang C: Strukturierte Entity-Relationship-Modelle zur Darstellung von Datenstrukturen

Strukturierte Entity-Relationship-Modelle (SERM) stellen ein Beschreibungsmittel für konzeptuelle Datenschemata dar.[453] Ein konzeptuelles Datenschema besteht aus einer Menge von Objekttypen, die durch Attribute charakterisiert und durch Beziehungen miteinander verbunden sind. SERM verwenden die in Abbildung C-1 dargestellten Beschreibungselemente.

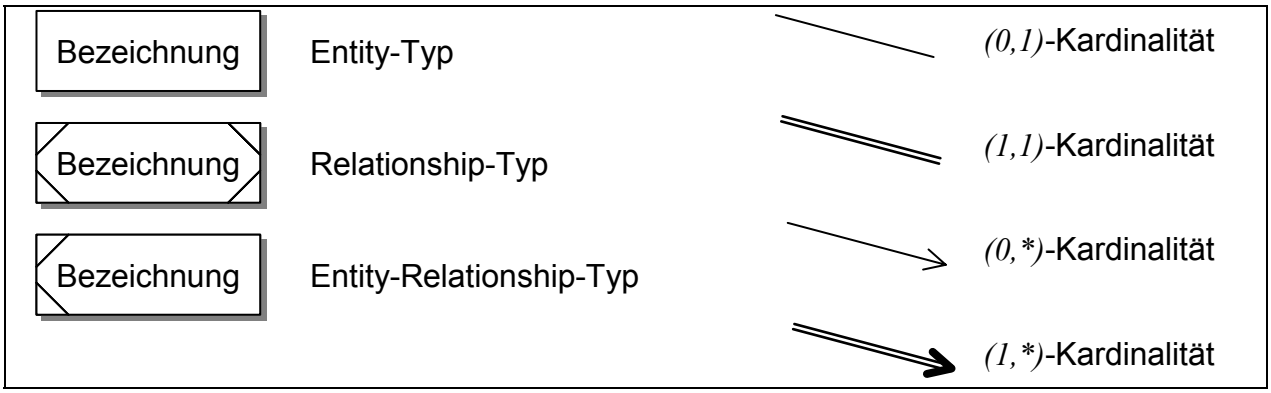



**Abbildung C-1: Beschreibungselemente im SERM**

Bei der Modellierung sind folgende Regeln zu beachten:

- ⇨ Als Entity-Typen werden Verallgemeinerungen von Objekten der Realsphäre modelliert, die durch dieselben Attribute charakterisiert werden können. Entity-Typen stehen links in einem SER-Diagramm, d.h. aus einem Entity-Typ darf keine Kante nach links austreten.
- ⇨ Als Relationship-Typen werden Verallgemeinerungen von Beziehungen zwischen den Objekten der Realsphäre modelliert. Daher ist jeder Relationship-Typ Zielknoten von mindestens zwei Kanten. Aus einem Relationship-Typ darf keine Kante nach rechts austreten.
- ⇨ Als Entity-Relationship-Typen werden Verschmelzungen je eines Entity- und eines Relationship-Typs modelliert, sofern sie in einer *(1,1)*-Beziehung zueinander stehen – d.h. sofern jedes Objekt des einen Typs mit genau einem Objekt des anderen Typs in Beziehung steht (und umgekehrt).

[453] Vgl. zu den Ausführungen in diesem Anhang: SINZ (1988), S. 196 f.

- ⇨ Alle Kanten werden von links nach rechts zwischen zwei Objekttypen gezeichnet und als Existenzabhängigkeit des rechten (Zielknoten) von dem linken Objekttypen (Startknoten) interpretiert.
- ⇨ Eine Kante repräsentiert eine *(min,max)*-Beziehung zwischen zwei Objekttypen, $A$ und $B$, wobei ein Objekt vom Typ $A$ mit mindestens *min* und maximal *max* Objekten vom Typ $B$ in Beziehung steht.
- ⇨ Zwischen zwei Objekttypen sind mehrere Kanten zulässig.
- ⇨ Man kann die Darstellung der Objekttypen durch Angabe ihrer Attribute detaillieren. Schlüsselattribute werden von links nach rechts vererbt. Dabei unterscheidet man Beziehungen zu einem Relationship- und zu einem Entity-Relationship-Typ:
  - → Der Primärschlüssel eines Relationship-Typs wird aus den Primärschlüsseln seiner Startknoten gebildet.
  - → Der Primärschlüssel eines Entity-Relationship-Typs wird:
    - entweder nur aus den Primärschlüsseln seiner Startknoten gebildet
    - oder – falls für eine Kombination der Objekte der Startknoten mehrere Beziehungen zulässig sein sollen – aus den Primärschlüsseln seiner Startknoten sowie mindestens einem weiteren Schlüsselattribut gebildet
    - oder aus einem eigenen Attribut gebildet. In diesem Fall werden die Primärschlüssel der Startknoten als Fremdschlüssel in den Entity-Relationship-Typ aufgenommen.

# Anhang D: Data Dictionary zu den Realdaten aus Abschnitt 6.2

Die in Tabelle D-1 dargestellten Metadaten wurden aus VAN DER PUTTEN ET AL. (2000) entnommen.[454] Verwendet wurden in Abschnitt 6.2 nur die kaufverhaltensbezogenen Attribute *44* bis *86*. Deren Bezeichnungen wurden hier verändert.

| Nr. | Attributname | Beschreibung |
|---|---|---|
| *44* | *c_third_party* | Contribution private third party insurance |
| *45* | *c_third_party* | Contribution third party insurance (firms) |
| *46* | *c_third_party* | Contribution third party insurane (agriculture) |
| *47* | *c_car* | Contribution car policies |
| *48* | *c_van* | Contribution delivery van policies |
| *49* | *c_motorcycle* | Contribution motorcycle/scooter policies |
| *50* | *c_lorry* | Contribution lorry policies |
| *51* | *c_trailer* | Contribution trailer policies |
| *52* | *c_tractor* | Contribution tractor policies |
| *53* | *c_aggricultural* | Contribution agricultural machines policies |
| *54* | *c_moped* | Contribution moped policies |
| *55* | *c_life* | Contribution life insurances |
| *56* | *c_private_accidents* | Contribution private accident insurance policies |
| *57* | *c_family_accidents* | Contribution family accidents insurance policies |
| *58* | *c_disability* | Contribution disability insurance policies |
| *59* | *c_fire* | Contribution fire policies |
| *60* | *c_surfboard* | Contribution surfboard policies |
| *61* | *c_boat* | Contribution boat policies |
| *62* | *c_bicycle* | Contribution bicycle policies |
| *63* | *c_property* | Contribution property insurance policies |
| *64* | *c_social_security* | Contribution social security insurance policies |

*(Fortsetzung nächste Seite)*

[454] Vgl. VAN DER PUTTEN ET AL. (2000), S. 5.

| | | |
|---|---|---|
| *65* | *third_party* | Number of private third party insurance |
| *66* | *third_party* | Number of third party insurance (firms) |
| *67* | *third_party* | Number of third party insurane (agriculture) |
| *68* | *car* | Number of car policies |
| *69* | *van* | Number of delivery van policies |
| *70* | *motorcycle* | Number of motorcycle/scooter policies |
| *71* | *lorry* | Number of lorry policies |
| *72* | *trailer* | Number of trailer policies |
| *73* | *tractor* | Number of tractor policies |
| *74* | *agricultural* | Number of agricultural machines policies |
| *75* | *moped* | Number of moped policies |
| *76* | *life* | Number of life insurances |
| *77* | *private_accident* | Number of private accident insurance policies |
| *78* | *family_accidents* | Number of family accidents insurance policies |
| *79* | *disability* | Number of disability insurance policies |
| *80* | *fire* | Number of fire policies |
| *81* | *surfboard* | Number of surfboard policies |
| *82* | *boat* | Number of boat policies |
| *83* | *bicycle* | Number of bicycle policies |
| *84* | *property* | Number of property insurance policies |
| *85* | *social* | Number of social security insurance policies |
| *86* | *caravan* | Number of mobile home policies |

**Tabelle D-1: In Abschnitt 6.2 verwendete Attribute der Caravan-Daten**