Rekonstruktionsbasierte Selektion relevanter Einflussgrößen

*D. Schauten *, H. Kiendl *, J. Meyer ** und D. H. Mache ***

* Lehrstuhl für Elektrische Steuerung und Regelung
Fakultät für Elektrotechnik und Informationstechnik
Universität Dortmund
44221 Dortmund
Tel.: 0231/755–4621 Fax: 0231 /755–2752
E–Mail: {Schauten, Kiendl}@esr.e–technik.uni–dortmund.de

** Angewandte Mathematik
Fachbereich Elektro– und Informationstechnik
Technische Fachhochschule Georg Agricola
44787 Bochum
Tel.: 0234/968–3212 Fax: 0234/968–3346
E–Mail: Jennifer.Meyer@freenet.de, Mache@thf–bochum.de

Zusammenfassung

In diesem Beitrag wird ein Verfahren vorgestellt, das im Rahmen einer Datenvoranalyse für eine nachgeschaltete datenbasierte Modellierung aus einer gegebenen Menge von potenziellen Einflussgrößen einen Satz relevanter und nichtredundanter Einflussgrößen selektiert. Hierdurch wird der Suchraum und somit auch die Komplexität für ein nachgeschaltetes Modellierungsverfahren erheblich reduziert. Im Gegensatz zu den meisten etablierten Selektionsverfahren bewertet das Verfahren nicht nur die Relevanz einzelner Einflussgrößen, sondern auch die von gesamten Sätzen von Einflussgrößen. Die Leistungsfähigkeit des Verfahrens wird an einem Demonstrationsbeispiel verdeutlicht. Des Weiteren wird exemplarisch die Auswirkung dieser Datenvoranalyse auf eine nachgeschaltete Fuzzy–Modellierung von bekannten Benchmarkbeispielen diskutiert.

1 Einführung

Zur Analyse, Weiterentwicklung und Automatisierung technischer Prozesse bilden häufig Modelle dieser Prozesse die Grundlage. Insbesondere bei komplexen Prozessen kann eine ausreichend genaue wissensbasierte oder physikalisch–mathematische Modellierung sehr zeitaufwändig oder aufgrund unzureichenden Expertenwissens gar unmöglich sein. In solchen Fällen besteht ein Lösungsansatz darin, die Modellierung des Prozesses datenbasiert auf der Grundlage von Messdaten $\boldsymbol{z}_j$, $j = 1, ..., N$ vorzunehmen. Dabei besteht ein Messpunkt $\boldsymbol{z}_j = (\boldsymbol{x}_j, y_j)$ aus einem Eingangsvektor $\boldsymbol{x}_j$, der die Werte von den beobachteten

n Eingangsgrößen (den potenziellen *Einflussgrößen*) des zu modellierenden Prozesses zusammenfasst und aus dem Wert y_j der Ausgangsgröße.

Erfahrungsgemäß sind nicht immer alle gemessenen oder zur Verfügung stehenden Einflussgrößen notwendig, um die zugrunde liegenden Prozesszusammenhänge zu modellieren. Eine vorhergehende Selektion der relevanten und nichtredundanten Einflussgrößen wirkt sich aus mehreren Gründen oftmals vorteilhaft aus:

- Der Arbeitsaufwand zum Messen und zum Speichern der Daten wird reduziert.
- Der Rechenaufwand für die Modellerstellung wird reduziert, da die Laufzeit vieler Modellierungsverfahren mit zunehmender Anzahl berücksichtigter Einflussgrößen stark ansteigt.
- Das generierte Modell ist häufig leichter zu interpretieren.
- Eine Elimination irrelevanter oder redundanter Einflussgrößen erhöht meistens die erzielbare Generalisierungsfähigkeit der generierten Modelle deutlich.

Eine Einteilung der aus der Literatur bekannten Selektionsverfahren kann zum einen anhand der verwendeten *Bewertungsfunktion* und zum anderen auf der Basis der verfolgten *Suchstrategie* vorgenommen werden. Dabei ist es die Aufgabe der Bewertungsfunktion, einzelnen Einflussgrößen oder einem ganzen Satz solcher Größen — auch *Merkmalssatz* genannt — einen Relevanzwert zuzuordnen. Die verwendete *Suchstrategie* beschreibt die Vorgehensweise, wie im Raum aller möglichen Merkmalssätze ein im Sinne der verwendeten Bewertungsfunktion möglichst günstigster Merkmalssatz gesucht wird.

1.1 Bewertungsfunktion

In der Literatur werden bezüglich der Wahl der Bewertungsfunktion zwei grundsätzlich verschiedene Ansätze [1, 2, 3] voneinander unterschieden. Bei dem sogenannten *Filteransatz* wird ein im Allgemeinen einfach und schnell zu berechnendes Maß zur Bewertung eines Merkmalssatzes oder einer einzelnen potenziellen Einflussgröße verwendet. Zur Relevanzbewertung einzelner Einflussgrößen sind beispielsweise die aus der Statistik bekannte Korrelation [4] und die aus der Informationstheorie bekannte Transinformation [5] etabliert. Eine gesamtheitliche Relevanzbewertung von Merkmalssätzen wird in [6] eingeführt. Beim sogenannten *Wrapperansatz* (engl. *wrapper* = einwickeln) wird die Relevanz eines Satzes von Einflussgrößen dadurch ermittelt, dass mit diesem Satz von Einflussgrößen anhand der verfügbaren Messdaten eine Modellierung mit Hilfe des vorgesehenen Modellierungsverfahrens vorgenommen und die resultierende Modellgüte als Maß für die Relevanz angesehen wird. Ein wesentlicher Vorteil dieses Ansatzes besteht darin, dass der *Bias* der Bewertungsfunktion mit dem des hinterher verwendeten Modellierungsverfahrens übereinstimmt. Bei einem Filteransatz kann dagegen nicht gewährleistet werden, dass eine im Sinne der Bewertungsfunktion als günstig bewertete Lösung sich auch als günstig für das nachgeschaltete Modellierungsverfahren erweist. Der wesentliche Vorteil des Filteransatzes hingegen liegt in dem meist vergleichsweise wesentlich geringeren Rechenaufwand.

In diesem Beitrag wird eine Bewertungsfunktion vorgestellt, die die Relevanz eines Merkmalssatzes in Anlehnung an [6] bewertet. Als maßgebend hierfür wird die Genauigkeit angesehen, mit der die einzelnen Ausgangsgrößenwerte y_j der Messdaten anhand von „benachbarten Messdaten“ rekonstruierbar sind (Abschnitt 2). Dabei wird der Abstand zweier Punkte $\boldsymbol{x}_j$ und $\boldsymbol{x}_i$ im Raum der berücksichtigten Einflussgrößen gemessen und über eine analytische Gewichtungsfunktion zur abstandsabhängigen Gewichtung des Einflusses der Punkte $\boldsymbol{x}_i$ bei der Rekonstruktion genutzt.

1.2 Suche von relevanten Sätzen von Einflussgrößen

Für die Suche eines Merkmalssatzes, der im Sinne der verwendeten Bewertungsfunktion als möglichst relevant bewertet wird, werden in der Literatur verschiedene Ansätze vorgeschlagen. Die vollständige Durchmusterung des Raumes aller Kombinationen von vorhandenen Einflussgrößen mittels einer *vollständige Suche* findet garantiert das globale Optimum. Aufgrund der exponentiell mit der Anzahl vorhandener Einflussgrößen anwachsenden Größe des Suchraums ist diese aber in der Praxis nur selten anwendbar. Daher werden meistens sogenannte *Greedy–Algorithmen* oder auch *Genetische Algorithmen* verwendet [7, 8].

Bekannte *Greedy–Algorithmen* sind die sogenannte *Rückwärts-* und *Vorwärtsselektion*. Bei der Vorwärtsselektion wird zunächst die Relevanz einer jeden potenziellen Einflussgröße separat ermittelt und die Einflussgröße mit der höchsten Relevanz als erste Größe des Merkmalssatzes übernommen. Darauf aufbauend wird der Merkmalssatz iterativ um diejenige noch nicht ausgewählte potenzielle Einflussgröße erweitert, deren Hinzufügung zum bereits selektierten Merkmalssatz die größte Relevanzsteigerung erbringt. Dieser sequentielle Erweiterungsprozess wird beendet, wenn keine oder keine „signifikante“ Relevanzsteigerung durch Hinzufügung einer weiteren Einflussgröße erzielt werden kann. Bei der Festlegung, was als „signifikant“ anzusehen ist, hat der Nutzer eine Wahlfreiheit. Er kann entscheiden, welchen Kompromiss er zur Berücksichtigung der gegenläufigen Ziele „hohe Relevanz“ und „möglichst wenig Einflussgrößen“ schließen will. Die *Rückwärtsselektion* hingegen startet mit der Relevanzanalyse des vollständigen Satzes aller potenziellen Einflussgrößen und verkleinert diesen iterativ immer um diejenige Einflussgröße, die bei Entfernung aus dem bestehenden Merkmalssatz die größte Relevanzsteigerung (oder auch geringste Relevanzverminderung) auslöst. Falls sich bei der Relevanzanalyse eines verkleinerten Merkmalssatzes in jedem Fall eine Verschlechterung der Relevanz ergibt, wird derjenige Merkmalssatz ausgewählt, der zur geringsten Relevanzverschlechterung führt, sofern diese als Preis für die Verkleinerung des Merkmalssatzes noch akzeptabel ist. In Abschnitt 3 wird die Wirkungsweise verschiedener Suchstrategien auf der Basis eines synthetischen Demonstrationsbeispiels illustriert.

2 Rekonstruktionsbasierte Bewertungsfunktionen

2.1 Rekonstruktionsbasierte Relevanzbewertung

Dem im Folgenden vorgestellten Verfahren zur Relevanzbewertung liegt die Idee der Rekonstruktion eines Datenpunktes auf der Basis aller anderen Datenpunkte zu Grunde [6]. Das in Abbildung 1 schematisch dargestellte Verfahren kann durch eine entsprechende Wahl der Bewertungsfunktion für eine Relevanzanalyse potenzieller Einflussgrößen sowohl von Approximationsproblemen (Abschnitt 2.1.1) als auch von Klassifikationsproblemen (Abschnitt 2.1.2) angewendet werden. Alle hierfür konzipierten Bewertungsfunktionen beruhen auf dem Prinzip „leave one out cross–validation“.

Die Grundlage für alle Bewertungsfunktionen bildet ein Rekonstruktionsalgorithmus, der für jeden Eingangsvektor $\boldsymbol{x}_j$ einen Ausgangsgrößenwert $\hat{y}_j$ auf der Basis der Ausgangsgrößenwerte y_i und den Abständen der zugehörigen Eingangsvektoren $\boldsymbol{x}_i$ aller anderer Datenpunkte $\boldsymbol{z}_i$ schätzt. Dabei werden die Abstände zwischen $\boldsymbol{x}_j$ und den $\boldsymbol{x}_i$ nur dann im vollen Eingangsraum gemessen, wenn alle Einflussgrößen berücksichtigt werden. Andernfalls werden sie in dem Unterraum gemessen, der von den berücksichtigten Einflussgrößen aufgespannt wird. Die so ermittelten Abstände werden über eine Gewichtungsfunktion zur Gewichtung des Einflusses der Punkte $\boldsymbol{z}_j$ bei der Rekonstruktion verwendet. Aus den tatsächlichen Ausgangsgrößenwerten y_j und den rekonstruierten Ausgangsgrößenwerten $\hat{y}_j$ wird der Rekonstruktionsfehler berechnet. Je kleiner dessen Wert ist, für desto relevanter wird der zugrunde liegende Satz von Einflussgrößen angesehen. Einflussgrößen, die bei Entfernung aus dem Satz den Rekonstruktionsfehler nicht oder nicht nennenswert vergrößern, werden als redundant angesehen.

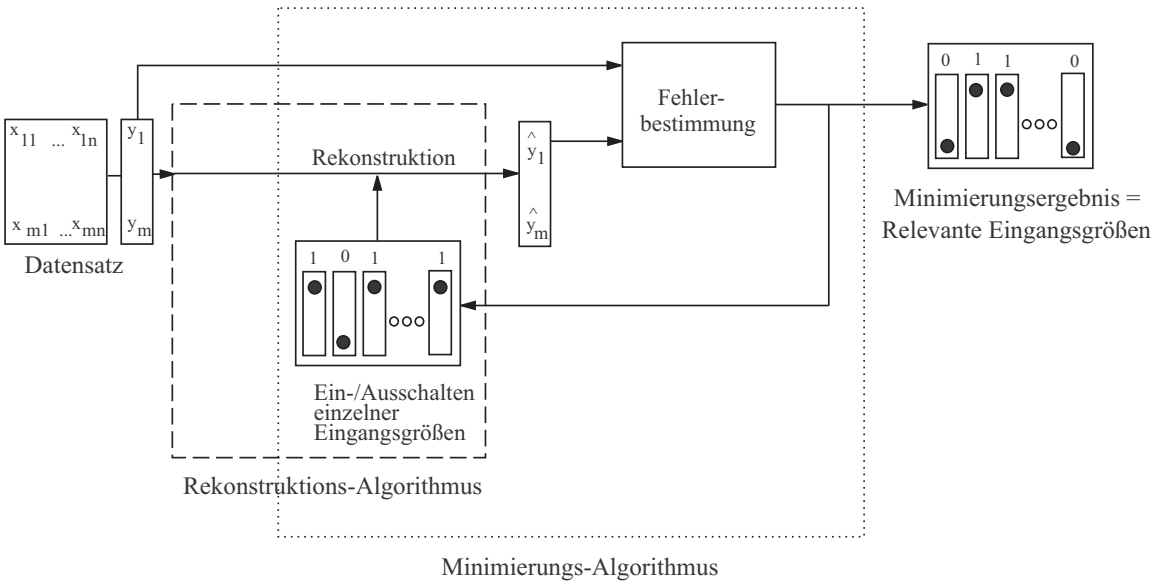

Abbildung 1: Konzept zur Selektion relevanter nichtredundanter Einflussgrößen.

Ein mit dem Rekonstruktionsalgorithmus gekoppelter Minimierungsalgorithmus zielt darauf ab, denjenigen Satz von Einflussgrößen zu bestimmen, der zu dem geringsten Rekon-

struktionsfehler führt oder einen guten Kompromiss zwischen den Zielen „möglichst geringer Rekonstruktionsfehler“ und „möglichst geringe Anzahl von Einflussgrößen“ herstellt. Im Folgenden werden verschiedene Rekonstruktionsstrategien und Bewertungsfunktionen vorgestellt.

2.1.1 Relevanzbewertung für Approximationsprobleme

Die rekonstruktionsbasierte Relevanzbewertung eines Satzes von Einflussgrößen basiert bei Approximationsproblemen auf der Analyse, wie fehlerfrei für jeden Datenpunkt $\boldsymbol{z}_j = (\boldsymbol{x}_j, y_j)$ der Ausgangsgrößenwert y_j durch die abstandsabhängig gewichteten Ausgangsgrößenwerte y_i aller anderen Datenpunkte $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i)$ (dabei $i \neq j$) rekonstruiert werden kann. Dabei wird der gewichtsbestimmende Abstand dadurch ermittelt, dass die Punkte $\boldsymbol{x}_i$ und $\boldsymbol{x}_j$ in den Raum R der berücksichtigten Einflussgrößen projiziert werden.

Hierzu ermittelt die Strategie APPROX1 für jeden Datenpunkt $\boldsymbol{z}_j$ einen Rekonstruktionswert $\hat{y}_j$ in Form einer gewichteten Mittelwertbildung:

$$\hat{y}_j = \frac{\sum\limits_{i \neq j} g_{_R}(i,j) \cdot y_i}{\sum\limits_{i \neq j} g_{_R}(i,j)} \ . \tag{1}$$

Dabei wird die jeweilige Gewichtung $g_{_R}(i,j)$ der Werte y_i in Abhängigkeit von den Abständen zwischen den Eingangsvektoren $\boldsymbol{x}_i$ und $\boldsymbol{x}_j$, gemessen im Raum R der berücksichtigten Einflussgrößen, vorgenommen.

Für die abstandsabhängige Gewichtung der Ausgangsgrößenwerte y_i wird eine Gauß'sche Gewichtungsfunktion verwendet, um Datenpunkte $\boldsymbol{x}_i$ um so stärker zu berücksichtigen, je mehr sie im betrachteten Eingangsraum R dem Punkt $\boldsymbol{x}_j$ benachbart sind:

$$g_{_R}(i,j) = e^{-p \, \cdot \, d_{_R}(i,j)} \ . \tag{2}$$

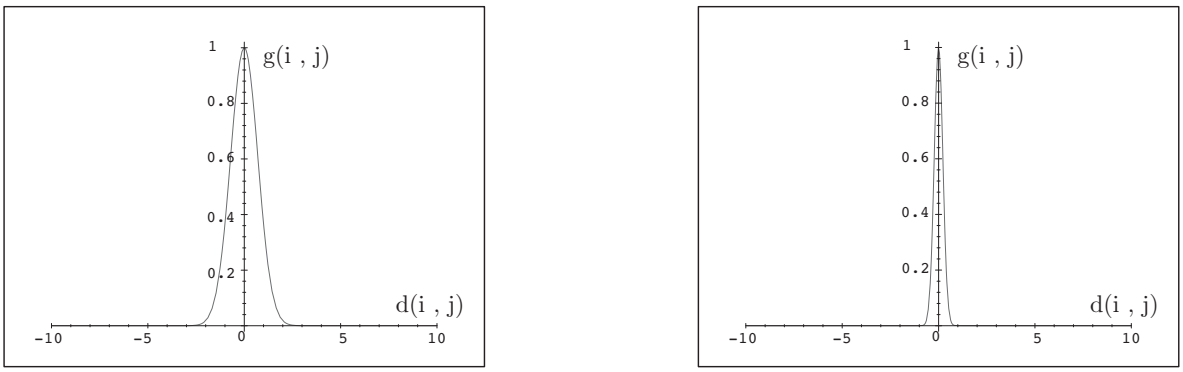

Abbildung 2: Breite der Gewichtungsfunktion $g(i,j)$ für $p = 1$ (links) und $p = 10$ (rechts).

Der Parameter p ist frei wählbar und beeinflusst die Breite der Glockenfunktion (Abbildung 2). Für $d_{_R}(i,j)$ wird das Quadrat des euklidischen Abstands der Punkte $\boldsymbol{x}_i$ und $\boldsymbol{x}_j$ verwendet:

$$d(i,j) = \|\boldsymbol{x}_j - \boldsymbol{x}_i\|^2_{2,R} \,. \tag{3}$$

Der Index R zeigt dabei an, dass dieser Abstand im Raum der berücksichtigten Einflussgrößen gemessen wird. Durch Anwendung der Rekonstruktion (1) auf alle Datenpunkte z_j erhält man einen Vektor $\widehat{\boldsymbol{y}}$, dessen Komponenten aus den rekonstruierten Werten $\hat{y}_j$ bestehen. Zur Quantifizierung der Relevanz des betrachteten Merkmalssatzes wird der mittlere quadratische Fehler gemäß

$$e = \frac{\sqrt{(\boldsymbol{y} - \widehat{\boldsymbol{y}})^T \cdot (\boldsymbol{y} - \widehat{\boldsymbol{y}})}}{N} \tag{4}$$

herangezogen, wobei der Vektor $\boldsymbol{y}$ aus den tatsächlichen Ausgangsgrößenwerten y_i des Datensatzes $\boldsymbol{z}_j$, $j = 1, 2, ..., N$ aufgebaut ist.

Alternativ zur Rekonstruktionsstrategie (1) wird die Strategie Approx2

$$\hat{y}_j = \frac{\sum\limits_{i=1}^{N} g(i,j) \cdot y_i}{\sum\limits_{i=1}^{N} g(i,j)} \tag{5}$$

verwendet. Darin geht bei der Bestimmung des Rekonstruktionswertes $\hat{y}_j$ auch der Wert y_i ein. Zur Berechnung des Rekonstruktionsfehlers e_j werden die gewichteten Abstände aller Funktionswerte vom rekonstruierten Wert zunächst summiert (Abbildung 3).

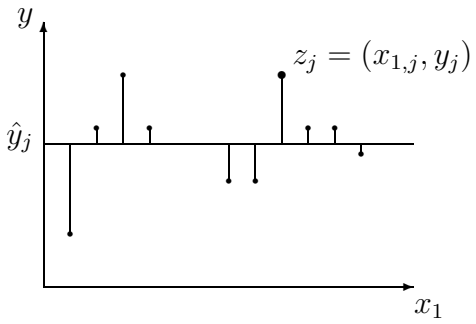

Abbildung 3: Zur Fehlerberechnung von Approx2 für den Sonderfall nur einer betrachteten Einflussgröße x_1.

Das Resultat wird durch die Summe aller Gewichtungen $g_{_R}(i,j)$, $i = 1, ..., N$ dividiert:

$$e_j = \frac{\sum\limits_{i=1}^{N} |y_i - \hat{y}_j| \cdot \, g_{_R}(i,j)}{\sum\limits_{i=1}^{N} g_{_R}(i,j)} \,. \tag{6}$$

Hierdurch wird ein gleichberechtigter Einfluss von isoliert liegenden Punkten und Punkten, die in Clustern liegen, gewährleistet. Für die Quantifizierung der Relevanz des betrachteten Merkmalssatzes wird die Summe aller Einzelfehler e_j verwendet:

$$E = \sum_{j=1}^{N} e_j \ . \tag{7}$$

2.1.2 Relevanzbewertung für Klassifikationsprobleme

Im Gegensatz zu den Approximationsproblemen ist es im Falle von Klassifikationsproblemen erforderlich, bei der Rekonstruktion eines Ausgangsgrößenwertes y_j ausschließlich eine der in den Daten auftretenden η diskreten Ausgangsgrößenwerte $y_1, y_2, \ldots, y_\eta$ als rekonstruierten Wert auszugeben.

Dazu ermittelt die Strategie KLASS einen Rekonstruktionswert auf der Basis aller anderen Datenpunkte $\boldsymbol{z}_i$ $(i \neq j)$. Hierzu werden diese Datenpunkte in Klassen $\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_\eta$ derart eingeteilt, dass alle Daten einer Klasse $\mathcal{K}_s$ den gleichen Ausgangsgrößenwert y_s aufweisen. Für jede Klasse $\mathcal{K}_s$ wird die gewichtete mittlere Häufigkeit

$$H_j(y_s) = \frac{\sum g_R(i,j)}{\mu(s)} \quad \forall i \text{ mit } \boldsymbol{z}_i \in \mathcal{K}_s \tag{8}$$

berechnet. Dabei ist $\mu(s)$ die Anzahl der Elemente in der Klasse $\mathcal{K}_s$. Der Rekonstruktionswert $\hat{y}_j$ ergibt sich schließlich durch Zuweisung desjenigen Klassenwertes y_s, für den die gewichtete mittlere Häufigkeit $H(y_s)$ maximal wird:

$$\hat{y}_j = \{y_s | H_j(y_s) = \max\{H_j(y_1), H_j(y_2), \ldots, H_j(y_\eta)\}\} \ . \tag{9}$$

Zur Quantifizierung der Relevanz des betrachteten Merkmalssatzes wird der relative Klassifikationsfehler

$$e = \frac{1}{N} \cdot \sum_{i=1}^{N} \text{sign}(|y_i - \hat{y}_j|) \tag{10}$$

verwendet, der sich bezüglich des gesamten Datensatzes $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N$ ergibt.

2.1.3 Einfluss der parametrisierbaren Gewichtungsfunktion

Es zeigt sich sowohl für die Rekonstruktion bei Approximationsproblemen nach APPROX1 (Gleichung (1)) und APPROX2 (Gleichung (5)) als auch für die Rekonstruktion bei Klassifikationsproblemen mit KLASS (Gleichung (9)), dass die Wahl des Parameters p der Gewichtungsfunktion einen wesentlichen Einfluss auf die Ergebnisse hat.

Um dies zu demonstrieren, werden für 250 äquidistant verteilte Stützstellen im Wertebereich von -10 bis 10 die Werte der Funktion $y(x) = x^2 sin(2x)cos(x)$ ermittelt.

Abbildung 4 zeigt für den so erzeugten Datensatz, dass die Güte der Rekonstruktion der Ausgangsgrößenwerte mit Hilfe von APPROX1 stark von der Wahl von p abhängig ist.

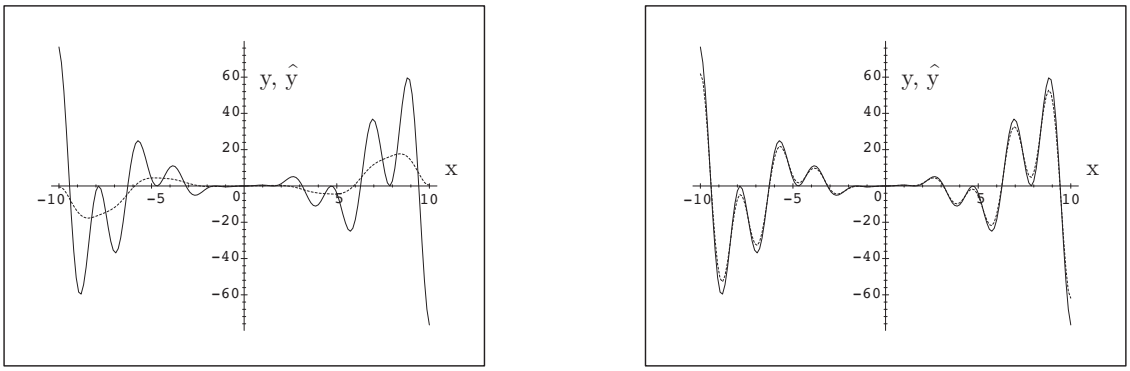

Abbildung 4: Einfluss des Parameters p auf die Güte der Rekonstruktion (gestrichelt dargestellt) von Werten der Funktion $y(x) = x^2 sin(2x)cos(x)$. Links $p = 1$, rechts $p = 10$. Die $\hat{y}$–Werte von auf der x–Achse benachbarten Punkten wurden zur grafischen Ausgabe durch Geradenstücke verbunden.

2.2 Individuelle Aktivierung von Einflussgrößen

Zur Relevanzanalyse unterschiedlicher Merkmalssätze ist es erforderlich, einzelne Einflussgrößen individuell ein- bzw. auszuschalten. Hierzu wird die Gewichtungsfunktion $g(i, j)$ geeignet parametrisiert. Zunächst wird dazu der quadrierte euklidische Abstand zweier Punkte $\boldsymbol{x}_i$ und $\boldsymbol{x}_j$ im Raum aller n potenziellen Einflussgrößen betrachtet:

$$d(i,j) = (x_{j,1} - x_{i,1})^2 + (x_{j,2} - x_{i,2})^2 + \ldots + (x_{j,n} - x_{i,n})^2 \; . \tag{11}$$

Durch Einsetzen von $d(i, j)$ für $d_{_R}(i, j)$ auf der rechten Seite von Gleichung (2) ergibt sich

$$e^{-p \cdot d(i,j)} = e^{-p \cdot (x_{j,1} - x_{i,1})^2} \cdot e^{-p \cdot (x_{j,2} - x_{i,2})^2} \cdot \ldots \cdot e^{-p \cdot (x_{j,n} - x_{i,n})^2} \; . \tag{12}$$

Ersetzt man darin die Terme $e^{-p \cdot (x_{j,k} - x_{i,k})^2}$ durch

$$e^{-p \cdot (x_{j,k} - x_{i,k})^2} = (1 - \lambda_k) + \lambda_k \cdot e^{-p \cdot (x_{j,k} - x_{i,k})^2} \; , \tag{13}$$

so kann man mit den hierdurch eingeführten Parametern λ_k jede Einflussgröße individuell durch Setzen von $\lambda_k = 1$ aktivieren oder mit $\lambda_k = 0$ deaktivieren.

Die Gewichtungsfunktion (2) erlangt hiermit die Form:

$$g_{R}(i,j) = \prod_{k=1}^{n} \left((1-\lambda_k) + \lambda_k \cdot e^{-p\cdot(x_{j,k}-x_{i,k})^2} \right) \ . \tag{14}$$

Dabei weist der Index R auf den Raum hin, der durch die aktivierten Einflussgrößen aufgespannt wird. Die Parametrisierung von $g_{R}(i,j)$ birgt im übrigen auch die Möglichkeit, die Einflussgrößen „teilweise“ zu aktivieren bzw. zu deaktivieren. In den meisten Anwendungen ist man jedoch ausschließlich an vollständig aktivierten bzw. deaktivierten Einflussgrößen interessiert.

2.3 Ein synthetisches Demonstrationsbeispiel

Die Arbeitsweise und Leistungsfähigkeit des hier vorgestellten Verfahrens mit der Rekonstruktionsvorschrift (1) und der zugehörigen Bewertungsfunktion (4) werden im Folgenden anhand der in Abbildung (5) angegebenen Testfunktion demonstriert.

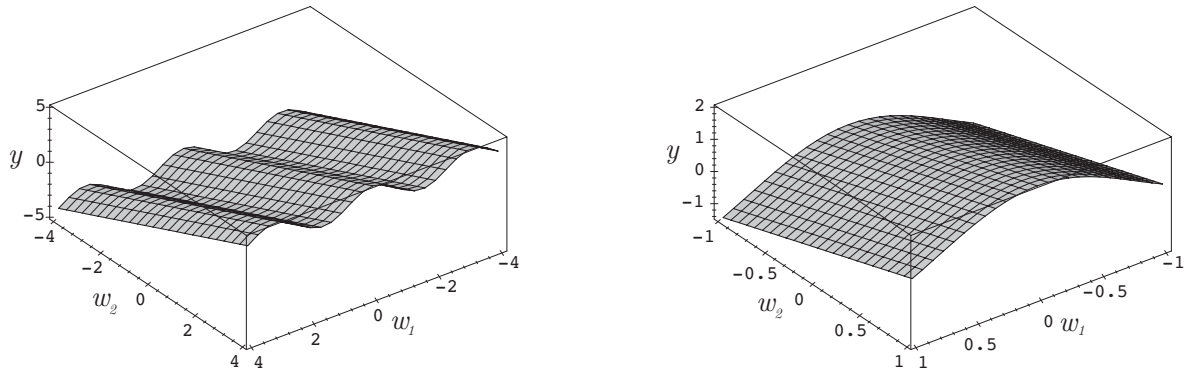

Abbildung 5: 3D–Plot der Testfunktion $y(w_1, w_2) = cos(2 \cdot w_1) + w_2$ für die Bereiche $|w_1| \leq 4\ , |w_2| \leq 4$ (links) und $|w_1| \leq 1\ , |w_2| \leq 1$ (rechts).

Mittels dieser von zwei Variablen abhängigen Testfunktion wird ein Datensatz von 441 Punkten der Form $\boldsymbol{z}_j = (\boldsymbol{x}_j, y_j)$ unter Verwendung eines vierdimensionalen Eingangsvektors $\boldsymbol{x}_j = (x_{j,1}, x_{j,2}, x_{j,3}, x_{j,4})$ geschaffen. Dabei werden die Größen $x_{j,1}$ und $x_{j,4}$ zufällig gleichverteilt im Bereich $[-1, 1]$ erzeugt, so dass diese nicht mit dem Ausgangsgrößenwert y_j korrelieren. Die Werte von $x_{j,2}$ und $x_{j,3}$ werden aus einem regelmäßigen Gitternetz des Bereichs $[-1, 1] \times [-1, 1]$ gewählt, anhand derer sich die zugehörigen Funktionswerte $y_j = y(x_{j,2}, x_{j,3})$ ergeben. Somit stehen nur die zweite und dritte Komponente des Eingangsvektors $\boldsymbol{x}_j$ in einem Zusammenhang mit y_j. Der Datensatz hat demnach die Form

$$\boldsymbol{z}_j = (rand(-1..1), x_{j,1}, x_{j,2}, rand(-1..1), y(x_{j,1}, x_{j,2}))\ .$$

Bei Anwendung des Verfahrens auf diesen Datensatz ergeben sich für die 16 möglichen Sätze der Einflussgrößen die in Tabelle 1 angegebenen Fehlerwerte.

x_1	x_2	x_2	x_4	Fehler
0	0	0	0	0.6997
1	0	0	1	0.6361
1	0	0	0	0.6355
0	0	0	1	0.6340
1	1	0	0	0.5520
1	1	0	1	0.5494
0	1	0	1	0.5488
0	1	0	0	0.5478
1	0	1	0	0.4282
0	0	1	1	0.4276
1	0	1	1	0.4270
0	0	1	0	0.4266
1	1	1	1	0.2004
1	1	1	0	0.1928
0	1	1	1	0.1902
0	**1**	**1**	**0**	**0.1875**

Tabelle 1: Fehlerwerte für alle Kombinationen von Einflussgrößen.

Daraus wird die hohe Relevanz der Einflussgrößen x_2 und x_3 ersichtlich. Ausschließlich für diejenigen Sätze von Einflussgrößen, die diese beiden Größen enthalten, ergeben sich die mit Abstand kleinsten Fehlerwerte. Zudem führt die Berücksichtigung der irrelevanten, zufällig erzeugten Einflussgrößen x_1 und x_4 zu einer geringfügigen, aber dennoch signifikanten Verschlechterung des kleinsten Fehlerwertes, der sich durch ausschließliche Berücksichtigung der Größen x_2 und x_3 ergibt. Die Fehlerwerte zeigen somit, dass die Größen x_1 und x_4 nicht relevant sind, während die dem funktionalen Zusammenhang tatsächlich zugrunde liegenden Einflussgrößen als relevant erkannt werden.

3 Selektion relevanter Einflussgrößen

Um mit Hilfe der angegebenen Bewertungsfunktionen letztendlich eine sinnvolle Teilmenge relevanter nichtredundanter Einflussgrößen auswählen zu können, benötigt man noch eine geeignete Suchstrategie. Im Folgenden wird hierzu ein kurzer Überblick über unterschiedliche Suchstrategien gegeben und es wird ihre Leistungsfähigkeit anhand des Testbeispiels aus Abschnitt 2.3 illustriert.

3.1 Vollständige Suche

Bei der *Vollständigen Suche* werden sukzessive für alle möglichen Kombinationen von Einflussgrößen durch Auswertung der Bewertungsfunktion die zugehörigen Fehlerwerte ermittelt. Derjenige Satz von Einflussgrößen, der zum minimalen Fehlerwert führt, ist eine sinnvolle Wahl für den gesuchten Merkmalssatz.

Die Anwendung einer *Vollständigen Suche* auf das Beispiel aus Abschnitt 2.3 liefert den minimalen Fehlerwert der Bewertungsfunktion APPROX1 für den Merkmalssatz

$$\mathbf{0\ 1\ 1\ 0} \quad \text{, Fehler: } 0.1875\ ,$$

was somit genau den tatsächlich bestehenden funktionalen Zusammenhang widerspiegelt. Bei dieser Suchstrategie wächst die erforderliche Laufzeit gemäß $\mathcal{T}(n) = 2^n = \mathcal{O}(2^n)$ exponentiell mit der Dimension n des Einflussgrößenraumes. Unter Berücksichtigung der Anzahl N vorhandener Datenpunkte zur Auswertung der Bewertungsfunktion APPROX1 ergibt sich damit eine meist nicht akzeptable Laufzeit von $\mathcal{O}(N^2 2^n)$. Bei sehr vielen Einflussgrößen werden daher die nachfolgend beschriebenen Suchstrategien eingesetzt.

3.2 Zerstörende lokale Verbesserungsstrategie

Dieser Algorithmus startet mit der Analyse des vollständigen Merkmalsraumes durch Aktivierung aller Einflussgrößen. Anschließend wird nacheinander *eine einzige* Einflussgröße deaktiviert. Erhöht sich dabei der Fehler signifikant (diese Entscheidung kann durch Einfügung eines Toleranzparameters geschaffen werden), so wird diese Einflussgröße als relevant erklärt und ansonsten als irrelevant verworfen. Die Laufzeit dieser Suchstrategie beträgt $\mathcal{T}(n) = n + 1 = \mathcal{O}(n)$ und somit in Verbindung mit der Bewertungsfunktion $\mathcal{O}(N^2 n^2)$.

Bei Anwendung dieser Strategie auf das obige Beispiel ergeben sich der Reihe nach folgende Merkmalssätze und zugehörige Fehlerwerte:

1	1	1	1	0.2004
0	1	1	1	0.1902
1	0	1	1	0.4279
1	1	0	1	0.5494
1	1	1	0	0.1928
0	**1**	**1**	**0**	**0.1875**

Ersichtlich wird dasselbe optimale Ergebnis wie bei der *Vollständigen Suche* gefunden.

3.3 Bewahrende lokale Verbesserungsstrategie

Auch diese Strategie beginnt mit allen aktivierten Einflussgrößen und deaktiviert systematisch einzelne Einflussgrößen. Im Falle einer signifikanten Verbesserung im Sinne der Bewertungsfunktion wird diese Veränderung des Merkmalsraumes jedoch sofort beibehalten. Im anderen Fall wird die Veränderung sofort wieder rückgängig gemacht. Dies wird sukzessive für jede Einflussgröße durchgeführt. Genau wie bei der *Zerstörenden lokalen* Strategie beträgt die Laufzeit auch hier insgesamt $\mathcal{O}(N^2n^2)$. Im Demonstrationsbeispiel findet diese Strategie ebenfalls den Merkmalssatz mit dem minimalen Fehler:

1	1	1	1	0.2004
0	1	1	1	0.1902
0	0	1	1	0.4276
0	1	0	1	0.5488
0	**1**	**1**	**0**	**0.1875**

3.4 Vollständige Greedy–Suche

Bei dieser Suchstrategie wird zunächst jede Einflussgröße einzeln bewertet und diejenige Größe, die den im Sinne der Bewertungsfunktion niedrigsten Fehlerwert erbringt, ausgewählt. Mit der so gefundenen Einflussgröße werden alle verbleibenden Einflussgrößen einzeln kombiniert und anschließend diejenige Einflussgröße, die zur größten Reduktion des Fehlers führt, als weitere relevante Einflussgröße ausgewählt. Dieser Vorgang wird so lange fortgesetzt, bis sich durch Hinzufügen einer weiteren Einflussgröße keine weitere signifikante Minimierung des Fehlers erzielen lässt.

1	0	0	0	0.6355
0	1	0	0	0.5478
0	0	1	0	0.4266
0	0	0	1	0.6340
1	0	1	0	0.4282
0	1	1	0	0.1875
0	0	1	1	0.4276
1	1	1	0	0.1928
0	1	1	1	0.1902
0	**1**	**1**	**0**	**0.1875**

Aus der Ergebnistabelle wird ersichtlich, dass auch dieser Suchalgorithmus den Merkmalssatz, der zum günstigsten Fehlerwert führt, findet.

Die in diesem Beispiel fast vollständig ausgeschöpfte worst–case Laufzeit beträgt beim *Greedy–Algorithmus* $\mathcal{T}(n) = \frac{1}{2} \cdot (n^2 + n) = \mathcal{O}(n^2)$. In Verbindung mit der Bewertungsfunktion ergibt sich demnach insgesamt eine Laufzeit von $\mathcal{O}(N^2n^3)$.

3.5 Genetischer Algorithmus

Die Grundlage zu Genetischen Algorithmen wurden von J. H. Holland in den 60er Jahren gelegt [9, 10]. Dieses nichtdeterministische populationsbasierte Verfahren ist besonders zur Optimierung hochdimensionaler Kombinationsprobleme geeignet. Die zu optimierenden Parametersätze repräsentieren dabei sogenannte *Individuen*. Für die hier vorgestellte Suche von relevanten Merkmalssätzen entspricht ein Individuum dem Vektor $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)$.

Ausgehend von einer zufällig initialisierten Startpopulation von Individuen — den *Eltern* — werden über die Verfahrensschritte Kreuzung und Mutation *Nachkommen* der Population hinzugefügt, die aufgrund des veränderten genetischen Codes andere Merkmalssätze repräsentieren. Im Verfahrensschritt Selektion wird eine Teilmenge der Population als Eltern der nächsten Generation ausgewählt, die mit großer Wahrscheinlichkeit für den weiteren Optimierungsprozess von Nutzen sind. Die Grundlage dieser Auswahl bildet eine Bewertung der Individuen anhand der Bewertungsfunktion, die eine Maß für die Relevanz des Merkmalssatzes darstellt.

Aufgrund der niedrigen Dimensionalität des Demonstrationsbeispiels wird hier auf die Anwendung und Auswertung einer genetischen Suche verzichtet.

4 Anwendung auf Bechmarkprobleme

Im Folgenden wird ein Überblick über die Ergebnisse gegeben, die durch Anwendung des vorgestellten Verfahrens für einige aus der Literatur bekannte und mit dem Fuzzy–ROSA–Verfahren [11, 12, 13, 14, 15] ausgiebig behandelte Klassifikations- und Approximationsprobleme erzielt worden sind. Bei den niederdimensionalen Bechmarkproblemen werden zudem die mit *Vollständiger Suche* und *Greedy–Suche* erzielten Ergebnisse miteinander verglichen. Zusätzlich werden die erzielten Ergebnisse, die sich bei einer erneuten Modellierung mit dem Fuzzy–ROSA–Verfahren auf der Basis des jeweils selektierten relevanten Merkmalssatzes einstellen, mit den bisher besten Modellierungsergebnissen verglichen. Dabei werden jeweils diejenigen Einstellungen des Fuzzy–ROSA–Verfahrens, mit denen das bisher beste Ergebnis erzielt wurde, beibehalten. Auf eine Adaption von Strategieparametern des Fuzzy–ROSA–Verfahrens zur weiteren Verbesserung der Modellierungsgüte wurde verzichtet, um die Auswirkung der Merkmalsselektion als Instrument einer Datenvoranalyse getrennt von der eigentlichen Modellierung erkennen zu können.

4.1 Der IRIS–Datensatz

Die Aufgabe des Klassifikationsproblems IRIS [16, 17] besteht darin, auf der Grundlage von vier Einflussgrößen (*Kelchblattlänge*, *Kelchblattbreite*, *Blütenblattlänge* und *Blütenblattbreite*) die zugehörige Schwertlilienart (*Setosa*, *Versicolor* oder *Virginica*) zu bestimmen.

Die in Tabelle 2 dokumentierten Ergebnisse zeigen für beide angewandten Suchverfahren ein identisches Ergebnis. Dabei werden jeweils die beiden Größen *Blütenblattlänge* (x_3)

x_1	x_2	x_3	x_4	Fehler
0	0	0	0	0.4987
0	0	0	1	0.0467
0	0	1	0	0.0533
0	1	0	0	0.4533
1	0	0	0	0.2933
0	**0**	**1**	**1**	**0.0400**
0	1	0	1	0.0667
0	1	1	0	0.0667
1	0	0	1	0.1400
1	0	1	0	0.1067
1	1	0	0	0.2200
0	1	1	1	0.0467
1	0	1	1	0.0533
1	1	0	1	0.1200
1	1	1	0	0.1067
1	1	1	1	0.0667

x_1	x_2	x_3	x_4	Fehler
1	0	0	0	0.2933
0	1	0	0	0.4533
0	0	1	0	0.0533
0	0	0	1	0.0467
1	0	0	1	0.1400
0	1	0	1	0.0667
0	**0**	**1**	**1**	**0.0400**
1	0	1	1	0.0533
0	1	1	1	0.0467

Tabelle 2: Getestete Einflussgrößenkombinationen mit zugehörigen Fehlerwerten für IRIS bei *Vollständiger Suche* (links) und *Greedy–Suche* (rechts).

und *Blütenblattbreite* (x_4) als relevante Einflussgrößen klassifiziert. Bei Wahl dieser beiden Einflussgrößen ergibt sich ein Klassifikationsfehler von 4%. Während die *Vollständige Suche* alle 16 Einflussgrößenkombinationen testet, ermittelt die *Greedy–Suche* schon nach 9 getesteten Einflussgrößenkombinationen das identische Suchergebnis.

Eine Modellierung mit dem Fuzzy–ROSA–Verfahren auf der Basis der beiden ermittelten Einflussgrößen x_3 und x_4 ergibt im Mittel einer 2–fachen Kreuzvalidierung einen Regelsatz von $\hat{R} = 6$ Regeln, der einen relativen Klassifikationsfehler von $\hat{\epsilon}_{lern} = 2.6\%$ auf Lerndaten und von $\hat{\epsilon}_{vali} = 4.1\%$ bezüglich der Validierungsdaten aufweist.

Damit kann das in [15] unter Verwendung aller Einflussgrößen beste erzielte Ergebnis — ein Modell mit $R = 9$ Regeln, $\epsilon_{lern} = 4.1\%$ und $\epsilon_{vali} = 4.1\%$ — sogar noch leicht verbessert werden.

4.2 Der WINE–Datensatz

Der WINE–Datensatz [18] betrifft die chemische Analyse von Weinen, die aus der gleichen Region Italiens stammen, aber von drei unterschiedlichen Weinbauern angebaut wurden. Das Ergebnis der chemischen Analyse der Weine ist die jeweils gefundene Menge von den 13 betrachteten Inhaltsstoffen *Alcohol*, *Malic Acid*, *Ash*, *Alcalinity of Ash*, *Magnesium*, *Total Phenols*, *Flavanoids*, *Nonflavanoid Phenols*, *Proanthocyanins*, *Color Intensity*, *Hue*, *OD280/OD315 of Dilluted Wines* und *Proline*, anhand derer die Herkunft der Weine klassifiziert werden soll.

Die in Tabelle 3 dargelegten Ergebnisse für eine *Greedy–Suche* zeigen, dass bereits in der dritten Stufe die Einflussgrößen *Alcohol* (x_1), *Flavanoids* (x_7) und *OD280/OD315 of*

Stufe	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	Fehler
1	0	0	0	0	0	0	1	0	0	0	0	0	0	0.2135
2	1	0	0	0	0	0	1	0	0	0	0	0	0	0.1124
3	1	0	0	0	0	0	1	0	0	0	0	1	0	0.0955
4	1	0	1	0	0	0	1	0	0	0	0	1	0	0.0955
4	1	0	0	0	0	0	1	1	0	0	0	1	0	0.0955
4	1	0	0	0	0	0	1	0	0	0	1	1	0	0.0955
3	**1**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**	**0**	**0.0955**

Tabelle 3: Beste Einflussgrößenkombinationen mit zugehörigen Fehlerwerten einer jeden durchlaufenen Analysestufe der *Greedy–Suche* bei WINE.

Dilluted Wines (x_{12}) als relevant eingestuft werden. Der zugehörige Fehler liegt bei 0.0955. In der anschließenden vierten Stufe wird drei mal eine weitere Einflussgröße gefunden, die zu einem identischen Fehler führt, so dass die in Stufe 3 gefundene Lösung als Ergebnis ausgegeben wird.

Die Durchführung einer *Vollständigen Suche* liefert in diesem Anwendungsbeispiel mit einem Fehler von 0.0843 ein anderes Suchergebnis. Dabei wird zusätzlich zu den bei der Greedy–Suche gefundenen Größen die Einflussgrößen *Ash* (x_3) und *Hue* (x_{11}) für relevant befunden. Dieses globale Optimum bedeutet im Hinblick auf die gefundene Lösung durch die *Greedy–Suche* nur eine geringfügige Verbesserung des Fehlerwertes, die zudem mit einem sehr hohen Aufwand von 8192 getesteten Einflussgrößenkombinationen — im Gegensatz zu 46 getesteten Einflussgrößenkombinationen der *Greedy–Suche* — ermittelbar ist.

In Anbetracht dieses großen zusätzlichen Zeitaufwands und der relativ geringen Verbesserung des Fehlerwertes liefert somit auch hier die *Greedy–Suche* mit einem Klassifikationsfehler von 9.5% ein sehr gutes Ergebnis. Auf der Basis dieser drei Merkmale wurde bei einer Modellierung mit dem Fuzzy–ROSA–Verfahren im Mittel einer 2–fachen Kreuzvalidierung ein Regelsatz von $\hat{R} = 15$ Regeln generiert, der einen relativen Klassifikationsfehler von $\hat{\epsilon}_{lern} = 6.7\%$ auf Lerndaten und von $\hat{\epsilon}_{vali} = 9.6\%$ bezüglich der Validierungsdaten aufweist.

Damit wird das beste in [15] auf der Basis des vollständigen Merkmalsraumes erzielte Modell, bestehend aus $R = 141$ Regeln mit $\epsilon_{lern} = 3.2\%$ und $\epsilon_{vali} = 6.2\%$, hier nicht erreicht. Allerdings ist das hier gefundene Modell aufgrund der deutlich geringeren Regelanzahl viel besser interpretierbar.

4.3 Der GENE–Datensatz

Bei diesem Benchmarkproblem geht es um die Klassifikation von Intron–Exon–Verbindungen in Nukleotidsequenzen [19]. Ein DNA–Sequenz–Fenster von 60 DNA-Nukleotidsequenzelementen liefert die Einflussgrößen $x_1, x_2, \ldots, x_{60}$, anhand derer entschieden werden soll, ob sich in der Mitte des Sequenz–Fensters ein Intron–Exon–Übergang (Donator), ein Exon–Intron–Übergang (Akzeptor) oder keines von beiden befindet. Die Durchführung

einer Einflussgrößenselektion mit einer *Greedy–Suche* liefert in diesem Anwendungsbeispiel einen Merkmalssatz, bestehend aus den Einflussgrößen $\{x_{28}, x_{29}, x_{30}, x_{31}, x_{32}, x_{35}\}$ mit einem Fehlerwert von 0.1465.

Bemerkenswert dabei ist, dass sich die ermittelten sechs Einflussgrößen ausschließlich auf Elemente um die Mitte des betrachteten Fensters der Nukleotidsequenz beziehen. Dieses Phänomen ist bereits in ähnlicher Form bei der Fuzzy–Modellierung mit dem Fuzzy–ROSA–Verfahren in [14] festgestellt worden. Eine Analyse der generierten Regelbasis hatte aufgedeckt, dass die mittleren Einflussgrößen $\{x_{28}, \ldots, x_{35}\}$ am häufigsten in den Regelprämissen vertreten waren.

Eine Modellierung mit dem Fuzzy–ROSA–Verfahren unter Berücksichtigung des ermittelten relevanten Merkmalsraumes ergibt im Mittel einer 2–fachen Kreuz–Validierung einen Regelsatz von $\hat{R} = 132$ Regeln, der einen relativen Klassifikationsfehler von $\hat{\epsilon}_{lern} = 5.0\%$ auf Lerndaten und von $\hat{\epsilon}_{vali} = 7.1\%$ bezüglich der Validierungsdaten aufweist. Damit wird zwar das bisher beste erzielte Ergebnis auf der Basis des vollständigen Merkmalsraumes von $R = 221$ Regeln mit $\epsilon_{lern} = 5.1\%$ und $\epsilon_{vali} = 5.8\%$ nicht ganz erreicht. Bezüglich eines von [14] vorgenommenen Modellgütenvergleichs mit 33 anderen aus der Literatur bekannten Lernverfahren bedeutet diese etwas geringere Modellgenauigkeit, dass das hier erzielte Ergebnis im Ranking statt auf den dritten auf den siebten Platz einzuordnen ist.

Demgegenüber wird aber wegen der deutlich geringeren Regelanzahl die Interpretierbarkeit des generierten Modells deutlich gesteigert. Zusätzlich muss angemerkt werden, dass der Referenzregelsatz aus [14] aus einem aufwändigen und nichtdeterministischen Regelgenerierungsprozess resultiert, der kaum ohne Expertenwissen vorgenommen werden kann. Eine von [15] konzipierte systematische Vorgehensweise für das Fuzzy–ROSA–Verfahren ergab unter Verwendung aller Einflussgrößen für dieses Benchmarkproblem bei näherungsweise identischen Modellgüten einen Regelsatz von $R = 1567$ Regeln.

4.4 Die MACKEY–Glass–Zeitfolge

Die chaotische Mackey–Glass–Zeitfolge [20] wird in der Literatur häufig als Benchmarkproblem für Approximationsverfahren genutzt. Die diskrete Version der Mackey–Glass–Zeitfolge lässt sich beschreiben durch

$$x(t+1) = (1-a)x(t) + \frac{bx(t-\tau)}{1+x^{10}(t-\tau)} \; .$$

Als Parameter werden $a = 0.1$, $b = 0.2$ und $\tau = 17$ gewählt, mit $x(t) = 0$ für $t < 0$ und $x(0) = 1.2$. Die Aufgabe besteht darin, auf Basis der Werte $x(t-18)$, $x(t-12)$, $x(t-6)$ und $x(t)$ den Wert $x(t+6)$ vorauszusagen. Für das hier vorgestellte Verfahren wurden 1000 Datenpunkte zufällig erzeugt.

Die ermittelten Ergebnisse für die Relevanzanalyse mit *Vollständiger Suche* und *Greedy–Suche* sind in Tabelle 4 zusammengefasst. Ähnlich wie beim Klassifikationsproblem IRIS finden beide Suchstrategien für dieses Approximationsproblem den gleichen Satz von Einflussgrößen.

x_1	x_2	x_3	x_4	Fehler
0	0	0	0	0.2151
0	0	0	1	0.1955
0	0	1	0	0.1653
0	1	0	0	0.1230
1	0	0	0	0.1580
0	0	1	1	0.1541
0	1	0	1	0.1164
0	1	1	0	0.1083
1	0	0	1	0.1406
1	0	1	0	0.1043
1	1	0	0	0.1123
0	1	1	1	0.0992
1	0	1	1	0.1029
1	1	0	1	0.0961
1	1	1	0	0.0757
1	**1**	**1**	**1**	**0.0730**

x_1	x_2	x_3	x_4	Fehler
1	0	0	0	0.1580
0	1	0	0	0.1230
0	0	1	0	0.1653
0	0	0	1	0.1955
1	1	0	0	0.1123
0	1	1	0	0.1083
0	1	0	1	0.1164
1	1	1	0	0.0757
0	1	1	1	0.0992
1	**1**	**1**	**1**	**0.0730**

Tabelle 4: Getestete Einflussgrößenkombinationen mit zugehörigen Fehlerwerten für MACKEY bei *Vollständiger Suche* (links) und *Greedy–Suche* (rechts).

Während auch in diesem Beispiel die *Vollständige Suche* alle 16 Einflussgrößenkombinationen testen muss, ermittelt die *Greedy–Suche* in diesem worst–case Szenario schon nach 10 getesteten Einflussgrößenkombinationen alle vorhandenen Einflussgrößen als Ergebnis. Da das Verfahren in diesem Fall keine Einflussgröße ausscheidet, erübrigt sich hier eine nachgeschaltete Fuzzy–Modellierung mit dem Fuzzy–ROSA–Verfahren.

4.5 Das BOSTON Housing Problem

Das Boston Housing Problem [21] hat zwei Zielsetzungen: Zum einen soll ein Modell für den Wohnungspreis in der Gegend von Boston anhand einiger ausgewählter Merkmale erstellt, zum anderen soll ein Modell für die Schadstoffbelastung anhand der gleichen Merkmale bestimmt werden. Hier wird nur die erste Zielsetzung behandelt. Das betrachtete Approximationsproblem besitzt die in Tabelle 5 aufgelisteten Einflussgrößen, die alle für den Menschen direkt verständliche Inhalte beschreiben. Ursprünglich sind 506 Datensätze für die Modellierung erhoben worden. 16 dieser Datensätze scheinen aber zensiert zu sein[1] und wurden daher für die Relevanzanalyse nicht betrachtet.

Auch bei diesem komplexeren Approximationsproblem liefern die *Greedy–Suche* und die *Vollständige Suche* das gleiche Ergebnis, wobei die *Greedy–Suche* in diesem Fall nur 91 von 8192 möglichen Einschaltkombinationen testet und somit wiederum eine erhebliche Einsparung von Rechenzeit erbringt.

Des Weiteren ist aus den in Tabelle 6 zusammengefassten Ergebnissen erkennbar, dass sich ab acht ausgewählten Eingangsgrößen der Fehlerwert nur geringfügig ändert. Im Prinzip

[1] Vergleich www.cs.toronto.edu/~delve/data/boston/bostonDetail.html

x_1	Kriminalitätsrate
x_2	Anteil von Grundstücken > 25000 sq.ft.
x_3	Anteil Industrie
x_4	Flussnähe (Charles River)
x_5	Schadstoffbelastung
x_6	Durchschnittliche Anzahl Räume pro Haus
x_7	Anteil Gebäude erbaut vor 1940
x_8	Gewichteter Abstand zu fünf Arbeitszentren
x_9	Index für die Erreichbarkeit der Highways
x_{10}	Steuerrate
x_{11}	Verhältnis Anzahl Lehrer zu Anzahl Schüler
x_{12}	Anteil schwarzer Bevölkerung
x_{13}	Armutsanteil

Tabelle 5: Betrachtete Einflussgrößen für das BOSTON Housing Problem.

hätte man an dieser Stelle die Suche abbrechen können. Bei dieser Realisierung war jedoch keine Mindestverbesserung im Sinne einer Toleranzschwelle vorgegeben, so dass jede noch so kleine Verbesserung berücksichtigt wurde.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	Fehler
1	0	0	0	0	0	0	0	0	0	1	0	0	0	4.9399
2	0	1	0	0	0	0	0	0	0	1	0	0	0	4.8764
3	0	1	0	0	0	0	1	0	0	1	0	0	0	4.8468
4	0	1	0	0	0	0	1	0	0	1	0	1	0	4.8317
5	0	1	0	0	0	0	1	0	0	1	0	1	1	4.8272
6	0	1	1	0	0	0	1	0	0	1	0	1	1	4.8259
7	1	1	1	0	0	0	1	0	0	1	0	1	1	4.8254
8	1	1	1	0	0	0	1	0	0	1	1	1	1	4.8250
9	1	1	1	0	0	1	1	0	0	1	1	1	1	4.8249
9	1	1	1	0	0	0	1	1	0	1	1	1	1	4.8249
10	1	1	1	0	0	1	1	1	0	1	1	1	1	4.8248
11	1	1	1	1	0	1	1	1	0	1	1	1	1	4.8248
11	1	1	1	0	1	1	1	1	0	1	1	1	1	4.8248
12	1	1	1	1	1	1	1	1	0	1	1	1	1	4.8248
13	1	1	1	1	1	1	1	1	1	1	1	1	1	4.8254
12	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**1**	**4.8248**

Tabelle 6: Verlauf der *Greedy–Suche* für BOSTON mit zugehörigen Fehlerwerten für jede durchlaufene Analysestufe.

Für eine erneute Modellierung mit dem Fuzzy–ROSA–Verfahren sind demzufolge die acht Einflussgrößen $\{x_1, x_2, x_3, x_7, x_{11}, x_{12}, x_{13}\}$ verwendet worden. Dabei hat sich ein Regelsatz von $\hat{R} = 155$ Regeln ergeben, der einen relativen Klassifikationsfehler von $\hat{\epsilon}_{lern} = 2.2$ auf Lerndaten und von $\hat{\epsilon}_{vali} = 2.8$ auf Validierungsdaten aufweist. Damit wird zwar das bisher beste erzielte Ergebnis auf der Basis des vollständigen Merkmalsraumes von $\epsilon_{lern} = 2.3$

und $\epsilon_{vali} = 2.5$ nicht ganz erreicht. Bezüglich eines von [22] vorgenommenen Modellgütenvergleichs mit vier anderen aus der Literatur bekannten Lernverfahren ändert dieser Verlust an Modellgenauigkeit jedoch nicht die Einordnung auf Platz 2 im Ranking.

5 Zusammenfassung und Ausblick

In diesem Beitrag wird ein Verfahren zur Selektion relevanter Sätze nichtredundanter Einflussgrößen vorgestellt. Das Verfahren quantifiziert dabei die Relevanz eines Satzes von Einflussgrößen, einem sogenannten *Merkmalssatz*, gesamtheitlich. Anhand eines synthetischen Demonstrationsbeispiels wird die Arbeitsweise und Leistungsfähigkeit des Verfahrens unter Verwendung verschiedener Suchstrategien demonstriert. Des Weiteren wird das Verfahren auf etablierte Benchmarkprobleme angewendet, um einerseits die Effizienz der *Greedy–Suche* im Vergleich zu einer *Vollständigen Suche* abzuschätzen und andererseits das Verfahren im Sinne einer Datenvoranalyse für eine Fuzzy–Modellierung zu nutzen.

Bezüglich der Suchstrategien wird in allen bearbeiteten Beispielen mittels einer *Greedy–Suche* in wesentlich kürzerer Zeit als mit einer *Vollständiger Suche* eine näherungsweise gleich gute Lösung gefunden, womit die Ergebnisse aus [7, 8] bestätigt werden können.

Die auf den ausgewählten Größen aufsetzende Anwendung des Fuzzy–ROSA–Verfahrens führt — bis auf die Ausnahme des Beispiels WINE — zu Fuzzy–Systemen mit Modellierungsgüten, die über oder sehr nahe an den bisher besten erzielten Ergebnissen liegen. Dies ist um so bemerkenswerter, als dass die Fuzzy–Modellierung eine Granularisierung der verwendeten Einflussgrößen erfordert, die durch das Verfahren weder optimiert noch berücksichtigt werden. Des Weiteren ist es durchaus vorstellbar, dass bei der Fuzzy–Modellierung auf der Basis der ausgewählten Einflussgrößen durch eine andere Parametrisierung des Lernverfahrens noch besserer Ergebnisse erzielt werden können.

Zusätzlich illustrieren die Ergebnisse, dass der zur Modellierung verwendete selektierte Merkmalssatz von relevanten und nichtredundanten Einflussgrößen die Interpretierbarkeit der generierten Fuzzy–Systeme meist deutlich erhöht. Deshalb und in Anbetracht der eher schlechten Ergebnisse bei WINE erscheint es verfolgungswert, das Verfahren durch eine Berücksichtigung oder gar Optimierung von Granularisierungen zu erweitern, um damit die Güte der nachgeschaltete Fuzzy–Modellierung noch zu erhöhen.

Danksagung

Diese Arbeit wurde von der Deutschen Forschungsgemeinschaft im Rahmen des Sonderforschungsbereiches 531 *„Computational Intelligence“* der Universität Dortmund gefördert.

Literatur

[1] John, G. H. ; Kohavi, R. ; Pfleger, K.: Irrelevant features and the subset selection problem. In: *Proc. International Conference on Machine Learning*, 1994. – Journal version in AIJ, S. 121–129

[2] Kohavi, K. ; John, G. H.: Wrappers for feature subset selection. In: *Artificial Intelligence* 97 (1997), Nr. 1–2, S. 273–324

[3] Kojadinovic, I. ; Wottka, T.: Comparision between a filter and a wrapper approach to variable subset selection in regression problems. In: *Proc. European Symposium on Intelligent Techniques (ESIT)*, 2000

[4] Hartung, J. ; Elpelt, B. ; Kloesener, K.-H.: *Statistik, 10. Auflage.* München : Oldenbourg Verlag, 1995

[5] Mathar, R.: *Informationstheorie.* Stuttgart : Teubner Verlag, 1996

[6] Xion, N.: *Designing Compact and Comprehensible Fuzzy Controllers Using Genetic Algorithms (Entwurf kompakter und interpretierbarer Fuzzy Controller mittels Genetischer Algorithmen).* Aachen : Shaker Verlag, 2001 (Berichte aus der Automatisierungstechnik)

[7] Imam, I. ; Vafaie, H.: An empirical comparison between global and greedy-like search for variable selection. In: *Florida AI Research Symposium (FLAIRS)*, 1994

[8] Deng, K. ; Moore, A.: On the greediness of feature selection algorithms. In: *Proc. International Conference on Machine Learning (ICML)*, 1998

[9] Holland, J. H.: *Adaptation in Natural and Artificial Systems.* Michigan, USA : Universität Michigan Press, Ann Arbor, 1975

[10] Goldberg, D. E.: *Genetic Algorithms in Search, Optimization and Machine Learning.* Reading, USA : Addison Wesley, 1989

[11] Kiendl, H. ; Krabs, M.: Ein Verfahren zur Generierung regelbasierter Modelle für dynamische Systeme. In: *at – Automatisierungstechnik* 37 (1989), Nr. 11, S. 423–430

[12] Krabs, M.: *Das ROSA–Verfahren zur Modellierung dynamischer Systeme durch Regeln mit statistischer Relevanzbewertung.* Düsseldorf : VDI Verlag, 1994 (Fortschritt–Berichte VDI, Reihe 8, Nr. 404)

[13] Kiendl, H.: *Fuzzy Control methodenorientiert.* München : Oldenbourg, 1997

[14] Krone, A.: *Datenbasierte Generierung von relevanten Fuzzy–Regeln zur Modellierung von Prozesszusammenhängen und Bedienstrategien.* Düsseldorf : VDI Verlag, 1999 (Fortschritt–Berichte VDI, Reihe 10, Nr. 615)

[15] Slawinski, T.: *Analyse und effiziente Generierung von relevanten Fuzzy–Regeln in hochdimensionalen Suchräumen.* Düsseldorf : VDI Verlag, 2001 (Fortschritt–Berichte VDI, Reihe 10, Nr. 686)

[16] Anderson, E.: The IRISes of the Gaspe Peninsula. In: *Bull. Amer. IRIS Soc.* 59 (1935), S. 2–5

[17] Pal, N. R. ; Pal, K. ; Bezdek, J. C.: A Mixed c–Means Clustering Model. In: *Proceedings of the Sixth IEEE International Conference on Fuzzy Systems (FUZZ–IEEE '97), Barcelona, Spanien, 1997* Bd. 1. Piscataway, NJ : IEEE Press, 1997, S. 11–21

[18] Corcoran, A. L. ; Sandip, S.: Using Real–Valued Genetic Algorithms to Evolve Rule Sets for Classifications. In: *Proceedings of the First IEEE Conference on Evolutionary Computation (ICEC '94), Orlando, USA, 1994* Bd. 1. Piscataway, NJ : IEEE Press, 1994, S. 120–124

[19] Prechelt, L.: PROBEN 1 – A Set of Neural Network Benchmark Problems and Benchmarking Rules / Fakultät für Informatik, Universität Karlsruhe. 1994 (21). – Forschungsbericht

[20] Mackey, M. ; Glass, L.: Oscillation and Chaos in Physiological Control Systems. In: *Science* 197 (1977), S. 287–289

[21] Harrison, D. ; Rubinfeld, D. L.: Hedonic Prices and the Demand for Clean Air. In: *Economics & Managment* 5 (1978), S. 81–102

[22] Krause, P.: *Datenbasierte Generierung von transparenten und genauen Fuzzy–Modellen für mehrdeutige Daten und komplexe Systeme.* Düsseldorf : VDI Verlag, 2001 (Fortschritt–Berichte VDI, Reihe 10, Nr. 691)