Qualitätsvergleiche
bei Kreditausfallprognosen

Professor Dr. Walter Krämer
Fachbereich Statistik, Universität Dortmund

1. Qualitative versus quantitative Prognosen

Wirtschaftsdaten als Objekte von Prognosen sind meist metrischer Natur: Arbeitslosenzahlen, Aktienkurse, Umsätze, Erlöse usw., alle sind quantitative Variable, bei denen sich Prognosen und realisierte Werte, wie auch konkurrierende Prognosen, leicht vergleichen lassen. Anders die Lage bei qualitativen, speziell dichotomen 0-1-Variablen, die im Zentrum der folgenden Überlegungen stehen. Hier ist der Vergleich von Prognosen und realisierten Werten, wie auch der Qualitätsvergleich konkurrierender Prognosen, erheblich schwerer.

Das folgende Kapitel diskutiert diese Problematik anhand von Kreditausfallprognosen. Unter Hintanstellung von Problemen, die mit der Definition von "Kreditausfall" verbunden sind, gibt es hier zwei Möglichkeiten: (i) der Kredit fällt aus und (ii) der Kredit fällt nicht aus, und die zahlreichen Verfahren, die es gibt – Diskriminanzanalyse, Logit- und Probit-Modelle, Neuronale Netze und Klassifikationsbäume – dieses Ereignis vorherzusagen (siehe Arminger et al. 1997 oder Blum et al. 2003 für eine Übersicht) müssen mit zwei Arten von Fehlern leben: Bei der Prognose "Kein Ausfall" tritt dennoch ein Ausfall ein – der Alpha-Fehler – oder bei einer Prognose von "Ausfall" tritt kein Ausfall ein – der Beta-Fehler. Je nach Bewertung und Wahrscheinlichkeit von Alpha- und

---

Research supported by Deutsche Forschungsgemeinschaft (DFG) under SFB 475.

Beta-Fehler lassen sich konkurrierende Prognosen dann hinsichtlich ihrer Prognosequalität vergleichen. Die einschlägigen Methoden sind seit langem wohlbekannt (siehe etwa Oehler und unser 2001, Kapitel III.2) und müssen hier nicht weiter erörtert werden. Die folgende Diskussion konzentriert sich vielmehr auf Prognosen, die nur die Wahrscheinlichkeiten für das interessierende Ereignis betreffen: Die Ausfallwahrscheinlichkeit bei Kredit X beträgt Y% mit $0 < Y < 100\%$. Dergleichen Wahrscheinlichkeitsprognosen haben in der Meteorologie und in der Medizin eine lange Tradition (siehe etwa DeGroot und Fienberg 1983, Redelmann et al. 1991 oder Winkler 1996), sind aber mit der wachsenden Bedeutung von Ratings und Rating-Agenturen im modernen Wirtschaftsleben auch dort in letzter Zeit vermehrt ins Rampenlicht getreten. Nimmt man noch die im Kielwasser von Basel II auf alle Geschäftsbanken zukommende Verpflichtung zur Belegung aller vergebenen Kredite mit Ausfallwahrscheinlichkeiten hinzu, so werden Wahrscheinlichkeitsprognosen in naher Zukunft zu den häufigsten Wirtschaftsprognosen überhaupt gehören.

Anders als bei Prognosevergleichen metrischer Variablen, steckt die einschlägige Methodologie aber hier noch in den Kinderschuhen. Im weiteren werden zunächst die gängigen Verfahren vorgestellt und auf ihre Eignung für die Praxis abgeklopft. Ferner werden Implikationsbeziehungen zwischen den durch die verschiedenen Qualitätskriterien definierten Halbordnungen abgeleitet, diskutiert und ausgewählte skalarwertige Gütemaße vorgestellt.

2. Trennschärfe und Kalibrierung

Angenommen, 2% aller Kredite eines größeren Portfolios fallen erfahrungsgemäß binnen eines festen Zeitraums, etwa eines Jahres, aus. Eine Rating-Agentur A, um eine Bewertung der Kredite dieses Portfolios gebeten, versieht jeden davon mit dem Etikett "Ausfallwahrscheinlichkeit 2%".

Diese Prognose ist "kalibriert" (synonym auch "valide" = valid oder "zuverlässig" = reliable, siehe Sanders 1963 oder Murphy 1973). Kalibriert bedeutet: Unter allen Krediten mit dem Etikett "Ausfallwahrscheinlichkeit x%" fallen langfristig x% tatsächlich aus.

Trotzdem ist dieses Rating wertlos – es liefert keine neuen Informationen, das alles hat man vorher schon gewußt. Oder anders ausgedrückt: Kalibrierung ist eine notwendige, aber keine hinreichende Bedingung für eine "gute" Wahrscheinlichkeitsprognose.

Agentur B teilt das Portfolio in zwei Gruppen auf: die erste mit Ausfallwahrscheinlichkeit 1%, die zweite mit Ausfallwahrscheinlichkeit 3%. Auch diese Bewertung sei kalibriert: In der ersten Gruppe fallen tatsächlich 1%, in der zweiten 3% der Kredite aus. Dann ist Agentur B ganz offensichtlich "besser" als Agentur A.

Das Rating von B heißt auch "trennschärfer" als das von A (synonym auch "sharper" oder "more refined", siehe Sanders 1963 oder DeGroot und Fienberg 1983). Trennschärfe ist ein Maß für das "Spreizen" der Wahrscheinlichkeitsprognosen in Richtung 0 bzw. 100%. Die trennschärfste Wahrscheinlichkeitsprognose läßt nur zwei Aussagen zu: "Ein Kredit fällt sicher aus" (Prognose 100%), oder "ein Kredit fällt sicher nicht aus" (Prognose 0%). Ist eine solche extrem trennscharfe Prognose außerdem noch kalibriert, dann ist sie absolut perfekt: Das Rating sagt jeden Kreditausfall mit Sicherheit exakt voraus.

Auch bei kalibrierten, aber nicht maximal trennscharfen Prognosen ist es sinnvoll, nachzufragen: Welches von mehreren kalibrierten Rating-Systemen kommt dem Ideal einer maximal trennscharfen Prognose am nächsten? In obigem Beispiel ist System B trennschärfer als A. Und nochmals trennschärfer sind zwei Systeme, C und D, welche die Kredite in die Ausfallklassen 0,5%, 1,5% und 4,5% bzw. 0,5%, 1% und 3% aufteilen.

Tabelle 1 zeigt eine mit Kalibrierung verträgliche Verteilung der Kredite auf die verschiedenen Ausfallklassen in diesen vier Prognosesystemen:

Tabelle 1: Prognostizierte Ausfallwahrscheinlichkeiten und ihre Verteilung auf die Gesamtzahl der Kredite

Prognostizierte Ausfallwahrscheinlichkeit	Verteilung der Kredite auf die prognostizierten Ausfallwahrscheinlichkeiten			
	A	B	C	D
0,5%	0	0	0,25	0,2
1%	0	0,5	0	0,25
1,5%	0	0	0,5	0
2%	1	0	0	0
3%	0	0,5	0	0,55
4,5%	0	0	0,25	0

Mathematisch ist "trennschärfer" bei kalibrierten Prognosen dadurch definiert, daß sich die trennschwächere Prognose in gewissem Sinn aus der trennschärferen ableiten läßt. Das ist bei einem Vergleich von A und B ganz offenbar der Fall: Unabhängig vom B-Etikett erhalten alle Kredite unter A die Prognose 2%. Aber auch die B-Prognose läßt sich ihrerseits aus der C-Prognose ableiten: Alle Kredite mit der C-Prognose 0,5% und eine zufällig ausgewählte Hälfte aller Kredite mit der C-Prognose 1,5% erhalten das Etikett 1%, die übrigen das Etikett 3%. Das Ergebnis ist eine kalibrierte Prognose mit der gleichen Trennschärfe wie B.

Die B-Prognose läßt sich aber auch aus der D-Prognose ableiten: Alle D-Prognosen 0,5% und 1% sowie ein zufällig ausgewähltes Elftel der D-Prognosen 3% erhalten das Etikett 1%, die übrigen das Etikett 3%. Das Ergebnis ist wieder eine kalibirierte Prognose mit der gleichen Trennschärfe wie B.

Die Prognosen C und D lassen sich jedoch in diesem Sinne nicht vergleichen: Weder ist D trennschärfer als C, noch C trennschärfer als D. Die Trennschärfe erzeugt also keine vollständige Ordnung, sondern nur eine Halbordnung unter allen kalibrierten Wahrscheinlichkeitsprognosen; es gibt kalibrierte Wahrscheinlichkeitsprognosen, die nach dem Kriterium der Trennschärfe nicht vergleichbar sind.

Formal: Seien $0 = p_1 < p_2 < \ldots p_k = 1$ die vorhergesagten Ausfallwahrscheinlichkeiten und $q^A(p_i)$ bzw. $q^B(p_i)$ die relativen Häufigkeiten, mit denen diese Vorhersagen getroffen werden. Dann ist A trennschärfer als B genau dann, wenn

$$q^B(p_i) = \sum_{j=1}^{k} M_{ij} q^A(p_j) \qquad \text{und} \tag{1}$$

$$p_j q^B(p_i) = \sum_{j=1}^{k} M_{ij} p_j q^A(p_j) \tag{2}$$

mit einer K×K Markoff-Matrix M (einer Matrix mit nichtnegativen Elementen, deren Spalten sich zu 1 addieren (siehe DeGroot und Fienberg 1983). Dabei formalisiert Gleichung (1) das auf A's Prognosen angewandte Randomisieren, und Gleichung (2) garantiert die Kalibrierung dieser so entstandenen neuen Prognosen.

3. Weitere Halbordnungen von Wahrscheinlichkeitsprognosen

Unabhängig von Trennschärfe und Kalibrierung ist es sinnvoll, beim Vergleich zweier Ratingsysteme A und B zu fragen: "Welches der beiden Systeme gibt den ausgefallenen Kredite die schlechtesten Bewertungen?" Diese Frage führt zum Begriff der "Ausfalldominanz" (Vardeman und Meeden 1983): Ein Ratingsystem A ist besser als ein Ratingsystem B im Sinne der Ausfalldominanz, falls A die ausgefallenen Kredite systematisch schlechter einstuft als B.

Formal: Sei $q_A(p_i|1)$ der Anteil der ausgefallenen Kredite, die von System A in die durch die prognostizierte Ausfallwahrscheinlichkeit p_i (i=1,...,K) definierte Ratingklasse einsortiert worden sind. Analog $q_B(p_i|1)$ usw. Dann ist A besser als B im Sinne der Ausfalldominanz, falls

$$\sum_{i=1}^{j} q_A(p_i\,|1) \;\le\; \sum_{i=1}^{j} q_B(p_i\,|1) \quad \text{für alle } j = 1,\dots,K. \tag{3}$$

In kalibrierten Ratingsystemen errechnen sich die $q_A(p_i|\,1)$ durch

$$q_A(p_i\,|1) = \frac{p_i \times q_A(p_i)}{p} \;, \tag{4}$$

mit p als Gesamtausfallwahrscheinlichkeit für alle Kredite insgesamt.

Analog läßt sich auch in Bezug auf die nicht ausgefallenen Kredite fragen, ob eines von zwei zu vergleichenden Rating-Systemen diese systematisch besser bewertet. Sei dazu $q_A(p_i\,|\,0)$ der Anteil der nicht ausgefallenen Kredite, die von System A in die verschiedenen Ratingklassen $p_i(i=1,\dots,K)$ eingeordnet worden sind. Analog $q_B(p_i\,|1)$. Dann ist A besser als B im Sinn der Nichtausfall-Dominanz, falls

$$\sum_{i=1}^{j} q_A(p_i \mid 0) \geq \sum_{i=1}^{j} q_B(p_i \mid 0) \text{ für alle } j = 1,...,K. \tag{5}$$

In kalibrierten Rating-Systemen errechnen sich die $q_A(p_i \mid 0)$ als

$$q_A(p_i \mid 0) = \frac{(1-p_i) \times q_A(p_i)}{1-p} \; . \tag{6}$$

In der Sprache der Mathematik handelt es sich hier um einen Vergleich von Wahrscheinlichkeitsverteilungen über Ratingklassen. System A ist in dieser Sprache besser als System B im Sinne der Ausfalldominanz, wenn die bedingte Verteilung von A, gegeben Ausfall, diejenige von B stochastisch dominiert. Und A ist besser als B im Sinne der Nichtausfall-Dominanz, wenn die bedingte Verteilung von B, gegeben kein Ausfall, diejenige von A stochastisch dominiert.

Analog läßt sich auch der Trennschärfe-Vergleich aus Abschnitt 2 in die Sprache der stochastischen Dominanz übertragen (DeGroot und Eriksson 1985): Ein kalibiriertes System A ist genau dann trennschärfer als ein kalibriertes System B, wenn die unbedingte Verteilung der Kredite auf die Ratingklassen unter A diejenige von B stochastisch in 2. Ordnung dominiert.

Ausfall-Dominanz und Nichtausfall-Dominanz sind je für sich sehr leicht zu erzeugen: Durch hinreichende Erhöhung der vorhergesagten Ausfallwahrscheinlichkeiten (und damit natürlich unter Preisgabe einer vorher vielleicht vorhandenen Kalibrierung) wird jedes System einen vorgegebenen Konkurrenten letztendlich im Sinne der Ausfalldominanz schlagen. Gleiches gilt für die Nichtausfall-Dominanz, wenn man die vorhergesagten Ausfallwahrscheinlichkeiten hinreichend reduziert.

Schwieriger ist dagegen eine Qualitätsverbesserung sowohl im Sinne der Ausfall- als auch im Sinne der Nichtausfall-Dominanz. Ein System, welches ein anderes in diesem Sinne dominiert, heißt im weiteren auch “besser im Sinne der doppelten Ausfallordnung".

Die doppelte Ausfallordnung ist ein sehr anspruchsvolles Kriterium. Wie man sich leicht überzeugt, ist sie für kein einziges Paar der in Tabelle 1 aufgelisteten Wahrscheinlichkeitsprognosen gegeben. Ganz allgemein läßt sich zeigen (siehe Krämer 2002, Satz 1), daß die doppelte Ausfallordnung mit Kalibrierung im wesentlichen unverträglich ist: Wenn für zwei kalibrierte Wahrscheinlichkeitsprognosen A und B gilt: $q_A(0) = q_B(0) = 0$, so ist die Ausfallordnung ausgeschlossen. Und für $q_A(1) = q_B(1) = 0$ ist die Nichtausfallordnung ausgeschlossen.

Unabhängig von Kalibrierung nennen DeGroot Und Fienberg (1983) deshalb eine Prognose A "suffizient" für B, wenn B's bedingte Verteilungen der Kredite auf die Ausfallwahrscheinlichkeiten p_i, gegeben sowohl Ausfall als auch Nicht-Ausfall, aus denen von A durch Randomisieren abgeleitet werden können:

$$q^B(a_i|\theta) = \sum_{j=1}^{k} M_{ij} q^A(p_j|\theta)\,, \quad i = 1, ..., k; \quad \theta = 0,1 \tag{7}$$

mit einer Markoff-Matrix M. Für kalibrierte Prognosen stimmt die so induzierte Halbordnung mit der durch den Trennschärfe-Vergleich induzierten Halbordnung überein.

Ein weiteres, von Kalibrierung unabhängiges und in der Praxis gern benutztes Qualitätskriterium (siehe z. B. Falkenstein et al. 2000) gründet sich auf dem Polygonzug durch die Punkte

$$(0,0)\,,\left(\sum_{i=0}^{j-1} q(p_{k-i}) \quad , \quad \sum_{i=0}^{j-1} q(p_{k-i}|1)\right), \quad j = 1, ..., k \tag{8}$$

Diese Kurve heißt in der angelsächsischen Literatur auch *power curve*, *cumulated accuracy profile* oder *gini curve* und sei im weiteren als Gini-Kurve bezeichnet. Eine Wahrscheinlichkeitsprognose A ist dann besser als eine Wahrscheinlichkeitsprognose B in diesem, dem Gini-Sinne, wenn A's Gini-Kurve nirgends unterhalb von der von B verläuft.

Ein System, das in jeder Rating-Klasse die gleichen prozentualen Ausfallanteile hätte, hat als Gini-Kurve die Diagonale. Dieses System liefert keine Informationen und ist in diesem Sinne das schlechtest mögliche.

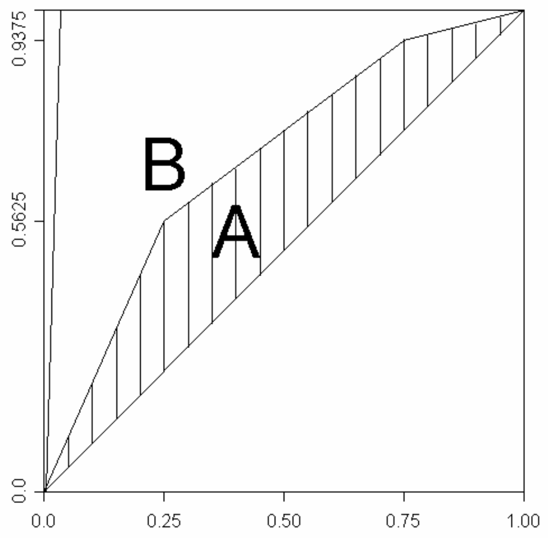
Angenommen, im Beispiel aus Abschnitt 2 seien insgesamt 800 Kredite zu bewerten. Agentur C prognostiziert für 200 davon eine Ausfallwahrscheinlichkeit von 0,5%, für 400 eine Ausfallwahrscheinlichkeit von 1,5%, und für 200 eine Ausfallwahrscheinlichkeit von 4,5%. Agentur C ist kalibriert, d.h. in der ersten Gruppe fällt im Mittel 1 Kredit (= 0,5% von 200) tatsächlich aus, in der zweiten Gruppe fallen 6 Kredite aus (= 1,5% von 400), in der dritten Gruppe 9 (= 4,5% von 200). Insgesamt gibt es 16 Ausfälle (2% von 800). Im weiteren sei der Einfachheit halber unterstellt, daß die erwarteten Ausfälle mit den tatsächlichen Ausfällen übereinstimmen. Gruppiert man die Kredite von schlecht nach gut, und stellt ihnen die kumulierten Anteile an den Ausfällen gegenüber, ergibt sich folgende Tabelle:

Tabelle 2: Bonität vs. Ausfallanteile für ein ausgewähltes Prognoseverfahren

Anteil an der Gesamtzahl der bewerteten Kredite	Anteile an der Gesamtzahl der Ausfälle
0	0/16
0,25	9/16
0,75	15/16
1	16/16

Diese Punkte, in ein 2-dimensionales Koordinatensystem übertragen und durch Geraden verbunden, erzeugen die in Abbildung 1 wiedergegebene Gini-Kurve der Prognose C. Ebenfalls eingezeichnet ist die optimale Gini-Kurve eines Ratingsystems, das alle 16 Ausfälle, und nur diese, in die schlechteste Bonitätsklasse aufgenommen hätte. Diese begrenzt zusammen mit der Winkelhalbierenden die Fläche B.

Abbildung 1:
Eine beispielhafte Ginikurve von Kreditausfällen



Das Verhältnis der Fläche A zur Fläche B heißt auch "Trefferquote" ("accuracy ratio"). Je höher die Trefferquote, desto näher kommt ein Rating-System an die in obigem Sinn optimale Prognose heran.

Alternativ betrachtet man oft auch die ROC-Kurve (ROC = "Receiver Operating Characteristic"), die durch die Punkte

$$(0,0), \left(\sum_{i=0}^{j-1} q(p_{k-i}|0) \quad , \quad \sum_{i=0}^{j-1} q(p_{k-i}|1) \right), \quad j = 1,\dots,k \tag{9}$$

und verbindende Geraden gegeben ist. ROC-Kurven sind vor allem in der medizinischen Diagnostik seit langem als Werkzeug zum Qualitätsvergleich konkurrierender Diagnosesysteme wohlbekannt (siehe Zweig und Campell 1993 oder Hajian-Tilaki und Henley 2002 für eine Übersicht). Da sich aber zwei Gini-Kurven genau dann schneiden, wenn sich die zugehörigen ROC-Kurven schneiden, sind die durch diese Kurven induzierten Halbordnungen äquivalent (Krämer 2002, Theorem 3). Außerdem ist die oft als skalares Qualitätskriterium genutzte Fläche unter der ROC-Kurve numerisch identisch zu der aus der Gini-Kurve abgeleiteten "Trefferquote" (siehe etwa Engelmann et al. 2003; diese Einsicht ist aber auch bei vielen anderen Autoren zu finden). Die ROC-Kurve liefert daher keine zusätzlichen Informationen und bleibt im weiteren außer Betracht.

Sowohl die Gini-Kurve als auch die ROC-Kurve sind invariant gegenüber monotonen Transformationen der vorhergesagten Ausfallwahrscheinlichkeiten. Nimmt die tatsächliche Ausfallwahrscheinlichkeit mit schlechter werdender Ratingklasse zu, heißt ein System auch "semikalibriert". Bei einem semikalibrierten Ratingsystem ist die Gini-Kurve konvex.

Man kann zeigen (siehe Krämer 2002, Theorem 5), daß für semikalibrierte Wahrscheinlichkeitsprognosen eine Ordnung bezüglich Suffizenz das Gini-Kriterium impliziert. Die Umkehrung gilt nicht. Analog folgt auch aus einer Überlegenheit im Sinne der doppelten Ausfallordnung eine Überlegenheit im Sinne des Gini-Kriteriums, unabhängig davon, ob die Prognosen semikalibriert sind oder nicht. Auch hier kann man durch einfache Gegenbeispiele zeigen, daß die Umkehrung nicht gilt.

4. Skalarwertige Abweichungsmaße

Eine alternative Möglichkeit zur Beurteilung der Qualität von Wahrscheinlichkeitsprognosen ist der direkte Vergleich von Prognosen und tatsächlich eingetretenen Ereignissen. Insgesamt gebe es n zu bewertende Kredite. Sei $p^j \in \{p_1,...,p_k\}$ die Prognose für Kredit j, und sei $\theta j = 1$ bei Ausfall und $\theta j = 0$, wenn kein Ausfall eintritt. Dann ist das Brier-Maß ("Brier-Score", nach G.W.Brier 1950) definiert als

$$B = -\frac{1}{n}\sum_{j=1}^{n}(p^j - \theta^j)^2. \qquad (10)$$

Das Brier-Maß ist das bekannteste Maß zur Bewertung von Wahrscheinlichkeitsprognosen. Er wurde und wird bislang vor allem zum Qualitätsvergleich von Wettervorhersagern eingesetzt, ist aber grundsätzlich in allen Kontexten einsetzbar, in denen Wahrscheinlichkeitsprognosen zu vergleichen sind.

Je kleiner das Brier-Maß, desto schlechter die Wahrscheinlichkeitsprognose. Der schlechtest mögliche Wert von B = -1 ergibt sich für eine Prognose von immer nur 0 oder 100% Wahrscheinlichkeit für Ausfall, bei der stets das Gegenteil des Vorhergesagten eintritt. Der bestmögliche Wert von 0 ergibt sich für eine Prognose von immer nur 0% oder 100% für Ausfall, bei der stets das Vorhergesagte tatsächlich eintritt.

Bei einem Gesamtausfall-Anteil p hat die Trivialprognose "Ausfallwahrscheinlichkeit von p für jeden Kredit" das (erwartete) Brier-Maß

$$\overline{B} = -p(1-p)^2 - (1-p)p^2. \qquad (11)$$

Dieser Ausdruck strebt für $p \rightarrow 0$ ebenfalls gegen 0 (dito für $p \rightarrow 1$). Das ist bei Anwendungen wie Kreditausfallprognosen, mit sehr kleinen Wahrscheinlichkeiten für das fragliche Ereignis, ein Problem. Es empfiehlt sich daher in den Anwendungen auf jeden Fall, einen realisierten Brier-Score relativ zu dem Trivialwert (11) zu sehen. Derart adaptierte Abweichungsmaße werden auch "skill-scores" genannt (Winkler 1994).

Es ist leicht zu überprüfen (De Groot und Fienberg 1983), daß ein Anwender sein subjektiv erwartetes Brier-Maß immer dann minimiert, wenn er als Prognose für die Ausfallwahrscheinlichkeit seine wahre subjektive Ausfallwahrscheinlichkeit einsetzt. Insofern belohnt das Brier-Maß „ehrliches“ Verhalten. Abweichungsmaße mit dieser Eigenschaft heißen in der angelsächsischen Literatur auch "proper scoring rules" (Winkler 1996).

Ein deutscher Ausdruck dafür wäre „anreizkompatible Abweichungsmaße“. Ein weiteres anreizkompatibles Abweichungsmaß ist die Mittlere logarithmische Abweichung (Good 1952)

$$L = \frac{1}{n} \sum_{j=1}^{n} - \log\left(\left|p^{j} + \theta^{j} - 1\right|\right) \tag{12}$$

Anreizkompatible Abweichungsmaße wie das Brier-Maß oder die Mittlere logarithmische Abweichung bieten sich als Entlohnungskriterium für Kreditsachbearbeiter an: Es lohnt sich, die wahren subjektiven Ausfallwahrscheinlichkeiten offenzulegen. Untertreibungen oder Übertreibungen der subjektiv für richtig gehaltenen Ausfallwahrscheinlichkeiten verschlechtern den subjektiven Erwartungswert des Abweichungsmaßes und werden insofern bestraft.

Angesichts der Vielzahl der in der Literatur vorgeschlagenen Abweichungsmaße ist es sinnvoll, nach einem Kriterium zu fragen, welches garantiert, daß zwei Wahrscheinlichkeitsprognosen bezüglich <u>aller</u> anreizkompatiblen Abweichungsmaße die gleiche Reihung erfahren. Dazu seien die vorhergesagten Ausfallwahrscheinlichkeiten mit den empirisch beobachteten Ausfallraten gleichgesetzt. Per definitionem sind damit die Prognosen kalibriert, und es lässt sich zeigen (siehe Krämer 2003b), daß eine Prognose A eine Prognose B <u>für alle</u> anreizkompatiblen Abweichungsmaße genau dann dominierten, wenn sie, unter Verwendung dieser vorhergesagten Wahrscheinlichkeiten, trennschärfer ist als B.

Unabhängig von der Art des verwendeten Abweichungsmaßes stellt sich ferner das Problem seiner stochastischen Eigenschaften. Ist ein Prognosesystem "systematisch" besser als eine Zufallsprognose (d.h. ist die Trefferquote "signifikant" größer als Null)? Ist eine Rating-Agentur tatsächlich "besser" als die Konkurrenz, oder geht ein Vorsprung, etwas gemessen durch die Trefferquote oder den Brier-Score, nur auf zufällige Abweichungen der Stichprobe von den "wahren" Populationsparametern zurück? Hier gibt es erste Ansätze (siehe etwa Redelmeier et al. 1991 oder Engelmann et al. 2003, die einen Signifikanztest für die Trefferquote entwickeln), aber im großen und ganzen steht eine Antwort auf diese Fragen im Augenblick noch aus.

Literatur

Arminger, G., Enache, D. and Bonne, T. (1997): "Analyzing credit risk data: A comparison of logistic discrimination, classification tree analysis, and feedforward networks." *Computational Statistics* 12, 293 – 310.

Blume, C., Overbeck, L. and Wagner, L. (2003): *An introduction to credit risk modelling*, Boca Raton (Chapman \& Hall/CRC).

Brier, G.W. (1950): "Verification of forecasts expressed in terms of probability." *Monthly Weather Review* 78, 1 – 3.

DeGroot, M. und Fienberg, S.E. (1983): "The comparison and evaluation of forecasters." *The Statistician* 32, 12 – 23.

DeGroot, M. und Eriksson, E.A. (1985): “Probability forecasting, stochastic dominance, and the Lorenz curve,“ in: S. S. Gupta und J. O. Berger (Hrsg): *Statistical decision theory and related topics III*, Vol 1, New York (Academic Press), S. 291-314.

Engelmann, B., Hayden, E. und Tasche, D. (2003): "Testing rating accuracy." *Risk* 16, 82 – 86.

Falkenstein, E., Boral, A. und Kocagil, A.E. (2000): “RiskCalc for private companies II: More results and the Australian Model.“ *Moody’s Investor Services*, Report No. 62265.

Good, I.J. (1952): "Rational decisions." *Journal of the Royal Statistical Society B* 14, 107 – 114.

Hajian-Tilaki, K. and Henley, J.A. (2002): “Comparison of three methods for estimating standard error of the area under the curve in ROC analysis of quantiative data.“ *Academic Radiology* 9, 1278 – 1285.

Krämer, W. (2002): "On the ordering of probability forecasts." SFB 475 Diskussionspapier 50/02, Dortmund.

Krämer, W. (2003a): "Die Bewertung und der Vergleich von Kreditausfallprognosen." *Kredit & Kapital* 36, 395 – 410.

Krämer, W. (2003b): "Evaluating probability forecasts in terms of refinement and strictly proper scoring rules." SFB 475 Diskussionspapier 24.

Murphy, A.H. (1973): "A new vector partition of the probability score." *Journal of Applied Meteorology* 12, 595 – 600.

Oehler, A. und Unser, M. (2001): *Finanzwirtschaftliches Risikomanagement*, Berlin (Springer) .

Redelmeier, D.A., Block, D.A. und Hickam, D.H. (1991): "Assessing predictive accuracy: How to compare Brier scores." *Journal of Clinical Epidemiology* 44, 1141 – 1146.

Sanders, F. (1963): "On subjective probability forecasting." *Journal of Applied Meteorology* 2, 191 – 201.

Vardeman, S. und Meeden, G. (1983): "Calibration, sufficiency and domination considerations for Bayesian probability assessors." *Journal of the American Statistical Association* 78, 808 – 816.

Winkler, R.L. (1994): "Evaluating probabilities: Asymmetric scoring rules." *Management Science* 40, 1395 – 1405.

Winkler, R.L. (1996): "Scoring rules and the evaluation of probabilities." *Test* 5, 1 -- 60.

Zweig, Mark H. and Campbell, Gregory (1993): "Receiver-Operating Characterisic (ROC) plots: A fundamental evaluation tool in clinical medicine." *Clinical Chemistry* 39, 561 – 577.