September 2003

Ein Vergleich ausgewählter Klassifikationsverfahren im Kontext von Finanzdienstleistungen

Ralph Langner
Dresdner Bank AG

Paul Alpar
Philipps-Universität Marburg, alpar@wiwi.uni-marburg.de

Markus Pfuhl
Philipps-Universität Marburg

In: Uhr, Wolfgang, Esswein, Werner & Schoop, Eric (Hg.) 2003. *Wirtschaftsinformatik 2003: Medien - Märkte - Mobilität*, 2 Bde. Heidelberg: Physica-Verlag

ISBN: 3-7908-0111-9 (Band 1)
ISBN: 3-7908-0116-X (Band 2)

Ein Vergleich ausgewählter Klassifikationsverfahren im Kontext von Finanzdienstleistungen

Ralph Langner
Dresdner Bank AG

Paul Alpar, Markus Pfuhl
Philipps-Universität Marburg

Zusammenfassung: Verfahren des Data Mining sind in der Literatur oft anhand allgemein zugänglicher Datensätze verglichen worden. Diese Vergleiche sind jedoch von begrenztem Nutzen, wenn ein Verfahren für ein konkretes Anwendungsumfeld ausgewählt werden soll. Der Beitrag beschreibt deswegen die Auswahl eines Verfahrens für die Klassifikation von Privatkunden im Bereich Finanzdienstleistungen.

Schlüsselworte: Klassifikation, Scoring, Entscheidungsbäume, Regression, Metalernverfahren, ROC-Diagramme, Data Mining, Database Marketing

1 Einleitung

Klassifikation wird heute als eine der zentralen Aufgaben des Data Mining angesehen, wozu auch Prognose, Clustering, Abhängigkeitsanalyse und Abweichungsanalyse gehören. Data Mining ist Bestandteil eines interaktiven Prozesses, der als „Wissensentdeckung in Datenbanken“ (WED)[1] bezeichnet wird und aus den Schritten Selektion, Vorverarbeitung, Transformation, Data Mining[2] und Evaluation/Interpretation besteht [Fay+96, S. 41 f.].

Ziel der Klassifikation ist es, eine Regel zu finden, mit deren Hilfe Objekte aufgrund von beobachtbaren Merkmalen möglichst zutreffend einer von *k* zuvor festgelegten Klassen zugeordnet werden können. Zur Erstellung eines solchen Klassi-

---

1 Fayyad et al. definieren diesen als „nichttrivialen Prozess der Identifizierung valider, neuer, potentiell nützlicher und schließlich verständlicher Muster in Daten“ [Fay+96, S. 40 f.].

2 Da erst der Data Mining-Schritt ein messbares und meist anschaulich visualisierbares Ergebnis liefert, etablierte sich der Begriff Data Mining in der Praxis auch als Bezeichnung für den gesamten Prozess [AlNi00, S. 4].

fikators wird eine Beispielmenge von Objekten verwendet, für die neben den Merkmalsausprägungen auch die wahre Klassenzugehörigkeit bekannt ist.

Bei der Anwendung des Data Mining in der Praxis steht man häufig vor der Frage, welches Verfahren aus der Menge der zur Lösung einer Aufgabenstellung geeigneten Verfahren verwendet werden soll. Die Leistungsfähigkeit der Verfahren wird oft durch Anwendung auf standardisierte Testdaten beurteilt. Eine umfassende Beschreibung der gängigen Testdaten[3] aus den verschiedensten Anwendungsgebieten gibt [Mi+94].

Die so ermittelten Ergebnisse können aber nur teilweise auf tatsächliche Fragestellungen des Data Mining übertragen werden. Am Beispiel der Datensammlung „Credit Management“ [Mi+94, S. 132] zur Überprüfung von Klassifikationsalgorithmen sei ein Aspekt dieses Problems erläutert: Die ursprüngliche Datensammlung besteht aus je 20.000 Datensätzen pro Klasse („kreditwürdig“ bzw. „nicht kreditwürdig“). Da bei diesem Ergebnis die gängigen Entscheidungsbaumalgorithmen (wie z.B. CART) schlecht abschneiden, wurde die Datensammlung umgeschichtet, so dass nun 1.000 Datensätze der Klasse „nicht kreditwürdig“ und 19.000 Datensätze der Klasse „kreditwürdig“ in der Datensammlung „Credit Management“ enthalten sind.

Die Folge dieser Vorgehensweise ist, dass alle Vergleiche, die mit Hilfe dieser Datensammlung durchgeführt werden, nur Aussagekraft für diese idealisierte Datensammlung haben und dass deren Ergebnis nicht ohne Einschränkung auf beliebige Anwendungsfelder übertragen werden kann. Bei dieser Art der Datenvorbereitung wird Experten- bzw. Apriori-Wissen über die Verteilung der Klassifikationsergebnisse in einer Datensammlung vorrausgesetzt. Liegt kein Wissen über die Verteilung der Klassen vor, muss die Auswahl des zu verwendenden Klassifikationsalgorithmus auf einer anderen Grundlage bestimmt werden. Hierzu bieten sich Testdaten an, die aus dem untersuchten Anwendungs-feld stammen und für jede zu unterscheidende Klasse die gleiche Zahl von Daten-sätzen enthalten. Beim Vergleich auf Grundlage dieser Daten wird sichergestellt, dass kein Algorithmus ausgewählt wird, der zwar auf angereicherten Testdaten gute Ergebnisse liefert, aber im tatsächlichen Anwendungsumfeld schlechter abschneidet als andere Algorithmen.

Wie beschrieben setzen Testdaten einerseits Apriori-Wissen voraus, welches nicht immer vorliegt. Andererseits wird vorliegendes Wissen eventuell nicht berücksichtigt. Im später beschriebenen Anwendungsumfeld spielte z. B. der Verlauf des Deutschen Aktienindex (DAX) eine wichtige Rolle bei der Datenauswahl, obwohl ein Merkmal DAX-Wert weder in den Daten vorhanden war noch als ein Klassifikationsmerkmal benötigt wurde.

---

[3] Vgl. http://www.liacc.up.pt/ML/statlog/datasets.html, Abruf: 12.02.2003.

Gleiches gilt für die Auswahlentscheidung bzgl. der Datenquellen. Standardisierte Testdaten enthalten meistens keine Felder, deren Beitrag sich auf eine Klassifikationsentscheidung störend auswirkt. Aber gerade auf die Fähigkeit, mit verrauschten Daten zu arbeiten, ist im praktischen Einsatz von Data-Mining nicht zu verzichten. Aus diesem Grund sollte die Auswahl eines Algorithmus auf Grundlage von Daten erfolgen, die aus dem späteren Anwendungsfeld stammen oder eng mit diesem verwandt sind.

Im vorliegenden Aufsatz wird die Klassifikationsgüte verschiedener Klassifikationsverfahren deswegen im für die spätere Anwendung relevanten Umfeld der Finanzdienstleistungen verglichen.

2 Klassifikation

2.1 Bewertungsmöglichkeiten

2.1.1 Beurteilung eines Klassifikators

In der Regel wird ein Objekt durch die bekannten Eigenschaften nicht vollständig charakterisiert, so dass meist kein Klassifikator existiert, der eine hundertprozentig korrekte Klassenzuordnung erreicht. Selbst wenn der Klassifikator für jede auftretende Kombination aus Merkmalswerten eine eigene Regel besitzt, so können verschiedene Objekte mit gleichen Merkmalswerten zu unterschiedlichen Klassen gehören.

Zur Ermittlung der Klassifikationsgüte ist es zunächst nahe liegend, den Klassifikator auf die Objekte der Beispielmenge anzuwenden und die so erhaltenen Klassenprognosen mit den bekannten wahren Klassenzugehörigkeiten zu vergleichen. Das Ergebnis lässt sich in einer Konfusionsmatrix darstellen [WiFr01, S. 148].

		Vorhersage des Klassifikators	
		1	0
Tatsächliche Klasse	1	Wahre Positive (TP)	Falsche Negative (FN)
	0	Falsche Positive (FP)	Wahre Negative (TN)

Tabelle 1: Konfusionsmatrix für den Fall zweier Klassen

Setzt man nun die fehlerhaft klassifizierten Fälle (FN+FP) ins Verhältnis zu allen Beispielfällen, so erhält man die Resubstitutionsfehlerrate [Br+84, S. 11], die al-

lerdings eine sehr optimistische Fehlerschätzung liefert. Je feiner ein Klassifikator auf die Detailstruktur der Beispieldaten eingeht, desto weniger generalisierungsfähig ist er.

Um diese Überanpassung zu erkennen ist es sinnvoll, die Beispieldaten zunächst in eine *Trainings*- und eine *Testmenge* zu unterteilen. Die Erstellung des Klassifikators erfolgt dann ausschließlich mittels der Trainingsmenge, während die Testmenge der Schätzung der Klassifikationsgüte vorbehalten bleibt. Die Aufteilung kann dabei geschichtet erfolgen (*Stratifikation*), so dass die relativen Klassenhäufigkeiten gleich sind oder denen der Gesamtmenge entsprechen. Eine ungünstige Aufteilung könnte sonst dazu führen, dass alle Instanzen einer Klasse in der Testmenge liegen und folgerichtig nie vorhergesagt würden.

In der Praxis ist die Ermittlung der wahren Klassenzugehörigkeiten aber oft mit hohen Kosten verbunden, so dass nur eine relativ kleine Beispielmenge zur Verfügung steht. Eine weitere Verkleinerung dieser Menge durch Abspaltung der Testmenge gefährdet in diesem Fall die Repräsentativität und verschlechtert die Qualität des Klassifikators erheblich. Es bieten sich dann zwei weitere Möglichkeiten an, die Fehlerrate zu schätzen, die allerdings auch mehr Rechenaufwand erfordern:

- Bei der *v-fachen Kreuzvalidierung* wird die Beispielmenge in v disjunkte Mengen B_i (i=1,...,v) möglichst gleicher Größe aufgeteilt. Nun werden nacheinander v Klassifikationen mit der Trainingsmenge $\bigcup_{j \neq i} B_j$ (i=1,...,v) durchgeführt und auf der Testmenge B_i (i=1,...,v) validiert [Br+84, S. 12].
- Eine Möglichkeit, bei der die Fallzahl der Beispielmenge zum Training nicht eingeschränkt werden muss, eine unabhängige Testmenge zur Verfügung steht und das Klassifikationsverfahren nicht mehrfach angewendet zu werden braucht, stellt das *Bootstrapping* dar [WiFr01, S. 137]. Dabei wird die Trainingsmenge durch n-maliges Ziehen mit Zurücklegen aus der n Instanzen umfassenden Beispielmenge ermittelt. Nicht gezogene Instanzen bilden die Testmenge.

2.1.2 Vergleichsmöglichkeiten zwischen Klassifikationsverfahren

Neben einem Verfahrensvergleich über die testmengenbasierte Fehlerrate können auch Kriterien wie Geschwindigkeit, Skalierbarkeit, Robustheit und Interpretierbarkeit eine wichtige Rolle spielen [HaKa01, S. 137].

Die *Geschwindigkeit* hängt stark von der konkreten Implementierung eines Verfahrens und von der eingesetzten Hardware ab. So kann etwa ein Verfahren, das viele Sortieroperationen erfordert, durch einen zu gering dimensionierten Arbeitsspeicher deutlich gebremst werden. Die Geschwindigkeit ist insofern bedeutsam, als dass in der Praxis trotz immer leistungsfähiger werdender Hardware die Komplexität einiger Verfahren nur die Analyse auf einer Stichprobe erlaubt.

Eng damit verbunden ist die *Skalierbarkeit*, also die Fähigkeit, auch mit großen Datenmengen gute Ergebnisse zu erzielen. Die Messung kann für einen gegebenen Klassifikator über die Entwicklung der benötigten I/O-Operationen bei zunehmender Größe der Datenmenge erfolgen.

Bei der *Robustheit* werden die Auswirkungen einer sukzessiven Einführung von Rauschen auf die Fehlerrate gemessen.

Das Maß der *Interpretierbarkeit* ist stark subjektiv geprägt, kann aber durch die Messung der Komplexität des Klassifikators (z.B. Anzahl der Endknoten in einem Baum) bewertet werden.

2.1.3 Berücksichtigung asymmetrischer Kosten und schiefer Klassenverteilungen

Das bislang betrachtete Gütemaß Klassifikationsgenauigkeit[4] ist das in der Praxis am häufigsten verwendete und wird von den meisten Klassifikationsverfahren als Maximierungskriterium verwendet. Allerdings basiert es auf zwei Annahmen, die bei den wenigsten Problemstellungen erfüllt sein dürften:

Einerseits wird davon ausgegangen, dass die a priori-Wahrscheinlichkeiten für die einzelnen Klassen annähernd gleich sind. Ist diese Forderung verletzt, so kann sich das sehr ungünstig auf den Aufbau des Klassifikators auswirken: Soll etwa untersucht werden, ob bestimmte Merkmalskombinationen auf einen drohenden Kredit-Ausfall schließen lassen und sind innerhalb der betrachteten Personengruppe 3% der Kredite ausgefallen, so käme der einfache und zugleich nutzlose Klassifikator, der nie einen Kreditausfall prophezeit, auf eine sehr gute Klassifikationsgenauigkeit von 97%.

Zum anderen unterscheiden sich die Kosten für eine Fehlklassifikation oft erheblich: So verursacht ein Kreditausfall meistens erheblich höhere Kosten als der durch einen nicht gewährten Kredit entgangene Gewinn.

Eine naheliegende Lösung des ersten Problems ist die künstliche Herstellung gleicher Klassenhäufigkeiten durch Weglassen (*Downsampling*) oder Duplizieren (*Upsampling*) von Fällen. Während bei einer großen Zahl von Trainingsbeispielen das Downsampling eine adäquate Problemlösung darstellen dürfte, muss beim Upsampling im Falle weniger Trainingsdaten mit einer deutlichen Verzerrung durch die Duplizierung von Ausreißern gerechnet werden.

Die folgenden Betrachtungen beschränken sich auf den in der Praxis häufigsten Fall k=2, mit den Klassen *positiv* (p) und *negativ* (n).[5] Die Behandlung unterschiedlicher Klassifikationskosten im Verfahren selbst ist dann ebenfalls über eine künstliche Veränderung der Klassenverteilung möglich: Werden etwa positive

[4] Bzw. die Fehlerrate als Komplement

[5] Viele Klassifikationsverfahren setzen eine Klassengröße k=2 voraus.

Fälle übergewichtet, so führen irrtümlich als negativ klassifizierte Fälle durchschnittlich zu einer höheren Fehlerrate und umgekehrt. Da die meisten Klassifikationsverfahren aber versuchen, die Fehlerrate zu minimieren, lassen sich auf diese Weise die kostenintensiveren Fehler eher vermeiden [DrHo00b, S. 239; WiFr01, S. 154 f.; Br+84, S. 114 f.]. [6]

Viele Klassifikationsverfahren liefern nicht nur eine binäre Entscheidung für oder gegen eine Klasse, sondern geben dazu noch eine Schätzung der a posteriori-Wahrscheinlichkeit an, mit der diese Entscheidung richtig ist. Damit ist es möglich, den Klassifikator zunächst kostenunabhängig und mit der Option der Schaffung gleicher Klassenhäufigkeiten zu erstellen und erst anschließend die Kosten der Fehlentscheidungen zu berücksichtigen. Provost und Fawcett beschreiben, wie bei bekannten Fehlerkosten eine kostenoptimale Klassenzuordnung erfolgen kann, indem ein Schwellenwert berechnet wird, bis zu dem die nach absteigender a posteriori-Wahrscheinlichkeit geordneten Fälle als positiv klassifiziert werden [PrFa97, S. 44]. Allerdings werden dabei nur variable Kosten berücksichtigt.

2.1.4 ROC-Kurven

Bei ROC-Kurven[7] wird anstelle des eindimensionalen Maßes der Klassifikationsgenauigkeit eine Kombination aus zwei Kennzahlen betrachtet: Der Irrtums-Positivrate (False Positive Rate)

$$FPR = \frac{FP}{FP + TN} \tag{2.a}$$

auf der x-Achse wird die wahre Positivrate (True Positive Rate)

$$TPR = \frac{TP}{TP + FN} \tag{2.b}$$

auf der y-Achse gegenübergestellt.

Zur Erstellung des Diagrammes werden zunächst wieder die Instanzen der Testmenge nach absteigender a posteriori-Wahrscheinlichkeit geordnet. Dann werden schrittweise die Klassifikatoren betrachtet, die genau die ersten i Instanzen positiv klassifizieren. Ein Klassifikator entspricht somit einem Punkt im ROC-Diagramm. Durch Interpolation all dieser Klassifikatoren erhält man die Kurve zu einer Methode. Das Ergebnis eines Verfahrens ohne Angabe von a posteriori-

---

6 Offenbar liegen zur Bewertung dieser Art der Kostenbehandlung noch kaum empirische Ergebnisse vor. Dieser Ansicht sind z.B. auch Provost und Fawcett [PrFa97, S. 44].

7 Das aus der Signalerkennungstheorie stammende Akronym ROC steht für Receiver Operating Characteristic (Empfängerbetriebseigenschaften) und beschreibt dort die Abwägung zwischen Trefferrate und Fehlalarmrate.

Wahrscheinlichkeiten lässt sich als Punkt darstellen. Abbildung 1 stellt ein ROC-Diagramm mit drei verschiedenen Methoden A, B und C dar.

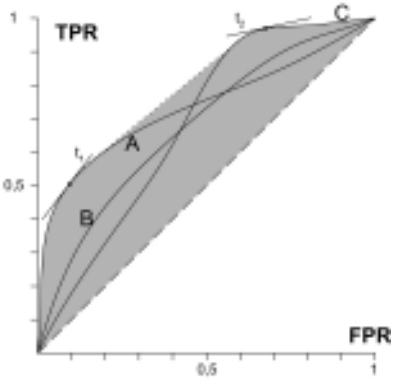

Abbildung 1: Beispiel für ein ROC-Diagramm

Die Punkte (0,0) bzw. (1,1) entsprechen der Strategie, eine Instanz niemals, bzw. immer als positiv zu bewerten. Der Punkt (0,1) kennzeichnet den perfekten Klassifikator, der alle positiven Instanzen als solche klassifiziert und keine Fehler macht. Die gestrichelte Verbindungslinie der Punkte (0,0) und (1,1) steht für ein zufälliges Raten der Klassenzugehörigkeiten. Allgemein ist ein Klassifikator umso besser, je näher er am optimalen Punkt (0,1) liegt, je höher also seine TPR und je niedriger seine FPR ist.

In Abbildung 1 dominiert keine der Kurven vollständig eine andere. Allerdings wird Kurve B in jedem Punkt dominiert und ist somit in keinem Punkt optimal. Ob nun Kurve A oder Kurve C besser ist, lässt sich aus dem Diagramm nicht ablesen, da es losgelöst von Kosten und Klassenverteilung ist.

Mit der ROC-Konvexe-Hüllen-Methode schaffen Provost und Fawcett einen Ansatz, um sowohl die Klassen- als auch die Kostenverteilung in das ROC-Modell zu integrieren [PrFa97]. Bezeichnen p(p) und p(n) die a priori-Wahrscheinlichkeiten der Klassen und c(1|n) und c(0|p) die durchschnittlichen Fehlklassifikationskosten für die Fälle FP bzw. FN, so besitzen zwei Punkte (TPR_1, FPR_1) und (TPR_2, FPR_2) genau dann die selben Kosten, wenn gilt:

$$\frac{TPR_2 - TPR_1}{FPR_2 - FPR_1} = \frac{p(n) \cdot c(1 \mid n)}{p(p) \cdot c(0 \mid p)} \tag{3}$$

Durch diese Gleichung wird die Steigung einer *Iso-Performance-Linie* definiert, so dass alle auf ihr liegenden Punkte dieselben erwarteten Kosten aufweisen. Weiter nordwestlich liegende Linien besitzen dabei wieder die niedrigeren Kosten. Die schraffierte Fläche in Abbildung 1 kennzeichnet die konvexe Hülle der Kurven. Dabei ist jeder Punkt des Bereiches, der von der gepunkteten Linie und den

Kurven A und C eingeschlossen wird, durch geeignete Kombinationen der Methoden A und C erreichbar. Um potentiell optimal zu sein, muss ein Punkt auf dem nordwestlichen Rand der konvexen Hülle liegen. t_1 und t_2 stehen für unterschiedliche Verteilungs-Kosten-Kombinationen. Entsprechend sind für die gegebenen Kurven die beiden optimalen Klassifikatoren gekennzeichnet.

Mit der ROC-Konvexe-Hüllen-Methode können die Ergebnisse verschiedener Klassifikationsverfahren beurteilt, verglichen und visualisiert werden. Soll dies aber zunächst ohne eine Kostenbewertung geschehen, so lässt sich die beste Methode nur bei einer globalen Dominanz bestimmen.

2.2 Verfahren zur Klassifikation

2.2.1 Entscheidungsbäume

Grundgedanke der Entscheidungsbaum-Verfahren ist es, die Klassifikation eines Objektes nicht in einem Schritt vorzunehmen, sondern aufgrund einer Reihe von hierarchisch angeordneten Tests, die jeweils durch einen Baumknoten repräsentiert werden. Dabei wird aufgrund des Ergebnisses eines Tests entschieden, ob und gegebenenfalls welche weiteren Tests noch erforderlich sind oder ob genügend Informationen vorliegen, um eine Prognose abzugeben.

Dem Aufbau von Entscheidungsbäumen liegt ein „teile und herrsche - Verfahren“ einer „gierigen Merkmalsauswahl“ zugrunde [BoKr98, S. 78]: Zu einer gegebenen Menge von klassifizierten Fallbeschreibungen werden die bedingten Häufigkeitsverteilungen der Klassen unter zur Trennung nutzbaren Attributen mit Hilfe eines Auswahlmaßes, etwa eines Reinheitsmaßes, bewertet.

Beim Aufbau von Entscheidungsbäumen können zwei Extreme auftreten: Während die Aussagekraft eines Baumes mit wenigen Knoten meist gering ist (*Underfitting*), liefern große, stark an die spezielle Struktur der Trainingsdaten angepasste Bäume bei der Anwendung auf andere Daten häufig schlechte Ergebnisse (*Overfitting*). Die meisten Entscheidungsbaum-Verfahren bauen zunächst einen verhältnismäßig großen Baum auf, der in einer darauf folgenden *Pruning-Phase* wieder gestutzt wird. Als Abbruchkriterium kann z. B. eine Mindestanzahl an Trainingsfällen innerhalb eines Endknotens verwendet werden.

2.2.2 Regressionsanalytische Verfahren

Aufgabe der Regression ist es, den Zusammenhang zwischen einer abhängigen Variablen y und einer oder mehreren unabhängigen Variablen $x_1,...,x_m$ zu analysieren. Bei der multiplen linearen Regression wird unterstellt, dass eine lineare Beziehung

$$y = \sum_{j=0}^{m} \beta_j \cdot x_j + \varepsilon, \qquad \mathrm{E}(\varepsilon) = 0 \tag{4}$$

besteht, bei der die Fehlervariable ε nicht gemessen werden kann. Die unbekannten Modellparameter $\beta_j \; (j = 0,..., m)$ sind dabei möglichst gut aus den Trainingsdaten zu schätzen. Obwohl sowohl die abhängige, als auch die unabhängigen Variablen als numerisch vorausgesetzt werden, lässt sich die Regression unter Zuhilfenahme geeigneter Transformationen zur Lösung des Klassifikationsproblems einsetzen.

Weiterhin wird durch den Einsatz statistischer Tests eine Auswahl relevanter Attribute möglich: Bei der *schrittweisen Regression* wird, beginnend mit einer leeren Attributmenge, solange das jeweils am besten geeignete Attribut hinzugefügt, bis ein Abbruchkriterium erfüllt ist, wobei Variablen, die sich zu einem späteren Zeitpunkt als nicht signifikant erweisen, auch wieder aus dem Modell entfernt werden können [FaHa96, S. 121].

2.2.3 Überwachte Künstliche Neuronale Netze

Die aus dem Bereich der künstlichen Intelligenz stammenden Künstlichen Neuronalen Netze (KNN) besitzen den Vorteil, dass kein expliziter funktionaler Zusammenhang zwischen verschiedenen Variablen unterstellt werden muss, wie das etwa bei der linearen Regression der Fall ist. Damit sind KNN besonders für die Entdeckung nichtlinearer Wirkungszusammenhänge geeignet und weisen eine hohe Toleranz gegenüber verrauschten Daten auf. Allerdings muss dafür mit langen Trainingszeiten gerechnet und auf eine anschauliche Erklärbarkeit der Ergebnisse verzichtet werden [Na$^+$98, S. 13; HaKa01, S. 303].

In Analogie zum menschlichen Gehirn besteht ein neuronales Netz aus einer Vielzahl von miteinander verbundenen Neuronen, die jeweils eine Reihe von Eingangssignalen mittels einer gewichteten Summe und einer anschließenden Transformation in ein Ausgangssignal umwandeln. Während unüberwachte KNN sich selbst organisieren, werden bei überwachten KNN im Zuge einer Trainingsphase die Gewichte adjustiert [HaKa01, S. 303; Ro$^+$94, S. 85 ff.].

2.2.4 Nächste-Nachbarn-Verfahren

Im Gegensatz zu regelbasierten Lernverfahren, bei denen aus der Trainingsmenge Entscheidungsregeln generiert werden, wird bei den Nächste-Nachbarn-Verfahren, als instanzbasierten Lernverfahren, zur Klassifikation eines neuen Objekts jeweils

der gesamte Trainingsbestand benötigt.[8] Voraussetzung ist dabei die Repräsentation der Objekte mittels numerischer Attribute in einem *m*-dimensionalen Raum, wobei davon ausgegangen wird, dass Objekte, die bezüglich eines vorzugebenden Abstandsmaßes nahe beieinander liegen, tendenziell der gleichen Klasse angehören. Die Klassifikation eines neuen Objekts erfolgt deshalb durch Zuweisung der Klasse, die das oder die nächstgelegenen Objekte aufweisen, bzw. aus einer daraus abgeleiteten Klasse [EsSa00, S. 119 ff.].

2.2.5 Weitere Verfahren aus der klassischen Statistik

Neben den Verfahren des Data Mining und der Regressionsanalyse bietet die klassische Statistik weitere Verfahren, die eine Klassifikation ermöglichen.

Bayes-Klassifikatoren optimieren für gegebene Hypothesen, a priori-Wahrscheinlichkeiten und Daten die Wahrscheinlichkeit einer korrekten Klassifikation [EsSa00, S. 112]. Da der optimale Bayes-Klassifikator bei sehr vielen Attributen eine praktisch nicht verfügbare Anzahl von Trainingsfällen benötigen würde, wird beim naiven Bayes-Klassifikator die vereinfachende Annahme getroffen, dass die Attribute für jede auftretende Klasse bedingt unabhängig sind. Dadurch verfälschen redundante Merkmale die Klassifikationsergebnisse oft erheblich.

Die Diskriminanzanalyse als Verfahren der multivariaten Statistik untersucht die Abhängigkeit einer nominal skalierten Gruppierungsvariablen von den metrisch skalierten Merkmalsvariablen der untersuchten Objekte [Ba+00, S. 146]. Durch mehrfache Anwendung der Diskriminanzanalyse mit unterschiedlichen Gruppierungsvariablen kann geprüft werden, wie groß der Beitrag einzelner Variablen zur Unterscheidung der Objekte ist. Der so gewonnene Merkmalskatalog kann als Grundlage einer Klassifikationsentscheidung dienen.

Auch das Verfahren der Clusteranalyse kann zur Klassifikation von Objekten eingesetzt werden. Ziel des Verfahrens ist es, die Gesamtheit aller untersuchten Objekte in Gruppen (Cluster) aufzuteilen. Dabei soll die Ähnlichkeit zwischen Objekten eines Clusters möglichst groß sein und gleichzeitig die Ähnlichkeit zwischen Objekten in unterschiedlichen Clustern möglichst gering. Als Ergebnis erhält man eine Klassifikation aller Objekte, wobei die Anzahl der Gruppen im vorhin festgelegt werden kann.

2.2.6 Meta-Lernverfahren

Zumal es kein global dominantes Klassifikationsverfahren gibt, welches in allen Situationen die besten Ergebnisse liefert, sondern verschiedene Verfahren in un-

[8] Um eine Fallklassifikation auch bei großen Datenbeständen performant durchführen zu können, ist die Verwendung effizierter räumlicher Indexstrukturen, beispielsweise Baumstrukturen, erforderlich [HaKa01, S. 314 f.].

terschiedlichen Situationen ihre Stärken oder Schwächen aufweisen, liegt die Überlegung nahe, die Stärken mehrerer Verfahren zu kombinieren. Dazu werden zunächst *s* verschiedene Klassifikationsverfahren, so genannte *Level-0-Verfahren*, auf eine Trainingsmenge L_0 angewendet und damit Klassifikatoren $K_q\,(q=1,...,s)$ aufgebaut. [WiFr01, S. 282 ff.] Daraufhin wird jeder dieser Klassifikatoren auf eine von L_0 unabhängige weitere Trainingsmenge L_1 angewendet. Die Vorhersagen der einzelnen Level-0-Verfahren ergeben Vektoren $V_q\,(q=1,...,s)$. Fasst man jeden dieser Vektoren als ein Merkmal auf, so erhält man durch die Matrix $\left(V_1,...,V_s,V^{True}\right)$, bei der V^{True} den Vektor mit den tatsächlichen Klassen der Trainingsmenge L_1 enthält, eine weitere Trainingsmenge $\tilde{L}_1$ für ein *Level-1-Verfahren*, also ein beliebiges Klassifikationsverfahren, das dann auf den Ergebnissen anderer Klassifikationsverfahren aufbaut.

Weitere Metalernverfahren, die allerdings jeweils das gleiche Verfahren mehrfach verwenden, sind das Boosting [WiFr01, S. 277 ff.] und das Bagging [Br96].

3 Klassifikation von Kunden im Bereich Finanzdienstleistungen

3.1 Anforderungen aus der Praxis

Wie in anderen Bereichen besteht auch bei Finanzdienstleistungen der Bedarf nach Klassifikation verschiedener Objekte, z. B. Kunden, Produkte, Filialen. Wir konzentrieren uns hier auf Kunden und zwar Privatkunden. Typische Klassifikationsanwendungen in diesem Bereich sind

- die Selektion von Kunden für ein Mailing.
- die Zuordnung von Kunden zu Affinitätsklassen bzgl. bestimmter Finanzprodukte.
- die Einstufung von Kunden nach Abwanderungsgefährdung im Rahmen eines Retention-Konzepts.
- die Präsentation personenspezifischer Inhalte und Angebote im Webportal.

Dabei ist neben der Frage, ob ein Klassifikationsansatz technisch erfolgversprechend erscheint, nicht zuletzt vor dem Hintergrund aktueller Sparmaßnahmen ein Kosten-Nutzen-Kalkül über die Rentablität erforderlich. Im Falle des Mailings besteht zwar regelmäßig ein enormes Einsparpotential durch Ansprache ausschließlich der erfolgversprechenden Kandidaten. Allerdings erhöhen sich dadurch

die aktionsspezifischen Fixkosten noch um die Kosten für die Analyse. Hilfreich ist in diesem Zusammenhang die Angabe von a posteriori-Wahrschein-lichkeiten, mittels derer sich etwa 50.000 Personen mit der höchsten erwarteten Reaktionswahrscheinlichkeit selektieren lassen.

Im Falle des individualisierten Portalauftritts lässt sich ein monetärer Nutzen nicht bestimmen. Vielmehr ist die Frage, ob und mit welchem Budget ein derartiges Vorhaben realisiert wird, in Abhängigkeit von der Unternehmensstrategie zu entscheiden.

Besonders bei den zuerst genannten Beispielen ist neben der Klassenzuordnung auch die *Beschreibung* der Klassen von Interesse, die je nach Verfahren durch die Angabe der Einflüsse der Prediktor-Variablen erfolgen kann. Eine einfache Interpretierbarkeit und Visualisierbarkeit der Ergebnisse, etwa durch eine Scorecard, erleichtert dabei nicht zuletzt die Kommunikation und Umsetzung der Resultate.

Wie bereits ausgeführt ist die Anwendung eines Klassifikationsverfahrens in einen WED-Prozess eingebettet, so dass das Ergebnis auch von den übrigen Schritten abhängt. Je nach anzuwendendem Verfahren ist etwa eine bestimmte Art der Datentransformation erforderlich. Andererseits werden Resultate häufig sehr zeitnah gefordert, so dass ein für ähnliche Aufgabenstellungen geeigneter Rahmenprozess, der auf ein möglichst universell einsetzbares und schnell konfigurierbares Klassifikationsverfahren ausgerichtet ist, die Schaffung einer Lösung deutlich beschleunigen kann. Weiterhin ist die fachliche Selektion potentiell relevanter Attribute meist sehr zeitintensiv und erfolgt oft nur in grober Form. Redundante Merkmale sollen deshalb möglichst im Prozess als solche erkannt werden. Ein robustes Verfahren, dass durch solche Merkmale nur unwesentlich beeinflusst wird, ist von Vorteil.

Kosteneffizienz, Einflussbestimmung der Prediktoren, universelle Anwendbarkeit und Robustheit sind in der Praxis damit meist wichtiger als die reine mathematische Qualität des Verfahrens.

3.2 Verfahren im Praxistest

3.2.1 Aufbau eines repräsentativen Testumfelds

Der erste Schritt besteht in der Festlegung einer konkreten Zielsetzung, die die wesentlichen, in Abschnitt 3.1 genannten Anforderungen beinhaltet.

Diese bildet dann die Basis für eine fachlich und wirtschaftlich sinnvoll erscheinende Datenvorauswahl. Dabei ist zu entscheiden, ob der bereits vorliegende interne Datenbestand, der noch von offensichtlich irrelevanten Merkmalen zu befreien ist, um weitere bedeutsam erscheinende Daten erweitert werden soll. Bei dieser Entscheidung ist neben einem Kosten-Nutzen-Kalkül auch darauf zu ach-

ten, dass die Dauer der Datenbeschaffung die Erfüllung der Aufgabe nicht gefährdet und dass die zeitliche Konsistenz im Datensatz gewahrt bleibt. Daraufhin sind die Daten zusammenzuführen und mit geeigneten Methoden vorzuverarbeiten.

Die Auswahl der zu vergleichenden Verfahren selbst erfolgt unter Berücksichtigung der in Abschnitt 3.1 genannten praktischen Anforderungen. Um die Robustheit verschiedener Verfahren zu testen, werden aus dem entstehenden Analysedatenbestand unterschiedlich verrauschte Teilmengen als Test-Datensätze herangezogen.

Mit Hilfe von Transformationsverfahren werden die Daten schließlich gemäß den Anforderungen eines konkreten Klassifikationsverfahrens in eine verarbeitbare Form überführt.

Für den Vergleich werden sowohl die Klassifikationsgenauigkeiten, als auch ROC-Diagramme eingesetzt.

3.2.2 Eine konkrete Problemstellung

Im Zuge des Aktienbooms bis Anfang 2000 hielten viele Anleger eine strukturierte Beratung für unnötig, da ihr Depot auch ohne diese erhebliche Wertzuwächse verzeichnete. Daher erschien ihnen ein preisgünstigeres Depot bei einer Direktbank vorteilhafter, bei der diese Beratungsleistung nicht mitbezahlt werden musste. Nach erfolgter Konsolidierung auf dem Aktienmarkt und aufgrund sinkenden Zinsen ist aber wieder mit einem stärkeren Beratungsbedarf im Wertpapierbereich zu rechnen. Da die Umstellung eines Depots mit einem gewissen organisatorischen Aufwand verbunden ist und viele Kunden im Beratungsgespräch nicht von sich aus angeben werden, dass sie ihr Depot bei einem Konkurrenzinstitut führen, wäre es für den Berater vorteilhaft, wenn ihm eine a priori-Einschätzung der Affinität des Kunden zum Wertpapierbereich verfügbar wäre, insbesondere wenn dabei ein hohes Depotvolumen zu erwarten ist.

Im ersten Halbjahr 2001 entfiel auf 20% der Depot-Kunden durchschnittlich rund 83% des gesamten Depotvolumens im Privatkundengeschäft der Dresdner Bank. Dies gab Anlass zur Suche nach einem Instrument, mit Hilfe dessen Kunden mit hohem Topsegment-Potential identifiziert werden können, die aktuell kein Depot bei der Dresdner Bank führen.

3.2.3 Beschreibung der verwendeten Datensätze

Bevor auf die Detailfragen der Datenauswahl eingegangen wird, ist zunächst ein Zeitrahmen abzugrenzen, innerhalb dessen alle Daten zumindest näherungsweise Gültigkeit besitzen sollten. Bei der Wahl dieses Zeitrahmens ist auf eine gewisse Konstanz der Zielgröße Depotwert zu achten, die angibt, ob ein Kunde zum Top-Segment zählt. Offensichtlich wird die Zugehörigkeit zu einer der Klassen gerade bei einem Engagement in sehr volatilen Märkten starken Schwankungen unterlie-

gen. Die Zuordnung der Kunden zu den Klassen erfolgte daher aufgrund eines Durchschnitts über die jeweiligen Monatsultimos für das erste Halbjahr 2001, in dem im Gegensatz zum zweiten Halbjahr 2001 der Kursverlauf des DAX nur geringe Schwankungen aufwies und die Kurse am ersten und letzten Handelstag relativ dicht beieinander lagen.[9] Unter der Annahme, dass diese Entwicklung repräsentativ für die Wertentwicklung der Kundendepots ist und sich die überwiegende Zahl der Kunden demzufolge nicht zu ungewöhnlich hohen Neu- oder Desinvestitionen veranlasst sah, kann deshalb mit einer ausreichenden Stabilität der Zielgröße gerechnet werden.

Praktisch wird es nur in Ausnahmefällen möglich sein, Merkmale zu finden, die zu einer fehlerfreien Klassifikation führen. Es fehlt also das Wissen über die bestmöglich erreichbare Klassifikationsgüte. Damit lässt sich auch die Frage nach der Güte einer Merkmalsauswahl nicht absolut beantworten, sondern es ist zunächst eine fachliche Entscheidung darüber zu treffen, welche verfügbaren Merkmale Relevanz besitzen könnten. Nur offensichtlich redundante oder irrelevante Merkmale sowie solche, deren Erhebung aus zeitlichen, ökonomischen oder anderen Gründen nicht zweckmäßig erscheint, sollten vom Analyseprozess ausgeschlossen werden. Im Zuge der Datenvorverarbeitung bietet sich dann noch Gelegenheit, mittels mathematisch-statistischer Verfahren weitere nicht benötigte Merkmale zu identifizieren und von der weiteren Betrachtung auszunehmen.

Merkmale können in verschiedenen Funktionen von Bedeutung für den Analyseprozess sein:

- Das *Klassenmerkmal* ist nur für einen Teil der Objekte bekannt. Die Aufgabe der Klassifikation besteht darin, dieses Wissen in Verbindung mit den Segmentierungsmerkmalen für eine möglichst gute Schätzung der unbekannten Merkmalswerte zu verwenden. Es ist genau ein Klassenmerkmal festzulegen.
- Die Objekte werden durch eine Reihe von Merkmalen repräsentiert, anhand derer ein Klassifikator auf das unbekannte Klassenmerkmal schließen soll. Es werden nominale und kategoriale *Segmentierungsmerkmale* unterschieden.
- Häufig soll nur ein Kundensegment untersucht werden, das sich durch bestimmte Merkmalsausprägungen auszeichnet. Um die Menge aller in einer Tabelle erfassten Kunden einzuschränken, werden nur diejenigen Kunden ausgewählt, bei denen die *Filterattribute* bestimmten Eigenschaften genügen. Auch Segmentierungsmerkmale können als Filterattribute verwendet werden.
- Meist werden Daten aus verschiedenen Quellen verwendet, die durch *Schlüsselattribute* verknüpft werden. Im Falle eines Primärschlüssels eignet sich dieses Attribut bei einigen Verfahren zur späteren Identifikation von Fällen.

[9] Der DAX fiel von 6289,82 Punkten am 02.01.01 auf 6058,38 Punkte am 29.06.01 [Dr02].

Die Durchführung von Klassifikationsverfahren erfordert die Überführung der Daten in eine zweidimensionale Tabelle, wobei meist jeweils ein Kunde durch eine Zeile und ein Merkmal durch eine Spalte repräsentiert wird. Eine solche Projektion kann mit Informationsverlust verbunden sein und erfordert eine fachliche Entscheidung, wenn verschiedene Attribute in eine Zelle zusammenzuführen sind (Aggregationsvorschrift). Verfügt ein Kunde beispielsweise über zwei Sparkonten und ein Kontokorrentkonto mit positivem Saldo, die alle in die Spalte Einlagen einzutragen sind, so würde die Funktion *Summation* eine geeignete Möglichkeit der Zusammenführung bieten. Problematisch wird es etwa, wenn eine Aggregation über ein Feld *Region* erfolgen soll und die zwei Konten eines Kunden in Zweigstellen unterschiedlicher Regionen geführt werden. Eine fachliche Entscheidung wäre hier beispielsweise die Verwendung der Region, in der der Kunde seinen Hauptbetreuer hat.

Die Analysetabelle enthält üblicherweise das Klassenmerkmal sowie die ausgewählten Segmentierungsmerkmale auf Kundenebene. Werden häufig ähnliche Klassifikationsprobleme betrachtet, so kann ein automatisierter Aufbau einer solchen zentralen Kundensicht den Analyseprozess deutlich beschleunigen. Dabei bietet sich die Beibehaltung von Filterkriterien an, die einfache Einschränkungen auf die für eine konkrete Aufgabenstellung benötigten Informationen zulassen.

Um die Robustheit der Verfahren bei unterschiedlich verrauschten Daten zu testen, erfolgte eine Unterteilung der Segmentierungsmerkmale in drei Datenpools:

- Gruppe *DP1* umfasst eine größere Anzahl an zugekauften sozio-demographischen Daten, wobei sich die Angaben nicht auf eine Person selbst, sondern auf ihr Wohnumfeld beziehen.
- Gruppe *DP2* beinhaltet solche bankinternen Daten, die auch für zahlreiche Nicht-Kunden ermittelbar sind (Alter, Geschlecht, Berufsgruppe etc.)
- Gruppe *DP4* enthält bankinterne Daten, welche nur im Rahmen einer Kunde-Bank-Beziehung nutzbar sind (Produktnutzungs- und Umsatzinformationen, Kundenbindungsdauer etc.)

Die verwendeten Datensätze ergeben sich nun gemäß Tabelle 2 als Kombinationen dieser Datenpools[10], wobei ein der Datensatznummer vorangestelltes „d“ die vorherige Diskretisierung mit dem Verfahren „FUSINTER“ kennzeichnet und der Datensatz „df“ Ergebnis der Anwendung des nur für diskretisierte Merkmale einsetzbaren Attributauswahlverfahrens „MIFS“ ist.[11]

---

[10] Die Datensatznummer entspricht der Summe der Nummern der beteiligten Datenpools.

[11] Für die Anwendung der Verfahren „FUSINTER“ und „MIFS“ wurde jeweils die Implementierung im Softwarepaket Sipina 3.0 verwendet.

Datensatz	Zusammensetzung	Anteil an allen Attributen	Vermuteter Informations-gehalt[12]	Vermutetes Rauschen[13]
DS1/DSd1	DP1	87 %	--	++
DS3/DSd3	DP1∪DP2	93 %	+	++
DS6/DSd6	DP2∪DP4	13 %	++	-
DS7/DSd7	DP1∪DP2∪DP4	100 %	++	++
DSdf	MIFS(DP1∪DP2∪DP4)	20 %	++	--

Tabelle 2: Charakterisierung der verwendeten Datensätze

3.2.4 Vorbereitungen der Analyse

Zur Anwendung der Klassifikationsverfahren müssen die ausgewählten Daten mittels der folgenden Schritte noch in eine geeignete Form gebracht werden:

- Die *Datenintegration* überführt die oft noch in verschiedenen Tabellen vorliegenden Daten in eine Tabelle, in der jeder Kunde durch eine Zeile und jedes Merkmal durch eine Spalte repräsentiert wird. Dabei sind eventuell unterschiedliche Aggregationsniveaus und Skalierungen zu beachten.
- Die Aufgabe der *Datenbereinigung* besteht darin, fehlende und fehlerhafte Daten zu erkennen und gegebenenfalls zu behandeln. Allerdings sollte bei fehlenden Daten nach Möglichkeit die Ursache für das Fehlen geklärt werden, zumal dieses oft nicht zufällig auftritt. So beinhaltet etwa das Feld *Kreditvolumen* implizit die Information, ob jemand einen Kredit in Anspruch nimmt. Eine automatisierte Ersetzung von fehlenden Werten, beispielsweise durch den Mittelwert würde diese Information zerstören und das Ergebnis verfälschen. Bei sehr dünn besetzten Merkmalen sollte ein Ausschluss erwogen werden. Im verwendeten Datensatz wurden fehlende Werte fallabhängig durch den Mittelwert geschätzt oder als *missing* gekennzeichnet.
- Verfahren zur *Datenreduktion* können an zwei Stellen ansetzen: Zum einen ist die Verwendung einer Stichprobe möglich. Hier wurden drei

12 Aufgrund der Korrelationen der Attribute zur Zielvariable geschätzer Anteil am gesamten Informationsgehalt.

13 Aufgrund der Korrelationen der Attribute zur Zielvariable geschätzer Anteil am gesamten Rauschen.

Stichproben[14] mit jeweils 10.000 Datensätzen erhoben, wobei eine geschichtete Auswahl mit gleichen Klassenhäufigkeiten vorgenommen wurde. Auf der anderen Seite kann die Attributmenge verkleinert werden, wodurch oft zusätzlich Störeinflüsse vermieden werden. Datensatz DSdf beruht auf dem Merkmalsauswahlverfahren *Mutual Information Feature Selection* (MIFS). Neben der Elimination von Attributen lassen sich Attributgruppen durch ein aggregiertes Attribut ersetzen. Beim verwendeten Datensatz erfolgte die Bildung des arithmetischen Mittels über die einzelnen Monatsultimobestände für verschiedene Volumensarten.

- Allgemein hängen Anwendbarkeit, Güte und Geschwindigkeit von Klassifikationsverfahren vom Skalenniveau der Daten ab. Alle hier verwendeten Verfahren können mit numerischen Daten arbeiten. Allerdings führte die Arbeit mit diskretisierten Datensätzen zu deutlich schnelleren Laufzeiten. Obwohl mit einer Diskretisierung ein Informationsverlust verbunden ist, können bei einigen Verfahren erst hierdurch nichtlineare Abhängigkeiten aufgespürt werden. Durch Transformationen jeweils einer kategorialen oder diskretisierten Variablen mit *k* Ausprägungen in *k* Indikatorvariablen[15] lassen sich alle Variablen in numerische Variablen umwandeln. Als überwachte Variante der Diskretisierung kam die *FUSINTER*-Methode [Zi+98] zur Anwendung (Datensätze DSd*).

3.2.5 Eingesetzte Verfahren und Software

Tabelle 3 gibt Auskunft über die verglichenen Methoden und die eingesetzte Software. Wir haben uns dabei auf Entscheidungsbaum- und Regressionsverfahren konzentriert. Das (naive) Bayes-Verfahren wurde als ein Repräsentant der klassischen Statistik herangezogen. Ein Meta-Lernverfahren kam ebenfalls zur Anwendung. Bei der Auswahl spielte auch die in Abschnitt 3.1 diskutierte Anforderung an die leichte Erklärbarkeit der Ergebnisse der Klassifikation, nicht unbedingt des Klassifikationsverfahrens, eine wichtige Rolle.

14 Trainingsmenge, Testmenge und zweite Trainingsmenge für Meta-Lernverfahren

15 Eine Indikatorvariable besitzt jeweils den Wert 1, falls eine Merkmalsausprägung auftritt und ansonsten den Wert 0.

Methode	Software	Bemerkung
Logistische Regression	SAS 8.1 (Proc Logistic)	schrittweise Regression
Binäre lineare Regression	SAS + ScoreXpert (ARA)	schrittweise Regression; Transformation aller Variablen in Indikatorvariablen
PLUS 1.0.3 Beta	PLUS	Entscheidungsbaum: Polytomous Logistic Regression Trees with Unbiased Splits [Li00]
QUEST 1.8.16	QUEST	Entscheidungsbaum: Quick Unbiased, Efficent Statistical Trees [LoSh97]
CART	Sipina 2.5[16]	Entscheidungsbaum: Classification and Regression Trees [Br+84], Gini-Rule und Twoing-Rule
C4.5	Sipina 2.5	Entscheidungsbaum [Qu93]
ChAID	Sipina 2.5	Entscheidungsbaum: Chisquared Automatic Interaction Detection [Ka80]
Meta-Verfahren	Excel, Sipina 2.5	Aufbereitung der Meta-Trainingsmenge in Excel, anschließend Anwendung des Verfahrens C4.5 [WiFr01, S. 282 ff.]
Naive Bayes	Sipina 3.0 beta	
FUSINTER	Sipina 3.0 beta	Diskretisierung [Zi+98]
MIFS	Sipina 3.0 beta	Attributauswahl

Tabelle 3: Eingesetzte Verfahren und Software

4 Ergebnisse und Schlussfolgerungen

Eine Gegenüberstellung der Klassifikationsgüten (Abb. 2) bestätigt den bereits nach der Charakterisierung der Datensätze (Tabelle 2) zu vermutenden relativ geringen Erklärungsgehalt von Datensatz DS1, der insbesondere bei den nicht-

[16] Eine funktionsfähige Implementierung der Verfahren CART, C4.5 und ChAID lag in Sipina 3.0 beta leider noch nicht vor.

diskretisierten Datensätzen[17] zu gleichmäßig schwachen Klassifikationsgüten führt. Die Hinzunahme von internen Kundendaten (DS3 und DS7) führt zu erheblich besseren Ergebnissen. Erst hier, insbesondere bei DS7, sind die Unterschiede der Klassifikationsgüten zwischen den Verfahren deutlich.[18] Bei einer Beschränkung auf die internen Daten (DS6) wird eine mangelnde Robustheit der Verfahren Naive-Bayes und PLUS offenbar, indem die Abnahme des Rauschens zu deutlich positiveren Ergebnissen führt. Bei der Diskretisierung stehen den schnelleren Laufzeiten leichte bis deutliche Verschlechterungen der Güte gegenüber. Die anschließende Merkmalsauswahl mittels *MIFS*, die ebenfalls mit einem Mehraufwand verbunden ist, bringt neben der verkürzten Laufzeit des Klassifikationsverfahrens nur beim Verfahren Naive Bayes einen spürbaren Vorteil, dem aber eine Verschlechterung beim Verfahren PLUS gegenübersteht. Auffällig ist der deutliche Leistungsabfall bei der logistischen Regression, bei der die als numerisch vorausgesetzen unabhängigen Variablen nach der Diskretisierung nur noch wenige Ausprägungen aufweisen, während für die Erklärung der Varianz aber nur die gleiche Anzahl an Gewichtungskoeffizienten zur Verfügung steht. Eine Aufspaltung der diskretisierten Variablen in Indikatorvariablen nebst Anwendung einer linearen Regression mit einer deutlich größeren Zahl an Gewichtungskoeffizienten verbessert die Ergebnisse wiederum deutlich (ScoreXpert).

Neben dem Vergleich der Klassifikationsgüten stellt die Betrachtung der ROC-Diagramme ein wichtiges Beurteilungsmittel dar, wobei hier die Diagramme der Datensätze DS1 und DS7 als besonders aussagekräftig erscheinen (Abb. 3). Während sich bei einem geringen Informationsstand die Ergebnisse ähneln, lässt sich bei genügend Informationen ein differenzierteres Bild mit relativ glatten Linien erkennen.

Insgesamt lässt sich auf eine Überlegenheit der Regressionsansätze auf verschiedenartigen Datenstrukturen schließen, wobei das Vorhandensein von Rauschen das Ergebnis offensichtlich nicht verschlechtert. Bei genügend Informationen zeigt das ROC-Diagramm außerdem eine gleichmäßige Güte. Das Regressionsverfahren ScoreXpert ermöglicht auch die Erstellung einer leicht verständlichen und nutzbaren Scorecard.

Beim Meta-Verfahren führt der erhebliche Mehraufwand zur Durchführung der Klassifikation nicht zu den erhofften Verbesserungen im Ergebnis.

---

17 Die Klassifiktationsgüten (Abb. 2) sind zur besseren Übersicht auf zwei Diagramme aufgeteilt, wobei das Obere die diskretisierten Datensätze und das Untere die nicht-diskretisierten Datensätze nebst dem nur diskretisiert generierbaren DS df darstellt.

18 Es lässt sich zeigen, dass das Schwankungsintervall bei einem Stichprobenumfang von 10.000 Datensätzen und einem Konfidenzniveau von α=0,1 abhängig vom Vorhersagewert bei ca. 1,5% liegt.

Bezüglich der Geschwindigkeit lieferten alle Verfahren akzeptable Ergebnisse, außer CART, das zur Erstellung eines Klassifikators auf Basis der 10.000 Trainingsdatensätze einige Stunden benötigte.

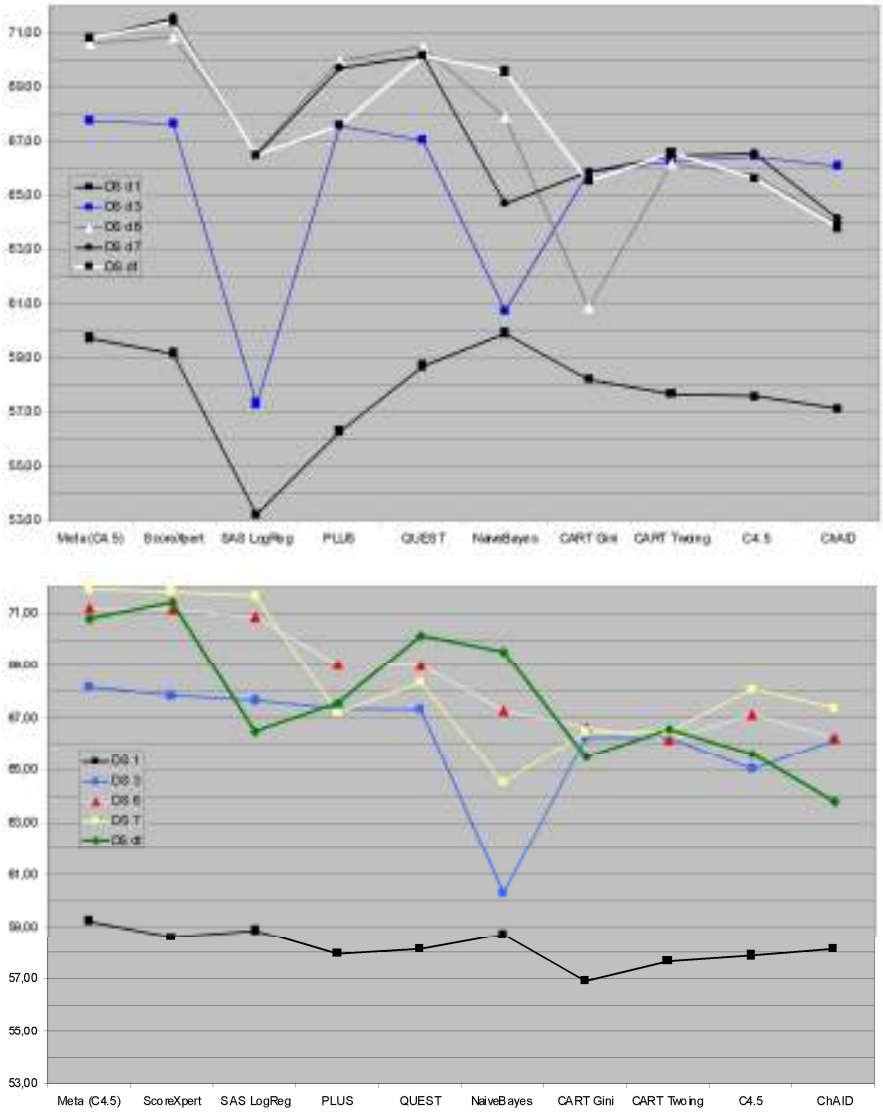

Abbildung 2: Übersicht der Vorhersagegüten der verwendeten Klassifikationsverfahren auf den verschiedenen Datensätzen

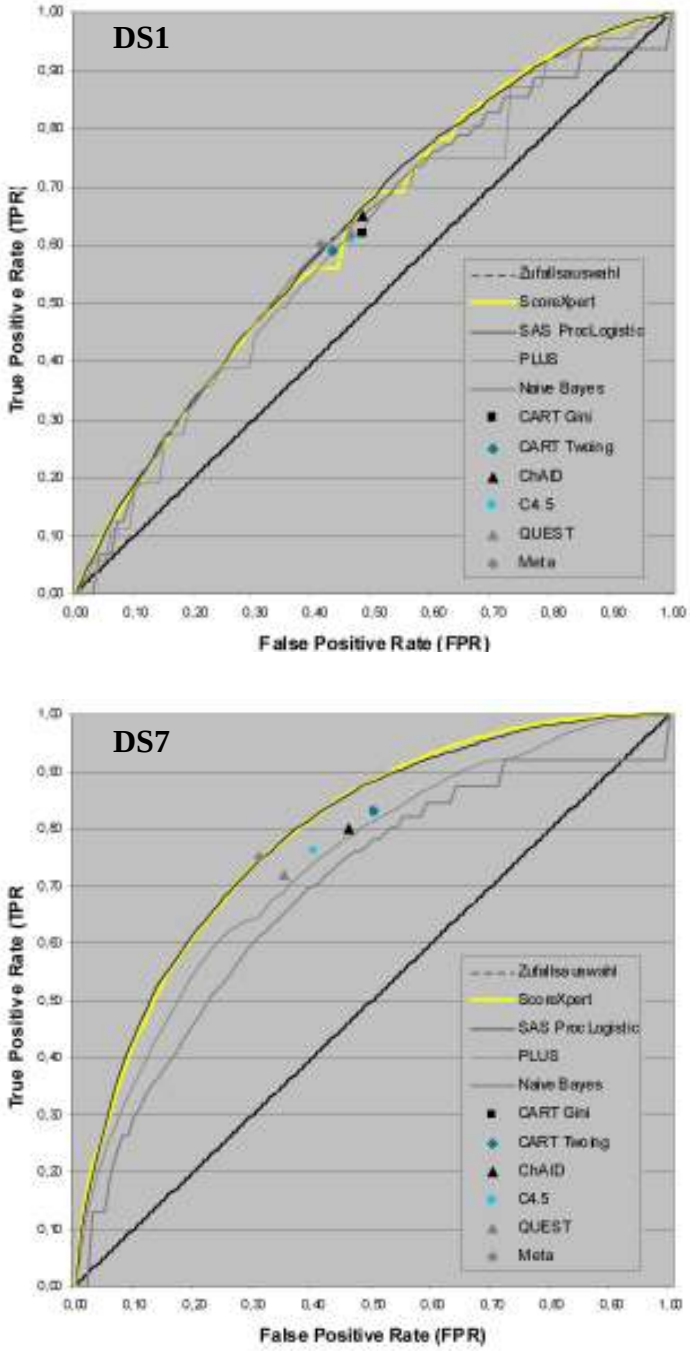

Abbildung 3: ROC-Diagramme für die Datensätze DS1 und DS7

Literatur

[AlNi00] Alpar, P.; Niedereichholz, J.: Einführung zu Data Mining. In: Alpar, P. und Niedereichholz, J. (Hrsg.): Data Mining im praktischen Einsatz. Vieweg, Braunschweig u.a., 2000, S. 1-27.

[BoKr98] Borgelt, C.; Kruse, R.: Attributauswahlmaße für die Induktion von Entscheidungsbäumen: Ein Überblick. In: Nakhaeizadeh, G. (Hrsg.): Data Mining – Theoretische Aspekte und Anwendungen. Physica, Heidelberg, 1998, S. 99-108.

[Br+84] Breiman, L. et al.: Classification and Regression Trees. Wadsworth International Group, Belmont, 1984. [Reprint bei: Chapman & Hall, Boca Raton u.a., 1998].

[Br96] Breiman, L.: Bagging Predictors. In: Machine Learning. 1996, 24(2), S. 123-140. ftp://ftp.stat.berkeley.edu/pub/users/breiman/bagging.ps.gz, 1996, Abruf am 2002-03-17.

[Dr02] Marktinformationssystem der Dresdner Bank. 2002. http://www.mis.dresdner-bank.de; Abruf: 02.05.2002.

[DrHo00a] Drummond, C.; Holte, R. C.: Explicitly Representing Expected Cost: An Alternative to ROC Representation. In: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2000, S. 198-207.

[DrHo00b] Drummond, C.; Holte, R. C.: Exploiting the Cost of (In)sensitivity of Decision Tree Splitting Criteria. In: Proceedings of the 17th International Conference on Machine Learning. Morgan Kaufmann, San Francisco, 2000, S. 239-246.

[EsSa00] Ester, M.; Sander, J.: Knowledge Discovery in Databases: Techniken und Anwendungen. Springer, Berlin u.a., 2000.

[FaHa96] Fahrmeir, L.; Hamerle, A.; Tutz, G.: Multivariate statistische Verfahren. 2. Aufl., de Gruyter, Berlin u.a., 1996.

[Fay+96] Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.: From Data Mining to Knowledge Discovery in Databases. In: AI Magazine, 1996, Vol. 17(3), S. 37-54.

[HaKa01] Han, J.; Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco u.a., 2001.

[Ka80] Kass, G.: An Exploratory Technique for Investigation Large Quantities of Categorical Data. In: Applied Statistics, 1980, Vol. 29(2), S. 119-127.

[Li00] Lim, T. S.: Polytomous Logistic Regression Trees. Dissertation, University of Wisconsin-Madison, 2000.]

[LoSh97] Loh, W. Y.; Shih, Y. S.: Split Selection Methods for Classification Trees. In: Statistica Sinica, 1997, Vol. 7, S. 815-840.

[Mi+94] Michie, D.; Spiegelhalter, D: J.; Taylor, C. C.: Machine Learning, Neural and Statistical Classification. Ellis Horwood, Ney York u.a., 1994.

[Na+98] Nakhaeizadeh, G.; Reinartz, T.; Wirth, R.: Wissensentdeckung in Datenbanken und Data Mining: Ein Überblick. In: Nakhaeizadeh, G. (Hrsg.): Data Mining – Theoretische Aspekte und Anwendungen. Physica, Heidelberg, 1998, S. 1-33.

[PrFa97] Provost, F.; Fawcett, T.: Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions. In: Proceedings of the Third International Conference on Knowledge Discovery and Data Mining. AAAI Press, Menlo Park, 1997, S. 43-48.

[Qu93] Quinlan, J. R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, 1993.

[Ro+94] Rohwer, R.; Wynne-Jones, M.; Wysotzki, F.: Neural Networks. In: Michie, D.; Spiegelhalter, D. J.; Taylor, C. C. (Hrsg.): Machine Learning, Neural and Statistical Classification. Ellis Horwood, 1994, S. 84-106.

[WiFr01] Witten, I. H.; Frank, E.: Data Mining: Praktische Werkzeuge und Techniken für das maschinelle Lernen. Carl Hanser, München u.a., 2001.

[Zi+98] Zighed, D.A., Rabaseda, S.; Rakotomalala, R.: FUSINTER: A Method for Discretization of Continuous Attributes. In: International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 1998, Vol. 6(3), S. 307-326.