Support Vektor Maschinen als Analyseinstrument im Marketing

Dissertation

zur Erlangung des akademischen Grades Dr. rer. pol.
der Fakultät für Wirtschaftswissenschaften
der Universität Bielefeld

Dipl.-Math. Katharina Monien

Bielefeld, im November 2005

1. Gutachter Prof. Dr. Reinhold Decker

2. Gutachter Prof. Dr. Joachim Frohn

Tag der mündlichen Prüfung: 27.4.2006

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Nomenklaturverzeichnis

Arabische Symbole:

a	Kernparameter
$a_i = (a_{i1}, \ldots, a_{iK})$	Codewort bei K Klassen
$A = (A_1, \ldots, A_L)$	Code der Länge L
$\boldsymbol{A}$	Matrix der Eingabedaten
$\boldsymbol{A'}$	Matrix der gewichteten Eingabedaten
$b,\ b^{[k]},\ b^{[k_1 k_2]}$	Biasterm
$\tilde{b}$	Kernparameter
$B = (B_1, \ldots, B_L)$	Code der Länge L
c_+, c_-, c_k	klassenbezogener Gewichtungsparameter
c_i	individueller Gewichtungsparameter
C	Kostenparameter
C_i	individueller Kostenparameter
$\boldsymbol{C}, \tilde{\boldsymbol{C}}$	Kovarianzmatrix
d	Kernparameter
$d(\cdot, \cdot)$	euklidischer Abstand
$d(\boldsymbol{x}; \boldsymbol{w}, b)$	Abstand eines Vektors $\boldsymbol{x}$ zur durch $\boldsymbol{w}$ und b bestimmten Ebene
$d_i^{\boldsymbol{x}}$	Teilabstand bei der Berechnung von $D(\cdot)$ mit $i = 1, \ldots, N$
$D(\cdot)$	Abstandsmaß
$D(\cdot, \cdot)$	Distanzmaß
$D_1,\ D_2,\ D_3$	Trennebenen
$D_{+/-}$	zu maximierende Spanne
$D^{[k]}, D^{[k_1 k_2]}$	zu bestimmende Hyperebene bei $K > 2$ Klassen
$\boldsymbol{e}_j$	j-ter Einheitsvektor
Err	zu erwartender Fehler bei der LOO-Methode
$f(\cdot),\ f_\beta(\cdot)$	Entscheidungsfunktion bei Biklassifikation
$f^{[k]}(\cdot)$	Entscheidungsfunktion bei OAA-Trennung der Klasse k
$f^{[k_1 k_2]}(\cdot)$	Entscheidungsfunktion zur Trennung der Klassen k_1 und k_2
$f_{a_i}(\cdot)$	Entscheidungsfunktion bei ECOC für Codewort a_i
$F(\cdot)$	Wert der Entscheidungsfunktion bei Biklassifikation
$F^{[k]}(\cdot)$	Wert der Entscheidungsfunktion bei OAA-Trennung der Klasse k
$F^{[k_1 k_2]}(\cdot)$	Wert der Entscheidungsfunktion bei Trennung der

	Klassen k_1 und k_2
$F'(\cdot)$	$= \lvert F(\cdot) \rvert$
FN	False Negative
FP	False Positive
$g(\cdot)$	integrierbare Funktion bei der Bedingung von *Mercer*
$\boldsymbol{g} = (g_1, \ldots, g_n)$	Gewichtsvektor bei der Gewichtung von n Merkmalen
$\boldsymbol{G}$	Gewichtsmatrix
h	VC-Dimension
H	Merkmalsraum
$H(\cdot, \cdot)$	Hamming-Distanz
$J(\cdot)$	Rayleigh-Koeffizient
k	Index der Klassen bei Multiklassifikation
$\tilde{k}$	Index der Kunden im Markovmodell
K	Anzahl vorliegender Klassen
$K(\cdot, \cdot)$	Kernfunktion
$\boldsymbol{K} = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{ij}$	Kernmatrix
l	Anzahl der Trainingsdaten
$l^{[k]}$	Anzahl der Trainingsdaten in Klasse k
l'	Anzahl der Trainingsdaten in einer Teilmenge bei OAO-Trennung
l_t	Anzahl an Testdaten
$l_t^{[k]}$	Anzahl an Testdaten in Klasse k
L	Anzahl an Codewörtern bei ECOC
$L(\cdot, \cdot; \cdot)$	Lagrange-Funktion
m	Index eines Eigenvektors bei KPCA
$m_k(\boldsymbol{x})$	Membership-Wert von Vektor $\boldsymbol{x}$ bei Fuzzy-SVM bzgl. Klasse k
$m_{kj}(\boldsymbol{x})$	Membership-Wert von Vektor $\boldsymbol{x}$ bei Fuzzy-SVM
$m'_k(\boldsymbol{x})$	angepasster Membership-Wert von Vektor $\boldsymbol{x}$
$\boldsymbol{m}_k$	Mittelwerte bei KFD
M	Minimum bei BERI-Einteilung
$\tilde{M}$	Margin
M_{ext}	Maß zur Beurteilung der Klassifikationsgüte
M_{med}	Maß zur Beurteilung der Klassifikationsgüte
$\boldsymbol{M}$	Matrix innerhalb der KFD
n, n'	Merkmalsanzahl
$\tilde{n}$	Index für die Anzahl der Käufe im Negativ-Binomial-Modell
N	Anzahl der Teilstücke bei der Berechnung von $D(\cdot)$
$\boldsymbol{N}$	Matrix innerhalb der KFD
NEG	Anzahl Beobachtungen in Klasse „-1“
$p_{\tau\tilde{k}}^{[K]}$	Wahrscheinlichkeit für einen Kauf in Periode τ von Kunde $\tilde{k}$
$p_{\tau\tilde{k}}^{[NK]}$	Wahrscheinlichkeit für einen Nichtkauf in Periode τ von Kunde $\tilde{k}$
$\boldsymbol{p}_1, \boldsymbol{p}_2$	Datenvektoren von Beobachtungen

$P(\cdot)$	Verteilungsfunktion
$P(\cdot)_T$	Wahrscheinlichkeit im Negativ-Binomial-Modell
POS	Anzahl Beobachtungen in Klasse „+1“
$Prec$	Genauigkeit bei der Multilabel-Klassifikation
q	Einteilung bei Kreuzvalidierung
r	Kernparameter
$\tilde{r}$	Anzahl Käufe in der Startperiode im Negativ-Binomial-Modell
$\boldsymbol{r} = (r_1, \ldots, r_n)$	Relevanz eines Merkmals i bei der Merkmalsreduktion
$\boldsymbol{r}^{[\tilde{s}]} = (r_1^{[\tilde{s}]}, \ldots, r_n^{[\tilde{s}]})$	Relevanz eines Merkmals i bei SVM $\tilde{s}$
R^2	Radius der kleinsten Sphäre
R_{emp}	empirisches Risiko
R_{erw}	erwartetes Risiko
s	Kernparameter
$\tilde{s}$	Index der SVM
$\boldsymbol{s}_i^{[\boldsymbol{x}]}$, $\boldsymbol{s}$	Hilfsvektor bei der Berechnung von $D(\cdot)$ mit $i = 1, \ldots, N$
$score(\cdot)$	Genauigkeitswert bei Multilabel-Klassifikation
$\boldsymbol{S}_B$	Zwischen-Klassen-Varianz
$\boldsymbol{S}_W$	Inner-Klassen-Varianz
Se	Sensitivität
Sp	Spezifität
S_{SVM}	Anzahl an SVM
t	Iteration im NLIRM-Algorithmus
T	maximale Anzahl an Käufen beim Negativ-Binomial-Modell
$T(l)$	Anzahl richtig klassifizierter Daten
TN	True Negative
TP	True Positive
TQ	Trefferquote bei Biklassifikation
TQ_k	Trefferquote in Klasse k
TQ_{alt}	Trefferquote innerhalb des NLIRM-Algorithmus
$\boldsymbol{U}^{[\tilde{k}]}$	Übergangsmatrix für Kunde $\tilde{k}$
$\boldsymbol{v}, \tilde{\boldsymbol{v}}$	Eigenvektor
$\boldsymbol{w}$	Normalenvektor bei der Biklassifikation
$\boldsymbol{w}^{[k]}$	Normalenvektor bei OAA-Trennung von Klasse k
$\boldsymbol{w}^{[k_1 k_2]}$	Normalenvektor bei Trennung der Klassen k_1 und k_2
$\boldsymbol{x}_i$	n-dimensionaler Eingabevektor
$\boldsymbol{x}_i^{[k]}$	Datenvektor i aus Klasse k
$\boldsymbol{x}_i^T$	transponierter Vektor $\boldsymbol{x}_i$
$\bar{\boldsymbol{x}}^{[1]}$, $\bar{\boldsymbol{x}}^{[2]}$	Mittelwertvektor der Klasse 1 bzw. 2
x_{ik}^{neu}	normierter Eintrag des Vektors $\boldsymbol{x}_i$ an Stelle k
y_i	Klassenzugehörigkeit mit $y_i \in \{1, -1\}$ oder $y_i \in \{1, \ldots, K\}$
$\boldsymbol{y}_i$	Klassenzugehörigkeitsvektor
Y	Anzahl an Käufen in der Folgeperiode im Negativ-Binomial-Modell

Griechische Symbole:

α_i	Lagrange-Multiplikator des Eingabevektors $\boldsymbol{x}_i$
α_{max}	maximal realisierter Wert der Lagrange-Multiplikatoren
$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_l)$	Vektor der Lagrange-Multiplikatoren
β	Parameter zur Kennzeichnung der zu bestimmenden Funktionenklasse
$\boldsymbol{\beta}$	Vektor bei KPCA
$\boldsymbol{\beta}^{[1]}, \ldots, \boldsymbol{\beta}^{[l]}$	Eigenvektoren bei KPCA
γ	Kernparameter (Radialbasis-Funktion)
$\gamma_1,\ \gamma_2$	Parameter des Negativ-Binomial-Modell
ϵ	Intervallbreite beim Gradientenverfahren
ζ	Konfidenzlevel
η	Hilfsvariable im NLIRM-Algorithmus
θ^+	Parameter zur Bestimmung von M_{ext}
θ^-	Parameter zur Bestimmung von M_{ext}
$\lambda_i(\boldsymbol{x})$	Winkel bei der Gradientenmethode an der Stelle $\boldsymbol{x}$
$\lambda, \lambda_1, \ldots, \lambda_l$	Eigenwert
μ^+	Parameter zur Bestimmung von M_{med}
μ_i^+	Parameter zur Berechnung von FCS
μ^-	Parameter zur Bestimmung von M_{med}
μ_i^-	Parameter zur Berechnung von FCS
ξ_i	Schlupfvariable bei der Biklassifikation
$\xi_i^{[k]}$	Schlupfvariable bei der OAA-Trennung von Klasse k
$\xi_i^{[k_1 k_2]}$	Schlupfvariable bei der Trennung der Klassen k_1 und k_2
σ	Kernparameter
$\sigma_i^{2(+)}$	Standardabweichung des Merkmals i in Klasse „+1“
$\sigma_i^{2(-)}$	Standardabweichung des Merkmals i in Klasse „-1“
τ	Kaufperiode im Markovmodell
$\Phi(\cdot)$	Abbildung
$\boldsymbol{\varphi}$	zu berechnender Vektor bei KFD
χ^2_{krit}	tabellierter χ^2-Wert

Mengen:

$\mathcal{B}_0, \mathcal{B}_1, \mathcal{B}_{i'}$	Basis von Vektoren für Online-Training (in Iterationsschritt $0, 1, i'$)
$\mathcal{E}_{i'}$	Menge der Fehler beim Online-Training in Iterationsschritt i'
$\mathcal{I}_\epsilon$	Indexmenge beim Gradientenverfahren
$\mathcal{M}$	Datenmenge bei Online-Training
$\mathbb{N}$	Menge der natürlichen Zahlen

$\mathcal{P}_i$	Menge prognostizierter Labels
$\mathcal{R}$	Merkmalsmenge im NLIRM-Algorithmus
$\mathbb{R}$	Menge der reellen Zahlen
$\mathbb{R}_+$	Menge der reellen Zahlen größer 0
$\mathcal{S}$	Merkmalsmenge im NLIRM-Algorithmus
$\mathcal{S}_{alt}$	Merkmalsmenge im NLIRM-Algorithmus
$\mathcal{SV}$	Menge der Indizes der Support Vektoren
$\mathcal{T}_1, \ldots, \mathcal{T}_t$	Datenmengen beim Online-Learning
$\mathcal{Y}_i$	Menge der wahren Klassenzugehörigkeiten

Abkürzungsverzeichnis

ANOVA	Analysis of Variance
AUC	Area under the Curve
BERI	Business Environment Risk Institute
bzgl.	bezüglich
bzw.	beziehungsweise
CHAID	Chi-square Automatic Interaction Detector
CRM	Customer Relationship Management
d.h.	das heißt
DAG	Directed Acyclic Graph
ECOC	Error-Correcting Output-Coding
etc.	et cetera
FCS	Fisher Criterion Score
GfK	Gesellschaft für Konsumforschung
HH	Haushalt
IRRM	Incremental Regularized Risk Minimization
KFD	Kernel-Fisher-Diskriminanzanalyse
KKT	Karush-Kuhn-Tucker
KPCA	Kernel Principal Component Analysis
LDA	Lineare Diskriminanzanalyse
LEH	Lebensmitteleinzelhandel
LIBSVM	Library for Support Vector Machines
LM	Lebensmittel
LOO	Leave-One-Out
MLP	Multi-Layer Perceptron
NB	Negativ-Binomial
NLIRM	Non Linear Incremental Risk Minimization
NLRFE	Non Linear Recursive Feature Elimination
OAA	One-Against-All
OAO	One-Against-One
OTC	Over the Counter
PCA	Principal Component Analysis
PDA	Personal Digital Assistant
Poly	Polynomiell
POS	Point of Sale
RBF	Radialbasis-Funktion

RFE	Recursive Feature Elimination
ROC	Receiver Operating Characteristics
SOM	Self Organizing Maps
SV	Support Vektor
SVM	Support Vektor Maschinen
TK	Tiefkühl
TQ	Trefferquote
u.a.	unter anderem
u.U.	unter Umständen
usw.	und so weiter
vgl.	vergleiche
WG	Warengruppe
z.B.	zum Beispiel
ZUMA	Zentrum für Umfragen, Methoden und Analysen

Kapitel 1

Einleitung

Das Verstehen von Zusammenhängen in Marktstrukturen bildet die Basis für ein erfolgreiches Zielgruppenmarketing. Für die kundenorientierte Ausrichtung von Marketingaktivitäten werden Informationen über die Kunden eines Unternehmens gesammelt, um Marketingentscheidungen zu erleichtern und somit zu einer effizienten Behandlung des Marktes zu gelangen. Die dafür benötigte Identifikation der gewünschten Zielgruppen kann mit Hilfe von Klassifikationsverfahren erreicht werden. Die Kundenklassifikation ermöglicht durch die Zuordnung von Kunden zu unterschiedlichen Kundengruppen somit einen Übergang von der Bedienung eines Massenmarktes zum individualisierten Marketing und wird in dieser Arbeit durch den Einsatz von Support Vektor Maschinen durchgeführt.

1.1 Motivation der Arbeit

In vielen Bereichen wurde das ursprüngliche Massenmarketing mit dem Ziel, ungeachtet der am Markt vorliegenden Strukturen und Kundenwünsche mit niedrigen Kosten und Preisen einen möglichst großen Marktanteil zu erreichen, durch das Database Marketing und damit der kundenorientierten Marketingausrichtung abgelöst (*Petrison et al.* (1997)). Dies wurde insbesondere durch die automatisierte Datenerhebung intensiviert. Die Kunden stehen im Mittelpunkt des unternehmerischen Handelns, und die Orientierung an den Wünschen der Kunden ist bei vielen Unternehmen Inhalt der Ausrichtung ihrer Aktivitäten. Durch den Einsatz einer Kundendatenbank ist es möglich, Informationen über Kunden zu sammeln, zukünftiges Verhalten zu prognostizieren und damit die Personalisierung jeder Interaktion mit dem Kunden zu erzielen (*Aaker et al.* (2004)). Aufgrund eines zunehmenden Kostendrucks, verbunden mit sinkender Loyalität und sinkender Bindungsbereitschaft auf Seiten der Kunden, erscheint die effiziente Ausgestaltung der Kundenbetreuung für die Unternehmen unabdingbar. Ziel bei der Behandlung eines Marktes sollte daher die Sicherung der Kundenzufriedenheit sein. Diese bildet einen Schlüsselfaktor bei der Bindung des Kunden an das Unternehmen und damit für den Unternehmenserfolg. Zufriedenheit von Seiten der Kunden ist dann

gegeben, wenn die Bedürfnisse eines Kunden erkannt und befriedigt werden, die Erwartungen des Einzelnen also erfüllt werden (*Hennig-Thurau, Hansen* (2001)). Nach *Diller* (1995) wird in einem angemessenen Umgang mit den Kunden, also der Erhöhung der Zufriedenheit der Kunden, ein strategischer Wettbewerbsvorteil gesehen. Dieser liegt u.a. in einem langfristig gesicherten Umsatz, hervorgerufen durch loyale Kunden, begründet. Weiterhin kann eine Reduzierung der Kosten bei der Konzentration auf loyale Kunden erfolgen, da nur diejenigen Kunden eine intensive Betreuung erfahren sollten, bei denen dies auch Erfolg versprechend ist. Ein weiterer wichtiger Punkt bei der Sicherung des Unternehmenserfolgs liegt in der Anwerbung von Neukunden. Da eine Gewinnung von Neukunden deutlich kostspieliger sein kann als die Sicherung von Stammkunden (*Kotler, Bliemel* (2001)), ist es umso wichtiger, dass auch hier gezielt vorgegangen wird und das Marketingbudget ebenfalls effektiv eingesetzt wird. Die adäquate Behandlung der Kunden kann somit mehr Kundennähe und mehr Kundenzufriedenheit und damit eine stärkere Kundenbindung schaffen (*Trommsdorff, Drüner* (2001)). Eine marktorientierte Denk- und Handlungsweise verbunden mit dem wirtschaftlich sinnvollen und effektiven Einsatz des Marketingbudgets kann somit durch eine stärkere Kundenbindung zu Wettbewerbsvorteilen führen.
Um diese Zielsetzungen angemessen umzusetzen, hat sich das Prinzip des Customer Relationship Managements (CRM) durchgesetzt. Ein Ziel des CRM liegt in der individuellen Ansprache der für ein Unternehmen wichtigen Kunden, um eine Ausweitung des Kreises der Stammkunden und eine engere Bindung der Kunden an das Unternehmen zu ermöglichen. Nach *Homburg, Sieben* (2005) umfasst CRM „die Planung, Durchführung, Kontrolle sowie Anpassung aller Unternehmensaktivitäten, die zu einer Erhöhung der Profitabilität der Kundenbeziehung und damit zu einer Optimierung des Kundenportfolios beitragen."
Um unterschiedliche Strukturen, beispielsweise im Kaufverhalten, aufzudecken, müssen die Kunden als Individuen mit eigenen Verhaltensweisen und Bedürfnissen begriffen werden. Bei der Differenzierung der Kunden geht es zum einen um die Selektion der für das Unternehmen attraktiven Kunden und zum anderen um deren adäquate Behandlung. So sollten, etwa im Rahmen des Vertriebs bestimmter Produkte, die zahlreichen potenziellen Kunden sorgfältig für geplante Vertriebsaktivitäten ausgewählt werden. Diese Aktivitäten könnten beispielsweise einen verstärkten Einsatz von Außendienstmitarbeitern zum Gewinn von Neukunden umfassen. Die Identifikation der für das Unternehmen interessanten und damit zu bindenden Kunden bildet ein Kernproblem für den wirksamen Einsatz finanzieller Mittel.
Analoge Probleme der Identifikation relevanter Kunden ergeben sich ebenso innerhalb des Direktmarketings. Nach *Link* (2001) wird durch Direktmarketing die Möglichkeit zur Erstellung eines persönlichen und bedarfsgerechten Informations- und Leistungsangebots und dem Dialog mit dem Kunden geboten. Somit steht der Kunde als Individuum im Mittelpunkt der Bemühungen. Durch die gezielte Ansprache von bestimmten Personengruppen soll eine hohe Rücklaufquote und langfristig die Bindung eines Kunden an das Unternehmen erreicht werden. Nach *Holland* (2004) entwickelt sich aber der Konsument aufgrund massenhaft verbreiteter, sich

häufig ähnelnder Kommunikationsmaßnahmen zu einem Informationsverweigerer. Daher erscheint eine individuelle Ansprache der Zielgruppe unverzichtbar, um die Aufmerksamkeit der Kunden zu gewinnen.
Die übereinstimmende Zielsetzung dieser Bereiche ist die Realisierung einer möglichst hohen Erfolgsquote der kundenorientierten Ausgestaltung der Marketingaktivitäten bei möglichst geringem finanziellen und personellen Einsatz. Die Klassifikation von Kunden als relevant oder nicht relevant für das Unternehmen spielt dabei eine wichtige Rolle.

Erst die systematische Sammlung relevanter Daten ermöglicht die Anwendung von Verfahren zur Gewinnung der für die individuelle und gewinnbringende Ausgestaltung der Kundenbeziehungen notwendigen Informationen. Bei der Interaktion eines Unternehmens mit seinen Kunden fallen in der Regel eine Vielzahl von kundenindividuellen Informationen an, die dazu genutzt werden können, die oben genannten Ziele zu verwirklichen. Aus Kundenbindungsprogrammen gewonnene oder extern vorliegende Kundendaten bilden eine wichtige Grundlage bei dieser zielgruppenorientierten Ausrichtung. So kann beispielsweise durch den Einsatz von Kundenkarten das Kaufverhalten der Kunden bei dem eigenen Unternehmen abgebildet werden. Die Kundeninformationen können verschiedener Art und unterschiedlichen Umfangs sein. Neben soziodemografischen Merkmalen oder Informationen über das bisherige Kaufverhalten können beispielsweise auch Lifestyledaten aus Umfragen zum Einsatz kommen. Im Business-to-Business Bereich wäre die Identifikation potenzieller Neukunden zur Erweiterung des bisherigen Außendienstbereiches auf Basis objektiv ermittelbarer Merkmale aus Firmendatenbanken (Grundkapital oder Anzahl Mitarbeiter) oder subjektiv von Außendienstmitarbeitern einzuschätzenden Merkmalen, wie etwa die Möglichkeit der Ausweitung von späteren Geschäftsbeziehungen, denkbar. Je nach Branche können weitere Datenquellen genutzt werden. So können Pharmaunternehmen beispielsweise auf den Einsatz von eDetailing zurückgreifen. Dabei handelt es sich um eine neue Möglichkeit der Vermittlung von Informationen an Ärzte über das Internet als Ergänzung der traditionellen Besuche des Außendienstes (vgl. *Baier et al.* (2004)), was durch interaktive Komponenten als zusätzliche Informationsquelle für Pharmaunternehmen genutzt werden kann. Je nach Zielsetzung fällt daher die Merkmalsauswahl zur Beschreibung der Beobachtungen unterschiedlich aus.

Am Anfang der Marktbearbeitung steht somit die Frage nach den Bedürfnissen und Wünschen der Kunden. Die Prognose des Kaufverhaltens, der Bestellmenge oder Ähnlichem und die Auswertung der für die jeweiligen Zielsetzungen zur Verfügung stehenden Daten hinsichtlich der Bestimmung kundenindividueller Marketingstrategien sind allerdings häufig mit herkömmlichen Datenbankabfragen oder statistischen Verfahren nicht möglich. Hier kommen Instrumente des Data Mining, insbesondere aus dem Bereich der Klassifikation, zum Einsatz. Neben der Regression oder Segmentierung ist die Klassifikation eine der wichtigen Funktionen des Data Mining (*Fayyad et al.* (1996)). Der Einsatz von Klassifikationsverfahren basiert auf einer vorangegangenen Einteilung der Datenbasis in mehrere, in

sich homogene Gruppen. Diese Festlegung der Gruppen kann durch Segmentierungsverfahren durchgeführt werden, oder sie ergibt sich automatisch aus dem jeweiligen Anwendungskontext. Darauf aufbauend können nun Klassifikationsinstrumente zum Einsatz kommen, die auf a priori definierte Gruppen bzw. Klassen angewiesen sind. Bei der Auswertung kundenrelevanter Daten werden die aus bisherigen Kundenbeziehungen gewonnenen Informationen dazu verwendet, in dem Kundenstamm homogene Gruppen zu finden, die der jeweils vorliegenden Zielsetzung gerecht werden. Durch Klassifikationsverfahren können so genannte Klassifikationsfunktionen bestimmt werden, mit deren Hilfe neue Objekte den a priori definierten Gruppen zugeordnet werden können. Etwaige Neukunden können so auf Basis bisher vorliegender Kundendaten (oder allgemeiner gesprochen: Merkmalsvektoren) den zuvor definierten Klassen zugewiesen werden und eine ihrer Klassenzugehörigkeit entsprechende Behandlung erfahren. Somit wird eine gezielte Ansprache der relevanten Zielgruppen ermöglicht. Durch den Einsatz von Klassifikationsverfahren könnten die zu kontaktierenden Kunden identifiziert werden, um mit einem Neuprodukt Erfolg zu haben. Weiterhin können durch den Einsatz entsprechender Klassifikationsverfahren beispielsweise responsewillige Kunden einer Direktmarketingaktion bestimmt werden.

Um möglichst optimale Ergebnisse bei der Kundenklassifikation zu erzielen, sind leistungsstarke Verfahren, die den jeweils verfolgten Zielen gerecht werden, unerlässlich. Ein klassisches Verfahren im Bereich des Marketings bildet die Diskriminanzanalyse. Dabei wird sowohl der diagnostische Ansatz zur Ermittlung vorliegender Unterschiede zwischen den zu untersuchenden Gruppen, als auch der prognostische Ansatz zur Prognose von Klassenzugehörigkeiten genutzt. Weiterhin erhalten Methoden des Data Mining, wie neuronale Netze oder Entscheidungsbäume, zunehmende Relevanz. Eine in jüngster Zeit ebenfalls mehr an Bedeutung gewinnende Methodenklasse sind die Support Vektor Maschinen (SVM). Diese aus dem Umfeld des maschinellen Lernens stammende Verfahrensklasse bildet bereits in einigen Bereichen ein leistungsstarkes Instrument zur Klassifikation.
Maschinelles Lernen befasst sich „mit der computergestützten Modellierung und Realisierung von Lernphänomenen“ (*Wrobel et al.* (2000)). Bei SVM handelt es sich um ein überwachtes Lernverfahren, welches von *Vapnik* (*Boser et al.* (1992)) Anfang der 90er Jahre entwickelt wurde. SVM generieren auch für besonders große Datensätze schnelle und zuverlässige Klassifikationsergebnisse. In der Standardliteratur (vgl. z.B. *Schölkopf, Smola* (2002)) werden SVM als eine der nennenswerten Neuerungen im Bereich der Klassifikation in den letzten 10 Jahren genannt, was mit den herausragenden Ergebnissen in den sich herauskristallisierten Stammanwendungsgebieten dieses Verfahrens begründet wird. Zu diesen zählen u.a. unterschiedliche Formen von Bilderkennungsanwendungen, bei denen Eigenschaften der jeweiligen Bilder als Merkmale in die Klassifikation einfließen. Die Analyse medizinischer Daten zur Krebsdiagnostik (*Guyon et al.* (2002)) und Handschriftenerkennung (z.B. *Schölkopf* (1997), *Bahlmann et al.* (2002)) gehören mittlerweile ebenfalls zu den klassischen Anwendungsgebieten von SVM, bei denen insbesondere die Möglichkeit der Behandlung hochdimensionaler Daten ausgenutzt

wird. Auch im Bereich der Bioinformatik finden SVM, etwa zur Genklassifikation mittels Mikroarrays (z.B. *Furey et al.* (2000)), zunehmend Verwendung. Ein weiteres wichtiges Gebiet ist die Textklassifikation, die in (*Joachims* (2002)) vertieft wird. Mittlerweile sind SVM zu einem anerkannten Klassifikationsinstrument in mehreren Bereichen avanciert und können in den Stammanwendungsgebieten als „state-of-the-art"-Verfahren zur Klassifikation bezeichnet werden (*Guyon et al.* (2002)). SVM bieten neben der linearen Trennung auch die Möglichkeit, durch den Einsatz so genannter Kernfunktionen nicht lineare Strukturen innerhalb von Daten zu entdecken und könnten somit die für das Marketing relevanten und geeigneten Werkzeuge sinnvoll ergänzen. SVM bilden ebenfalls die Grundlage für so genannte kernbasierte Lernalgorithmen (vgl. *Schölkopf, Smola* (2002)). Ziel von SVM ist die Trennung a priori vorgegebener Klassen mittels einer im allgemeinen nicht linearen (Hyper-)Ebene, um so die Klassenzugehörigkeit neuer Objekte prognostizieren zu können. Der Einsatz von Kernfunktionen ermöglicht eine flexible Anpassung an die zugrunde liegenden Daten. Auf Basis der daraus resultierenden Entscheidungsfunktionen können Objekte mit unbekannter Gruppenzugehörigkeit den getrennten Klassen zugeordnet werden. Dabei wird die Trennung sowohl durch Bi- als auch durch Multiklassifikation ermöglicht (vgl. *Hsu, Lin* (2002)).
SVM werden in anderen Bereichen bereits seit längerer Zeit erfolgreich eingesetzt. Im Umfeld des Marketing ist diese Methode noch weitestgehend unbekannt, was sich in einer sehr geringen Anzahl an Publikationen zu diesem Thema widerspiegelt. Erst jetzt befassen sich auch zwei Artikel in einer der renommierten Zeitschriften im Marketing mit diesem Verfahren. Während in *Cui, Curry* (2005) SVM als Alternative zur Prognose im Marketing vorgestellt und mit der Leistungsfähigkeit des Multinomial-Logit-Modells verglichen werden, werden SVM in *Evgeniou et al.* (2005) bereits zur Modifizierung traditioneller Verfahren (hier der Conjoint-Analyse) eingesetzt. Dies zeigt die Aktualität dieser Thematik auch im Marketing und die Notwendigkeit der intensiven Auseinandersetzung mit dieser Verfahrensklasse. In dieser Arbeit werden daher SVM als mögliche neue Methode im Marketing diskutiert und ihre Vorteile derart genutzt, dass die Identifikation der für die jeweils zu verfolgenden Marketingstrategien relevanten Kunden ermöglicht wird. Das Verfahren wird hierzu den jeweiligen Anforderungen bei der Kundenklassifikation angepasst, um ein leistungsstarkes Instrument zu erhalten.

1.2 Angestrebtes Erkenntnisziel

In vielen Bereichen haben sich SVM gegenüber alternativen Verfahren bewährt und stellen ein leistungsfähiges Instrument zur Klassifikation dar. Die steigende Anzahl an Publikationen, Anwendungsbereichen und Beiträgen auf wissenschaftlichen Konferenzen verschiedenster Fachgebiete rechtfertigt daher einen genaueren Blick auf SVM hinsichtlich ihres Einsatzes im Marketing. Auch in diesem Bereich ist eine hohe Präzision bei der Vorhersage des Kaufverhaltens oder der Bedürfnisse von Kunden wünschenswert und notwendig, um die Marketingaktivitäten besser auf die individuellen Interessen der Kunden einzustellen und somit kundenorientiert und Ressourcen schonend handeln zu können.
Da es sich bei SVM um eine noch recht junge Methode zur Klassifikation handelt, besteht ein erstes Ziel der Arbeit in der Einführung des Verfahrens in den Bereich des Marketings. Neben der Vorstellung der Methodik bildet die Adaption des Instrumentariums an bestehende Probleme im Marketing ein wichtiges Element. Die Leistungsstärke von SVM zeigt sich überwiegend an der sehr guten Prognosegüte im Vergleich zu anderen Verfahren. Die Zielsetzungen und Anforderungen an ein Klassifikationsinstrument im betriebswirtschaftlichen Kontext gehen jedoch über die Betrachtung der Genauigkeit der Prognosen, die dieses Instrument liefert, hinaus. Einen entscheidenden Aspekt bei der Beurteilung eines Analyseinstrumentes im Marketing bildet die Anwenderfreundlichkeit. Für einen erfolgreichen Einsatz in der Praxis muss die Methodik leicht zu handhaben und bedienerfreundlich gestaltet sein, sodass keine großen Zeitverluste in der Anwendung zu verzeichnen sind. Dieser Aspekt betrifft insbesondere die Wahl möglicher Parameter, die vom Benutzer festzulegen sind. Diese Einstellungen erfordern vom Benutzer ein zum Teil tief gehendes Verständnis der Arbeitsweise von SVM. Insbesondere bei der nicht linearen Trennung treten Schwierigkeiten auf, da auch kernabhängige Parameter festzulegen sind. Ein Teilbereich befasst sich daher mit der Auswahl guter Parameter.
Weiterhin wird eine Anpassung der Basismethodik an spezielle marketingspezifische Anforderungen berücksichtigt. Dies umfasst u.a. die unterschiedlichen Gewichtungen verschiedener Kundenklassen und Merkmale. Diese Gewichtungen können sich beispielsweise aus der Bedeutung der Kunden für das Unternehmen ergeben. Somit können die Ergebnisse durch die Anpassung der Methodik in gewissem Maße in eine gewünschte Richtung gelenkt werden.
Neben der Eröffnung eines Anwenderzugangs ist die Darstellung der Ergebnisnutzung ein Kernelement der Arbeit. Dabei findet insbesondere die Interpretation und die Weiterverwendung der ermittelten Ergebnisse Berücksichtigung. Die bisher zur Verfügung stehenden Informationen, die mittels SVM generiert werden, sind aus Marketingsicht nicht zufrieden stellend, da sie neben der Ermittlung von Trefferquoten keine weiteren Aussagen generieren. Daher werden die Ergebnisse von SVM in dieser Arbeit intensiv betrachtet. Eine unterschiedliche Intensität in der Zuweisung der Beobachtungen zu einzelnen Klassen ermöglicht eine weitere Differenzierung der Kundenansprache und trägt somit zur weiteren Kundenorientierung bei. Die Verwendung von mehreren Klassen wird dabei ebenfalls berücksichtigt.

Somit liefert diese Art der Ergebnisnutzung in dieser Arbeit einen neuen, wichtigen Ansatz im Marketing, der zur adäquaten Ansprache der Kunden beitragen kann. Die Entdeckung diskriminatorisch relevanter Merkmale, die die Beobachtungen beschreiben, bildet hier ein weiteres wichtiges Element der anwendungsorientierten Betrachtung eines Klassifikationsinstruments. Ein Ziel liegt in der Reduktion der relevanten Merkmale zur Reduzierung des Kostenaufwands bei der Generierung der notwendigen Information. Es werden unterschiedliche Verfahren zur Erreichung dieses Ziels vorgestellt und den Anforderungen entsprechend erweitert und ergänzt. Es stellt sich heraus, dass die bereits existierenden und die in dieser Arbeit vorgenommenen Erweiterungen bestehender Algorithmen leistungsstarke Alternativen zu bisherigen Methoden der Merkmalsreduktion bilden.
Die gesamtheitliche Betrachtung von SVM unter Berücksichtigung vieler einzelner Facetten dieser Methodik und insbesondere die Adaption an marketingspezifische Anforderungen bilden einen neuen Beitrag zur Klassifikation im Marketing.

Die Arbeit bildet den Transfer zwischen dem methodisch geprägten Bereich des maschinellen Lernens und der praxisbezogenen Anwendungsseite. Der Einsatz von SVM dient in dieser Arbeit dazu, Wissen über den Markt bzw. über die Kunden zu generieren, um Marketingentscheidungen zu unterstützen und zu vereinfachen. Es wird die Anpassung und Erweiterung der SVM als Analyseinstrument im Marketing vorgenommen und die Weiterentwicklung zu einem anwenderfreundlichen Tool fokussiert. Dabei werden die methodischen Grundlagen auf marketingspezifische Probleme und Anforderungen übertragen. Durch die Betrachtung und Erweiterung wichtiger Aspekte, die die SVM charakterisieren, werden die Unterschiede zu traditionellen, alternativ einzusetzenden Klassifikationsverfahren herausgearbeitet. Es stellt sich heraus, dass SVM über eine Reihe von Vorteilen gegenüber herkömmlichen Verfahren verfügt, sodass diese Methode das bisherige Instrumentarium sinnvoll ergänzt. Die Potenziale von SVM im Rahmen des Marketings werden in dieser Arbeit identifiziert und die Leistungsfähigkeit der Methodik anhand von empirischen Beispielen überprüft.

Die Arbeit ist wie folgt gegliedert: Zunächst werden in Kapitel 2 die grundlegenden Ansätze zur Klassifikation mittels SVM vorgestellt. Da neben der Betrachtung von zwei a priori definierten Klassen insbesondere bei der Kundenklassifikation mehrere Gruppen zum Einsatz kommen können, wird zusätzlich auf die Möglichkeiten zur Multiklassifikation mittels SVM eingegangen.
Anwendungsorientierte Aspekte finden in Kapitel 3 Berücksichtigung. Um diesem Gesichtspunkt im Rahmen des betriebswirtschaftlichen Einsatzes Rechnung zu tragen, werden dabei die Möglichkeiten zur Einbindung von a priori-Wissen vorgestellt. Ein weiterer Fokus liegt auf der Beurteilung und Nutzung der generierten Ergebnisse. Dabei wird sowohl auf die lineare als auch auf die nicht lineare Trennung eingegangen und die hier vorgeschlagene Vorgehensweise auf die Multiklassifikation erweitert.
In Kapitel 4 wird anhand empirischer Beispiele die Leistungsfähigkeit der SVM im Bereich des Marketing untersucht. Bei der Anwendung der SVM auf reale Daten, die

aus verschiedenen Bereichen des Marketings entstammen, werden unterschiedliche Situationen der Klassifikation abgebildet und mittels SVM gelöst. Neben der grundlegenden Methodik finden die im Kapitel zuvor vorgeschlagenen Erweiterungen Verwendung. Neben der Zuordnung von Kunden zu unterschiedlichen Klassen sind auch weitere marketingbezogene Einsatzbereiche für die Klassifikation denkbar. So kann beispielsweise eine Klassifikation verschiedener Konzepte zur Vermarktung von Neuprodukten eingesetzt werden. Innerhalb dieser Arbeit soll der Fokus allerdings auf der Kundenklassifikation liegen. Das langfristige Ziel liegt darin, SVM als leistungsstarke Alternative oder Ergänzung zu traditionellen Verfahren zu etablieren.
Als Alternative zur Auswertung empirischer Daten könnten zur Einschätzung der Qualität von SVM simulierte Daten im Rahmen kontrollierter Experimente herangezogen werden. Dadurch kann die Wirkung bestimmter Strukturcharakteristika der Daten auf das Verhalten des Verfahrens untersucht werden. In dieser Arbeit liegt hingegen der Fokus auf der Auswertung realer Daten. Dadurch liegen Daten vor, die bei Klassifikationsanwendungen in Unternehmen auch zur Verfügung stehen und die dort gegebene Situation besser abbilden, so dass die Vor- und Nachteile bei dieser Art der Anwendung angemessen aufgedeckt werden können.
Die Arbeit wird mit einer zusammenfassenden Betrachtung der gewonnenen Erkenntnisse über den Einsatz und die Potenziale von SVM in Kapitel 5 abgeschlossen.

Kapitel 2

Methodische Grundlagen

Das vorliegende Kapitel bietet einen kompakten Überblick über die der SVM zugrunde liegende Methodik. Dazu werden im Folgenden zunächst die Basisansätze der SVM erläutert, bevor die Erweiterung der Biklassifikation auf den Mehrklassenfall vorgestellt wird. Abschließend wird kurz auf zwei weitere kernbasierte Verfahren eingegangen. Es wird die Grundlage geschaffen, auf der in Kapitel 3 bisher in der Literatur vorliegende Erweiterungen vorgestellt und neue vorgeschlagen werden können. Dieses Kapitel bildet somit einen Überblick über das Ziel und die Arbeitsweise kernbasierter Verfahren.

2.1 Basisansätze

Ziel der SVM-Methodik ist es, in vorgegebenen Daten Abhängigkeiten und Strukturen zu erkennen. Dabei sollen die bekannten Daten nicht nur genau beschrieben werden, sondern es soll auch eine gute Generalisierungsfähigkeit für die Prognose neuer Beobachtungen erzielt werden. Dazu werden Trennfunktionen bestimmt, die die vorgegebenen Klassen voneinander separieren sollen und somit den Zusammenhang zwischen den die Beobachtungen beschreibenden Merkmalen und den jeweiligen Klassenzugehörigkeiten erklären. Beobachtungen werden hierbei als Vektoren in einem n-dimensionalen Raum aufgefasst, die durch (Hyper-)ebenen[1] getrennt werden sollen. Der Methodik liegen die drei wesentliche Elemente Dualität, Kerne und Support Vektoren zugrunde (*Bennett, Campbell* (2000)). Durch Verwendung eines dualen Optimierungsproblems wird eine Trennebene bestimmt. Die nicht lineare Trennung wird durch den Einsatz so genannter Kernfunktionen ermöglicht. Die Support Vektoren sind diejenigen Vektoren, die die Lage der Trennebene beeinflussen. Dieses verteilungsfreie Verfahren der SVM ist insbesondere dann gut einsetzbar, wenn die Anzahl an beschreibenden Merkmalen sehr groß ist (*Wrobel et al.* (2000)) und eine große Anzahl an Beobachtungen analysiert werden soll.

[1] Im Folgenden werden die Bezeichnungen Hyperebene und Ebene synonym verwendet, sodass auch in einem hochdimensionalen Raum von einer Trennebene gesprochen wird.

Es werden im Folgenden die grundlegenden mathematischen Schritte erarbeitet, die zur Bestimmung einer derartigen Entscheidungsfunktion zur Klassifikation von Objekten notwendig sind. Zunächst wird dazu auf das hinter den SVM stehende Prinzip der Risikominimierung eingegangen.

2.1.1 Prinzip der Risikominimierung

Für die folgenden Betrachtungen seien Trainingsvektoren $\boldsymbol{x}_i \in \mathbb{R}^n$, $i \in \{1, \ldots, l\}$, mit Klassenzugehörigkeiten $y_i \in \{-1, +1\}$ gegeben, die einer Verteilung $P(\boldsymbol{x}, y)$ genügen. Diese Trainingsdaten bilden die Grundlage für die spätere Berechnung der SVM. Es wird zudem zwischen Trainings- und Testdaten unterschieden, wobei die Testdaten der Überprüfung der Prognosequalität dienen und die gleiche Struktur wie die Trainingsdaten aufweisen. Die Klassenzugehörigkeiten entsprechen den Ausgabewerten der zu berechnenden Funktion $f : \mathbb{R}^n \to \{-1, 1\}$. Um Prognosen für unbekannte Testdaten abgeben zu können, muss eine Funktion $f_\beta(\boldsymbol{x})$ gefunden werden, die das erwartete Risiko minimiert. Dieser Wert gibt damit den zu erwartenden Fehler an, der mittels der berechneten Trennfunktion bei der Klassifikation neuer Beobachtungen auftreten würde. Dieses Risiko ist wie folgt definiert (vgl. *Schölkopf et al.* (1999)):

$$R_{erw} = \int \frac{1}{2} |y - f_\beta(\boldsymbol{x})| dP(\boldsymbol{x}, y). \tag{2.1}$$

Die Funktion $f_\beta(\cdot)$ wird durch den Parameter β gekennzeichnet, wobei dieser stellvertretend für die zu bestimmenden Parameter der trennenden Funktion steht. Demnach umfasst β alle variablen Größen, die die Trennfunktion festlegen und diese somit charakterisieren. R_{erw} kann im Allgemeinen nicht berechnet werden, da die Verteilung $P(\boldsymbol{x}, y)$ der Eingabedaten unbekannt ist. Um dennoch etwas über die Güte der gefundenen Funktion aussagen zu können, wird das empirische Risiko herangezogen, das die Anzahl der auftretenden Klassifikationsfehler innerhalb der Trainingsdaten beinhaltet und somit das erwartete Risiko approximiert:

$$R_{emp} = \frac{1}{2l} \sum_{i=1}^{l} |y_i - f_\beta(\boldsymbol{x}_i)|,$$

wobei l die Anzahl der Trainingsvektoren ist. Im Falle der Trennung von zwei Klassen zählt der Ausdruck lR_{emp} also die Anzahl der fehlklassifizierten Vektoren (*Orsenigo, Vercellis* (2003)). Nachteil der Zielsetzung, nur das empirische Risiko zu minimieren, ist die fehlende Generalisierungsfähigkeit, da nur die Anpassung an die Trainingsdaten optimiert wird, was aufgrund dieser einseitigen Betrachtungsweise zu einer Überanpassung, dem so genannten Overfitting, führt.
Um neben der Anpassung der Trennfunktion an die Trainingsdaten auch deren Komplexität zu berücksichtigen, die bei der Minimierung des empirischen Risikos recht hoch wäre, wird das strukturelle Risiko betrachtet. Mit Hilfe der strukturellen Risikominimierung kann dem Problem der hohen Komplexität der Trennfunktion ent-

gegen gewirkt werden. Ziel ist es, das erwartete Risiko (2.1) nach oben zu begrenzen. Dazu ist die Einführung der VC-Dimension[2] erforderlich. Die VC-Dimension einer Funktionenklasse bezeichnet die maximale Anzahl an Beobachtungen, die unabhängig von ihrer Klassenzugehörigkeit fehlerfrei getrennt werden können (vgl. *Schölkopf, Smola* (2002)). So entspricht die VC-Dimension der Funktionenklasse der Geraden im $\mathbb{R}^2$ gerade drei, wie leicht mit Hilfe von Abbildung 2.1 zu erkennen ist. Unabhängig von der Zuweisung zu den zwei Klassen, können hierbei die beiden so definierten Gruppen durch eine Gerade getrennt werden, was bei Heranziehung von vier Beobachtungspunkten im $\mathbb{R}^2$ und beliebiger Klassenzugehörigkeit nicht mehr möglich ist.

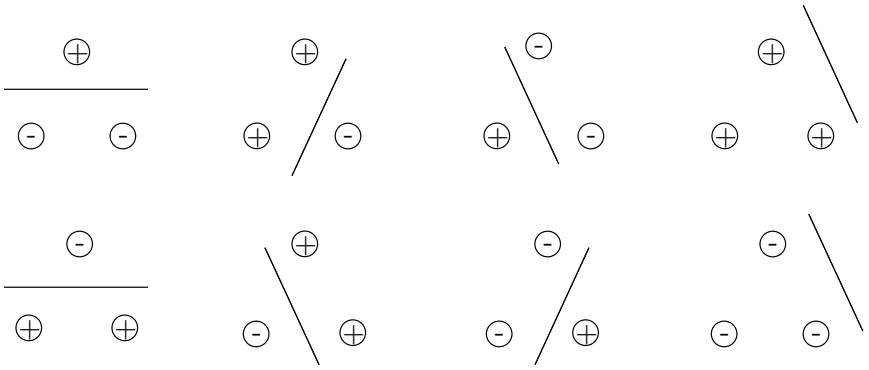

Abbildung 2.1: Lineare Trennung von drei Punkten im $\mathbb{R}^2$ unabhängig von ihren Klassenzugehörigkeiten

Nach *Vapnik* gilt mit $0 \leq \zeta \leq 1$ die folgende Abschätzung (*Burges* (1998))

$$R_{erw} \leq R_{emp} + \sqrt{\frac{h(\log(\frac{2l}{h}) + 1) - \log(\frac{\zeta}{4})}{l}}, \tag{2.2}$$

mit einer Wahrscheinlichkeit von $1 - \zeta$, wobei h die VC-Dimension der jeweiligen Funktionenklasse bezeichnet. Der zweite Summand der rechten Seite wird auch die VC-Konfidenz genannt und ist eine monoton steigende Funktion der VC-Dimension. Vorteil dieser Darstellung ist die Unabhängigkeit des erwarteten Risikos von $P(\boldsymbol{x}, y)$. Das Prinzip des strukturellen Risikos sieht die Minimierung der rechten Seite von Ungleichung (2.2) vor (vgl. *Müller et al.* (2001)), sodass damit das erwartete Risiko minimiert wird. Die skizzierte Situation ist in Abbildung 2.2 wiedergegeben.

[2]Die VC-Dimension wurde benannt nach *Vapnik* und *Chervonenkis*.

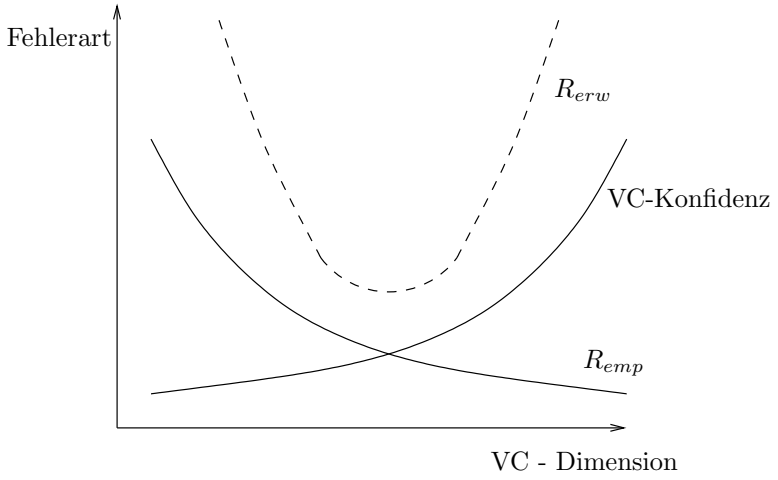

Abbildung 2.2: Einfluss der VC-Dimension auf R_{emp}, R_{erw} und die VC-Konfidenz (in Anlehnung an *Müller et al.* (2001))

Gesucht wird demnach eine Funktion, die zwei vorgegebene Klassen so trennt, dass sowohl das empirische Risiko, gemessen anhand der Trainingsdaten, als auch die VC-Konfidenz klein gehalten werden und zu einem Minimum des erwarteten Risikos R_{erw} (vgl. Abbildung 2.2) führt. Beide Ausdrücke werden durch die VC-Dimension beeinflusst, die bei gleichzeitiger Minimierung der Fehler nach oben begrenzt werden sollte. Nach *Vapnik* (1995) kann die VC-Dimension nach oben durch einen Ausdruck begrenzt werden, der umgekehrt proportional von der Spanne der Daten abhängig ist. Somit wird durch die im folgenden Abschnitt vorgestellte Maximierung dieser Spanne die VC-Dimension möglichst klein gehalten und demnach das Prinzip der Risiko-Minimierung umgesetzt. Dadurch sind SVM auch unempfindlicher gegenüber Overfitting, wie *Veropoulos et al.* (1999) ausführen.

2.1.2 Lineare Support Vektor Maschinen

Bei der Berechnung der Entscheidungsfunktion zur Trennung vorgegebener Klassen soll zunächst angenommen werden, dass die Daten linear und ohne Fehler trennbar sind. Dazu werden die Daten als Vektoren im $\mathbb{R}^n$ aufgefasst, die durch n Merkmale[3] beschrieben sind. Sie gehören zu zwei Klassen, die durch eine Ebene getrennt werden sollen.
Ziel ist es, eine Ebene zu konstruieren, die eine optimale Trennung zwischen den beiden zu untersuchenden Klassen „+1“ und „-1“ ermöglicht, d.h. eine Trennung mit möglichst wenigen, fehlklassifizierten Vektoren. Diese Ebene wird durch die Menge $\{\boldsymbol{x}|\boldsymbol{w}\boldsymbol{x} + b = 0\}$ beschrieben[4]. Der Vektor $\boldsymbol{w}$ ist ein Normalenvektor dieser Ebene und durch $\frac{|b|}{\|\boldsymbol{w}\|}$ wird mit $b \in \mathbb{R}$ und $\boldsymbol{w} \in \mathbb{R}^n$ der Abstand der Ebene zum Ursprung angegeben.[5] Diese Ebene soll nun so bestimmt werden, dass damit zwei

[3] Die Merkmale werden synonym auch als Variablen bezeichnet.
[4] Dabei bezeichnet $\boldsymbol{w}\boldsymbol{x}$ das Skalarprodukt zwischen den Vektoren $\boldsymbol{w}$ und $\boldsymbol{x}$.
[5] $\|\boldsymbol{w}\|$ bezeichne die euklidische Norm des Vektors $\boldsymbol{w}$, also $\|\boldsymbol{w}\| = \sqrt{\boldsymbol{w}\boldsymbol{w}}$.

Ziele erreicht werden: die Maximierung des Abstands zwischen den Klassen und die Minimierung der Fehler in den Trainingsdaten. Dies entspricht dem empirischen Risiko. Die Maximierung der Spanne zwischen den Klassen bewirkt eine Minimierung der VC-Dimension (*Müller et al.* (2001)). Um dies umzusetzen, wird die Ebene so durch die Daten gelegt, dass der Abstand der jeweils nächsten Punkte beider Klassen zur Ebene maximiert wird. Hier kommt ein wesentlicher Unterschied z.B. zur Diskriminanzanalyse zum Ausdruck, bei der die Streuung innerhalb des gesamten eingehenden Datensatzes einer Klasse für die Trennung relevant ist. Diese Streuung spielt bei SVM keine Rolle, es kommt hier lediglich auf die Spanne zwischen den jeweils nächsten Vektoren einer Klasse an. Die Bestimmung der optimalen Hyperebene durch diese Bedingung gewährleistet eine optimale Generalisierungsfähigkeit (vgl. z.B. *Inoue, Abe* (2001)), sodass nur in wenigen Fällen Overfitting auftritt. Die Wahl einer optimalen Hyperebene D_1 wird in Abbildung 2.3 aufgezeigt. Die durch die fett gezeichnete Linie markierte Ebene wäre in diesem Fall die optimale. In den anderen Fällen ist die Spanne zwischen den Punkten nicht maximal, wie die eingezeichneten Abstände der jeweils nächsten Punkte zu den jeweiligen Ebenen D_1, D_2 und D_3 verdeutlichen.

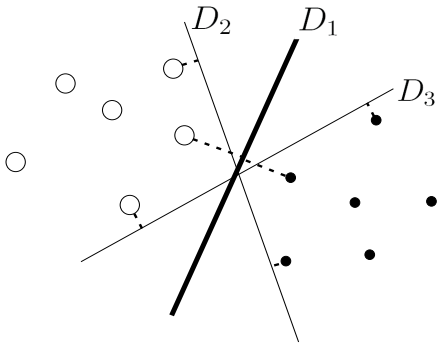

Abbildung 2.3: Wahl der optimalen Hyperebene

Um das oben beschriebene Ziel zu erreichen, werden zwei Hilfsebenen $\{\boldsymbol{x}|\boldsymbol{w}\boldsymbol{x}+b=1\}$ und $\{\boldsymbol{x}|\boldsymbol{w}\boldsymbol{x}+b=-1\}$ eingeführt, die den gleichen, maximalen Abstand zur Trennebene haben und so positioniert werden, dass die jeweils nächsten Punkte der beiden Klassen auf diesen Hilfsebenen liegen. Die Situation ist in Abbildung 2.4 wiedergegeben[6]. Das zweite zu erreichende Ziel besteht in der Richtigklassifikation der Beobachtungen. Bei Vorgabe der Definitionen der Hilfsebenen müssen die Eingabevektoren die folgenden Bedingungen erfüllen:

$$\begin{aligned} \boldsymbol{w}\boldsymbol{x}_i + b &\geq 1 \quad \text{für } y_i = 1 \\ \boldsymbol{w}\boldsymbol{x}_i + b &\leq -1 \quad \text{für } y_i = -1 \end{aligned}$$

für alle $i = 1, \ldots, l$. Diese beiden Bedingungen können zu einer zusammengefasst werden:

$$y_i(\boldsymbol{w}\boldsymbol{x}_i + b) - 1 \geq 0, \quad \forall i \in \{1, \ldots, l\}.$$

[6]Eine Maximierung der Spanne entspricht dabei einer Minimierung der VC-Dimension (*Schölkopf et al.* (1999b)).

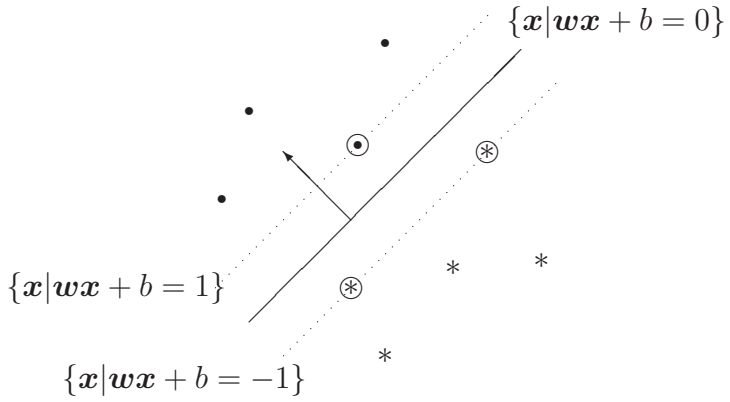

Abbildung 2.4: Trennung der Eingabedaten durch separierende (Hilfs-)Ebenen

Die zu maximierende Spanne lässt sich durch den Abstand $D_{+/-}$ der beiden Hilfsebenen berechnen. Dieser Abstand kann ausgedrückt werden durch:

$$D_{+/-} = \min_{i:y_i=1}(d(\boldsymbol{x}_i;\boldsymbol{w},b)) + \min_{i:y_i=-1}(d(\boldsymbol{x}_i;\boldsymbol{w},b)),$$

wobei $d(\boldsymbol{x}_i;\boldsymbol{w},b) = \frac{|\boldsymbol{w}\boldsymbol{x}_i+b|}{||\boldsymbol{w}||}$ den Abstand eines Punktes $\boldsymbol{x}_i$ zu der durch $\boldsymbol{w}$ und b bestimmten Ebene bezeichnet. Ziel ist es, diesen Abstand zu maximieren. Es ergibt sich daher:

$$\begin{aligned}\max D_{+/-} &= \max(\min_{i:y_i=1}(d(\boldsymbol{x}_i;\boldsymbol{w},b)) + \min_{i:y_i=-1}(d(\boldsymbol{x}_i;\boldsymbol{w},b)))\\ &= \max(\min_{i:y_i=1}\tfrac{|\boldsymbol{w}\boldsymbol{x}_i+b|}{||\boldsymbol{w}||} + \min_{i:y_i=-1}\tfrac{|\boldsymbol{w}\boldsymbol{x}_i+b|}{||\boldsymbol{w}||})\\ &= \max(\tfrac{1}{||\boldsymbol{w}||} + \tfrac{1}{||\boldsymbol{w}||})\\ &= \max\ \tfrac{2}{||\boldsymbol{w}||}.\end{aligned}$$

Somit resultiert das folgende Optimierungsproblem, um die Spanne zwischen den Klassen zu maximieren[7]:

$$\min\ \frac{1}{2}\,||\boldsymbol{w}||^2 \tag{2.3}$$

unter den Nebenbedingungen

$$y_i(\boldsymbol{w}\boldsymbol{x}_i + b) \geq 1 \qquad \forall\ i \in \{1,\ldots,l\}. \tag{2.4}$$

Um die numerische Berechnung zu vereinfachen und auch nicht lineare Trennung einführen zu können, wird zum dualen Optimierungsproblem gewechselt (vgl. *Burges* (1998)).
Dazu wird zunächst die Lagrange-Funktion zu (2.3) und (2.4) gebildet, indem Lagrange-Multiplikatoren $\alpha_1,\ldots,\alpha_l$ mit $\alpha_i \geq 0$ für alle i eingeführt werden:

$$L(\boldsymbol{w},\ b,\ \boldsymbol{\alpha}) := \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{l}\alpha_i y_i(\boldsymbol{w}\boldsymbol{x}_i + b) + \sum_{i=1}^{l}\alpha_i, \tag{2.5}$$

[7]Da die Eindeutigkeit des Ergebnisses gewährleistet werden soll, wird hierbei das Quadrat $||\boldsymbol{w}||^2$ verwendet.

mit $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_l)$. Gesucht ist nun der Sattelpunkt dieser Funktion, das heißt, $L(\boldsymbol{w},\ b,\ \boldsymbol{\alpha})$ muss hinsichtlich $\boldsymbol{w}$ und b maximiert und hinsichtlich der Lagrange-Multiplikatoren $\alpha_1, \ldots, \alpha_l$ minimiert werden.
Dazu müssen die Gradienten der Funktion $L(\boldsymbol{w},\ b,\ \boldsymbol{\alpha})$ bezüglich $\boldsymbol{w}$ und b verschwinden. Dies führt nach wenigen Umformungen dazu, dass die folgenden Bedingungen erfüllt sein müssen:

$$\frac{\partial L(\boldsymbol{w},\ b,\ \boldsymbol{\alpha})}{\partial\ \boldsymbol{w}} = 0 \Leftrightarrow \boldsymbol{w} = \sum_{i=1}^{l} \alpha_i y_i \boldsymbol{x}_i \tag{2.6}$$

$$\frac{\partial L(\boldsymbol{w},\ b,\ \boldsymbol{\alpha})}{\partial\ b} = 0 \Leftrightarrow \sum_{i=1}^{l} \alpha_i y_i = 0. \tag{2.7}$$

Wird nun Bedingung (2.6) in Gleichung (2.5) eingesetzt und zusätzlich Gleichung (2.7) ausgenutzt, so resultiert das duale Optimierungsproblem

$$\max \Big\{ \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i \boldsymbol{x}_j \Big\} \tag{2.8}$$

unter den Nebenbedingungen

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \qquad \text{und} \qquad \alpha_i \geq 0 \quad \forall\ i. \tag{2.9}$$

Dabei sind die Karush-Kuhn-Tucker-Bedingungen (KKT-Bedingungen)

$$\alpha_i (y_i (\mathrm{w}\boldsymbol{x}_i + b) - 1) = 0 \tag{2.10}$$

für alle $i \in \{1, \ldots, l\}$ einzuhalten.

Die berechneten Werte für α_i geben einen Anhaltspunkt über den Einfluss des jeweiligen Eingabevektors $\boldsymbol{x}_i$ auf die Lage der berechneten Hyperebene. Gilt $\alpha_i > 0$ für ein i, so heißt der zugehörige Vektor $\boldsymbol{x}_i$ Support Vektor. Diese Vektoren bilden die Menge der Daten, die für das Verfahren charakteristisch sind. Sie sind die einzigen Vektoren, die in die letztendliche Berechnung der Ebene eingehen. Für die übrigen Vektoren gilt $\alpha_i = 0$. Somit treten diese im rechten Teil der Relation (2.6) zur Berechnung des Normalenvektors der Ebene nicht mehr auf. Dies wird später in Gleichung (2.15) noch deutlicher. In Abbildung 2.4 sind die Support Vektoren mit einem Kreis gekennzeichnet. Nur diese Vektoren haben einen Einfluss auf die Lage der Ebene, sodass eine erneute Optimierung unter Berücksichtigung lediglich der Menge der Support Vektoren exakt die gleiche Lösung erzeugen würde. Die Anzahl der Support Vektoren gibt die Anzahl an Beobachtungen an, die bei fest vorgegebener Parameterkonstellation nötig sind, um die Eingabedaten zu repräsentieren und die Trennebene zu bestimmen. Ein Ziel in der Anwendung von SVM liegt darin, diese Anzahl auf ein Minimum zu beschränken. Liegen die Support Vektoren exakt auf den Hilfsebenen, so spricht man von so genannten Margin-Vektoren.

2.1.3 Lineare SVM auf linear nicht trennbaren Daten

Bisherige Voraussetzung für die Berechnung der Trennebene war die lineare Trennbarkeit der Eingabedaten. Dies bedeutet, dass eine lineare Hyperebene gefunden werden kann, die die Daten ohne Fehler voneinander trennt. Werden auch Fehler zu einem gewissen Grad zugelassen, so verändert sich das ursprüngliche Optimierungsproblem (2.3) mit (2.4). Diese Veränderungen können wie folgt ausgedrückt werden (*Cortes, Vapnik* (1995)):

$$\min \ \frac{1}{2} ||\boldsymbol{w}||^2 + C \sum_{i=1}^{l} \xi_i \tag{2.11}$$

unter den Nebenbedingungen

$$y_i(\boldsymbol{w}\boldsymbol{x}_i + b) \geq 1 - \xi_i \qquad \forall \ i \in \{1, \ldots, l\}, \tag{2.12}$$

wobei ξ_i eine Schlupfvariable ist, die Information über Fehler wie folgt beinhaltet. Falls eine Fehlklassifikation, also eine Verletzung der in Gleichung (2.4) geforderten Bedingung vorliegt, so ist $\xi_i > 1$. Gilt $0 \leq \xi_i \leq 1$, so liegt der Eingabevektor $\boldsymbol{x}_i$ zwischen den Hilfsebenen. Für $\xi_i = 0$ ist der Vektor auf der ihm zugewiesenen Seite der Hyperebene zu finden und damit richtig klassifiziert.
Die beschriebene Situation ist in Abbildung 2.5 veranschaulicht. Bei Vektor $\boldsymbol{x}_1$ liegt kein Fehler vor, aber dennoch eine Verletzung der ursprünglich geforderten Bedingung (2.4) vor, sodass $0 < \xi_1 < 1$ gilt. Die Vektoren $\boldsymbol{x}_2$ und $\boldsymbol{x}_3$ sind falsch klassifiziert. In beiden Fällen ist die zugehörige Variable ξ_2 bzw. ξ_3 größer als Eins.

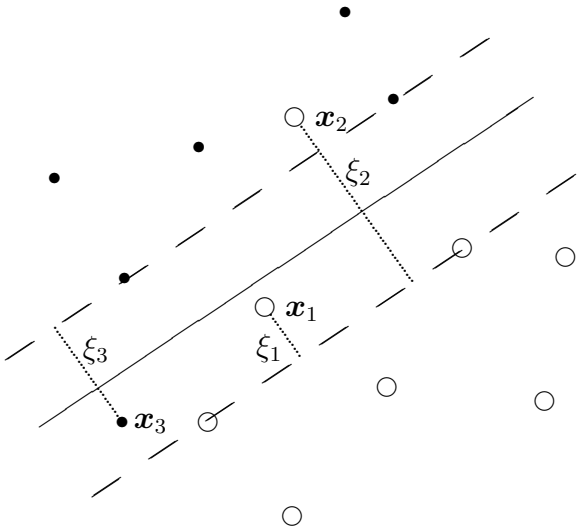

Abbildung 2.5: Schlupfvariablen mit unterschiedlichen Ausprägungen

Wird auch in diesem Fall zur Darstellung mittels Lagrange-Multiplikatoren übergegangen, so resultiert das Optimierungsproblem:

$$\max \Big\{ \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i \boldsymbol{x}_j \Big\} \tag{2.13}$$

unter den Nebenbedingungen

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \qquad \text{und} \quad 0 \leq \alpha_i \leq C \quad \forall\, i. \tag{2.14}$$

Weiter reichende Ausführungen sind etwa in *Schölkopf, Smola* (2002) enthalten. Man erkennt, dass sich die Optimierungsprobleme (2.8) mit (2.9) und (2.13) mit (2.14) nur durch die obere Schranke C für die gesuchten Werte α_i unterscheiden. Dabei bewirkt dieser Parameter, dass der Einfluss jedes einzelnen Punktes auf die Lage der Ebene nach oben beschränkt wird (vgl. *Schölkopf, Smola* (2002)). Mit Hilfe der a priori zu wählenden Obergrenze C kann der Anwender somit zwischen der Maximierung der Spanne zwischen den Hilfsebenen und der Minimierung der auftretenden Fehler variieren. Je größer der Wert für C, desto eher steht die Minimierung der Fehler im Vordergrund der Optimierung, wie in Ausdruck (2.11) ersichtlich ist.

Nachdem die Werte für die Lagrange-Multiplikatoren berechnet worden sind, kann durch den rechten Teil der Relation (2.6) der Normalenvektor $\boldsymbol{w}$ bestimmt werden. Für die Berechnung des Biasterms b wird die KKT-Bedingung (2.10) ausgenutzt. Für einen beliebigen Vektor $\boldsymbol{x}_i$, für den $\alpha_i > 0$ gilt, kann nun b aus folgender Gleichung bestimmt werden (vgl. *Burges* (1998)):

$$y_i(\boldsymbol{w}\boldsymbol{x}_i + b) - 1 = 0$$

im linear trennbaren Fall, bzw.

$$y_i(\boldsymbol{w}\boldsymbol{x}_i + b) - 1 + \xi_i = 0$$

im nicht linear trennbaren Fall. In beiden Fällen wird die Entscheidungsfunktion $f : \mathbb{R}^n \rightarrow \{-1, +1\}$ zur Klassifikation eines neuen Datenpunktes $\boldsymbol{x}$ wie folgt bestimmt:

$$f(\boldsymbol{x}) = sign(\sum_{i \in \mathcal{SV}} \alpha_i y_i \boldsymbol{x}_i \boldsymbol{x} + b). \tag{2.15}$$

Dabei bezeichnet $\mathcal{SV}$ die Menge der Indizes der berechneten Support Vektoren. Die Entscheidung darüber, in welche Klasse ein neuer Datenvektor einzuordnen ist, ist damit nur von den Support Vektoren abhängig, was bei großen Datenmengen eine Reduzierung des Rechenaufwandes bei der Neuklassifikation mit sich bringt. Die Reduzierung der Eingabedaten auf die für die Trennung relevanten Vektoren, den Support Vektoren, gibt dem Verfahren den Namen. Die übrigen Vektoren spielen bei der letztendlichen Klassifikation keine Rolle mehr.

2.1.4 Nicht lineare Trennung mittels Kernfunktionen

Neben der einfachen Berechnung liegt ein weiterer Vorteil des dualen Optimierungsproblems darin, dass die Eingabevektoren nur in Form von Skalarprodukten in die

Optimierung eingehen. Um eine nicht lineare Trennung zu ermöglichen, werden Kernfunktionen eingeführt, die dieses Skalarprodukt ersetzen. Dieser „Kerntrick“ kann ebenfalls bei anderen Verfahren, in denen ausschließlich Skalarprodukte auftreten, zum Beispiel bei der Hauptkomponentenanalyse zur Nichtlinearisierung der Verfahren genutzt werden, worauf in Abschnitt 2.3 eingegangen wird.

Die Grundidee bei der nicht linearen Trennung basiert darauf, die Datenvektoren durch eine Abbildung $\Phi : \mathbb{R}^n \to H$ in einen höherdimensionalen Merkmalsraum H abzubilden, um sie dort linear zu trennen (*Burges* (1998)).
Das ursprüngliche Optimierungsproblem ändert sich dahingehend, dass nicht mehr das Skalarprodukt $\boldsymbol{x}_i\boldsymbol{x}_j$ der Eingabevektoren $\boldsymbol{x}_i$ und $\boldsymbol{x}_j$ berechnet wird, sondern dieses durch das Skalarprodukt $\Phi(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_j)$ der Bildvektoren ersetzt wird. Damit stellt sich das in (2.13) und (2.14) gestellte Optimierungsproblem als

$$\max\Big\{\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j \Phi(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_j)\Big\} \tag{2.16}$$

unter den Nebenbedingungen

$$\sum_{i=1}^{l}\alpha_i y_i = 0 \qquad \text{und} \qquad 0 \le \alpha_i \le C \quad \forall\, i. \tag{2.17}$$

Ein Beispiel für eine derartige Abbildung ist die folgende: $\Phi : \mathbb{R}^2 \to \mathbb{R}^6$ wobei

$$(x_1, x_2) \mapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1). \tag{2.18}$$

Ziel ist es nun, die Eingabedaten im Merkmalsraum $\mathbb{R}^6$ linear zu trennen. Allgemein kann $\Phi(\cdot)$ als eine Abbildung der obigen Art die folgende Form haben.

$$\Phi(x_1, \ldots, x_n) = (x_1^2, \ldots, x_n^2, \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_{n-1}x_n, \sqrt{2}x_1, \ldots, \sqrt{2}x_n, 1)$$

Hätte der Eingaberaum die Dimension $n = 100$, so hätte der Merkmalsraum hier bereits die Dimension $100 + \frac{100\cdot 99}{2} + 100 + 1 = 5151$, wobei es sich, wie sich zeigen wird, eher um eine einfache Form der zur Verfügung stehenden Abbildungen $\Phi(\cdot)$ handelt.

Dies bringt zum Ausdruck, dass durch die Transformation der Eingabedaten in einen höher dimensionalen Merkmalsraum zwar die Eigenschaft der nicht linearen Trennung gewonnen wird, dies allerdings zu Lasten des zu bewältigenden Rechenaufwands geht.
Die Situation der nicht linearen Trennung im höher dimensionalen Raum wird durch Abbildung 2.6 beschrieben. Die Daten werden mittels einer Abbildung Φ in einen Merkmalsraum transferiert und können dort durch eine Ebene getrennt werden. Diese lineare Entscheidungsfläche entspricht einer nicht linearen Trennung im Eingaberaum.

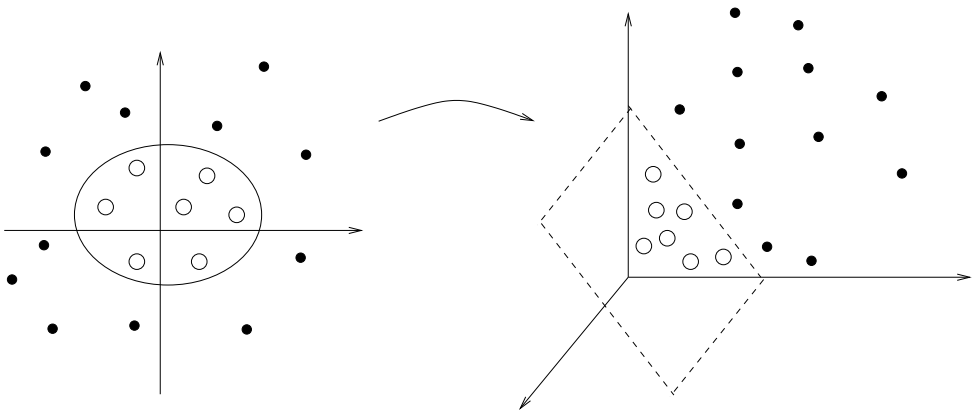

Abbildung 2.6: Abbildung nicht linearer Strukturen in einen höher dimensionalen Raum (Quelle: *Müller et al.* (2001))

Durch diese Darstellung werden die zu berechnenden Skalarprodukte unter Umständen sehr groß, sodass hier vom „Fluch der Dimensionen“ (*Schölkopf et al.* (1999b)) gesprochen wird. An dieser Stelle kommt die Bedingung von *Mercer* zum Einsatz (*Mercer* (1909)), welche die Umgehung dieses Problems durch Einsatz so genannter Kernfunktionen erlaubt. Nach *Cristianini, Shawe-Taylor* (2000) ist ein Kern eine Funktion $K : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$, sodass für alle $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbb{R}^n$ die folgende Beziehung gilt:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \Phi(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_j),$$

wobei Φ eine Abbildung vom Eingaberaum $\mathbb{R}^n$ in einen Merkmalsraum $\mathbb{R}^m$ mit $n,\ m \in \mathbb{N}$ ist.
Kernfunktionen ermöglichen eine implizite Berechnung von Skalarprodukten in einem höher dimensionalen Raum und werden im Rahmen von SVM zur Bestimmung nicht linearer Trennfunktionen eingesetzt. Unter welchen Umständen eine solche Auswechslung vorgenommen werden kann, gibt die Bedingung von *Mercer* an[8] (vgl. *Burges* (1998)):
Es gibt eine Abbildung Φ *und eine Erweiterung (Kernfunktion)* $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_k \Phi(\boldsymbol{x}_i)_k \Phi(\boldsymbol{x}_j)_k$ *genau dann, wenn für jede integrierbare Funktion* $g(\boldsymbol{x})$*, für die* $\int g(\boldsymbol{x})^2 d\boldsymbol{x}$ *endlich ist, gilt:*

$$\int K(\boldsymbol{x}_i, \boldsymbol{x}_j) g(\boldsymbol{x}_i) g(\boldsymbol{x}_j) d\boldsymbol{x}_i d\boldsymbol{x}_j \geq 0.$$

Dabei garantiert diese Bedingung unter den gegebenen Voraussetzungen die Existenz eines Skalarproduktes in einem höher dimensionalen Raum, sie gibt allerdings keine Auskunft über die Gestalt dieses Skalarproduktes bzw. die der Abbildung.
Das letztendliche Optimierungsproblem ergibt sich aus dem Ersatz des Skalarproduktes der abgebildeten Eingabevektoren im Ausdruck (2.16) durch eine Kernfunktion:

$$\max \Big\{ \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \Big\} \tag{2.19}$$

[8] Für weitere Ausführung zum Theorem von *Mercer* siehe *Schölkopf, Smola* (2002).

unter den Nebenbedingungen (analog zu Gleichung (2.17))

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \qquad \text{und} \quad 0 \leq \alpha_i \leq C \quad \forall\, i. \tag{2.20}$$

Wird wieder das in (2.18) gezeigte Beispiel der Abbildung Φ betrachtet, so ergibt sich bei der Bestimmung des Skalarproduktes die folgende Rechnung:

$$\begin{aligned}
\Phi(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_j) &= \Phi(x_{i1}, x_{i2})\Phi(x_{j1}, x_{j2}) \\
&= (x_{i1}^2, x_{i2}^2, \sqrt{2}x_{i1}x_{i2}, \sqrt{2}x_{i1}, \sqrt{2}x_{i2}, 1) \\
&\quad (x_{j1}^2, x_{j2}^2, \sqrt{2}x_{j1}x_{j2}, \sqrt{2}x_{j1}, \sqrt{2}x_{j2}, 1) \\
&= x_{i1}^2 x_{j1}^2 + x_{i2}^2 x_{j2}^2 + 2x_{i1}x_{i2}x_{j1}x_{j2} + 2x_{i1}x_{j1} + 2x_{i2}x_{j2} + 1 \\
&= (x_{i1}x_{j1} + x_{i2}x_{j2} + 1)^2 \\
&= (\boldsymbol{x}_i\boldsymbol{x}_j + 1)^2
\end{aligned}$$

Das Skalarprodukt im Merkmalsraum entspricht einer funktionalen Beziehung der Vektoren im Eingaberaum. Die möglicherweise aufwändige Berechnung des Skalarproduktes $\Phi(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_j)$ kann hierbei durch ein Polynom zweiten Grades ersetzt werden:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i\boldsymbol{x}_j + 1)^2.$$

Somit sind Berechnungen im Merkmalsraum, dessen Dimension sehr schnell sehr groß werden kann, nicht nötig. Diese können durch den Einsatz von Kernfunktionen ersetzt werden.
Um Datenvektoren nun mittels der Entscheidungsfunktion f zu klassifizieren, wird analog zu (2.15) vorgegangen:

$$f(\boldsymbol{x}) = sign\left(\sum_{\boldsymbol{x}_i \in \mathcal{SV}} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b\right). \tag{2.21}$$

Hierbei kann eine Berechnung des Normalenvektors $\boldsymbol{w}$ umgangen werden, die aufgrund der Abbildung $\Phi(\cdot)$ unter Umständen sehr umständlich sein kann bzw. unmöglich ist. Für homogene Polynome vom Grad d ($K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}\boldsymbol{y})^d$) und einen n-dimensionalen Eingaberaum (die Beobachtungen werden durch n Merkmale beschrieben) ist der zugehörige Raum, in dem die Daten linear getrennt werden, von der Dimension $\binom{n+d-1}{d}$ (*Burges* (1998)). Das bedeutet, dass der Merkmalsraum für ein Polynom vom Grad drei und 15-dimensionalen Eingabedaten bereits die Dimension $\binom{17}{3} = \frac{17!}{3!14!} = 680$ aufweist. Steigt die Anzahl der Merkmale, so wächst die Dimension des Merkmalsraums in diesem Fall von der Ordnung $O(n^d)$ und kann nicht mehr einfach oder gar nicht mehr bearbeitet werden.

Sind bereits Kerne vorgegeben, so können durch einfache Operationen, zum Beispiel durch Addition oder skalarer Multiplikation weitere Kerne entwickelt werden. Erlaubt sind alle Verknüpfungen, die die positiv-Definitheit der Kerne erhalten. Näheres zur Entwicklung neuer Kerne ist beispielsweise in *Schölkopf, Smola* (2002) zu finden.

Nr.	Kernfunktion	Bemerkung	Quelle
1	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i \boldsymbol{x}_j$		*Schölkopf, Smola* (2002)
2	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i \boldsymbol{x}_j)^d$	$d \in \mathbb{N}$	
3	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i \boldsymbol{x}_j + 1)^d$	$d \in \mathbb{N}$	*Vapnik* (1995)
4	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\frac{\boldsymbol{x}_i \boldsymbol{x}_j}{a} + \tilde{b})^d$	$a \neq 0,\ \tilde{b}, d \in \mathbb{N}$	
5	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{1-(\boldsymbol{x}_i \boldsymbol{x}_j)^d}{1-(\boldsymbol{x}_i \boldsymbol{x}_j)}$	$-1 < \boldsymbol{x}_i \boldsymbol{x}_j < 1,\ d \in \mathbb{N}$	*Saunders et al.* (1998)
6	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp(-\gamma \lVert \boldsymbol{x}_i - \boldsymbol{x}_j \rVert^2)$	$\gamma \in \mathbb{R}^+$	*Vapnik* (1995)
7	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp(-\frac{\lVert \boldsymbol{x}_i - \boldsymbol{x}_j \rVert^2}{\sigma^2})$	$\sigma \in \mathbb{R} \backslash \{0\}$	*Schölkopf, Smola* (2002)
8	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = a(\exp(\frac{\gamma}{\lVert \boldsymbol{x}_i - \boldsymbol{x}_j \rVert^2 + \sigma^2}) - 1)$	$\sigma, \gamma \in \mathbb{R} \backslash \{0\},\ a \in \mathbb{R}$ konstant	*Ayat et al.* (2002)
9	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp(-\gamma D(\boldsymbol{x}_i, \boldsymbol{x}_j))$	$\gamma \in \mathbb{R}$, $D(\cdot, \cdot)$ Distanzmaß	*Bahlmann et al.* (2002)
10	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = tanh(r \boldsymbol{x}_i \boldsymbol{x}_j - s)$	$r, s \in \mathbb{R}$	*Schölkopf, Smola* (2002)
11	$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{\pi}{2\gamma} \frac{cosh(\frac{\pi - \lVert \boldsymbol{x}_i - \boldsymbol{x}_j \rVert}{\gamma})}{sinh(\frac{\pi}{\gamma})}$	$0 \leq \lVert \boldsymbol{x}_i - \boldsymbol{x}_j \rVert \leq 2\pi,\ \gamma \neq 0$	*Vapnik* (1998)

Tabelle 2.1: Eine Auswahl möglicher Kerne

Tabelle 2.1 enthält eine Auswahl möglicher Kerne, wobei diese mit einigen Ausnahmen in empirischen Analysen nur sehr selten eingesetzt werden. Üblicherweise kommen drei Arten von Kernen in der Anwendung von SVM zum Einsatz (*Burges* (1998)), die bereits gute Ergebnisse liefern und daher der Einsatz weiterer Kerne nicht nötig ist. Weiterhin sind in den gängigen Software-Tools lediglich diese drei Kerne verfügbar. Dies sind ein homogenes bzw. inhomogenes Polynom vom Grad d (Nr. 2 bzw. 3 in Tabelle 2.1), eine Radialbasis-Funktion mit Kernparameter γ (Nr. 6) und eine sigmoide Funktion, die ein neuronales Netz repräsentiert (Nr. 10), bei dem zwei Parameter r und s festgelegt werden müssen. Im Falle des neuronalen Netzes kommen nur bestimmte Kombinationen der beiden zu wählenden Parameter in Frage, damit die Funktion einen nach *Mercer*s Bedingung zulässigen Kern bildet (vgl. *Burges* (1998)).
Häufig werden für bestimmte Anwendungen, die besondere Eigenschaften fordern, spezielle Kerne entwickelt, was ein Grund für deren seltenen Einsatz in der Praxis bildet. So wurde beispielsweise in *Bahlmann et al.* (2002) ein Kern gesucht, der ein Vergleich von Objekten erlaubt, die nicht die gleiche Dimension aufweisen (vgl. Nr. 9 in Tabelle 2.1). Dies liegt bei diesem speziellen Fall in der Repräsentation von Handschriften begründet. Diese Buchstaben werden durch Sequenzen dargestellt, die unterschiedlicher Länge sein können, was den Einsatz eines speziellen Distanzmaßes $D(\cdot, \cdot)$ erfordert. Dieses ermöglicht den Vergleich derartiger Objekte. Angewendet wird dies bei der (online) Handschriftenerkennung z.B. bei einem PDA. So werden zur Integrierung spezieller Daten neue Kerne entworfen. Die traditionellen liefern dennoch gute Ergebnisse.
Zu jedem Kern gibt es Parameter, zum Beispiel den Grad des Polynoms, die a priori festgelegt werden müssen. In der Literatur gibt es allerdings nur wenige Hinweise dazu, wie diese Parameter gewählt werden sollten, sodass dieser Aspekt in Abschnitt 3.3 untersucht wird.

Der Einsatz von Kernen ermöglicht die Erkennung nicht linearer Strukturen innerhalb der zu klassifizierenden Daten und kann daher zu besseren Analyseergebnissen, verglichen mit linearen SVM, beitragen.

2.2 Multiklassifikation

Bisher wurde nur der Fall betrachtet, in dem Datenvektoren aus zwei verschiedenen Klassen voneinander getrennt werden sollten. Im Marketing liegt dies etwa bei der Unterscheidung von Käufern und Nichtkäufern vor. Allerdings können im Marketingkontext auch mehrere Käufersegmente Gegenstand der Untersuchung sein. Deshalb soll in diesem Kapitel ein Überblick über die Möglichkeiten gegeben werden, die im Rahmen von SVM zur Klassifikation bei Vorliegen mehrerer Gruppen zur Verfügung stehen. Hier kann zwischen den direkten Verfahren und denjenigen, die auf der Trennung zweier Klassen basieren, unterschieden werden. Letztere werden zwar auf das Lösen mehrerer binärer Klassifikationsprobleme zurückgeführt, sind allerdings aufgrund geringerer Rechenzeit häufig effektiver als

die direkten Verfahren zur Multiklassifikation (vgl. *Hsu, Lin* (2002)).

Analog zu Abschnitt 2.1 seien in allen Multiklassifikationsverfahren wiederum Datenvektoren $\boldsymbol{x}_i \in \mathbb{R}^n$ gegeben mit $i \in \{1, \ldots, l\}$ und entsprechenden Klassenzugehörigkeiten $y_i \in \{1, \ldots, K\}$, wobei K die Anzahl der zu untersuchenden Klassen angibt. Es soll eine Entscheidungsfunktion bestimmt werden, die einen neuen Datenpunkt $\boldsymbol{x}$ einer der K Klassen zuordnet.
Welche unterschiedlichen Möglichkeiten zur Bestimmung dieser Entscheidungsfunktion zur Verfügung stehen, wird in den folgenden Abschnitten vorgestellt. Hier wird auf die lineare Trennung zurückgegriffen, wobei die Vorgehensweisen ebenso für die nicht lineare Trennung gültig sind.

2.2.1 One-Against-All

Bei diesem Multiklassifikationsverfahren (vgl. *Kreßel* (1999)) wird jede Klasse für sich betrachtet und von den übrigen getrennt. Um den Ansatz der Biklassifikation benutzen zu können, werden die restlichen Klassen zu einer zusammengefasst, sodass insgesamt K Optimierungsprobleme bei K zu untersuchenden Klassen zu lösen sind.

Bei dem One-Against-All-Verfahren (OAA) sind in jedem zu lösenden binären Entscheidungsproblem zwei Klassen gegeben: zum einen die interessierende Klasse k und zum anderen eine Klasse, die aus den Datenvektoren besteht, die ursprünglich den restlichen $K - 1$ Klassen zugewiesen wurden. Damit hat dann das k-te Optimierungsproblem mit $k \in \{1, \ldots, K\}$ im Falle der linearen Trennung folgende Gestalt:

$$\min \ \frac{1}{2} ||\boldsymbol{w}^{[k]}||^2 + C \sum_{i=1}^{l} \xi_i^{[k]} \tag{2.22}$$

unter den Nebenbedingungen

$$\begin{aligned} \boldsymbol{w}^{[k]}\boldsymbol{x}_i + b^{[k]} &\geq 1 - \xi_i^{[k]} \quad , \text{falls } y_i = k \\ \boldsymbol{w}^{[k]}\boldsymbol{x}_i + b^{[k]} &\leq -1 + \xi_i^{[k]} \quad , \text{falls } y_i \neq k \\ \xi_i^{[k]} &\geq 0, \ i \in \{1, \ldots, l\}, \ k \in \{1, \ldots, K\}, \end{aligned} \tag{2.23}$$

wobei die Bezeichnungen aus der Methodik der Biklassifikation übernommen werden und der Index k die K verschiedenen Optimierungsprobleme kennzeichnet. Der Vektor $\boldsymbol{w}^{[k]}$ ist ein Normalenvektor der zu bestimmenden Hyperebene und wird entsprechend (2.6) im linearen Fall bestimmt.

Bei der Berechnung einer Klassifikationsfunktion gibt es verschiedene Vorgehensweisen. Nach K-maliger Lösung des Optimierungsproblems (2.22) und (2.23) entstehen zunächst insgesamt K Entscheidungsfunktionen der Form

$$f^{[k]}(\boldsymbol{x}) = sign(\boldsymbol{w}^{[k]}\boldsymbol{x} + b^{[k]}) \quad \forall \ k \in \{1, \ldots, K\}.$$

Falls $K = 3$ Klassen vorliegen, so liegt die durch Abbildung 2.7 veranschaulichte Situation vor. Für jede der drei Klassen $k = 1, 2, 3$ muss eine Trennebene $D^{[k]}$ bestimmt werden, die diese Klasse vom Rest der Datenvektoren trennt. Dabei ist $D^{[k]} := \{\boldsymbol{x} | \boldsymbol{w}^{[k]}\boldsymbol{x} + b^{[k]} = 0\}$. Auf Basis dieser Ebenen kann dann eine Zuweisungsvorschrift verwendet werden, die einem neuen Datenvektor $\boldsymbol{x}$ eine dieser Klassen zuordnet. Häufig findet diese Klassenzuweisung auf Basis der Signum-Funktion statt.

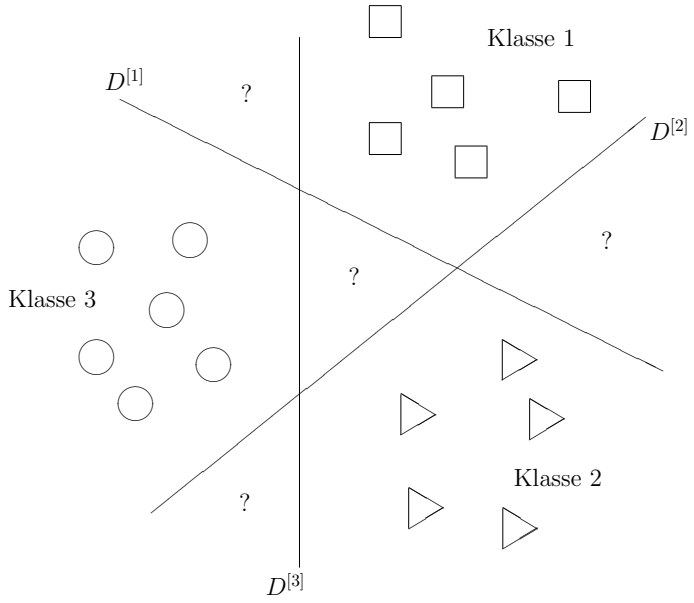

Abbildung 2.7: One-Against-All Trennung bei drei Klassen

Es wird also diejenige Klasse zugewiesen, deren zugehörige Entscheidungsfunktion $f^{[k]}$ den Wert 1 liefert. Diese Zuordnung ist nicht immer eindeutig, wie schon in Abbildung 2.7 zu erkennen ist. Falls der zu klassifizierende Datenvektor in eine mit einem Fragezeichen gekennzeichnete Fläche fällt, so kann keine eindeutige Zuordnung vorgenommen werden, da $f^{[k]} = 1$ für mindestens zwei Werte für k gilt.
Dieses Problem wird dadurch umgangen (*Kreßel* (1999)), dass die so genannte Winner-takes-all Methode anwendet wird, die den Eingabewert für die Signum-Funktion als Entscheidungsgrundlage verwendet. Die Klassenzuweisung y_{neu} eines Vektors $\boldsymbol{x}_{neu}$ erfolgt nun derart, dass dieser Vektor derjenigen Klasse zugeordnet wird, deren zugehörige Entscheidungsfunktion den größten Wert liefert:

$$y_{neu} = \arg \max_{k=1,\ldots,K} (\boldsymbol{w}^{[k]}\boldsymbol{x}_{neu} + b^{[k]}) \tag{2.24}$$

Die bei diesem auch als Tie-breaking bezeichneten Vorgehen entstehenden Trennebenen sind in Abbildung 2.8 wiedergegeben.

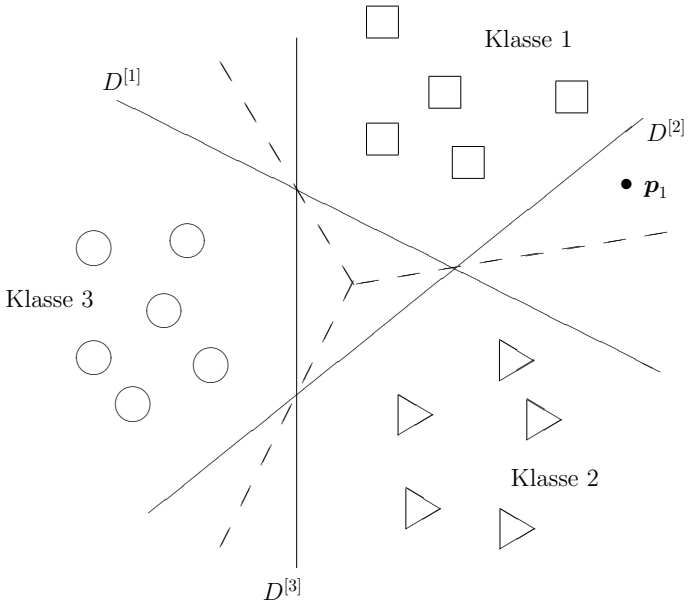

Abbildung 2.8: Tie-Breaking bei der One-Against-All Trennung bei drei Klassen

Hier ist zu erkennen, dass der Punkt $\boldsymbol{p}_1$, der sich im unscharfen Bereich befindet, mittels Tie-breaking aufgrund des größeren positiven Entscheidungswertes zu Klasse „1“ geordnet wird, sodass eine eindeutige Zuordnung aller Bereiche möglich wird.
Einen Nachteil einer derartigen Vorgehensweise bilden die ungleich großen Klassenumfänge. Sind die K zu trennenden Klassen in etwa gleich groß, so ergibt sich bei jeder Berechnung einer SVM ein Verhältnis von 1 zu $K-1$.

2.2.2 One-Against-One

Diese auf der Biklassifikation basierende Vorgehensweise ist von *Kreßel* (1999) vorgestellt worden, um die Gebiete, die zu einer nicht eindeutig klassifizierbaren Region gehören, zu entfernen oder zu verkleinern und die Unausgewogenheit der SVM bei der OAA-Trennung zu umgehen.
Für je zwei Klassen wird eine Ebene berechnet, die diese beiden Klassen unabhängig von den übrigen Klassen optimal trennt. Dazu sind $\frac{K(K-1)}{2}$ Paarvergleiche notwendig, was den Aufwand gegenüber dem OAA-Verfahren deutlich erhöht. Dies wird allerdings durch die niedrigere Anzahl an Trainingsbeispielen pro Klasse wieder ausgeglichen (*Schölkopf, Smola* (2002)). Aufgrund der Vorgehensweise wird dieses Verfahren auch One-Against-One (OAO) genannt.
Werden die Klassen k_1 und k_2 verglichen, so muss das folgende Optimierungsproblem gelöst werden:

$$\min \ \frac{1}{2}\|\boldsymbol{w}^{[k_1k_2]}\|^2 + C\sum_{i=1}^{l'}\xi_i^{[k_1k_2]}$$

unter den Nebenbedingungen

$$\boldsymbol{w}^{[k_1k_2]}\boldsymbol{x}_i + b^{[k_1k_2]} \geq 1 - \xi_i^{[k_1k_2]} \qquad \text{, falls } y_i = k_1$$

$$\boldsymbol{w}^{[k_1k_2]}\boldsymbol{x}_i + b^{[k_1k_2]} \leq -1 + \xi_i^{[k_1k_2]} \qquad \text{, falls } y_i = k_2$$

$$\xi_i^{[k_1k_2]} \geq 0,\ i \in \{1, \ldots, l'\},\ k_1, k_2 \in \{1, \ldots, K\},$$

wobei $l' \leq l$ die Anzahl der Vektoren aus den beiden zu betrachtenden Klassen k_1 und k_2 bezeichnet.
Um nun die berechneten Hyperebenen für eine Entscheidungsfunktion zu nutzen, gibt es verschiedene Möglichkeiten. Üblicherweise wird die Voting-Strategy verfolgt (vgl. *Kreßel* (1999)). Dazu wird für einen neu zu klassifizierenden Vektor $\boldsymbol{x}_{neu}$ der Wert jeder der $\frac{K(K-1)}{2}$ Entscheidungsfunktionen $f^{[k_1k_2]}(\boldsymbol{x}_{neu}) = \text{sign}(\boldsymbol{w}^{[k_1k_2]}\boldsymbol{x}_i + b^{[k_1k_2]})$ berechnet. Wird der Vektor zu Klasse k_1 zugeordnet, so erhält Klasse k_1 einen Punkt, wird er durch die Entscheidungsfunktion zu Klasse k_2 zugeordnet, so wird Klasse k_2 ein Punkt zugeschrieben. So werden also $\frac{K(K-1)}{2}$ Punkte auf K Klassen verteilt, wobei $\boldsymbol{x}_{neu}$ nun derjenigen Klasse mit der höchsten Punktanzahl zugeordnet wird. Erhalten zwei Klassen gleich viele Punkte, so ist die Zuordnung zwischen diesen Klassen nicht eindeutig. In diesem Fall wird meist die Klasse mit dem kleineren Index gewählt.
Die gegebene Situation bei drei Klassen ist in Abbildung 2.9 wiedergegeben, wobei die Trennebenen $D^{[12]}$, $D^{[13]}$ und $D^{[23]}$ die Trennung zwischen den jeweils involvierten Klassen vornehmen.

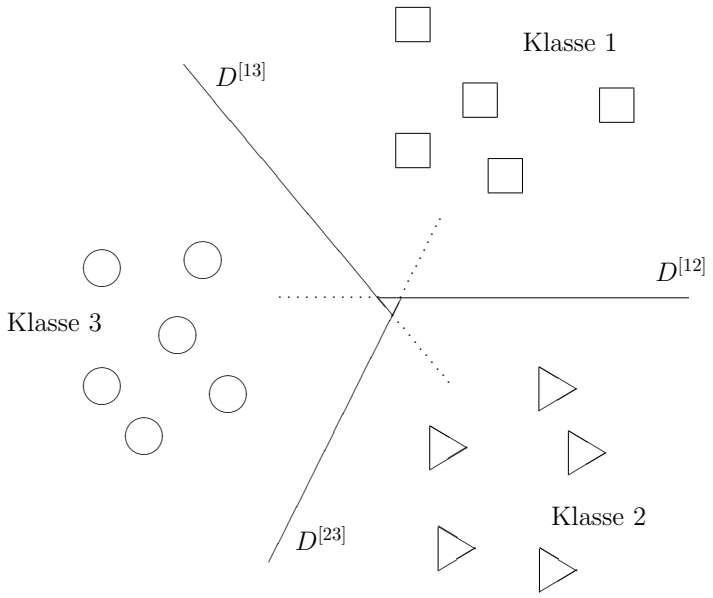

Abbildung 2.9: One-Against-One Trennung bei drei Klassen

Die nicht eindeutig klassifizierbaren Gebiete sind im Vergleich zu Abbildung 2.7 kleiner geworden, existieren aber immer noch. In Abbildung 2.9 wird dieses Gebiet

durch das kleine Dreieck in der Mitte gebildet. Eine eindeutige Zuordnung ist in diesem Fall wiederum nicht ohne weiteres möglich. Jede der drei Klassen würde bei den $\frac{3 \cdot 2}{2} = 3$ zu berechnenden Ebenen jeweils einen Punkt erhalten.

Die beiden in den Absätzen 2.2.1 und 2.2.2 behandelten Verfahren bieten eine einfache Möglichkeit, ohne viel Aufwand eine Multiklassifikation auf Basis von binären Trennungen vorzunehmen. Der ersten Methode liegt das Problem zugrunde, unsymmetrische Klassengrößen zu behandeln, da jeweils $K - 1$ Klassen zu einer zusammengefasst werden. Beide Verfahren haben den Nachteil, dass unscharfe Bereiche unterschiedlicher Größe entstehen können. Um auch diese nicht klassifizierbaren Gebiete zu umgehen, können Directed Acyclic Graph oder Fuzzy-SVM verwendet werden, die in den folgenden Absätzen vorgestellt werden.

2.2.3 Directed Acyclic Graph

Das Directed Acyclic Graph (DAG)-Verfahren wurde von *Platt et al.* (2000) entwickelt. Es beruht auf der Vorgehensweise der One-against-one-Trennung, die mit einem gerichteten Graphen verknüpft wird. Dieser Graph enthält K innere Knoten, an denen jeweils eine SVM berechnet werden muss. Die Endknoten dienen der Klassenzuweisung von Testdaten, die durch den zu erstellenden Graph klassifiziert werden. Das generelle Vorgehen kann an Abbildung 2.10 veranschaulicht werden.

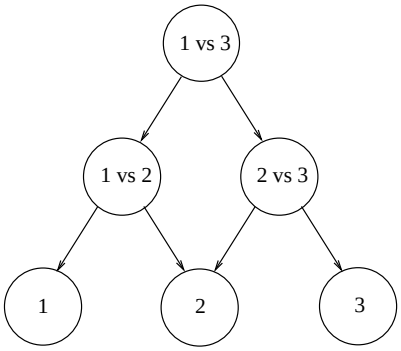

Abbildung 2.10: Vorgehensweise beim DAG-Verfahren

Es soll hierbei wiederum die Vorgehensweise bei der Klassifikation mit drei zugrunde liegenden Klassen verdeutlicht werden. Man startet am oberen Knoten des Graphen und bestimmt auf Basis einer zu berechnenden Ebene die vorläufige Klassenzugehörigkeit des Testdatenpunktes. So wird der Testvektor bei Vorlage von Abbildung 2.10 zu Beginn entweder nicht zu Klasse „1“ oder nicht zu Klasse „3“ zugeordnet. Es erfolgt also keine direkte Klassifikation, sondern es wird nach dem Ausschlussprinzip vorgegangen. Im nächsten Schritt wird die Klassenzugehörigkeit zu den verbleibenden Klassen überprüft. Im Fall $K = 3$ ist dieser Schritt bereits der letzte. Es werden sukzessive Klassen ausgeschlossen, bis letztlich eine übrig bleibt, in die dann der

Testdatenvektor einzuordnen ist. Ein beispielhaftes Vorgehen zeigt Abbildung 2.11, wobei die Beobachtung $\boldsymbol{p}_2$ eine der drei Klassen zugeordnet werden soll.

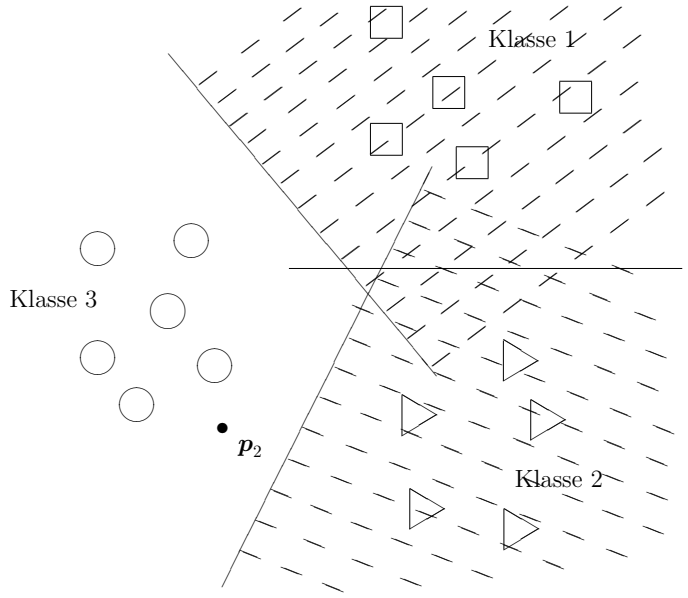

Abbildung 2.11: Klassifikation einer Beobachtung mittels DAG

Wird der Baum aus Abbildung 2.10 gewählt, so wird bei der ersten SVM festgestellt, dass die Beobachtung nicht zu Klasse „1“ gehört und somit nicht in den oberen rechten schraffierten Bereich fällt. Die nächste Überprüfung ergibt, dass ebenfalls keine Zugehörigkeit zu Klasse „2“ vorliegt, die Beobachtung somit in Klasse „3“ klassifiziert wird. Die schraffierten Bereiche werden Schritt für Schritt ausgeschlossen. Hierbei werden die Vorteile von DAG deutlich: zum einen wird die Beobachtung eindeutig der Klasse „1“ zugeordnet und zum anderen wird das Ergebnis der Entscheidungsfunktion, die Klasse „1“ und Klasse „2“ trennt, nicht benötigt. Dieser Vorteil kann insbesondere bei der Analyse von Datensätzen verwendet werden, die auf vielen Klassen basieren, da dann große Teile des Graphen in Abbildung 2.10 nicht benötigt werden. Um diesen Graph zu generieren, sind zunächst allerdings ebenso viele SVM nötig wie beim OAO-Verfahren.
Ein großer Nachteil ist darin zu sehen, dass die letztendliche Klassenzugehörigkeit vom Aufbau des Graphen abhängig ist. Liegt der zu klassifizierende Vektor beispielsweise in dem mittleren Dreieck in Abbildung 2.9, so wird der Vektor bei Vorgabe dieser drei Klassen zu Klasse „2“ zugeordnet, falls das Vorgehen aus Abbildung 2.10 gewählt wird. Werden zu Beginn (im obersten Knoten des Baumes) jedoch die Klassen „1“ und „2“ verglichen, so wird der Vektor letztlich Klasse „3“ zugewiesen.
Mittels Fuzzy-SVM wird versucht, auch diesen Nachteil zu umgehen, und eine eindeutige, nicht von zufälligen Gegebenheiten abhängige Zuordnung zu ermöglichen.

2.2.4 Fuzzy Support Vektor Maschinen

Die Fuzzy-SVM gehen auf *Abe, Inoue* (2002) zurück. Die Idee dieses Ansatzes ist die Reduzierung bzw. Beseitigung der nicht eindeutig zuordenbaren Gebiete bei den binärbasierten Ansätzen der Multiklassifikation. Hier ist nicht mehr die Entscheidungsfunktion von Bedeutung, sondern es werden so genannte Membership-Funktionen eingeführt, mit deren Hilfe eine Klassifikation erfolgen kann. Hierbei kann das Vorgehen entweder auf dem OAA- oder auf dem OAO-Verfahren (*Inoue, Abe* (2001), *Abe, Inoue* (2002)) beruhen.

Aufbauend auf den bei der OAA-Multiklassifikation zu bestimmenden K Ebenen $D^{[k]}$ für $k \in \{1, \ldots, K\}$ werden Membership-Funktionen eingeführt, mit deren Hilfe die Klassenzugehörigkeit in unscharfer, aber eindeutiger Weise bestimmt werden kann. Für eine Klasse k werden folgende Kennzahlen für einen zu klassifizierenden Eingabevektor $\boldsymbol{x}$ berechnet (vgl. *Inoue, Abe* (2001)).
Sei $k, j \in \{1, \ldots, K\}$, dann ist

für $k = j$:

$$m_{kj}(\boldsymbol{x}) = m_{kk}(\boldsymbol{x}) = \begin{cases} 1 & \text{, falls } F^{[k]}(\boldsymbol{x}) > 1 \\ F^{[k]}(\boldsymbol{x}) & \text{, sonst} \end{cases}$$

und für $k \neq j$:

$$m_{kj}(\boldsymbol{x}) = \begin{cases} 1 & \text{, falls } F^{[j]}(\boldsymbol{x}) < -1 \\ -F^{[j]}(\boldsymbol{x}) & \text{, sonst.} \end{cases}$$

Die Werte $F^{[k]}(\boldsymbol{x})$ geben die reellwertige Ausgabe der Entscheidungsfunktion $f^{[k]}$ an der Stelle $\boldsymbol{x}$ an:

$$F^{[k]}(\boldsymbol{x}) := \boldsymbol{w}^{[k]}\boldsymbol{x} + b^{[k]}.$$

Die Bedeutung einer Klasse für den Vektor $\boldsymbol{x}$ bestimmt sich durch:

$$m_k(\boldsymbol{x}) = \min_{j=1,\ldots,K} m_{kj}(\boldsymbol{x}).$$

$m_k(\boldsymbol{x})$ gibt dabei die Zugehörigkeit des Vektors $\boldsymbol{x}$ zur Klasse k an. Je höher dieser Wert ist, desto eher wird der Vektor in die betreffende Klasse eingeordnet.
Die resultierende Situation für den Fall der linearen Trennung von drei Klassen ist in Abbildung 2.12 gegeben. Hierbei werden die durch die SVM mittels OAA-Verfahren ermittelten Trennebenen durch die durchgezogenen Linien veranschaulicht (bezeichnet mit $D^{[1]}$, $D^{[2]}$ und $D^{[3]}$). Durch die Einführung der Membership-Funktionen ergeben sich für jede Klasse Bereiche, in denen der Wert der jeweiligen Membership-Funktion konstant ist. Für die Werte $m_1 \in \{1, 0, -1\}$ sind diese Bereiche für die Klasse 1 zusätzlich eingezeichnet (gepunktete Linien). Je näher dieser Wert an 1 liegt, desto eher wird ein Vektor der betreffenden Klasse zugeordnet. Somit ergeben sich neue Entscheidungsflächen (bzw. -geraden), gekennzeichnet durch die gestrichelten Linien, die eine eindeutige Zuordnung von Vektoren zu den drei Klassen

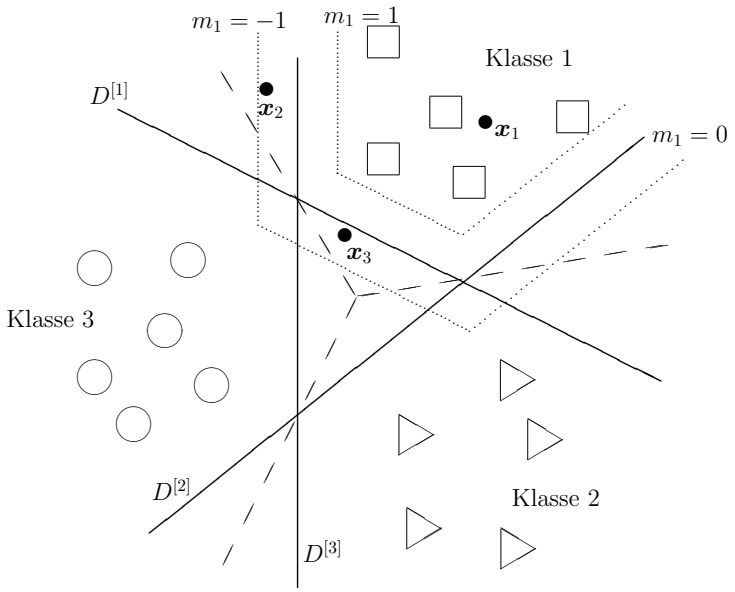

Abbildung 2.12: Aus der Fuzzy Klassifikation resultierende Trennebenen (OAA)

ermöglichen. Unscharfe Bereiche werden vollständig eliminiert. Die Zuordnung eines Vektors $\boldsymbol{x}$ zu einer der K Klassen ergibt sich nun aus

$$y = \arg \max_{k=1,\ldots,K} m_k(\boldsymbol{x}). \tag{2.25}$$

Ist die Zuordnung zu einer Klasse k für einen Punkt $\boldsymbol{x}$ eindeutig, so ergibt sich $m_{kj} = 1$ für alle $k,\ j$. Es resultiert eine eindeutige Zuordnung zu Klasse k, da $\max\ m_k(\boldsymbol{x}) = 1$ gilt.
Für die zusätzlich in Abbildung 2.12 eingefügten Vektoren ergeben sich die in der Tabelle 2.2 zusammengefassten Werte der Membership-Funktionen und die daraus resultierenden Klassenzuweisungen. Dabei werden für jede der drei Beobachtungen zunächst die Membership-Werte, bezogen auf den Vergleich zweier Klassen (m_{ij}), aus Sicht aller drei Klassen bestimmt, bevor daraus der letztendliche Membership-Wert für jede Klasse bestimmt werden kann. Die Zuordnung findet nun entsprechend Vorschrift (2.25) statt. Jeder der in Tabelle 2.2 aufgeführten Vektoren wird Klasse „1“ zugeordnet, allerdings auf Basis sehr unterschiedlicher Membership-Werte. So wird der Vektor $\boldsymbol{x}_1$ klar mit einem Membership von 1 zur Klasse „1“ geordnet. Die beiden anderen Vektoren $\boldsymbol{x}_2$ und $\boldsymbol{x}_3$ erhalten lediglich negative Membership-Werte, die dennoch zu einer Zuordnung zu Klasse „1“ führen, da sie die größten der jeweils erreichten Werte sind. Die Beobachtung $\boldsymbol{x}_2$ erhält mit nur $m_1 = -0,8$ den geringsten Zugehörigkeitswert.
Nach *Abe* (2003) sind die bei Fuzzy-SVM resultierenden Entscheidungsebenen in der OAA-Situation identisch zu denen beim so genannten Tie-breaking. Daher ergibt sich in Abbildung 2.12 ein ähnliches Bild wie in Abbildung 2.8. Es ist erkennbar, dass die Bereiche, in denen ein Vektor nicht eindeutig zuordenbar ist, nicht mehr existieren.

Vektor	Klasse 1	Klasse2	Klasse 3	Zuordnung
$\boldsymbol{x}_1$	$m_{11} = 1$	$m_{21} = -2$	$m_{31} = -2$	
	$m_{12} = 1$	$m_{22} = -2$	$m_{32} = 1$	
	$m_{13} = 1$	$m_{23} = 1$	$m_{33} = -3$	
	$\mathbf{m_1 = 1}$	$m_2 = -2$	$m_3 = -3$	1
$\boldsymbol{x}_2$	$m_{11} = 1$	$m_{21} = -1,5$	$m_{31} = -1,5$	
	$m_{12} = 1$	$m_{22} = -5$	$m_{32} = 1$	
	$m_{13} = -0,8$	$m_{23} = -0,8$	$m_{33} = 0,8$	
	$\mathbf{m_1 = -0,8}$	$m_2 = -5$	$m_3 = -1,5$	1
$\boldsymbol{x}_3$	$m_{11} = -0,4$	$m_{21} = 0,4$	$m_{31} = 0,4$	
	$m_{12} = 1$	$m_{22} = -1,7$	$m_{32} = 1$	
	$m_{13} = 1$	$m_{23} = 1$	$m_{33} = -1,5$	
	$\mathbf{m_1 = -0,4}$	$m_2 = -1,7$	$m_3 = -1,5$	1

Tabelle 2.2: Beispielhafte Berechnung der Membership-Funktionen

Die Zuweisung von Zugehörigkeitswerten erfolgt bei zugrunde liegender OAO-Trennung ähnlich (vgl. *Abe, Inoue* (2002)). Dazu werden für eine Klasse k folgende Werte ermittelt:

$$m_{kk'}(\boldsymbol{x}) = \begin{cases} 1 & \text{, falls } F^{[kk']}(\boldsymbol{x}) > 1 \\ F^{[kk']}(\boldsymbol{x}) & \text{, sonst} \end{cases}$$

mit $k \neq k'$. Die reellen Werte der Entscheidungsfunktion zwischen beiden Klassen k und k' an der Stelle $\boldsymbol{x}$ werden hierbei durch $F^{[kk']}(\boldsymbol{x})$ angegeben. Damit gibt $m_{kk'}(\boldsymbol{x})$ den Zugehörigkeitswert zu Klasse k bei Vergleich der Klassen k und k' an. Die letztendliche Ermittlung der Zugehörigkeit eines Vektors $\boldsymbol{x}$ zu einer Klasse k werden analog zur OAA-Trennung berechnet und entsprechen an dieser Stelle

$$m_k(\boldsymbol{x}) = \min(1, \min_{k' \neq k, k'=1,\ldots,K} F^{[kk']}(\boldsymbol{x})).$$

Eine Beobachtung $\boldsymbol{x}$ wird nun der Gruppe $\arg\max_{k=1,\ldots,K} m_k(\boldsymbol{x})$ zugeordnet. Wenn die Zugehörigkeit eines Vektors zu einer Klasse k bei Vergleich von Klassen k und k' eindeutig nicht vorliegt, so erhält dieser einen niedrigen Wert $m_{kk'}$. Damit ist die Wahrscheinlichkeit, dieser Klasse k zugewiesen zu werden, ziemlich gering, da das Maximum aller ermittelten Membership-Werte verwendet wird. Große Membership-Werte werden hingegen erreicht, wenn sich der betreffende Vektor nicht in einem unscharfen Bereich befindet und die Zuweisung damit eindeutig ist.
Die resultierenden Trennebenen, die sich auf Basis der Fuzzy-SVM bei OAO für den unscharfen Bereich ergeben, zeigt Abbildung 2.13.

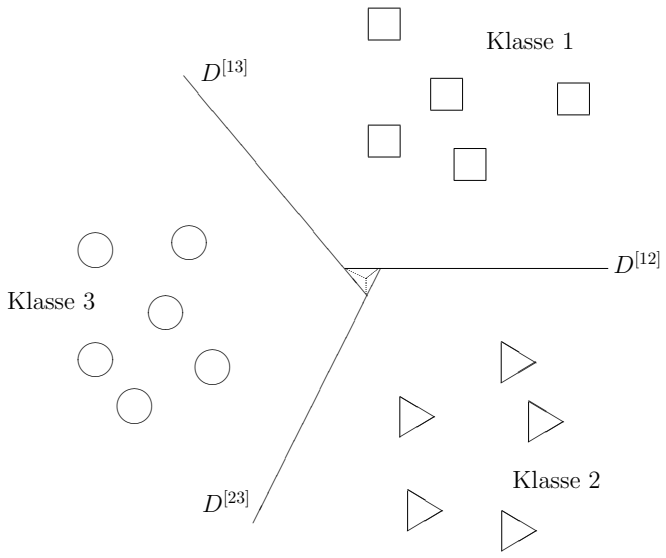

Abbildung 2.13: Aus der Fuzzy Klassifikation resultierende Trennebenen (OAO)

Die Trennebenen ersetzen in diesem Fall die zuvor durch SVM ermittelten Ebenen, die mit $D^{[kk']}$ bezeichnet werden, und sollen zu einer präziseren Zuordnung der Beobachtungen führen. Der unscharfe Bereich in der Mitte der Abbildung wird durch die Einführung der Membership-Werte geschärft, sodass eine eindeutige Zuordnung möglich wird. Dies wird in Abbildung 2.13 durch die Teilung des mittleren Dreiecks in drei Bereiche in Abhängigkeit der Membership-Werte verdeutlicht.
Ob die intuitiv einsichtige Vorgehensweise auch in der Praxis erfolgreich ist, wird in Abschnitt 4.4.2 untersucht.

2.2.5 Error-Correcting Output-Codes

Das Verfahren des Error-Correcting Output-Coding (ECOC) ist von *Dietterich, Bakiri* (1995) entwickelt worden und dient ebenfalls der Multiklassifikation. Dabei wird davon ausgegangen, dass das maschinelle Lernen als ein Prozess der Übertragung von Daten aufgefasst werden kann, bei dem Fehler auftreten können. Um diese Fehler abzufangen, werden Codes eingeführt. Die wesentliche Idee dieses Ansatzes liegt darin, dass die ursprünglich zu trennenden Klassen jeweils einen Code zugewiesen bekommen, der aus $L \in \mathbb{N}$ Codewörtern besteht. Die Codewörter erhalten Einträge mit Ausprägungen 1 oder 0. Für jedes Codewort von der Länge der Anzahl der Klassen wird nun eine SVM trainiert.

Tabelle 2.3 gibt eine mögliche Situation wieder. Es werden sechs Codewörter verwendet, um die vier Klassen zu beschreiben. Jedes Codewort kann als eine zusätzliche Eigenschaft verstanden werden, die gleich mehrere Klassen auszeichnet. Wird beispielsweise das erste Codewort a_1 betrachtet, so gibt Tabelle 2.3 an, dass diese Eigenschaft von Klasse 1 und 2 erfüllt wird, von den Klassen 3 und 4 hingegen

Klasse	Codewörter					
	a_1	a_2	a_3	a_4	a_5	a_6
1	1	0	0	1	0	1
2	1	1	1	0	0	0
3	0	1	0	0	1	1
4	0	0	1	1	1	0

Tabelle 2.3: Exemplarische Codematrix für $K = 4$ Klassen und $L = 6$ Codewörter

nicht. Bei jeder der L zu berechnenden SVM werden nun diejenigen Klassen, die die Eigenschaft des Codewortes a_i haben, von den übrigen Klassen getrennt. Anders ausgedrückt bedeutet dies, dass jede SVM eines der Codewörter lernen muss. In der in Tabelle 2.3 beschriebenen Situation bedeutet dies für das erste Codewort a_1, dass die zu bestimmende Entscheidungsfunktion f_{a_1} folgendes Kriterium erfüllen muss[9]:

$$f_{a_1}(\boldsymbol{x}) = \begin{cases} 1 & \text{, falls } \boldsymbol{x} \in \text{ Klasse 1 oder 2} \\ 0 & \text{, sonst.} \end{cases}$$

An dieser Stelle sei angemerkt, dass jedes Codewort unterschiedliche Ausprägungen haben muss, was bedeutet, dass die Spalten paarweise disjunkt sind. Das gleiche muss auch für die Zeilen gelten. Anderenfalls könnten die Klassen nicht voneinander unterschieden werden. Für einen neu zu klassifizierenden Vektor muss der Wert jeder der L ermittelten SVM berechnet werden, sodass für diesen Eingabevektor ein Code der Länge L resultiert, der im Idealfall exakt einem Code einer der K Klassen entspricht, sodass er eindeutig zugeordnet werden kann.
Ist dies nicht der Fall, so kommt die Hamming-Distanz $H(\cdot,\cdot)$ zum Einsatz. Sind $A = (A_1, \ldots, A_L)$ und $B = (B_1, \ldots, B_L)$ binäre Codes der Länge L mit $A_i, B_i \in \{0, 1\}$, dann gilt $H(A, B) = \sum_{i=1}^{L} |A_i - B_i|$. Diese Distanz berechnet den Abstand zweier Codes und gibt an, an wie vielen Stellen ein Code geändert werden müsste, damit die beiden Codes übereinstimmen. Ein Vektor wird der Klasse mit dem geringsten Abstand zugewiesen. Der Code, der sich für den zu klassifizierenden Vektor ergibt, gibt demnach an, welche der L Eigenschaften dieser Vektor enthält.

Die Anzahl der zu berechnenden SVM entspricht genau der Anzahl der vom Benutzer frei zu wählenden Codewörter. *Dietterich, Bakiri* (1995) geben an, wie die Codes in Abhängigkeit der Anzahl der Klassen bestimmt werden können, wobei sie darauf hinweisen, dass die Bestimmung einer Methode zur Festlegung der Codewörter in Abhängigkeit der Anzahl der Klassen ein noch offenes Problem darstellt. Bei K Klassen gibt es bei einer binären Codierung 2^K mögliche Codewörter. Die Hälfte bilden komplementäre Codewörter, die keine zusätzliche Information enthalten. Weiterhin gibt es ein Codewort, was nur aus 0 bzw. nur aus 1 besteht und somit nicht zur Diskriminierung der Klassen beitragen kann. Somit resultiert eine maximale, noch sinnvolle Codelänge von $2^{K-1} - 1$ Einträgen (*Dietterich, Bakiri*

[9]Die Zuweisung der Klasse „-1“ wie bisher wird hierbei durch die Zuweisung von „0“ ersetzt, was lediglich kennzeichnende Aufgaben erfüllt.

(1995)). Sollen beispielsweise handgeschriebene Buchstaben erkannt werden, so liegt für jeden Buchstaben eine Klasse vor, die von den übrigen separiert werden muss. Demnach könnten nach obigem Vorgehen in diesem Fall $2^{26-1} - 1 = 33.554.431$ Codewörter bestimmt werden. Da dies aufgrund der Rechenzeit keine akzeptable Lösung darstellt, besteht die Möglichkeit, den Codewörtern eine inhaltliche Bedeutung zuzuweisen. So könnte bei der Erkennung von Buchstaben ein Codewort das Vorliegen bestimmter Eigenschaften eines Buchstabens angeben, wie etwa das Auftreten eines horizontalen, vertikalen oder diagonalen Strichs, einer geschlossenen oder zu einer bestimmten Seite offenen Kurve. Die Codes für eine Klasse ergeben sich damit auf natürliche Weise.[10] Dadurch kann die Anzahl der Codewörter und damit die Anzahl der zu berechnenden SVM drastisch reduziert werden. Um die Trennfähigkeit zwischen den Klassen zu gewährleisten, sollten die gewählten Codewörter weiterhin einen bestimmten Mindestabstand, gemessen durch die Hamming-Distanz, nicht unterschreiten.

Der in *Kikuchi, Abe* (2003) vorgenommene Vergleich der beiden in Abschnitt 2.2.4 und 2.2.5 vorgestellten Methoden zeigt, dass ECOC gegenüber Fuzzy SVM (OAA) häufig keine signifikante Verbesserung der Ergebnisse liefert, sodass diese beiden Methoden als gleichwertig anzusehen sind. Eine Verbesserung der Ergebnisgüte von ECOC ist nach *Kikuchi, Abe* durch eine verbesserte Struktur der Codewörter möglich.

2.2.6 Direkte Klassifikation

Die bisher vorgestellten Multiklassifikationsverfahren basieren alle auf der Klassifikation von zwei Klassen, die kombiniert werden und eine Zuordnung zu $K > 2$ Klassen ermöglichen. Bereits *Vapnik* (1998) stellte eine Methode vor, die eine direkte Betrachtung aller Datenvektoren, insbesondere aller Klassen, in einem einzigen Optimierungsproblem ermöglicht. Die resultierende Entscheidungsbasis hat allerdings eine ähnliche Form wie die bei der One-against-All Vorgehensweise (vgl. Gleichung (2.24)). Der wesentliche Unterschied ist in dem zu lösenden Optimierungsproblem zu sehen, welches alle Klassen auf einmal betrachtet. Nach *Weston, Watkins* (1999) muss dazu folgender Ausdruck minimiert werden:

$$\min \frac{1}{2} \sum_{k=1}^{K} ||\boldsymbol{w}^{[k]}||^2 + C \sum_{i=1}^{l} \sum_{k \neq y_i} \xi_i^{[k]}$$

unter den Nebenbedingungen

$$\boldsymbol{w}^{[y_i]} \boldsymbol{x}_i + b^{[y_i]} \geq \boldsymbol{w}^{[k]} \boldsymbol{x}_i + b^{[k]} + 2 - \xi_i^{[k]} \quad \text{mit} \quad \xi_i^{[k]} \geq 0,$$

wobei $k \in \{1, \ldots, K\} \backslash y_i$ mit $y_i \in \{1, \ldots, K\}$. Die Entscheidungsfunktion im Falle der direkten Klassifikation lautet

$$f(\boldsymbol{x}) = \arg \max_k (\boldsymbol{w}^{[k]} \boldsymbol{x} + b^{[k]})$$

[10] Beispielsweise kann der Buchstabe „L“ durch das Aufweisen eines vertikalen und eines horizontalen Strichs charakterisiert werden.

mit $k \in \{1, \ldots, K\}$.
Außer diesem direkten Ansatz gibt es weitere Möglichkeiten zur Multiklassifikation, die unter anderem in *Crammer, Singer* (2001) oder *Hsu, Lin* (2002) vorgestellt werden, auf die an dieser Stelle aus praxisrelevanten Gründen nicht weiter eingegangen werden soll.

Alles in allem gibt es keine Methode zur Multiklassifikation, die die übrigen in der Trainingszeit und an der Generalisierungsfähigkeit übertrifft. Bei der Wahl eines Verfahrens spielen mehrere Faktoren wie etwa die Anzahl der Klassen oder die Anzahl an Beobachtungen eine entscheidende Rolle, sodass die Auswahl einer Methode individuell getroffen werden muss. Auf Basis ausführlicher Auswertungen von zehn frei zugänglichen und bereits untersuchten Datensätzen kommen *Hsu, Lin* (2002) zu dem Schluss, dass das OAO-Verfahren oder DAG-SVM gegenüber OAA- oder direkten Multiklassifikationsverfahren in praktischen Anwendungen aufgrund der Trainingszeit vorzuziehen ist[11].
Die Schwierigkeiten der Biklassifikation treten bei der Multiklassifikation erneut auf. So gilt es auch hier, Kerne sowie Parameterkonstellationen zu finden, sodass eine gute Prognose ermöglicht wird. Hierbei tritt die zusätzliche Schwierigkeit auf, dass für jede der eingesetzten SVM bei Multiklassifikationsverfahren, die auf Binärklassifikation beruhen, der gleiche Kern mit den gleichen Einstellungen eingesetzt wird. Weiterhin sei auf die Bestimmung der Klassen hingewiesen. Im Gegensatz zu Segmentierungsverfahren können bei der Klassifikation keine Gruppen innerhalb der Daten erkannt werden. Eine eher schlechte Prognose bei der Multiklassifikation könnte daher darauf hindeuten, dass die zugrunde liegende Gruppierung der Daten nicht adäquat ist, sondern evtl. neue Klassen gebildet werden sollten.

2.3 Weitere kernbasierte Methoden

Innerhalb der SVM ist nur die Eigenschaft der Darstellung des Optimierungsproblems mit Hilfe von Skalarprodukten dafür entscheidend, dass die nicht lineare Trennung mittels Kernen eingeführt werden kann. Es stellt sich die Frage, ob dies auch bei anderen linearen Methoden umsetzbar ist. Solange eine Methode nur Skalarprodukte berechnet, und die Eingabedaten nicht in einer anderen Weise in den Berechnungen verknüpft sind, kann aus diesen Methoden durch Einsatz von Kernen eine nicht lineare Form erzeugt werden. Obwohl diese Eigenschaft bekannt war, ist diese mit Ausnahme der SVM im Bereich des maschinellen Lernens erst genutzt worden (*Schölkopf et al.* (1999c)), als die nicht lineare Variante der Hauptkomponentenanalyse und der Diskriminanzanalyse entwickelt wurden.
Im Folgenden soll kurz auf diese beiden modifizierten Verfahren eingegangen werden, auch wenn sie nicht Gegenstand der empirischen Untersuchungen sind. Die Hauptkomponentenanalyse (im Folgenden PCA genannt (Principal Component Analy-

[11] Fuzzy-SVM und ECOC wurden bei diesem Vergleich allerdings nicht betrachtet.

sis)) wird seit langer Zeit im Bereich des Marketing eingesetzt, um Wirkungszusammenhänge bei Vorliegen vieler Variablen auf die wichtigsten Komponenten zurückzuführen. Ebenso wie die PCA gehört auch die Diskriminanzanalyse (im Folgenden mit LDA abgekürzt (Lineare Diskriminanzanalyse)) zum traditionellen Instrumentarium im Bereich des Marketing. Daher kann bei beiden Verfahren die kernbasierte Erweiterung eine sinnvolle Ergänzung zur Behandlung nicht linearer Strukturen innerhalb von Marketingdaten darstellen.

2.3.1 Kern-Hauptkomponentenanalyse

Ziel der PCA ist es, die Varianz innerhalb vorliegender Daten zu erklären. Dazu werden Hauptkomponenten aus den ursprünglichen Daten extrahiert, die die größte Varianz erklären, anhand derer dann die Daten neu im Raum positioniert werden können, um beispielsweise Ähnlichkeiten von Objekten zu erkennen oder um eine Reduktion der Dimension vorzunehmen. Der Unterschied der von *Schölkopf* (1997) vorgestellten Kern-PCA zur linearen PCA ist nun, dass die eigentliche Analyse im höher dimensionalen Raum durchgeführt wird, ähnlich wie bei SVM. Die Extraktion von Hauptkomponenten in diesem Raum entspricht nicht linearen Hauptkomponenten im Eingaberaum. So können auch nicht lineare Strukturen innerhalb der Daten extrahiert werden. Um analog zu SVM vorgehen zu können, wird eine Darstellung des innerhalb der linearen PCA zu lösenden Problems gesucht, in der die Trainingsdaten nur mittels Skalarprodukten miteinander verknüpft sind. Das bei herkömmlicher PCA zu lösende Eigenwert-Problem stellt sich dar als

$$\lambda \boldsymbol{v} = \boldsymbol{C}\boldsymbol{v} \tag{2.26}$$

für $\lambda \geq 0$ und nicht negative Eigenvektoren $\boldsymbol{v} \in \mathbb{R}^n \backslash \boldsymbol{0}$. Hierbei ist $\boldsymbol{C} = \frac{1}{l}\sum_{i=1}^{l} \boldsymbol{x}_i \boldsymbol{x}_i^T$ die Kovarianzmatrix der Eingabevektoren $\boldsymbol{x}_i$, die hierbei unabhängig von ihrer Klassenzugehörigkeit untersucht werden. Es liegt daher ein unüberwachtes Lernverfahren vor.
Bei der nicht linearen Form der PCA werden die Daten in einen höher dimensionalen Raum abgebildet, in dem die Faktorbestimmung durchgeführt wird (vgl. *Schölkopf et al.* (1999c)). Gleichung (2.26) wird dann geschrieben als

$$\lambda \tilde{\boldsymbol{v}} = \tilde{\boldsymbol{C}}\tilde{\boldsymbol{v}} = \frac{1}{l}\sum_{j=1}^{l} (\Phi(\boldsymbol{x}_j)\tilde{\boldsymbol{v}})\Phi(\boldsymbol{x}_j),$$

wobei $\tilde{\boldsymbol{C}} = \frac{1}{l}\sum_{j=1}^{l} \Phi(\boldsymbol{x}_j)\Phi(\boldsymbol{x}_j)^T$ die Kovarianzmatrix in diesem Raum bildet. Da der Eigenvektor $\tilde{\boldsymbol{v}}$ hier in der lineare Hülle der Vektoren $\Phi(\boldsymbol{x}_j)$ liegt, ergibt sich mit $\tilde{\boldsymbol{v}} = \sum_{j=1}^{l} \alpha_j \Phi(\boldsymbol{x}_j)$ das äquivalente Gleichungssystem

$$\lambda \Phi(\boldsymbol{x}_i)\tilde{\boldsymbol{v}} = \Phi(\boldsymbol{x}_i)\tilde{\boldsymbol{C}}\tilde{\boldsymbol{v}}, \quad \forall\, i \in \{1, \ldots, l\}.$$

Durch Einsatz eines Kerns und der dazugehörigen Kernmatrix
$\boldsymbol{K}_{ij} = (\Phi(\boldsymbol{x}_i)\Phi(\boldsymbol{x}_j)) = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ resultiert in gekürzter Schreibweise

$$l\lambda\boldsymbol{\beta} = \boldsymbol{K}\boldsymbol{\beta}.$$

Das zu lösende Eigenwertproblem kann somit mit Hilfe des „Kerneltricks“ in eine nicht lineare Variante umgewandelt werden. Dabei ist $\boldsymbol{\beta}$ ein Vektor mit Einträgen β_i für $i \in \{1, \ldots, l\}$. Es resultieren die Eigenwerte $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_l$ von $\boldsymbol{K}$ und Eigenvektoren $\boldsymbol{\beta}^{[1]}, \ldots, \boldsymbol{\beta}^{[l]}$, die nicht linear im Eingaberaum sind und zu den Eigenvektoren $\tilde{\boldsymbol{v}}$ im höherdimensionalen Raum korrespondieren. Um die Projektionen eines Vektors $\boldsymbol{x}$ auf einen Eigenvektor $\tilde{\boldsymbol{v}}^{[m]}$ in diesem Raum zu erhalten, ist wiederum der Einsatz der Kernfunktion bei

$$\tilde{\boldsymbol{v}}^{[m]}\Phi(\boldsymbol{x}) = \sum_{i=1}^{l} \beta_i^{[m]}\Phi(\boldsymbol{x}_i)\Phi(\boldsymbol{x}) = \sum_{i=1}^{l} \beta_i^{[m]} K(\boldsymbol{x}_i, \boldsymbol{x})$$

erforderlich.
Somit wird durch den „Kerntrick“ eine Extraktion nicht linearer Hauptkomponenten aus vorgegebenen Daten ermöglicht. Weitere Ausführungen dazu sind zum Beispiel in *Schölkopf, Smola* (2002) dargestellt.

2.3.2 Kern-Fisher-Diskriminanzanalyse

Die Kern-Fisher-Diskriminanzanalyse (KFD) ist eine von *Mika et al.* (1999) entwickelte nicht lineare Variante der linearen Fisher Diskriminanzanalyse und arbeitet analog zu SVM auch mit Kernen. Das ursprüngliche lineare Optimierungsproblem lautet:

$$\max J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_W \boldsymbol{w}},$$

wobei $\boldsymbol{w}$ der zu bestimmende Gewichtsvektor ist und mit $\boldsymbol{S}_B$ bzw. $\boldsymbol{S}_W$ die Zwischen-Klassen-Varianz bzw. die Inner-Klassen-Varianz bezeichnet wird. Diese Werte werden berechnet durch $\boldsymbol{S}_B = (\bar{\boldsymbol{x}}^{[1]} - \bar{\boldsymbol{x}}^{[2]})(\bar{\boldsymbol{x}}^{[1]} - \bar{\boldsymbol{x}}^{[2]})^T$ und $\boldsymbol{S}_W = \sum_{k=1}^{2} \sum_{i=1}^{l^{[k]}} (\boldsymbol{x}_i^{[k]} - \bar{\boldsymbol{x}}^{[k]})(\boldsymbol{x}_i^{[k]} - \bar{\boldsymbol{x}}^{[k]})^T$, wobei $l^{[k]}$ die Anzahl der Beobachtungen in Klasse $k \in \{1, 2\}$ angibt und $\bar{\boldsymbol{x}}^{[k]} = \frac{1}{l^{[k]}} \sum_{j=1}^{l^{[k]}} \boldsymbol{x}_j^{[k]}$ ist. Ziel dieser Optimierung ist es, einen Richtungsvektor $\boldsymbol{w}$ zu finden, der die Varianz zwischen den gegebenen Klassen maximiert und gleichzeitig die Varianz innerhalb der Klassen minimiert.

Um eine nicht lineare Richtung zu bestimmen, werden die Trainingsdaten mittels einer Abbildung Φ in einen höher dimensionalen Raum abgebildet und die auftretenden Skalarprodukte durch Kernfunktionen ersetzt. Der Richtungsvektor $\boldsymbol{w}$ wird ähnlich zu SVM ausgedrückt durch

$$\boldsymbol{w} = \sum_{i=1}^{l} \alpha_i \Phi(\boldsymbol{x}_i),$$

mit $\alpha_i \in \mathbb{R}$. Verwendet man zusätzlich die Abbildung der Eingabedaten in den höher dimensionalen Raum und die dortigen Berechnungen der Mittelwerte $\boldsymbol{m}_k$ der Klassen, so erhält man durch $\boldsymbol{w}\boldsymbol{m}_k$ einen Ausdruck, in dem das auftretende Skalarprodukt durch eine Kernfunktion ersetzt werden kann. Mittels mehrerer Umformungen

gelangt man zu dem zu maximierenden Rayleigh-Koeffizient[12]:

$$\max J(\boldsymbol{\varphi}) = \frac{\boldsymbol{\varphi}^T \boldsymbol{M} \boldsymbol{\varphi}}{\boldsymbol{\varphi}^T \boldsymbol{N} \boldsymbol{\varphi}},$$

wobei die Matrizen $\boldsymbol{M}$ und $\boldsymbol{N}$ nun von den Kernfunktionen der Eingabedaten abhängen und φ der zu bestimmende Vektor ist.

Wichtig ist hierbei, dass innerhalb der beiden Verfahren SVM und KFD verschiedene Ziele verfolgt werden. Die KFD zielt auf die Maximierung der Zwischen-Klassen-Varianzen, während die SVM das Ziel der Spannenmaximierung zwischen den beiden Klassen verfolgt. Die Mittelwerte spielen bei letzterem Ansatz keine Rolle, was in der Fokussierung auf die Support Vektoren resultiert. Dennoch liegen hier Verfahren vor, die bei Klassifikationsproblemen alternativ verwendet werden können.

2.4 Zusammenfassung

Im Folgenden wird ein kurzer Überblick über die vorgestellten Grundlagen gegeben, auf denen die in Kapitel 3 angegebenen Erweiterungen basieren, wobei sowohl die Bi- als auch die Multiklassifikation von Bedeutung sein wird. Die Ausführungen beschränken sich lediglich auf SVM. Die weiteren kernbasierten Verfahren liefern Ansatzpunkte für weitere Forschungsarbeiten.

Da nicht davon ausgegangen werden kann, dass reale Datensätze fehlerfrei separiert werden können, wird im Folgenden die Trennung mit Kostenparameter C aus Abschnitt 2.1.3 vorgenommen, der die Möglichkeit der Gewichtung einer möglichst fehlerfreien Trennung und der Maximierung der Spanne eröffnet. Es wird sowohl die lineare als auch die nicht lineare Trennung verwendet. Beim Einsatz von Kernfunktionen zur nicht linearen Trennung kommen die Radialbasis-Funktion sowie vereinzelt auch das Polynom zum Einsatz. Daher ist neben dem Kostenparameter C auch der Kern sowie ein Kernparameter in den Anwendungen im Vorhinein zu wählen. Bei Einsatz des neuronalen Netz muss die Eigenschaft berücksichtigt werden, dass nur bestimmte Kombinationen der beiden zu wählenden Parameter einen zulässigen Kern bilden. Da dieser Kern daher in realen betriebswirtschaftlichen Anwendungen eher selten verwendet werden wird, wird von einem Einsatz in dieser Arbeit abgesehen.
Multiklassifikationsprobleme werden herkömmlicherweise mit zwei Verfahren gelöst: OAO und OAA. Diese werden auch in dieser Arbeit hauptsächlich verwendet und durch die in Tabelle 2.4 dokumentierten Verfahren stellenweise ergänzt. Die Auswahl mehrerer Verfahren zur Multiklassifikation beruht dabei auf dem Bestreben nach einer maximalen Trefferquote. Außerdem sollen Fuzzy-SVM dazu eingesetzt

[12]Die exakte Herleitung dieses Koeffizienten kann u.a. in *Mika et al.* (1999) oder *Mika* (2002) nachvollzogen werden.

Methode	Charakteristika
OAO	$\frac{K(K-1)}{2}$ SVM benötigt balancierte Auswertung uneindeutige Zuordnung von Vektoren Entscheidung auf Basis von vergebenen Punkten
OAA	K SVM benötigt nicht balancierte Auswertung uneindeutige Zuordnung von Vektoren Entscheidung auf Basis der Entscheidungswerte
DAG	$\frac{K(K-1)}{2}$ SVM benötigt balancierte Auswertung eindeutige Zuordnung der Vektoren Entscheidung abhängig von der gewählten Baumstruktur
Fuzzy-SVM (OAA)	K SVM benötigt nicht balancierte Auswertung eindeutige, aber unscharfe Zuordnung der Vektoren Entscheidung auf Basis der Membership-Werte
Fuzzy-SVM (OAO)	$\frac{K(K-1)}{2}$ SVM benötigt balancierte Auswertung eindeutige, aber unscharfe Zuordnung der Vektoren Entscheidung auf Basis der Membership-Werte
ECOC	Codelänge bestimmt die Anzahl der benötigten SVM balancierte Auswertung, falls Codewörter so gewählt eindeutige Zuordnung der Vektoren Entscheidung auf Basis der Hamming-Distanz

Tabelle 2.4: Charakteristika der einzusetzenden Multiklassifikationsverfahren

werden, die Interpretation der Ergebnisse zu bereichern. Es werden somit mehrere Eigenschaften ausgenutzt, um den im Marketing gesetzten Zielen möglichst effektiv nachzukommen.
Die Methoden unterscheiden sich durch verschiedene Aspekte. Zum einen variiert die Anzahl der für die Trennung nötigen und zu trainierenden SVM, die bei zugrunde liegendem OAO-Verfahren, wie auch bei DAG- oder Fuzzy-SVM, mit $\frac{K(K-1)}{2}$ am höchsten ist. K gibt die Anzahl der Klassen an. Bei ECOC ist diese Anzahl abhängig von dem gewählten Design der Codematrix. Da bei OAA nur jeweils eine Klasse vom Rest getrennt wird, unterscheiden sich die Umfänge der eingehenden Klassen deutlich voneinander. Hieraus ergibt sich ein nicht balanciertes Problem. Je nach eingesetztem Verfahren unterscheiden sich auch die Vorgehensweisen, nach denen die Klassenzugehörigkeit eines Vektors bestimmt wird. Soweit möglich wird auf Basis der Entscheidungswerte entschieden, wie bei OAA oder abgeändert auch bei Fuzzy. Können die Entscheidungswerte allein nicht herangezogen werden, wie bei OAO, DAG oder ECOC, so werden alternative Formen der Zuweisung herangezogen. Je nach eingesetztem Verfahren ergeben sich Entscheidungen, die

in Kapitel 3 näher untersucht und erweitert sowie in Kapitel 4 auf reale Daten angewendet werden sollen.

Kapitel 3

Anwendungsbezogene Aspekte

Ziel dieses Kapitels ist die Eröffnung eines Anwenderzugangs. Es wird gezeigt, wie sich die in Kapitel 2 vorgestellte, grundlegende Methodik der SVM modifizieren lässt, um eine im Sinne des beabsichtigten Untersuchungsziels sinnvolle Anwendung zu ermöglichen. Weiterhin werden die mittels SVM erzeugbaren Ergebnisse kritisch durchleuchtet und wichtige Erweiterungen vorgestellt. Es lassen sich mehrere Anwendungsgebiete von SVM innerhalb des Marketings finden, wie z.B. in Abschnitt 4.2 ausgeführt wird. Hier soll jedoch das Paradeanwendungsgebiet - die Kundenklassifikation - zugrunde gelegt werden, da in diesem Bereich eine Vielzahl von Aspekten verdeutlicht werden können. Das bei der Kundenklassifikation intendierte Ziel ist die Zuordnung von (potenziellen Neu-)Kunden zu a priori definierten Kundenklassen, um ihnen aufgrund bestimmter, die jeweilige Gruppe auszeichnender Eigenschaften eine besondere Behandlung zukommen zu lassen. Dies kann z.B. in der Bindung umsatzstarker Kunden an das Unternehmen durch intensive Kundenbetreuung bestehen. Die Kunden eines Unternehmens teilen sich bei Betrachtung des Umsatzes oder der Intensität der Kundenbeziehung in natürlicher Weise in verschiedene Bereiche. So bildet der Kundenstamm eines Unternehmens in sich keine homogene Menge, sondern untergliedert sich in unterschiedliche Segmente, die es in adäquater Weise zu behandeln gilt. Bei der Kundenklassifikation müssen derartige Segmente zunächst aufgedeckt bzw. definiert werden, um dann eine Zuordnung (von neuen Kunden beispielsweise) mittels SVM vornehmen zu können. Eine solche Zuordnung kann durch eine angemessene Einteilung des Marketingbudgets auf den Kundenstamm motiviert sein, um eine effiziente Allokation der Marketingressourcen zu gewährleisten.
Die in Abbildung 3.1 enthaltenen Aspekte, die die Qualität einer Methode zu Klassifikation im Marketing u.a. beeinflussen, werden dazu in den folgenden Abschnitten thematisiert.

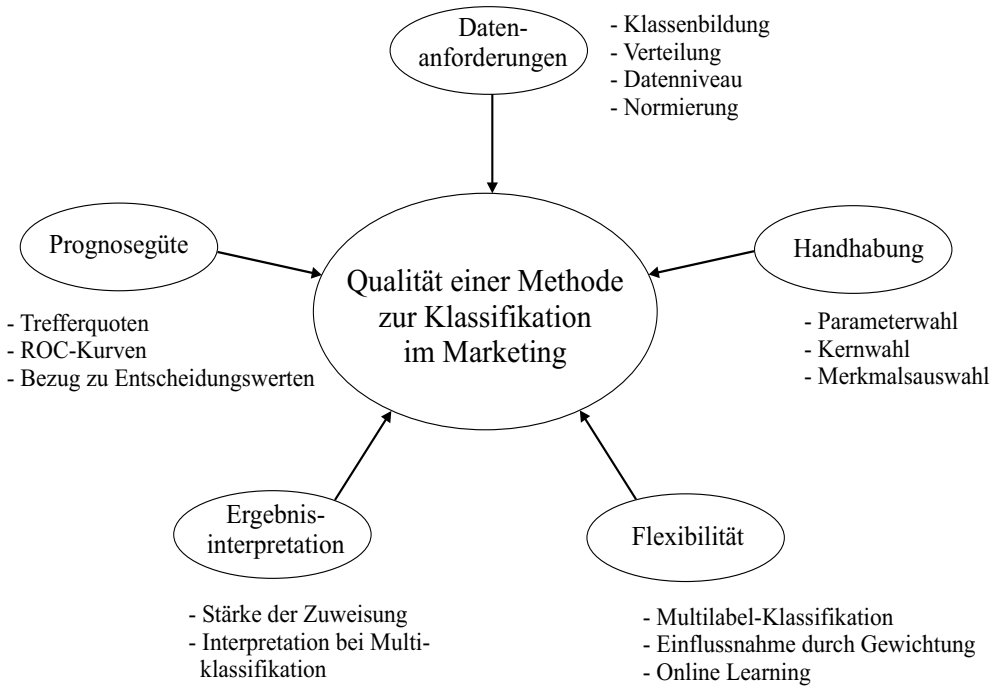

Abbildung 3.1: Die Qualität eines Klassifikationsinstrumentes charakterisierende Eigenschaften

Die behandelten Punkte umfassen neben der Diskussion der bei der Anwendung von SVM auftretenden Datenproblematik (Abschnitt 3.1) die Handhabung und die Flexibilität der Methodik. Die Handhabung betrifft bei SVM insbesondere die Wahl der Kerne und die damit verbundene Festlegung der Parameter. Beides sollte möglichst intuitiv und einfach gestaltet sein. Um die Datengrundlage zu minimieren, können ebenfalls Hinweise zur Reduktion von Merkmalen hilfreich sein. Diese Aspekte werden in den Abschnitten 3.3 und 3.5 thematisiert. Die Flexibilität beinhaltet neben der Berücksichtigung der Multiklassifikation, deren Ansätze bereits in Kapitel 2 vorgestellt wurden, auch die Möglichkeit zur Zuweisung mehrerer Klassen (die so genannte Multilabel-Klassifikation (vgl. Abschnitt 3.6)). Ebenso sollte die Einflussnahme auf die Trennung der Klassen ermöglicht sein. Dies umfasst die Gewichtung von Klassen und Merkmalen, die in Abschnitt 3.2 behandelt werden. Die Anpassung an unterschiedliche Gegebenheiten, wie etwa das Online Learning (Abschnitt 3.4), ist ebenso wünschenswert. Ist die optimale Trennebene berechnet worden, so ist insbesondere im Marketing die Auswertung der generierten Ergebnisse von Interesse, was Inhalt von Abschnitt 3.7 ist. Zu den wichtigsten Kriterien bei der Wahl eines Klassifikationsinstrumentes zählt die erzielte Prognosegüte, die in Abschnitt 3.8 behandelt wird. Dabei wird u.a. auf Trefferquoten, ROC-Kurven und die Bewertung des Ergebnisses anhand der Entscheidungswerte eingegangen.
Die in Abbildung 3.1 aufgeführten Aspekte werden fast alle anschließend in Kapitel 4, dem empirischen Teil der Arbeit, angewendet.

3.1 Anforderung an die Datengrundlage

Mit dem Einsatz von SVM wird die Möglichkeit geboten, a priori klassifizierte Daten zu trennen und somit die Basis zur Prognose von Zugehörigkeiten nicht klassifizierter Beobachtungen zu bilden. Dazu müssen einige Anforderungen an die Datengrundlage gestellt werden. Diese umfassen die generelle Struktur der Daten sowie das Datenniveau, die Behandlung fehlender Werte und die Merkmalsausprägungen. Diese Aspekte werden im Folgenden erörtert.

Der Untersuchungsgegenstand muss so gestaltet sein, dass sich eine sinnvolle Klasseneinteilung innerhalb der Daten ergibt. Die vom Benutzer vorzugebende Einteilung der Daten in mehrere Klassen kann auf unterschiedliche Weise vorgenommen werden. Zum einen ist eine manuelle Einteilung auf Basis vorhandenen Wissens denkbar. So ist es möglich, dass sich bei der Kundenklassifikation eine Einteilung des vorhandenen Kundenstammes in mehrere Klassen je nach intendiertem Untersuchungsziel ergibt. Bei der Biklassifikation könnte dies in einer durch den Umsatz natürlich gegebenen Einteilung in wichtige und unwichtige Kunden bestehen. Einen weiteren möglichen Ansatz zur Einteilung der Kunden liefert die ABC-Analyse, bei der die Kunden auf Basis des erzielten Umsatzes zu den drei Klassen A-, B- und C-Kunden zugeteilt werden (vgl. etwa *Krafft, Albers* (2000)).
Alternativ kann der vorangehende Einsatz von Segmentierungsverfahren dazu dienen, in einer Menge von nicht klassifizierten Beobachtungen möglichst homogene Segmente zu identifizieren. Hierzu zählt vor allem die Clusteranalyse, aber auch neuronale Netze in Form von selbstorganisierenden Karten (SOM) können diesem Ziel gerecht werden. Mittels Entscheidungsbäumen wie dem CHAID-Algorithmus (*Temme* (2002)) kann sowohl eine Zuordnung der Beobachtung in einzelne Segmente als auch eine Beschreibung derselbigen anhand der vorliegenden Merkmale erreicht werden, was zu einem besseren inhaltlichen Verständnis der gewonnenen Einteilung beitragen kann (vgl. *Monien, Decker* (2004a)). Dies ist bei SOM beispielsweise nicht gegeben und eine Charakterisierung der Segmente muss manuell vorgenommen werden.
Der allgemeine Ablauf eines Segmentierungs- bzw. Klassifikationsprozesses bei der Zuordnung eines Kunden zu einem Kundensegment respektive einer Klasse ist in Abbildung 3.2 dargestellt.

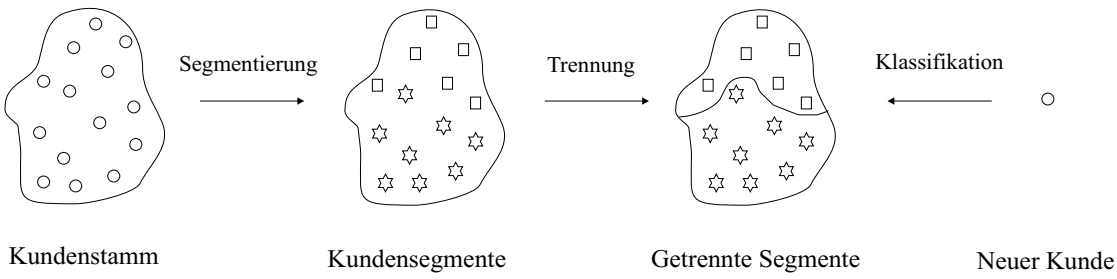

Abbildung 3.2: Ablauf der Schritte bei der Kundenklassifikation

Der Kundenstamm bildet die Ausgangsbasis, die es zu trennen gilt. Mittels geeigneter Segmentierungsverfahren werden die Kunden in zwei (oder mehr) Segmente untergliedert, die in der Abbildung mit einem Stern bzw. einem Quadrat gekennzeichnet sind. Diese bilden sodann die Voraussetzung für den Einsatz von Klassifikationsinstrumenten wie SVM. Der Klassifikationsprozess teilt sich in die Trennung der vorgegebenen Segmente und die daran anschließende Prognose der Klassenzugehörigkeit eines potenziellen Neukunden auf. Auf Basis der erhobenen Merkmale kann nun eine adäquate Behandlung dieses Kunden erfolgen.

Bei der Auswahl der beschreibenden Merkmale wird ein Zusammenhang zum Untersuchungsgegenstand, also hier den Klassenzugehörigkeiten, unterstellt. Weiterhin muss für die Klassifikation mittels SVM gewährleistet werden, dass die verwendeten Merkmale metrisch oder zumindest metrisch interpretierbar sind. Die Annahme einer bestimmten Verteilung der Daten ist hingegen nicht notwendig, wie es bei vielen anderen Verfahren der Fall ist. Somit muss SVM nicht auf Datensätze eingeschränkt werden, die einer bestimmten Verteilung genügen. Die Methodik der SVM erfordert aufgrund der Berechnung in einem euklidischen Raum mit Distanzmaßen die Darstellung der Daten als Vektoren mit reellwertigen Einträgen (*Bennett, Campbell* (2000)). Dies ist bei vielen Anwendungen gerade im Bereich des Marketing nicht gegeben. Eine zu analysierende Datenbasis besteht hier häufig aus Antworten, die mittels eines Fragebogens ermittelt worden sind, bei dem Skalen nicht metrischen Niveaus eingesetzt werden. So können kategoriale Daten nicht ohne weiteres verwendet werden. Eine Möglichkeit der Transformation der Daten liegt in der binären Darstellung einer kategorialen Variablen (*Hsu et al.* (2003)). Somit wird jede dieser Kategorien durch einen binär kodierten Eintrag im Merkmalsvektor repräsentiert. Das bedeutet für eine kategoriale Variable mit drei Ausprägungen (z.B. „niedrig“, „mittel“, „hoch“), dass die Darstellung dieses Merkmals mittels dreier Einträge im Merkmalsvektor der Form $(1,0,0)$ (=„niedrig“), $(0,1,0)$(=„mittel“) oder $(0,0,1)$ (=„hoch“) erfolgt. Eine Eins an Stelle i ($i \in \{1,2,3\}$) bedeutet, dass die jeweilige Kategorie i vorliegt. Da eine sinnvolle Interpretation der resultierenden Trennebene, sowie der erzeugten Distanzen im Merkmalsraum bei Verwendung kategorialer Variablen nicht möglich ist, wird vom Einsatz dieser eher abgeraten. Binärkodierte Daten könnten im Einzelfall derart verwendet werden, dass Werte zwischen 0 und 1 als Anteile interpretiert werden, zu denen das jeweilige Merkmal vorliegt. Als metrisch interpretierbar hingegen gelten beispielsweise Ratingskalen, die häufig innerhalb der Marketingforschung eingesetzt werden.
Ein weiteres ebenfalls häufig im Data Mining und insbesondere im Marketing auftretendes Problem sind fehlende Werte (*Hippner, Wilde* (2001)). Werden Antworten bei der Erhebung von Daten mittels Fragebögen verweigert, so treten fehlende Werte auf, welche auf den Einsatz etwa zu persönlicher Fragen zurückzuführen sind. Eine Auswahl an Möglichkeiten des Umgangs mit fehlenden Werten liefert z.B. *Wagner et al.* (1998). Das Problem kann entweder ignoriert werden, d.h. die betreffenden Beobachtungen werden komplett aus der Datenbasis gelöscht, bzw. die betreffenden Merkmale werden entfernt, oder zufällig fehlende Werte werden

durch gängige Imputationsverfahren ersetzt. Der Einsatz von SVM macht eine vollständige Datenbasis erforderlich, bei der bei jeder Beobachtung ein Eintrag zu jedem Merkmal vorhanden sein muss. Neben der Fragebogenerhebung können fehlende Werte insbesondere bei manueller Erhebung von Daten auftreten. Unter Umständen muss durch die notwendige Vollständigkeit der Daten somit eine Vielzahl von Beobachtungen gelöscht werden, wodurch wichtige Informationen verloren gehen würden. Wünschenswert für zukünftige Anwendungen ist daher eine Anpassung der Methodik, die die Verwendung von Daten mit fehlenden Werten erlaubt. Bei Vorliegen von fehlenden Werten kann eine Reihe von Vorarbeiten nötig sein, um vollständig beschriebene Beobachtungen zu erhalten und diese mittels SVM analysieren zu können.

Liegen l vollständige Daten vor, so sollten diese auf ein Intervall $[0;1]$ oder $[-1;1]$ normiert werden. Eine Normierung verhindert eine Dominanz von Merkmalen, die hohe Ausprägungen haben, aber weniger gut diskriminieren als andere Merkmale. Dem wird durch eine Normierung der Ausprägungen aller Merkmale auf ein einheitliches Intervall entgegengewirkt. Dies bewirkt zum einen, dass Merkmale mit unterschiedlichen Skalen gleichen Einfluss ausüben können. So kann z.B. in der Kundenklassifikation mit den eingehenden Merkmalen „Einkommen“, gemessen in Euro, und „Anzahl der Kinder“ verhindert werden, dass das Merkmal Einkommen aufgrund der Ausprägungen das zweite Merkmal dominiert und somit allein Einfluss auf die Lage der Ebene ausübt. Zum anderen werden durch eine derartige Normierung Schwierigkeiten bei der numerischen Berechnung der Skalarprodukte vermieden (*Hsu et al.* (2003)) und die Anzahl der Iterationen bei der Konvergenz des Algorithmus reduziert, was zu einer verkürzten Rechenzeit führen kann. Bei den in Kapitel 4 eingesetzten Daten wird die folgende Normierung gewählt:

$$x_{ik}^{neu} = \frac{x_{ik} - \min_j x_{jk}}{\max_j x_{jk} - \min_j x_{jk}},$$

wobei x_{ik} die Merkmalsausprägung des k-ten Merkmals der i-ten Beobachtung angibt. Die neuen Beobachtungswerte x_{ik}^{neu} liegen somit im Intervall $[0,1]$, womit ein systematischer Einfluss von einzelnen Merkmalen verhindert werden kann.

3.2 Einbindung von a priori-Wissen

In diesem Abschnitt soll gezeigt werden, wie die bisher vorgestellte Methodik so verändert werden kann, dass die Ziele, die in empirischen Untersuchungen verfolgt werden (z.B. die Identifizierung der für ein Direktmailing interessantesten Kunden), besser erreicht werden. Die Intention liegt daher darin, a priori-Wissen über die Daten mit in die Optimierung einfließen zu lassen, um somit Möglichkeiten zur effektiven Zielerreichung zu diskutieren. Dieses Wissen kann etwa darin bestehen, dass bestimmte Merkmale für die Trennung von Gruppen eine wichtigere Rolle spielen oder spielen sollen als andere. Weiterhin ist es denkbar, dass die zu untersuchenden Gruppen unterschiedliche Bedeutung für ein Unternehmen haben. Um ein aus

Sicht des Anwenders optimales Ergebnis hinsichtlich dieser Aspekte zu erzielen, sollte dieses Wissen mit in die Optimierung einfließen. Es wird geklärt, wie Merkmale, Klassen oder Beobachtungen gewichtet werden können, um ihnen eine höhere Bedeutung innerhalb der Analyse zukommen zu lassen.

3.2.1 Merkmalsgewichtung

In vielen Anwendungen von Klassifikationsverfahren spielen die eingesetzten Merkmale unterschiedliche Rollen. So kann einem Merkmal eine besondere Bedeutung zukommen, wenn es in einer speziellen Beziehung zum Untersuchungsgegenstand steht, sei es in positiver oder auch negativer Form. Gerade im Marketing tritt der erste Fall sehr häufig auf, etwa in der Kundenklassifikation, bei der der bisherigen Kaufhistorie im Gegensatz z.B. zum Alter eines Kunden meist eine höhere Bedeutung bei der Klassifikation zukommt. Bei der herkömmlichen Methode der SVM besteht keine Möglichkeit, diesem Aspekt Rechnung zu tragen, da bei der Berechnung der Hyperebene aufgrund der vorgenommenen Normierung alle Merkmale gleich stark in die Optimierung eingehen. Im Folgenden wird untersucht, inwieweit die grundlegende Methodik modifiziert werden kann, um eine zusätzliche Gewichtung von hervorzuhebenden Merkmalen zu ermöglichen. Dazu stehen grundsätzlich zwei verschiedene Möglichkeiten zur Verfügung, die als implizite und explizite Berücksichtigung von Wissen bezeichnet werden können. Zum einen können die ursprünglichen Daten so verändert werden, dass eine Gewichtung der Merkmale erfolgt und das Wissen somit implizit mit einfließt. Zum anderen kann die bestehende Methodik der SVM explizit in ihrer Form variiert werden. Beide Ansätze werden hier diskutiert.

Sollen die Daten vor der Optimierung verändert werden, so werden zunächst die Merkmale ausgewählt, die in die Klassifikation eingehen. Im zweiten Schritt folgt die Definition der zu wählenden Verhältnisse der Wichtigkeiten der festgelegten Merkmale. Hierbei handelt es sich nun um die eigentliche Gewichtung. Falls bestimmte Merkmale bei der Klassifikation eine höhere Rolle spielen oder stärker ins Gewicht fallen sollen als andere, so werden sie in diesem Schritt anhand der Verhältnisse gewichtet. Es wird demnach ein Vektor erstellt, dessen Länge der Anzahl der heranzuziehenden Merkmale entspricht und der die Gewichte der jeweiligen Merkmale enthält. Durch die multiplikative Verknüpfung der Gewichte mit den Daten resultiert ein neuer Datensatz:

$$\boldsymbol{AG} = \boldsymbol{A'} \quad \text{mit } \boldsymbol{A} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{l1} & x_{l2} & \dots & x_{ln} \end{pmatrix} \quad \text{und } \boldsymbol{G} = \begin{pmatrix} g_1 & 0 & \dots & 0 \\ 0 & g_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & g_n \end{pmatrix}$$

wobei $\boldsymbol{A}$ die ursprüngliche Datenmatrix bezeichnet und $\boldsymbol{G}$ eine Diagonalmatrix ist, deren Einträge die Gewichte g_j der einzelnen Merkmale $j = 1, \dots, n$ bilden. Somit resultiert die Matrix $\boldsymbol{A'}$, die die gewichteten Merkmalsausprägungen aller

Beobachtungen enthält.
Das beschriebene Vorgehen ähnelt dem Verfahren der Scoring-Methode (vgl. *Krafft, Albers* (2000)), die sich in mehrere Schritte zur Bestimmung des Kundenwertes aufteilt. Folglich werden Kriterien festgelegt, nach denen die Kunden bewertet werden sollen. Entsprechend der Gewichtung der Kriterien und der Ausprägung des Merkmals erhält jeder Kunde für jedes Kriterium respektive Merkmal einen Wert. Im Rahmen der Scoring-Methode wird anschließend für jeden zu untersuchenden Kunden ein Score durch Addition der zugewiesenen und gewichteten Merkmalsausprägungen ermittelt, der den eigentlichen Kundenwert repräsentiert. Dieser Schritt entfällt an dieser Stelle.

Die intuitive Vorgehensweise der manuellen Gewichtung der Merkmale erscheint aufgrund des enormen Aufwandes recht umständlich, da für jede Veränderung der Gewichtung ein neuer Datensatz erstellt werden muss, der in die Optimierung der SVM eingeht. Daher soll im Folgenden vorgestellt werden, wie die Methodik modifiziert werden kann, um eine Erstellung neuer Datensätze zu umgehen. Dabei spielt der auch bei der Scoring-Methode zu erstellende Gewichtsvektor eine entscheidende Rolle.
Bei der Berechnung der Hyperebene gehen die Eingabevektoren nur in Form von Skalarprodukten ein, was den Einsatz von Kernfunktionen ermöglicht. Das Skalarprodukt zweier n-dimensionaler Vektoren $\boldsymbol{x}_i$ und $\boldsymbol{x}_{i'}$ mit $i, i' \in \{1, \ldots, l\}$, definiert durch $\boldsymbol{x}_i \boldsymbol{x}_{i'} = \sum_{j=1}^{n} x_{ij} x_{i'j}$, bildet den Ansatzpunkt der Integration von Merkmalsgewichtungen. Statt des Faktors 1 vor jedem Summanden $x_{ij} x_{i'j}$ des Skalarproduktes können individuelle Faktoren für jedes Merkmal $j \in \{1, \ldots, n\}$ eingesetzt werden. Diese zusätzlichen Faktoren bewirken, dass einem besonders hoch gewichteten Merkmal ein großer Einfluss auf die Lage der Ebene im Vergleich zu weniger stark gewichteten Merkmalen zugewiesen werden kann. So lässt sich das zu optimierende Problem (2.13) im linearen Fall wie folgt verändern:

$$\max \Big\{ \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{i'=1}^{l} \alpha_i \alpha_{i'} y_i y_{i'} \sum_{j=1}^{n} g_j x_{ij} x_{i'j} \Big\}$$

unter den Nebenbedingungen

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \qquad \text{und} \qquad 0 \leq \alpha_i \leq C \quad \forall\, i.$$

Die Gewichte g_j bilden den Gewichtsvektor $\boldsymbol{g} = (g_1, \ldots, g_n)$, dessen Einträge die Hauptdiagonale der Gewichtsmatrix $\boldsymbol{G}$ bilden. Dieser Vektor ist a priori vom Benutzer festzulegen. Für die praktische Umsetzung einer SVM ist es von entscheidendem Vorteil, dass jetzt nur der Gewichtsvektor variiert werden muss, um die Gewichtung der Merkmale zu verändern. Dies kann gerade bei einer großen Anzahl von Merkmalen bedeutsam sein[1].

[1] Ein Vergleich der beiden Möglichkeiten zur Merkmalsgewichtung mittels Daten- und Methodenmodifikation zeigt, dass beide zu gleichen Ergebnissen kommen. Hier muss berücksichtigt wer-

Nicht nur die lineare Form der SVM kann auf diese Art und Weise modifiziert werden. Auch innerhalb nicht linearer Kernfunktionen lassen sich Veränderungen vornehmen, um die unterschiedliche Hervorhebung der Merkmale zu realisieren. Bei Wahl eines polynomiellen Kerns ist eine zum linearen Fall analoge Modifikation möglich:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = (\sum_{j=1}^{n} g_j x_{ij} x_{i'j} + 1)^d$$

mit $g_j \geq 0$ für alle j. In der Literatur sind bisher nur wenige Ansätze zu dieser Modifikation zu finden. Im Rahmen der Parameterbestimmung wird bei *Chapelle et al.* (2002) ein Polynom vom Grad 2 der Form $K(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = (1 + \sum_{j=1}^{n} \frac{x_{ij} x_{i'j}}{g_j^2})^2$ verwendet[2]. Ziel war hier allerdings die automatische Bestimmung möglichst vieler Parameter, und nicht wie das hier eigentlich intendierte Ziel der a priori-Gewichtung von Merkmalen. Diese Modifikation kann in einfacher Art auf höhere Polynome obiger Form ausgedehnt werden.
Die Auswirkungen einer Hervorhebung des zweiten Merkmals bei zweidimensionalen Daten auf die Lage der Ebene zeigt Abbildung 3.3.

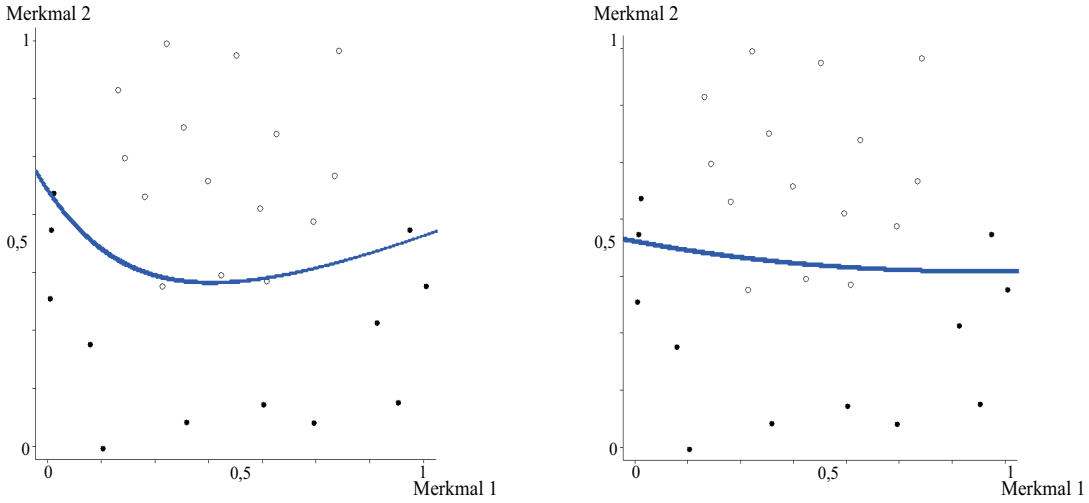

Abbildung 3.3: Trennung von Daten durch ein Polynom zweiten Grades bei gleichgewichteten Merkmalen (links) und bei zusätzlicher Hervorhebung des zweiten Merkmals ($g_1 = 0,2$ und $g_2 = 0,8$) (rechts)

Gegenüber der herkömmlichen Trennung wurde das zweite Merkmal im rechten Teil der Abbildung höher gewichtet als das erste. Es ist deutlich zu erkennen, dass bei der Berechnung der Ebene im rechten gewichteten Fall das erste Merkmal keinen

den, dass das Merkmal j bei der manuellen Modifikation der Daten mittels Gewicht g'_j zweimal bei der Berechnung eines Skalarproduktes auftritt, sodass für einen Vergleich bei Einsatz z.B. des polynomiellen Kerns $g'_j = \sqrt{g_j}$ gelten muss.

[2]Die quadratische Gewichtung sichert die Positivität der Gewichte. Die Gewichtung eines Merkmals mit einem negativen Gewicht ist nicht zugelassen.

großen Einfluss hat, da die Ebene hauptsächlich durch einen Ausdruck gebildet wird, der stark vom zweiten Merkmal abhängig ist. Hierbei steht lediglich die Auswirkung der Veränderung der Gewichte im Vordergrund. Die Prognosegüte spielt in diesem Beispiel eine untergeordnete Rolle.
Eine weitere Möglichkeit der Einbindung von a priori-Wissen über die Wichtigkeit der Merkmale bildet ein modifizierter Radialbasis-Kern (*Chapelle et al.* (2002))

$$K(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = \exp\left(-\sum_{j=1}^{n} \frac{(x_{ij} - x_{i'j})^2}{2g_j^2}\right)$$

bzw. in der Form des in dieser Arbeit verwendeten Radialbasis-Kerns

$$K(\boldsymbol{x}_i, \boldsymbol{x}_{i'}) = \exp\left(-\sum_{j=1}^{n} g_j (x_{ij} - x_{i'j})^2\right) \quad \text{mit } g_j \geq 0 \quad \forall j.$$

In der ursprünglichen, nicht gewichteten Form gilt $g_1 = \ldots = g_n = \gamma$, wobei γ der zu wählende Kernparameter ist. Auch hier kann durch g_j eine Hervorhebung der einzelnen Merkmale durchgeführt werden, da in der Berechnung des Kerns nur Verknüpfungen von Merkmalsausprägungen der gleichen Dimension berücksichtigt werden.
Wie die einzelnen Parameter allerdings zu wählen sind, ist ein offenes Problem (vgl. Abschnitt 3.3). Hierfür wird in *Chapelle et al.* (2002) eine Lösung vorgeschlagen, die zur Ermittlung der optimalen Trennebene die Gewichte automatisch innerhalb des Optimierungsalgorithmus anpasst. Eine Alternative bei der Bestimmung der Gewichte liegt in der manuellen Gewichtung der Merkmale z.B. auf Basis von Expertenwissen.

Die Datenmodifikation weist den nicht zu unterschätzenden Nachteil auf, dass für jede Gewichtung ein neuer Datensatz generiert werden muss. Dies ist unter Umständen bei Datenbanken enormen Umfangs nur schwer oder gar nicht zu realisieren. Die Modifikation der Methodik erfordert demgegenüber lediglich einen Vektor $\boldsymbol{g}$ der Länge n (Anzahl der Merkmale), in dem die vorzunehmenden Gewichtungen gespeichert sind. Somit können schneller die Auswirkungen mehrerer Gewichtungen auf die Lage der Ebene und das Klassifikationsergebnis getestet werden, um eine optimale Gewichtung der Merkmale zu erreichen.

3.2.2 Klassengewichtung

Eine Gewichtung durch den Parameter C bewirkt, dass bei festen Kernparametern die Trefferquote bei steigendem Kostenparameter C zunächst nur im Trainingsdatensatz, unabhängig von der Klassifikationsgüte innerhalb der Testdaten, verbessert wird. Anschaulich bedeutet dies, dass der Parameter C dazu dient, eine Information über die Wichtigkeit der Richtigklassifikation zu geben. Je höher dieser Wert ist, desto eher wird die Richtigklassifikation im Gegensatz zur Maximierung der Spanne zwischen den Klassen fokussiert. Der Parameter kann also interpretiert werden als

Kosten, die für die Zuordnung einer Beobachtung zur falschen Klasse entstehen. Es wird bisher nicht unterschieden, aus welcher Klasse die jeweilige Beobachtung stammt. Gerade bei betriebswirtschaftlichen Anwendungen z.B. im Rahmen einer Kundenklassifikation ist es jedoch von Vorteil, neben der Hervorhebung einzelner Merkmale auch die zu separierenden Klassen unterschiedlich gewichten zu können. Dies ist immer dann der Fall, wenn die richtige Klassifikation bestimmter Beobachtungen wichtiger ist als die anderer. Eine falsche Zuordnung derjenigen Kunden, die für das Unternehmen interessant sind, ist schwer wiegender als der umgekehrte Fall. Im ersteren Fall erfährt der ursprünglich wichtige Kunde aufgrund der Zuordnung zu den eher uninteressanten Kunden eine seiner bisherigen Kaufhistorie nicht adäquate Behandlung, und somit geht dem Unternehmen eventuell ein wertvoller Kunde verloren. Wird hingegen ein ehemals für das Unternehmen uninteressanter Kunde den wichtigen Kunden zugeordnet, so führt dies zwar zu einem möglichen Fehleinsatz des Marketingbudgets, allerdings ist dies nicht so folgenschwer wie die falsche Behandlung oder gar der Verlust eines wichtigen Kunden. Im Zusammenhang mit der in jüngster Zeit intensiv diskutierten Kundenbindung (z.B. *Krafft, Götz* (2004)) erhält die Richtigklassifikation der wichtigen Kunden eine besondere Bedeutung. Im Rahmen von Kundenbindungsprogrammen, wie z.B. Kundenkarten, wird versucht, die bestehenden, wichtigen Kunden zu halten, da der Aufwand, Neukunden zu gewinnen, sehr viel höher und kostspieliger ist.
Neben der Betonung der Richtigklassifikation von wichtigen Kunden kann dieses Konzept in vielen Bereichen zum Einsatz kommen. So ist es beispielsweise in der Fertigungsindustrie wichtig, fehlerhafte Produktteile zu erkennen und auszusortieren. Dies kann automatisch durch Bilderkennungsverfahren erfolgen, in denen SVM zum Einsatz kommen können. Einen weiteren wichtigen Anwendungsbereich bildet die Medizin bei der Diagnose von schweren Krankheiten, die frühzeitig für eine erfolgreiche Behandlung erkannt werden müssen. Auch hier kann durch eine Hervorhebung der Richtigklassifikation der Klasse der positiv getesteten Objekte die Erfolgsquote in der Heilung der Krankheit durch Anwendung der angemessenen Therapie u.U. erhöht werden.

Die in Kapitel 2 vorgestellte Methodik erlaubt diese verschiedenartige Gewichtung noch nicht und erfordert zusätzliche Modifikationen. Bisher steht die Richtigklassifikation aller Punkte bei gleichzeitiger Maximierung der Spanne im Vordergrund. Im Optimierungsproblem (2.11) und (2.12) ist am konstanten Gewichtungsparameter C zu erkennen, dass die Beobachtung ungeachtet ihrer Klassenzugehörigkeit behandelt werden. Im Folgenden wird auf Modifikationen dieses Parameters für den Fall der Biklassifikation eingegangen[3].

Bisher gibt es einige Ansätze und Anwendungen der beschriebenen Modifikation, von denen viele sich ähneln, deren Zielsetzungen aber teilweise voneinander abweichen. So führen *Lin et al.* (2002) SVM für Nicht-Standardsituationen ein, indem sie sowohl verschiedene Kosten für Fehlklassifikationen innerhalb der beiden zu

[3]Die vorgestellten Veränderungen lassen sich entsprechend auf die Multiklassifikation übertragen und werden in Abschnitt 4.4.2 angewendet.

trennenden Klassen als auch die möglicherweise abweichende Verteilung innerhalb des Trainingsdatensatzes und der Grundgesamtheit berücksichtigen. Hierbei sei angemerkt, dass bei der Standardversion der SVM die zugrunde liegende Verteilung im Gegensatz zu anderen Verfahren wie der Diskriminanzanalyse bei der Berechnung der Hyperebene keine Rolle spielt. Dennoch wird diese in *Lin et al.* verwendet, um eine erhöhte Trefferquote durch repräsentative Daten zu erhalten. *Ma, Ding* (2002) benutzen ebenfalls verschiedene Kostenparameter für die Fehlklassifikation bei der Erkennung von Gesichtern, um die richtige Identifizierung eines Gesichts zu verbessern. In den Arbeiten von *Huang, Liu* (2002) und *Lin, Wang* (2002) wird eine Unterscheidung zwischen den Fehlklassifikationen bestimmter Punkte eingeführt. Das Ziel hierbei ist es, mittels Fuzzyfizierung den Einfluss von Ausreißern auf die Lage der zu bestimmenden Ebene zu begrenzen, was durch die Herabsetzung der jeweiligen Kosten erfolgte. Bei *Morik et al.* (1999) dient die Kostengröße C dazu, das möglicherweise ungleiche Verhältnis der Klassengrößen auszugleichen, was beim praktischen Einsatz ebenfalls angewendet werden sollte.

Im Einzelnen bedeutet eine Variation des Parameters C eine Aufsplittung in beobachtungsabhängige Kostengrößen. Das Ausgangsproblem ist äquivalent zu (2.11):

$$\min \ \frac{1}{2}\,||\boldsymbol{w}||^2 + \sum_{i=1}^{l} C_i \xi_i \tag{3.1}$$

unter den Nebenbedingungen

$$y_i(\boldsymbol{w}\boldsymbol{x}_i + b) \geq 1 - \xi_i \qquad \forall\ i \in \{1, \ldots, l\}, \tag{3.2}$$

wobei die Kostengröße nun nach folgender Vorschrift für jede Klasse einzeln gewichtet werden kann:

$$\begin{array}{l} C_i = c_+ C \quad \text{, falls } y_i = 1 \\ C_i = c_- C \quad \text{, falls } y_i = -1 \end{array} \quad \text{mit } C, c_+, c_- \in \mathbb{R}^+.$$

Dabei seien c_+ und c_- die individuellen Gewichtungsparameter, C ist der in Abschnitt 2.1.3 eingeführte, feste Kostenparameter. Hierbei wird C so gewählt wie im herkömmlichen Fall, c_+ bzw. c_- werden verändert, um eine zusätzliche Hervorhebung bestimmter Klassen zu ermöglichen.

Wird auch hier zum dualen Problem gewechselt, so ergibt sich der folgende Optimierungsansatz, welcher sich nur durch den Gewichtungsparameter C_i von (2.19) und (2.20) unterscheidet:

$$\max \Big\{ \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \Big\} \tag{3.3}$$

unter den Nebenbedingungen

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \qquad \text{und} \quad 0 \leq \alpha_i \leq C_i \quad \forall\ i. \tag{3.4}$$

Mit den verschiedenen Gewichtungsparametern c_+ und c_- kann nun der Fokus auf eine bestimmte, besonders interessierende Klasse gelegt werden. Bezogen auf die oben erwähnte Kundenklassifikation würde $c_+ > c_-$ gewählt werden, wenn die Klasse der gewinnbringenden, also interessanten Kunden mit „+1“ bezeichnet ist. Auf die Möglichkeiten der Bestimmung dieser Parameters wird in Abschnitt 3.3 eingegangen. An dieser Stelle sei bereits darauf hingewiesen, dass der Wert dieser Variablen stark von den Daten und deren Trennbarkeit abhängig ist.
Die Möglichkeit der unterschiedlichen Behandlung verschiedener Klassen kann ebenfalls hinzugezogen werden, wenn ein nicht balancierter Datensatz zugrunde liegt. Dies ist dann der Fall, wenn die Klassen unterschiedliche Größe haben (*Morik et al.* (1999)). Das Verhältnis der Klassengrößen kann durch $\frac{c_+}{c_-}$ ausgedrückt werden (vgl. dazu die Betrachtungen auf Seite 63) und somit kann einer ungleichen Verteilung, die zur Fehlklassifikation ganzer Klassen führen kann, Rechnung getragen werden. Damit wird einer Fehleinschätzung der Klassifikationsgüte bei stark asymmetrischen Klassengrößen vorgebeugt, wie in Abschnitt 3.8 gezeigt wird.
Wird C nun aufgespalten in einen positiven und einen negativen Teil, wird die Richtigklassifikation der Vektoren aus Klasse „+1“ durch die Erhöhung von c_+ hervorgehoben.

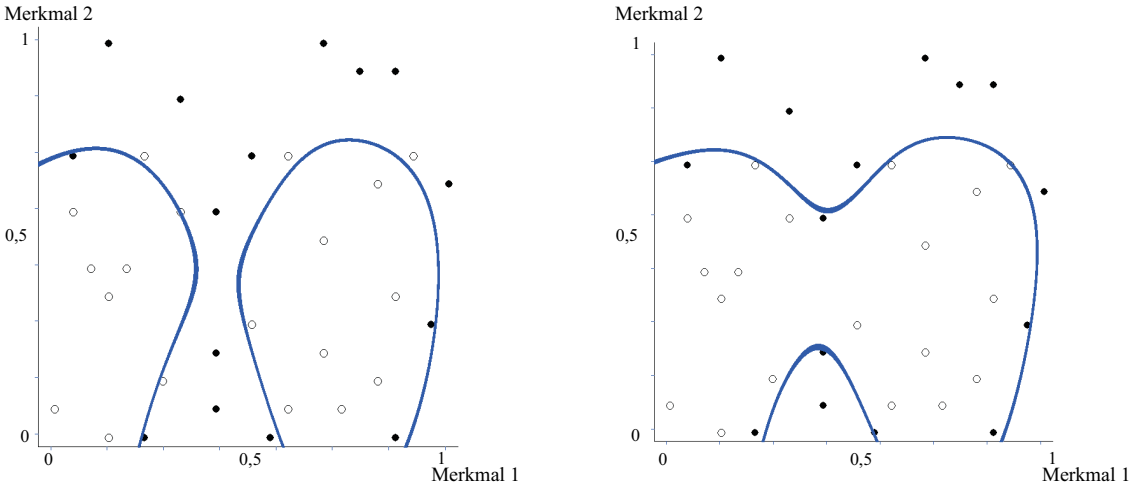

Abbildung 3.4: Trennung zweier Klassen mit $\gamma = 0,03$ und $C = 1$ bei gleichgewichteten Klassen (links) und unterschiedlichen Klassengewichten $c_+ = 1,14$ und $c_- = 1$ (rechts)(Klasse „+1“ ist durch ∘, Klasse „-1“ durch • gekennzeichnet.

In Abbildung 3.4 werden die Daten zweier, durch zwei Merkmale beschriebenen Klassen durch eine nicht lineare Trennung mittels Radialbasis-Kern getrennt. Klasse „+1“ wird durch das Symbol ∘, Klasse „-1“ durch das Symbol • gekennzeichnet. Im Vergleich zur linken Trennung, bei der die beiden Klassen gleich gewichtet wurden ($C = 1$), ist in der rechten Abbildung zu erkennen, dass die Erhöhung des Gewichts die Klasse „+1“ zu einer besseren Klassifikation der entsprechenden Beobachtungen führt. Wurden im linken Bild vier Beobachtungen aus Klasse „+1“ falsch klassifiziert, so wird diese Klasse in der rechten Abbildung vollständig

richtig erkannt, wobei die Güte der Klassifikation der anderen Klasse vernachlässigt und somit im vorliegenden Fall verringert wird. Es ist anzumerken, dass drei der Vektoren aus Klasse „+1“ sehr nah an der Ebene positioniert und somit nur knapp richtig klassifiziert werden.
Hier muss beachtet werden, dass insbesondere beim Einsatz des Radialbasis-Kerns die Parameter derart gewählt werden können, dass alle Vektoren richtig klassifiziert werden, bei denen dies auch beabsichtigt ist, und die Richtigklassifizierung der übrigen Punkte eine untergeordnete Rolle spielt. Dies kann trivial durch Zuweisung aller Punkte zu einer Klasse erreicht werden, was den Einsatz eines Klassifikationsinstrumentes allerdings überflüssig macht. Häufig sind aber weitere Einschränkungen gegeben, z.B. Vorgaben über finanzielle Mittel, die eine solche Zuordnung nicht erlauben und eine Richtigklassifikation möglichst vieler Beobachtungen aus beiden Klassen erforderlich machen.

Soll die Fokussierung auf eine beliebige Untermenge der Eingabedaten unabhängig von der Klassenzugehörigkeit der jeweiligen Punkte ausgedehnt werden, so besteht die Möglichkeit der Einführung individueller Gewichtungsparameter. Dies kann als Erweiterung der Klassengewichtung auf eine Gewichtung von einzelnen **Beobachtungen** verstanden werden. Dieses Gewicht kann eine funktionale Beziehung widerspiegeln, in der beispielsweise Kosten für eine Fehlklassifikation verarbeitet werden. Ein weiteres Ziel bei der Verwendung von beobachtungsabhängigen Kostenparametern kann die Herabsetzung des Einflusses von Ausreißern sein (vgl. *Huang, Liu* (2002)). Unabhängig von der Klassenzugehörigkeit können die Kosten für diese Beobachtungen minimiert werden. Obwohl die Gewichtung von Beobachtungen zur ex post Klassifikation keine Bedeutung hat, stellt sie dennoch ein für das Marketing interessanten Ansatz dar. So können Ausreißer mit einem geringen Gewicht in die Analyse integriert werden, ohne sie vollständig aus dem Datensatz zu entfernen. Weiterhin könnten etwa Opportunitätskosten im Rahmen betriebswirtschaftlicher Anwendungen in die Analyse eingearbeitet und bei der Bestimmung der Trennebene berücksichtigt werden.
Der zu analysierende Datensatz wird um ein weiteres Merkmal erweitert werden, welches Einfluss auf die Veränderung der Hyperebene durch verschiedenartige Gewichtung der Datenvektoren nimmt. Eine Menge $\{(\boldsymbol{x}_i, y_i, c_i); i = 1, \ldots, l\}$ mit $c_i > 0$ bildet hierbei die Eingabedaten.

Das zu lösende Optimierungsproblem hat in der primalen Form die gleiche Form wie in (3.1) und (3.2). Hierbei wird jedoch C_i nicht mehr durch die Klassenzugehörigkeit der einzelnen Punkte bestimmt, sondern durch die zusätzliche Eingabegröße c_i:

$$C_i = c_i C \quad \text{, mit } c_i \in \mathbb{R}_+.$$

Das zugehörige duale Optimierungsproblem hat die gleiche Form wie (3.3) und (3.4). Die Auswirkung einer Veränderung des Kostenparameters für nur einzelne Vektoren zeigt Abbildung 3.5. Dabei werden die im rechten Teil der Abbildung mit $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ und $\boldsymbol{x}_3$ bezeichneten Vektoren mit $c_i = 1,5$ für $i = 1, 2, 3$ gewichtet, wohingegen den übrigen Vektoren die Gewichtung in Abhängigkeit ihrer Klassenzugehörigkeit

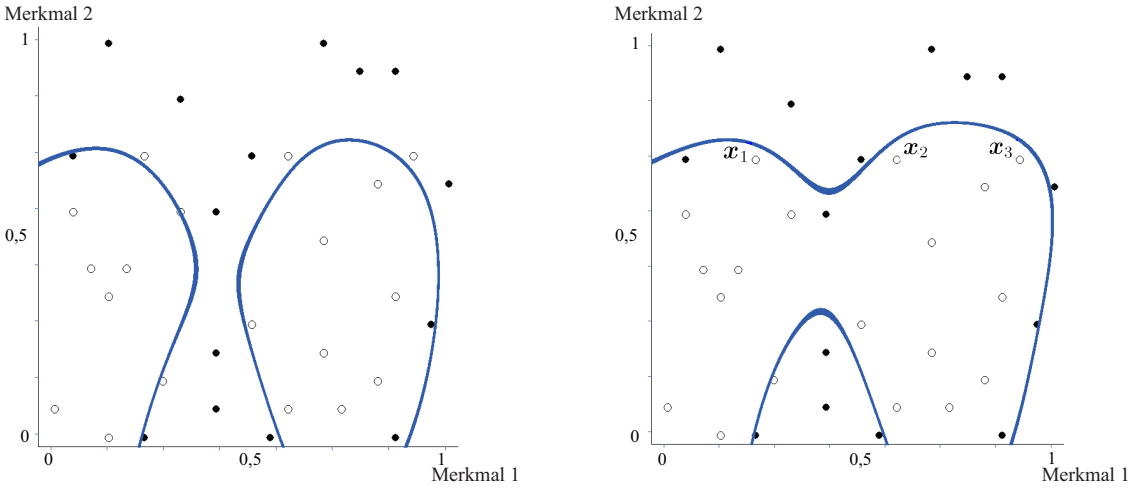

Abbildung 3.5: Trennebene mit $\gamma = 0,03$ und $C = 1$ (links) und mit $\gamma = 0,03$, $C = 1$, $c_+ = 1,14$ und $c_i = 1,5$ (rechts) für die drei beschrifteten Vektoren

(analog zu Abbildung 3.4) zugewiesen wird, also $c_+C = 1,14$, falls $y_i = +1$ und $c_-C = 1$, falls $y_i = -1$. Wie in Abbildung 3.4 werden die Beobachtungen aus Klasse „+1“ mit ∘ und diejenigen aus Klasse „-1“ durch • gekennzeichnet. Im Vergleich zur gleichmäßigen Gewichtung im linken Teil der Abbildung ist zu erkennen, dass SVM die Ebene so konstruiert, dass die Vektoren mit den höheren Kosten zu der ihnen zugewiesenen Klasse zugeordnet werden können. Ohne die zusätzliche Gewichtung der drei Vektoren ergibt sich die Ebene aus dem rechten Teil von Abbildung 3.4. Die weitere Erhöhung der Gewichte für diese Vektoren bewirkt ihre eindeutige Richtigklassifikation, allerdings unter in Kaufnahme von Fehlklassifikationen aus Klasse „-1“. Diese Eigenschaft kann insbesondere im Marketing genutzt werden, falls bei der Kundenklassifikation das Augenmerk auf besonders interessante Kunden innerhalb bestimmter Klassen gelegt werden soll.

Die beiden vorgestellten Ansätze zur Gewichtung von Merkmalen und Klassen respektive Beobachtungen bilden eine Möglichkeit zur Einbindung von a priori-Wissen und somit zur Steuerung der SVM im Sinne des intendierten Untersuchungsziels. So können Informationen über die Daten und strukturelle Zusammenhänge bei der Optimierung berücksichtigt werden, um somit die Güte der Klassifikation zu verbessern.

3.3 Parameterwahl

Die Nutzung von SVM für Klassifikationsaufgaben erscheint aufgrund guter Analyseergebnisse in verschiedenen Untersuchungen viel versprechend. So werden häufig gleichwertige oder sogar bessere Ergebnisse verglichen mit alternativen Klassifikationsverfahren, wie Entscheidungsbäumen oder neuronalen Netzen, erzielt. Die Ergebnisse basieren auf einer guten Wahl der Parameter, die der Benutzer a priori zu

treffen hat. Bei herkömmlichen SVM sind in der Regel etwa zwei bis drei Parameter im Vorhinein festzulegen. Zum einen muss die Ausprägung des Kostenparameters C (bzw. der Kostengrößen c_+, c_- oder c_i) und zum anderen müssen je nach zuvor gewähltem Kern die Werte des oder der Kernparameter festgelegt werden. Die Parameter sollten so gewählt werden, dass ein möglichst kleiner Generalisierungsfehler resultiert. Bisher ist diese Wahl der Parameter zur Erlangung eines möglichst guten Klassifikationsergebnisses auf Testdaten ein noch offenes Problem der Forschung (*Bennett, Campbell* (2000), *Schölkopf, Smola* (2002)). Es ist ebenfalls nicht möglich, Intervalle von Parameterausprägungen anzugeben, in denen gute Ergebnisse zu erwarten sind. Auf der anderen Seite ist aber eine angemessene Wahl beispielsweise des richtigen Kerns unabdingbar für eine gute Leistung von SVM (*Chapelle et al.* (2002)). Daher ist die Einschränkung der möglichen Parameterausprägungen auf ein Minimum erforderlich, um die Suche nach geeigneten Parametern anwenderfreundlich zu gestalten. Um zu Klassifikationsresultaten zu gelangen, die mit den Ergebnissen anderer Verfahren vergleichbar sind, bedarf es bei der nicht linearen Trennung u.U. eines großen Zeitaufwands. Daher liegt das Ziel dieses Abschnittes in der Diskussion der Möglichkeiten zur effizienten Bestimmung der Parameter, um SVM möglichst anwenderfreundlich zu gestalten.
Abbildung 3.6 zeigt, dass die Wahl der optimalen Parameter sehr stark vom zugrunde liegenden Datensatz abhängig ist. Es wurden dazu vier unterschiedliche Datensätze ausgewählt und die Trennung unter Einsatz des Radialbasis-Kerns durchgeführt, sodass die Parameter C und der Kernparameter γ bestimmt werden müssen. Die verwendeten Datensätze unterscheiden sich im Umfang der Beobachtungen, der Anzahl der Merkmale, der Anzahl der Klassen sowie dem Verhältnis der Klassenumfänge voneinander. Der inhaltliche Zusammenhang der Daten weicht ebenfalls voneinander ab. So handelt es sich bei der oberen linken Abbildung um Ergebnisse auf den bekannten Irisdaten von *Fisher* (1936). Der Datensatz „Kredit“ beinhaltet Merkmale zur Klassifikation von Kreditnehmern, wohingegen es sich bei den unteren beiden um marketingspezifische Daten zur Klassifikation von Kunden handelt[4]. Es ist zu erkennen, dass die Regionen, in denen optimale Parameter zu erwarten sind, voneinander abweichen.

[4]Quellenangaben zum Datensatz „Kredit“ finden sich auf Seite 105. Der Datensatz „Pharma“ wird im empirischen Teil in Abschnitt 4.3.2 näher untersucht. Bei dem Datensatz „CRM“ handelt es sich um einen fiktiven Datensatz, der in *Decker, Monien* (2003) verwendet wird.

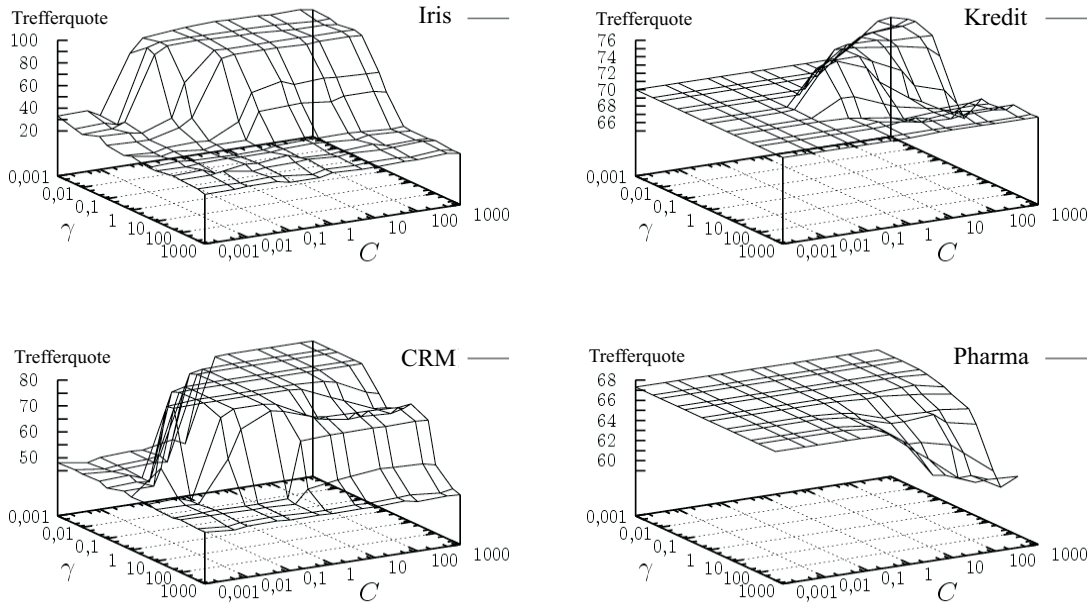

Abbildung 3.6: Darstellung der Trefferquoten für vier verschiedene Datensätze in Abhängigkeit der Parameter γ und C bei Einsatz des Radialbasis-Kerns

Weiterhin liegen die Bereiche einer hohen Trefferquote in Abbildung 3.6 zum Teil sehr nah an denen, die eine schlechte Trefferquote erzielen. Dies ist z.B. beim Datensatz „CRM“ bei den Ausprägungen $C = 0,01$ oder $C = 0,05$ bei $\gamma = 5$ der Fall. Dabei ergeben sich die stark unterschiedlichen Trefferquoten von 46,9% bzw. 77,4%. Dies zeigt, dass SVM empfindlich auf kleinste Änderungen der Parameter reagieren, was in *Ou et al.* (2003) bestätigt wird. Insgesamt erscheint jedoch die Kombination aus einem kleinen Wert für γ und einem großen Wert für C als eine gute Wahl zur ersten Prognose.
Von optimalen Parametern eines Datensatzes kann nur auf die Parameterkonstellationen eines neuen Datensatzes geschlossen werden, wenn dieser den bekannten Daten strukturell sehr ähnlich ist. Ansonsten bleibt eine umfassende Suche nach einer guten Konfiguration nicht aus.

Um eine gute Prognosegüte zu erreichen, müssen die Parameter so ausgewählt werden, dass sie eine hohe Trefferquote auf adäquaten Testdaten ermöglichen, d.h. ein Overfitting der Trainingsdaten soll vermieden werden. Overfitting tritt immer dann auf, wenn die Anpassung eines Modells, in diesem Fall die Anpassung der Hyperebene an die Trainingsdaten, sehr gut ist, aber diese Anpassung nicht generalisiert werden kann, d.h. dass die zufällig auftretenden Störungen als systematisch und bedeutsam aufgefasst werden und bei der Berechnung der Hyperebene berücksichtigt werden.

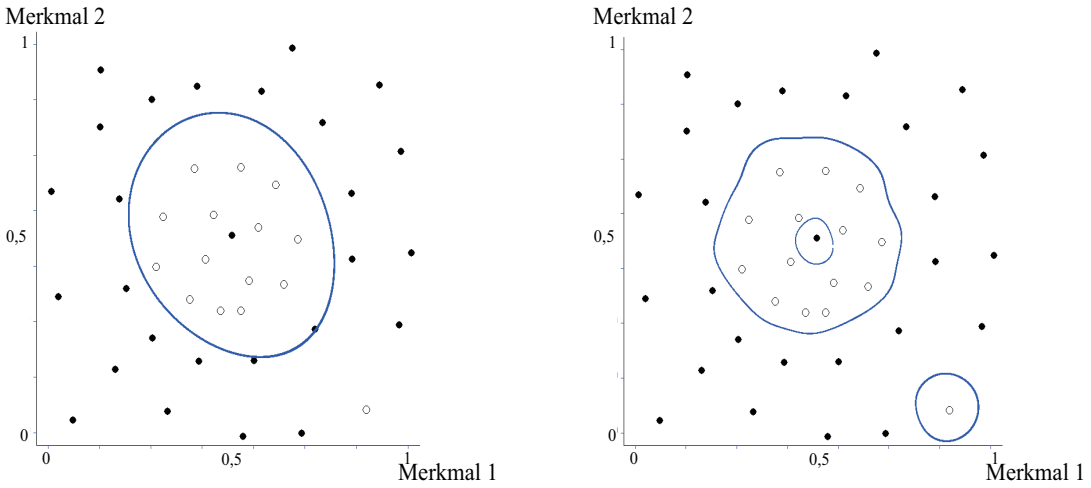

Abbildung 3.7: Trennung von Daten mittels zweier verschiedener Parameterkonstellationen

In Abbildung 3.7 wird die unterschiedliche Gestalt der Ebene in Abhängigkeit des Kernparameters verdeutlicht. Um die beiden abgebildeten Klassen zu trennen, würde in diesem Fall die linke Lösung der rechten vorgezogen werden. Im rechten Bild ist die Anpassung der Ebene trotz Richtigklassifikation aller Beobachtungen als unangemessen zu beurteilen, da angenommen werden kann, dass sich das vorliegende Muster lediglich auf den Kreis in der Mitte beschränkt und die beiden falsch klassifizierten Beobachtungen als Ausreißer zu bewerten sind. Die Richtigklassifikation aller in die Berechnung von SVM einfließenden Beobachtungen kann daher nicht als Bewertungskriterium herangezogen werden. Vielmehr sollte im Sinne der auf die Durchführung von SVM folgenden Marketingimplikationen auf eine gute Prognosegüte geachtet werden.
Ein weiterer Grund für das Auftreten von Overfitting neben falscher Parameterwahl oder fehlenden Testdaten kann eine zu kleine Anzahl an Trainingsdaten in Verbindung mit einer vergleichsweise hohen Anzahl an beschreibenden Merkmalen sein, auf denen die Modelle basieren. Diese können somit nicht angemessen verallgemeinert werden.
Von Overfitting kann insbesondere dann ausgegangen werden, wenn eine Trennebene auf Trainingsdaten bessere Ergebnisse liefert, hingegen bei Testdaten schlechter abschneidet als eine vergleichbare Trennebene. Die Auswahl des Datensatzes erfolgt meistens in der Art, dass die ursprünglichen, zur Generierung der Trennebene heranzuziehenden Daten zufällig in einen Trainings- und einen Testdatensatz geteilt werden. Um sich nicht auf einen zwar zufällig ausgewählten aber eventuell nicht repräsentativen Testdatensatz beschränken zu müssen, bietet sich hier die Kreuzvalidierung an. Bei der q-fachen Kreuzvalidierung wird der komplette Datensatz in q (mit $q \leq l,\ q \in \mathbb{N}$) Teile der Größe $\frac{l}{q}$ geteilt[5], wovon in q Wiederholungen

[5] l bezeichnet hier die Anzahl der im kompletten Datensatz vorhandenen Beobachtungen. Falls $\frac{l}{q} \notin \mathbb{N}$ wird so gerundet, dass alle Beobachtungen berücksichtigt werden.

jeweils einer der q Teile als Testdatensatz dient und die restlichen $q-1$ Teile zum Training verwendet werden. Dadurch kann vermieden werden, dass es zu Fehlern bei der Angabe der Prognosegüte kommt, die auf das zufällige Ziehen eines Testdatensatzes zurückzuführen sind. Als Prognosegüte bzw. Trefferquote wird hierbei der Mittelwert der ermittelten q Trefferquoten angegeben. Um die Zuverlässigkeit der gewonnenen Ergebnisse zu erhöhen, wird häufig die einfache Kreuzvalidierung durch eine q-malige Wiederholung einer q-fachen Kreuzvalidierung ersetzt, sodass das gemittelte Ergebnis aussagekräftiger wird (*Witten, Frank* (2000)).

Eine weitere Möglichkeit, Overfitting zu vermeiden, ist die so genannte Leave-One-Out (LOO)-Methode (z.B. *Joachims* (2002)). Dabei werden bei l vorhandenen Beobachtungen im Datensatz l verschiedene SVM auf Basis von $l-1$ Beobachtungen berechnet. Die letzte fehlende Beobachtung wird dann mittels der Entscheidungsfunktion klassifiziert, sodass jeder der l Vektoren einmal klassifiziert wird und somit eine Trefferquote für die gewählte Parameterkonstellation angegeben werden kann. Diese Methode entspricht einer l-fachen Kreuzvalidierung. Vorteil dieses Verfahrens ist die Unabhängigkeit von zufälligen Ziehungen und die Berücksichtigung einer maximalen Menge an Trainingsdaten. Dem entgegen steht allerdings der enorme Aufwand, den die für ein Ergebnis benötigten SVM mit sich bringen, und der bei großen Datensätzen nur schwer zu bewältigen ist.

Eine systematische Vorgehensweise bei der Bestimmung der Parameter auf Basis von Kreuzvalidierung oder LOO bildet das so genannte Gridsearch (*Hsu et al.* (2003)). Dabei bilden die Ausprägungen der Parameter ein gleichmäßiges Netz, bei dem in jedem Knoten, also jeder Kombination der vorher festgelegten Parameterausprägungen, eine Trefferquote ermittelt wird. In Abbildung 3.6 wurde dies für die Kombinationen der Parameter γ und C mit den Ausprägungen $\gamma, C \in \{0,001; 0,005; 0,01; 0,05; 0,1; 0,5; 1; 5; 10; 50; 100; 500; 1000\}$ durchgeführt. Jeder Wert der Trefferquote auf der Oberfläche entspricht dem gemittelten Ergebnis 6-maliger 6-fach-Kreuzvalidierung, um Effekte, die auf das zufällige Ziehen eines Testdatensatzes zurückzuführen sind, auszuschließen. Wird nur ein Parameter variiert, so liegt ein Linesearch vor, da kein Netz, sondern eine Gerade von den möglichen Ausprägungen des Parameters aufgespannt wird. Eine grobe Parameterauswahl kann durch ein detailliertes Gridsearch verfeinert werden, wenn sich bei der groben Suche Erfolg versprechende Regionen herauskristallisiert haben. Die Suche sollte solange verfeinert werden, bis eine sehr gute Parameterkonstellation gefunden wurde. Für jede Kombination wird die Kreuzvalidierung durchgeführt, um zufällige Auswahlfehler weitestgehend auszuschließen. Werden beim Gridsearch die optimalen Ausprägungen zweier Parameter (etwa C und Kernparameter γ) gesucht, so ergeben sich bei jeweils 10 Ausprägungen und einer 5-fachen Kreuzvalidierung bereits $10 \times 10 \times 5 = 500$ zu berechnende SVM, um die grobe Region für potentiell optimale Parameter zu bestimmen. Dies zeigt, dass sich Gridsearch zwar durch ein sauberes Vorgehen und eine relativ sichere Bestimmung der optimalen Parameter auszeichnet, allerdings durch die vielen zu berechnenden SVM sehr zeitaufwändig ist. Dabei ist der Umfang von 10 Ausprägungen bei einer 5-fachen

Kreuzvalidierung eher als gering einzuschätzen. Ein weiterer Nachteil liegt in der Bestimmung der Trefferquote. Häufig ist die Aufsplittung der Trefferquote in klassenbezogene Trefferquoten notwendig. Beim herkömmlichen Gridsearch wird dies allerdings vernachlässigt und eine gesamte Trefferquote maximiert. Dies kann dazu führen, dass bei ungleich großen Klassen eine weniger umfangreiche Klasse vernachlässigt wird und zu Gunsten einer insgesamt höheren Trefferquote komplett falsch klassifiziert wird. Die zu maximierende Trefferquote kann dadurch unter Umständen zwar erhöht werden, aber das eigentliche Ziel der Untersuchung wird möglicherweise verfehlt. Nähere Ausführungen zu diesem Aspekt finden sich in Abschnitt 3.8.

Aufgrund des enormen Zeitaufwandes, den ein Gridsearch in Verbindung mit der Kreuzvalidierung mit sich bringt, sind Möglichkeiten zur Beschleunigung dieses Verfahrens sowie Alternativen dazu von großem Interesse. In den letzten Jahren haben sich einige Wissenschaftler mit dieser Fragestellung auseinander gesetzt, von denen an dieser Stelle einige erwähnt seien. So können *Chapelle et al.* (2002) etwa die Anzahl der zu berechnenden SVM deutlich verringern, indem sie statt der Trefferquote auf einem festgelegten Testdatensatz die Grenze des Generalisierungsfehlers minimieren. Das folgende in *Vapnik* (1998) erzielte Resultat dient dazu als Grundlage. Es gilt, dass die Fehlerquote Err von SVM durch die Ungleichung $Err \leq \frac{R^2}{\tilde{M}^2}$ abgeschätzt werden kann. Dabei bezeichnet $\tilde{M}$ die Spanne zwischen den zu trennenden Klassen im Merkmalsraum, und R^2 beschreibt den Radius der kleinsten Kugel, die die Daten im Merkmalsraum umfasst.[6] Diese Werte sind beide von den Parametern abhängig, sodass sie als Funktion derselben dargestellt werden können. Eine große Spanne bei der Berechnung der Hyperebene zwischen den Klassen ermöglicht eine bessere Klassifikation von Testdaten (*Schölkopf, Smola* (2002)). Durch ein Gradientenabstiegsverfahren ermitteln *Chapelle et al.* (2002) den minimalen Wert des Quotienten $\frac{R^2}{\tilde{M}^2}$ und somit die Werte der Parameter. Die mit den Resultaten der Kreuzvalidierung vergleichbaren Klassifikationsergebnisse erreichen sie durch eine wesentlich kleinere Anzahl an Iterationen. So sind in der von ihnen gewählten Anwendung nicht mehr 500 Berechnungen von SVM erforderlich, sondern nur noch etwa 15. Eine vergleichbare Vorgehensweise ist ebenfalls bei *Keerthi* (2002) zu finden.
Eine gleichfalls auf obigen Ansatz zurückgehende Möglichkeit bieten *Teow, Loe* (2000). In ihren Analysen stellt sich heraus, dass sich das obige Kriterium aufgrund der hohen Abweichung des generierten Fehlers im Vergleich zur Berechnung der Trefferquoten von Testdaten nicht gut zur Festlegung der Kernparameter eignet. Sie integrieren daher zusätzlich die bei der Diskriminanzanalyse zugrunde liegende Idee, die Inner-Klassen-Varianzen zu minimieren bei gleichzeitiger Maximierung der Zwischen-Klassen-Varianzen, um eine Verbesserung der Prognosegüte zu erreichen. Die Berücksichtigung der Varianzen spielt bei der klassischen Form der SVM keine Rolle und wird von *Teow, Loe* zur Verbesserung der Prognosegüte eingeführt. Sie legen dazu ein erweitertes Kriterium bei der Bestimmung der Parameter

[6] In *Vapnik* (1998) wird gezeigt, dass R^2 mit Hilfe der Kernfunktion berechnet werden kann.

zugrunde, welches durch eine Kombination der Klassenvarianzen und des obigen Kriteriums gebildet wird. Dadurch erreichen sie eine geringere Fehlerquote als bei der herkömmlichen Variante. Eine Zeitersparnis bringt diese Methode nicht.
Eine auf Gridsearch zurückgreifende Art der Beschleunigung des Verfahrens stellen *Ou et al.* (2003) vor. Hierbei wird nicht die eigentliche Methode modifiziert, sondern der Datensatz verkleinert. Dies erfolgt in der Art, dass nur diejenigen Vektoren selektiert werden, die auch einen Einfluss auf die resultierende Hyperebene haben. Vektoren, die einen großen Abstand zu ihrem „nächsten Gegner“ haben, werden als unwichtig für die Klassifikation eingestuft. Als „nächster Gegner“ ist derjenige Vektor der jeweils anderen Klasse zu verstehen, der am nächsten, gemessen durch die euklidische Distanz, am betrachteten Vektor gelegen ist. Die Rolle der übrigen Vektoren ist mit denen der Support Vektoren vergleichbar. Kritisch ist hierbei die Verbindung der Datenreduktion mit der Kreuzvalidierung zu sehen. Dieses Verfahren sollte nur eingesetzt werden, wenn die Menge der zur Verfügung stehenden Daten hinreichend groß ist. Durch die Reduktion der Daten werden einzelne Berechnungen innerhalb des Gridsearch beschleunigt.
Zusammenfassend kann festgehalten werden, dass Gridsearch eine systematische Vorgehensweise bei der Suche nach guten Parameterkombinationen bildet, die sich bei realen Anwendungen allerdings als sehr aufwändig gestaltet.

Eng verbunden mit der Suche nach guten Parametern ist die Wahl des einzusetzenden Kerns. In vielen Veröffentlichungen werden Polynom, die Radialbasis-Funktion und sigmoide Funktionen als mögliche Kerne genannt. Allerdings können a priori keine Empfehlungen dahingehend gegeben werden, welcher Kern bei vorliegenden Daten die besten Ergebnisse erwarten lässt. Dies sollte im Zusammenspiel mit der Parameterwahl mittels Kreuzvalidierung ermittelt werden (*Müller et al.* (2001)). Falls eine nahezu lineare Abhängigkeitsstruktur vermutet wird, kann als Vergleichsmaß das Ergebnis der linearen SVM dienen, da diese sehr einfach zu berechnen sind. Trotz einer breiten Auswahl an Kernen kommt in den meisten Veröffentlichungen zur Anwendung von SVM in unterschiedlichsten Bereichen lediglich die Radialbasis-Funktion zum Einsatz. Dies liegt darin begründet, dass damit neben sehr guten Klassifikationsergebnissen eine schnelle Approximation innerhalb der Optimierung gegeben ist, und sich der Radialbasis-Kern daher für den Einsatz auf umfangreichen Datensätzen besonders gut eignet. Auch bei den im Rahmen dieser Arbeit durchgeführten Analysen hat sich die Leistungsstärke dieses Kerns gezeigt, sodass dieser für die Praxis empfohlen wird, solange die Anwendung nicht den Einsatz eines speziellen Kerns erfordert. Daher soll im Folgenden kurz auf den Einfluss des zu wählenden Kernparameters γ eingegangen werden.
Die obigen Ausführungen zur Parameterwahl basieren alle auf einer umfangreichen Untersuchung des jeweils zugrunde liegenden Datensatzes, um eine möglichst gute Konfiguration der Parameter zu finden. Die Ausprägungen der Kern- und Kostenparameter hängen sehr stark vom Umfang und einer möglichen Normierung der Merkmale sowie der Trennbarkeit der Daten ab (vgl. Abbildung 3.6). Abbildung 3.8 zeigt dazu die Wirkung einer systematischen Veränderung des Kernparameters γ der Radialbasis-Funktion auf die Trefferquote eines exemplarischen Trainings-

und Testdatensatzes bei festem Kostenparameter C.[7]

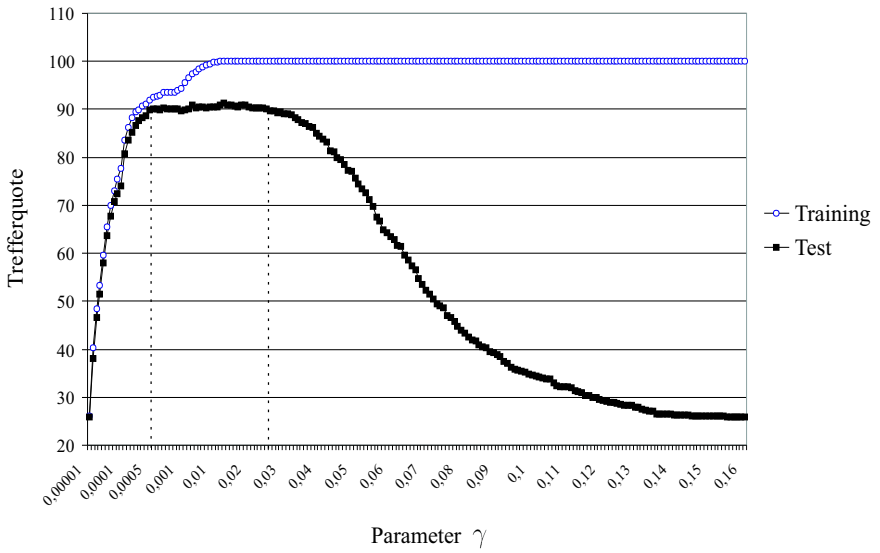

Abbildung 3.8: Auswirkung der Erhöhung des Radialbasis-Kernparameters auf die Trefferquote (bei $C = 1$)

Mit zunehmendem Wert für γ verbessert sich die Anpassung der Ebene an die Daten, was sich in einer erhöhten Trefferquote für die Trainingsdaten ausdrückt. Bei etwa $\gamma = 0,015$ erreicht die obere Kurve ihr Maximum von 100%, was einer perfekten Anpassung der Ebene an die vorliegenden Daten entspricht. Die eigentlich interessanten Entwicklungen hingegen liegen bei den Testdaten vor. Die Prognosegüte verbessert sich zunächst, nimmt jedoch mit wachsendem Kernparameter wieder ab, was auf ein zunehmendes Overfitting zurückzuführen ist. Bei der Wahl eines geeigneten Parameters gilt es also, eine möglichst gute Anpassung an die Daten bei gleichzeitig hoher Prognosegüte zu finden. Dieser Bereich liegt im vorliegenden Beispiel etwa zwischen $\gamma = 0,0005$ und $\gamma = 0,024$ (markierter Bereich innerhalb Abbildung 3.8). Wie entsprechende Analysen gezeigt haben, sind ähnliche Strukturen im Verlauf der Kurven auch bei anderen Datensätzen zu erwarten.
Neben den Kernparametern muss die Kostenvariable C festgesetzt werden. Die Durchführung des Gridsearch kann dadurch vereinfacht werden, dass bei vorgegebenen Kernparametern der Kostenparameter C so groß gewählt wird, dass die Trainingsdaten fehlerfrei separierbar sind, soweit dies möglich ist. Falls die Daten nur sehr schwer trennbar sind, so scheidet diese Option zur Wahl der Parameter aus, da hier die Berechnung einer optimal trennenden Hyperebene für große Werte für C sehr zeitaufwändig ist. Im fehlerfrei trennbaren Fall kann der maximal erreichte Wert für die Lagrange-Multiplikatoren α_i bei festem Kernparameter als Obergrenze für den zu wählenden Parameter C herangezogen werden. Wird ein höherer Kostenparameter gewählt, so hat dies keinen Einfluss auf die Lage der Ebene, sodass unnötige

[7] Der hier zugrunde liegende Datensatz wird in Abschnitt 4.4.4 ausführlich ausgewertet.

Berechnungen innerhalb des Gridsearch vermieden werden können. In Abbildung 3.9 wird beispielhaft eine Trennung von zwei durch zwei Merkmale beschriebenen Klassen vorgenommen, um die Auswirkung einer Variation der Kostengröße C auf die Lage der Ebene zu visualisieren. Der für die Trennung der simulierten Daten eingesetzte Kern ist hier ein Polynom zweiten Grades.

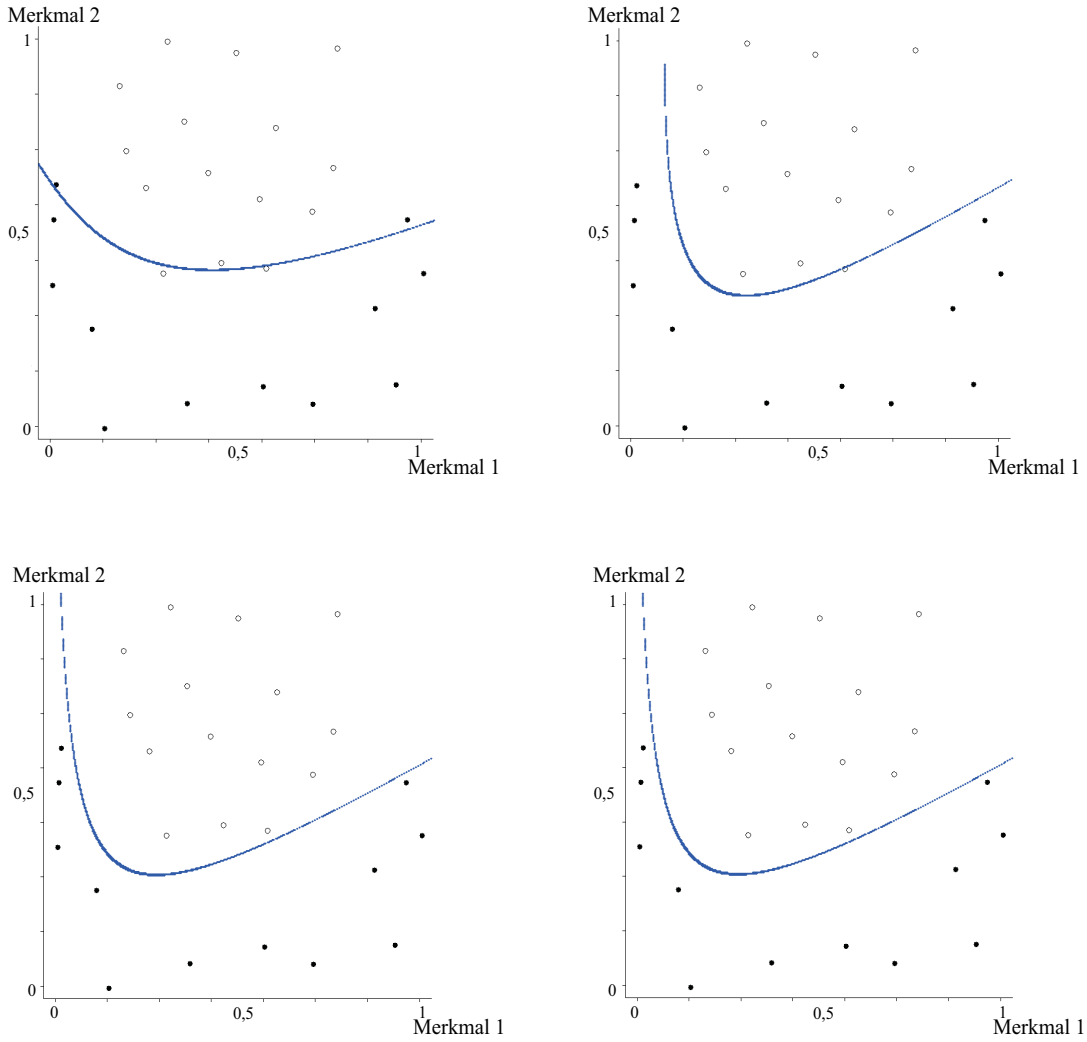

Abbildung 3.9: Einfluss der Variation des Kostenparameters C auf die Lage der Ebene bei festem Kernparameter $d = 2$ der polynomiellen Trennung ($C = 10,\ 100,\ 1327,\ 1500$)

Der Kostenparameter C wird hierbei von $C = 10$ über $C = 100$ und $C = 1327$ bis zu $C = 1500$ variiert. Mit zunehmendem Wert des Kostenparameters passt sich die Lage der Ebene besser den Daten an. Bereits bei $C = 100$ ist eine gute Anpassung der Ebene an die Daten erreicht. Der maximale Wert α_{max} für α_i wird in diesem Beispiel bei hinreichend großem C bei $\alpha_{max} \approx 1327$ angenommen. Wird $C > \alpha_{max}$ gewählt, so hat dies keinen Einfluss mehr auf die Lage der Ebene, was beim Vergleich der unteren beiden Bilder ($C = 1327$ und $C = 1500$) zu erkennen ist, sodass

diese Bildelemente die gleichen Trennebenen beinhalten. Bei realen Anwendungen tritt dieser Fall aufgrund fehlerhaft klassifizierter Daten allerdings nur selten auf. Daher wird dies im empirischen Teil der Arbeit (Kapitel 4) nicht betrachtet, da davon ausgegangen wird, dass bei den Datensätzen mindestens eine Fehlklassifikation vorliegt und somit immer $\alpha_{max} = C$ gilt.
Die Wahl der Kostengrößen c_+, c_- und c_i hängt sehr stark von der jeweiligen Fragestellung ab und kann allgemein nicht vereinfacht werden. Steht allerdings bei der Verwendung der ungleichen Kosten die Behandlung ungleicher Klassengrößen im Vordergrund, so sollte das Verhältnis $\frac{c_+}{c_-}$ dem Verhältnis $\frac{NEG}{POS}$ entsprechen, wobei POS bzw. NEG die Anzahl der Beobachtungen in Klasse „+1“ bzw. in Klasse „−1“ bezeichnen (vgl. *Morik et al.* (1999)). Bei einer Multiklassifikation sollten die Verhältnisse analog entsprechend der jeweiligen Klassengrößen gewählt werden, sodass ein Ungleichgewicht durch die Kostengrößen ausgeglichen wird. Dies kann etwa so geschehen, dass der Gewichtungsfaktor c_k von Klasse k bei der OAA-Trennung als

$$c_k = \frac{l - l^{[k]}}{l^{[k]}}$$

gewählt wird, wobei $l^{[k]}$ die Anzahl der Beobachtungen in Klasse k angibt. Soll zusätzlich zu dieser Angleichung eine Klasse besonders hervorgehoben werden, so muss der entsprechende Kostenparameter erhöht werden. Inwieweit diese Erhöhung vorgenommen werden muss, kann ebenfalls mittels Gridsearch oder Linesearch herausgefunden werden, da diese Hervorhebung ebenso wie die Wahl der Kernparameter sehr stark an die Struktur der Daten gebunden ist.

Zusammenfassend wird in Tabelle 3.1 ein Überblick über die empfohlene Vorgehensweise bei der Parameterbestimmung gegeben.

Schritt	Vorgehen
1	Festlegung der zu überprüfenden Kerne (empfohlen wird RBF)
2	Festlegung der Kosten- und Kernparameter Falls eine ungleiche Klassenverteilung vorliegt, Bestimmung der Kostengrößen im Verhältnis $\frac{c_+}{c_-} = \frac{POS}{NEG}$
3	Durchführung des groben Gridsearch nicht linearer SVM für jeden ausgewählten Kern
4	Festlegung einer Erfolg versprechenden Region innerhalb des jeweils untersuchten Bereiches für jeden Kern
5	Verfeinerte Suche innerhalb dieser Regionen
6	Auswahl des besten Kerns und der Parameter auf Basis der erreichten Klassifikationsgüte
7	Training der SVM auf Basis des ausgewählten Kerns und der gefundenen Parameter

Tabelle 3.1: Mögliches Vorgehen bei der Parameterwahl bei nicht linearer Trennung

Nach Festlegung der zu überprüfenden Kerne und entsprechender Parameter (Schritt 1 und 2) wird in Schritt 3 zunächst ein grobes Gridsearch durchgeführt. Nachdem neben der Auswahl des Kerns auch die Kernparameter in einem feinen Gridsearch festgelegt wurden (Schritt 4-6), erfolgt das eigentliche Training der SVM auf Basis dieser ermittelten Konstellationen (Schritt 7). Es sollte darauf geachtet werden, dass die letztendlich verwendeten Trainings- und Testdaten den gleichen Umfang wie die bei Gridsearch verwendeten Daten haben, um eine Veränderung der optimalen Parameterkonstellation, die auf ungleich große Datensätze zurückzuführen ist, auszuschließen. Bei veränderter Größe des Trainingsdatensatzes sind bei gleichen Parameterwerten andere Ergebnisse zu erwarten.

3.4 Online Learning

Gerade im Marketing ist es wünschenswert, dass die Möglichkeit zum „Online-Lernen“ besteht. Darunter wird die Anpassung eines Lernalgorithmus an sich im Zeitablauf verändernde, in diesem Fall vergrößernde, Datengrundlagen verstanden. Diese Variante sollte dann eingesetzt werden, wenn die zu klassifizierenden Daten nicht vollständig gegeben sind, sondern sich sukzessive innerhalb mehrerer Perioden erweitern und ein wiederholtes Training der SVM nötig wäre, um die Prognosegüte zu optimieren. Da diese Situation insbesondere bei der Erfassung von Warenkörben durch POS-Scannerdaten gegeben ist, soll im Folgenden diese Art der Erweiterung der SVM zur Nutzung innerhalb des Marketing erläutert werden. Bei SVM wird diese Möglichkeit der schrittweisen Verarbeitung von Beobachtungen bisher nur sehr selten verwendet, obwohl die Eigenschaft, die Datenbasis auf nur wenige Support Vektoren zu reduzieren, gut ausgenutzt werden kann.
Ein möglicher Ansatz wird bei *Syed et al.* (1999) verfolgt. Der ursprüngliche Datensatz wird in gleich große Teile $\mathcal{T}_1, \ldots, \mathcal{T}_t$ geteilt, um SVM zunächst nur auf $\mathcal{T}_1$ zu trainieren. Aus $\mathcal{T}_1$ werden nun die Support Vektoren ausgewählt. Diese Menge wird zu $\mathcal{T}_2$ hinzugefügt. Auf diesen Daten wird trainiert und daraufhin wiederum die resultierenden Support Vektoren ausgewählt. Dieses iterative Vorgehen wird bis $\mathcal{T}_t$ fortgesetzt. Die Autoren zeigen, dass diese Art der sukzessiven Erweiterung des Trainingsdatensatzes nur zu einem geringen Verlust an der durch die Trefferquote gemessenen Genauigkeit führt. Diesen Ansatz erweitert *Rüping* (2001) dahingehend, dass er die Vorgehensweise mit der Hinzunahme von individuellen Kosten kombiniert. Support Vektoren aus vorangegangenen Teilen werden bei einer Fehlklassifikation mit einem zusätzlichen Kostenparameter belegt und damit höher bestraft. In beiden Arbeiten wird allerdings nicht das eigentliche „Online Learning“, wie es hier verstanden werden soll, durchgeführt, sondern lediglich ein schrittweises Training einer SVM vollzogen.
In *Lau, Wu* (2003) wird das „Online Learning“ explizit durchgeführt. Im Vordergrund stehen die KKT-Bedingungen, die beim grundlegenden Optimierungsproblem im regulären Fall vorliegen.

Aus diesen folgt:

$$\alpha_i = 0 \Leftrightarrow y_i F(\boldsymbol{x}_i) > 1 \tag{3.5}$$

und

$$0 < \alpha_i \leq C \Leftrightarrow y_i F(\boldsymbol{x}_i) \leq 1. \tag{3.6}$$

Die rechte Seite der Äquivalenzrelation (3.5) sagt aus, dass die Beobachtung $\boldsymbol{x}_i$ richtig klassifiziert worden ist und zudem einen Entscheidungswert größer eins hat, also jenseits der Hilfsebene liegt. Diese Vektoren gehören nicht zu den Support Vektoren und erhalten demnach den Wert $\alpha_i = 0$. Im anderen Fall (Relation (3.6)) kann entweder eine Falschklassifikation der Beobachtung $\boldsymbol{x}_i$ vorliegen, oder $\boldsymbol{x}_i$ ist zwar in die richtige Klasse eingeordnet worden ($y_i F(\boldsymbol{x}_i) > 0$), liegt aber innerhalb der Spanne, die ursprünglich maximiert wurde ($|F(\boldsymbol{x}_i)| < 1$). Die hier gegebene Situation wurde bereits in Abbildung 2.5 in Abschnitt 2.1.3 erläutert.
Der von *Lau, Wu* vorgestellte Algorithmus ist nur für die binäre Klassifikation anwendbar. Daher wird er im Folgenden für die Verwendung innerhalb der Multiklassifikation angepasst. Die beiden Gleichungen (3.5) und (3.6) charakterisieren die Beobachtungen, die für den Algorithmus von Bedeutung sind. Die Basis bildet eine Menge $\mathcal{B}_0$ von Vektoren, wobei jeder der vorliegenden K Klassen mindestens ein Vektor zugeordnet sein muss. Für alle Vektoren $\boldsymbol{x}_{neu}$, die neu hinzukommen und für die die wahre Klassenzugehörigkeit $y_{neu} = k$ bereits bekannt ist, werden neue SVM berechnet, falls die rechte Seite von Gleichung (3.6) für mindestens eine der zu berechnenden Trennebenen erfüllt ist. Falls jedoch für jede dieser SVM $y_{neu} F^{[kk']}(\boldsymbol{x}_{neu}) > 1$ (für alle $k' \neq k$) bei der OAO-Trennung bzw. $y_{neu} F^{[k]}(\boldsymbol{x}) > 1$ für die OAA-Trennung gilt, so ist die Beobachtung richtig klassifiziert. Weiterhin befindet sie sich nicht zwischen den jeweiligen Hilfsebenen. Demnach ist diese Beobachtung nicht zu den Support Vektoren zu zählen, wenn die SVM auf allen Daten gerechnet werden würde. Daher ist in diesem Fall keine neue Berechnung erforderlich, und der Datensatz behält seine ursprüngliche Zusammensetzung bei. So kann eine sukzessive Verarbeitung stückweise ankommender Daten auch bei SVM effektiv verarbeitet werden, ohne kontinuierlich neue Trainingsdatensätze zu bilden, die redundante Informationen enthalten können. Zusätzlich zu den neu hinzukommenden Beobachtungen wird bei den in vorangegangenen Iterationen als richtig klassifizierten Datenpunkten dieser Status ebenfalls überprüft. Hat sich die Ebene (bzw. die Ebenen) innerhalb der Iterationen derart verändert, dass diese Beobachtungen nun falsch klassifiziert werden, so müssen diese Beobachtungen im nächsten Schritt erneut mit in die Optimierung aufgenommen werden. Die folgende Tabelle 3.2 gibt den Pseudo-Code für die Vorgehensweise beim Online Learning in Anlehnung an *Lau, Wu* (2003) an, der für den allgemeineren Fall der Multiklassifikation erweitert wurde. Dabei bezeichne $\mathcal{M}$ die Menge der Beobachtungen, die mittels dieses Algorithmus in die Optimierung integriert werden sollen.

Schritt	Aktion
1.	$\mathcal{B}_0 := \{(\boldsymbol{x}_1, y_1),\ (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_K, y_K), \ldots\}$ bildet die Basis (aus jeder der K Klassen liegt mindestens eine Beobachtung vor)
2.	Bestimme alle SVM auf Basis von $\mathcal{B}_0$; $\mathcal{E}_0 := \{\}$
3.	Setze $i = 1$
4.	Solange $\mathcal{M}$ nicht leer ist
4.1.	Ziehe neue Beobachtung $(\boldsymbol{x}_{neu}, y_{neu})$ aus $\mathcal{M}$
4.2.	$\mathcal{M} = \mathcal{M} \setminus \{\boldsymbol{x}_{neu}\}$
4.3.	Falls $y_{neu} F(\boldsymbol{x}_{neu}) \leq 1$ für mindestens eine betreffende SVM: $\mathcal{B}_i := \mathcal{B}_{i-1} \cup \boldsymbol{x}_{neu} \cup \mathcal{E}_{i-1}$ sonst: $\mathcal{B}_i := \mathcal{B}_{i-1} \cup \mathcal{E}_{i-1}$
4.4.	Falls $\mathcal{B}_i \neq \mathcal{B}_{i-1}$: Berechne die betreffenden SVM auf Basis von $\mathcal{B}_i$
4.5.	$\mathcal{B}_i := \{\boldsymbol{x}_j \vert \alpha_j \neq 0$ bei mindestens einer SVM in Iteration $i\}$
4.6.	Bestimme $\mathcal{E}_i$
4.7.	Setze $i := i + 1$
5.	Berechne SVM auf Basis der letzten Menge $\mathcal{B}_i \cup \mathcal{E}_i$

Tabelle 3.2: Pseudo-Code für Online Learning bei Multi-SVM in Anlehnung an *Lau, Wu* (2003)

Innerhalb des Pseudo-Codes werden die Beobachtungen, die bei der erneuten Berechnung der SVM als falsch klassifiziert werden, aber bisher bei der Optimierung nicht berücksichtigt wurden, zu der Menge $\mathcal{E}_i$ zusammengefasst. Da die Vorgehensweise auch für die Trennung bei der Multiklassifikation gültig sein soll, muss in Schritt 4.3 für jede der die Beobachtung $\boldsymbol{x}_{neu}$ betreffenden SVM die Bedingung überprüft werden. Im Falle der Trennung mittels OAA sind dies alle K durchzuführenden SVM (bei K Klassen), da eine Beobachtung entweder der betreffenden Klasse oder den restlichen Beobachtungen angehört. Bei einer OAO-Trennung ist hingegen nur ein kleinerer Teil der $\frac{K(K-1)}{2}$ SVM in Abhängigkeit von y_{neu} betroffen[8]. Weiterhin ist die Neuberechnung der Ebenen in Schritt 4.4. nur dann erforderlich, wenn die Klasse y_{neu} involviert ist oder die Klassenzugehörigkeiten der Beobachtungen aus $\mathcal{E}_{i-1}$ tangiert werden. Abschließend erfolgt eine Berechnung aller erforderlichen SVM auf Basis der zuletzt berechneten Mengen $\mathcal{B}_i \cup \mathcal{E}_i$, um eine endgültige Lösung zu erhalten.
In Abbildung 3.10 werden die erste Schritte des Algorithmus aus Tabelle 3.2 dargestellt. Dazu ist die Menge $\mathcal{B}_0 = \{\boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \boldsymbol{x}_3\}$ gegeben, wobei jede Beobachtung einer anderen Klasse angehört.

[8]Beispiel: Bei $y_{neu} = 3$ und $K = 5$ sind nur die SVM „1 vs 3“, „2 vs 3“, „3 vs 4“ und „3 vs 5“ betroffen, nicht aber die SVM „1 vs 2“ „1 vs 4“ „2 vs 4“ „2 vs 5“ und „4 vs 5“.

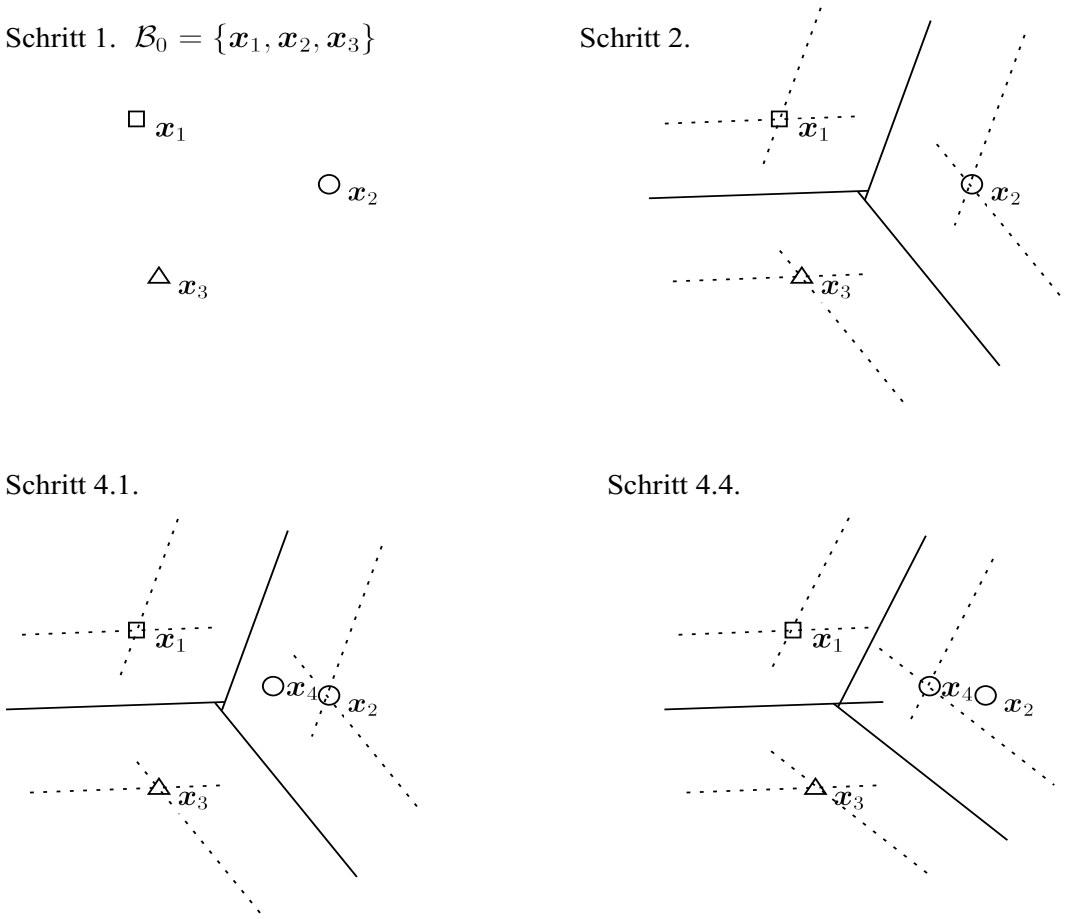

Abbildung 3.10: Darstellung des Ablaufs des Algorithmus beim Online-Learning

In Schritt 2. werden die Trennebenen auf Basis des OAO-Verfahrens gebildet. Eine neue Beobachtung $\boldsymbol{x}_4$ mit der gleichen Klassenzugehörigkeit wie $\boldsymbol{x}_2$ wird im dritten, den Schritt 4.1. repräsentierenden Teil der Abbildung hinzugefügt. Der letzte Teil der Abbildung enthält die eigentlichen Veränderungen, die durch die neue Beobachtung vorgenommen werden müssen. Da $\boldsymbol{x}_4$ sowohl bezüglich der Trennung von $\boldsymbol{x}_1$ und $\boldsymbol{x}_2$ als auch bezüglich der Trennung von $\boldsymbol{x}_2$ und $\boldsymbol{x}_3$ zwischen den eingezeichneten Hilfsebenen liegt, müssen diese beiden SVM neu berechnet werden (Abfrage 4.3. in Tabelle 3.2). Die dritte Trennebene ist von der Vorgabe der Klassenzugehörigkeit des vierten Objekts nicht betroffen und wird daher in Abbildung 3.10 nicht verändert. Zunächst verändert sich die Datenbasis zu $\mathcal{B}_1 = \mathcal{B}_0 \cup \boldsymbol{x}_4 = \{\boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \boldsymbol{x}_3,\ \boldsymbol{x}_4\}$. Auf dieser Grundlage werden die beiden Trennebenen neu berechnet. Es ergibt sich die in der Abbildung unten rechts gezeigte Situation. Da $\boldsymbol{x}_2$ nun weder bei der Trennung von $\boldsymbol{x}_1$ noch bei der Trennung von $\boldsymbol{x}_3$ zu den Support Vektoren gehört, wird diese Beobachtung aus der Datenbasis $\mathcal{B}_1$ gelöscht. Im vorliegenden Beispiel liegen keine fehlklassifizierten Beobachtungen vor, somit ist die Menge $\mathcal{E}_i$ leer, sodass bei Hinzunahme einer weiteren neuen Beobachtung $\mathcal{B}_1 = \{\boldsymbol{x}_1,\ \boldsymbol{x}_3,\ \boldsymbol{x}_4\}$ als Datengrundlage dient (Schritt 4.5.).

Da im Laufe der Iterationen zu erwarten ist, dass sich die Lage der Trennebene mit neuen Datenpunkten ändert, umfasst das erneute Ausrichten des Modells an den Daten in Schritt 4.4. insbesondere die Parameterbestimmung. Eine Orientierung an den optimalen Kombinationen der vorangegangenen Iterationen ist empfehlenswert, um den Aufwand zu minimieren. Die Parameter können ebenfalls vollständig übernommen werden. Dazu sind Daten zur Kalibrierung nötig, anhand derer die Güte der Anpassung angegeben werden kann. Neue Beobachtungen werden auf Basis des berechneten Modells klassifiziert und gehen bei Vorliegen der wahren Klassenzugehörigkeit als neuer Datenvektor in die Datenbasis ein. Weiterhin sei anzumerken, dass die Menge $\mathcal{M}$ nicht a priori vorliegen muss, sondern sich im Laufe der Anwendung des Algorithmus, z.B. bei POS-Scannerdaten, ständig erweitern kann (z.B. durch täglich hinzukommende Warenkörbe), sodass ein anderes Abbruchkriterium verwendet werden muss. Die Klassifikation könnte in diesem Fall in der Zuordnung der Kunden zu unterschiedlichen Gruppen bestehen, auf deren Basis eine zielgerichtete Verteilung von Coupons erfolgen könnte. Eine abschließende Optimierung der Anpassungsgüte (*Lau, Wu* (2003)) muss nach Abschluss des Verfahrens ebenfalls vorgenommen werden.

Durch die sukzessive Hinzunahme einzelner Beobachtungen wird unter Ausnutzung der Reduzierung des Algorithmus auf die Support Vektoren eine gute Möglichkeit bereit gestellt, ein Online Learning auch bei SVM durchführen zu können. Die Vorgehensweise kann mit wachsenden neuronalen Netzen verglichen werden (z.B. *Decker* (2005)), wo die Anzahl der die Daten repräsentierenden Neuronen mit zunehmender Datenmenge verändert bzw. angepasst wird. Die Minimierung der Anzahl der eingehenden Support Vektoren entspricht also etwa der Minimierung der Anzahl der nötigen Prototypen bzw. Neuronen.
Diese Online-Variante von SVM kann insbesondere dann verwendet werden, wenn es sich um die Entwicklung von Echtzeit-Anwendungen beispielsweise im Rahmen des E-Commerce handelt. In Abhängigkeit der Zugehörigkeit zu vorhandenen Kundengruppen oder Käuferprofilen können direkt bei Abschluss eines Kaufvertrages Empfehlungen für weitere Produkte gegeben werden und der jeweilige Kunde direkt mit in die Datenbasis aufgenommen werden. Dies stellt folglich eine Ergänzung der herkömmlichen Methoden wie Assoziationsregeln oder reine Häufigkeitsauszählungen im Rahmen von Recommender-Systemen dar.

3.5 Merkmalsauswahl

Die im Rahmen einer Kundenklassifikation zu klassifizierenden Kunden werden häufig durch eine Vielzahl von Merkmalen beschrieben. Diese können etwa durch den Einsatz von Kundenkarten und die somit generierten Kaufhistorien zustande kommen oder aber als Variablen in einer Kundendatenbank zur Verfügung stehen. Somit kann bei einer Kundenklassifikation unter Umständen auf eine große Anzahl an charakterisierenden Merkmalen zurückgegriffen werden, was zu einer verbesserten Beschreibung der Kunden und somit zu einer besseren Trennung der Kundenklassen

beiträgt. Auf der anderen Seite impliziert der Einsatz dieser Merkmale allerdings auch, dass für eine Beschreibung und letztendlich eine zuverlässige und gültige Zuordnung von bisher nicht klassifizierten Kunden die Ausprägung jedes der vorher ausgewählten Merkmale vorhanden sein muss. Um die Kosten und den Aufwand bei der Erhebung dieser Merkmale zu verringern, sollte die Menge an Variablen auf die nötigsten, für eine Klassifikation mit hoher Generalisierungsfähigkeit relevanten Merkmale reduziert werden. Im Folgenden werden daher Möglichkeiten zur Merkmalsreduktion vorgestellt, die im Rahmen der Klassifikation mittels SVM durchgeführt werden können. Es wird u.a. die Frage beantwortet, wie mittels SVM die für die Klassifikation wichtigen Merkmale identifiziert werden können.
Es gibt zwei Möglichkeiten, wie die Menge der zur Klassifikation heranzuziehenden Merkmale bestimmt werden kann (vgl. *Weston et al.* (2000)). Wird eine maximale Anzahl $n' < n$ an zu verwendenden Merkmalen vorgegeben, so sollen zum einen diejenigen n' der n zur Verfügung stehenden Merkmale gefunden werden, die beim vorliegenden Datenmaterial zu dem geringsten Generalisierungsfehler führen. Die zweite Möglichkeit besteht in der Vorgabe eines maximalen Generalisierungsfehlers, sodass die Merkmale derart ausgewählt werden, dass die Anzahl möglichst klein ist und die vorgegebene Schranke für den Fehler nicht überschritten wird. Dies setzt voraus, dass einer der beiden Parameter (Anzahl der zu extrahierenden Merkmale oder der maximale Fehler) bereits bekannt ist.
Die Auswahl der relevanten Merkmale sollte automatisch vorgenommen werden, da eine manuelle Überprüfung der Relevanz sehr schnell zu einer zeitintensiven Analyse auch bei insgesamt wenigen Variablen führen kann. Um beispielsweise 6 Merkmale aus 16 möglichen auszuwählen, bedarf es bereits 8008 SVM, um für jede mögliche Kombination von sechs Merkmalen die resultierende Trefferquote zu bestimmen.
Die folgenden Verfahren bilden auf Basis einer bereits auf allen zur Verfügung stehenden Merkmalen durchgeführten SVM eine Rangliste der Merkmale. Für jedes Merkmal i wird ein Score r_i in Abhängigkeit seiner Relevanz für die Klassifikation berechnet, sodass ein Vektor der ermittelten Scorewerte $\boldsymbol{r} = (r_1, \ldots, r_n)$ resultiert. Daraufhin können diejenigen Merkmale extrahiert werden, die für die Trennung der vorgegebenen Klassen von größter Bedeutung sind. Diese Verfahren zählen zum Backward-Verfahren, da zunächst alle möglichen Merkmale zur Trennung herangezogen werden, um dann sukzessive oder auf einmal reduziert zu werden.

Eine Möglichkeit zur Merkmalsauswahl respektive zur Bestimmung des Einflusses der verwendeten Merkmale auf die Trennung der Klassen ist bei der linearen Trennung mittels SVM gegeben. Bei der berechneten Entscheidungsfunktion $f(\boldsymbol{x}) = sign(\boldsymbol{w}\boldsymbol{x} + b)$ übernimmt der Normalenvektor eine entscheidende Rolle. Dies wird anhand des folgenden Beispiels deutlich, das in Abbildung 3.11 veranschaulicht wird. Für die Trennung der beiden Klassen ist lediglich das zweite Merkmal relevant. Hier gilt $\boldsymbol{w} = \binom{0}{1}$. Die Merkmale bei einer linearen Trennung fließen mit unterschiedlichen Gewichten in die Trennung ein, die anhand von $\boldsymbol{w}$ abgelesen werden können. So erhält das zweite Merkmal in diesem Beispiel ein Gewicht von 1, wohingegen das erste Merkmal nicht relevant ist. Allgemein

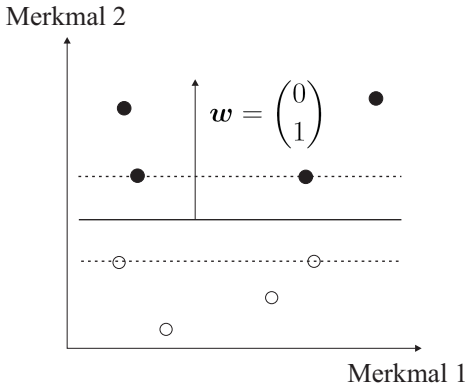

Abbildung 3.11: Beispielhafte Trennung zweier Klassen

gesprochen gilt: je höher der Eintrag w_i des Normalenvektors $\boldsymbol{w} = (w_1, \ldots, w_n)$, desto höher der Einfluss des Merkmals i auf die vorliegende Trennung, sodass hier einfach $\boldsymbol{r} = (r_1, \ldots, r_n) = \boldsymbol{w}$ gilt. Diese Art der Merkmalsgewichtung kann nach abgeschlossener Analyse Aufschluss über die Wichtigkeit unterschiedlicher Merkmale geben, sodass die gering gewichtigen aus darauf folgenden Analysen ausgeschlossen werden können. Das Einschränken der ursprünglich zur Verfügung stehenden Merkmale auf die wenigen wichtigen kann daher auch als eine Art abschließendes Pruning von SVM verstanden werden.
Eine solche Interpretation der Einträge des Normalenvektors ist allerdings nur bei linearer Trennung möglich. Bei nicht linearer Trennung werden die Daten zunächst in einen höher dimensionalen Raum transformiert, um dort linear getrennt zu werden. Da die Dimension dieses Raumes im Allgemeinen nicht bekannt ist, ist auch der Normalenvektor nicht bekannt. Da eine Darstellung des Normalenvektors $\boldsymbol{w}$ somit nicht ohne die implizit verwendete aber unbekannte Abbildung $\Phi(\cdot)$ möglich ist, kann der Einfluss der Merkmale nicht immer direkt abgelesen werden wie im linearen Fall. Bei der multiplikativen Verknüpfung von Merkmalen zur Berechnung der Entscheidungsfunktion lässt sich weder die Relevanz eines Merkmals ablesen, noch eine Rangfolge der Merkmale hinsichtlich ihrer Wichtigkeit bilden.
Eine mögliche analoge Erweiterung des linearen Ansatzes zur Ermittlung des Merkmalseinflusses auf die nicht lineare Trennung wurde von *Hermes, Buhmann* (2000) gegeben. Dort wird mittels Winkelberechnung die Abhängigkeit eines Richtungsvektors von den jeweiligen Einheitsvektoren ermittelt. Formal wird der Winkel $\lambda_j(\boldsymbol{x})$ zwischen dem Gradienten $\nabla F(\boldsymbol{x}) = \sum_{i \in \mathcal{SV}} \alpha_i y_i \nabla_{\boldsymbol{x}} K(\boldsymbol{x}_i, \boldsymbol{x})$ der Entscheidungsfunktion an der Stelle $\boldsymbol{x}$ und dem Einheitsvektor $\boldsymbol{e}_j$ ermittelt. Bei $\lambda_j(\boldsymbol{x}) \approx \frac{\pi}{2}$ ist die Trennung an der Stelle $\boldsymbol{x}$ vom Merkmal j unabhängig, d.h. j hat nur einen sehr geringen bzw. gar keinen Einfluss. Im Gegensatz dazu deuten geringe Werte für $\lambda_j(\boldsymbol{x})$ auf einen höheren Einfluss von j an der entsprechenden Stelle hin. In Abbildung 3.12 ist die Trennung von 2-dimensionalen Vektoren abgebildet. Der besonders gekennzeichnete Bereich zeigt, dass der Winkel λ_1 größer als λ_2 für den Vektor $\boldsymbol{x}$ ist. Demnach leistet das zweite Merkmal an dieser Stelle einen höheren

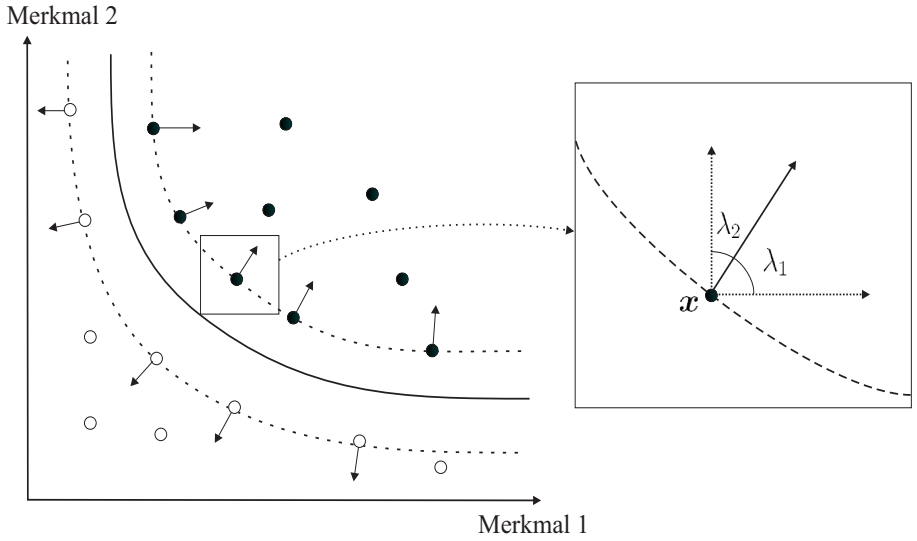

Abbildung 3.12: Nichtlineare Trennung von zwei Klassen mit Gradienten an Support Vektoren

Erklärungsbeitrag als das erste.[9]
Um eine generelle Aussage über die Wichtigkeit der einzelnen Merkmale für die gesamte Trennebene zu ermöglichen, wird eine Menge von Vektoren herangezogen, die für die Trennung der beiden zu betrachtenden Klassen relevant sind. Diese wird durch die Margin-Vektoren gebildet, die genau auf den beiden Hilfsebenen (vgl. Abschnitt 2.1) liegen und somit die Lage der Ebene unmittelbar beeinflussen. Um eine kritische Menge von Beobachtungen zusammenfassen zu können, werden nicht nur die Margin-Vektoren herangezogen, sondern alternativ alle Vektoren innerhalb einer ϵ-Umgebung um die Hilfsebenen betrachtet. Um ein gemitteltes Gewicht für Merkmal j zu erhalten, wird Folgendes berechnet:

$$r_j = 1 - \frac{2}{\pi} \frac{\sum_{i \in \mathcal{I}_\epsilon} \lambda_j(\boldsymbol{x}_i)}{|\mathcal{I}_\epsilon|}. \tag{3.7}$$

Hierbei bezeichnet $\mathcal{I}_\epsilon$ die Menge der Vektoren, die innerhalb dieser ϵ-Umgebung positioniert sind (*Hermes, Buhmann* (2000)). Je höher der Wert r_j für ein Merkmal j ausfällt, desto höher ist der Einfluss, den dieses Merkmal auf die Klassifikation, bzw. die Lage der Trennebene hat. Bei einer möglichen Merkmalsreduzierung würden demnach zunächst diejenigen Merkmale mit dem geringsten Wert ausgespart.
Aufgrund der Mittelbildung in Gleichung (3.7) gehen möglicherweise wichtige Informationen verloren, wie das folgende Beispiel verdeutlichen soll: Angenommen ein Unternehmen, welches hochpreisigen Schmuck verkauft, klassifiziert seine Kunden in potenzielle Käufer und Nichtkäufer eines seiner neuen Produkte. Die Kunden seien durch ihr Einkommen, die bisherigen Kauferfahrungen mit diesem Unternehmen und einige weitere Eigenschaften charakterisiert. Da die gut verdienenden Kunden

[9] In Abbildung 3.12 wird $\lambda_j(\boldsymbol{x})$ der Einfachheit halber mit λ_j abgekürzt.

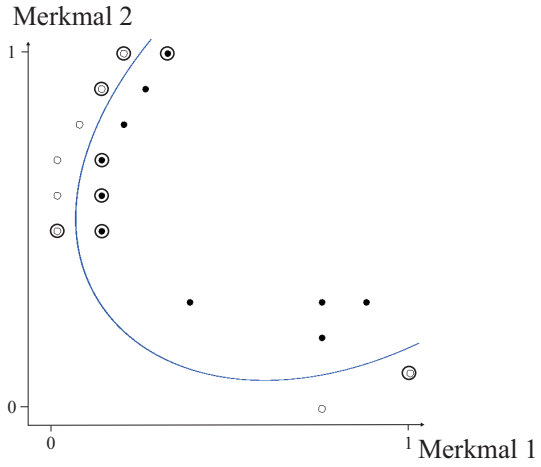

Abbildung 3.13: Auswirkung der Anzahl an Support Vektoren auf die Gewichtung von Merkmalen

aufgrund der Produktpalette ohnehin den typischen Kreis der Kundschaft bilden, wird sich bei einem an eine Klassifikation mittels SVM anschließenden Verfahren zur Merkmalsauswahl das Einkommen als eine eher unwichtige Größe herausstellen. Liegt jedoch ein Ziel des Unternehmens in der Richtigklassifikation gerade der kleineren Gruppe von weniger gut verdienenden Kunden, die sich dennoch ein Luxusgut gönnen möchten, so wäre es wichtig, das Einkommen als Merkmal im Datensatz zu belassen.
Abbildung 3.13 zeigt, dass es gerade bei gut trennbaren Daten aufgrund einer geringen Anzahl von Support Vektoren zu Informationsverlusten bei der Mittelbildung in Gleichung (3.7) kommen kann. In Bereichen mit wenigen Support Vektoren (in obigem Beispiel die weniger gut verdienenden Kunden) gehen die berechneten Winkel auch entsprechend mit einem geringeren Gewicht in die Berechnung des Gesamtgewichts eines Merkmals ein, als in Bereichen, die durch eine hohe Anzahl an Support Vektoren gekennzeichnet sind. Dies führt dann dazu, dass Merkmale, die höher gewichtet werden könnten, ein nur mittleres Gewicht erhalten. In Abbildung 3.13 lässt die Lage der Trennebene darauf schließen, dass beide Merkmale etwa gleich gewichtet werden. Allerdings erhält das erste Merkmal mittels des Gradientenverfahrens mit $r_1 = 0,85$ einen deutlich höheren Wert als das zweite Merkmal mit $r_2 = 0,15$. Dies ist auf die höhere $(7 > 1)$ Anzahl an Support Vektoren zurückzuführen, die für die Berechnung des höheren Einflusses von Merkmal 1 von Bedeutung sind. Diese Support Vektoren sind in Abbildung 3.13 mit einem Kreis gekennzeichnet. Im rechten unteren Bereich der Abbildung, der für die Gewichtung des zweiten Merkmals verantwortlich ist, befindet sich lediglich ein Support Vektor, was in einem deutlich geringeren Gewicht resultiert.
Ein Ausweg ist die von *Hermes, Buhmann* (2000) vorgeschlagene Berücksichtigung einer ϵ-Umgebung um den wichtigen Entscheidungswert $|F(\boldsymbol{x})| = 1$. Bei guter Generalisierungsfähigkeit wie im vorliegenden Fall müsste allerdings ϵ recht groß gewählt werden. Zusätzlich könnte neben dem mittleren Gewicht der an den

Vektoren innerhalb dieser Umgebung minimal erreichte Winkel λ_j berücksichtigt werden. Ist dieser verglichen mit dem durchschnittlichen Winkel an einigen Support Vektoren recht klein, so deutet dies auf einen hohen Einfluss des betreffenden Merkmals innerhalb einer bestimmten Region hin. In der Regel passen sich die Entscheidungsfunktionen bei der nicht linearen Trennung so den Daten an, dass ein globales Urteil über die Relevanz einzelner Merkmale nur schwer möglich ist. Daher könnte bei einer möglichen Reduzierung der Merkmale je nach Zielsetzung neben den ermittelten Score-Werten unter Umständen auch die Spanne der erreichten Winkel in Betracht gezogen werden.

Eine weitere Möglichkeit zur Bestimmung der relevanten Merkmale, die nicht nur auf den Einsatz von SVM ausgerichtet ist, ist durch den Fisher Criterion Score (FCS) gegeben (vgl. *Pavlidis et al.* (2001), *Guyon et al.* (2002)). Ausgehend von den einzelnen Merkmalen $i = 1, \ldots, n$ wird dazu die Wichtigkeit eines Merkmals i hinsichtlich der Trennung von zwei vorgegebenen Klassen durch

$$r_i = \frac{(\mu_i^+ - \mu_i^-)^2}{\sigma_i^{2(+)} + \sigma_i^{2(-)}}$$

ermittelt. Dabei wird durch μ_i^+ bzw. μ_i^- der Mittelwert der Ausprägungen des Merkmals i in Klasse „+1“ bzw. „-1“ und durch $\sigma_i^{2(+)}$ bzw. $\sigma_i^{2(-)}$ die zugehörige Standardabweichung in der entsprechenden Klasse angegeben. Dies ähnelt der Vorgehensweise bei der Berechnung des F-Wertes der ANOVA (*Guyon, Elisseeff* (2003)). Je größer der Einfluss eines Merkmals i auf die Trennung ist, desto höher fällt wiederum der Wert r_i aus.

Die bisher vorgestellten Methoden basieren auf einer Trennung zweier Klassen, sodass für die Merkmalsreduktion bei einer Multiklassifikation das Auswahlkriterium angepasst werden muss. Falls Verfahren zur Trennung mehrerer Klassen eingesetzt werden, die auf der Kombination mehrerer binärer SVM (z.B. OAO oder OAA) basieren, so ergibt sich für jede SVM eine Liste bzw. eine Rangfolge von Merkmalen, die für die jeweilige SVM entscheidend ist. Wird die Trennung jeder einzelnen Klasse untersucht, so können diese Listen zur Beurteilung und zur Auswahl der relevanten Merkmale herangezogen werden. Ist jedoch eine allgemeine Aussage zur Wichtigkeit der Merkmale bei der Trennung aller Klassen von Interesse, so müssen diese Listen angemessen kombiniert werden, um zu einem globalen Gewicht für jedes Merkmal zu gelangen. Dazu bieten sich mehrere Möglichkeiten an. In allen Fällen ist für jede zu berechnende SVM $\tilde{s}$ (mit $\tilde{s} = 1, \ldots, S_{SVM}$) ein Vektor mit Merkmalsgewichtungen $\boldsymbol{r}^{[\tilde{s}]} = (r_1^{[\tilde{s}]}, \ldots, r_n^{[\tilde{s}]})$ für die Merkmale $i = 1, \ldots, n$ gegeben, der aus oben vorgestellten Verfahren resultiert. Zum einen kann nun das Gewicht r_i für Merkmal i durch die Addition aller verfügbaren Gewichte bestimmt werden:

$$r_i = \sum_{\tilde{s}=1}^{S_{SVM}} r_i^{[\tilde{s}]}.$$

Zum anderen kann auch wie in *Hermes, Buhmann* (2000) ein über alle SVM gemittelter Wert der Relevanzen verwendet werden, was die Reihenfolge der Wich-

tigkeiten nicht verändert. Um diese Werte mit den Ergebnissen anderer Trennungen auf Grundlage der gleichen Merkmale und Klassen vergleichbar zu machen, kann der resultierende Vektor $\boldsymbol{r}$ durch $\boldsymbol{r}\frac{1}{\|\boldsymbol{r}\|}$ auf die Länge eins normiert werden. Zum anderen besteht die Möglichkeit, eine Kombination aus den für die einzelnen SVM als wichtig beurteilten Merkmalen zu generieren. So ist die Heranziehung der pro SVM jeweils N wichtigsten Merkmale denkbar.

Falls eine Auswahl von mehreren Merkmalen vorgenommen werden soll, so ist die bisher beschriebene Vorgehensweise, die auf den Ergebnissen einer berechneten SVM beruhen, möglicherweise nicht adäquat. Eine Verbesserung des Ergebnisses kann u.a. durch den Einsatz des Algorithmus des Recursive Feature Elimination (RFE) erreicht werden (*Guyon et al.* (2002)). Danach wird nach jedem Trainieren des Modells durch die Einträge des Normalenvektors ein Merkmal bestimmt, das im weiteren Verlauf des Algorithmus aufgrund seiner geringen Relevanz ausgeschlossen wird. Dies wird solange durchgeführt bis eine Reihenfolge der Wichtigkeit der Variablen bestimmt ist. Neben der in *Guyon et al.* (2002) vorgeschlagenen Verwendung des Normalenvektors kann dies auf die nicht lineare Trennung ausgeweitet werden. Dazu wird nach jedem Training beispielsweise mittels des Gradientenverfahrens die Reihenfolge der Wichtigkeit der Merkmale bestimmt. Dies soll im Folgenden mit NLRFE (nicht lineare RFE) bezeichnet werden.
Derartige Vorgehensweisen bilden insbesondere für große Datensätze, die durch eine große Anzahl von Merkmalen beschrieben sind, eine sehr aufwändige Variante, bei der eine nicht zu unterschreitende Trefferquote vorgegeben werden muss, um letztendlich die Anzahl der zu verwendenden Merkmale zu bestimmen. Eine weitere Möglichkeit ist der in *Guyon et al.* (2002) vorgeschlagene Schrankenwert für ein zu berechnendes Ranking-Kriterium. Dazu werden alle Merkmale verwendet, deren Score höher als ein festgelegter Schrankenwert ist. Dies findet in der Praxis neben der Vorgabe einer bestimmten Anzahl an Merkmalen aufgrund der Zeitersparnis wohl am häufigsten Verwendung. Allerdings bildet die a priori Festlegung des Schrankenwertes ein analoges Problem wie bei der Parameterwahl. Daher soll im Folgenden kurz eine Möglichkeit vorgestellt werden, wie dieser Wert festgelegt werden kann. Dazu kann eine Art Ellbogenkriterium eingesetzt werden, was mit der Vorgehensweise bei der Bestimmung der zu bildenden Gruppen bei der Clusteranalyse (vgl. *Backhaus et al.* (2003)) vergleichbar ist. Die Merkmale werden zunächst entsprechend ihrer Score-Werte r_i geordnet, und diejenigen Merkmale verwendet, die einen Score oberhalb einer Grenze G besitzen, nach der die Score-Werte drastisch fallen (Ellbogen). Eine mögliche Verteilung von Werten ist in Abbildung 3.14 gegeben. Hier stehen acht Merkmale zur Verfügung, von denen nach dem Ellbogenkriterium aber nur zwei verwendet werden würden[10]. Diese Vorgehensweise liegt darin begründet, dass zu erwarten ist, dass die Trefferquote für gering gewichtete Merkmale nur geringfügig bei Entfernen dieser Merkmale fällt, jedoch sensibel auf die Löschung wichtiger Variablen reagiert. Liegen innerhalb

[10]Die Datengrundlage bildet hier die „Pima Indians Diabetes Database“ Daten. Diese umfassen 798 Beobachtungen, die positiv oder negativ auf Diabetes getestet wurden.
(Quelle: http://www.ics.uci.edu/∼mlearn/databases/pima-indians-diabetes/, Zugriff: 12.8.2004)

einer derartigen Darstellung mehrere Ellbogen vor, so sollte zusätzlich mittels der Ermittlung der resultierenden Trefferquote zwischen den Alternativen entschieden werden.

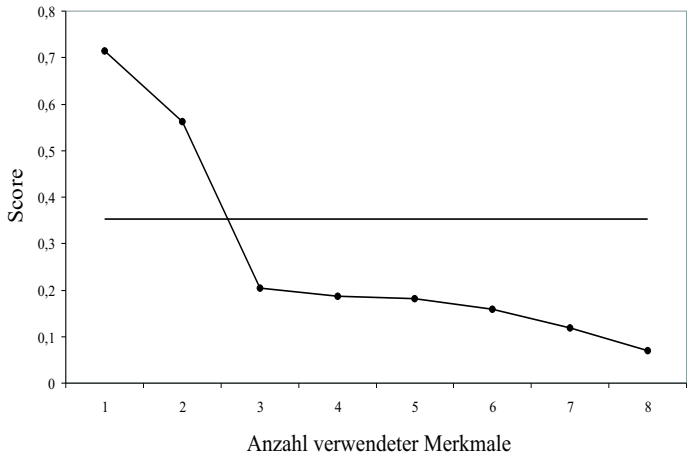

Abbildung 3.14: Bestimmung der Anzahl der zu extrahierenden Merkmale mit Hilfe des Ellbogenkriteriums

Ergänzend zum Ellbogenkriterium kann der zu erwartende Score für ein Merkmal bei gleichverteilten Relevanzen der Merkmale herangezogen werden. Wenn die Länge des Vektors $\boldsymbol{r}$ auf Eins normiert wird und man davon ausgeht, dass alle Merkmale den gleichen Einfluss auf die Trennung haben, so ergibt sich der dann zu erwartende Score r_i für alle Merkmale i aus:

$$\begin{aligned} & 1 = \sqrt{\textstyle\sum_{i=1}^{n} r_i^2} \\ \Leftrightarrow \quad & 1 = \sqrt{n r_i^2} \quad \forall\ i \\ \Leftrightarrow \quad & r_i = \tfrac{1}{\sqrt{n}} \quad \forall\ i. \end{aligned}$$

Eine mögliche Regel wäre: Wähle diejenigen Merkmale, die einen Score $r_i > \frac{1}{\sqrt{n}}$ aufweisen und bei nach absteigenden Score-Werten geordneter Darstellung „vor“ einem möglichen Ellbogen liegen. In Abbildung 3.14 ist der entsprechende zu erwartende Score durch die durchgezogene konstante Linie gekennzeichnet. Bei der nichtlinearen Trennung kann neben dem Ellbogenkriterium alternativ zum zu erwartenden Score der Mittelwert der erreichten Score-Werte herangezogen werden. Eine weitere Möglichkeit, die zum definitiven Ausschluss bestimmter Merkmale führt, wird von *Oukhellou et al.* (1998) gegeben. Die Autoren schlagen vor, die Datengrundlage um eine weitere, zufällig verteilte Variable zu ergänzen. Alle Merkmale, deren Score-Werte kleiner sind als der der zufälligen Variable, sollten von der Untersuchung ausgeschlossen werden. Somit kann die Anzahl zumindest geringfügig reduziert werden.

In den bisherigen Ausführungen zum Ranking und zur Auswahl von Merkmalen wurde lediglich auf einzelne Merkmale zurückgegriffen. Es kann hingegen

vorkommen, dass Merkmale separat betrachtet keine ausgeprägte Diskriminierungseigenschaft besitzen, aber in Kombination mit anderen Merkmalen relevant sind. Daher ist der Einsatz von Verfahren zur Auswahl von Teilmengen der ursprünglichen Merkmalsmenge sinnvoll, welche eine vorher festgelegte Anzahl n' an Merkmalen aus den ursprünglich n zur Verfügung stehenden Variablen auswählt. Ein Algorithmus dazu wird in *Fröhlich, Zell* (2004) unter dem Namen Incremental Regularized Risk Minimization (IRRM) vorgestellt. Aufbauend auf einer initialen Rangfolge aller Merkmale wird die Zusammenstellung der Menge der auszuwählenden Merkmale durch sukzessives Hinzufügen einzelner Merkmale so lange verändert, bis das bestmögliche Ergebnis erreicht worden ist. Diese Vorgehensweise wird in Abbildung 3.15 verdeutlicht[11].

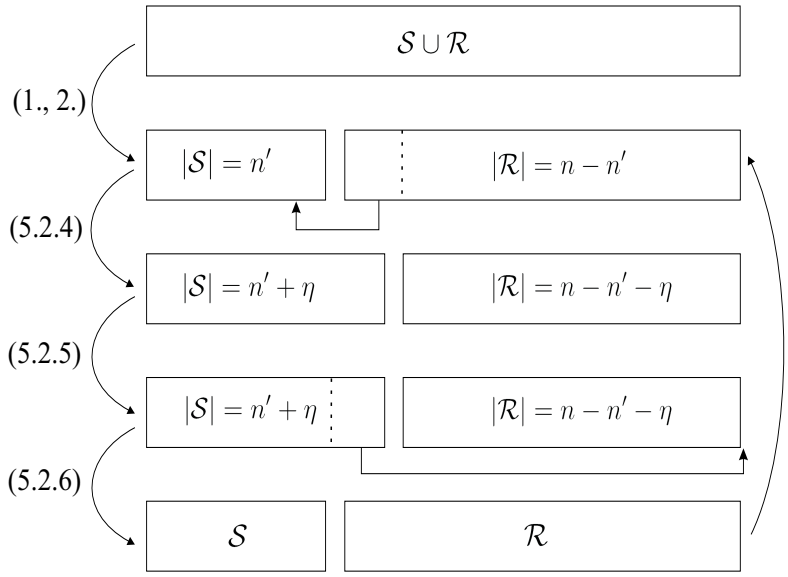

Abbildung 3.15: Idee des IRRM-Algorithmus in Anlehnung an *Fröhlich, Zell* (2004)

Dazu werden aus einem vorgegebenen Ranking zunächst die wichtigsten n' der insgesamt n Merkmale zu einer Menge $\mathcal{S}$ zusammengefasst. Die Merkmale seien in der Grafik nach Wichtigkeit absteigend geordnet. Die Menge der verbleibenden Merkmale werden mit $\mathcal{R}$ bezeichnet. Dann werden sukzessive η Merkmale[12] zwischen $\mathcal{R}$ und $\mathcal{S}$ ausgetauscht und auf deren gemeinsame Diskriminierungseigenschaft in $\mathcal{S}$ überprüft. Nach erneutem Sortieren von $\mathcal{S}$ werden die schlechtesten η Merkmale wieder der Menge $\mathcal{R}$ angehängt. Dies erfolgt solange, bis eine maximale Trefferquote auf Basis der Merkmale aus $\mathcal{S}$ erreicht ist. Um die Stabilität des Ergebnisses zu sichern, wird die Konvergenz durch wiederholte Iterationen gewährleistet.
Dieser Algorithmus wird nun dahingehend erweitert, dass die in dem Algorithmus nötigen Sortierungen der Merkmale durch das Gradientenverfahren von *Hermes,*

[11] Die hinzugefügten Nummern geben die Schritte in dem auf Seite 78 dargestellten Algorithmus wieder.

[12] Die Anzahl der zu überprüfenden Merkmale kann im einfachsten Fall auf $\eta = 1$ gesetzt werden.

Buhmann (2000) durchgeführt werden. Bei IRRM geschieht dies mit Hilfe des Wertes, der die Abweichung in der bei SVM zu maximierenden Spanne bei Entfernen eines Merkmals angibt. Ist diese sehr groß, so gilt das Merkmal für die vorliegende Trennaufgabe als wichtig. Es wird dabei davon ausgegangen, dass sich die Menge der eingehenden Support Vektoren bei Entfernen eines weniger wichtigen Merkmals nicht ändert. Ein einfaches Beispiel zeigt jedoch, dass dies im Vergleich zu wichtigen Merkmalen nicht zwingend gegeben ist. In Abbildung 3.16 wird die Trennung der durch zwei Merkmale beschriebenen fiktiven Daten derart durchgeführt, dass die gezeigte Trennebene mit acht Support Vektoren resultiert. Dem Merkmal 2 würde intuitiv eine höhere Bedeutung zugesprochen werden als dem Merkmal 1.

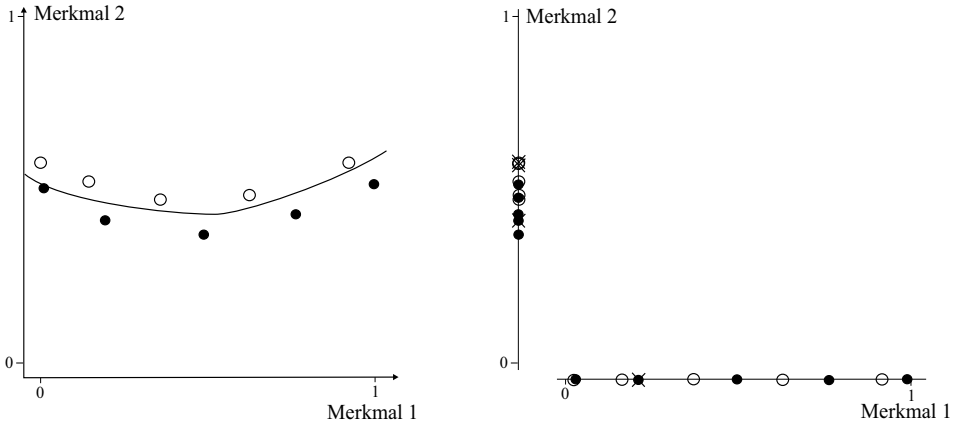

Abbildung 3.16: Veränderung der Menge der Support Vektoren bei Löschen jeweils eines Merkmals

Wird nun erste bzw. die zweite Dimension der Daten entfernt, also jeweils ein Merkmal gelöscht, so ergeben sich die rechts enthaltenen Grafiken. Durch die Kreuze werden diejenigen Vektoren markiert, bei denen sich an der Zugehörigkeit zur Menge der Support Vektoren durch das Löschen des jeweiligen Merkmals etwas geändert hat, sei es durch neues Hinzufügen oder durch Herausfallen aus der Menge. Bei Löschen des weniger wichtigen Merkmals (Merkmal 1) ergeben sich drei Kreuze, wohingegen das Löschen von Merkmal 2 zu lediglich einem Kreuz führt. In diesem speziellen Fall würde sich die Menge der Support Vektoren bei Löschen des weniger wichtigen Merkmals mehr ändern. Aus diesem Grund und um speziell nicht lineare Trennung intensiver zu betrachten, wird das Gradientenverfahrens innerhalb des Algorithmus zur Auswahl der Merkmale eingesetzt. Da der Algorithmus weiterhin auf die Multiklassifikation angewendet werden soll, wird er dahingehend abgeändert, dass zur Überprüfung der Prognosegüte statt des regulierten Risikos die Trefferquote auf vorher definierten Testdaten verwendet wird. Der durch diese Veränderungen geprägte Algorithmus sei im Folgenden durch NLIRM (Non Linear Incremental Risk Minimization) bezeichnet. In Tabelle 3.3 wird der modifizierte Algorithmus zur Bestimmung einer Menge $\mathcal{S}$ an n' Merkmalen angegeben[13].

[13]Die Umfang n' der zu bestimmenden Menge $\mathcal{S}$ muss a priori vorgegeben werden.

Schritt	Aktion
1.	Führe eine Merkmalsgewichtung auf Basis des Gradientenverfahrens der vorliegenden Merkmale durch
2.	Bei n' auszuwählenden Merkmalen konstruiere die Menge $\mathcal{S}$ mit $\|\mathcal{S}\| = n'$ und die Menge $\mathcal{R}$ mit $\|\mathcal{R}\| = n - n'$
3.	Berechne die Trefferquote TQ auf Basis von $\mathcal{S}$
4.	Setze t:=0
5.	Solange $\mathcal{S}$ sich ändert oder $t \leq 1$ gilt, führe aus:
5.1	$\mathcal{S}_{alt} := \mathcal{S}$
5.2	Bis die Trefferquote konvergiert, führe aus:
5.2.1	$TQ_{alt} := TQ$
5.2.2	Bestimme die Trefferquote TQ für die Merkmale aus $\mathcal{S}$
5.2.3	Falls $TQ \leq TQ_{alt}$, dann $\mathcal{S} := \mathcal{S}_{alt}$
5.2.4	Füge die ersten (wichtigsten) η Merkmale aus $\mathcal{R}$ an $\mathcal{S}$ an.
5.2.5	Bestimme die η unwichtigsten Merkmale aus $\mathcal{S}$ mittels Gradientenverfahren
5.2.6	Füge diese als letzte Merkmale in $\mathcal{R}$ an
5.3	Führe eine Merkmalsgewichtung von $\mathcal{R}$ auf Basis von NLRFE durch
5.4	Setze $t := t + 1$
6.	Gib gefundene Lösung $\mathcal{S}$ zurück

Tabelle 3.3: NLIRM-Algorithmus auf Basis des IRRM-Algorithmus von *Fröhlich, Zell* (2004)

An jeder Stelle, an der eine Trefferquote berechnet wird, muss zunächst die entsprechende Optimierung mit SVM vorgenommen werden. Die anfangs erforderliche Gewichtung der Merkmale erfolgt hierbei bereits auf Basis des Gradientenverfahrens. Die entscheidenden Schritte 1., 2., 5.2.4, 5.2.5 und 5.2.6 wurden bereits in Abbildung 3.15 verdeutlicht. Die erneute Sortierung der in $\mathcal{R}$ enthaltenen Merkmale in Schritt 5.3 wird wiederum unter Einsatz des Gradientenverfahrens durch NLRFE (vgl. Ausführungen auf Seite 74) durchgeführt. Da hierbei auch auf die Multiklassifikation zurückgegriffen wird, werden die Gewichtungen der Merkmale zunächst für jede einzelne SVM ermittelt und dann ein gemittelter Wert herangezogen. Die Ausgabe des Algorithmus bildet die letztlich optimal zusammengestellte Menge $\mathcal{S}$.
Die Bestimmung der zu verwendenden Merkmale n' hängt allerdings auch hier immer vom zugrunde liegenden Klassifikationsproblem und der Anzahl der insgesamt betrachteten Merkmale ab. Daher sollten die hier vorgestellten Ansätze lediglich als Richtlinie verstanden werden. Dies gilt insbesondere, wenn die maximale Merkmalsanzahl sehr klein ist.
Weiterhin sollten die durch die beschriebenen Verfahren ermittelten Einflüsse der Merkmale auf die Trennung der Ebene mit der in Abschnitt 3.2.1 vorgestellten Möglichkeit zur a priori-Gewichtung der Merkmale verglichen werden, um eine vorschnelle, evtl. falsche Reduzierung von Gewichten zu verhindern.

3.6 Multilabel-Klassifikation

Im Folgenden soll auf die Möglichkeit zur Zuweisung von Mengen von Klassen, die so genannte Multilabel-Klassifikation, eingegangen werden. Können Beobachtungen nicht eindeutig einer Klasse zugewiesen werden, so wird diesem Fall bei der Betrachtung von mehreren Klassen z.B. mit Fuzzy-SVM Rechnung getragen. Die Ursache für die Nichteindeutigkeit kann zum einen darin liegen, dass die betreffenden Beobachtungen einer neuen, bisher nicht betrachteten Klasse angehören und somit die vorhandenen Klassen das Klassifikationsproblem nicht ausreichend beschreiben. Ein weiterer möglicher Grund ist die gleichzeitige Zugehörigkeit zu mehreren Klassen. In diesem Fall ist eine eindeutige Zuordnung, wie sie mittels Fuzzy-SVM erreicht wird, nicht wünschenswert. Vielmehr sollte jeder Beobachtung eine Menge von Klassen zugeordnet werden, was im Folgenden thematisiert wird.
Diese Multilabel-Klassifikation kann etwa in der Textklassifikation auftreten, bei der einzelne Texte sowohl z.B. zum Bereich „Politik“, als auch zum Bereich „International“ gehören können. Diese Zuweisung von Mengen von Klassen zu den Beobachtungen ist immer dann nützlich, wenn die a priori definierten Klassen nicht eindeutig voneinander abzugrenzen sind, sondern sich in manchen Bereichen überlappen. Im Marketing kann diese Art der Klassifikation ebenfalls auftreten. So können im Bereich der Recommender-Systeme Kunden auf Basis ihres bisherigen Surf- und Kaufverhaltens Empfehlungen aus unterschiedlichen Kategorien erhalten, die sich nicht widersprechen, aber mit der herkömmlichen Klassifikation nicht zu vereinbaren sind, bzw. getrennte Analysen erfordern. Multilabel-Klassifikation erlaubt die Zuweisung von mehreren Empfehlungen zu einzelnen Kunden, sodass der Kunde z.B. beim Besuch eines Internetbuchhandels auf spezielle Produkte bzw. Bücher aus den Bereichen Krimi, Ratgeber und Reise hingewiesen werden kann, was mit in Abschnitt 2.2 beschriebener Multiklassifikation nicht möglich wäre. In diesem Fall wäre beispielsweise lediglich die Zuweisung zum Bereich „Reise“ möglich.

Multilabel-Klassifikationen können mittels SVM auf verschiedene Weise umgesetzt werden (*Joachims* (2002)). Eine Möglichkeit besteht darin, die auch den Trainingsdaten zugewiesenen Mengen von Klassen direkt mit in das Optimierungsproblem einfließen zu lassen (*Elisseeff, Weston* (2002)), was an dieser Stelle nicht weiter ausgeführt wird.
Eine andere Möglichkeit, die im Folgenden verwendet wird, besteht darin, das Problem auf das Lösen mehrerer binärer Probleme zurückzuführen und die Ausgaben so zu nutzen, dass eine Zuweisung einer Menge von Klassen möglich wird (*Boutell et al.* (2003)). Formal liegt die folgende Ausgangssituation vor: gegeben sind Eingabedaten $\boldsymbol{x}_i \in \mathbb{R}^n$ mit Klassenzugehörigkeiten $\boldsymbol{y}_i \in (0,1)^K$, d.h. jedem Eingabevektor ist eine Menge von Klassen zugeordnet, denen er angehört. Diese Menge Y_i wird repräsentiert durch einen binären Vektor $\boldsymbol{y}_i$, bei dem der Eintrag 1 angibt, dass die betreffende Klasse zu der Menge gehört und 0, dass der Vektor der betreffenden Klasse nicht zugeordnet ist. Beim Training einer SVM wird auf das OAA-Verfahren (vgl. Abschnitt 2.2.1) zurückgegriffen, um die Abspaltung jeweils einer Klasse von den restlichen Beobachtungen auszunutzen. Dies führt

dazu, dass eine Beobachtung, die k' Klassenlabels besitzt, auch k'-fach innerhalb des Optimierungsprozesses auftaucht. *Boutell et al.* (2003) weisen darauf hin, dass die Daten bei der Optimierung lediglich in der jeweils vom Rest zu trennenden Klasse auftauchen, nicht aber im restlichen Teil als negative Beobachtung. Somit wird vermieden, dass es zu unnötigen Komplikationen kommt, da eine Beobachtung gleichzeitig beiden zu trennenden Klassen angehört und damit nur schwer die optimale Hyperebene gefunden werden kann. Weiterhin wird bei diesem Vorgehen vermieden, dass es zu sehr dünn besetzten Klassen kommt. Dies tritt insbesondere dann auf, wenn jede Zuweisung zu mindestens zwei Klassen eine weitere Klasse bilden würde. D.h. die Zuordnung zu Klasse 2 und 3 bei vier möglichen Klassen würde eine neue, fünfte Klasse bilden, was zu einer stark erhöhten Anzahl an Klassen führen kann.
Es gibt mehrere Möglichkeiten, neben der Zuweisung von genau einer Klasse auch die Zuweisung von Mengen von Klassen bei Testdaten zu ermöglichen (vgl. *Boutell et al.* (2003)). Hier soll die Zuweisung auf Basis der Entscheidungswerte vorgenommen werden. Liegt eine Beobachtung sehr nah an der berechneten Hyperebene, so erhält sie einen geringen Wert der entsprechenden Entscheidungsfunktion. Demnach wird eine Beobachtung denjenigen Klassen k zugewiesen (mit $k = 1, \ldots, K$), für die $F^{[k]}(\boldsymbol{x})$ groß ist. Im Folgenden wird eine Zuweisung dann vorgenommen, wenn der jeweilige Entscheidungswert positiv ist (vgl. P-Kriterium in *Boutell et al.*(2003)). Dieses Vorgehen ermöglicht eine Zuweisung von einer oder mehreren Klassen zu einer Beobachtung. Dies wird ebenfalls durch das in *Boutell et al.* (2003) beschriebene C-Kriterium erreicht, bei dem die Entscheidungswerte der Klassen, zu denen eine Beobachtung zugewiesen wird, nah beieinander liegen müssen. Diese Zuweisung der Testdaten kann auch auf Basis der herkömmlichen SVM erfolgen, ohne dass zuvor ebenfalls eine Optimierung auf Basis der Multilabel-Klassifikation erfolgte und bietet somit eine weitere Möglichkeit, die Daten bzw. Zuweisungen sinnvoll zu interpretieren.
Bei bereits vorliegender mehrfacher Klassenzuweisung muss die Berechnung der Trefferquote in Abhängigkeit aller zugewiesenen Klassen erfolgen. Dazu kann wiederum die Hamming-Distanz $H(\cdot, \cdot)$ zum Einsatz kommen, bei der die Abweichung einzelner Einträge zweier Vektoren $\boldsymbol{y}$ und $\boldsymbol{y}' \in \mathbb{R}^K$ untersucht wird (in Anlehnung an *Joachims* (2002)):

$$H(\boldsymbol{y}, \boldsymbol{y}') = \sum_{k=1}^{K} (|y_k - y'_k|.$$

Damit wird die Anzahl der Abweichungen innerhalb der beiden Vektoren aufaddiert. Somit kann berücksichtigt werden, dass eine Menge von zugewiesenen Klassenlabels der Menge der wahren Klassenlabels ähnelt, aber dennoch nicht identisch ist, und eine geringe aber dennoch von 0 verschiedene Hamming-Distanz aufweist.
Eine weitere, etwas differenziertere Möglichkeit zur Berechnung der Prognosegüte schlagen *Boutell et al.* (2003) vor. Die Autoren geben folgende Formel für die Genauigkeit $Prec$ an:

$$Prec = \frac{1}{l'} \sum_{i=1}^{l'} score(\mathcal{P}_i),$$

wobei l' die Anzahl an Beobachtungen des zu überprüfenden Datensatzes und $\mathcal{P}_i$ die Menge der vorhergesagten Klassenlabels angibt. $score(\cdot)$ ist eine reellwertige Funktion, die sowohl die realen (Menge $\mathcal{Y}_i$) als auch die vorhergesagten (Menge $\mathcal{P}_i$) Klassenlabels berücksichtigt:

$$score(\mathcal{P}_i) = \left(\frac{|\mathcal{Y}_i \cap \mathcal{P}_i|}{|\mathcal{Y}_i \cup \mathcal{P}_i|} \right).$$

Somit wirkt sich eine Zuweisung vieler nicht richtiger Labels genauso nachteilig auf die Genauigkeit der Klassifikation aus wie das Nichterkennen von wahren Klassenlabels. Würde der Vektor $\boldsymbol{y}$ der wahren Klassenzugehörigkeiten einer Beobachtung folgende Form haben: $\boldsymbol{y} = (1, 0, 0, 1, 0)$, die Beobachtung i also zwei der möglichen fünf Klassen angehören, so ergibt sich $\mathcal{Y}_i = \{1, 4\}$. Wird durch $\boldsymbol{y}' = (1, 0, 0, 0, 1)$ der vorhergesagte Klassenvektor angegeben, so ist $\mathcal{P}_i = \{1, 5\}$. Damit resultiert für diese Beobachtung ein Wert von

$$score(\mathcal{P}_i) = \left(\frac{|\{1, 4\} \cap \{1, 5\}|}{|\{1, 4\} \cup \{1, 5\}|} \right) = \frac{|\{1\}|}{|\{1, 4, 5\}|} = \frac{1}{3}.$$

Bei einer perfekten Erkennung der vorliegenden Klassenzugehörigkeiten ergibt sich ein Wert von 1.
Allein aufgrund des Wertes für $score(\cdot)$ kann noch nicht auf die Zusammensetzung der beiden Mengen innerhalb der Multiklassenzuweisung geschlossen werden. So kann ein Wert von $\frac{1}{3}$ durch die folgenden Szenarien entstanden sein: Eine Klasse ist richtig erkannt worden und

- zwei weitere Prognosen waren falsch
- zwei weitere wahre Klassen sind nicht erkannt worden
- eine weitere Klasse ist nicht erkannt worden und eine weitere Prognose war falsch

Je besser allerdings die wahren Werte bzw. Klassen erkannt worden sind, desto höher ist der Wert für *score* (vgl. *Boutell et al.* (2003)) und damit auch die Genauigkeit *Prec*.
Die Multilabel-Klassifikation ist mit der alternativen Methode der Diskriminanzanalyse nur bedingt durchführbar. Da eine Rückführung auf die OAA-Trennung dem Wesen der Diskriminanzanalyse widerspricht, muss eine traditionelle Mehr-Gruppen-Analyse durchgeführt werden. Hierbei sind eher schlechte Klassifikationsresultate zu erwarten, da die Beobachtungen mit mehreren Gruppenzugehörigkeiten auch mehrfach in die Analyse eingehen. Die Klassifikation neuer Objekte könnte mit Hilfe des Distanzkonzeptes vorgenommen werden (vgl. *Backhaus et al.* (2003)). Um die Zuweisung zu mehreren Klassen zu ermöglichen, ist die Einführung eines Grenzwertes denkbar, den die Distanz für eine Zuweisung zu der betreffenden Gruppe nicht unterschreiten darf.

Der Vorteil der Multilabel-Klassifikation liegt in der gleichzeitigen Betrachtung vieler möglicher Klassen, was in Kapitel 4.4 bei der Analyse von Paneldaten umgesetzt wird.

3.7 Interpretation der Entscheidungswerte

Ein Schwachpunkt der SVM liegt in der Schwierigkeit der inhaltlichen Interpretation der Ausgabe. Im Rahmen der Kundenklassifikation interessiert bei der Zuordnung von Kunden in unterschiedliche Klassen beispielsweise nicht nur, ob ein Kunde relevant ist, also in die Klasse der wertvollen Kunden fällt, sondern auch dessen relative Wertigkeit, um etwa eine Differenzierung entsprechender Kundenbetreuungsmaßnahmen innerhalb der Klasse der wertvollen Kunden zu ermöglichen. Die vorgestellte Methodik erlaubt bisher nur eine binäre Zuordnung zu den a priori definierten Klassen. Einem Testdatenvektor $\boldsymbol{x}$ wird dazu mittels der Entscheidungsfunktion der Wert 1 oder -1 zugeordnet und somit seine Klassenzugehörigkeit bestimmt. Im Folgenden wird erörtert, auf welche Art und Weise zusätzlich die Stärke dieser Zuweisung ermittelt werden kann, um die Ergebnisse der SVM im Marketing sinnvoll nutzen zu können.

3.7.1 Theoretische Zusammenhänge

Als Entscheidungsbasis dient die Klassifikationsfunktion (2.15), mittels derer diese differenzierte Betrachtung durchgeführt werden kann. Statt die Zuweisung auf die rein binäre Zuordnung zu beschränken, wird das Ergebnis ohne die Signum-Funktion verwendet[14]:

$$F(\boldsymbol{x}) = \sum_i \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b.$$

In diesem Abschnitt wird zunächst lediglich die lineare Trennung betrachtet, sodass hier $K(\boldsymbol{x}_i, \boldsymbol{x}) = \boldsymbol{x}_i \boldsymbol{x}$ gilt. Die Interpretation der Ausgabe einer nicht linearen Trennung wird in Abschnitt 3.7.3 betrachtet.
Es werden also Ausgaben $F(\boldsymbol{x}) \in \mathbb{R}$ generiert, die Aussagen über die Stärke der jeweiligen Zuweisung machen können. *Zadrozny, Elkan* (2002) bezeichnen diesen Term auch als „confidence in prediction“. Bei der Multiklassifikation wird diese Idee ansatzweise bei der Winner-takes-all-Methode umgesetzt (vgl. Seite 24), bei der die Zugehörigkeit zu einer Klasse vom absoluten Wert der Entscheidungsfunktion abhängig ist.
Je größer der Wert $|F(\boldsymbol{x})|$ ist, desto weiter ist $\boldsymbol{x}$ von der Trennebene entfernt und desto sicherer ist die Klassifikation des Vektors zu bezeichnen. Umgekehrt gilt: je näher der Vektor an der Trennebene positioniert ist, desto kleiner ist $|F(\boldsymbol{x})|$ und desto unwahrscheinlich ist die Richtigklassifikation. Eine nur sehr kleine Variation

[14] Bei der Multiklassifikation werden die betreffenden Entscheidungswerte mit $F^k(\boldsymbol{x})$ bei OAA-Trennung bzw. mit $F^{k_1 k_2}(\boldsymbol{x})$ bei der OAO-Trennung bezeichnet.

in der Lage der Ebene führt in diesem Fall zu einem anderen Klassifikationsergebnis. Dies ist analog zum Vorgehen bei der Bestimmung der optimalen Hyperebene zu sehen. Dort wird die Spanne zwischen den beiden Klassen vor dem Hintergrund maximiert, einen möglichst großen Abstand zu den Eingabedaten zu erzeugen (vgl. Abschnitt 2.1.2).
Bei den folgenden Abschnitten wird die Klassifikation von Testdaten bzw. a priori unklassifizierter Beobachtungen untersucht. Dies bedeutet, dass bereits das Training einer SVM abgeschlossen wurde und neue Beobachtungen klassifiziert werden sollen. Neben der Klassenzugehörigkeit, die durch (2.15) ermittelt werden kann, werden Aussagen über die Treffsicherheit dieser Zugehörigkeit getroffen. Bei einer Zuordnung einer Beobachtung zu Klasse „+1“ bedeutet dies noch nicht, dass diese Zuordnung auch der wahren Klassenzugehörigkeit dieser Beobachtung entspricht. Daher wird im Folgenden die Interpretation der generierten Entscheidungswerte im Hinblick auf mögliche Marketinganwendungen verfeinert. Somit werden Aussagen über die Sicherheit der Richtigklassifikation getroffen.

3.7.2 Kategorisierung am Beispiel der Kundenklassifikation

Ziel des folgenden Abschnittes ist die Ausrichtung der SVM an den im Marketing intendierten Klassifikationszielen, um eine innerhalb dieses Kontextes angemessene Auswertung der Ergebnisse zu ermöglichen. In diesem Zusammenhang kann bei der Kundenklassifikation von einer sinnvollen Anwendung gesprochen werden, wenn Kunden entsprechend ihrer Wichtigkeit für das Unternehmen innerhalb der a priori gewählten Klassen der Intensität der einzusetzenden Marketingaktivitäten zugeordnet werden können. Für eine ökonomische Verteilung von Ressourcen ist es nicht nur von Interesse, zu wissen, ob ein Kunde zu den wichtigen Kunden innerhalb des Kundenstammes gehört, sondern auch wie wichtig er ist. Um diesem Anspruch gerecht zu werden, bieten SVM auf Basis der generierten Entscheidungswerte $F(\cdot)$ die Möglichkeit, die Ausgabe dem vorgegebenen Ziel entsprechend zu analysieren und die Güte der Klassifikation zu beurteilen. Es sei angemerkt, dass nicht der ökonomische Wert eines Kunden ermittelt wird, sondern bei einer Klassifikation diejenigen heraus gefiltert werden sollen, die potentiell wertvoll für das Unternehmen sein könnten. Eine aus den Ergebnissen einer SVM beispielhaft resultierende Situation wird in Abbildung 3.17 beschrieben. Dabei wird davon ausgegangen, dass die zu klassifizierenden Kunden in zwei Gruppen eingeteilt sind, in die für das Unternehmen interessanten Kunden (Klasse „+1“) und die eher uninteressanten Kunden (Klasse „-1“). Die relevante Hyperebene wird durch die durchgezogene Linie verdeutlicht. Die vertikale Dimension beinhaltet die Ausprägung der jeweiligen Entscheidungswerte, die horizontale Verteilung der Beobachtungen dient hier lediglich der übersichtlicheren Darstellung.

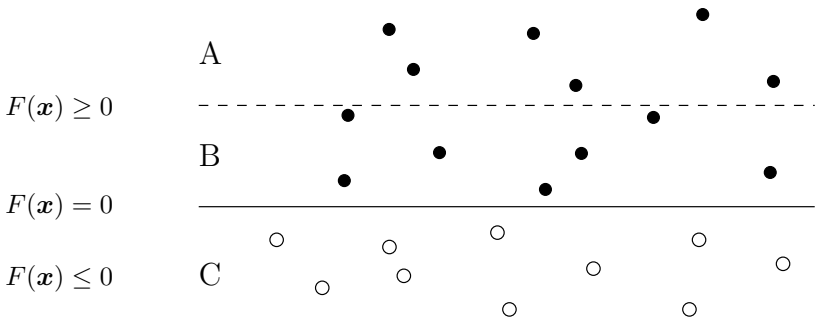

Abbildung 3.17: Mögliche Einteilung der zu klassifizierenden Testdaten in die drei Bereiche der A-, B- und C-Kunden

Die Entscheidungswerte können gut zur Bewertung der Klassifikationsgüte herangezogen werden, da sie die für die Kundenklassifikation relevanten Informationen enthalten. Da zur Entwicklung von Marketingstrategien jedoch nur eine diskrete Anzahl an Maßnahmen realisiert werden kann, muss eine kleine Anzahl an Bereichen geschaffen werden. Daher ist eine Einteilung des insgesamt aufgespannten Intervalls der Entscheidungswerte sinnvoll. Dies kann in Anlehnung an die Kundensegmentierung mittels ABC-Analyse (vgl. *Krafft, Albers* (2000)) durchgeführt werden. Dazu werden die interessanten Kunden hier zusätzlich in die beiden Bereiche A und B eingeteilt, die durch die gestrichelte Linie voneinander abgegrenzt werden. Die eher als uninteressant eingestuften Kunden bilden demnach die Klasse der C-Kunden. Dabei spielt etwa der Umsatz des jeweiligen Kunden im Gegensatz zur herkömmlichen Idee der ABC-Analyse hier keine Rolle. Die Zuordnung erfolgt lediglich auf Basis der aus SVM gewonnenen Entscheidungswerte[15]. Ein hoher Wert $|F(\boldsymbol{x})|$ führt zu einer sicheren Zuordnung zu der jeweiligen Klasse. Dementsprechend kann bei den Kunden, die einen Wert $F(\boldsymbol{x}) \gg 0$ zugewiesen bekommen, davon gesprochen werden, dass sie mit hoher Wahrscheinlichkeit zu den interessanten Kunden gehören und daher zum Bereich A zugeordnet werden. Ein „kleiner“ positiver Entscheidungswert führt zu einer Zuweisung zur Gruppe B, wohingegen ein negativer Wert $F(\boldsymbol{x})$ eine Klassifikation als C-Kunde nach sich zieht. Von kleinen Entscheidungswerten kann hier gesprochen werden, wenn diese absolut gesehen kleiner als eins sind.

Eine solche Erweiterung der Klassifikation eröffnet die Möglichkeit der gezielteren Ansprache der Kunden, liefert allerdings nur ein zusätzliches Segment im Vergleich zur herkömmlichen Klassifikation der Kunden in die a priori gegebenen zwei Klassen. Um die Einteilung des Kundenstammes weiter zu verfeinern, bietet sich die Orientierung an den im Finanzbereich verwendeten Ratingdefinitionen für Insurer Financial Strength-Ratings zur Bewertung von Versicherungen von *Standard & Poor's* (2004) an. Die weitere Differenzierung der beiden Klassen in jeweils vier Bereiche und deren eigentliche Bedeutung nach *Standard & Poor's* sowie

[15]Die Entscheidungswerte werden möglicherweise vom Umsatz beeinflusst, da dieser in die Menge der die Kunden beschreibenden Merkmale eingehen kann.

die Übertragung auf die Kundenklassifikation wird in Tabelle 3.4 vorgestellt. Die verwendete Einteilung in sichere bzw. finanzstarke Unternehmen (Kategorien AAA, AA, A und BBB) und anfällige, bzw. finanzschwache Unternehmen (Kategorien BB, B, CCC und CC) kann demnach auf die Biklassifikation übertragen werden und durch weitere zusätzliche Abstufungen innerhalb einer Kategorie, z.B. B+ oder C-, weiter verfeinert werden.
Bei einer derartigen Übertragung dieses Konzepts ergeben sich aufgrund der Entscheidungswerte für jede der beiden Kundenklassen „interessante, wichtige Kunden“ und „uninteressante, unwichtige Kunden“ vier Wertigkeitsbereiche. Es wird davon ausgegangen, dass die vier Bereiche pro Klasse beispielsweise mit der Intensität der Betreuung der betreffenden Kunden korrespondieren. Analog zum *Standard & Poor's* Rating umfasst der Bereich AAA die für das Unternehmen interessantesten Kunden, die damit zu den Topkunden zu zählen sind. Auf der anderen Seite setzt sich der Wertigkeitsbereich CC aus den für das klassifizierende Unternehmen am wenigsten interessanten Kunden zusammen. Die hierbei angenommene Analogie ist in der Interpretation der Entscheidungswerte als sichere bzw. nicht sichere Klassifikation der Kunden in die beiden Klassen begründet. Die Nähe eines einen Kunden beschreibenden Vektors zu der berechneten Ebene und die damit verbundene unsichere Klassifikation mit einer gering zu erwartenden Trefferquote findet sich in den Bereichen BBB und BB wieder. Eine Zuweisung der Kunden zu den unterschiedlichen Wertigkeitsbereichen kann eine differenzierte und ökonomische Aufteilung von Marketingressourcen ermöglichen.
In welchen Abständen die Einteilung vorgenommen wird, ist dem Anwender überlassen und hängt sehr stark vom zugrunde liegenden Datenmaterial ab. Neben dem Verfahren von *Standard & Poor's* kann dazu das vom BERI Institut entwickelte Ratingverfahren zur Bewertung von Länderrisiken herangezogen werden. Einem Land werden unterschiedliche Indizes zugeordnet, die zu einem Gesamtindex summiert werden (vgl. *Meyer* (1987)), der Werte zwischen 0 und 100 erreichen kann. Dieses Intervall wird wie oben auch in acht in sich homogene Bereiche geteilt, deren Intervallbreite allerdings voneinander abweichen. Eine daran angepasste Einteilung könnte die in Tabelle 3.5 vorgeschlagene sein. Hierbei wird durch

$$M = \min(|\max_{i=1,\ldots,l} F(\boldsymbol{x}_i)|, |\min_{i=1,\ldots,l} F(\boldsymbol{x}_i)|)$$

der kleinere der maximal realisierten Absolutwerte der Entscheidungsfunktion in beiden Klassen auf Basis der Trainingsdaten ($i = 1, \ldots, l$) bezeichnet. Eine derartige Differenzierung liegt darin begründet, dass in den klaren Abschnitten, also bei der Zuweisung von sehr wenigen oder sehr vielen Punkten, die Intervalle von größerer Breite sind als in den Bereichen, in denen keine klare Zuordnung vorgenommen werden kann (35-65 Punkte). Dies kann sehr gut auf die Situation bei SVM übertragen werden, bei der schmale Intervalle bei absolut gesehen sehr großen Entscheidungswerten ebenfalls nicht erforderlich erscheinen und eine Differenzierung in der Breite der Intervalle demnach begründet ist.

Abhängig vom Entscheidungswert kann eine Beobachtung einer der acht definierten Bereiche zugewiesen werden. Darauf basierend kann z.B. im Rahmen des Direktmar-

	Bedeutung	
	nach *Standard & Poor's*	bei der Kundenklassifikation
AAA	herausragende finanzielle Stabilität; höchstes zu vergebenes Rating	Topkunden Wahrscheinlichkeit der Zugehörigkeit zu Klasse „+1“ ist sehr hoch
AA	ausgezeichnete finanzielle Stabilität; nur geringe Unterschiede zu AAA	sehr gute Kunden Wahrscheinlichkeit der Zugehörigkeit zu Klasse „+1“ ist hoch
A	sehr gute finanzielle Stabilität; Wahrscheinlichkeit der negativen Beeinflussung der Stabilität ist etwas größer	gute Kunden Wahrscheinlichkeit der Zugehörigkeit zu Klasse „+1“ ist eher gering
BBB	gute finanzielle Stabilität; Wahrscheinlichkeit der negativen Beeinflussung der Stabilität ist größer	unsichere Kandidaten
BB	marginale finanzielle Stabilität; ungünstige Entwicklungen können zur Nichterfüllung der Verpflichtungen führen	unsichere Kandidaten
B	schwache finanzielle Stabilität; ungünstige Entwicklungen führen mit großer Wahrscheinlichkeit zur Nichterfüllung der Verpflichtungen	eher unwichtige Kunden Wahrscheinlichkeit der Zugehörigkeit zu Klasse „-1“ ist gering
CCC	sehr schwache finanzielle Stabilität; Erfüllung der Verpflichtungen nur unter günstigen Bedingungen möglich	unwichtige Kunden Wahrscheinlichkeit der Zugehörigkeit zu Klasse „-1“ ist hoch
CC	extrem schwache finanzielle Stabilität; Erfüllung der Verpflichtungen mit hoher Wahrscheinlichkeit nicht möglich	ignorierbare Kunden Wahrscheinlichkeit der Zugehörigkeit zu Klasse „-1“ ist sehr hoch

Tabelle 3.4: Einteilung der a priori definierten Kundenklassen in je vier Bereiche

Klasse „+1“

Bezeichnung	AAA	AA	A	BBB
BERI Punkte	65-100	60-65	55-60	50-55
SVM $F(\boldsymbol{x})$	$[0,3M;M]$	$[0,2M;0,3M]$	$[0,1M;0,2M]$	$[0;0,1M]$

Klasse „-1“

Bezeichnung	BB	B	CCC	CC
BERI Punkte	45-50	40-45	35-40	0-35
SVM $F(\boldsymbol{x})$	$[0;\text{-}0,1M]$	$[\text{-}0,1M;\text{-}0,2M]$	$[\text{-}0,2M;\text{-}0,3M]$	$[\text{-}0,3M;\text{-}M]$

Tabelle 3.5: Einteilung der Abstände der Entscheidungsbereiche in Anlehnung an das Prinzip des BERI-Indexes

ketings das einem Kunden zuteil werdende Kundenbetreuungsprogramm festgelegt werden. In Tabelle 3.5 werden die Bezeichnungen der Klassen AAA, . . . ,CC aus Tabelle 3.4 übernommen, wobei sich die inhaltliche Bedeutung der acht Bereiche bei *Standard & Poor's* und bei BERI nicht stark voneinander unterscheiden. Der Abschnitt AAA bzw. CC wird je nach Realisierung der Größe M variiert[16], um den gesamten Bereich der Entscheidungswerte abzudecken. Die Anpassung der Trennebene erfolgt bei SVM auf Basis der Gruppenränder (Support Vektoren) und ignoriert die jeweiligen Gruppenmittel. Dies bewirkt, dass die maximal erreichten Entscheidungswerte der beiden Klassen sich stark voneinander unterscheiden können. Eine Anpassung der Bereiche pro Klasse, sodass die Verhältnisse der Intervallbreiten zueinander gleich blieben, die absoluten Breiten der Intervalle jedoch in beiden Klassen unterschiedlich wären, würde dem Prinzip der SVM der Maximierung der Spanne zwischen den Gruppen widersprechen.
Eine weitere Annahme der oben einzuteilenden Bereiche ist die Repräsentativität der Daten. Die Beobachtungen sollten so gewählt werden, dass sie die Grundgesamtheit repräsentieren und mit der Zeit durch neue Daten erweitert und somit an eventuell neu auftretende Strukturen angepasst werden. Neue Beobachtungen, bei denen Entscheidungswerte größer als M realisiert werden, werden wie bei den Trainingsdaten ebenfalls den Bereichen AAA bzw. CC zugewiesen. Die jeweils äußeren Intervalle sind demnach eher als offene Intervalle $[0,3M;\infty]$, bzw. $[-\infty;-0,3M]$ zu verstehen. Je nach Trenngüte des Modells kann eine stärkere Differenzierung der Länge der Intervalle empfehlenswert sein. Liegen sehr viele der Beobachtungen nah an der Entscheidungsebene, so kann es vorteilhafter sein, wenn dieser Bereich gestaucht wird (Abschnitte BBB (bzw. BB) von $[0;0,1M]$ (bzw. $[0;-0,1M]$) auf $[0;0,075M]$ (bzw. $[0;-0,075M]$)) und die angrenzenden Bereiche entsprechend gestrafft werden[17].

[16] Falls beispielsweise $M = |\max_{i=1,\ldots,l}(F(\boldsymbol{x}_i))|$, so wird der Bereich CC vergrößert, sodass alle Vektoren in diesen Bereich fallen, für die $F(\boldsymbol{x}) \in [-0,3M;\min_{i=1,\ldots,l}(F(\boldsymbol{x}_i))]$ gilt.

[17] Hier müssten also die Bereiche B und A entsprechend vergrößert werden und die Straffung an die weiteren Bereiche weitergegeben werden, sodass die äußeren Abschnitte AAA und CC (um $0,025M$) größer werden.

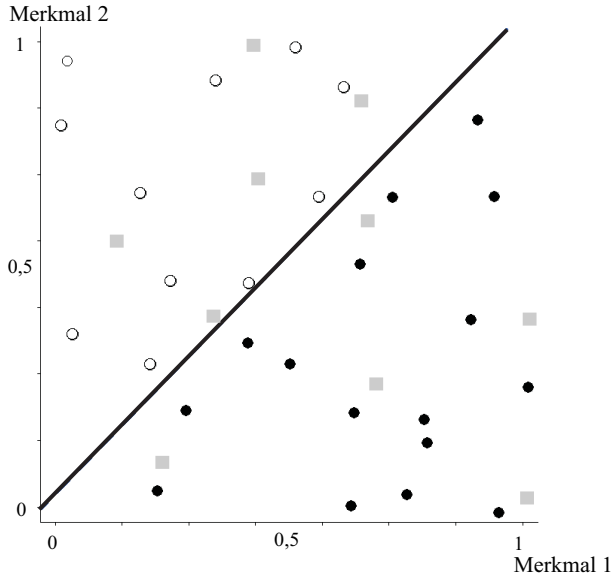

Abbildung 3.18: Trennung fiktiver Daten zur Veranschaulichung der Interpretation der Entscheidungswerte

Die obige Idee wird anhand eines fiktiven Beispiels verdeutlicht. Gegeben seien dazu 2-dimensionale Eingabedaten, die zwei Klassen zugeteilt sind. In Abbildung 3.18 wird die beispielhafte lineare Trennung der gegebenen Trainingsdaten verdeutlicht. Die schwarz markierten ausgefüllten (bzw. nicht ausgefüllten) Kreise repräsentieren die Trainingsdaten der Klasse „-1“ (bzw. Klasse „+1“). Die Testdaten, die a priori keiner Klasse zugeordnet wurden, sind durch graue Quadrate dargestellt. Auf Basis der berechneten Entscheidungsfunktion ergibt sich zunächst

$$\begin{aligned} M &= \min(|\max_{i=1,\ldots,26}(F(\boldsymbol{x}_i))|, |\min_{i=1,\ldots,26}(F(\boldsymbol{x}_i))|) \\ &= \min(|8,857|, |-11,125|) = 8,857. \end{aligned}$$

Daraus resultiert die in Abbildung 3.19 wiedergegebene Situation für die Trainings- und Testdaten. Die Beobachtungen wurden gemäß der Ausprägung ihres Entscheidungswertes angeordnet. Die horizontale Verteilung spielt neben der Übersichtlichkeit wiederum keine Rolle. Aufgrund der Verteilung dieser Werte bietet sich eine Einteilung in unterschiedliche Bereiche in Anlehnung an *Standard & Poor's* dergestalt an, dass die sich ergebenden acht Bereiche jeweils Intervalle der Länge $0,886$ (Bereiche AA, ...,CCC) und $6,3$ (Bereiche AAA und CC) abdecken. Diese Art der Darstellung trägt dem Zustand Rechnung, dass bei eher mittelmäßigen Trennungen von zwei Klassen die Datenpunkte häufig sehr nah an der Ebene positioniert sind und demnach eine feinere Einteilung gegenüber den weiter entfernteren Bereichen erfordern. Im obigen Beispiel werden die neuen Beobachtungen (Testdaten) jeweils einem der acht Bereiche zugeordnet und können somit eine Behandlung erfahren, die in der Zuordnung zu den jeweiligen Bereichen begründet ist.

Darüberhinaus bietet diese Form der Interpretation eine Möglichkeit zur vi-

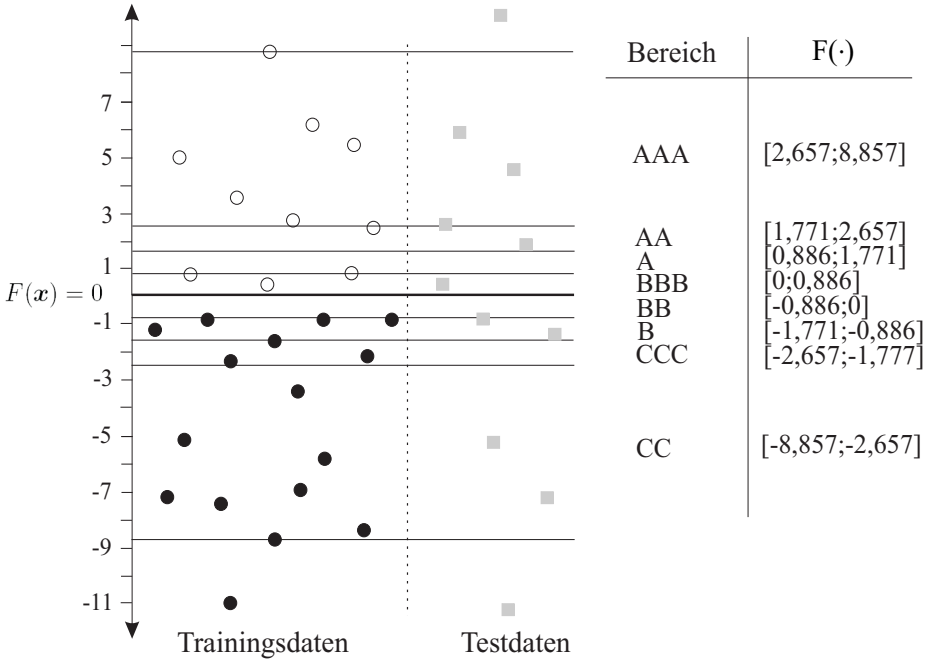

Abbildung 3.19: Einteilung in Bereiche in Anlehnung an *Standard & Poor's* und BERI

suellen Darstellung der Ergebnisse. Es wird eine Projektion in den ein- (bzw. zwei-)dimensionalen Raum vorgenommen, um auf diese Weise die Ergebnisse der Klassenzugehörigkeitsprognose zu veranschaulichen. Falls eine grafische Ausgabe aufgrund der Hochdimensionalität (für $n \geq 3$) der Daten nicht möglich ist, werden durch eine derartige Abbildung zumindest die Entscheidungswerte sichtbar gemacht, sodass anhand dieser Verteilung ebenfalls Aussagen zur Güte der Klassifikation gemacht werden können. Liegen beispielsweise eine Reihe falsch klassifizierter Beobachtungen vor, ist die Trennung dennoch als akzeptabel zu bezeichnen, wenn die entsprechenden Absolutwerte der Entscheidungsfunktion sehr gering und die Beobachtungen somit sehr nahe an der Ebene positioniert sind, was anhand derartiger Abbildungen schnell ersichtlich ist. Demnach bildet diese Visualisierung ebenfalls eine Hilfestellung bei der Bewertung der Klassifikationsgüte von SVM.

3.7.3 Interpretation bei nicht linearer Trennung

Die lineare Trennung von Eingabedaten beruht auf dem Prinzip der Risikominimierung, welches durch die Maximierung der Spanne bei der Berechnung der Trennebene verfolgt wird. Dabei erhalten die Vektoren $\boldsymbol{x}$, die auf den resultierenden Hilfsebenen liegen, einen Entscheidungswert von $F(\boldsymbol{x}) = 1$ (vgl. Abbildung 2.4). Je weiter die Vektoren von dieser Hilfsebene entfernt sind, desto größer wird der eingehende Wert der Entscheidungsfunktion. Die gleiche Situation tritt im Merkmalsraum bei der nicht linearen Trennung auf. Diese lineare Beziehung zwischen der Distanz und

dem Entscheidungswert ist bei der nicht linearen Trennung allerdings nicht im Eingaberaum gegeben. Dies kann anhand eines fiktiven Datensatzes erläutert werden, dessen zwei Klassen in Abbildung 3.20 sowohl linear[18] (links) mit $C = 10$ als auch nicht linear (rechts) durch eine Radialbasis-Funktion (mit $\gamma = 0,1$ und $C = 10$) getrennt werden.

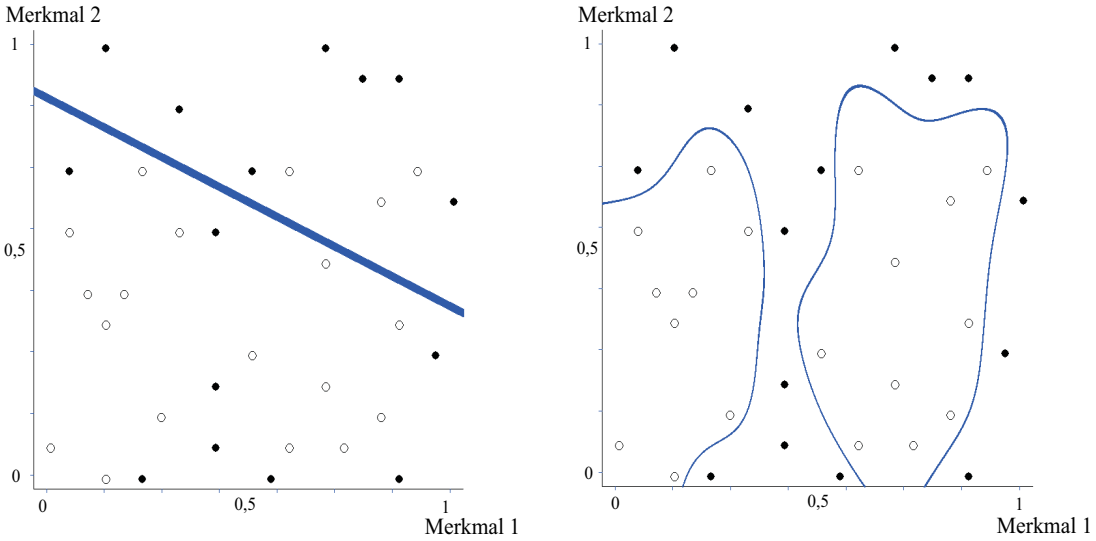

Abbildung 3.20: Trennung fiktiver Daten mittels linearer und nicht linearer Entscheidungsfunktion

Die dargestellten Vektoren können als Kunden eines Unternehmens aufgefasst werden, die durch zwei Dimensionen, etwa bisheriger Umsatz und Anzahl bisheriger Retouren, charakterisiert werden.[19] Die beiden Klassen können durch die „Wiederkäufer“ und “andere Kunden“ gebildet werden, die durch Kreise und Punkte symbolisiert werden. Werden nun die auf Basis der berechneten (Hyper-)Ebenen entstehenden Entscheidungswerte für die durch die Eingabedaten aufgespannte Fläche als dritte Dimension über diesem Raum aufgefasst, so ergibt sich für die lineare Trennung die Darstellung in Abbildung 3.21, bei der die Visualisierung der Entscheidungswerte umso heller wird, je höher der absolute Wert ausfällt. Die lineare Abhängigkeit zwischen dem Abstand eines Punktes zu der Trennebene und den auf Basis der Entscheidungsfunktion resultierenden Entscheidungswerten $F(\boldsymbol{x})$ im Falle der linearen Trennung ist gut zu erkennen. Abbildung 3.22 verdeutlicht, dass dies im nicht linearen Fall nicht gegeben ist. Würde diese lineare Abhängigkeit wie bei der linearen Trennung auch hier vorliegen, so müssten sich beispielsweise

[18] Der Vorstellung des folgenden Konzeptes wird ein zwei-dimensionaler Eingaberaum zugrunde gelegt, um die Daten grafisch darstellen zu können. Weiterhin ist die bei der linearen Trennung entstehende große Anzahl an Klassifikationsfehlern für die folgenden Betrachtung unerheblich, da sich die Ausführungen im Folgenden nur auf die Struktur der Entscheidungswerte beziehen.

[19] Hier sei angemerkt, dass die Daten in dieser Arbeit vor einer Auswertung mit SVM immer auf das Intervall $[0, 1]$ normiert werden und die Merkmalsausprägungen somit nicht direkt dem Inhalt der Merkmalsbeschreibung, wie „Anzahl an bisherigen Retouren“ entsprechen.

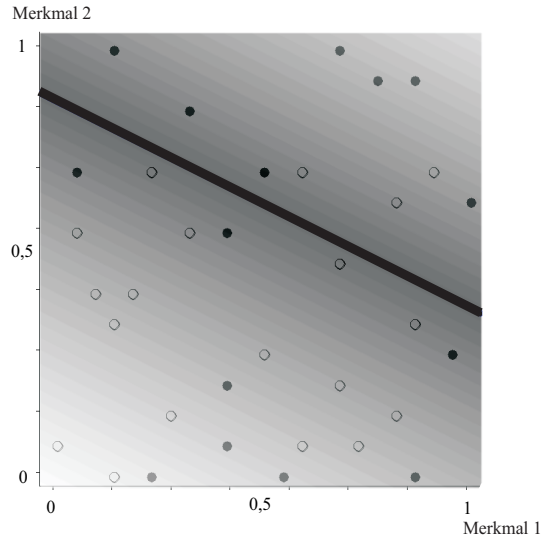

Abbildung 3.21: Visualisierung von F für den linearen Fall

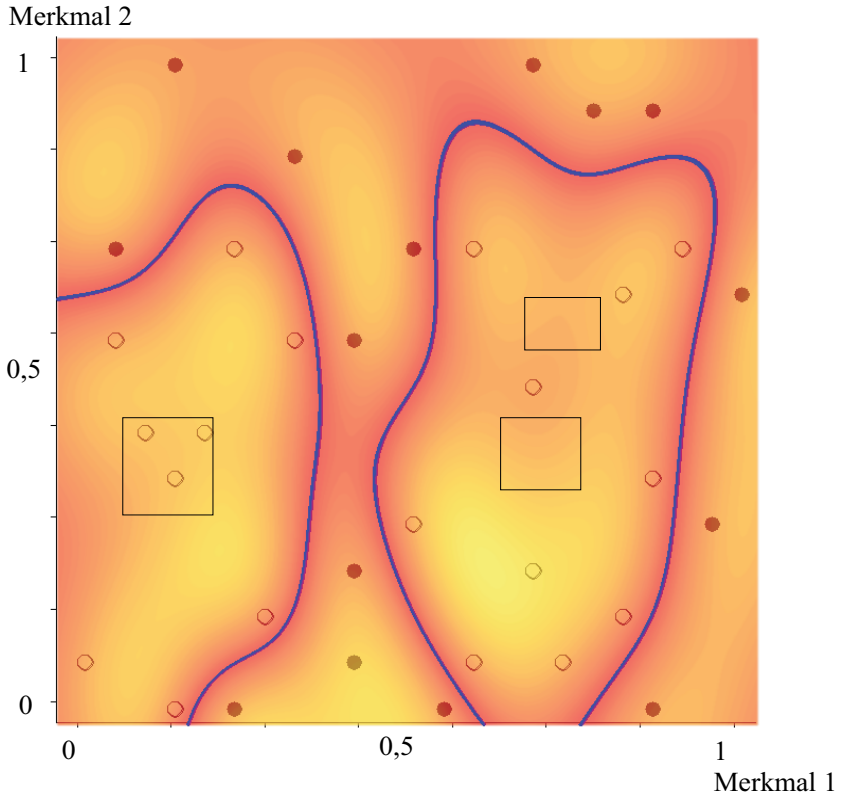

Abbildung 3.22: Visualisierung von F bei Trennung mittels einer Radialbasis-Funktion

innerhalb der eingezeichneten Rechtecke hellere Bereiche ergeben.
Bei der Klassifikation im Rahmen des Marketings interessieren vornehmlich die Eigenschaften der berechneten Trennfunktion im Eingaberaum, da eine Aussage über die Güte einer Klassenzuweisung eines noch nicht klassifizierten Vektors in Abhängigkeit seiner Lage im Eingaberaum getroffen werden soll. Bei einer nicht linearen Trennung lässt ein niedriger Wert $|F(\boldsymbol{x})|$ für einen Vektor $\boldsymbol{x}$ nur darauf schließen, dass sich das Bild des Vektors unter der Abbildung Φ nah an der berechneten Ebene befindet, nicht aber der eigentliche Eingabevektor $\boldsymbol{x}$ selbst. Es kann so etwa der Fall auftreten, dass ein hoher absoluter Wert für F nicht einem hohen Abstand zur Trennebene im Eingaberaum entspricht. Innerhalb der in Abbildung 3.22 durch Rechtecke markierten Bereiche befinden sich Abschnitte, in denen Vektoren existieren, die bei gleichzeitig geringeren Entscheidungswerten einen größeren Abstand zur Ebene aufweisen als näher positionierte Vektoren[20]. Auf Basis der im $\mathbb{R}^2$ visualisierten Trennebene würden Vektoren des betreffenden Bereichs als sicher klassifiziert gelten, da sie weit von der Entscheidungsebene im Eingaberaum entfernt liegen, die auf Basis von Trainingsdaten fest vorgegeben ist. Auf Basis der geringen Entscheidungswerte jedoch würde dieses Urteil anders ausfallen und die Vektoren im Vergleich zu Anderen als unsicher klassifiziert beurteilt werden.
Ziel ist daher die Einführung eines neuen Maßes für den Einsatz im Marketing, welches dazu dient, neben den durch SVM ermittelten Entscheidungswerten auch die Abstände der Eingabedaten zur berechneten, nicht linearen Hyperebene im Eingaberaum zu berücksichtigen. Es soll ein Pendant zu F im Eingaberaum geschaffen werden. Damit kann analog zu Abschnitt 3.7.2 ermittelt werden, welche Vektoren sicher oder unsicher klassifiziert worden sind. Somit soll eine entsprechende Zuordnung zu verschiedenen Wertigkeitsbereichen (AAA bis CC) auch bei der nicht linearen Trennung unter Berücksichtigung der Situation im Eingaberaum ermöglicht werden. Ziel der Ausführungen dieses Abschnitts ist die Generierung von Aussagen über die Sicherheit der Klassifikation von Vektoren vor dem Hintergrund der vorliegenden SVM sowie eine bessere Veranschaulichung der Resultate im Eingaberaum.

Der zweidimensionale Eingaberaum wird wie in Abbildung 3.22 mit der zusätzlichen Dimension $F'(\cdot) = |F(\cdot)|$ als ein dreidimensionaler Raum aufgefasst[21]. Dieses Vorgehen kann anschaulich als die Betrachtung einer topografischen Karte mit Höhenlinien aufgefasst werden. Die Höhenlinien werden durch verschiedene Werte von F' erzeugt. Grundlegende Idee des Ansatzes ist die Berücksichtigung der Ausprägungen von F' entlang der kürzesten Verbindungslinie des Vektors $\boldsymbol{x}$ mit der Hyperebene bei der Berechnung eines neuen Maßes $D(\boldsymbol{x})$. Abbildung 3.23 zeigt einen Querschnitt eines solchen dreidimensionalen Bildes. Verdeutlicht wird hierbei der Bereich zwischen dem Punkt $\boldsymbol{s}_0^{[\boldsymbol{x}]}$ auf der Trennebene und einem zu klassifizierenden Vektor $\boldsymbol{x}$. Dabei ist $\boldsymbol{s}_0^{[\boldsymbol{x}]}$ der Vektor, der die folgende Bedingung

[20] Hierbei handelt es sich nicht unbedingt um die eingezeichneten Trainingsvektoren, sondern um zu klassifizierende neue Beobachtungen mit entsprechenden Merkmalsausprägungen.

[21] Es werden lediglich die Absolutwerte betrachtet, da nur die Ausprägung, nicht aber die Klasse von Bedeutung ist.

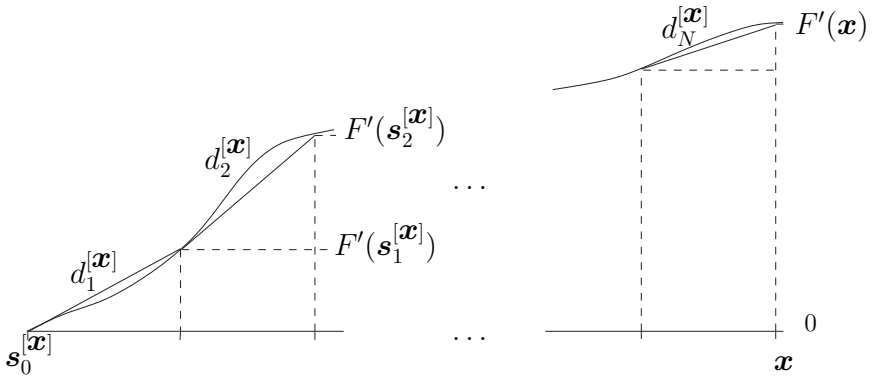

Abbildung 3.23: Querschnitt des zu untersuchenden Raumes bei Vorliegen zweidimensionaler Daten

erfüllt:

$$\boldsymbol{s}_0^{[\boldsymbol{x}]} = \arg \min_{\boldsymbol{s}:F'(\boldsymbol{s})=0} \{d(\boldsymbol{s};\boldsymbol{x})\} \tag{3.8}$$

Demnach ist $\boldsymbol{s}_0^{[\boldsymbol{x}]}$ der Vektor auf der Ebene, der den geringsten Abstand zu $\boldsymbol{x}$ aufweist. Nun wird die Länge der Kurve bestimmt, also die Länge der Verbindung zwischen $\boldsymbol{s}_0^{[\boldsymbol{x}]}$ und $F'(\boldsymbol{x})$, die durch die Ausprägungen der jeweiligen Entscheidungswerte definiert ist. Die eigentliche Länge der in Abbildung 3.23 angedeuteten, von F induzierten Kurve wird hierbei durch N lineare Teilstücke $\boldsymbol{d}_i^{[\boldsymbol{x}]}$ approximiert. Diese ergeben sich aus den Entscheidungswerten $F'(\boldsymbol{s}_i^{[\boldsymbol{x}]})$ und $F'(\boldsymbol{s}_{i-1}^{[\boldsymbol{x}]})$, wobei $N \in \mathbb{N}$ und $N \geq 1$. Das eigentliche Maß $D(\boldsymbol{x})$ zur Beurteilung der Lage eines Vektors $\boldsymbol{x}$ ergibt sich nun aus:

$$D(\boldsymbol{x}) = \sum_{i=1}^{N} d_i^{[\boldsymbol{x}]}.$$

Um einen zu großen Einfluss der eingehenden Distanzen auf D zu verhindern, und den Ausprägungen von F' größeres Gewicht zu verleihen, ist eine Normierung der Abstände von Vektoren, die sich innerhalb des durch die Trainingsdaten bestimmten Raumes befinden, möglich. Die Werte für die Abstände $d(\boldsymbol{x};\boldsymbol{w},b)$ eines Vektors $\boldsymbol{x}$ zur Ebene werden auf Werte im Intervall $[0, \max_{\boldsymbol{x}'}(F'(\boldsymbol{x}'))]$ normiert[22]. Dies bewirkt, dass die Bereiche, in denen ein hoher Wert für F' zu erwarten ist, auch zu den Gebieten gehören, in denen der entsprechende Wert für D hoch ist.
Die in Abbildung 3.23 angedeutete und durch F' bestimmte Kurve ist weiterhin nicht notwendigerweise streng monoton steigend. Es kann eine Situation wie in Abbildung 3.24 auftreten, in der zwei Beobachtungen mit gleichem Abstand zur Ebene aber unterschiedlichen Entscheidungswerten $F'(\boldsymbol{x})$ und $F'(\boldsymbol{x}')$ den gleichen Wert für D erhalten. Bei Auftreten von lokalen Maxima würde bei der Ermittlung von $D(\boldsymbol{x})$ für einen Vektor $\boldsymbol{x}$ das Abfallen von F' nach diesem Optimum als Steigung gewertet werden. Um dieses Problem zu umgehen und der Beobachtung $\boldsymbol{x}'$ aufgrund des höheren Entscheidungswertes mehr Gewicht zu verleihen, kann statt der Länge der erzeugten Kurve das zugehörige Integral, ausgedrückt durch den

[22]Das Maximum wird innerhalb des durch die Trainingsdaten aufgespannten Raumes bestimmt.

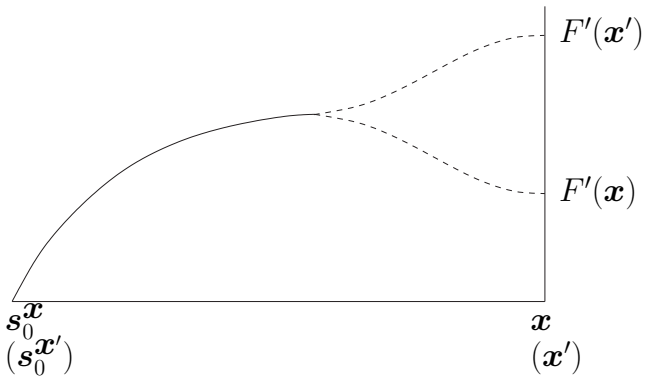

Abbildung 3.24: Berechnung von D bei Vorliegen eines lokalen Maximums

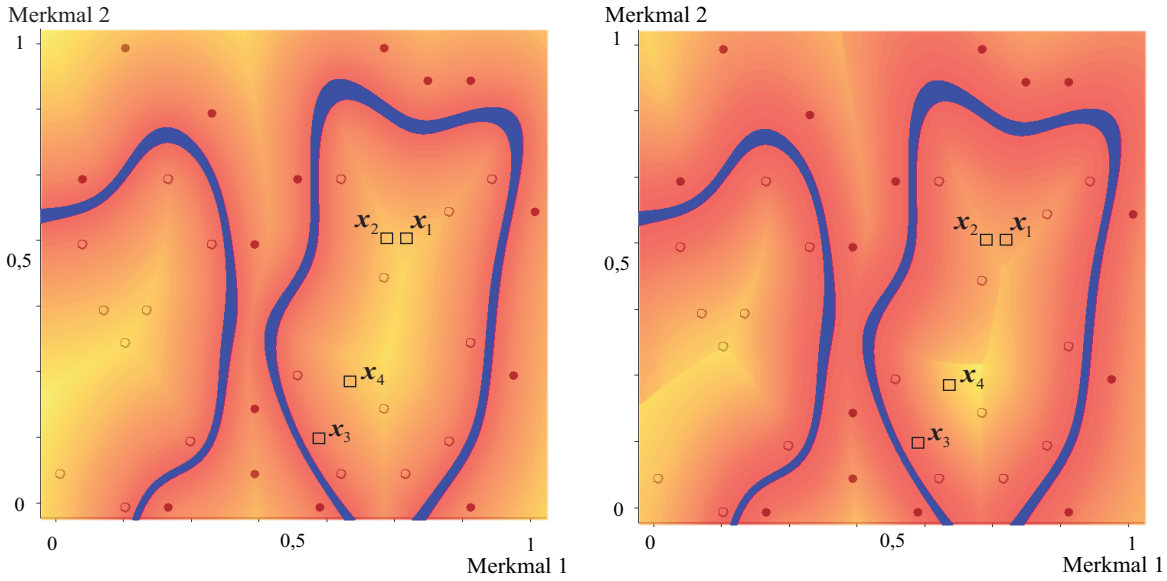

Abbildung 3.25: Resultierende Verteilung von D bei Berücksichtigung der Kurvenlänge (links) und Integralbetrachtungen (rechts)

Flächeninhalt, herangezogen werden. Dazu wird analog zum Vorgehen in Abbildung 3.23 die Strecke zwischen $\boldsymbol{s}_0^{\boldsymbol{x}}$ und $\boldsymbol{x}$ in N Teilstücke zerlegt, mit deren Hilfe der Flächeninhalt approximativ ermittelt wird.
Werden nun statt der ursprünglichen Entscheidungswerte die neu berechneten Werte für D zugrunde gelegt und außerdem diese Werte mit Hilfe des Integrals berechnet, so ergeben sich die beiden Darstellungen in Abbildung 3.25. Der linke Teil der Abbildung zeigt die resultierende Verteilung von D bei der Umsetzung der in Abbildung 3.23 verdeutlichten Idee. Analog zu Abbildung 3.22 werden hohe Werte mit heller und niedrige Werte mit dunklerer Schattierung dargestellt. Der Effekt der oben vorgeschlagenen Transformation der ursprünglichen Entscheidungswerte in den Wert D ist gut zu erkennen. Im Vergleich zu Abbildung 3.22 wird neben dem Einfluss der Entscheidungswerte auch der Einfluss des jeweiligen euklidischen Abstandes zur Trennebene deutlich, was an den entsprechenden Schattierungen

Vektor	F'	D (Kurve)	D (Integral)
$\boldsymbol{x}_1$	1,166	4,459	1,430
$\boldsymbol{x}_2$	1,206	4,035	1,418
$\boldsymbol{x}_3$	0,999	1,476	0,918
$\boldsymbol{x}_4$	2,388	4,380	2,535

Tabelle 3.6: Vergleich von F' und D für ausgewählte Vektoren

zu erkennen ist: Ausgehend von einem Bereich auf der Trennebene nimmt mit zunehmenden Abstand auch der Wert für D zu. Mit anderen Worten existieren in diesen Bereichen keine „Täler“ mehr (wie in Abbildung 3.24) sondern nur ein „Berg“ (wenn der Wert D als dritte Dimension aufgefasst wird). Dem Abstand eines Vektors zur Trennebene kommt im linken Teil der Abbildung große Bedeutung bei der Berechnung von D zu. Die Berechnung auf Basis des Integrals bewirkt hingegen, dass der Entscheidungswert einen höheren Einfluss erhält und somit z.B. der Bereich um die Beobachtung $\boldsymbol{x}_4$ rechts heller, also der Wert $D(\boldsymbol{x})$ für Vektoren aus diesen Bereich höher wird. Der Bereich um die Beobachtungen $\boldsymbol{x}_1$ und $\boldsymbol{x}_2$ ist gegenüber der linken Darstellung dunkler, also der Wert $D(\boldsymbol{x})$ für die entsprechende Vektoren kleiner. Werden für die vier in Abbildung 3.25 zusätzlich eingefügten Vektoren die entsprechenden Werte für F' und D gegenüber gestellt, so resultiert Tabelle 3.6. Es ist zu erkennen, dass sich das Phänomen, welches sich wie oben beschrieben bei F' ergeben kann, hier umkehrt, da nun die euklidische Distanz mit in die Berechnung des Maßes einfließt. Der höhere Entscheidungswert F' von Vektor $\boldsymbol{x}_2$ im Vergleich zu Vektor $\boldsymbol{x}_1$ wird aufgrund der geringeren euklidischen Distanz zur Trennebene in einen niedrigeren Wert für D umgewandelt. Vektor $\boldsymbol{x}_1$ ist im Vergleich zu Vektor $\boldsymbol{x}_2$ weiter von der Trennebene entfernt. Dies drückt sich in einem knapp höheren Wert für D sowohl auf Basis der Kurve als auch auf Basis des Integrals aus. Dennoch werden die vorher bestehenden Situationen bei nicht kritischen Vektoren beibehalten, wie bei den Vektoren $\boldsymbol{x}_3$ und $\boldsymbol{x}_4$ zu sehen ist. Sowohl bei F' als auch bei D erhält $\boldsymbol{x}_3$ im Gegensatz zu $\boldsymbol{x}_4$ aufgrund der geringen Distanz zur Trennebene den deutlich geringeren Wert für alle hier aufgeführten Maße. Dies zeigt, dass sowohl die Distanz im Eingaberaum als auch die durch die Trennebene induzierten Entscheidungswerte in das neue Maß einfließen. Mittels D können nun Aussagen zur Güte der Klassifikation analog zu Abschnitt 3.7.2 generiert werden. Je höher der Wert für $D(\boldsymbol{x})$, desto wahrscheinlicher ist die Richtigklassifikation des Vektors $\boldsymbol{x}$ im Vergleich zu einem anderen Vektor $\boldsymbol{x}'$ mit geringerem Wert $D(\boldsymbol{x}')$. Dies liegt darin begründet, dass ein hoher Wert $D(\boldsymbol{x})$ impliziert, dass der Vektor $\boldsymbol{x}$ entweder sehr weit von der Ebene im Eingaberaum entfernt ist, oder einen sehr hohen Wert $F'(\boldsymbol{x})$ erhält. Beides lässt auf eine sichere Klassifikation des Vektors schließen. Im Gegensatz zu den herkömmlichen Entscheidungswerten können keine Aussagen bezüglich des absoluten Wertes gemacht werden, sodass $D(\boldsymbol{x}) = 1$ nicht unbedingt bedeuten muss, dass der entsprechende Vektor auf einer der Hilfsebenen positioniert ist, wie es bei der Betrachtung der Entscheidungswerte der Fall ist. Mittels D kann somit lediglich die Sicherheit der Zuweisung bei Vergleich mehrerer

Vektoren bewertet werden.

Diese Vorstellung von D als ein Bewertungsmaß soll an dieser Stelle als eine von mehreren Möglichkeiten verstanden werden, die Distanz zusätzlich zum Entscheidungswert bei der Bewertung der Sicherheit der Zuweisung zu berücksichtigen. Eine Alternative besteht darin, $D(\boldsymbol{x})$ für einen Vektor $\boldsymbol{x}$ in leicht abgewandelter Form zu verwenden. Statt der Verwendung der minimalen Distanz zur Trennebene in Gleichung (3.8) könnte die Minimierung von $D(\boldsymbol{x})$ über alle Vektoren auf der Trennebene herangezogen werden. Dies würde die bei der Berechnung des Integrals in Abbildung 3.25 entstandenen Sprungstellen relativieren und zu einer ebenmäßigeren Darstellung führen.
Es sei darauf hingewiesen, dass diese Erweiterung im empirischen Teil der Arbeit nicht berücksichtigt werden kann. Dies liegt darin begründet, dass das hier beschriebene Phänomen bei Beschränkung auf zwei der später auftretenden Merkmale nicht existiert. Daher wurde die Vorgehensweise hier lediglich anhand dieses fiktiven Datensatzes erläutert.

3.7.4 Ergebnisinterpretation bei Multiklassifikation

Die bisherigen Darstellungen zur Interpretation der Entscheidungswerte basieren auf der Annahme, dass eine binäre Klassifikation, also eine Einteilung der Daten in lediglich zwei Klassen, vorliegt. Diese Vorgehensweise ist leicht auf die Multiklassifikation übertragbar, wobei die Verfahrensweise bei der Fuzzy-Multiklassifikation genutzt werden kann. Wie in Abschnitt 2.2.4 beschrieben, werden dabei pro Beobachtung Zugehörigkeitswerte berechnet, die sich aus den ursprünglichen Entscheidungswerten ergeben. Bei der Visualisierung der Daten sollten möglichst Ergebnisse aller berechneten SVM berücksichtigt werden, sodass die Zugehörigkeit einer Beobachtung zu den unterschiedlichen Klassen leicht zu erkennen ist. Um eine Intensität der Zuweisung zu einer Klasse auch außerhalb der zu maximierenden Spanne angeben zu können, muss die Ausgabe der Fuzzy-SVM derart modifiziert werden, dass auch Membership-Werte größer als 1 zugelassen werden. Dazu wird ein neuer Membership-Wert m'_k, der die Zugehörigkeit zu Klasse k angibt, im Falle einer OAO-Trennung bestimmt durch

$$m'_k(\boldsymbol{x}) = \begin{cases} \min_{i \neq k, i=1,\ldots,K} F^{[ki]}(\boldsymbol{x}) & \text{, falls } m_k(\boldsymbol{x}) = 1 \\ m_k(\boldsymbol{x}) & \text{, sonst} \end{cases}$$

und bei Vorliegen der OAA-Trennung durch

$$m'_k(\boldsymbol{x}) = \begin{cases} \min_{i=1,\ldots,K} |F^{[i]}(\boldsymbol{x})| & \text{, falls } m_k(\boldsymbol{x}) = 1 \\ m_k(\boldsymbol{x}) & \text{, sonst.} \end{cases}$$

Dadurch wird die Angabe der Intensität der Zugehörigkeit analog zur Biklassifikation ermöglicht. Die neuen Membership-Werte für die Zugehörigkeit zu Klasse 1 eines fiktiven Datensatz mit drei Klassen bei einer OAO-Trennung werden in Abbildung 3.26 dargestellt. Die Quadrate bilden die betreffenden Beobachtungen der Klasse

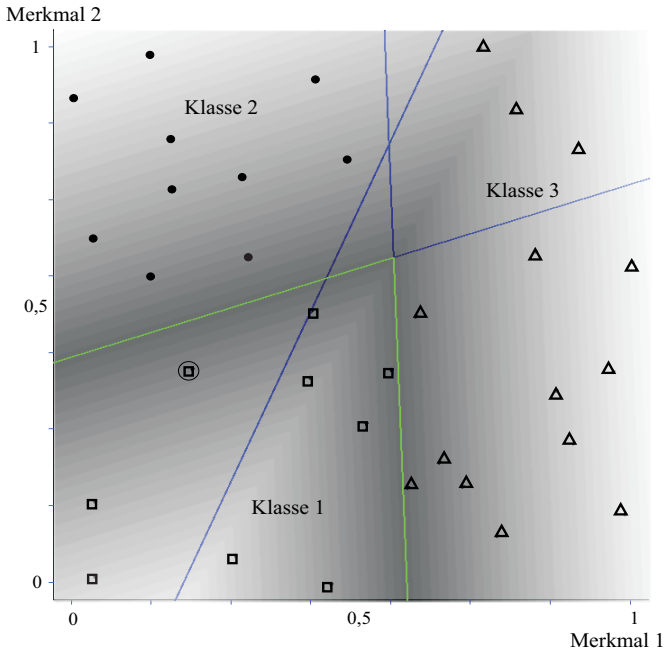

Abbildung 3.26: Visualisierung der Zugehörigkeit zu Klasse 1 bei OAO-Trennung

1. Je dunkler die Schattierung dargestellt wird, desto niedriger ist der jeweilige Membership-Wert m_1' in dem betreffenden Bereich. Es sind lediglich die Ergebnisse der Trennung von Klasse 1 gegen 2 und Klasse 1 gegen 3 relevant. Die dritte Trennebene geht hier aufgrund der theoretischen Hintergründe nicht in die Berechnung der Membership-Werte ein. Die klassifizierten Daten können analog zu Abschnitt 3.7.2 bewertet werden. Werden zusätzlich die Membership-Werte m_2' und m_3' betrachtet, so bietet sich eine Visualisierung der Werte durch Parallelkoordinaten an. Dieses Verfahren wurde von *Inselberg* (1985) entwickelt, und wurde 2000 erstmals im Marketing zur Abbildung konkurrierenden Preisverhaltens eingesetzt (*Klemz, Dunne* (2000)). Die Idee besteht in der gleichzeitigen Visualisierung von vielen Merkmalsausprägungen, die die Beobachtungen beschreiben (*Ankerst* (2000)). In diesem Fall werden die Vektoren visualisiert, die durch die Zugehörigkeit zu den einzelnen Klassen charakterisiert werden. Dabei werden die Ausprägungen an parallel verlaufenden Koordinaten veranschaulicht und pro Beobachtung jeweils miteinander verbunden. Die Daten der Abbildung 3.26 etwa werden in Abbildung 3.27 mittels Parallelkoordinaten dargestellt. Eine Linie, die die drei Achsen auf verschiedenen Höhen schneidet, repräsentiert eine der in Abbildung 3.26 visualisierten Beobachtungen. Die schwarze, fett markierte Verbindung wird durch einen höheren Wert bezüglich Klasse 1 und niedrige Werte bezüglich der Klassen 2 und 3 gekennzeichnet und würde demnach zu Klasse „1“ zugeordnet werden. Die betreffende Beobachtung ist in Abbildung 3.26 mit einem Kreis gekennzeichnet. Die eindeutige Nichtzuordnung zu Klasse 3 durch den sehr geringen Membership-Wert m_3 ist ebenfalls anhand der Abbildung zu erkennen. Je weiter die Ausprägungen in der Mitte einer Parallelkoordinate positioniert sind, desto weniger stark ist ihre Zugehörigkeit zu der jeweiligen Klasse

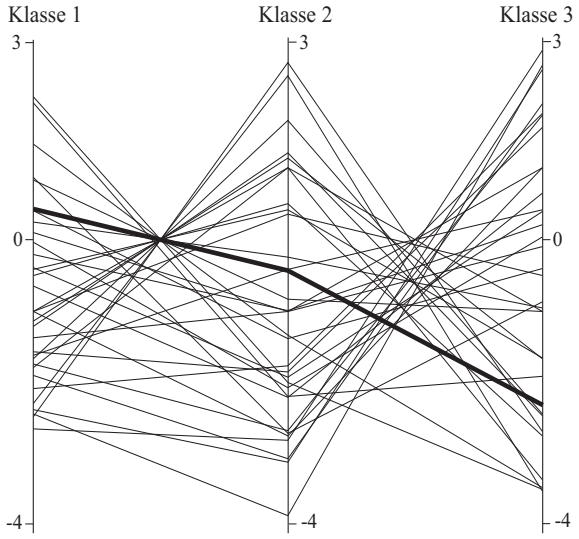

Abbildung 3.27: Veranschaulichung der Zugehörigkeiten mittels Parallelkoordinaten

ausgeprägt. Werden die Klassen nahezu eindeutig getrennt, so gibt es nur wenige Beobachtungen, die in der Mitte bei einem Membership-Wert nahe 0 positioniert sind. Aufgrund dieser guten Trennung der Klassen weisen die Beobachtungen bei jeweils einer Klasse einen hohen, bei den anderen beiden Klassen jeweils einen niedrigen Membership-Wert m'_k auf. Analog zur Biklassifikation können nun die Intervalle der Membership-Werte durch Einführung unterschiedlicher Bereiche für spätere Marketingmaßnahmen aufgeteilt werden. Dies wird für jede Klasse einzeln festgelegt. Steht nur eine Klasse im Fokus der Analyse, so wird analog zur Biklassifikation vorgegangen, also nur die Membership-Werte bezüglich der betreffenden Klasse betrachtet und ausgewertet. Dies wird insbesondere durch die Zulassung von Membership-Werten größer eins ermöglicht, wodurch wichtige Informationen hinzu gewonnen werden, die die Intensität der Zuordnung betreffen. Die Entscheidungswerte können analog zu Abschnitt 3.7.2 behandelt werden, sodass auch bei der Multiklassifikation, für jede Klasse einzeln betrachtet, unterschiedliche Bereiche entstehen. Stehen mehrere Klassen im Focus der Analyse, so müssen Prioritäten gesetzt werden. Dies liegt darin begründet, dass bei schlechterer Trennung mehrere positive Membership-Werte für Klassen auftreten können, die aufgrund entsprechender Bereichseinteilungen in widersprüchlichen Strategien resultieren können, wie auf Seite 152 in Abschnitt 4.4.2 zu sehen sein wird.
Die Festlegung der Bereichsgrenzen kann sich zwischen den Klassen unterscheiden, um klassenindividuelle inhaltliche Interpretation zu ermöglichen. Dies drückt sich in einer individuellen Bereichseinteilung pro Parallelkoordinate, also pro Membership-Wert aus.
Ein Nachteil dieser Art der Visualisierung liegt in der Darstellung selbst. Bei nur wenigen Datenpunkten wird die Übersichtlichkeit der Grafik stark beeinträchtigt. Dennoch wird somit ein erster Eindruck der Verteilung innerhalb aller Klassen ermöglicht und Parallelkoordinaten sollten zur Beurteilung der Klassifikation von

mehreren Klassen herangezogen werden.
Um kundenindividuelle Marketingstrategien zu entwickeln, werden die hier vorgestellten Aspekte in der Anwendung von SVM in Abschnitt 4.4.2 durchgeführt.

3.8 Beurteilung der Güte der Klassifikation

Um die Güte eines Modells zur Klassifikation zu bewerten und mit anderen vergleichen zu können, müssen Maße zur Beurteilung der Ergebnisse herangezogen werden. Viele Veröffentlichungen bewerten die Ergebnisse von SVM lediglich anhand der jeweils erreichten Trefferquoten, was einen Großteil der durch die Klassifikation erzielten Ergebnisse nicht beinhaltet. Daher bietet dieser Abschnitt einen Überblick der gerade bei SVM zur Verfügung stehenden Maße. Dazu werden die Ergebnisse zunächst anhand der Trefferquote eingeschätzt. Für den Zwei-Klassen-Fall besteht ferner die Möglichkeit des Einsatzes von ROC-Kurven. Speziell bei Anwendung der SVM kann auf Gütemaße zurückgegriffen werden, die deren Ergebnisse berücksichtigen. Nach einem kurzen Überblick über die Definition und Anwendung dieser Maße werden diese anhand eines exemplarischen Datensatzes verglichen.

3.8.1 Trefferquoten

Beim bekannten Konzept der Trefferquoten wird untersucht, ob die wahren mit den prognostizierten Klassenzugehörigkeiten übereinstimmen, um die Prognosegüte eines Modells anzugeben. Die Gesamttrefferquote eines Klassifikationsinstrumentes lässt sich durch den Anteil der richtig klassifizierten Beobachtungen an allen vorliegenden Beobachtungen berechnen:

$$TQ = \frac{T(l)}{l},$$

wobei $T(l) \leq l$ die Anzahl der richtig klassifizierten Beobachtungen bei einem Datensatz im Umfang von l Beobachtungen angibt. Auf Basis von Ergebnissen eines Testdatensatzes oder nach Durchführung einer Kreuzvalidierung kann mit TQ die Wahrscheinlichkeit angegeben werden, dass das Verfahren eine ungesehene Beobachtung richtig klassifiziert.
Eine restriktive Betrachtung der Gesamttrefferquote ist bei vielen Anwendungen nicht adäquat. Liegt bei der Kundenklassifikation etwa eine Aufteilung der wichtigen und unwichtigen Kunden in 30% und 70% vor, so ist es wichtiger, die ersteren richtig zu klassifizieren, als eine hohe Gesamttrefferquote zu Lasten der richtig klassifizierten wichtigen Kunden zu erreichen. Daher sollte der Anteil der Treffer für jede Klasse einzeln betrachtet werden, um mögliche Missverhältnisse aufzudecken. Dazu dient die in Tabelle 3.7 abgebildete Klassifikationsmatrix für den Zwei-Klassen-Fall, in der die verschiedenen Typen von Zuordnungen dargestellt werden. Hierbei werden zur Abkürzung die englischen Bezeichnungen gewählt: true positive (TP), true negative (TN), false positive (FP) und false negative (FN). Die klassenbezogenen

		Prognostizierte Klasse	
		1	−1
Wahre Klasse	1	richtig (TP)	falsch (FN)
	−1	falsch (FP)	richtig (TN)

Tabelle 3.7: Klassifikationsmatrix für die Biklassifikation

Trefferquoten ergeben sich dann aus

$$TQ_{\text{Klasse } 1} = \frac{TP}{TP + FN}$$

und

$$TQ_{\text{Klasse } -1} = \frac{TN}{TN + FP}.$$

Im Mehrklassenfall ergibt sich die Trefferquote für eine Klasse k entsprechend aus

$$TQ_k = \frac{T(l^{[k]})}{l^{[k]}},$$

wobei durch $l^{[k]}$ die Anzahl der Beobachtungen in Klasse k bezeichnet wird. So kann eine differenzierte Beurteilung der Klassifikationsgüte erfolgen.
Die Trefferquote wird zur Bestimmung guter Parameter (vgl. Abschnitt 3.3) und zur Beurteilung der Prognosegüte eines Modells eingesetzt. Bei welcher Trefferquote von einer guten Klassifikation gesprochen wird, hängt von verschiedenen Faktoren, z.B. dem Verhältnis der einzelnen Klassenumfänge, ab. Dazu kann neben dem Gleichverteilungskriterium auch das Kriterium der größten Klasse (Größte-Gruppen-Kriterium) verwendet werden (vgl. *Temme* (2002)), nach dem eine Trefferquote als gut beurteilt wird, wenn sie größer als der Anteil der Beobachtungen der größten Klasse an den gesamten Beobachtungen ist. Dies hat den Hintergrund, dass ohne Vorliegen eines Klassifikationsinstrumentes alle Beobachtungen als zu der größten Gruppe gehörig klassifiziert werden würden, um eine hohe Trefferquote zu erreichen. Eine Klassifikationsmethode müsste besser sein als bei diesem trivialen Vorgehen. *Morrison* (1969) verwendet außerdem das „Proportional-Chance-Kriterium“, welches die Anteile der jeweiligen Klassen am Gesamtumfang der Daten berücksichtigt. Demnach ist ein Klassifikator als gut einzuschätzen, wenn er eine Trefferquote von

$$TQ \geq \sum_{k=1}^{K} \left(\frac{l^{[k]}}{l} \right)^2$$

bei K gegebenen Klassen und insgesamt l Beobachtungen erreicht.
Zusammenfassend lässt sich festhalten, dass die Trefferquote als ein einfaches Vergleichsmaß zur Einordnung der Güte eines Klassifikationsinstrumentes geeignet ist.

3.8.2 Einbeziehung der Werte der Entscheidungsfunktion

Neben der Trefferquote stellen *Guyon et al.* (2002) ein weiteres Maß zur Beurteilung von SVM vor, welches die Ausprägungen der Entscheidungswerte $F(\boldsymbol{x})$ im Zwei-Klassen-Fall[23] mit einbezieht. Mit $\theta^+ = \min_{\boldsymbol{x}:y=1} F(\boldsymbol{x})$ und $\theta^- = \max_{\boldsymbol{x}:y=-1} F(\boldsymbol{x})$ ist dieses mit M_{ext} bezeichnete Maß definiert als (vgl. *Valentini et al.* (2004))

$$M_{ext} = \frac{\theta^+ - \theta^-}{\max F(\boldsymbol{x}) - \min F(\boldsymbol{x})}. \tag{3.9}$$

Aufgrund der Definition von θ^+ und θ^- gilt

$$-1 \leq M_{ext} \leq 1.$$

M_{ext} wird positiv, sobald es keine fehlklassifizierte Beobachtung gibt, oder die Werte für $F(\boldsymbol{x})$ der fehlklassifizierten Beobachtungen aus Klasse „+1“ (bzw. Klasse „-1“) absolut gesehen kleiner sind als die Werte für $F(\boldsymbol{x})$ der richtig klassifizierten Vektoren aus Klasse „-1“ (bzw. Klasse „+1“). Daher gilt: je höher der Wert für M_{ext}, desto sicherer („more confident“) ist die Klassifikation (*Valentini et al.* (2004)). Dabei sagt dieses Maß noch nichts über die Prognosegüte aus.

Gilt $\min F(\boldsymbol{x}) = \max F(\boldsymbol{x})$ für Vektoren einer Klasse, also $F(\boldsymbol{x}) = 1$ für alle Beobachtungen $\boldsymbol{x}$ einer Klasse, so ist $M_{ext} = 1$. Das Minimum $M_{ext} = -1$ hingegen wird genau dann erreicht, wenn $\min F(\boldsymbol{x}) - \max F(\boldsymbol{x}) = \theta^+ - \theta^-$ gilt, also wenn der minimale und der maximale Wert der pro Klasse von fehlklassifizierten Beobachtungen der jeweils anderen Klasse angenommen wird.[24]
Ein formaler Zusammenhang zwischen der Trefferquote und M_{ext} kann nicht angegeben werden. So zieht eine hohe Trefferquote nicht notwendigerweise einen hohen Wert für M_{ext} nach sich, wie die folgende Abbildung verdeutlicht.

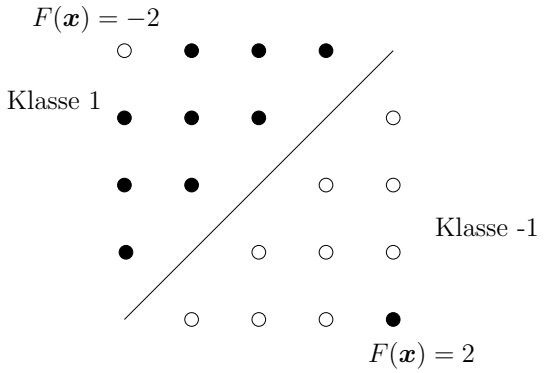

Abbildung 3.28: Verdeutlichung des fehlenden Zusammenhangs zwischen der Ausprägung von M_{ext} und der Trefferquote

[23] Wie die Vorgehensweise auf die Multiklassifikation erweitert werden kann, ist bisher offen.
[24] Dies ist leicht durch Fallunterscheidung zu zeigen.

Innerhalb der Trennung der Daten in Abbildung 3.28 gibt es in den beiden Klassen jeweils eine falsch klassifizierte Beobachtung, die den maximalen Wert von $F(\boldsymbol{x})$ der ihr durch die Trennebene zugewiesenen Klasse annimmt. Damit liegt die Trefferquote bei 90%, allerdings gilt $M_{ext} = -1$. Andererseits deutet ein Wert von $M_{ext} = 1$ auf eine Trefferquote von 100% hin. Alle Beobachtungen bekommen den Wert $|F(\boldsymbol{x})| = 1$ zugewiesen, da $M_{ext} = 1$ nicht anders erreicht werden kann. Dies bildet einen großen Nachteil des Maßes. Bei $M_{ext} = -1$ kann nicht darauf geschlossen werden, dass eine schlechte Klassifikation vorliegt.
Um diesem Umstand Rechnung zu tragen und unempfindlicher gegenüber Ausreißern zu sein, kann statt der Extrema θ^+ und θ^- jeweils der Median herangezogen werden (*Valentini et al.* (2004)):

$$M_{med} = \frac{\mu^+ - \mu^-}{\max F(\boldsymbol{x}) - \min F(\boldsymbol{x})},$$

mit μ^+ (bzw. μ^-) als dem Median der Werte von $F(\boldsymbol{x})$ von Beobachtungen aus Klasse $+1$ (bzw. aus Klasse -1). Je höher der Wert für M_{med}, desto höher ist der Grad der insgesamt erreichten Zuverlässigkeit der Klassifikation (vgl. *Valentini et al.* (2004)). Empirische Analysen zeigen (*Guyon et al.* (2002)), dass ein größerer Zusammenhang zwischen den Werten von M_{med} und der erzielten Trefferquote besteht als bei Verwendung von M_{ext}, bei dem die Ausreißer einen zu hohen Einfluss ausüben. Angestrebt wird also ein Wert, der möglichst nahe „1“ gelegen ist. Zu einer zuverlässigen Beurteilung sollten aber weitere Informationen herangezogen werden. Daher ist dieses Maß lediglich als Ergänzung für den Einsatz im Marketing sinnvoll.

3.8.3 Receiver Operating Characteristics

Eine übliche Methode zur Einordnung von Klassifikationsergebnissen unter Berücksichtigung der beiden Fehler FP und FN bilden die receiver operating characteristics (ROC), die häufig bei der Bewertung von Modellen im Bereich der Medizin Verwendung finden (*Metz* (1978)) und teilweise in Verbindung mit SVM zum Einsatz kommen (vgl. z.B. *Bazzani et al.* (2001)). So bilden die beiden Fehler bei der Anwendung im medizinisch-diagnostischen Bereich, einen Gesunden als krank zu diagnostizieren, oder die Krankheit bei Vorliegen nicht zu entdecken, die Grundlage der ROC-Analyse. Da ähnliche Probleme, wie bereits in Abschnitt 3.2.2 angesprochen, ebenso bei der Klassifikation von Kunden auftreten, soll die ROC-Analyse zur möglichen Beurteilung der Klassifikationsgüte von SVM im Marketing erläutert werden.

Das zugrunde liegende Ziel bei der Durchführung einer ROC-Analyse ist die Visualisierung der Leistung eines Klassifikationsinstrumentes mittels der so genannten ROC-Kurve, um verschiedene Modelle vergleichen zu können. Dafür sind zwei Maße notwendig, die in der Literatur mit Spezifität (Sp) und Sensitivität (Se) bezeichnet werden. Sie geben den Anteil der richtig klassifizierten Beobachtungen aus Klasse „+1“ respektive Klasse „-1“ an und können durch die in Tabelle 3.7

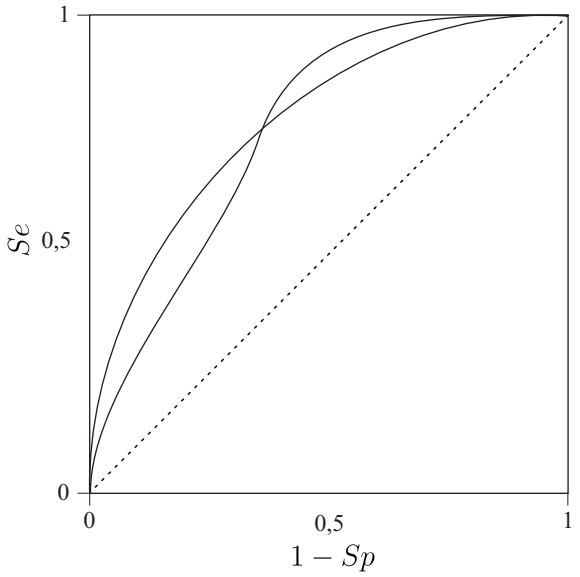

Abbildung 3.29: Beispielhafte ROC-Kurven determiniert durch Spezifität und Sensitivität

verwendeten Abkürzungen berechnet werden:

$$Se = \frac{TP}{TP + FN}$$

und

$$Sp = \frac{TN}{TN + FP}.$$

Die Bezeichnungen FP und FN können in Analogie zu den Fehlern erster und zweiter Art beim Testen von Hypothesen gesehen werden. Ein Fehler erster Art tritt bei der irrtümlichen Ablehnung einer zu prüfenden Hypothese auf. Ein Fehler zweiter Art liegt bei der unrechtmäßigen Annahme einer nicht zutreffenden Hypothese vor. Bei der Klassifikation in zwei Klassen „+1“ und „-1“ wird hier die Nullhypothese überprüft, dass eine Beobachtung zu Klasse „+1“ zugehörig ist. Die Ablehnung dieser Hypothese bei Vorliegen einer Klasse „+1“-Beobachtung resultiert in einem Fehler erster Art (FN). Der Fehler zweiter Art wird durch FP ausgedrückt.
Ähnlich zu der Möglichkeit, die Macht eines statistischen Tests durch die Operationscharakteristik (OC) zu visualisieren (vgl. *Decker, Wagner* (2002)), wird die ROC-Kurve verwendet. Der dazu benötigte zwei-dimensionale Raum wird nun von Sensitivität (Se) (y-Achse) sowie $1 - Sp$ (x-Achse) aufgespannt. Für Klassifikationsinstrumente mit reellwertigen Ausgabewerten ergibt sich dann eine ROC-Kurve durch die Verschiebung eines Grenzwertes zur Klassifikation der Beobachtungen (*Metz* (1978)). In Abbildung 3.29 sind zwei dieser Kurven exemplarisch aufgezeigt. Die Kurve, die eine gute Klassifikationsgüte besitzt, würde nahe der linken oberen Ecke verlaufen, da in diesem Fall sowohl eine hohe Sensitivität als auch eine hohe Spezifität vorliegt, also die Anzahl der richtig zugeordneten positiven Beobachtung

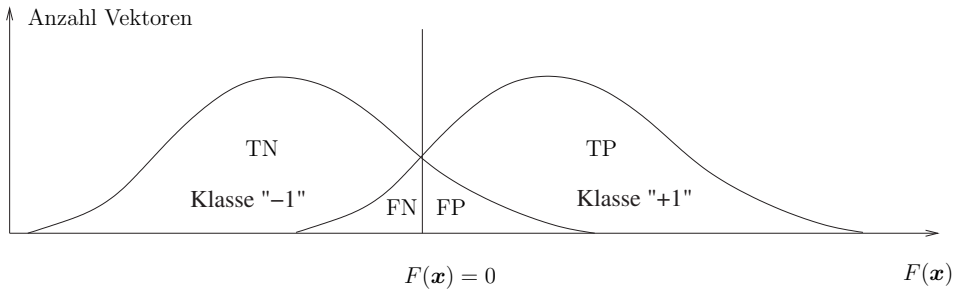

Abbildung 3.30: Mögliche Verteilung von F mit Kennzeichnung der auftretenden Fehler

hoch ist bei gleichzeitiger Minimierung der Anzahl falsch zugeordneter negativer Beobachtungen. Die Beobachtungen werden also in beiden Klassen sehr gut klassifiziert. Eine ROC-Kurve dominiert eine andere, wenn sie für alle Grenzwerte oberhalb der anderen verläuft. Die zugehörige Klassifikationsmethode würde dann als geeignet(er) charakterisiert werden. In Abbildung 3.29 wird wegen des Schnittpunktes der beiden Kurven keine der beiden Kurven von der anderen dominiert. Liegt der Fokus allerdings auf der Richtigklassifikation bei unterschiedlich gewichteten Klassen, so kann auch ein Teilbereich der Kurve von Interesse sein und die Kurve als gut bewertet werden, obwohl sie eine vergleichbare nicht dominiert. Die gestrichelte Linie deutet den Verlauf bei zufälliger Zuordnung zu zwei Klassen an. Von Interesse sind daher nur Kurven, die sich in der oberen Hälfte des aufgespannten Raumes befinden.
Die Erstellung einer ROC-Kurve bei Vorliegen der Ausgabe einer SVM verläuft auf Basis der in aufsteigender Reihenfolge von $\min F(\cdot)$ bis $\max F(\cdot)$ geordneten Ausgabewerte $F(\cdot)$. Eine mögliche Verteilung der Vektoren in Abhängigkeit der Ausgabewerte ist in Abbildung 3.30 gegeben. Bei dem bei SVM verwendeten Grenzwert zur Klassifikation von Beobachtungen von $F(\boldsymbol{x}) = 0$ ergeben sich die vier Bereiche TN, FN, TP und FP, wie in der Abbildung angedeutet. Diese Konstellation entspricht genau einem Punkt auf der zugehörigen ROC-Kurve. Wird dieser Grenzwert nun von links ($\min F(\boldsymbol{x})$) nach rechts ($\max F(\boldsymbol{x})$) verschoben, so ergeben sich die Koordinaten, die die ROC-Kurve bilden.
Alternativ zu ROC-Kurven können auch Recall-Precision-Kurven berechnet werden (*Witten, Frank* (2000)), die häufig im Rahmen des Information Retrieval Verwendung finden. Der Wert für den Recall entspricht dem der Sensitivität. Precision wird ebenso auf Basis der Bezeichnungen aus Tabelle 3.7 berechnet durch

$$Precision = \frac{TP}{TP + FP}.$$

Precision gibt den Anteil der richtig klassifizierten Beobachtungen aus der Klasse der positiven Objekte an allen Beobachtungen an, die als positiv klassifiziert worden sind. Analog zu ROC-Kurven lässt sich die Veränderung eines Schwellenwertes auch anhand einer Kurve darstellen. Es ist zu beachten, dass aufgrund der Definition von Recall und Precision die ROC- und Recall-Precision-Kurven einen anderen Verlauf

haben. Im letzteren Fall liegt die angestrebte Situation in der rechten oberen Ecke, in der Recall=Precision=1 gilt.
Ein ähnliches Konzept liegt den häufig im Marketing angewandten Lift Charts zugrunde, die auf probabilistischen Ausgaben eines Klassifikationsinstrumentes beruhen (*Witten, Frank* (2000)) und eine zu ROC vergleichbare Kurve hervorbringen.

Ein Maß zur Beurteilung der Prognosegüte einer Methode auf Basis einer ROC-Kurve liefert die so genannte „area under the curve“ (AUC), die Fläche unterhalb dieser Kurve. Diese kann Werte im Intervall $[0,5;1]$ annehmen, da die ROC-Kurve oberhalb der Diagonalen positioniert ist. Durch AUC wird somit ein weiteres Maß zur Beurteilung der Güte eines Klassifikationsinstrumentes bereit gestellt. Der Wert korrespondiert mit der Wahrscheinlichkeit, dass ein zufällig ausgewähltes Paar einer positiven und einer negativen Beobachtung richtig klassifiziert wird (*Hanley, McNeil* (1982)). So bedeutet $AUC = 1$, dass es sich um ein perfekt trennendes Verfahren handelt. Damit kann die visuelle Darstellung der Klassifikationsgüte eines Verfahrens auf ein reellwertiges Maß reduziert werden. Dennoch sollte dieser Wert nicht allein als Bewertungsmaßstab dienen, sondern immer in Kombination mit anderen Maßen gesehen werden. So ergibt sich möglicherweise ein höherer AUC-Wert für ein Verfahren, dessen ROC-Kurve sich mit der Kurve eines anderen Verfahrens schneidet. In diesem Fall kann aber nicht davon gesprochen werden, dass ein Verfahren das andere dominiert. Ein möglicher Vorteil in einem bestimmten Bereich der Kurve würde durch die ausschließliche Betrachtung des AUC Wertes allerdings nicht erkannt werden.

Die ROC-Analyse kann zwei verschiedene Zielsetzungen verfolgen. So kann sie zum einen dazu dienen, eine gute Parameterkonstellation für SVM während des Optimierungsprozesses zu finden, indem die ROC-Kurven der aus verschiedenen Kombinationen resultierenden SVM miteinander verglichen werden. Die mittels Gridsearch ermittelten Ergebnisse können durch entsprechende ROC-Kurven überprüft werden. Neben der Ermittlung einer optimalen Parameterkonstellation verfolgt die ROC-Analyse hauptsächlich das Ziel, die Leistung eines Klassifikationsinstrumentes nach der Optimierung zu beurteilen und mittels AUC zu bewerten.

3.8.4 Vergleich der Gütemaße

Anhand des bereits in Abbildung 3.6 verwendeten Datensatzes „Kredit“ werden nun die vorgestellten Gütemaße zur Beurteilung des Klassifikationsergebnisses verglichen. In diesem Datensatz liegen insgesamt 1000 Beobachtungen vor, die auf Basis von elf Merkmalen[25] zwei Klassen zugeordnet worden sind. Es handelt sich

[25]Hierbei handelt es sich um eine modifizierte Form des vom Institut für Statistik und Ökonometrie der Universität Hamburg stammende Datensatzes *German* (Quelle: http://www.liacc.up.pt/ML/statlog/datasets/german/german.doc.html, Zugriff: 6.6.2004), bei dem nur die 11 metrischen der ursprünglich 20 Variablen verwendet werden.

um die Kunden eines Kreditinstituts, die in die beiden Gruppen „good credit“ (700 Beobachtungen) bzw. „bad credit“ (300 Beobachtungen) eingeteilt worden sind. Diese Art von Daten können ebenfalls im Marketing relevant sein, wenn es sich beispielsweise um die Ausgestaltung von Zahlungsweisen im Versandhandel dreht. So können Waren bei einem guten Kunden auf Rechnung versendet werden, wohingegen bei Vorliegen der Klasse „bad credit“ nur auf Vorkasse geliefert wird. Zur Anwendung der Gütemaße wird der Datensatz zunächst zufällig in einen Trainings- und einen Testdatensatz eingeteilt, sodass die ursprüngliche Klassenverteilung in beiden Teilen erhalten bleibt. Da auf die in Abbildung 3.6 als gut ermittelte Parameterkonstellation ($\gamma = 1$ und $C = 1$) zurückgegriffen werden soll, muss das Verhältnis der Trainings- und Testdaten ebenfalls beibehalten werden. Für die Erstellung von Abbildung 3.6 ist eine sechsmalige 6-fach-Kreuzvalidierung durchgeführt worden, sodass aus dem ursprünglichen Datensatz 166 Beobachtungen für den Test zufällig ausgewählt wurden.

Die bei der obigen Parameterkonstellation ermittelte Trefferquote von $74,10\%$ innerhalb des Testdatensatzes weicht von der mittels sechsmaliger 6-fach-Kreuzvalidierung bestimmten durchschnittlichen Trefferquote von $75,22\%$ ein wenig ab, was auf die Auswahl genau eines Testdatensatzes zurückzuführen ist.

		Trainingsdaten Prognostizierte Klasse		
		1	-1	Summe
Wahre Klasse	1	554	30	584
	-1	136	114	250

	Testdaten Prognostizierte Klasse		
	1	-1	Summe
1	106	10	116
-1	33	17	50

Tabelle 3.8: Klassifikationsmatrizen für den Trainings- und Testdatensatz

Die in Tabelle 3.8 enthaltenen Klassifikationsmatrizen für beide Datensätze enthalten in den Zeilen die wahren Klassenzugehörigkeiten und in den Spalten die mittels SVM prognostizierten Werte. Die Gesamttrefferquote täuscht in diesem Fall über die eigentlich sehr schlechten Ergebnisse bezogen auf die Klasse „-1“ hinweg. Die unterschiedlichen Trefferquoten (im Testdatensatz) von $91,38\%$ für Klasse „+1“ bzw. 34% für Klasse „-1“ zeigen, dass diese Art der einfachen Ergebnisdarstellung wertvolle zusätzliche Informationen liefern kann. Liegt das Proportional-Chance-Kriterium zugrunde, so müsste für die methodenbedingte Trefferquote

$$TQ > \sum_{k=1}^{2} \left(\frac{l_t^{[k]}}{l_t} \right)^2 = (116/166)^2 + (50/166)^2 = 0,579$$

gelten, wobei $l_t^{[k]}$ den Umfang der Testbeobachtungen in Klasse k und l_t die Gesamtanzahl der Beobachtungen im Testdatensatz angibt. Dies ist im vorliegenden Fall erfüllt. Als weiteres Vergleichsmaß kann das Größte-Gruppen-Kriterium (vgl.

Morrison (1969)) hinzugezogen werden. Dabei muss

$$TQ > \frac{\max_k l_t^{[k]}}{l_t}$$

gelten, da angenommen wird, dass jede Beobachtung der größten Gruppe zugeordnet wird. Dies wird im vorliegenden Fall ebenfalls erfüllt, da

$$TQ = 0,741 > \frac{116}{166} = 0,699.$$

Trotz der eher als schlecht zu bewertenden Trefferquote für Klasse „-1“ ist nach den obigen Kriterien SVM auf Basis der gewählten Parameterkonstellation als gut einzustufen. Wie dies im Einzelfall zu beurteilen ist, muss auf Basis der zugrunde liegenden Fragestellung entschieden werden.

Werden nun neben der reinen Zuordnung auch die mittels SVM generierten Entscheidungswerte mit einbezogen, so resultiert nach Gleichung (3.9)

$$M_{ext} = \frac{\theta^+ - \theta^-}{\max F(\boldsymbol{x}) - \min F(\boldsymbol{x})} = \frac{-0,658 - 1,378}{1,555 + 1,882} = -0,592.$$

Da dieser Wert gemessen am Wertebereich von M_{ext} recht niedrig ist, kann gefolgert werden, dass diese Klassifikation als nicht zuverlässig einzustufen ist. Wird zusätzlich das gegenüber Ausreißern unempfindlichere Maß M_{med} verwendet, so resultiert:

$$M_{med} = \frac{\mu^+ - \mu^-}{\max F(\boldsymbol{x}) - \min F(\boldsymbol{x})} = \frac{1,03 - 0,49}{1,555 + 1,882} = 0,157.$$

Dieser Wert fällt deutlich höher aus. Anhand dieser beiden Gütemaße, die die Werte der Entscheidungsfunktion mit einbeziehen, kann darauf geschlossen werden, dass hier eine eher weniger zufrieden stellende Klassifikation vorliegt. In Abbildung 3.31 sind dazu zusätzlich die Quantile der Entscheidungswerte für Klasse „-1“ und Klasse „+1“ anhand von Boxplots für beide Klassen dargestellt. Es fällt auf, dass die Werte für Klasse „-1“ mit einem Median von 0,49 insgesamt sehr weit im positiven Bereich liegen, was auf viele Fehlklassifikationen hindeutet und die bisherigen Resultate damit bestätigt.

Im Folgenden wird die zugehörige ROC-Kurve untersucht. Werden die ROC-Kurven zu der als gut befundenen Parameterkonstellation von $\gamma = 1$ und $C = 1$ sowie zwei alternativer Kombinationen berechnet, so ergibt sich die Grafik in Abbildung 3.32. Obwohl die Trefferquote der ersten (oberen) Kurve darauf schließen lässt, dass die bisher ausgewählte Kombination eine dominante SVM gegenüber den übrigen ist, kann dies durch Darstellung der ROC-Kurven nicht bestätigt werden. Es ist deutlich zu erkennen, dass die erste Kurve, für die Parameterkombinationen $C = 1$ mit $\gamma = 1$, und die zweite (mit $C = 1$ mit $\gamma = 0,1$) sich schneiden. Lediglich gegenüber der dritten Kurve ist die erste dominant. Die zugehörigen Zuordnungen zu den beiden Klassen sind für alle drei Kombinationen nochmal in Klassifikationsmatrizen (Tabelle 3.9) zusammengefasst.

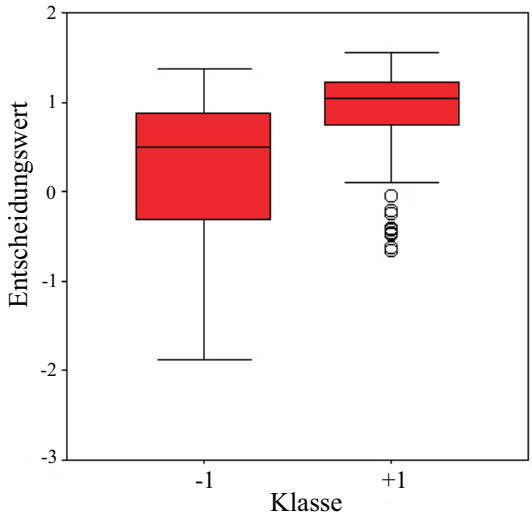

Abbildung 3.31: Boxplot der resultierenden Entscheidungswerte für beide Klassen

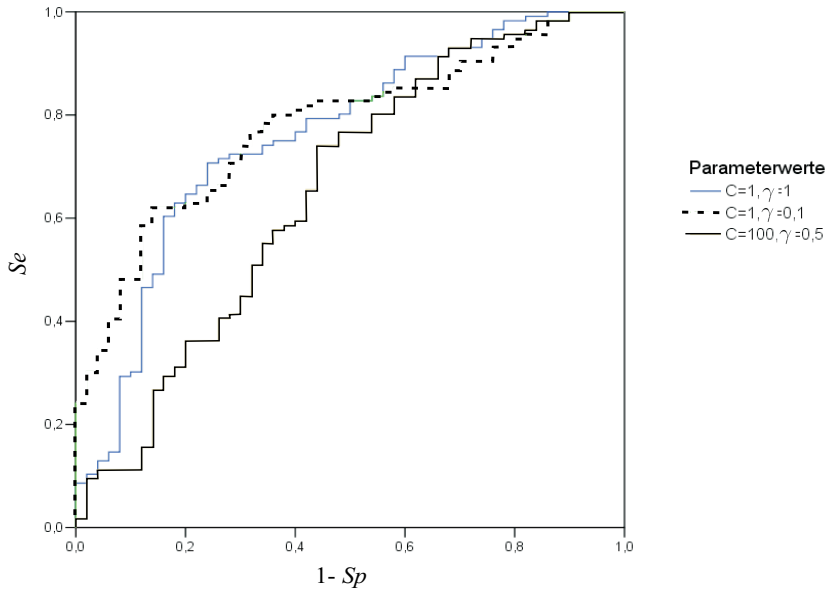

Abbildung 3.32: ROC-Kurven für ausgewählte Parameterkonstellationen

		Prognostizierte Klasse					
		$(C = 1,\ \gamma = 1)$		$(C = 1,\ \gamma = 0,1)$		$(C = 100,\ \gamma = 0,5)$	
		1	-1	1	-1	1	-1
Wahre	1	106	10	114	2	94	22
Klasse	-1	33	17	44	6	29	21

Tabelle 3.9: Klassifikationsmatrizen für unterschiedliche Parameterkonstellationen

Bemerkenswert ist die Dominanz aufgrund einer geringeren Fehlerzahl für Klasse „-1“ bei der dritten Parameterkombination ($C = 100,\ \gamma = 0,5$). Der Verlauf dieser Kurve ist daher auf die geringe Gesamttrefferquote zurückzuführen.
Als weiteres Maß ist der Wert der Fläche unterhalb der jeweiligen Kurve zu berechnen. Die zugehörigen Werte für AUC werden in Tabelle 3.10 angegeben, wobei zu Vergleichszwecken weiterhin die einzelnen Trefferquoten des Testdatensatzes bei Einsatz der unterschiedlichen Parameterkonstellationen angegeben sind.

Parameter		Wert für AUC	Trefferquoten		
C	γ		Gesamt	„+1“	„-1“
1	1	0,758	74,1%	91,4%	34,0%
1	0,1	0,778	72,3%	98,3%	12,0%
100	0,5	0,653	69,3%	81,0%	42,0%

Tabelle 3.10: Werte für die Fläche unter der ROC-Kurve (AUC) für unterschiedliche Parameter

Ein hoher Wert für AUC deutet auf eine gute Klassifikation hin. Es fällt auf, dass der Wert der zweiten Kombination am größten ist, obwohl die zugehörige Trefferquote nicht optimal ist. Der dritte Wert ist erwartungsgemäß recht klein, was durch die niedrigen Trefferquoten bestätigt wird. Hierbei sei allerdings angemerkt, dass es sich bei der dritten Parameterkonstellation um jene handelt, bei der die Trefferquoten der einzelnen Klassen zwar unterschiedlich sind, allerdings nicht so stark voneinander abweichen wie bei den anderen beiden Kombinationen.
Zusammenfassend kann geschlossen werden, dass aufgrund einer kleinen Anzahl an Gesamttreffern nicht davon ausgegangen werden kann, dass die zugehörigen ROC-Kurven und der Wert für AUC die übrigen dominiert. Ein Vergleich verschiedener Maße erscheint daher ratsam.

Die Trefferquoten zeigen zwar, dass die Anforderungen der üblicherweise verwendeten Kriterien durch die durchgeführte Klassifikation erfüllt werden, aber der Vergleich verschiedener Maße zur Beurteilung eines Klassifikationsergebnisses zeigt dennoch, dass die vorliegende Klassifikation lediglich als zufrieden stellend bezeichnet werden kann. Weiterhin sind die Ergebnisse nicht immer konsistent. Eine einseitige Betrachtung kann dazu führen, wichtige Informationen zu übersehen,

sodass der Einsatz mehrerer Maße als Ergänzung der Trefferquoten zur Beurteilung sinnvoll erscheint. Dies kann durch die mittels Boxplots visualisierte Verteilung der Entscheidungswerte oder durch die ROC-Kurven bestätigt werden.

Kapitel 4

Empirischer Einsatz von SVM

Im vorangegangenen Kapitel wurden Modifikationen und Ergänzungen des in Kapitel 2 vorgestellten Basisansatzes von SVM diskutiert und überwiegend anhand kleiner Beispiele erläutert. In diesem Kapitel werden nun die vorgestellten Ansätze auf reale Daten aus dem Bereich des Marketing angewendet. Es wird überprüft, inwieweit SVM eine sinnvolle Ergänzung der Menge der bisherigen Instrumente zur Klassifikation im Marketing bilden.
Dazu wird zunächst auf die allgemeine Vorgehensweise beim Einsatz von SVM eingegangen. Nach einem Literaturüberblick über die bisherigen Anwendungen von SVM in Verbindung mit speziell das Marketing betreffenden Fragestellungen erfolgt der Einsatz von SVM auf unterschiedlichen Datensätzen. Die Auswahl mehrerer, strukturell unterschiedlicher Datensätze liegt darin begründet, dass eine Anwendung der in Kapitel 3 behandelten Ansätze auf nur einem Datensatz nicht möglich wäre. So wird u.a. neben der Biklassifikation und der Multiklassifikation auch die Zuweisung einer Beobachtung zu mehreren Klassen (Multilabel-Klassifikation) sowie die jeweilige Interpretation der Ergebnisse betrachtet.

4.1 Allgemeine Vorgehensweise

Die aus den methodischen Grundlagen bereits hervorgehende grundsätzliche Vorgehensweise bei der Klassifikation mittels SVM wird hier noch einmal veranschaulicht, um die empirische Umsetzung zu unterstützen. Die Analysen in den folgenden Abschnitten unterscheiden sich durch den Einsatz optionaler Bereiche.
Der generelle Ablauf der Durchführung einer Klassifikation mittels SVM umfasst die Festlegung der Rahmenbedingungen, die Optimierung der SVM sowie die Klassifikation neuer, bisher unklassifizierter Beobachtungen. Um das Training einer SVM durchführen zu können, um also die Trennebene bestimmen zu können, werden bereits klassifizierte Trainingsdaten benötigt, die direkt in die Optimierung eingehen. Die trainierten SVM liefern die Basis zur Klassifikation der Testdaten, bzw. neuer, bisher nicht klassifizierter Daten. Die Elemente der einzelnen Schritte des Prozesses werden im Folgenden differenziert betrachtet.

Damit die Daten in die Klassifikation aufgenommen werden können, ist eine Zuweisung zu den zu untersuchenden Klassen notwendig. Diese Klassen können sich aufgrund des vorliegenden Untersuchungsgegenstandes ergeben. Bei der ABC-Kundenklassifikation entspricht diese Zuweisung der Betrachtung der bisherigen Kunden als A-, B- oder C- Kunden. Die Klassenzuweisung kann ebenfalls durch ein vorgeschaltetes Segmentierungsverfahren erstellt werden wie dies bereits in Abschnitt 3.1 diskutiert wurde. Daran anschließend wird eine Auswahl der Merkmale und Beobachtungen vorgenommen, die die Analyse bestimmen (Vorselektion). Die eingehenden Daten müssen vollständig sein, sie dürfen also keine fehlenden Werte aufweisen. Weiterhin muss festgelegt werden, um welche Art der Klassifikation es sich handelt: Bi-, Multi- oder Multilabel-Klassifikation. Weitere Rahmenbedingungen werden durch die Vorgabe einer unterschiedlichen Gewichtung der Klassen und der Merkmale gegeben. Liegen die benötigte Daten vor, so gilt es, eine möglichst gute Parameterkombination zu finden und die Auswahl der Merkmale auf die wichtigen zu beschränken. Um eine gute Parameterkombination zu finden, wird ein zweistufiges Gridsearch (vgl. Abschnitt 3.3) durchgeführt. Zunächst wird durch ein grobes Gitter die Region von Parameterkonstellationen bestimmt, in der gute Trefferquoten für den vorliegenden Datensatz zu erwarten sind. Dabei wird zur Gewährleistung der Generalisierungsfähigkeit auf die Kreuzvalidierung oder die LOO-Methode zurückgegriffen. Innerhalb des Erfolg versprechenden Bereiches wird die Suche zum feinen Gridsearch verfeinert, sodass dann eine Kombination von Parametern identifiziert wird, die zu möglichst guten Ergebnissen führt. Mittels dieser Kombination erfolgt das endgültige Training einer SVM. Die eingehenden Trainingsdaten sollten den gleichen Umfang wie die Daten haben, die eventuell bei der Kreuzvalidierung in das Training eingehen, um Verzerrungen zu vermeiden. Auf Grundlage der berechneten Lagrange-Multiplikatoren (vgl. Kapitel 2.1) erfolgt als letzter Schritt eine Zuordnung von neuen, bisher unklassifizierten Beobachtungen.
Wird innerhalb der Analyse durch verschiedene Verfahren deutlich, dass bestimmte Merkmale für die Trennung der vorgegebenen Klassen nicht von Bedeutung sind (vgl. Abschnitt 3.5), so können diese entweder von der Analyse ausgeschlossen werden (Merkmalsextraktion) oder entsprechend ihres Einflusses gewichtet werden (vgl. Abschnitt 3.2.1). Diese Gewichtung erfolgt analog zur Klassengewichtung und kann bereits innerhalb des Gridsearch oder nach erfolgreicher Durchführung des Trainings einer SVM ansetzen. Da sich die Daten bei einer Merkmalsextraktion und einer Gewichtung verändern, kann im letzteren Fall ein erneutes feines Gridsearch für die Bestimmung der guten Parameterkonstellation hilfreich sein. Das Gleiche gilt insbesondere für die Klassen.

Die eigentliche Anwendung von SVM erfolgt nach der Optimierungsphase auf Basis von neuen, nicht klassifizierten Daten, die z.B. durch etwaige Neukunden gebildet werden können. Die Interpretation der Ergebnisse schließt an die Klassifikation von Testdaten an. Sind neue Beobachtungen mittels SVM klassifiziert worden, so kann zusätzlich eine Beurteilung mittels der Entscheidungswerte F bzw. D (vgl. Abschnitt 3.7.2 und 3.7.4 bzw. 3.7.3) vorgenommen werden. Liegen mehrere Klassen

vor, so ist neben einer einfachen Klassifikation eine Zuweisung zu einer Menge von Klassen möglich, wie dies in Abschnitt 3.6 beschrieben wird. Weiterhin sollten die Ergebnisse angemessen (z.B. durch Parallelkoordinaten) visualisiert werden. Sind auch bei diesen Daten die wahren Klassenzugehörigkeiten bekannt, so können sie zur Erweiterung der bisherigen Trainingsdaten dienen. Da sich diese Daten wiederum von den ursprünglichen Trainingsdaten um die neu hinzugekommenen Beobachtungen unterscheiden, sollte ein erneutes feines Gridsearch zur Bestimmung der optimalen Parameter durchgeführt werden.

Die Auswertung der Daten innerhalb dieses Kapitels erfolgt auf Basis von LIBSVM, Version 2.36 (*Chang, Lin* (2001)), welches ein Software-Paket zur Klassifikation und Regression mittels SVM ist, das von *Chih-Jing Lin* und seinen Kollegen entwickelt wurde und mittlerweile sehr häufig in mehreren Veröffentlichungen eingesetzt wird. Da es sich als leistungsstarkes Tool herausgestellt hat (vgl. u.a. *Meyer et al.* (2003)), soll es auch innerhalb dieser Arbeit eingesetzt werden. Es ermöglicht bei der Multiklassifikation die Klassifikation auf Basis von OAO (vgl. Abschnitt 2.2.2), die der OAA-Methode vorgezogen wird, da sie die Trennung auf Basis gleich großer (gleichwertiger) Gruppen vollzieht. Gerade bei einer großen Anzahl von Gruppen macht sich die Größe der Klassen bemerkbar und kann durch den Einsatz von OAO vermieden werden. Diese Software wurde für die Zwecke dieser Arbeit derart erweitert, sodass mehrere Verfahren der Multiklassifikation durchgeführt werden können und eine individuelle Gewichtung der Beobachtungen und der Merkmale sowie die Modifikationen und Interpretationen der Ergebnisse ermöglicht werden. Es wurden ferner Multilabel-Klassifikation und die in Abschnitt 3.5 vorgestellten Ansätze zur Merkmalsreduktion umgesetzt. Weiterhin wurde die Software durch eine grafische Ausgabe der Ergebnisse ergänzt.

4.2 Einsatzbereiche im Marketing - ein Literaturüberblick

In den in jüngster Zeit zunehmenden Anwendungen von SVM finden sich auch Bereiche des Marketings wieder, bei denen das Verfahren direkt oder indirekt zum Einsatz kommt. Die folgenden Ansätze verwenden SVM unter unterschiedlichen Zielsetzungen, sodass die Vorteile sehr unterschiedlich genutzt werden können.
In *Curry, Cui* (2005) wird ein Überblick über SVM als mögliches Verfahren zur Prognose im Marketing gegeben. Die Anwendungen u.a. anhand eines preispolitischen Beispiels resultieren in deutlich besseren Trefferquoten als die dazu alternativ gewählte (lineare und quadratische) Diskriminanzanalyse. Basierend auf den Ergebnissen auf zwei-dimensionalen Daten kommen die Autoren zu dem Schluss, dass SVM in die Menge der modernen Verfahren zur Datenanalyse im Marketing aufgenommen werden sollten.
Das bereits mehrfach erwähnte „Stammanwendungsgebiet“ von Klassifikationsmethoden im Marketing bildet die Kundenklassifikation. In *Decker, Monien* (2003)

werden dazu bei der Identifikation potentieller Neukunden bessere Ergebnisse erzielt als mit der herkömmlichen Methode zur Klassifikation im Marketing, der Diskriminanzanalyse. Die Resultate werden allerdings auf Basis einer umfangreichen Parametersuche erzielt, wohingegen bei der alternativ einzusetzenden Diskriminanzanalyse nur wenige bzw. keine Einstellungen vorzunehmen sind.
Neben der Kundenklassifikation bilden die Recommender-Systeme geradezu einen klassischen Bereich, in dem Klassifikation sehr gut angewendet werden kann und damit auch SVM zunehmend an Bedeutung gewinnen werden. Ziel ist die Generierung von Empfehlungen auf Basis historischer Daten. Man unterscheidet hauptsächlich zwischen inhaltsbasierten und kollaborativen Filterverfahren in Abhängigkeit der Datenbasis, die als Grundlage für die Entwicklung von Empfehlungen eingeht. Für den Bereich des E-Commerce ist bei *Schafer et al.* (2001) ein kurzer Überblick zu finden. Mittlerweile gibt es einige Veröffentlichungen, die sich mit diesem Gebiet in Kombination mit SVM auseinandersetzen. So verwenden *Zhang, Iyengar* (2002) SVM dazu, um unterschiedliche Websites zu empfehlen. Dabei erzielen sie mit ihrem leicht modifizierten Algorithmus sehr gute Klassifikationsresultate, obwohl sie nur auf die lineare Variante von SVM zurückgreifen. Bei *Cheung et al.* (2003) werden bei der Entwicklung eines inhaltsbasierten Recommender-Systems die Entscheidungswerte dazu genutzt, Empfehlungen über Produkte abzugeben. Bei der Auswertung individueller Produktratings werden im Vergleich zu anderen, traditionellen Verfahren signifikant bessere Klassifikationsresultate erzielt. In *Bomhardt* (2004) wird ebenfalls ein Recommender-System entwickelt, bei dem SVM aufgrund ihrer Leistungsstärke im Vergleich zu alternativen Klassifikationsmethoden als Prognoseinstrument eingesetzt wird. Insbesondere durch die Möglichkeit der Multilabel-Klassifikation eignen sich SVM für den Einsatz innerhalb eines Recommender-Systems. Dadurch werden einem Benutzer mehrere Empfehlungen auf Basis seines bisherigen Kauf- und Surfverhaltens gegeben. Unterschiedliche Themengebiete könnten die Klassen bilden, zu denen ein Kunde eine mehr oder weniger stark ausgeprägte Affinität aufweist. Auf Basis von SVM könnte durch die Zuordnung zu mehreren Themenbereichen ein ganzes Produktbündel statt einzelner Produkte offeriert werden. Das Gebiet der Recommender-Systeme bildet im Rahmen des Marketings einen viel versprechenden Bereich. In *Huang et al.* (2005) werden SVM dazu genutzt, Kaufempfehlungen für die so genannten „cold seller“ abzugeben. So werden diejenigen Produkte bezeichnet, mit denen ein Unternehmen den wenigsten Umsatz im Vergleich zu der die 80/20 Regel erfüllenden Produkte macht. Statt der ungleichen Klassengewichte wird hier die Klasse der Käufer der cold seller künstlich vergrößert. Als Bewertungsmaß kommt bei dieser Anwendung AUC zum Einsatz.
Weitere Anwendungsmöglichkeiten von SVM im Bereich eines Marketingkontextes sind in *Bennett et al.* (1998) oder *Orsenigo, Vercellis* (2003) zu finden, wobei jeweils einander ähnliche Ansätze verwendet werden. *Bennett et al.* (1998) kombinieren SVM mit Entscheidungsbäumen, um so ein leistungsstarkes Instrument zur besseren Ansprache der Kunden innerhalb des Database Marketings zu entwickeln. Ziel war es, durch eine Reduzierung von Merkmalen die Struktur des resultierenden Entscheidungsbaumes im Vergleich zu herkömmlichen Bäumen zu vereinfachen und

gleichzeitig einem Overfitting entgegen zu wirken. Hierbei wird insbesondere die Eigenschaft der Dimensionsreduktion bei SVM ausgenutzt. Beim Einsatz von SVM in *Orsenigo, Vercellis* (2003), bei dem ebenfalls eine Kombination mit Entscheidungsbäumen zum Einsatz kommt, wird der Marketingkontext allerdings nur am Rande erwähnt und steht nicht im Mittelpunkt der Betrachtung. Die Autoren waren bestrebt, durch den Einsatz von SVM ein möglichst gutes Klassifikationsergebnis zu erzielen.
In *Crone et al.* (2004) nehmen die Autoren einen Vergleich von SVM zu den alternativen Verfahren der Vektor Quantisierung und MLP (Multi-Layer Perceptron) vor, um deren Performancestärke im Rahmen des Customer Relationship Managements gegenüber zu stellen. Anhand der durch ROC-Kurven vorgenommenen Bewertungen kommen die Autoren zu dem Schluss, dass sich SVM aufgrund der Ergebnisse gut als Klassifikationsinstrument bei betriebswirtschaftlichen Fragestellungen eignen.
Cui, Curry (2005) setzen SVM zur Prognose von Kaufentscheidungen der Konsumenten ein. Der Vergleich zum Multinomial-Logit-Modell zeigt, dass SVM einen viel versprechenden Ansatz in der Prognose im Marketing, z.B. im Rahmen des Data Mining, bilden, der die klassischen Methoden gut ergänzen kann. Einen ähnlichen Ansatz verfolgen *Evgeniou et al.* (2005). Sie nutzen de Idee der SVM-Methodik aus, um Nutzenfunktionen bei der Wahl von zwei Alternativen zu spezifizieren. Ihr Ansatz schneidet bei den Analysen besser bzw. äquivalent zu alternativen Verfahren (logistische Regression oder hierarchische Bayes Analyse) ab. Im Gegensatz zu traditionellen Methoden können durch SVM hochdimensionale Daten (in diesem Fall große Datensätze für Produkte mit einer großen Anzahl an Attributen) effizient verarbeitet werden. In den beiden folgenden Arbeiten treten SVM als Hilfsmittel auf. So verwenden *Viaene et al.* (2001) die Methodik zur iterativen Aufdeckung von Merkmalen, die für die Modellierung des Wiederholkaufverhaltens im Direktmarketing relevant sind. Bei der Merkmalsauswahl spielen die guten Klassifikationsergebnisse der SVM eine entscheidende Rolle. *Yang* (2002) setzt sich mit der kundenindividuellen Planung von Marketingstrategien auseinander, wobei die Vorgehensweise neben SVM auch auf das Prinzip des fallbasierten Schließens zurückgreift. SVM werden hierbei zusätzlich eingesetzt, um die Eigenschaft der Reduzierung der Eingabevektoren auf die Support Vektoren auszunutzen, die dann letztendlich als Fälle in das fallbasierte Schließen eingehen. Diese bilden eine Alternative zu den ebenfalls bestimmten Clusterzentroiden, die als typische Fälle verwendet werden. Somit dienen SVM nur als Mittel zum Zweck, wobei allerdings die essentielle Eigenschaft dieser Methodik ausgenutzt wird.
Fast allen diesen Veröffentlichungen ist gemeinsam, dass sie nicht nur die Trefferquote als Ergebnisbewertung zugrunde legen, sondern häufig ROC-Kurven oder ähnliches zur Bewertung der Ergebnisse heranziehen.

Nach dem kurzen Überblick über die in der Literatur vorliegenden Anwendungen von SVM in einem mehr oder weniger weit gefassten Marketingkontext wird in den folgenden Abschnitten des Kapitels eine umfassende empirische Betrachtung von SVM als Analyseinstrument im Marketing vorgenommen. Dabei sollen die zuvor in Kapitel 3 vorgestellten Eigenschaften und vorgenommenen Erweiterun-

gen umgesetzt werden. Dadurch werden Vor- und Nachteile einer Vielzahl von Eigenschaften von SVM beim Einsatz des Instruments im Bereich des Marketings herausgestellt.

4.3 Anwendung von SVM im Vertrieb

Der Außendienst als ein Instrument des Marketings zur Kundenbetreuung steht bei den folgenden Betrachtungen im Vordergrund, in dessen Rahmen SVM direkt oder indirekt zur Klassifikation eingesetzt werden können.
Die Möglichkeit zur Beeinflussung von Kunden durch persönliche Gespräche soll genutzt werden, um Kunden an das eigene Unternehmen zu binden. Da der Außendienst einen Großteil des Marketingbudgets verschlingt, erscheint es umso wichtiger, dass die entsprechenden Mitarbeiter an den richtigen Stellen eingesetzt werden. Eine pauschale Verteilung der Vertriebsressourcen auf den gesamten Kundenstamm ist in den meisten Fällen unökonomisch. Daher erscheint der Einsatz von Klassifikationsinstrumenten zur effektiven Aufteilung der Vertriebsressourcen sinnvoll. Die Kundenbasis eines pharmazeutischen Unternehmens wird so strukturiert, dass unterschiedliche Gruppen entstehen, die in Abhängigkeit ihrer Relevanz verschieden intensiv durch Mitarbeiter des Unternehmens betreut werden.
Ziel der Untersuchung ist die Klassifikation der Kunden pharmazeutischer Unternehmen, bei denen Kundenbesuche ein wesentliches Element bzw. Erfolgsfaktor in der Marktbearbeitung und Kundenbindung darstellen (*Baier et al.* (2004)). Den Großteil der Kunden eines solchen Unternehmens bilden Apotheken und niedergelassene Ärzte. Zu den Aufgaben des Außendienstes im Pharmamarketing gehört die Informationsbeschaffung über die Nachfrage, um das Angebot zu verbessern. Weiterhin ist die Informationsvermittlung zur Steuerung des Angebots der einzelnen Apotheken ein wichtiger Aspekt. Die dritte Aufgabe des Außendienstes ist die Kontaktpflege, um als Anbieter bestimmter Arzneimittel präsent zu sein, sodass diese öfter verschrieben und verkauft werden (*Gehrig* (1992)). Bei der Pflege der Kundenbeziehungen sollten demnach alle diese Punkte berücksichtigt werden. Da die zur Verfügung stehende Zeit der Außendienstmitarbeiter und die finanziellen Mittel beschränkt sind, können nicht alle Kunden gleich häufig besucht werden, sodass eine Auswahl getroffen werden muss. Als eine weitere Möglichkeit der Kontaktpflege mit den Kunden bietet sich das in jüngster Zeit häufig diskutierte Mittel des eDetailing an (*Heutschi et al.* (2003)). Wie bereits in Abschnitt 1.1 erwähnt, handelt es sich um die Möglichkeit, durch Einsatz internetbasierter Kommunikationsmethoden den traditionellen Außendiensteinsatz zu erweitern. Da diese durch den zusätzlichen Arbeitsaufwand eine kostenintensive Ergänzung darstellt, sollten nur diejenigen Kunden (Apotheken oder Ärzte) in den Genuss dieser Art der Kundenbetreuung kommen, bei denen dieser Einsatz auch gerechtfertigt ist.
Bei der Allokation der Vertriebsressourcen stellt sich also das Problem der Aufteilung der Ressourcen auf die Kunden (*Albers* (2002)). Es soll verhindert werden, dass zu viel Zeit und Geld in umsatzschwache Kunden investiert wird. Die Methodik der SVM kann hier einen wesentlichen Beitrag zur Lösung beitragen, bei dem zusätzliche

Optionen, wie eDetailing oder nicht persönliche Kommunikation durch Versendung von Prospektmaterial, ebenfalls berücksichtigt werden können.
Im Folgenden werden zwei Beispiele der Kundenklassifikation in der Pharmabranche behandelt. Ziel ist die Erhöhung des Absatzes bestimmter Produkte durch gezielte Verteilung der Vertriebsressourcen auf lohnende Kunden. Zunächst liegt in Abschnitt 4.3.1 der Fokus auf der Klassifikation von Apotheken als Kunden eines Pharmaunternehmens und die Gestaltung der Kundenbetreuung bei der Vermarktung von OTC-Produkten. In Abschnitt 4.3.2 wird auf die Anwendung hinsichtlich der Erhöhung der Verordnungen eines verschreibungspflichtigen Medikaments durch niedergelassene Ärzte eingegangen. Hierbei werden unterschiedliche Merkmale verwendet, die das Klientel des Einzugsgebietes eines Arztes berücksichtigen. In beiden Fällen werden die Vorteile, die SVM im Rahmen der Klassifikation zur Verbesserung der Kundenbetreuung liefern, herausgestellt und angewendet.

4.3.1 Klassifikation von Apotheken

Problemstellung und Datenbeschreibung

Grundlage der folgenden Ausführungen sind die Daten eines Unternehmens aus der pharmazeutischen Industrie, welches den Umsatz bzw. Abverkauf eines seiner nicht verschreibungspflichtigen, aber apothekenpflichtigen Medikamente steigern möchte. Dieses Medikament gehört somit zu den so genannten Over-the-Counter (OTC) Produkten. Der OTC-Markt bildet einen immer wichtiger werdenden Bereich, der 2004 ein Umsatzwachstum von 10% im Vergleich zu 2003 verzeichnete[1]. Solange der Kunde keine spezielle Vorliebe für einen Hersteller hat, werden die Kaufentscheidungen bezüglich der OTC-Produkte häufig vom Apotheker selbst getroffen, da dieser dem Kunden meist ein oder zwei Produkte zur Wahl anbietet und somit die Menge der möglichen Hersteller von vornherein auf wenige einschränkt. Dies liegt unter anderem darin begründet, dass nicht alle Hersteller in der Apotheke gelistet sind. Weiterhin bezieht laut *Dialego* (2004) der überwiegende Teil der Kunden Informationen zur Selbstmedikation von Mitarbeitern einer Apotheke. Dies unterstreicht die wichtige Rolle der Apotheker beim Verkauf von OTC-Produkten. Daher ist eine Erhöhung der Kundenbindung und damit die Sicherung der Kundentreue von Seiten der Apotheken ein wichtiges Ziel im Rahmen des Apothekenmarketings von Pharmaunternehmen zur Erhöhung des Umsatzes der Generika und der OTC-Produkte.
Zur gezielten Steigerung des Abverkaufs und Erhöhung des Umsatzes stehen mehrere Möglichkeiten der Kundenbetreuung zur Verfügung. Die kostensparsamste liegt sicherlich in der Versendung geeigneten Prospektmaterials über die zu vermarktenden Produkte. Wird zusätzlich die persönliche Kommunikation hinzugezogen, so kann eine Betreuung der Apotheken auch in telefonischen Kontakten bestehen, um die Eigenschaften von neuen Produkten zu besprechen. Die traditionelle und am weitesten verbreitete Form liegt allerdings in dem Einsatz eines Außendienstes als

[1] Quelle: Institut für medizinische Statistik, Frankfurt a.M., www.imshealth.de, Zugriff: 15.3.05

strategisches Instrument der Marktbearbeitung. Durch den direkten persönlichen Kontakt können Außendienstmitarbeiter eine enge Beziehung zu den Apothekern aufbauen und so auf die Präparatenauswahl in gewissem Maße Einfluss nehmen. Somit bildet die Steuerung des Außendienstes ein zentrales Element beim Vertrieb von OTC-Produkten. eDetailing bildet durch die Nutzung von Informationstechnologien zur Verkaufsförderung eine moderne und viel diskutierte zusätzliche Erweiterung dieses traditionellen Ansatzes (*Heutschi, Alt* (2003)). Diese Kommunikationsform umfasst die Mensch-zu-Mensch Interaktionen (*Heutschi et al.* (2003)), die sich auf die Verwendung des Internets durch Video detailing, Chats, Foren oder ähnlichem bezieht. Neue Technologien und ein erhöhter Kostendruck haben bereits dazu geführt, dass verstärkt Elemente des eDetailing zur persönlichen Produktpromotion bei der Betreuung von Ärzten eingesetzt werden. Ärzte bilden das zentrale Element bei der Vermarktung von verschreibungspflichtigen Arzneimitteln und sind daher die größere Zielgruppe. Aber auch bei der Betreuung von Apotheken könnte eDetailing bei der Vermarktung von OTC-Produkten eingesetzt werden. Das System dient mit einem flexiblen Einsatz vor Ort hauptsächlich der Informationsvermittlung bzgl. neuer Produkte für die Apotheker und deren Mitarbeiter. Dabei wird eDetailing den traditionellen Außendienst allerdings nicht ersetzen (*Baier et al.* (2004)), sondern ist lediglich als Ergänzung zu sehen.

Das Ziel dieses Abschnitts liegt darin, auf Basis der bereits bestehenden Kunden des pharmazeutischen Unternehmens ein Modell zu generieren, welches die Unterscheidung in für das Unternehmen wichtige und unwichtige Apotheken zulässt. Die hier verwendete Vorgehensweise kann dabei ebenfalls auf den Markt der Ärzte angewendet werden. Denkbar ist hier eine Situation, in der das klassifizierende Pharmaunternehmen seinen bisherigen Kundenstamm erweitern möchte (vgl. ähnliche Ausführungen in *Monien, Decker* (2004)). Da sowohl die finanziellen Mittel als auch die Kapazitäten des Außendienstes nicht ausreichen, um bei einer sehr umfangreichen Vergrößerung des Kundenstammes jede Apotheke zu besuchen, soll eine Klassifikation der potentiellen Neukunden erfolgen. Diese ermöglicht die Differenzierung der Art der Neukundenansprache. Durch die Zuweisung einer der oben genannten Kundenbetreuungsmaßnahmen kann somit eine der zu erwartenden Wichtigkeit der jeweiligen Apotheke entsprechende Kundenbetreuung realisiert werden. Bei den weniger wichtigen Apotheken erscheint die Versendung von Informationsmaterial in der Anfangsphase der Kundenbeziehung ausreichend, wohingegen sich die Maßnahmen der Kundenbetreuung von interessanten Kunden z.B. durch die Ausstattung mit der entsprechenden eDetailing-Infrastruktur deutlich aufwändiger gestalten.
Der Einsatz von SVM als Klassifikationsinstrument ermöglicht eine Identifikation der umsatzstarken bzw. umsatzschwachen Apotheken. Wie zu sehen sein wird kann allerdings auch eine Zuweisung der unterschiedlichen Arten der Kundenbetreuungen vorgenommen werden. Die beschreibenden Merkmale müssen so ausgewählt werden, dass sie auch für bisher nicht zum Kundenstamm gehörige Apotheken zu beschaffen sind und somit nicht auf Daten, die die bisherigen Kundenbeziehungen betreffen, zurückgreifen.

Die Datenbasis zum Training der SVM umfassen demnach Apotheken, die bereits zum Kundenstamm des Pharmaunternehmens gehören, für die also die Merkmalsausprägungen ausgewählter, für das Klassifikationsziel wichtiger Merkmale zur Verfügung stehen. Diese Merkmale umfassen Eigenschaften von Apotheken, die entweder in Branchenverzeichnissen vermerkt sind oder durch einen Außendienstmitarbeiter erhoben werden müssen. Dies sind die Anzahl der Mitarbeiter sowie die Größe der Verkaufsfläche der Apotheke. Eine Apotheke wird dann als wertvoll für das Unternehmen eingestuft, wenn sie einen hohen Umsatz mit einem führenden OTC-Präparat (OTC-Umsatz) aufweist. Dies kann als OTC-Potenzial angesehen werden und geht als Merkmal in die Klassifikation ein. Neben dieser speziellen Information über ein OTC-Produkt ist die Auskunft über den generellen Anteil von OTC-Präparaten am gesamten Sortiment (OTC-Anteil) von Interesse. Für die Prognose der Wichtigkeit eines Kunden für ein Unternehmen spielt die Lage der Apotheke eine entscheidende Rolle. Diejenigen Apotheken, die zentral in der Stadt gelegen und mit öffentlichen Verkehrsmitteln gut zu erreichen sind, haben aufgrund des zu erwartenden höheren Umsatzes mit OTC-Produkten in der vorliegenden Anwendung höhere Relevanz als Apotheken, die eher in ländlicheren Regionen situiert sind, wo noch eine stärkere Orientierung an den Empfehlungen des Hausarztes vorherrschen kann. Daher wird die Variable „Lage“ ebenfalls mit aufgenommen. Des Weiteren spielt die wirtschaftliche Situation des Einzugsgebietes der Apotheke zur Beurteilung der Wertigkeit eine Rolle. Dies wird in der Kaufkraft (Merkmal 6) und in der Kaufkraft pro Einwohner des Postleitzahlgebietes (Merkmal 7) ausgedrückt.
In Tabelle 4.1 werden die in den folgenden Analysen verwendeten Merkmale noch einmal aufgelistet. Jede zu bewertende Apotheke wird somit durch sieben Merkmale beschrieben. Alle Informationen können bei entsprechenden Marktforschungsinstituten erworben werden und stehen daher auch für neue Apotheken zur Verfügung. Die ausgewählten Merkmale sind zum einen dadurch gekennzeichnet, dass sie wertvolle Informationen über die betreffenden Apotheken liefern und zum anderen sicherstellen, dass sie ohne viel Aufwand auch für das Pharmaunternehmen neue bzw. bezüglich des betreffenden Produktes bisher noch nicht besuchte Apotheken erhoben werden können. Da bei den ersten vier Merkmalen bei neu einzustufenden Apotheken lediglich Schätzungen vorliegen, werden die Ausprägungen kategorisiert und diese danach metrisch interpretiert. Die übrigen Merkmale liegen in metrischer Form vor und können so direkt in die Optimierung eingehen.
Ziel der Untersuchung soll es sein, die für das Unternehmen interessanten Apotheken auf Basis der obigen Merkmale zu identifizieren. Dazu müssen SVM auf bereits vorliegenden Daten trainiert werden, die schon in die zu differenzierenden Klassen eingeteilt sind. Diese sind in diesem Fall die interessanten Kunden, die verstärkt kontaktiert werden sollten, und die übrigen. Die interessanten Kunden zeichnen sich hier durch einen hohen Umsatz mit Produkten im laufenden Jahr aus, die von dem Pharmaunternehmen angeboten werden. Je größer der Umsatz ausfällt, desto wertvoller und interessanter sind die betreffenden Apotheken für das Unternehmen. Bei einem gewissen Mindestumsatz im laufenden Geschäftsjahr werden die Kunden als interessant eingestuft. Diese Kunden gehören zu der Gruppe von Apotheken,

Nr.	Merkmal	Wertebereich
1	Anzahl Mitarbeiter	1: 1-2
		2: 3-6
		3: >6
2	Verkaufsfläche	1: < 30 m^2
		2: 31 - 60 m^2
		3: 61 - 100 m^2
		4: > 100 m^2
3	OTC-Anteil	1: $> 40\%$
		2: $20 - 40\%$
		3: $< 20\%$
4	Lage	1: sehr schlecht
		2: schlecht
		3: gut
		4: sehr gut
5	OTC-Umsatz	$\in \mathbb{R}$
6	Kaufkraft	$\in \mathbb{R}$
7	Kaufkraft pro Einwohner	$\in \mathbb{R}$

Tabelle 4.1: Merkmalsbeschreibung

die verstärkt besucht werden sollten und werden demnach in die Klasse „+1“ eingeordnet. Die übrigen Apotheken gehören der Klasse „-1“ an. Die vorliegenden 1474 Beobachtungen weisen einen Anteil von etwa 68% an positiv klassifizierten Apotheken auf.
Das Ziel zur Umsetzung der Kundenstammerweiterung liegt neben der Bestimmung der Klassenzugehörigkeit in der Bestimmung der Wichtigkeit der für das Unternehmen neuen Apotheken und eine Erstellung einer Rangfolge, um eine möglichst effektive Allokation des Vertriebsbudgets durch eine Differenzierung in der Neukundenkontaktierung zu erreichen. Daneben soll eine adäquate Ansprache realisiert werden, um keine potentiellen Kunden durch eine falsche Betreuung zu verlieren. Demnach lautet die in diesem Abschnitt zu untersuchende Fragestellung

Welche Art der Kundenbetreuung ist für welche Apotheke adäquat?

Die bereits beschriebenen Merkmale weisen sehr unterschiedliche Ausprägungen auf, wie etwa an den Werten für die Verkaufsfläche ($\in \{1, 2, 3, 4\}$) und dem Umsatz mit OTC-Produkten ($\in [1.762,64; 63.974,51]$) erkennbar ist. Da eine Verzerrung durch ungleiche Ausprägungen vermieden werden soll, werden die Daten pro Merkmal auf das Intervall $[0, 1]$ normiert, um den Einfluss der Merkmale auf die Trennung der Klassen zu vereinheitlichen.

Auswertung

Um die Bereiche festzulegen, in denen bei einer späteren Klassifikation gute Ergebnisse zu erwarten sind, wird ein Gridsearch bei Wahl des Radialbasis-Kerns durchgeführt. Nach einer groben Einteilung werden die Bereiche mit den höchsten Trefferquoten verfeinert. Die Ergebnisse sind in Abbildung 4.1 dargestellt.

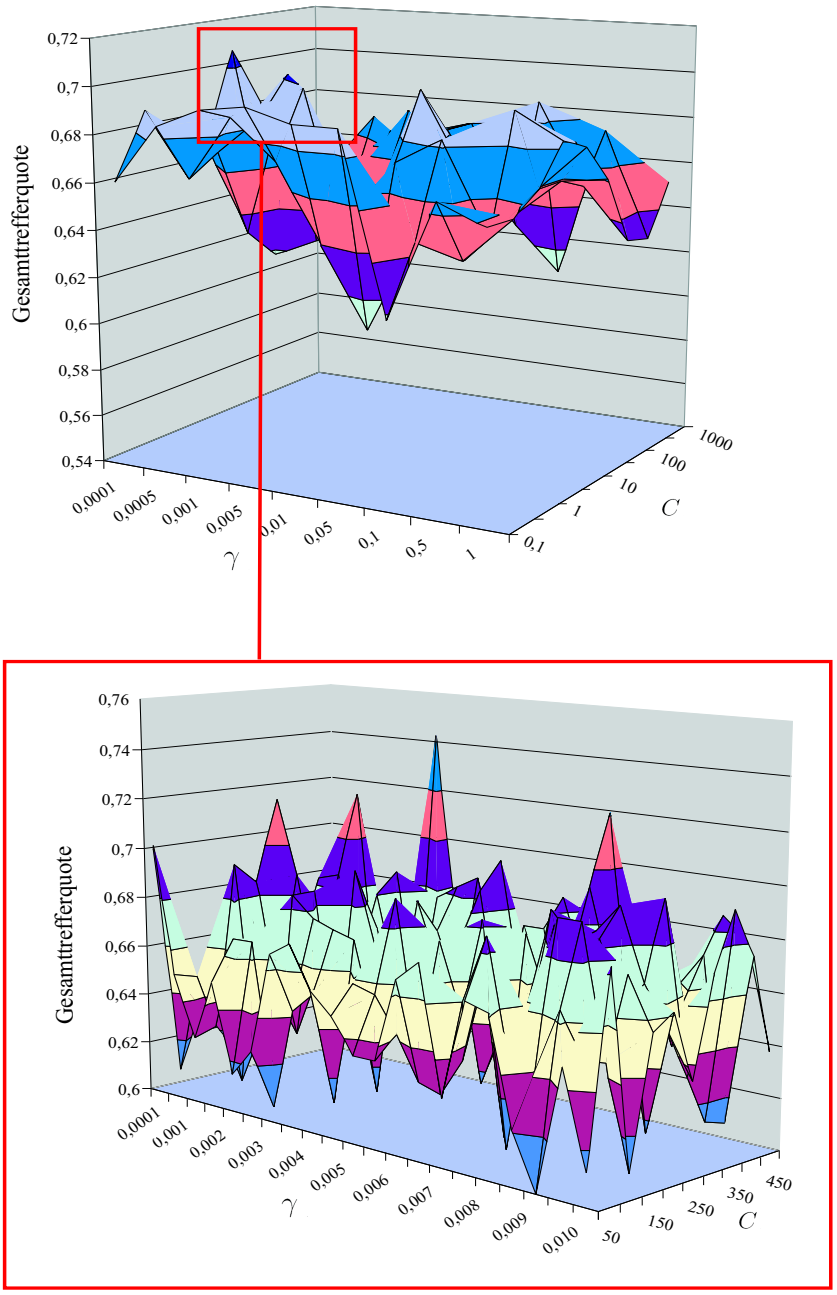

Abbildung 4.1: Darstellung der Gesamttrefferquote bei grobem und feinem Gridsearch

Im oberen Teil der Abbildung ist zu erkennen, dass geringe Werte für den Kostenparameter C und hohe Werte für den Kernparameter γ zu einer schlechten Trefferquote führen. Daher wird für die feinere Suche nach den optimalen Parameterkombinationen der markierte Bereich ausgewählt. Bei Wahl von γ aus dem Intervall [0,0001;0,01] bei gleichzeitiger Wahl von mittleren Werten für die Kosten ($C \in [50; 500]$) lässt die Klassifikation gute Ergebnisse erwarten. Es sind keine eindeutigen Tendenzen innerhalb des feinen Gridsearch zu erkennen, sodass mehrere Parameterkombinationen zu guten Ergebnissen führen können. Insgesamt lassen sich bei gleich gewichteten Klassen Gesamttrefferquoten von bis zu 72% erwarten.

Die Daten werden zufällig im Verhältnis 3 zu 1 in Trainings- und Testdaten aufgeteilt, sodass die Testdaten 369 Beobachtungen umfassen. In Tabelle 4.2 sind die Trefferquoten für die beiden Klassen der fest gewählten Testdaten sowie die Gesamttrefferquote separat aufgelistet. Die Gesamttrefferquote wird durch den Anteil der unabhängig von der wahren Klassenzugehörigkeit richtig zugeordneten Beobachtungen an den gesamten Testdaten berechnet.

Verfahren	Parameter					Trefferquoten		
	γ	d	C	c_+	c_-	Klasse +1	Klasse −1	Gesamt
SVM linear	-	-	500	1,2	1,9	74,00%	68,91%	72,36%
SVM RBF	0,0025	-	500	1,2	1,9	74,40%	68,07%	72,36%
SVM Polynom	-	2	500	1	1,9	71,6%	68,07%	70,46%
LDA	-	-	-	-	-	63,60%	78,15%	68,29%
C4.5	-	-	-	-	-	84,80%	37,82%	69,65%
MLP	-	-	-	-	-	86,40%	36,97%	70,46%

Tabelle 4.2: Trefferquoten einzelner Verfahren im Vergleich auf Testdaten

Für den Parameter γ wird hier ein Wert gewählt, der sich bei der Kreuzvalidierung als gut herausgestellt hat. Die Gewichtung der Klassen durch c_+ und c_- erfolgte zunächst in Abhängigkeit der Klassengrößen und wurde so angepasst, dass sowohl die Trefferquote für Klasse „+1“ und die Gesamttrefferquote hoch ist, als auch das Ergebnis für Beobachtungen aus Klasse „-1“ akzeptabel bleibt. Zusätzlich zum Radialbasis-Kern wurde der lineare Kern sowie ein polynomieller Kern zweiten Grades gewählt. Da sich die Güte eines Klassifikators an den Ergebnissen der Testdaten ermitteln lässt, sind hier die Ergebnisse der Trainingsdaten nicht aufgeführt. Diese weisen allerdings ähnliche Werte auf, sodass anhand der Ergebnisse die Anpassung des Modells an die Daten als gut bezeichnet werden kann.
Um die Güte der Klassifikation bewerten zu können, wurden alternative Verfahren ausgewählt, die sich in einer Vergleichsstudie als sehr leistungsstark herausgestellt haben (*Lim et al.* (2000)). Hier zeigte sich, dass die lineare Diskriminanzanalyse u.a. die besten Ergebnisse erzielte und der Entscheidungsbaum C4.5 (*Quinlan* (1993)) sich als besonders schnell erwies.[2] Der Vergleich von SVM mit diesen Verfahren

[2]Das neuronale Netz MLP war in dieser Studie nicht enthalten, wird hier aber dennoch als ein alternatives Verfahren zur Klassifikation zu Vergleichszwecken aufgenommen.

zeigt, dass mittels Radialbasis-Kern die besten Ergebnisse erreicht werden, wenn nur die Gesamttrefferquote beurteilt wird. Die Verfahren sind alle (bis auf die Lineare Diskriminanzanalyse (LDA)) nach dem Größte-Gruppen-Kriterium (Trefferquote größer als 67,75%) und nach dem Proportional-Chance-Kriterium (Trefferquote größer als 56,30%) als gut zu bewerten. Eine genauere Betrachtung der ebenso guten Ergebnisse des neuronalen Netzes MLP[3] oder des Entscheidungsbaumes C4.5 zeigt, dass hier die hohe Gesamttrefferquote zu Lasten der Treffer innerhalb der Klasse der weniger interessanten Apotheken geht. Da eine ausgewogene Verteilung der richtigen Zuordnungen beider Klassen verfolgt werden sollte, ist die Klassifikation mittels SVM hier vorzuziehen.

Eine Gewichtung von Klassen bewirkt, dass die Ergebnisse der Klassifikation in gewissem Maße gesteuert werden können. Würde auf die Gewichtung verzichtet werden, so würde dies eine vollständige Zuordnung aller Beobachtungen zur größeren Gruppe (Klasse „+1“) und damit eine Trefferquote von 67,75% zur Folge haben. Durch die Erhöhung des Gewichtes der kleineren Klasse wird dies umgangen. Innerhalb des Datensatzes ist Klasse „+1“ mit 1003 Beobachtungen (POS) und Klasse „-1“ mit 471 Beobachtungen (NEG) vertreten, sodass die Klasse „-1“ entsprechend höher gewichtet wird, um das Verhältnis

$$\frac{c_+}{c_-} = \frac{NEG}{POS} = \frac{471}{1003}$$

(vgl. Seite 63) zu erhalten. Das Gewicht von Klasse „+1“ wird auf 1 gesetzt, sodass ein Gewicht von $c_- = 2,1$ für Klasse „-1“ resultiert. Da die Klasse „+1“ diejenigen Beobachtungen bzw. Apotheken beinhaltet, die zu den für das klassifizierende Pharmaunternehmen interessanten Kunden gehören, ist es von Interesse, die Richtigklassifikation dieser Beobachtungen stärker zu forcieren, was mit einer Erhöhung des Gewichtes (hier $c_+ = 1,2$) bei gleichzeitiger Reduzierung des Gewichtes der anderen Klasse (hier $c_- = 1,9$) erreicht wird. Diese Wahl der Kostenparameter liegt in der Vermeidung unnötiger Besuche begründet. Um die richtigen Kunden (also diejenigen, die einen hohen Umsatz erwarten lassen) anzusprechen, ist es wichtiger, Beobachtungen aus Klasse „+1“ richtig zu klassifizieren. Dennoch ist stets die Trefferquote der anderen Klasse zu berücksichtigen, die etwa bei Anwendung von C4.5 oder MLP laut Tabelle 4.2 mit 36,97% bzw. 37,82% sehr gering ausfällt. Daher ist der Einsatz von SVM trotz der geringeren Trefferquote für Klasse „+1“ hier vorzuziehen.

Wie in Abschnitt 3.2.2 ausgeführt wurde, besteht bei SVM die Möglichkeit, die Gewichtung der Klassen auf individuelle Kosten C_i auszudehnen, sodass unter Umständen l verschiedene Gewichte in die Optimierung eingehen. Ein mögliches Ziel dieser Vorgehensweise kann die besondere Behandlung von Ausreißern sein.

[3]Bei Verwendung des MLP wird ein dreischichtiges Netz mit drei Hidden Units gewählt. Die Ausgabeschicht wird durch zwei Units gebildet. Die Zuweisung erfolgt anhand der jeweiligen Aktivierungsniveaus. Es ist die von *Christian Borgelt* erstellte Version von MLP verwendet worden (http://fuzzy.cs.uni-magdeburg.de/∼borgelt/software.html, Zugriff: 20.8.2004).

Diese wurden bei dem hier vorliegenden Datensatz ermittelt und durch besonders kleine Kosten gewichtet, um ihren Einfluss auf die Berechnung der Ebene minimal zu halten, sie aber dennoch im Datensatz zu belassen. Als Ausreißer gelten hierbei die Beobachtungen, die bezüglich mindestens einer der sieben Merkmale nicht unterhalb (bzw. oberhalb) des mittels Boxplots ermittelten größten (bzw. kleinsten) nicht extremen Wertes liegen. Werden diese Ausreißer nun mit $C_i = 1$ und die restlichen Beobachtungen mit den in Tabelle 4.2 angegebenen Klassengewichtungen von $C = 500$ und $c_+ = 1,2$ und $c_- = 1,9$ gewichtet, so ergibt sich bei Einsatz eines Radialbasis-Kerns mit $\gamma = 0,0025$ eine fast identische Trefferquote von 72,89% (vgl. entsprechende Zeile in Tabelle 4.2). Werden hingegen die Kosten für die nicht als Ausreißer identifizierten Beobachtungen auf $C = 1000$ zusammen mit den entsprechenden Klassengewichtungen verändert, so erhöht sich die Trefferquote auf 73,71%. Dies zeigt, dass eine besondere Behandlung einzelner Beobachtungen[4] sinnvoll sein kann, um die Trefferquote bei der Ermittlung der optimalen Trennebene zu erhöhen.
Die gesonderte Behandlung der Ausreißer führt hier zu einer nur geringfügigen Verbesserung der Trefferquote. Dennoch kann dieses Vorgehen zur Vermeidung der durch Ausreißer verursachten, ungewünschten Strukturen innerhalb der Modells verwendet werden. An dieser Stelle ist anzumerken, dass diese Gewichtung von Klassen oder einzelnen Beobachtungen nicht ohne Weiteres bei den alternativen Verfahren vorgenommen werden kann. Dies ist ein besondere Vorteil von SVM, der bei Anwendungen im Marketing ausgenutzt werden sollte.

Neben der Gewichtung der beiden Klassen kann auch den einzelnen Merkmalen (vgl. Abschnitt 3.2.1) höhere Bedeutung zukommen. Eine Analyse der Wichtigkeit der sieben Merkmale bei der linearen Trennung der beiden Klassen zeigt, dass der OTC-Umsatz, die Lage der Apotheke und die Kaufkraft des Postleitzahlgebietes stark mit der vorangegangenen Klasseneinteilung korrelieren. Diese drei Merkmale spielen die wichtigste Rolle bei der Trennung der beiden Klassen. Dies kann durch die schrittweise Diskriminanzanalyse gestützt werden. Bei der Generierung des Entscheidungsbaumes bei C4.5 spielen die Lage und die Kaufkraft ebenfalls eine entscheidende Rolle.
Wird davon ausgegangen, dass der Anteil an OTC-Präparaten und der Umsatz mit einem führenden OTC-Produkt in der Praxis als ein Erfolgsindikator bei der Beurteilung von Apotheken gesehen wird, so kann die Methodik entsprechend verändert werden, um den Einfluss dieser Merkmale zu erhöhen. Dazu kann, wie in Abschnitt 3.2.1 beschrieben, derart in die Klassifikation eingegriffen werden, dass diese Merkmale z.B. mit dem zehnfachen Gewicht gegenüber den übrigen Merkmalen belegt werden. Dies führt bei den sieben Merkmalen zu folgender Gewichtung

$$g_1 = g_2 = g_4 = g_6 = g_7 = 1,\ g_3 = g_5 = 10.$$

[4]Diese besondere Behandlung kann ebenfalls in der Erhöhung einzelner Gewichte bei besonderer Relevanz der jeweiligen Beobachtungen bestehen im Gegensatz zur Verminderung des Einflusses im vorliegenden Fall.

Eine derartige Gewichtung der Merkmale führt sowohl bei den Trainingsdaten als auch bei den Testdaten zu einer Verbesserung des Klassifikationsresultats. Es wird eine leichte Erhöhung der Trefferquote auf 73,2% erreicht. Aber auch nur geringfügige Verbesserungen in der Prognosegüte werden durch die hohen Kosten für Außendienstmitarbeiter gerechtfertigt. Demgegenüber bringt eine fünffache Gewichtung dieser Merkmale lediglich eine Verbesserung des Ergebnisses auf den Trainingsdaten. Die Trefferquote der Testdaten verschlechtert sich bei Klasse „-1“. Dies zeigt die Sensibilität der SVM bei Veränderungen der Parametereinstellungen, wie dies ebenfalls bereits bei der speziellen Gewichtung von Ausreißern zu sehen war. Die automatische Bestimmung der individuellen Merkmalsgewichte nach dem Vorgehen von *Chapelle et al.* (2002) kann daher hilfreich sein.

Neben der guten Trefferquote und der Möglichkeit, die Ergebnisse bzw. deren Güte in gewissem Maße durch Hervorhebung einzelner Klassen und Merkmale zu steuern, hat SVM die Eigenschaft, dass häufig nur ein relativ geringer Teil der Daten zur Klassifikation neuer Beobachtungen benötigt wird. Dieser Teil wird von den Support Vektoren gebildet. Im vorliegenden Fall umfasst die Menge der Support Vektoren etwa 73% der Daten, was leider nur einer geringfügigen Reduktion der ursprünglichen Ausgangsdaten entspricht. Diese Support Vektoren zeichnen sich im Vergleich zur Menge der Vektoren, die nicht zu den Support Vektoren gehören, durch niedrige Werte bei Merkmal 1 und 2 und eher höhere Werte bei Merkmal 3 aus. Die Support Vektoren gehören damit zu denjenigen Apotheken, die einen geringen OTC-Anteil umfassen und durch eine eher schlechtere Lage geprägt sind. Die Gruppe der nicht zu den Support Vektoren gehörigen Apotheken zeichnet sich durch einen sehr hohen Anteil (86,3%) an Beobachtungen aus, die zur Klasse der interessanten Apotheken gehören. Demnach sind wie zu erwarten die kritischen Vektoren[5] diejenigen, bei denen das Besuchsverhalten bzw. das Kundenbetreuungsprogramm nicht a priori klar ist, weil sie nicht deutlich zur Klasse der interessanten Apotheken gehören. Die Analyse der Support Vektoren (bzw. gerade der Vektoren, die nicht zur Menge der Support Vektoren gehören) kann im vorliegenden Fall also insbesondere zur praktischen Unterstützung des Außendienstmitarbeiters dienen, um die Unterschiede zwischen den einzelnen Gruppen sichtbar zu machen und diese für die Besuchsplanung und Kundenbetreuung zu nutzen.

Die folgende Abbildung 4.2 zeigt die ROC-Kurven, die mittels der vier Verfahren „lineare SVM“ und „nicht lineare SVM (RBF)“, „neuronales Netz MLP“ und „lineare Diskriminanzanalyse“ erreicht werden. An den in Tabelle 4.3 eingetragenen Werten für die jeweils erzielte Fläche unter der ROC-Kurve (AUC, vgl. Abschnitt 3.8.3) ist zu erkennen, dass SVM gegenüber dem jeweils vergleichbaren Verfahren leicht besser abschneidet und damit mittels AUC am besten beurteilt werden würde. Nach *Hosmer, Lemeshow* (2000) weist eine Trennung bei einem Wert zwischen 0,7 und 0,8 eine akzeptable Diskriminierung der Klassen auf, was bei allen vier Verfahren zutrifft.

[5] Als kritisch werden hier diejenigen Vektoren bezeichnet, die die Menge der Support Vektoren bilden und demnach einen Einfluss auf die Lage der Ebene haben.

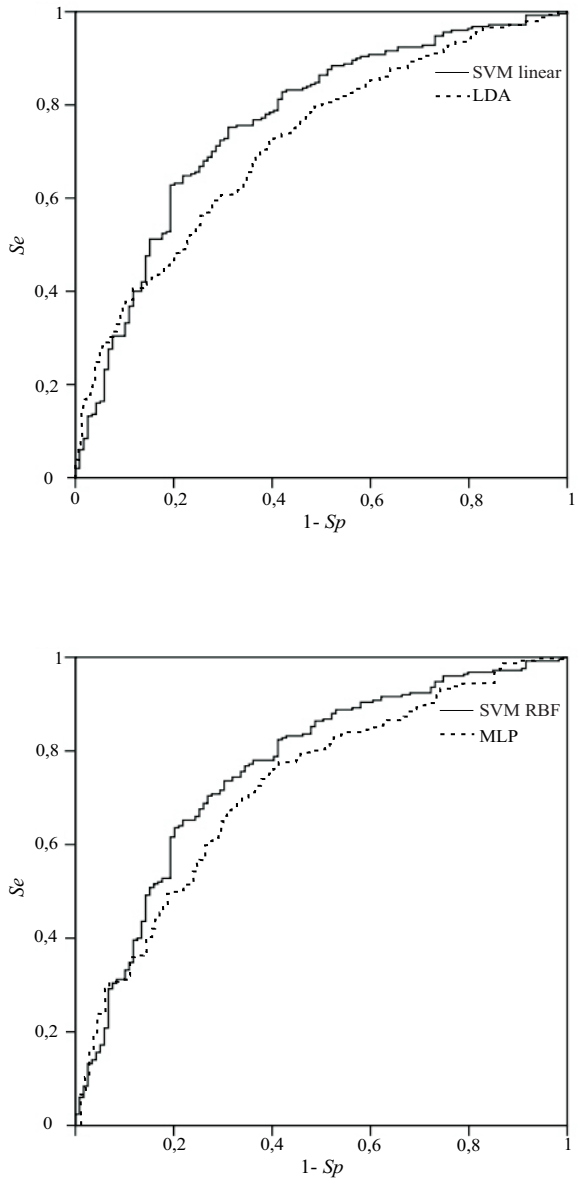

Abbildung 4.2: ROC-Kurven auf Basis von linearer (SVM linear und LDA) und nicht linearer (SVM RBF und MLP) Verfahren

	SVM linear	LDA	SVM RBF	MLP
AUC	0,761	0,718	0,762	0,726

Tabelle 4.3: Werte für die ermittelten Flächen unter der ROC-Kurve

In Abbildung 4.2 ist erkennbar, dass sowohl im linearen als auch im nicht linearen Fall die Kurve der SVM-Trennung die jeweils andere Kurve dominiert, was sich ebenfalls in den AUC-Werten widerspiegelt. Dies bedeutet, dass die Ergebnisse der Trennung basierend auf SVM verglichen zu alternativen Verfahren mittels ROC-Kurven als besser beurteilt werden können.
Insgesamt konnte durch den Einsatz von SVM ein zufrieden stellendes Klassifikationsergebnis erzielt werden, was durch die hohe Trefferquote und die Möglichkeit zur Einflussnahme auf die Gewichtung der Elemente Vorteile gegenüber alternativen Verfahren erkennen lässt.

Ziel der vorliegenden Anwendung von SVM bei der Klassifikation von Apotheken ist jedoch die Entscheidungsunterstützung bei der Zuweisung von Kundenbetreuungsmaßnahmen bei der Kontaktierung neuer Apotheken. Die hier gegebenen 369 Testdaten werden als Teil der insgesamt etwa 21400 Apotheken in Deutschland[6] interpretiert, für die persönlicher oder telefonischer Kontakt oder lediglich die Versendung von Prospektmaterial zur anfänglichen Kundenansprache zur Erweiterung des Kundenstammes geplant werden soll. Die Auswahl der Betreuungsmaßnahmen sollen sich an der Intensität der Klassifikation der Apotheken als potentiell gute bzw. eher umsatzschwache Kunden orientieren. In Abbildung 4.3 wird dazu die Verteilung der Entscheidungswerte gezeigt, die sich bei der Klassifikation der oben verwendeten Testdaten auf Basis einer linearen Trennung mit $C = 500$ und $c_+ = 1,2,\ c_- = 1,9$ ergeben. Die vertikale Achse verdeutlicht die Ausprägung der Entscheidungswerte. Auf eine Markierung der falsch klassifizierten Beobachtungen wird zugunsten einer übersichtlicheren und realistischeren Darstellung verzichtet.

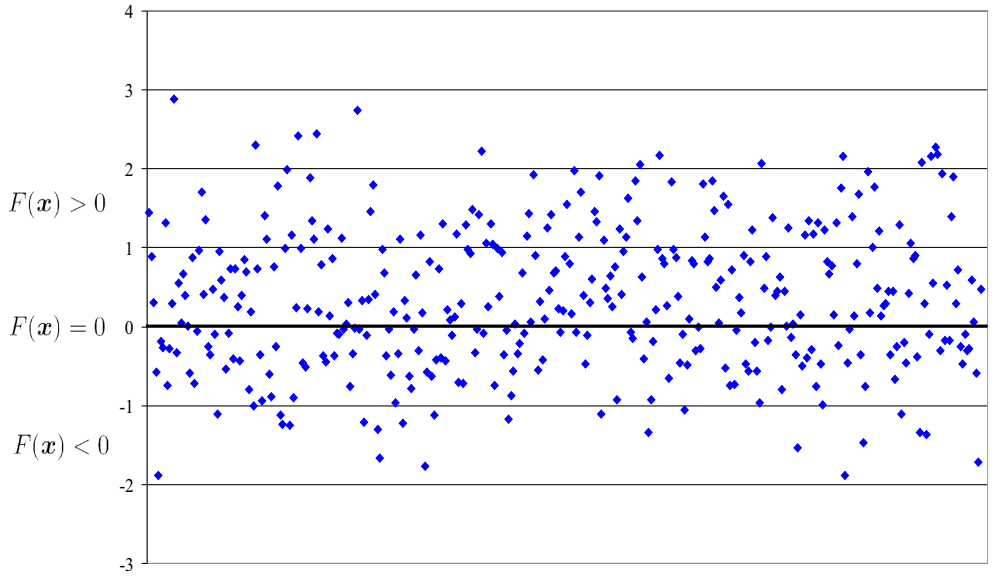

Abbildung 4.3: Aus der Trennung mittels linearem Kern (mit $C = 500$) resultierende Darstellung der Entscheidungswerte der Testdaten

[6]Quelle: Institut für medizinische Statistik, Frankfurt a.M., www.imshealth.de, Zugriff: 18.9.2005

Bei der vorliegenden Trennung kann eine Treffergenauigkeit von 72,36% erwartet werden (vgl. Tabelle 4.2). Da die Entscheidungswerte sich für Trennungen mittels RBF-Kern (mit $\gamma = 0,0025$ und $C = 500$) sowohl in der Klassifikationsgüte als auch in der Ausprägung der Werte („eher großer absoluter Entscheidungswert oder nicht“) sehr ähneln, wird im Folgenden lediglich das Ergebnis der linearen Trennung betrachtet.
Dies bildet die Grundlage für die Entwicklung von Marketingstrategien, die sich im Fall einer umfangreichen Erweiterung des Kundenstammes auf die Ausgestaltung der ersten Kundenkontakte beziehen. Entsprechend des jeweiligen Klassifikationsergebnisses bzw. entsprechend der Entscheidungswerte soll die Ausgestaltung der Kundenbeziehungen differenziert werden. Um eine derartige Differenzierung vornehmen zu können, werden wie in Abschnitt 3.7 beschrieben die Bereiche oberhalb und unterhalb der Trennebene in unterschiedliche Zonen eingeteilt. Eine Vorgehensweise zur Einteilung der Bereiche in Anlehnung an den BERI-Index (vgl. Abschnitt 3.7.2) erscheint aufgrund mehrerer Kriterien nicht adäquat. So werden lediglich vier Bereiche benötigt, die sich hinsichtlich der methodischen Vorgehensweise besser an den innerhalb des Verfahrens wichtigen Entscheidungswerten $F(\boldsymbol{x}) = 1$ und $F(\boldsymbol{x}) = -1$ orientieren sollten. Die Intervallgrenzen bei der Vorgehensweise in Anlehnung an den BERI-Index sind losgelöst von diesen Einscheidungswerten. Dies ist als größter Nachteil dieser Art der Einteilung zu sehen. Daher wird hier eine alternative Einteilung vorgezogen, bei der die Spanne zwischen den eingehenden Hilfsebenen in zwei Bereiche aufgeteilt wird. Es werden lediglich vier Abstufungen vorgenommen, um den oben erläuterten gesetzten Zielen dieser Anwendung gerecht zu werden. Tabelle 4.4 zeigt diese Einteilung der Wertigkeitsbereiche, bei der jede der 369 vorliegenden Apotheken einer der vier ausgewählten Bereiche zugeordnet wird.

Bereich	Intervall der Entscheidungswerte	Anzahl zugewiesener Kunden	Trefferquote
AA	[1;∞[	86	88,37%
A	[0;1[	136	80,15%
B	[-1;0[	124	53,23%
C	]−∞;-1[	23	69,57%

Tabelle 4.4: Definition und Belegungsdichte der Wertigkeitsbereiche

Die zusätzliche Angabe über die Trefferquote, die innerhalb jedes einzelnen Bereiches erreicht wird, zeigt deutlich, dass die Beobachtungen, die in den Bereich B fallen zu den eher unsicher klassifizierten Beobachtungen zu zählen sind. Der Bereich C weist eine bei diesem Modell durchschnittlich zu erwartende Trefferquote von 69,57% auf, wohingegen die Bereiche A und AA durch eine überdurchschnittlich hohe Trefferquote von 80,15% bzw. 88,37% gekennzeichnet sind. Bei der Allokation des Marketingbudgets kann ein besonderes Augenmerk auf die letzteren Beobachtungen gelegt werden, da hier eine überdurchschnittlich hohe Wahrscheinlichkeit der Richtigklassifikation zu der Klasse der potentiell interessanten Kunden besteht.

Eine mögliche Differenzierung der Marketingaktivitäten liegt durch die folgende Zuweisung der oben genannten vier Alternativen zu den Apotheken der jeweiligen Bereiche vor:

AA	Außendienstbesuche verbunden mit eDetailing
A	Außendienstbesuche
B	telefonische Kontaktierung
C	Versendung von Prospektmaterial

Im vorliegenden Fall besteht die Aufgabe der Marketingabteilung darin, für ein gegebenes Budget den Einsatz des Außendienstes so zu steuern, dass die Erfolg versprechenden Apotheken verstärkt besucht werden, wohingegen die eher uninteressanten Apotheken nicht besucht, sondern telefonisch kontaktiert oder mit Prospektmaterial versorgt werden. Die Zuweisung der Strategien zu den vier Bereichen liegt darin begründet, dass mit zunehmenden Entscheidungswerten die Zugehörigkeit zu den interessanten Kunden steigt. Da Besuche des Außendienstes einen wesentlichen Erfolgsfaktor in der Pharmabranche bilden (*Baier et al.* (2004)) werden diese Besuche als kosten- und zeitintensives Element nur den potenziell guten Kunden zugewiesen, um diese an das eigene Unternehmen zu binden. Als kostspielige Erweiterung ist das eDetailing zu sehen, was denjenigen Kunden zusätzlich zum Einsatz des Außendienstes angeboten werden soll, die mit einer sehr hohen Wahrscheinlichkeit auch zu den interessanten Kunden des Pharmaunternehmens gehören. Die angegebenen Marketingaktionen schließen sich nicht gegenseitig aus. So kann eine Apotheke aus Bereich A ebenfalls Prospektmaterial erhalten, allerdings sollten die Anstrengungen für die Apotheken aus Bereich C sich in der Anfangsphase der Markterschließung nur auf diese Marketingaktionen beschränken. Da keine Aussagen über die Güte der gewählten Aktionen getroffen werden können, kann die Bewertung des Modells nur anhand der in Tabelle 4.4 angegebenen Trefferquoten erfolgen. Da in Bereich AA mit einer Trefferquote von 88,37% die Apotheken mit den Marketingaktionen angesprochen werden, die einen hohen Umsatz erwarten lassen, so ist die Zuweisung von kostspieligen Außendienstbesuchen und der zusätzliche Einsatz eines eDetailing-Systems durchaus berechtigt. Die geringe Trefferquote von 69,57% in Bereich C lässt hingegen darauf schließen, dass hier viele Apotheken, die einen hohen Umsatz erwarten lassen, also zu Klasse „+1“ gehören, durch die Versendung von Informationsmaterial nicht adäquat angesprochen werden. Dennoch kann durch diese Art der Aufteilung ein erster Ansatzpunkt zur differenzierten Kundenansprache bei Neukunden gegeben werden, der ein besseres Ergebnis als die zufällige Zuweisung der ausgewählten Marketingaktionen erwarten lässt. Durch den Einsatz von SVM wird dem Risiko der nicht adäquaten Kundenansprache mit höheren Trefferquoten als bei der zufälligen Verteilung entgegen getreten.

Die hier verwendeten Daten umfassen lediglich 1475 Beobachtungen, für die die notwendigen Informationen der Apotheken vorliegen. In einer realen Anwendung kann das hier entwickelte Modell auf alle Apotheken Deutschlands angewendet werden, für die die benötigten Daten zugänglich sind. Diese Implikationen bauen

allerdings lediglich auf den Ergebnissen der vorliegenden Daten auf und können bei Hinzunahme anderer Beobachtungen oder Veränderungen innerhalb eines Jahres variieren.
Bei der letztendlichen Zuweisung der Marketingaktivitäten zu den Apotheken spielen die Kosten für einen einzelnen Besuch, die Kapazität des Außendienstes sowie das zur Verfügung stehende Budget für die gesamten Marketingaktivitäten eine entscheidende Rolle. Dies bedeutet, dass die Bestimmung der Bereichsgrenzen immer in Abhängigkeit der jeweils vorliegenden Situation erfolgen muss und nicht pauschal im vorhinein bestimmt werden kann.
Die vorgestellte Anwendung zeigt, dass sich die Eigenschaften einer SVM ausnutzen lassen, um diese Methodik effektiv im Rahmen der Steuerung der Kundenansprache eines pharmazeutischen Unternehmens einzusetzen. Dabei sind alternative resultierende Marketingstrategien denkbar. So könnte mit Hilfe der Entscheidungswerte ebenfalls eine Steuerung der jährlichen Besuchszahlen im Außendienst erzielt werden, wenn davon ausgegangen wird, dass jede der zu betrachtenden Apotheken besucht werden soll. Eine mögliche Einteilung wäre etwa durch die folgende beispielhafte Zuweisung von Häufigkeiten der Außendienstbesuche zu den vier Bereichen gegeben:

AA	7-8 Außendienstbesuche
A	5-6 Außendienstbesuche
B	3-4 Außendienstbesuche
C	1-2 Außendienstbesuche

Somit wären durch die Apotheken aus Bereich AA wiederum die Kunden mit dem höchsten Potenzial gegeben, denen eine verstärkte Aufmerksamkeit durch (in vertretbarem Maße) gehäufte Außendienstkontakte zukommt. Gerade bei diesem Beispiel zeigt sich, dass die Festlegung der Bereichsgrenzen individuell erfolgen muss, um eine Abstimmung mit den finanziellen Mitteln und dem Außendienstkontingent zu erzielen.

Bei der hier vorliegenden Auswahl an Merkmalen und dem gegebenen Ziel der Untersuchung bietet sich ebenfalls der Einsatz der Regressionsanalyse an, um auf Basis der Merkmale die Anzahl der durchzuführenden Besuche festzulegen. Daher ist SVM hier nur als eine alternative Möglichkeit zu sehen, die zusätzlich zu traditionellen Verfahren angewendet werden kann. Zudem können durch eine Diskretisierung der Ausgabe einer reellwertigen, abhängigen Variablen unter gewissen Umständen bessere Prognosewerte erzielt werden, wie *Bodapati, Gupta* (2004) ausführen. Dies spricht ebenfalls für einen Einsatz der SVM, der zusätzlich zur Regression durchgeführt wird.

Die Datengrundlage enthält Daten, die teilweise aus der bisherigen Zusammenarbeit mit dem Unternehmen resultieren. Bei der Vermarktung eines Neuproduktes kann sich die Ausrichtung des Vertriebs auch an bisherigen Gegebenheiten der Konkurrenz ausrichten, wie es im folgenden Abschnitt durchgeführt wird.

4.3.2 Klassifikation von Ärzten

Hier wird eine alternative Strategie bei der Vermarktung eines verschreibungspflichtigen Medikamentes herangezogen. Da es sich in diesem Abschnitt um ein zwar verschreibungspflichtiges aber häufig medizinisch nicht notwendiges Medikament handelt, soll bei der Steuerung des Außendienstes eine Orientierung an dem bisherigen Konkurrenzprodukt stattfinden. Ziel der Untersuchung ist die Identifikation derjenigen Ärzte, die mit hoher Wahrscheinlichkeit dieses Produkt verschreiben würden. Diese Ärzte sind dadurch gekennzeichnet, dass sie bereits das Konkurrenzprodukt verschreiben und somit eine hohe Prädisposition für die Verschreibung dieser Art von Medikamenten haben. Der Erfolg eines verschreibungspflichtigen Präparates, gemessen an der Verschreibungshäufigkeit, kann unter anderem von den Besuchen des betreffenden Pharmaunternehmens bei den entsprechenden Ärzten abhängen (*Albers* (2002)). Daher ist auch hier eine regelmäßige Betreuung durch Vertreter des Unternehmens wichtig. Durch diese Anwendung wird eine alternative Bestimmung der Steuerung des Außendienstes realisiert. Die Grundgesamtheit niedergelassener Ärzte ist auch in diesem Fall zu groß, um alle Ärzte häufig besuchen zu können. Die Identifikation der wichtigen Ärzte kann analog zum vorangegangenen Abschnitt im Rahmen einer Klassifikation mit Hilfe von SVM durchgeführt werden. Die zu untersuchende Fragestellung unterscheidet sich lediglich in der inhaltlichen Motivation der Klassifikation.

Im Folgenden werden die Ärzte in Abhängigkeit der Intensität, ein von dem betreffenden Pharmaunternehmen hergestelltes Medikament zu verschreiben, klassifiziert. Dazu liegen zwei Gruppen gleichen Umfangs vor, um die Auswirkungen einer Variation des Kostenparameters später besser zu zeigen[7]. Die für das Pharmaunternehmen besonders interessante Klasse „+1“ bilden diejenigen Ärzte, die das betreffende Medikament häufig verschreiben und somit auch als Absatzmittler für das neue, konkurrierende Produkt in Frage kommen. Die andere Gruppe wird durch die Ärzte gebildet, bei denen dieses spezielle Medikament nicht von der Bevölkerung des Einzugsgebietes nachgefragt wird und die Ärzte es entsprechend wenig verschreiben. Neben der inhaltlichen Ausrichtung unterscheidet sich der Datensatz zu dem in 4.3.1 verwendeten hauptsächlich durch die Auswahl der eingehenden Merkmale. Da sich die Ausrichtung der Marketingaktivitäten an der Verschreibungsintensität des Konkurrenzproduktes orientieren sollen und keine Daten über die jeweiligen Patienten vorliegen, werden die Ärzte durch Charakteristika ihres Einzugsgebietes beschrieben. Somit wird ebenfalls gewährleistet, dass neu zu klassifizierende Ärzte ebenfalls mit diesem Modell behandelt werden können. Im Einzelnen werden die Beobachtungen durch die jeweiligen Anteile an ausländischen Haushalten, an Haushalten mit niedrigem und mit hohem sozialen Status, an 1 oder 2-Familienhäusern und an 7- oder mehr-Familienhäusern sowie durch die Gesamtanzahl der Haushalte im Einzugsgebiet beschrieben. Es liegen insgesamt 1446 Ärzte vor, die einer der beiden oben aufgeführten Klassen zugeordnet sind

[7]Ein ähnlicher auf der gleichen Datengrundlage basierender Datensatz größeren Umfangs mit ungleich großen Klassen wurde bereits in *Monien, Decker* (2004) eingesetzt.

und durch 6 Merkmale beschrieben werden.

Tabelle 4.5 enthält die Ergebnisse, die auf zufällig ausgewählten Testdaten mittels unterschiedlicher Verfahren erzielt werden. Die Parameter bei SVM wurden durch eine vorangegangene Kreuzvalidierung ermittelt[8].

Verfahren	Parameter			Trefferquoten		
	γ	d	C	Klasse $+1$	Klasse -1	Gesamt
SVM linear ($c_- = 1,7$)	-	-	500	75,14%	74,59%	74,86%
SVM RBF ($c_- = 1,2$)	0,1	-	1000	77,90%	71,82%	74,86%
SVM Poly ($c_- = 1,2$)	-	3	100	81,77%	67,96%	74,86%
LDA	-	-	-	82,87%	61,33%	72,10%
C4.5	-	-	-	82,87%	65,75%	74,31%
MLP	-	-	-	71,82%	72,38%	72,10%

Tabelle 4.5: Trefferquoten der Testdaten einzelner Verfahren im Vergleich

Anhand der Gesamttrefferquote ist zu erkennen, dass SVM den übrigen Verfahren überlegen ist. Selbst die lineare SVM liefert bessere Ergebnisse als die nicht linearen Verfahren C4.5 und MLP[9], was die Leistungsfähigkeit der SVM bei Klassifikationsaufgaben dieser Art zeigt. Bei den vorliegenden Daten handelt es sich um balancierte Daten, was bedeutet, dass die beiden Klassen die gleiche Anzahl an Beobachtungen umfassen. Dies kann ein Grund dafür sein, dass die zu SVM vergleichbaren Verfahren in diesem Fall ebenfalls recht gut klassifizieren.
Weiterhin ist zu erkennen, dass fast alle Verfahren die Klasse derjenigen Ärzte besser klassifizieren, die das Medikament sehr häufig verschreiben. Die andere Klasse wird hingegen nicht so gut erkannt. Ein ausgewogenes Ergebnis, bei dem die Trefferquoten für beide Klassen in etwa gleich sind, lässt sich sowohl mit SVM durch Steuerung der Kostenparameter bei linearer oder nicht linearer Trennung als auch mittels MLP erzielen.
Das in Abschnitt 3.8.2 vorgestellte Maß zur Beurteilung des Klassifikationsergebnisses von SVM (M_{ext} bzw. M_{med}) unter Berücksichtigung der Entscheidungswerte ergibt im vorliegenden Fall die in Tabelle 4.6 aufgezeigten Werte.

Verfahren	M_{ext}	M_{med}
SVM linear	-0,573	0,223
SVM RBF	-0,566	0,239
SVM Poly	-0,501	0,186

Tabelle 4.6: Beurteilung der Güte der Klassifikation anhand von M_{ext} und M_{med}

[8]Das gewählte Verhältnis von Trainings- und Testdaten wurde bei der zufälligen Auswahl der Testdaten beibehalten.

[9]Bei Verwendung des MLP wird ein dreischichtiges Netz mit drei Hidden Units gewählt, wobei die Ausgabeschicht durch zwei Units gebildet wird.

Nr.	Merkmal j	lineare Trennung		nicht lineare Trennung	
		r_j	Rang	r_j	Rang
1	Anteil ausländischer Haushalte	5,45	1	0,59	1
2	Haushalte mit niedrigem Status	1,66	4	0,14	4
3	Haushalte mit hohem Status	0,71	6	0,05	6
4	1- oder 2-Familienhäuser	2,09	3	0,10	5
5	7- oder mehr-Familienhäuser	2,74	2	0,16	3
6	Gesamtanzahl der Haushalte	1,25	5	0,28	2

Tabelle 4.7: Wichtigkeit der Merkmale bei lineare und nicht linearer Trennung

Für die beiden Maße liefern die drei eingesetzten Kerne jeweils ähnliche Werte, die nur geringfügig voneinander abweichen. Werden jedoch die beiden Werte der Maße pro eingesetztem Kern verglichen, so unterscheiden diese sich sehr drastisch bis zu einer Abweichung von $0,239 - (-0,566) = 0,805$. Dies zeigt zunächst, dass die durch die drei Kerne ermittelten Ebenen sehr ähnliche Strukturen aufweisen. Hier ist ebenfalls der Einfluss von Ausreißern auf das Maß M_{ext} gut zu erkennen. Während M_{ext} für alle drei Kerne negative Werte zwischen -0,5 und -0,58 liefert, ist die Bewertung unter Einbeziehung des Medians positiv mit einem Wert von etwa 0,22. Zur Bewertung der Klassifikationsgüte sollte daher M_{med} herangezogen werden, um stabile Aussagen zu generieren. In diesem Fall kann aufgrund des positiven Wertes auf eine zufrieden stellende Klassifikation geschlossen werden.
Wird eine Analyse der Wichtigkeiten der einzelnen Merkmale mit dem Normalenvektorverfahren (für die lineare Trennung) und dem Gradientenverfahren (für die nicht lineare Trennung mittels Radialbasis-Kern) durchgeführt, so resultieren die in Tabelle 4.7 aufgeführten Ergebnisse. Die Absolutwerte der Relevanzen r_j eines Merkmals j haben hierbei keine Bedeutung. Es ist zu erkennen, dass mittels beider Verfahren der Anteil an ausländischen Haushalten als das Wichtigste bestimmt wird, wohingegen der Anteil an Haushalten mit hohem Status für die Trennung irrelevant erscheint. Bei der Ermittlung des nächsten wichtigen Merkmals unterscheiden sich die Verfahren jedoch. Während bei der linearen Trennung der Anteil an 7- oder mehr-Familienhäusern eine weitere wichtige Rolle übernimmt, hat die Gesamtanzahl der Haushalte hohe Relevanz für die Trennung mittels nicht linearer SVM. Im Weiteren wird lediglich die nicht lineare Trennung mittels Radialbasis-Kern betrachtet. Werden die Beobachtungen auf die beiden wichtigsten Merkmale (vgl. Tabelle 4.7) reduziert, so wird immer noch mittels Radialbasis-Kern eine Gesamttrefferquote von 73,4% erreicht. Bei Wahl der beiden Merkmale, die hinsichtlich des Gradientenverfahrens als nicht relevant für die Klassifikation charakterisiert wurden (Anteil der Haushalte mit einem hohen sozialen Status und Anteil der 1-2 Familienhäuser im Einzugsgebiet des jeweiligen Arztes), ergibt sich eine Reduzierung der Trefferquote auf 57,4%, was bei balancierten Daten kaum besser als eine zufällige Auswahl der Ärzte ist. Dies zeigt, dass der Einsatz von Verfahren zur Merkmalsreduzierung sinnvoll ist, um den Umfang der zu erhebenden Merkmale für neue Beobachtungen zu minimieren und gleichzeitig eine dennoch hohe Trefferquote zu gewährleisten.

In Abbildung 4.4 wird die resultierende Trennebene bei Einsatz des Radialbasis-Kerns mit $\gamma = 0,1$ und $C = 1000$ gezeigt (gestrichelte Linie), wobei lediglich die wichtigsten beiden Merkmale verwendet werden[10]. Die Beobachtungen der Klasse "+1" werden durch Kreise, die Beobachtungen aus Klasse "-1" durch Punkte repräsentiert. Durch die zusätzlich eingefügten Trennebenen (durchgezogenen Linien) wird verdeutlicht, wie sich eine Erhöhung des Kostenparameters für die Klasse der interessierenden Ärzte auf die Lage der Ebene auswirkt. Je höher die Kosten gewählt werden (Variation von $c_+ \in \{1,0; 1,1; 1,2; 1,4; 1,7\}$), desto weiter bewegt sich die Ebene in der Abbildung nach oben. Dies liegt darin begründet, dass die fehlklassifizierten Beobachtungen der Klasse „+1" am oberen Rand des Raumes zu hohen Kosten führen und somit möglichst richtig klassifiziert werden sollten, um die Zielfunktion (vgl. Abschnitte 2.1.3 und 2.1.4) zu minimieren.

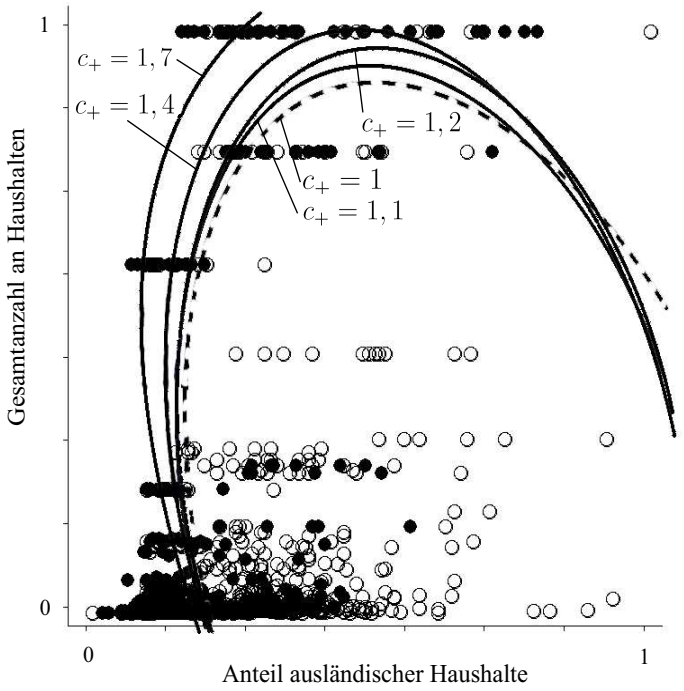

Abbildung 4.4: Aus der Trennung mittels RBF-Kern (mit $C = 1000$) resultierende Trennebenen bei Erhöhung der Kosten c_+ (∘: Klasse „+1"; •: Klasse „-1")

Durch die Trennung bei Wahl des Kostenparameters $c_+ = 1,7$ wird bei zwei Merkmalen immer noch eine Gesamttrefferquote von 71,0% erreicht. Dies wird allerdings durch eine deutlich verringerte Treffergenauigkeit von nur noch 55,8% auf Seiten der Klasse „-1" beeinflusst, die durch die besondere Hervorhebung der anderen Klasse (Trefferquote: 86,2%) vernachlässigt wird. Dies zeigt, dass die zusätzliche Betrachtung der klassenindividuellen Trefferquoten weitere wichtige

[10]Es sei an dieser Stelle angemerkt, dass die Merkmalsausprägungen jeweils auf das Intervall $[0, 1]$ normiert wurden (vgl. Abschnitt 3.1). Daher geben die Ausprägungen beispielsweise des Merkmals „Gesamtanzahl an Haushalten" nicht mehr die absoluten Werte der jeweiligen Beobachtung an.

Informationen liefern kann, die für die Festlegung eines Modells, sowie für die Entwicklung von Strategien entscheidend sein können.
Abbildung 4.4 lässt erkennen, dass innerhalb der Klasse der interessanten Kunden (Klasse „+1“) eine große Anzahl an Beobachtungen (also Ärzte) enthalten ist, deren Einzugsgebiet durch einen erhöhten Anteil an ausländischen Haushalten gekennzeichnet ist verglichen zur anderen Klasse. Hinsichtlich der Gesamtanzahl der Haushalte ist kein eindeutiger Unterschied zu erkennen. Dies lässt die Vermutung zu, dass diejenigen Ärzte, die verstärkt durch das pharmazeutische Unternehmen besucht werden sollten, zumeist in Orten oder Ortsteilen mit einem höheren Ausländeranteil tätig sind. So kann die Visualisierung der Trennebene im niedrig dimensionalen Raum zusätzlich als Entscheidungsunterstützung dienen.

Je nach Zielsetzung kann die eine oder die andere Klasse der Ärzte im Mittelpunkt der Untersuchungen stehen und durch die Kosten für die jeweilige Klasse durch c_+ bzw. c_- gesteuert werden. Gilt es, neue, bisher nicht besuchte Ärzte zu klassifizieren und die Besuchshäufigkeit etwa an die Ergebnisse der Klassifikation anzupassen, so steht die richtige Klassifikation der Beobachtungen der Klasse „+1“ im Vordergrund, um für das Unternehmen wichtige Ärzte, die eine hohe Verschreibungshäufigkeit erwarten lassen, nicht zu übersehen. Dies entspricht der Erhöhung der Kosten für die betreffende Klasse, hier also c_+. Eine erhöhte Besuchsfrequenz bei weniger wichtigen Ärzten kann hierbei als nicht so schwer wiegender Fehler gesehen werden. Geht es hingegen bei der Klassifikation darum, schon bekannte Ärzte einzuordnen und die Besuchshäufigkeiten zu planen, so kann eher der Fokus auf diejenigen Ärzte gelegt werden, die das Medikament eher weniger oft oder gar nicht verschreiben. So wird einer drohenden Abwendung eines Arztes von diesem Medikament durch erhöhte Zuwendung durch Besuche von Seiten des Unternehmens entgegengewirkt. Eine Differenzierung bei der Besuchshäufigkeit kann in Anlehnung an die auf Seite 130 vorgenommene Unterteilung des Wertigkeitsbereichs vorgenommen werden. Dies bildet die Basis, um die Außendienstaktivitäten auf die Ärzte, gemäß ihrer Wichtigkeit gemessen an der Verschreibungsintensität des Produktes, zu verteilen.

4.3.3 Zusammenfassung

In den vorangegangenen Abschnitten wurde der Einsatz von SVM im Rahmen der Planung und Steuerung von Marketing- und insbesondere von Außendienstaktivitäten untersucht und dazu insbesondere auf die Nutzung der generierten Ergebnisse eingegangen. Die Aufteilung des Entscheidungsraumes hat gezeigt, wie die Ergebnisse zur effektiven Allokation des Marketingbudgets beitragen können.
Je nach Zielsetzung kann die Ausrichtung der SVM variiert werden. Statt des erhöhten Aufwandes für die zu erwartenden guten Kunden, ist ebenfalls eine alternative Strategie denkbar. Im Fall der Klassifikation von Apotheken wäre

aufgrund der Gruppenbildung beispielsweise ebenfalls eine intensive Betreuung der schwächeren Apotheken denkbar, um den Umsatz zu fördern und gerade diese Kunden zu aktivieren.
Die Möglichkeit der Anpassung der Parameter an die Daten und an die jeweilige Zielsetzung zeigt weiterhin die Flexibilität der SVM, die dadurch einen großen Vorteil gegenüber traditionellen Verfahren wie LDA haben, wo diese Möglichkeit der Fokusverschiebung nicht gegeben ist. Diese Anpassungsfähigkeit wird durch den Einsatz unterschiedlicher Kerne ermöglicht.
Die Ergebnisse der Anwendungen in Abschnitt 4.3.1 konnten durch die Miteinbeziehung von speziellen Gewichtungen von individuellen Beobachtungen (hier Ausreißer) und Merkmalen leicht verbessert werden. Dies zeigt, dass die Möglichkeit zur Einbeziehung von a priori-Wissen bei SVM genutzt werden sollte, um optimale Ergebnisse zu erzielen. Demgegenüber kann diese Gewichtung bei alternativen Verfahren nur eingeschränkt angewendet werden, was die Überlegenheit von SVM diesbezüglich verdeutlicht.

Die hier verwendete Klassifikation von Apotheken und Ärzten als Kundengruppen eines pharmazeutischen Unternehmens sollten in realen Anwendungen als Teil einer Analyse des Marktes dienen, da die Klassifikation mittels SVM lediglich eine kleine Facette im Rahmen der betrieblichen Entscheidungsunterstützung bildet. Die Anwendung von SVM kann aber als effektive Unterstützung bei der Planung der Außendienstallokation oder Überprüfung der Resultate dienen. So kann die SVM zur Überprüfung der Ergebnisse oder zur Ergänzung herangezogen werden, wenn zuvor zur Ermittlung der Besuchszeiten und der Außendienststärke, wie etwa in *Männche* (2004), eine Responsefunktion ermittelt wurde. Durch ihre Flexibilität und Klassifikationsstärke liefern SVM häufig bessere Ergebnisse als traditionelle Verfahren zur Klassifikation und lassen sich, wie die Ausführungen gezeigt haben, durch die spezielle Interpretation der Entscheidungswerte im Rahmen der Gestaltung der Kundenbetreuung sehr gut einsetzen, da durch die Bildung unterschiedlicher Bereiche explizite Angaben über mögliche Behandlungen von Kunden durch die zusätzliche Differenzierung innerhalb einer Klasse gegeben werden können.

4.4 Anwendung von SVM in der Kaufverhaltensanalyse

Da Apotheken (und eingeschränkt auch Ärzte) als Absatzmittler bei der Vermarktung von pharmazeutischen Produkten auftreten, handelt es sich bei den bisherigen Anwendungen um eine Business-to-Business Sichtweise. Um zusätzlich die Fokussierung auf den Endverbraucher zu richten, wird im Folgenden ein weiterer Datensatz analysiert, bei dem der Konsument im Mittelpunkt steht und damit zur Business-to-Consumer Sichtweise gewechselt wird. Es sollen diejenigen Endverbraucher identifiziert werden, die für eine spezielle Marketingaktion in Frage

kommen und positive Resonanz erwarten lassen. Dazu werden wiederum SVM eingesetzt, um so zwischen unterschiedlichen Kundengruppen oder Käufertypen in Konsumgütermärkten zu differenzieren. Durch ein gezieltes Ausrichten der Marketingaktivitäten an den Wünschen der Kunden soll die Zufriedenheit auf Seiten der Kunden erhöht werden. Somit wird den Prinzipien des Relationship-Marketings als Grundsatz der modernen marktorientierten Unternehmensführung Rechnung getragen. Nach *Aaker et al.* (2004) bestimmen drei wesentliche Elemente den Erfolg des Relationship-Marketings. Den Grundstein bildet eine gepflegte und aktuelle Kundendatenbank, die Informationen über gegenwärtige und mögliche Kunden enthält. Das zweite Element ist die Differenzierung der Kundenansprache der unterschiedlichen Zielgruppen, die in diesem Abschnitt ebenfalls verfolgt wird. Weiters wird beim Relationship-Marketing durch die Aufzeichnung der Interaktionen mit den Kunden eine effektive Allokation der Marketingressourcen ermöglicht, die eine Messbarkeit des Erfolgs mit sich zieht.
Im Gegensatz zu den zu beschaffenden unternehmensexternen Daten in den vorangegangenen Abschnitten wird hier nun auf unternehmensinterne Daten zurückgegriffen, die aus den bisherigen Beziehungen zu den einzelnen Kunden resultieren. Die Konsumenten werden hierzu in dieser Analyse u.a. durch ihre bisherige Kaufhistorie charakterisiert. Die Ansprache der Kunden wird hier durch Direktmailing durchgeführt, um eine gezielte Interaktion mit dem Kunden zu gewährleisten und eine Bindung des Kunden an das Unternehmen zu realisieren.
Die Verfahren zur derartigen Behandlung des Marktes werden zum einen durch traditionelle Methoden wie dem Scoring gebildet. So wird in *Link, Hildebrand* (1997) Scoring zur Abschätzung der Kaufwahrscheinlichkeit durch Berücksichtigung von monetären und kaufrelevanten Merkmalen bei der Punktbewertung vorgeschlagen. Zur Realisierung der individuellen Kundenansprache auch im Massenmarketing kommen verstärkt Verfahren des Data Mining zum Einsatz (*Wilde* (2001)). Die Methoden können zunächst zur Strukturierung des Marktes eingesetzt werden. Auf Basis der auch in diesem Abschnitt verwendeten Daten werden etwa in *Decker* (2005a) mit Hilfe von wachsenden neuronalen Netzen Lifestyle-Segmente gebildet. Die Einstellungen der Kunden gegenüber ernährungsbezogenen Aspekten dienen der Segmentierung. Eine ähnliche Segmentbildung wird in *Lüdtke, Schneider* (2001) durchgeführt. Zur Bildung der Lifestyle-Segmente auf äquivalenten Einstellungsdaten und kaufrelevanten Merkmalen wird hier die Clusteranalyse herangezogen.
Neben der Marktsegmentierung können Verfahren des Data Mining in der Kaufverhaltensforschung zur Entwicklung von Recommender-Systemen dienen. Klassische Anwendungen finden sich hierbei im E-Commerce. Auf Basis des bisherigen Kaufverhaltens werden bei *Lawrence et al.* (2001) mittels Assoziationsregeln Kaufempfehlungen bzgl. anderer Produkte gegeben. Die vorgestellte Vorgehensweise kann ebenfalls auf die Situation bei Einsatz von Kundenkarten im Lebensmitteleinzelhandel (LEH) übertragen werden. Daneben können auch Entscheidungsbäume oder neuronale Netze sinnvoll zur Klassifikation bzw. Zielgruppenbestimmung im Rahmen des CRM eingesetzt werden (*Berry, Linoff* (2000)). Diese Klassifikationsverfahren bilden somit die Verbindung von Segmentierung und der Abgabe von Empfehlungen bzw. der individuellen Kundenansprache.

Als viel versprechendes und leistungsstarkes Klassifikationsverfahren sollen SVM auch in diesem Bereich verwendet werden, um die Möglichkeit der in Abschnitt 3.7 vorgestellten Ergebnisinterpretation bei der Zuordnung der Kunden zu Segmenten und der darauf folgenden Kundenansprache sinnvoll nutzen zu können. In diesem Abschnitt wird untersucht, inwieweit SVM eingesetzt werden können, um den gesetzten Zielen gerecht zu werden. Diese bestehen in der möglichst zuverlässigen Zuweisung adäquater Marketingaktionen zu den jeweils klassifizierten Kunden. Die dazu benötigten Informationen über das Kaufverhalten der Kunden können aus einem Haushaltspanel gewonnen oder durch Kundenkarten erhoben werden. Die Kundenkarten übernehmen nach *Homburg, Krohmer* (2003) zwei Funktionen: Zum einen liefern sie die für die Analyse wichtigen Daten über das Kaufverhalten und zum anderen binden sie den Kunden durch entsprechende Kundenclubs an das Unternehmen. Im vorliegenden Fall liegen Daten aus einem Haushaltspanel vor, wobei die Anwendung aufgrund der Datenauswahl auf die Situation eines Unternehmens mit Kundenkartensystem übertragen werden kann. Es wird die Ausrichtung der Marketingaktivitäten eines Lebensmitteleinzelhändlers betrachtet. Die Ausführungen bleiben aufgrund der Datenauswahl auf die Händlerperspektive beschränkt.[11] Aufgrund der Händlersicht ist die Kundenbindung hier als Bindung des Kunden an eine Filiale oder an eine Einkaufstätte zu verstehen. Nach *Gündling* (1997) können die Schritte der Wahrnehmung der individuellen Bedürfnisse der Kunden, die Umsetzung der Problemlösung zur Erfüllung der Bedürfnisse und die Überprüfung der Kundenzufriedenheit als Ziele für eine maximale Kundenorientierung gesehen werden. Bei der Anwendung von SVM kann der erste Schritt durch die Klassifikation des Kunden auf Basis seiner Kaufhistorie etc. zu Käufer- bzw. Ernährungstypen realisiert werden. Die Umsetzung der Problemlösung erfolgt durch Versenden der relevanten Werbung, die für diese Kundengruppe von Interesse sein kann. Der Kauf bzw. Nichtkauf gibt letztendlich Aufschluss darüber, ob die Bedürfnisse und Wünsche des Kunden erfüllt wurden. Somit bilden diese Kunden wieder neue Beobachtungen zur Entwicklung eines neuen SVM-Modells, auf dessen Basis weitere Kunden klassifiziert werden können und SVM somit iterativ angewendet wird.

Das Ziel der folgenden Abschnitte ist demnach die Zuordnung der Kunden zu unterschiedlichen Kundengruppen, für die Marketingaktivitäten optimiert werden sollen. Das langfristige Ziel ist es, den Kunden diejenigen Produkte anzubieten oder Informationen und Angebote derjenigen Produkte zukommen zu lassen, für die sie als potentielle Käufer in Frage kommen, um somit den Kundenbedürfnissen gerecht zu werden. Um diese Art der Kundenklassifikation mittels SVM umzusetzen, werden die Kunden durch soziodemografische Daten und ihre bisherige Kaufhistorie beschrieben, die eine Zugehörigkeit zu verschiedenen, durch eine Clusteranalyse identifizierten Typen vermuten lassen. Die potentiellen Zielpersonen sollen konkret ausgewählt und kontaktiert werden, um verschiedene Produkte, die den einzelnen Gruppen zuordenbar sind, mit dem größtmöglichen Erfolg zu vermarkten. Eine hohe

[11] Liegen entsprechende Informationen über die Kunden vor, so kann die folgende Auswertung auch auf die Interessen der Hersteller übertragen werden.

Generalisierbarkeit des verwendeten Modells ist daher für die korrekte Zuteilung der Direktwerbung von hohem Interesse. Der Einsatz von SVM ist in diesem Abschnitt als ein Teilschritt innerhalb des CRM zu sehen.

4.4.1 Problemstellung und Datenbeschreibung

Bei den hier vorliegenden Daten handelt es sich um Verbraucherpaneldaten, die auf dem ComsumerScan Haushaltspanel der GfK[12] basieren und von Zentrum für Umfragen, Methoden und Analysen in Mannheim (ZUMA) zur Verfügung gestellt wurden (*Papastefanou et al.* (2001)). Die Daten beinhalten Informationen über insgesamt 9064 Haushalte aus dem Jahr 1995. Zusätzlich zu den Erhebungen über die Käufe von Produkten des täglichen Gebrauchs stehen Informationen über soziodemografische Merkmale sowie Daten über die Einstellung der haushaltsführenden Personen zu Ernährung, Umwelt und Konsum zur Verfügung. Die an dem Panel teilnehmenden Haushalte wurden gebeten, ein Haushaltsbuch zu führen, in dem sie unter anderem detaillierte Informationen zu den gekauften Produkten, den genauen Kaufzeitpunkt und den Ort des Einkaufs dokumentieren sollten. Es gibt produktunabhängige Merkmale wie Datum des Kaufs, und produktspezifische Eigenschaften wie die Geschmacksrichtung oder die Sorte. Beide Arten sollen in der Untersuchung an späterer Stelle Berücksichtigung finden. Die Kaufhäufigkeiten sind je nach Warengruppe (z.B. Getränke gegenüber Reinigungsmittel) sehr unterschiedlich. Als Basis für die Erstellung des Datensatzes wurden für jeden Haushalt und für jede Warengruppe die Gesamtausgaben über das Jahr 1995 ermittelt, um so Informationen über das Kaufverhalten zu erhalten. Die Daten zur Einstellung der befragten Personen zu unterschiedlichen Themen wurden einmalig im Jahr 1995 erhoben und umfassen folgende Merkmale, die in den anschließenden Anwendungen auf verschiedene Weise in die Auswertungen mit einfließen:

- Ansichten zur Ernährung
  Hierbei handelt es sich um Items, in denen die Präferenzen beim Nahrungsmittelkauf erfasst werden (z.B. ob gesunde oder naturbelassene Produkte bevorzugt werden, in welcher Weise Kriterien wie Entdeckerfreude, Herkunft der Produkte, Vollwertkost, Frische, Hausmannskost, Markenartikel, Vitamine/Mineralstoffe oder Zubereitungskomfort etc. eine Rolle spielen).

- Ansichten zu Dingen des täglichen Lebens
  Hierzu gehören Items, mit denen das Ausmaß des Interesses an neuen Produkten, der traditionellen Lebensführung, der Erlebnisorientierung, der Qualität usw. erhoben wird.

Insgesamt wurden 61 Items aus den beiden Bereichen erhoben. Diese wurden von der GfK in einer Faktorenanalyse zu 20 Faktoren (Tabelle 4.8) zusammengefasst, die sich aus den zuvor erhobenen Einstellungsmerkmalen ergeben.

[12] Gesellschaft für Konsumforschung, Nürnberg

Nr.	Faktor	Nr.	Faktor
1	Schlankheitsorientierung	11	Pro Markenartikel
2	Medizinisch gesund	12	Pro Vitamine/Mineralstoffe
3	Naturbelassen	13	Unkritischer Ernährungsstil
4	Entdeckerfreude	14	Nostalgie
5	Pro deutsche Produkte	15	Misstrauen gegenüber Neuprodukten
6	Convenience-Orientierung	16	Convenience-orientiertes Kochen
7	Hausmannskost	17	Qualitätsorientierung
8	Vollwertkost	18	Traditionelle Lebensführung
9	Anspruchsvoll genießen	19	Innovationsneigung
10	Frische Orientierung	20	Erlebnisorientierung

Tabelle 4.8: Von der GfK erstellte Faktoren

So wird das Ausmaß der Frischeorientierung beispielsweise durch die Items

a. Ohne Fertigprodukte kann ich mir das Kochen kaum noch vorstellen

b. Bei Lebensmitteln kaufe ich ausschließlich frische Produkte anstelle von Konserven oder Tiefkühlkost.

c. Heutzutage schmecken mir Konserven genauso gut wie Frisches.

bestimmt[13]. Hohe Ausprägungen („stimme etwas zu“ oder „stimme voll und ganz zu“) bei Item b. und niedrige Werte bei Items a. und c.(„stimme eher nicht zu“ oder „stimme überhaupt nicht zu“) führen zu einer hohen Ausprägung („hoch“ oder „sehr hoch“) des Faktors Frischeorientierung.
Da in vielen Merkmalen fehlende Werte auftreten, die auf die mangelnde Bereitschaft, die Daten aufzuzeichnen, zurückgehen können, wurden die betreffenden Beobachtungen gelöscht. Weiterhin wurde aus Gründen der Vereinheitlichung nur eins von den zwei zur Verfügung stehenden Panels verwendet, um eine homogene Basis an Kaufhistorien zu gewährleisten. Daher verbleiben 4300 Haushalte, die in unterschiedlicher Art in die Untersuchungen eingehen.
Eine Möglichkeit, sinnvolle Klassenzugehörigkeiten zu gewinnen, liegt in der Bildung von Segmenten auf Basis der Einstellungsdaten, um eine Einteilung in unterschiedliche Ernährungstypen zu erhalten. Diese Klasseneinteilung wird in den Abschnitten 4.4.2 und 4.4.4 verwendet. Dazu bilden die Werte der oben beschriebenen Einstellungsdaten der am Haushaltspanel der GfK teilnehmenden Haushalte die Grundlage[14]. Es handelt sich dabei nicht um Paneldaten im typischen Sinne, sondern um Antworten zu Fragen zur Ernährung und Lebensführung, die lediglich einmal im Betrachtungszeitraum erhoben wurden. Das Ziel lag darin, möglichst

[13]Telefonische Kommunikation mit dem Zentrum für Umfragen, Methoden und Analysen, Mannheim, 2003.

[14]Eine genaue Auflistung der Items befindet sich im Anhang A.1 in Tabellen A.3 und A.4.

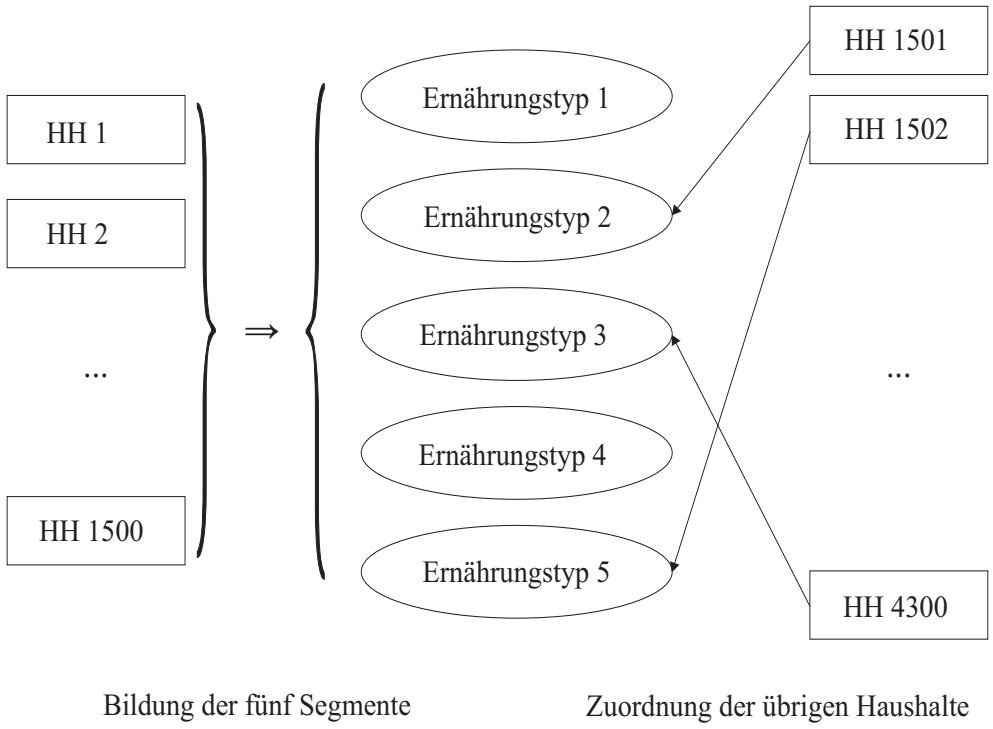

Abbildung 4.5: Visualisierung der Clusterbildung sowie Zuordnung der Haushalte

homogene Gruppen zu bilden, die sich hinsichtlich ihrer Ess- oder Kochgewohnheiten untereinander unterscheiden. Dazu werden mittels einer hierarchischen Clusteranalyse[15] Cluster erzeugt, die fünf in sich relativ homogene Gruppen bilden, die unterschiedliche Ernährungstypen repräsentieren. Es wurden lediglich 1500 der ursprünglich 4300 Beobachtungen bei der Bildung der Cluster berücksichtigt, die durch 61 erhobene, ernährungsbezogene Merkmale charakterisiert sind. Diese Zuordnung ist im linken Teil der Abbildung 4.5 verdeutlicht[16]. Die restlichen Beobachtungen werden durch Minimierung der euklidischen Distanzen den fünf Clustern zugeordnet, um Verzerrungen bei der späteren Klassifikation zu vermeiden. Dies wird durch den rechten Teil der Abbildung visualisiert. Die resultierenden Segmente unterscheiden sich sowohl hinsichtlich der Einstellungsmerkmale als auch der Struktur- und demografischen Daten, was im Folgenden kurz vorgestellt wird. In Klammern sind die Häufigkeiten bzw. Anteile der Items innerhalb der ursprünglichen 1500 Beobachtungen angegeben. Die nicht in Klammern vermerkten Angaben beziehen sich auf die Auswertung der gesamten 4300 Beobachtungen.
Das erste Cluster umfasst 433 (115) der insgesamt 4300 Kunden und kann im Gegensatz zu den übrigen Segmenten nicht eindeutig als ein typischer Käufer- oder Ernährungstyp charakterisiert werden. Eine charakteristische Eigenschaft ist

[15]Hier wurde das Ward-Verfahren zur Ermittlung der Distanzen eingesetzt.

[16]Die Bezeichnungen der Haushalte „HH 1“ bis „HH 4300“ dient lediglich der Verdeutlichung. Bei den Analysen wurden die Haushalte zufällig ausgewählt und die ursprüngliche Nummerierung nicht berücksichtigt.

jedoch die Vorliebe für neue Produkte. Im Gegensatz zum Durchschnitt (mit 19,6% (19%) Zustimmung) stimmen 43,2% (50,4%) der Haushalte zu, viele Artikel zu kaufen, die anderen noch unbekannt sind. Zustimmung wird in diesem Fall bei den Einstellungsdaten durch „stimme etwas zu“ und „stimme voll und ganz zu“ ausgedrückt. Weiterhin achtet diese Gruppe der Verbraucher auf eine fettarme und schlankheitsorientierte Ernährung (89,6% (90,4%) Zustimmung für „Ich achte auf meine Figur“ im Gegensatz zu durchschnittlichen 66% (64,7%) Zustimmung bzw. 91,5% (91,3%) Zustimmung für „Ich achte auf eine fettarme Ernährung“ im Vergleich zu 66,3% (65,4%) im Durchschnitt). Somit kann diese Gruppe als die Gruppe der schlankheitsbewussten Meinungsführer bezeichnet werden.
Das zweite Segment fällt mit 1201 (472) zugehörigen Haushalten deutlich größer aus und wird als die „traditionelle, deutsche, bürgerliche Küche“ bezeichnet. Die zugehörigen haushaltsführenden Personen sind zu 58,3% (64,2%) bereits 60 Jahre oder älter und sind zu 80,3% (81,8%) nicht (mehr) berufstätig. Im Vergleich zum Durchschnitt mit 32% (31,8%) sind in diesem Segment mit 53,2% (55,3%) sehr viele Rentner und Pensionäre vertreten. Die meist in 1-2 Personen-Haushalten lebenden Verbraucher innerhalb dieses Segments fühlen sich häufig zuhause am wohlsten. Sie treten dem Ausprobieren neuer Gerichte oder fremdländischen Spezialitäten eher verhalten entgegen und bevorzugen gewohnte und altbewährte Gerichte. Beim Kauf von altbekannten Produkte vertrauen sie auf deutsche Herkunft. Daher werden diese Beobachtungen dem „traditionellen“ Segment zugeordnet.
Da die Beobachtungen des dritten Segments eher „andere Interessen als die Küche“ haben (81,3% (76,5%) Zustimmung gegenüber durchschnittlichen 59,3% (59,4%) Zustimmung) und eine Vorliebe für einfache Gerichte deutlich an mehreren Fragen erkennbar ist, kann dieses Segment als das „Fast-Food“-Segment bezeichnet werden. Dies kann z.B. auch durch die Einstellung zum Kauf von Frischem bestätigt werden, wobei 20,3% (24,6%) dieser Aussage zustimmen, was deutlich unter dem Durchschnitt von 50% (51,6%) liegt. Die insgesamt 1120 (422) Beobachtungen dieses Segments zählen eher zu den jüngeren Panelteilnehmern, bei denen der Anteil an Kindern höher ist als im Durchschnitt und der Anteil derjenigen, die eine Mikrowelle besitzen mit 53,2% (55,5%) auch über den durchschnittlich erreichten 43,2% (43,8%) liegen. Im Vergleich zum Fast-Food-Segment, welches von der ZUMA erstellt wurde, kann hier die überdurchschnittliche Kaufkraft der betreffenden Beobachtungen allerdings nicht bestätigt werden.
Das vierte Segment umfasst insgesamt 732 (229) Haushalte, die eher preiswert einkaufen und nicht auf eine speziell gesunde Ernährung achten. Weiterhin ist in dieser Gruppe das fehlende Markenbewusstsein (Markenvorliebe) stark ausgeprägt. So ziehen lediglich 6,8% (7,5%) den Kauf von Markenartikeln vor, im Gegensatz zu durchschnittlichen 20,7% (23,0%). Die Lebensmittel bekannter und unbekannter Marken werden hinsichtlich ihrer Qualität eher als gleichwertig betrachtet. Insgesamt bildet diese Gruppe ein eher junges Segment an Verbrauchern, die zur Hälfte (49,8% (49,9%)) unter 44 Jahren sind. Da das Einkommen innerhalb dieses Segments höher als der Durchschnitt liegt, und das Kaufverhalten auf keine bestimmte Marken festgelegt ist, kann gefolgert werden, dass die Konsumenten durchaus mehr Geld in Lebensmittel einbringen könnten, als sie es zurzeit tun. Dies

Inhalt	Datenumfang			Merkmale	Klassen
	Training	Test	Anz.	Inhalt	
Direktmarketing Abschnitt 4.4.2	3225	1075	62	Ausgaben für 59 Warengruppen u. 3 demografische Merkmale	5 Ernährungstypen (vgl. Seite 141)
Multilabel-Klassifikation Abschnitt 4.4.3	999	407	34	Ausgaben für 24 Warengruppen u. 10 inhaltliche Informationen	11 Faktor-zuordnung
Merkmals-reduktion Abschnitt 4.4.4	2100	700	61	Items zur Ernährung	5 Ernährungstypen (vgl. Seite 141)

Tabelle 4.9: Beschreibung der Kaufverhaltensdaten

macht diese Gruppe für das Marketing zu einer interessanten und kaufkräftigen Gruppe, obwohl sie eher durch die Präferenz niedrigpreisiger Produkte geprägt ist und daher im Folgenden als „Preisbewusste“ bezeichnet werden.
Als die „bewussten Genießer“ kann das fünfte Segment mit insgesamt 814 (262) Beobachtungen bezeichnet werden. Die Gruppe wird durch gut verdienende, junge Kleinfamilien mit schulpflichtigen Kindern gebildet, die offen für neue Produkte und ausländische Spezialitäten sind. So stimmen 77,8% (63,3%) der Haushalte der Aussage „ich habe Spaß am Ausprobieren fremdländischer Gerichte“ zu im Gegensatz zu durchschnittlich nur 46,4% (46,1%). Die Bevorzugung von ökologisch wertvollen Lebensmitteln wird durch die Vorliebe für vegetarische Ernährung und Einsatz vollwertiger Nahrungsmittel deutlich.
Eine Varianzanalyse zur Überprüfung der Trennschärfe zwischen den Gruppen zeigt, dass sich die Gruppen bezüglich der eingehenden Einstellungsmerkmale signifikant unterscheiden.

Um diverse Facetten eines Einsatzes von SVM umzusetzen und Eigenschaften dieses Verfahren auszunutzen, werden die Merkmale und Klassenzugehörigkeiten unterschiedlich zusammengesetzt. Somit können unter anderem Multilabel-Klassifikation und die Reduktion der eingehenden Merkmale realisiert werden. Tabelle 4.9 zeigt, wie die in den folgenden Abschnitten eingesetzten Daten sich zusammensetzen. Die oben beschriebenen gebildeten Cluster spielen eine entscheidende Rolle und werden zur Identifizierung der Zugehörigkeit zu Ernährungstypen zweimal verwendet.

Die folgenden Abschnitte befassen sich mit der für die richtigen Ansprache der Kunden relevanten Zuordnung zu verschiedenen Kundensegmenten. In Abschnitt 4.4.2 steht für das Direktmarketing die Zuordnung zu den fünf oben beschriebenen Ernährungstypen auf Basis des bisherigen Kaufverhaltens im Vordergrund. Diese Zuordnung wird ebenfalls in Abschnitt 4.4.4 verwendet, wobei die Kunden dort durch ihre angegebene Einstellung zur Lebensweise und Ernährung gekennzeichnet werden und die Anzahl der in die Analyse einfließenden Merkmale reduziert werden soll. In Abschnitt 4.4.3 wird die Multilabel-Klassifikation durch die Zuordnung von Haushalten zu mehreren Klassen umgesetzt. Das Ziel der Analyse besteht in der Identifikation der Kombination von Zugehörigkeiten zu den Faktorausprägungen, sodass auf dieser Basis eine zielgerichtete Werbung realisiert werden kann.

4.4.2 Auswertung der Daten im Rahmen des Direktmarketings

Datengrundlage

Ziel der Untersuchung ist es, auf Basis des beobachtbaren Kaufverhaltens, dokumentiert durch die Ausgaben für spezielle Warengruppen, auf die Einstellung bezüglich der Ernährung zu schließen. Die hier zu beantwortende Frage ist also

Welchem Ernährungstyp gehört ein Kunde an und wie stark ist diese Zugehörigkeit ausgeprägt?

Dazu werden zur Beschreibung der Kunden die dokumentierten Einkäufe bezüglich unterschiedlicher Warengruppen herangezogen. Es wird ein Datensatz generiert, der Informationen zu den Käufen von 59 verschiedenen Warengruppen enthält, die in Tabelle A.1 in Anhang A.1 zusammengefasst sind. Diese Auswahl beschränkt sich auf Warengruppen, die auch tatsächlich von den Teilnehmern des zu untersuchenden Panels mindestens einmal gekauft wurden. Diese Auswahl der die Kunden beschreibenden Merkmale wird um die Merkmale „Haushaltsgröße“, „Alter der haushaltsführenden Person“ und „Anzahl der Kinder“ erweitert. Im Verlauf der Auswertungen zeigt sich, dass diese Merkmale mit zu den relevanten bei der Trennung der Klassen zu zählen sind.[17] Des Weiteren werden die insgesamt 4300 Beobachtungen für die Analyse pro Variable auf ein Intervall $[0, 1]$ normiert, um zu vermeiden, dass der Kauf von hochpreisigen Produkten einen höheren Einfluss bei der Berechnung der Trennebene erhält als der Kauf von billigeren Produkten.
Zwischen dem Erwerb von Lebensmittel der ausgewählten Warengruppen und der Zuweisung zu den oben beschriebenen Ernährungsclustern wird ein (linearer) Zusammenhang vermutet. So lässt zum Beispiel eine Zunahme an Ausgaben für Fertiggerichte bei gleichzeitigem Rückgang der Ausgaben für frische Artikel auf eine stärkere Zugehörigkeit zum Cluster der Fast-Food-Typen schließen.

[17] Dies wird ebenfalls durch die eingesetzte Diskriminanzfunktion bestätigt, die für die betreffenden Merkmale hohe Koeffizienten aufweist.

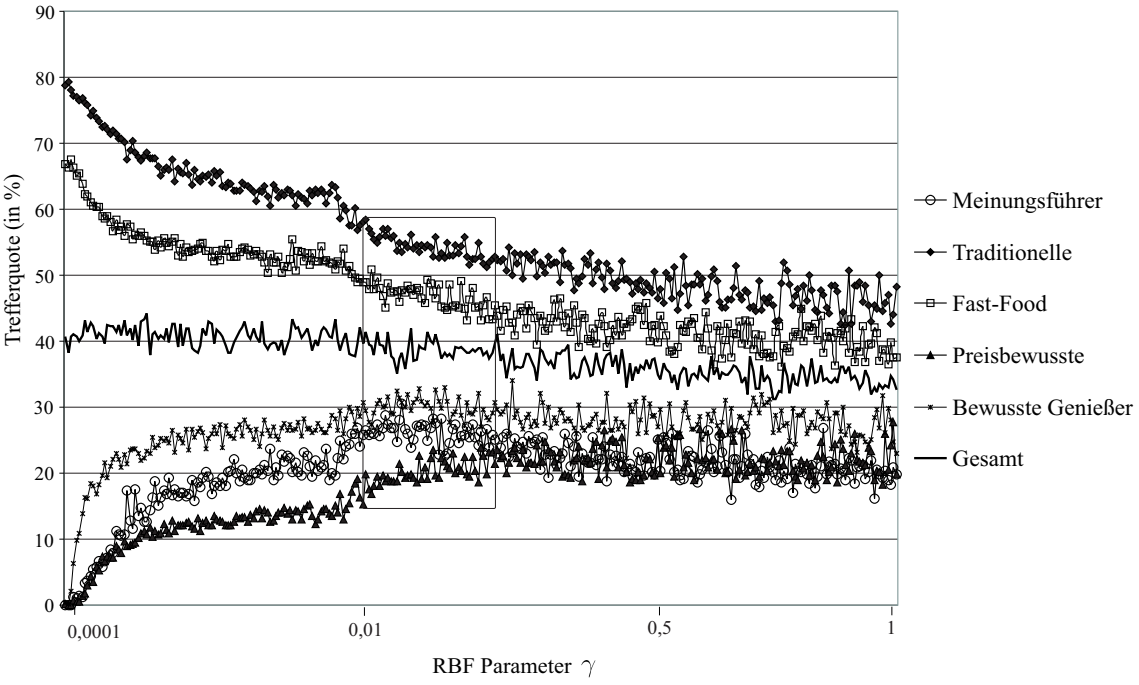

Abbildung 4.6: Ergebnisse bei 4-facher Kreuzvalidierung mit festem Kostenparameter $C = 100$ und Variation des Parameters γ

Auswertungen

Um später zu aussagekräftigen und verlässlichen Ergebnissen zu gelangen und einen Überblick über die generell zu erwartenden Trefferquoten zu erhalten, wird für die nicht lineare Trennung mittels RBF durch Gridsearch die Suche nach den optimalen Parametern durchgeführt. Abbildung 4.6 zeigt die Ergebnisse einer 4-fachen Kreuzvalidierung bei $C = 100$, wobei der Kernparameter γ von 0,0001 bis 1 variiert. Abgebildet werden die klassenindividuellen Trefferquoten, die bei Zuordnung zu den fünf Clustern erreicht werden, sowie die Gesamttrefferquote. Werte des Kernparameters γ außerhalb dieses Bereiches haben sich als nicht adäquat herausgestellt, wohingegen alternative Werte für den Kostenparameter C (wie etwa 50, 1000) zu ähnlichen Ergebnissen führen, sodass hier lediglich $C = 100$ betrachtet wird. Es zeigt sich, dass Werte zwischen 0,05 und 0,3 für den Kernparameter γ (in der Abbildung durch ein Quadrat gekennzeichnet) zu einer guten Gesamttrefferquote und akzeptablen Trefferquoten für die einzelnen Klassen führt. Bei der Aufteilung des Datensatzes in Trainings- und Testdaten bleibt die Verteilung der Klassen erhalten, wie Tabelle 4.10 zu entnehmen ist[18].

[18]Ein χ^2-Homogenitätstest zeigt, dass die beiden Datensätze Training und Test bei einem Signifikanzniveau von $\alpha = 0,05$ eine homogene Verteilung der Klassen aufweisen ($\chi^2 = 3,46 < 9,49 = \chi^2_{krit}$).

Methode/ Parameter	TQ 1	TQ 2	TQ 3	TQ 4	TQ 5	Gesamt
SVM OAO Linear						
$C = 100$	21,82%	62,41%	48,65%	19,89%	24,17%	40,00%
zus. Fuzzy	21,82%	62,06%	47,30%	21,59%	29,86%	40,93%
SVM OAA Linear						
$C = 100$	32,73%	48,23%	39,86%	28,41%	31,28%	37,77%
zus. Fuzzy	32,73%	48,23%	39,86%	28,41%	31,28	37,77%
ECOC (dense)	0,0%	81,21%	65,88%	2,27%	8,53%	41,49%
ECOC (sparse)	0,91%	79,43%	67,57%	2,84%	8,53%	41,67%
LDA	33,64%	47,52%	35,14%	27,27%	35,07%	36,93%
SVM OAO RBF						
$C = 100, \gamma = 0,085$	25,45%	61,35%	46,28%	18,75%	27,49%	39,91%
zus. Fuzzy	24,55%	59,93%	47,30%	19,89%	30,33%	40,47%
zus. DAG	24,55%	59,93%	47,30%	19,89%	30,33%	40,47%
ECOC (dense)	0,09%	77,30%	60,47%	10,80%	22,27%	43,16%
ECOC (sparse)	2,73%	72,70%	64,53%	7,95%	22,75%	42,88%
SVM OAA RBF						
$C = 10, \gamma = 0,53$	26,35%	51,06%	42,23%	27,84%	32,23%	38,70%
MLP	0,0%	61,70%	52,03%	2,84%	42,65%	39,35%
C4.5	20,91%	38,65%	29,73%	19,32%	22,75%	28,09%

Tabelle 4.11: Ergebnisse auf Basis unterschiedlicher SVM-Multiklassifikationsverfahren und vergleichbarer Methoden (nach linearem und nicht linearem Verfahren getrennt)

Datensatz	Klasse 1	Klasse 2	Klasse 3	Klasse 4	Klasse 5	Gesamt
Training	323	919	824	556	603	3226
Test	110	282	296	176	211	1075

Tabelle 4.10: Verteilung der Klassen innerhalb der Trainings- und Testdaten

Tabelle 4.11 enthält die Ergebnisse der Anwendung verschiedener Multiklassifikationsverfahren von SVM auf diesen zufällig ausgewählten Testdaten im Umfang von 1075 Beobachtungen[19] im Vergleich zu den alternativen Methoden LDA bei linearer Trennung und MLP und C4.5 bei nicht linearer Trennung. Hier werden die in Abschnitt 2.4 zusammengefassten Verfahren der Multiklassifikation verwendet, um die Leistungsstärke der einzelnen Verfahren zu dokumentieren. Im Gegensatz zur Biklassifikation erhöht sich damit die Anzahl der resultierenden Ergebnisse. Die Verfahren OAO, OAA, Fuzzy-SVM, ECOC sowie DAG (vgl. Abschnitt 2.2) kommen hier zum

[19]Das bei der Kreuzvalidierung gewählte Verhältnis 3 zu 1 wird hier beibehalten, sodass 4300/4=1075 Beobachtungen für den Testdatensatz resultieren.

Einsatz. Die Ergebnisse sind in der Tabelle nach linearer und nicht linearer Trennung geordnet und geben neben den klassenindividuellen Ergebnissen (Trefferquote für Klasse „1“ (TQ 1) bis Trefferquote für Klasse „5“ (TQ 5)) auch die insgesamt erreichte Trefferquote an. Es zeigt sich, dass die lineare Version von SVM bezogen auf die Gesamttrefferquote gegenüber dem alternativen Verfahren der LDA bessere Ergebnisse erreicht. Eine zusätzliche Anwendung der Fuzzy-SVM („zus. Fuzzy“) bewirkt zumindest bei OAO eine geringfügige Verbesserung der Trefferquoten, insbesondere der Gesamttrefferquote. Demgegenüber führt der Einsatz von ECOC sowohl mit einer dicht (dense) als auch mit einer dünn (sparse) besetzten Codematrix lediglich zu einer Erhöhung der Gesamttrefferquote, die in der drastischen Erhöhung der Anzahl der Treffer für die umfangreichen Klassen „2“ und „3“ begründet liegt. Hierbei wurden die bei LIBSVM empfohlenen ECOC-Matrizen verwendet, die im Anhang in den Tabellen A.5 und A.6 dargestellt sind. Eine inhaltliche Bedeutung der Codewörter liegt hier daher nicht vor. Das gute Abschneiden der Fuzzy-SVM wird durch vergleichbare Anwendungen in der Literatur gestützt, bei denen im Vergleich zu herkömmlicher SVM diese bessere Ergebnisse liefert (vgl. *Abe, Inoue* (2002) oder *Inoue, Abe* (2001)).
Bei der nicht linearen Trennung ist zu beobachten, dass SVM und MLP vergleichbare Ergebnisse erzielen[20]. Die Ausführungen zu der nicht linearen Trennung beziehen sich lediglich auf die Verwendung des Radialbasis Kerns, da mittels polynomiellen Kerns unter Verwendung unterschiedlicher Grade lediglich schlechtere Ergebnisse erzielt worden sind. Da die nicht lineare Trennung nicht zu einer deutlichen Verbesserung des Klassifikationsergebnisses beiträgt, kann die eigentliche Stärke der SVM hier nicht genutzt werden, was an der intuitiv einleuchtenden linearen Struktur der Daten liegt.
Es sei darauf hingewiesen, dass hier ausschließlich die gewichtete Variante von SVM verwendet wurde. Der Vergleich der klassenbezogenen Trefferquoten, die mittels SVM und MLP erreicht wurden, zeigt die Stärke der SVM, bei denen durch die zusätzliche Gewichtung schwächer besetzter Klassen eine ausgewogene Verteilung der Treffer gewährleistet werden kann. Eine Berechnung der Trennung auf Basis der Angabe von nur einem gruppenübergreifenden Kostenfaktor C würde zu der vollständigen Falschzuordnung der kleinsten Klasse führen, wie es bei MLP oder auch ECOC der Fall ist. Aufgrund der komplexen Struktur der Codematrizen kann für ECOC keine sinnvolle Gewichtung der Klassen für die gesamte Optimierung der SVM durchgeführt werden. Bei Einsatz von MLP geht die recht hohe Trefferquote zu Lasten der Genauigkeit innerhalb der kleinen Gruppen Klasse 1 und Klasse 4. Die in Tabelle 4.11 dargestellten Trefferquoten wurden auf Basis des Einsatzes gruppenindividueller Kosten erzielt. Diese sind in Tabelle 4.12 dokumentiert.

[20]Das neuronale Netz wird durch drei Hidden Units fünf Ausgabe-Units gebildet.

	Klassen				
	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
Umfang der Klassen	323	919	824	556	603
Gewichte c_k (OAA)	8,98	2,51	2,91	4,80	4,35
Gewichte c_k (OAO)	2,25	1	1,04	1,30	1,22

Tabelle 4.12: Verteilung der Klassen innerhalb der Trainingsdaten und die daraus resultierenden Gewichte

Diese Gewichte werden dazu analog zu der Empfehlung von *Morik et al.* (1999) auf die Multiklassifikation mittels OAA ausgeweitet. Demnach werden die Gewichte der Klassen den jeweiligen Verhältnissen der Klassen in den einzelnen SVM angepasst. So erhält Klasse 1 bei der OAA-Trennung ein Gewicht von 8,98 ($= \frac{3225-323}{323}$). Bei OAO ist dieses geradlinige Vorgehen nicht möglich. Da eine Klasse in mehreren SVM zum Einsatz kommt, wird hierbei das Mittel der auftretenden Verhältnisse gewählt. So ergibt sich beispielsweise das Gewicht c_1 für Klasse 1 aus den Verhältnissen der Klassenumfänge von Klasse 1 ($l^{[1]}$) verglichen mit jeweils allen übrigen Klassen ($l^{[k]}$ mit $k = 2, \ldots, 5$).

$$\begin{aligned} c_1 &= \tfrac{1}{4}(\max(1, \tfrac{l^{[2]}}{l^{[1]}}) + \max(1, \tfrac{l^{[3]}}{l^{[1]}}) + \max(1, \tfrac{l^{[4]}}{l^{[1]}}) + \max(1, \tfrac{l^{[5]}}{l^{[1]}})) \\ &= \tfrac{1}{4}(\tfrac{919}{323} + \tfrac{824}{323} + \tfrac{556}{323} + \tfrac{603}{323}) \\ &= 2,25 \end{aligned}$$

Entsprechende Auswertungen zeigen, dass diese Bestimmung der klassenindividuellen Gewichtungen, die als Kompromiss zwischen Gewichten bei der OAA-Trennung und keiner Gewichtung zu bezeichnen ist, in besseren Ergebnissen resultiert als bei fehlender Differenzierung der Gewichte oder Verwendung der Gewichte der OAA-Trennung. Dabei kann lediglich die Verteilung der Beobachtungen innerhalb der Trainingsdaten (3225 Beobachtungen) berücksichtigt werden. Es wird wiederum angenommen, dass sie eine repräsentative Stichprobe der Grundgesamtheit bilden, um die Testdaten oder unklassifizierte Beobachtungen verlässlich klassifizieren zu können. Bei der Gewichtung werden die minimalen Kosten von C beibehalten und durch die Gewichtung durch c_k lediglich angepasst.
Um die Ergebnisse mittels des Proportional-Chance-Kriteriums und des Größte-Gruppen-Kriteriums bewerten zu können, werden die Formeln von Seite 100 verwendet. Hier gilt für die Gesamttrefferquote bei beispielsweise linearer Trennung

$$TQ = 0,40 > 0,2204 = \left(\frac{110}{1075}\right)^2 + \left(\frac{282}{1075}\right)^2 + \left(\frac{296}{1075}\right)^2 + \left(\frac{176}{1075}\right)^2 + \left(\frac{211}{1075}\right)^2$$

bezüglich des Proportional-Chance-Kriteriums und

$$TQ = 0,40 > 0,2753 = \frac{296}{1075}$$

bezüglich des Größte-Gruppen-Kriteriums. Da der Einsatz beider Kriterien zeigt, dass die erzielten Ergebnisse trotz der relativ geringen absoluten Trefferquoten

als zufrieden stellend zu bezeichnen sind (vgl. Abschnitt 3.8.4), stellt der Einsatz von SVM eine gute Alternative im Rahmen der Kundenklassifikation dar. Eine lineare Struktur innerhalb der Daten ist dafür ursächlich, dass der Einsatz von nicht linearen Kernen bei der Berechnung des Modells keine deutliche Verbesserung hervorbringt. Bei der linearen Trennung (OAO), die eine relativ gute Trefferquote erreicht, gehören lediglich 5-9% der Support Vektoren zu denen, die auf den Hilfsebenen positioniert sind, und für die daher $0 < \alpha_i < C$ (bzw. $0 < \alpha_i < c_k C$) gilt. Die übrigen Support Vektoren sind entweder falsch klassifiziert, oder liegen innerhalb der beiden Hilfsebenen. Beides spricht nicht für eine gute Trennung und damit eher für die durchschnittliche Prognosegüte der Trennung.

Je nach Zielsetzung kann die Klassifikation mittels SVM noch durch den Benutzer beeinflusst werden. Denkbar ist beispielsweise eine Situation, in der ein Händler den Umsatz einer kalorienreduzierten Linie steigern möchte. Es soll nun Werbung für ein neues Produkt dieser Linie, etwa einem neuen, kalorienreduzierten Brotaufstrich auf Pflanzenbasis, gemacht werden, die speziell denjenigen Kunden präsentiert werden soll, für die dieses Produkt interessant ist. Diese zielgerichtete Kundenansprache kann im Direktmarketing durch die Versendung von Werbebriefen oder Produktproben realisiert werden. Dadurch kann die Rücklaufquote durch die Ansprache besonders responsewilliger Kunden erhöht werden. Dies sind insbesondere die Kunden aus dem ersten Segment, die zu den Meinungsführern bzw. Innovatoren zu zählen sind und sich durch den baldigen Kauf neuer Produkte auszeichnen. Weiterhin sind diese durch eine schlankheitsorientierte Ernährungsweise gekennzeichnet, was sie zu der Zielgruppe des neuen Produktes werden lässt. Desweiteren sind die Kunden des fünften Segments aufgrund ihres Interesses am Ausprobieren neuer Produkte und der Vorliebe für eine vegetarische Ernährung zur Zielgruppe zu zählen. Da diese Kunden nicht speziell auf kalorienreduzierte Artikel fokussiert sind, sondern auf eine gesundheitsbewusste Ernährung achten, steht die Richtigklassifikation zwar im Fokus der Untersuchung, aber schwächer als bei der Betrachtung des ersten Segments. Im Sinne einer Vermeidung der Informationsüberlastung und Anpassung der Werbemittel an die Kundschaft sollte neben einer Zuordnung der Segmente 1 und 5 die Richtigzuordnung des Segments 3 bei der Ausrichtung des Klassifikationsmodells berücksichtigt werden. Dies ist damit zu begründen, dass die Kunden dieses Segments für jenes Produkt mit hoher Wahrscheinlichkeit nicht als Käufer in Frage kommen, da der Ernährungsstil hauptsächlich durch den Konsum von Fast-Food und nicht speziell kalorienreduzierter Produkte gekennzeichnet ist. Daher sollte das Modell derart ausgerichtet werden, dass die Beobachtungen der Segmente 1, 3 und 5 richtig ihren jeweiligen Klassen zugeordnet werden und die Gewichte für diese Klasse entsprechend erhöht werden. Abbildung 4.7 enthält die Veränderungen der klassenabhängigen Trefferquote und die Gesamttrefferquoten in Abhängigkeit der vorgenommenen Gewichtungen bei einer linearen Trennung mit $C = 100$. Hierbei wird in der verstärkten Gewichtung nur von Klasse 1 (linker Teil von Abbildung 4.7), von Klasse 1 und 5 (mittlerer Teil) und in der Gewichtung aller interessierenden Klassen 1, 3 und 5 (rechter Teil) unterschieden. Die Gewichte werden jeweils in 0,1-Schritten verändert und

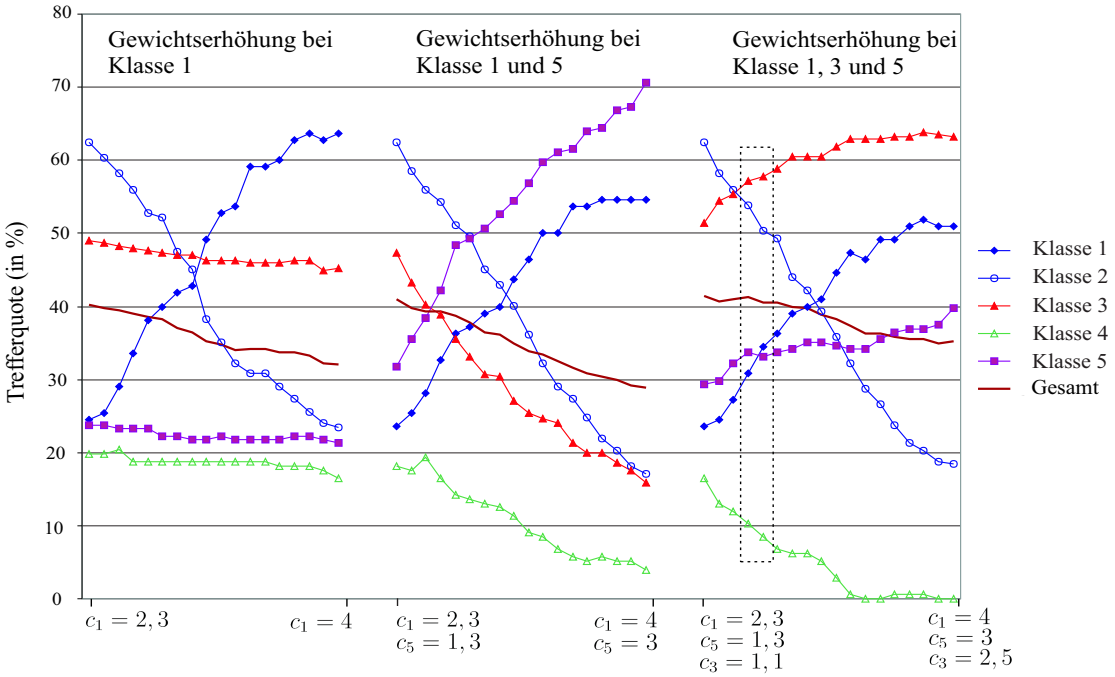

Abbildung 4.7: Veränderung der Trefferquote bei Gewichtung unterschiedlicher Klassen

liegen in den drei betrachteten Abschnitten, wie unter der Grafik angegeben, in den folgenden Intervallen:

$$c_1 \in [2,3;\ 4]$$
$$c_3 \in [1,1;\ 2,5]$$
$$c_5 \in [1,3;\ 3].$$

Diese Abbildung soll die Auswirkungen der unterschiedlichen Gewichtungen auf die klassenbezogenen Trefferquoten visualisieren, um diejenigen Beobachtungen zu identifizieren, die für eine zielgerichtete Kundenklassifikation in Frage kommen. Deutlich zu erkennen ist die Erhöhung der Trefferquote bei den jeweils besonders hervorgehobenen Klassen. So bewirkt etwa eine Erhöhung der Gewichte für die Segmente 1 und 5 (mittlerer Abschnitt in Abbildung 4.7) eine Erhöhung der Trefferquoten für diese Segmente, bei gleichzeitiger Abnahme der Anzahl der Treffer bei den übrigen drei Klassen. Die Gesamttrefferquote fällt bei allen drei Abschnitten unter das vorher erreichte Niveau. Da hier die Beobachtungen der Testdaten betrachtet werden, scheint durch die Erhöhung der Gewichte in den vorgegebenen Intervallen kein Overfitting vorzuliegen. Eine Erhöhung der Anzahl der richtig zugeordneten Beobachtungen der im vorliegenden Kontext interessierenden Gruppen geht zu Lasten der Trefferquoten der übrigen Klassen, die durch die Gewichtungen teilweise sehr stark (vgl. „TQ Klasse 2“ in allen drei Abschnitten) sinken. Steht allerdings die Richtigklassifikation der interessierenden Klassen 1, 3 und 5 im Vordergrund, kann ein Verlust der richtigen Ansprache der übrigen Kunden bei der Optimierung des Modells in Kauf genommen werden. Hinsichtlich des verfolgten Ziels, die schlankheits-

und gesundheitsbewussten Kunden richtig zu klassifizieren, um die Responserate der entsprechender Werbung zu erhöhen, sollte die Gewichtung so gewählt werden, dass neben der Erhöhung der entsprechenden klassenbezogenen Trefferquoten auch die Gesamttrefferquote berücksichtigt wird. Aus diesem Grund erscheint die Wahl des in Abbildung 4.7 durch das gestrichelte Quadrat markierten Bereichs trotz des Verlustes auf Seiten von Klasse 4 akzeptabel. Dies entspricht einer erhöhten Gewichtung der drei betreffenden Klassen von $c_1 = 2,6$, $c_3 = 1,4$ und $c_5 = 1,6$. Eine derartige Gewichtung führt im Einzelnen zu den in Tabelle 4.13 festgehaltenen Ergebnissen.

	TQ 1	TQ 2	TQ 3	TQ 4	TQ 5	Gesamt
Standardgewichtung	21,82%	62,41%	48,65%	19,89%	24,17%	40,00%
Erhöhte Gewichte	30,91%	53,90%	57,09%	10,23%	33,65%	41,30%
Veränderung	+9,09	-8,51	+8,45	-9,66	+9,48	+1,30

Tabelle 4.13: Veränderungen der Trefferquote bei Gewichtung der Klassen 1, 3 und 5

Neben der Erhöhung der Treffer in den interessierenden Gruppen bewirkt die Gewichtung ebenfalls eine Erhöhung der Gesamttrefferquote auf 41,3%.
Aufgrund unterschiedlicher Strukturen und nicht prognostizierbarer Reaktionen auf die Erhöhung einzelner Gewichte muss individuell entschieden werden, wie bei einer Klassifikation gewichtet werden soll, um die zu verfolgenden Ziele adäquat umzusetzen.

Analog zur binären Klassifikation kann eine Interpretation der Entscheidungswerte erfolgen, um die Behandlung der Kunden differenzieren zu können. Da bei der OAO-Trennung bereits 10 SVM zu bestimmen sind und demnach ebenso viele Entscheidungswerte für jede Beobachtung vorliegen, werden hier die in Abschnitt 3.7.4 verwendeten Membership-Werte herangezogen. Es ergeben sich pro Beobachtung fünf Werte, die Aufschluss über die Zugehörigkeit zu einer Klasse geben. Wird weiterhin davon ausgegangen, dass der kalorienreduzierte Brotaufstrich auf den Markt gebracht werden soll, so bilden Beobachtungen, die zu Klasse 1 zugewiesen werden, die vorrangige Zielgruppe der zu tätigenden Marketingaktivitäten. Auf diese Kundengruppe soll weiterhin der Fokus gerichtet werden. Wird eine lineare Trennung mit den bereits oben verwendeten Parametern und Klassengewichten durchgeführt, die zu den in Tabelle 4.13 dargestellten Trefferquoten führen, so resultiert die in Abbildung 4.8 verdeutlichte Verteilung der Membership-Werte, die Aussagen über die Zugehörigkeit eines Kunden zu Klasse „1“ geben. Diese Werte können nun analog zur binären Klassifikation so interpretiert werden, dass ein hoher Membership-Wert auf eine eindeutige Zuordnung zu der betreffenden Klasse hindeutet. Je kleiner dieser Wert ausfällt, desto geringer ist die Zugehörigkeit zu dieser Klasse. Ein Membership für einen Vektor $\boldsymbol{x}$ bezüglich Klasse „1“ , der größer als eins ist, kommt bei der Trennung der fünf Klassen dann zustande, wenn jeder der Entscheidungswerte $F^{[12]}(\boldsymbol{x})$, $F^{[13]}(\boldsymbol{x})$, $F^{[14]}(\boldsymbol{x})$ und $F^{[15]}(\boldsymbol{x})$ größer

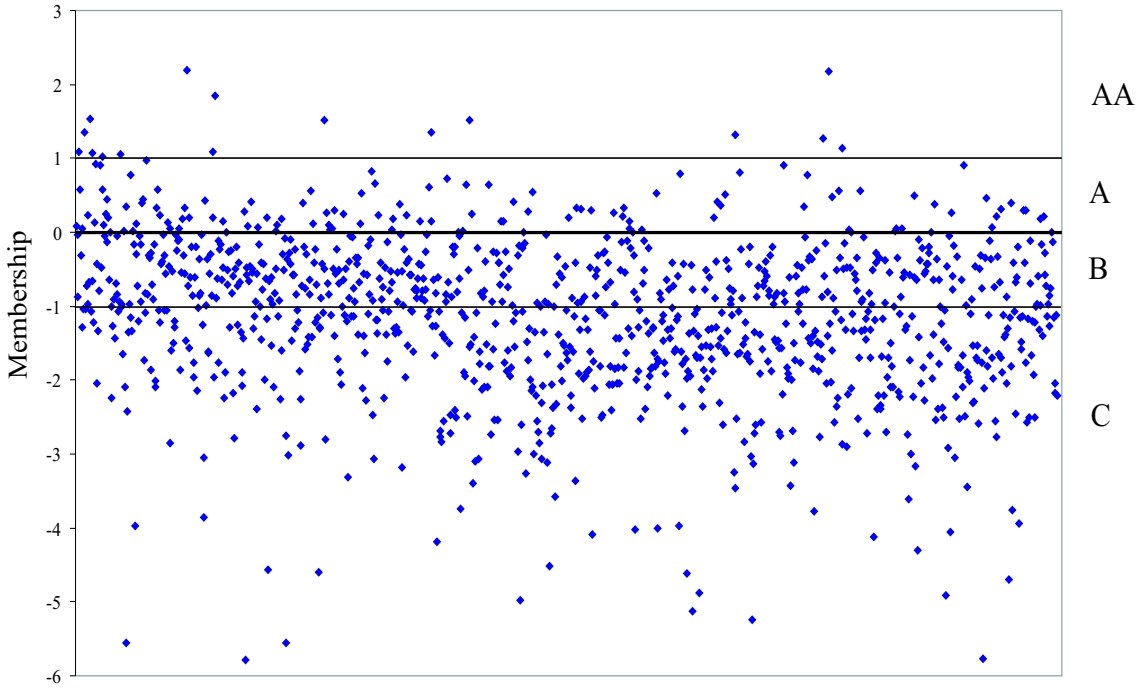

Abbildung 4.8: Darstellung der Membership-Werte bezüglich Klasse 1 bei linearer Trennung

als eins ist. Dies entspricht einer eindeutigen Zuordnung zu Klasse „1“ bezüglich aller betreffenden Entscheidungsfunktionen. Wie in Abbildung 4.8 zu erkennen ist, kann diese eindeutige Zuordnung zu dem Bereich, in dem der intensivste Einsatz von Marketingaktivitäten geplant ist, nur für sehr wenige Beobachtungen vorgenommen werden. Es werden im vorliegenden Beispiel vier Bereiche (AA, A, B und C) festgesetzt (vgl. Abbildung 4.8), innerhalb derer die Marketingaktivitäten differenziert werden. Mögliche Abstufungen in der Behandlung der Kunden können wie folgt vorgenommen werden.

C	keine Aktivität
B	die entsprechenden Kunden erhalten Werbung für das Produkt
A	die entsprechenden Kunden erhalten zusätzlich zur Werbung einen Gutschein über eine Produktprobe
AA	die entsprechenden Kunden erhalten zusätzlich zur Werbung und zum Gutschein über eine Produktprobe Rabatt beim Kauf des Produktes.

Diese Einteilung liegt in der Intensität der Zuweisung zu Klasse „1“ begründet. Bei Beobachtungen aus dem Bereich C kann davon ausgegangen werden, dass diese

eindeutig nicht zur ersten Klasse zugehörig sind. Die Kunden, die in den Bereich B fallen, weisen eine leichte Zugehörigkeit auf (wenn der Membership-Wert auch negativ ist), sodass diese Kunden zumindest Werbung für das Produkt erhalten sollten. In den Bereichen A und AA erhöht sich die Zugehörigkeit, ausgedrückt in positiven Membership-Werten. Hier sollte eine zusätzliche intensivere Aktivierung vorgenommen werden, da in diesen beiden Bereichen eine hohe Responsequote zu erwarten ist. In Tabelle 4.14 sind die resultierenden Häufigkeiten der Zuweisungen zu den einzelnen Bereichen sowie der Anteil an Beobachtungen aus Klasse „1“ enthalten.

Bereich	Anzahl Beobachtungen	davon aus Klasse 1	in %
AA	16	6	37,5%
A	115	26	22,6%
B	399	42	10,5%
C	545	36	6,6%
Summe	1075	110	–

Tabelle 4.14: Anzahl und Anteil der zugewiesenen Beobachtungen zu den einzelnen Bereichen auf Basis der Membership-Werte

Der größte Anteil der gesamten Beobachtungen wird dem Bereich C zugewiesen, was in dem insgesamt geringen Anteil (etwa 10,2%) an Beobachtungen begründet ist, die eine wahre Klassenzugehörigkeit zu Klasse „1“ aufweisen.
Da in den Bereichen A und AA deutlich mehr Geld in die Betreuung der Kunden investiert wird, sollte die zu erwartende Erfolgsquote entsprechend hoch sein. Anhand der letzten Spalte aus Tabelle 4.14 ist zu erkennen, dass der Anteil der Beobachtungen aus Klasse „1“ mit wachsender Zugehörigkeit zu dieser Klasse ebenfalls steigt. In Bereich C liegt ein unterdurchschnittlicher Anteil dieser Kunden von 6,6% vor, der bis auf 37,5% in Abschnitt AA ansteigt. Dies rechtfertigt die vorgenommene Einteilung und zeigt, dass eine differenzierte Kundenansprache auf Basis der ermittelten Membership-Werte eine Erfolg versprechende Alternative zur Gleichbehandlung aller Kunden bildet.
Sollen neben der Fokussierung auf Klasse „1“ ebenfalls die Klassen „5“ und „3“ bei der Entwicklung der Marketingstrategien berücksichtigt werden, so müssen die beiden entsprechenden Membership-Werte herangezogen werden. Falls unterschiedliche Strategien entwickelt werden sollen, muss dies in Abstimmung mit den hier durchgeführten Bereichseinteilung vorgenommen werden. Dies bedeutet, dass bei der expliziten Bestimmung der Marketingaktivitäten für einen Kunden Prioritäten gesetzt werden müssen. So kann es beispielsweise vorkommen, dass aufgrund der nicht eindeutigen Trennung einem Kunden sowohl für Klasse „1“ als auch für Klasse „3“ ein geringer, positiver Membership-Wert zugewiesen wird. Da die Informationsüberlastung bei Kunden aus Klasse „3“ vermieden werden sollte, stehen diese beiden Werte bzw. die daraus resultierenden möglichen Strategien im Widerspruch zueinander, da bereits bei einer geringen Zugehörigkeit zu Klasse

„1“ Werbung sowie eine Produktprobe versendet wird (vgl. Seite 152). Angesichts der Bedeutsamkeit der Klasse „1“ für die erfolgreiche Einführung des kalorienreduzierten Brotaufstrichs auf dem Markt, sollte die Versendung der Werbung dem Nichtwidmen dieses Kunden vorgezogen werden.
Das Beispiel der Kundenfokussierung bei der Werbung für einen neuen kalorienreduzierten Brotaufstrich hat gezeigt, wie Marketingstrategien zur differenzierten Behandlung des Marktes unter Berücksichtigung der vorliegenden Einstellung der Kunden mit Hilfe von SVM entwickelt werden können. Dazu kann zur gezielten Richtigklassifikation die Erhöhung der jeweiligen Kostenparameter herangezogen werden.

Immer dann, wenn die Daten wie im vorliegenden Beispiel nicht vollständig vorliegen, sondern stetig erweitert werden, bietet sich eine Umstellung des Systems auf das Online Learning (vgl. Abschnitt 3.4) an. In diesem Anwendungsbeispiel können die Daten neben der Panelerhebung auch durch ein Kundenkartensystem erhoben worden sein, sodass die Datenbasis bei neuen Kartenanträgen um neue Kunden erweitert wird, sobald diese durch den Kauf von (beworbenen) Lebensmitteln die Zugehörigkeit zu den hier gebildeten Ernährungsgruppen erkennen lassen. Allerdings sei hier angemerkt, dass sich der Einsatz von Online-SVM nur dann lohnt, wenn zu erwarten ist, dass viele der neu hinzukommenden Beobachtungen richtig den a priori definierten Klassen zuzuordnen sind. Nur dann kann ein erneutes Training der SVM bei einer neuen Beobachtung vermieden und somit Rechenzeit durch die in Abschnitt 3.4 erläuterte spezielle Vorgehensweise eingespart werden. Im vorliegenden Fall ist daher ein Online-Training aufgrund der vielen Fehlklassifikationen nicht uneingeschränkt empfehlenswert. Der Vorteil des Online-Learnings liegt in der schnelleren Anpassung der Ebene an die Daten, was hier nicht gut ausgenutzt werden kann. Die insgesamt eher mittelmäßige, aber dennoch akzeptable (vgl. Seite 148) Trennung der Verfahren führt zu einer nur geringen Anzahl von eindeutig zuordenbaren Vektoren mittels Fuzzy-SVM. Aufgrund der ähnlichen Leistung unterschiedlichster Verfahren ist das in diesem Fall wohl darauf zurückzuführen, dass die Daten nur bedingt für die hier zugrunde liegende Fragestellung geeignet sind.
Ein weiteres Problem in der vorliegenden Anwendung kann die Zuordnung der Beobachtungen zu den Klassen sein, die unter Umständen nicht eindeutig erfolgen muss. Stellt sich heraus, dass Kunden aufgrund ihres Kaufverhaltens zu mehreren Gruppen zuzuordnen sind, so bietet sich die Analyse mittels Multilabel-Klassifikation an, die im folgenden Abschnitt thematisiert werden soll.

4.4.3 Erweiterung auf Multilabel-Klassifikation

Die bisherigen Untersuchungen haben gezeigt, dass sich die Daten auf Basis der bereits vorgegebenen Ernährungscluster nicht perfekt trennen lassen und somit eine Prognose des Kaufverhaltens nur bedingt möglich ist. Dies kann darin begründet sein, dass ein Kunde nicht eindeutig genau einem Ernährungscluster zugewiesen

werden kann, sondern vielmehr mehreren Typen gleichzeitig ähnelt und daher eine Zuordnung zu genau einer Klasse die gegebene Situation nicht adäquat beschreibt. Daher soll im Folgenden eine alternative Vorgehensweise angewendet werden, bei der die modifizierte Multiklassifikation entsprechend Abschnitt 3.6 eine Zuweisung zu mehr als einer Klasse ermöglicht.

Datengrundlage

Bei dieser Anwendung werden wiederum die durch das Haushaltspanel erhobenen Daten eingesetzt. Die eingehenden Merkmale gehen in leicht abgeänderter Form in die Optimierung ein. Um die Konsumenten zu beschreiben, werden diese durch ihr Kaufverhalten bezüglich der Warengruppen charakterisiert, die bereits in Abschnitt 4.4.2 verwendet wurden. Die Ausgaben pro Warengruppe werden kritisch auf deren Aussagefähigkeit bezüglich des typischen Kaufverhaltens überprüft und notfalls von der Untersuchung ausgeschlossen. So leistet beispielsweise die Variable, die Auskunft über die Ausgaben für Mineralwasser gibt, keinen entscheidenden Beitrag zur Diskriminierung der Klassen. Zusätzlich zu den Ausgaben für bestimmte Warengruppen werden Angaben über kalorienreduzierte Artikel, Fertiggerichte, lose oder abgepackte Salate, sowie Informationen zu vollwertigen Lebensmitteln herangezogen. Eine vollständige Liste der eingehenden 34 Merkmale ist im Anhang in Tabelle A.2 zu finden.
Das Ziel dieses Abschnittes ist die Anwendung der Multilabel-Klassifikation im Rahmen der Kundenklassifikation, um auch die Zuweisung zu nicht klar voneinander abgrenzbaren und nicht disjunkten Gruppen bearbeiten zu können. Neben der bereits in Abschnitt 3.6 genannten Anwendung bei Recommender-Systemen lassen sich im betriebswirtschaftlichen Kontext weitere Beispiele der Anwendung von Multilabel-Klassifikation finden. So können etwa Kundensegmente vorliegen, die keine eindeutige Zuordnung der Kunden zu diesen Segmenten ermöglichen, sodass neben der Zuordnung zu beispielsweise dem „modernen“ Segment ebenfalls eine Zuordnung zum Segment der „Sparsamen“ vorliegen kann. Derartige Daten, die unterschiedliche, sich überschneidende Segmente beschreiben, waren nicht verfügbar. Da allerdings reale Situationen und die daraus resultierende Vorgehensweise und die Leistungsfähigkeit der SVM abgebildet werden sollen, wird hier eine alternative Möglichkeit herangezogen, die in realen Anwendungen aufgrund anderer zur Verfügung stehender Informationen eher weniger relevant ist. Dazu werden die Klassenzugehörigkeiten durch die Ausprägungen von Faktorwerten bestimmt. Die Klassenbildung wird auf Basis der Werte der durch die ZUMA ermittelten Faktoren (vgl. Tabelle 4.8) durchgeführt, sodass aufgrund der eingeschränkten Verfügbarkeit geeigneter Daten die Faktoren die zu trennenden Klassen bilden. Demnach stehen insgesamt 20 Klassen zur Verfügung. Diese Erweiterung auf die Faktoren wird vorgenommen, um die Situation in der Multilabel-Klassifikation aufgrund einer größeren Auswahl an Klassen besser abbilden zu können. Eine Beobachtung wird einer Klasse zugeordnet, wenn die jeweilige Faktorausprägung sehr hoch ist. Die Ausprägungen der Faktoren nehmen Werte zwischen 1 („sehr gering“) bis 5 („sehr hoch“) an. So bedeutet etwa eine 5 beim Faktor „Innovationsneigung“, dass der ent-

Faktor (Klasse)	Bezeichnung	Häufigkeit
1	Convenience-orientiertes Kochen	131
2	Qualitätsorientierung	164
3	Schlankheitsorientierung	112
4	Medizinisch gesund	225
5	Naturbelassen	215
6	Convenience-Orientierung	124
7	Vollwertkost	36
8	Anspruchsvoll genießen	95
9	Frische-Orientierung	150
10	Pro Vitamine	23
11	Unkritischer Ernährungsstil	131

Tabelle 4.15: Für die Multilabel-Klassifikation resultierende Klassen und die Häufigkeit ihrer Zuweisung

sprechende Haushalt bzw. die haushaltsführende Person sehr gerne neue Produkte ausprobiert oder neue Produkte eher kauft als Bekannte es tun. Wenn hier ein hoher Wert (Ausprägung 4 oder 5) vorliegt, so wird der Haushalt diesem Faktor, bzw. der Klasse „Innovationsneigung“ zugeordnet. Diese Klasse wird nun durch diejenige Käufer gebildet, die gerne und frühzeitig neue Produkte ausprobieren und immer auf der Suche nach Produkten sind, die ihren Bedürfnissen entsprechen. Die übrigen Faktoren werden so auch entsprechend ihres Inhalts interpretiert. Eine Zuweisung eines Haushaltes zu einer Klasse ist demnach als Zuweisung eines wahrscheinlich hohen Faktorwertes des entsprechenden Faktors zu verstehen. Da die hohen Faktorwerte bei mehreren Faktoren gleichzeitig auftreten können und demnach ein Haushalt in mehrere Klassen gleichzeitig eingeordnet werden würde, kann die Multiklassifikation an dieser Stelle auf die Multilabel-Klassifikation ausgeweitet werden, bei der die Zuweisung von einzelnen Beobachtungen zu mehreren Klassen gleichzeitig erlaubt ist. Da nicht alle der 20 Faktoren hinsichtlich der zur Verfügung stehenden Merkmale einen Zusammenhang zum dokumentierten Kaufverhalten vermuten lassen, aber eine sinnvolle Interpretation der Ergebnisse möglich bleibt, wird diese Anzahl auf 11 Klassen reduziert. Die ausgewählten Faktoren sind in Tabelle 4.15 festgehalten. Die letzte Spalte gibt die Häufigkeit an, mit der eine Zuweisung zu diesem Faktor vorgenommen wird. Insgesamt werden 1406 Beobachtungen realisiert, die faktisch 401 Haushalten entsprechen. Es wurden nur Beobachtungen berücksichtigt, die mindestens drei der elf Klassen zugewiesen werden, um die Situation bei Multilabel-Klassifikation besser abzubilden. Die Datengrundlage wurde also entsprechend den Anforderungen der Multilabel-Klassifikation modifiziert und interpretiert. In realen Anwendungen stehen einem Unternehmen möglicherweise Lifestyletypologien oder Ähnliches zur Verfügung, die teilweise ebenso im Rahmen der Multilabel-Klassifikation eingesetzt werden können.
Ziel dieser Klassenbildung ist der Ausbau der Möglichkeit zur gezielten Versendung von Werbung für unterschiedliche Produkte, die Kunden aus mehreren Kun-

Klasse	Bezeichnung	mögliche Produkte
3	Schlankheitsorientiert	Reformhausprodukte,
4	Medizinisch gesund	diätische Lebensmittel (LM),
10	Pro Vitamine	LM mit Zugabe von Vitaminen
1	Convenience-orientiertes Kochen	Fast Food,
6	Convenience-Orientierung	(Tiefkühl-)Fertiggerichte
11	Unkritischer Ernährungsstil	
2	Qualitätsorientierung	Natur-/Bio-Produkte,
5	Naturbelassen	(ausgewiesene Vitaminzugabe)
10	Pro Vitamine	

Tabelle 4.16: Beispielhafte Kombinationen der vorliegenden Klassen

densegmenten ansprechen, also nur eine bestimmte Zielgruppe bedienen. Die 11 Klassen können durch die Klassifikation von Beobachtungen durch trainierte SVM unterschiedlich kombiniert werden, sodass verschiedene Muster entstehen. Mögliche Kombinationen der vorliegenden Klassen im Rahmen der Multilabel-Klassifikation enthält Tabelle 4.16. Hier sind zusätzlich mögliche Produkte angegeben, die für die Haushalte, bei denen die entsprechenden Klassenzuweisungen auftreten, interessant sein könnten. Diese Auswahl orientiert sich an den Zusammenstellungen der jeweils kombinierten Klassen. Bei der ersten möglichen Kombination, die in der Zuweisung zu den Klassen „Schlankheitsorientiert“, „medizinisch gesund“ und „pro Vitamine“ besteht, kann davon ausgegangen werden, dass damit ein gesundheits- und figurbewusster Kunde charakterisiert wird. Demnach können bei Auftreten dieser Kombination diätische Lebensmittel oder Produkte mit ausdrücklicher Zugabe von Vitaminen empfohlen werden. Vermieden werden sollte z.B. eine Werbung für Lebensmittel aus dem Fast-Food-Bereich. Die Zuweisung eines Kunden zu den Klassen „Convenience-orientiertes Kochen“, „Convenience-Orientierung“ und „Unkritischer Ernährungsstil“ könnten hingegen dazu dienen, den betreffenden Kunden mit Werbung von dazu passenden Produkten, wie etwa neue Sorten Tiefkühlpizza oder andere neue Fertiggerichte, zu kontaktieren oder Sonderpreise innerhalb der betreffenden Warengruppen zu gewähren. Das Ziel einer derartigen Anwendung der Multilabel-Klassifikation liegt in der Erhöhung des Abverkaufs durch gezieltes Ansprechen der relevanten Zielgruppe. Es soll vermieden werden, die Kunden durch überflüssige Werbung zu überfrachten und möglicherweise sogar vom Kauf abzuhalten. Dabei kann eine Nichtzuweisung zu manchen Klassen ebenfalls von Belang sein. Das Nichtzuweisen einer Beobachtung zu beispielsweise der Klasse „Naturbelassen“, verbunden mit einem sehr niedrigem Entscheidungswert führt dazu, dass eher Fertiggerichte und nicht speziell vitaminreiche Produkte beworben werden sollten, um die Erfolgsquote einer solchen Werbeaktion zu erhöhen. Somit umfassen die folgenden Analysen die Beantwortung der Frage

Durch welche Kombination an Klassenzuweisungen ist ein Kunde geprägt und welche Kaufempfehlungen können ihm auf dieser Basis ausgesprochen werden?

Die Kombinationen an Zuweisungen ergeben sich aus den jeweiligen Klassifikationen und können sehr unterschiedlich sein.

Auswertung

Bei der Auswertung durch SVM wird nun mittels OAA-Verfahren jede der 11 Klassen von den jeweils übrigen getrennt, um somit zu einer Multiklassifikation mit Multilabel-Klassifikation zu gelangen. Manche Beobachtungen werden von der Untersuchung ausgeschlossen, falls bei der Trennung einer Klasse Beobachtungen aus dieser Klasse ebenfalls anderen Klassen zugewiesen worden sind. Diese Beobachtungen werden bei der Optimierung der entsprechenden SVM nicht berücksichtigt (vgl. *Boutell et al.* (2003)), um eine sinnvolle Trennung zu ermöglichen. So würde die Beobachtung $\boldsymbol{x}$ mit $\boldsymbol{y} = (1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0)$ bei der Trennung von Klasse „1“ vom Rest nur in Klasse „1“ auftreten, nicht aber im Rest, obwohl die Zugehörigkeit zu Klasse „3“ vorliegt. Bei der Trennung von Klasse „3“ vom Rest wird entsprechend andersherum verfahren. Dennoch führt die OAA-Trennung dazu, dass die zu trennende Klasse durch einen weitaus geringeren Klassenumfang gegenüber dem Rest gekennzeichnet ist. Dies tritt insbesondere bei der Trennung von vielen Klassen auf. Bei etwa gleich großen Klassen ergibt sich somit ein Verhältnis von „1 zu 10“ bei 11 zu analysierenden Gruppen. Dieses Missverhältnis wird zunächst unabhängig vom einzusetzenden Kern durch eine entsprechende Gewichtung der Klassen ausgeglichen.
Wie in vorangegangenen Auswertungen wird eine lineare Trennung und eine nicht lineare Trennung mittels Radialbasis-Kerns vorgenommen. Die Ergebnisse werden hier mittels der Genauigkeit, ermittelt durch $Prec$ (vgl. Seite 80), und durch die durchschnittlich erreichte Hamming-Distanz angegeben. Die herkömmliche Trefferquote ist hier nicht anwendbar. Unter Einsatz eines linearen Kerns mit $C = 50$ und den Umfängen der Klassen entsprechender Gewichtung der Gruppen wird eine Genauigkeit von $Prec = 0,356$ erreicht. Die Verwendung einer nicht linearen Trennung mittels eines RBF-Kerns[21] führt zu $Prec = 0,351$. Diese geringe Abweichung und das leicht bessere Abschneiden des linearen Kerns zeigt, dass bei den vorliegenden Daten wiederum ein linearer Zusammenhang zwischen den Merkmalen und den Klassenzuweisungen angenommen werden kann, sodass sich hier der Einsatz von SVM auf die lineare Form beschränken lässt.
Um die Güte dieser Ergebnisse bewerten zu können, wird hier ein Wert herangezogen, der analog zum Größte-Gruppen-Kriterium berechnet werden kann. Dazu wird jeder Beobachtung das häufigste Muster zugeordnet. Als ein Muster wird an dieser Stelle die Kombination von Klassenzuweisungen verstanden. Das am häufigsten vertretene Muster ist in diesem Fall die Zuweisung zu den Klassen „Medizinisch Gesund“, „Naturbelassen“ und „Frische Orientierung“. Dies resultiert in einer Genauigkeit von 0,316 und einer durchschnittlichen Hamming-Distanz von 4,49. Dies bedeutet, dass bei der trivialen Vorgehensweise die Vektoren der prognostizierten

[21] Die hierbei eingesetzten Parameter sind $C = 100$ und $\gamma = 0,004$.

Klassen von den wahren Werten in durchschnittlich 4,49 von 11 Stellen voneinander abweichen, was durch den Einsatz von SVM übertroffen wird. Bei der linearen Trennung wird eine durchschnittliche Hamming-Distanz von 3,87 erreicht[22]. Die nur geringfügige Verbesserung gegenüber einer trivialen Zuweisung der Klassen kann mehrere Gründe haben. Die vorliegenden Daten basieren auf Einstellungen, die die jeweils haushaltsführende Person aufweist. Die verwendeten Merkmale beschreiben demgegenüber jedoch die Ausgaben eines Haushaltes bezüglich der betrachteten Warengruppen, also die Einkäufe mehrerer Personen. Hierbei ist ein unterschiedliches Kaufverhalten der Haushaltsmitglieder durchaus vorstellbar, sodass der Zusammenhang zwischen dem dokumentierten Kaufverhalten und den aufgezeichneten Einstellungsdaten nicht notwendigerweise vorliegt. Zwar kann eine positive Einstellung gegenüber naturbelassenen Produkten angegeben worden sein, dies muss allerdings nicht zwangsläufig durch das Kaufverhalten bestätigt werden, was ein weiterer Grund für die weniger gute Erkennung der Zugehörigkeiten sein kann. Die Vielfalt der auftretenden Muster kann ebenfalls dazu beitragen. In den Trainingsdaten gibt es 120 verschiedene Muster, die vom Modell erkannt werden müssen.

Die durch die SVM generierten Entscheidungswerte können bei der Ausrichtung der durchzuführenden Marketingaktivitäten hilfreich sein. In Abbildung 4.9 werden dazu die berechneten Werte der Entscheidungsfunktion für die 11 Klassen mittels Parallelkoordinaten visualisiert.

[22] Je kleiner die Hamming-Distanz wird, desto besser werden die Beobachtungen den verschiedenen Klassen zugewiesen. Also weist ein kleiner Wert hier auf eine bessere Klassifikation hin.

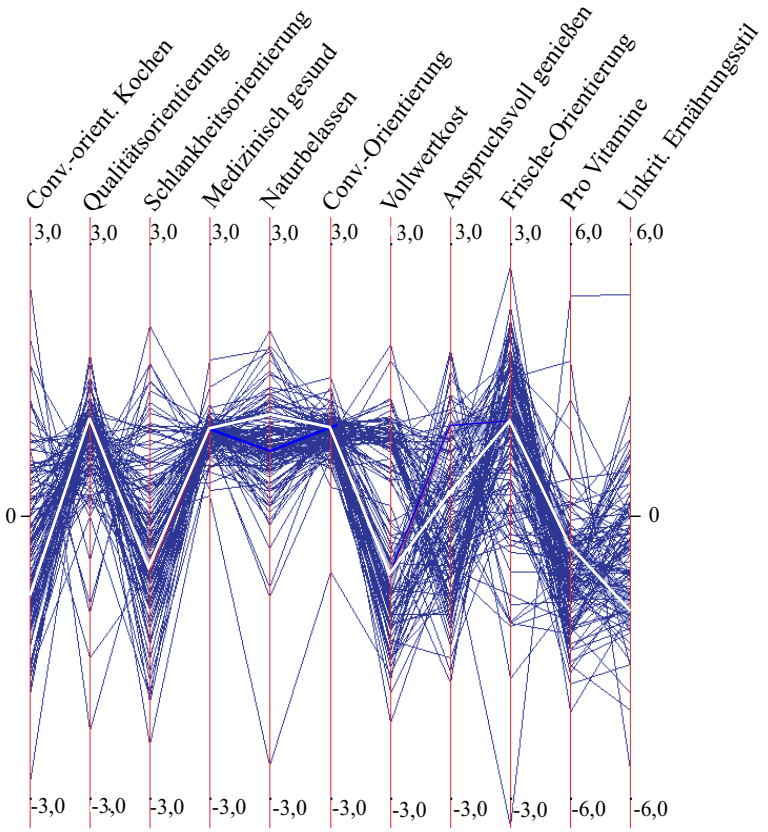

Abbildung 4.9: Visualisierung der Entscheidungswerte bei Multilabel-Klassifikation mittels Parallelkoordinaten

Jede Parallelkoordinate repräsentiert das Ergebnis bezüglich einer der 11 Klassen[23]. Die Schnittpunkte der Linien, die die Haushalte repräsentieren, mit den senkrechten die 11 Klassen repräsentierenden Linien, werden durch die Ausprägungen der Entscheidungswerte bezüglich der einzelnen Klassen bestimmt. Es ist deutlich zu erkennen, dass die Kombination der Zuweisung zu der Klasse „2“ („Qualitätsorientierung“) sehr häufig mit der Nichtzuweisung zu Klasse „1“ auftritt („Convenience-orientiertes Kochen“). Dies wird durch die entsprechenden hohen bzw. niedrigen Entscheidungswerte deutlich und ist inhaltlich nachvollziehbar: ein Kunde, der gesteigerten Wert auf ein schnelles Gericht legt, nimmt häufig Abstriche bei der Qualität zu Gunsten der Einfachheit beim Kochen in Kauf. Der in Abbildung 4.9 weiß markierte Haushalt erhält durch den Einsatz von SVM das Muster $\{2, 4, 5, 6, 9\}$, was der Zuweisung zu den Klassen „Qualitätsorientierung“, „Medizinisch gesund“, „Naturbelassen“, „Convenience-Orientierung“ und „Frische Orientierung“ entspricht. Der sehr geringe Wert hinsichtlich der beiden Klassen „Convenience-orientiertes Kochen“ sowie „Unkritischer Ernährungsstil“ macht deutlich, dass der vorliegende Kunde trotz der Zuweisung zur Klasse der Convenience-orientierten Kunden Wert auf eine gesunde, vollwertige Ernährung legt.

[23]Dargestellt werden die Ergebnisse der eingesetzten Testdaten.

Dementsprechend können diesem Kunden seinem Kaufverhalten entsprechende Produktangebote, z.B. Produkte mit dem Biosiegel oder ähnliches, unterbreitet werden.

Mit Hilfe der Entscheidungswerte können neben der Kombinationen von Klassenzuweisungen auch die Stärke der jeweiligen Zuordnung ermittelt werden und somit die Ansprache der Kunden an die jeweilige Intensität der Zugehörigkeit zu einer Klasse angepasst werden. Es wird analog zu dem in Abschnitt 3.7.4 vorgestellten Vorgehen verfahren, bei dem durch positive Membership-Werte ebenfalls Zugehörigkeiten einer Beobachtung zu mehreren Klassen auftreten können. Die Methodik und die Möglichkeit zur Umsetzung der Multilabel-Klassifikation sollte bei dieser Anwendung fokussiert werden, wohingegen die eher als schlecht zu bewertende Klassifikationsgüte im Hintergrund stand.

4.4.4 Auswahl relevanter Merkmale

Ziel dieses Abschnitts ist wiederum die Prognose von Zugehörigkeiten zu verschiedenen Ernährungstypen. Die Basis für die Klassenvariable bilden die bereits in Abschnitt 4.4.1 beschriebenen fünf Ernährungscluster. Die die Einstellung zu unterschiedlichen Ernährungsfragestellungen beschreibenden Variablen sollen als charakterisierende Merkmale in die Untersuchung eingehen. Aufgrund besserer und schnellerer Erhebung dieser Einstellungsmerkmale sollte ein möglicher Fragenkatalog nicht zu umfangreich sein. Demnach liegt das Ziel der Ausführungen dieses Abschnittes darin, die Merkmale auf wenige zu beschränken und dennoch eine akzeptable Generalisierungsfähigkeit zu gewährleisten. Ein mögliches Vorgehen liegt in der sukzessiven Überprüfung einzelner Merkmale auf ihre Diskriminierungsfähigkeit, wie es etwa in *Viaene et al.* (2001) verfolgt wird. Statt dieser sehr aufwändigen Vorgehensweise sollen im Folgenden unterschiedliche Möglichkeiten zur Reduktion der Merkmale diskutiert werden, die in Abschnitt 3.5 vorgestellt wurden.

Datenbeschreibung und Problemstellung

Zur Charakterisierung der Kunden werden die Einstellungsmerkmale herangezogen (vgl. Tabelle 4.9), die zur Zuordnung neuer Kunden zu den fünf Ernährungstypen ebenfalls erhoben werden müssen. Um den Umfang einer solchen Erhebung zu minimieren, ist es von Interesse, lediglich die für eine erfolgreiche und treffsichere Zuordnung zu den Ernährungsclustern wichtigen Merkmale auszuwählen. Der vorrangige Zweck dieser Untersuchung liegt demnach neben der Gewährleistung einer akzeptablen Generalisierbarkeit des zu berechnenden Modells in der Reduzierung der Merkmale, um somit auf Basis einer reduzierten Merkmalsauswahl eine Entscheidungsunterstützung bei der richtigen Behandlung von Kunden bieten zu können.
In Abschnitt 4.4.1 wurde die Klassenbildung auf Basis von 1500 Beobachtungen beschrieben, die in den fünf Gruppen „schlankheitsbewusste Meinungsführer“, „traditionell, deutsch bürgerliche Küche“, „Fast-Food“, „Preisbewusste“ und „bewusste

Methode	TQ 1	TQ 2	TQ 3	TQ 4	TQ 5	Gesamt
Linear						
$C = 1$	88,34%	92,97%	92,95%	**91,54%**	90,59%	92,14%
$C = 0,5$; gewichtet	90,11%	92,23%	93,46%	90,76%	**92,19%**	92,0%
RBF						
$\gamma = 0,0056$, $C = 0,5$	88,29%	**95,67%**	92,49%	87,88%	91,52%	93,0%
$\gamma = 0,00073$, $C = 5$	**91,72%**	93,63%	**94,11%**	88,26%	92,09%	**94,57%**
Polynom						
$d = 4$, $C = 50$	91,12%	89,86%	92,73%	89,83%	88,56%	92,71%

Tabelle 4.17: Ergebnisse bei Anwendung unterschiedlicher Kerne mit dem OAO-Verfahren

Genießer“ resultiert. Um Verzerrungen durch hohe Korrelationen zwischen den Merkmalen und den Klassen zu vermeiden, werden in diesem Abschnitt lediglich die verbleibenden 2800 Beobachtungen aus der Datenbasis verwendet. Dies wird durch den rechten Teil der Abbildung 4.5 verdeutlicht. Die Beobachtungen werden auf Basis der euklidischen Distanzen den fünf Ernährungstypen zugeordnet. Alle Beobachtungen werden durch 61 Merkmale beschrieben, die in den Tabellen A.3 und A.4 in Anhang A.1 dokumentiert sind. Die erforderlichen Daten werden somit durch die Paare $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^{61} \times \{1, 2, 3, 4, 5\}$ mit $i = 1, \ldots, 2800$ für die Analyse mittels SVM gestellt. Im Zentrum dieser Untersuchung steht demnach die Zuordnung von Kunden zu den unterschiedlichen Ernährungstypen auf Basis möglichst weniger Merkmale.
Der praktische Nutzung dieser Vorgehensweise liegt in der Reduzierung der für eine Klassifikation zu erhebenden Merkmale, was bei einem großen Reduktionsanteil zu erheblichen Kosten- und Zeiteinsparungen bei der Erhebung neuer Beobachtungen führen kann. Weiterhin sollte nach *Berekoven et al.* (2004) zur Erhöhung des Rücklaufs ein Fragebogen unter anderem so kurz wie möglich sein, sodass hier die Auswahl der für die Klassifikation wichtigsten Merkmale vorgenommen wird. Für die Modellbildung wären immer noch alle 61 Merkmale nötig, die Klassifikation neuer Beobachtungen, also die Befragung neuer Kunden, kann allerdings auf Basis einer deutlich reduzierten Menge von Merkmalen durchgeführt werden.

Auswertung

Mittels Kreuzvalidierung werden die optimalen Parameterwerte analog zu dem in Abschnitt 4.1 beschriebenen Vorgehen ermittelt. Tabelle 4.17 zeigt, dass sich sowohl mit einem linearen Kern als auch unter Einsatz von Radialbasis-Funktionen eine maximale Trefferquote von mindestens 92% erreichen lässt, sodass in diesem Fall zunächst wieder auf den Einsatz von linearen SVM zurückgegriffen werden kann. Die Abweichungen der Trefferquoten der einzelnen Klassen sind akzeptabel. Der Vergleich unterschiedlicher Multiklassifikationsmethoden ergab, dass die OAO-Multiklassifikation zu den besseren Ergebnissen gelangt. Hinsichtlich einer

eingesetzten Gewichtung erfolgt keine deutliche Verbesserung der Trefferquote, sodass die in der Tabelle 4.17 zusammengefassten Resultate als gut zu bezeichnen sind. Die fett markierten Ergebnisse kennzeichnen die jeweils besten Trefferquoten in den entsprechenden Gruppen. Die guten Ergebnisse von SVM können durch den Vergleich zu alternativen Verfahren bestätigt werden. So wird zum Beispiel mittels LDA lediglich eine Gesamttrefferquote von 86,86% erreicht, was durch den Einsatz linearer SVM deutlich verbessert werden kann. Es ist anzumerken, dass die identifizierten Gruppen sehr gut durch SVM getrennt werden können, da die auftretenden fehlklassifizierten Beobachtungen sich sehr nah an der Entscheidungsebene befinden und somit nur mit einem sehr geringen, absoluten Entscheidungswert klassifiziert wurden. Der niedrige Anteil von lediglich etwa 13% an durchschnittlich erreichten Support Vektoren bestätigt diese Beobachtung.
Die eigentliche Analyse der Daten und die spätere Reduzierung der Merkmale wird auf einer festen Einteilung in Trainings- und Testdaten vorgenommen, deren Verhältnis (3 zu 1) dem Verhältnis der Trainingsdaten zu den Testdaten bei der vorangegangenen Kreuzvalidierung entspricht. Die zufällige Zuweisung zu den beiden Datensätzen ergibt eine zum gesamten Datensatz ähnliche Verteilung der fünf Klassen, wie Tabelle 4.18 zeigt.[24]

Klasse	1	2	3	4	5	Summe
Training	10,95%	26,09%	25,24%	18,29%	19,43%	100%
Test	12,57%	25,86%	24,0%	17,0%	20,57%	100%
Gesamt	11,36%	26,04%	24,93%	17,96%	19,71%	100%

Tabelle 4.18: Verteilung der Klassen innerhalb der Trainings- und Testdaten sowie im gesamten Datensatz

Die zweite Klasse („traditionell, deutsch bürgerliche Küche“) nimmt wieder den größten Anteil ein. Ziel der nächsten Schritte ist die Reduzierung der Merkmale, sodass eine Zuordnung von Beobachtungen zu den fünf Klassen auch auf Basis weniger, aber entscheidender Merkmale mit einer akzeptablen Trefferquote möglich ist. Dazu kommen unterschiedliche Verfahren zur Merkmalsreduktion zum Einsatz, die bereits in Abschnitt 3.5 vorgestellt wurden. Die mittels der festen Auswahl eines Testdatensatzes ermittelten Trefferquoten liegen bei 91,43% bei linearer SVM mit $C = 1$ und bei 91,0% bei nicht linearer SVM mittels Radialbasis-Kern mit $\gamma = 0,00073$ und $C = 5$ und sollen im Folgenden als Vergleichsmaß herangezogen werden.
Da bei der vorliegenden Datengrundlage hauptsächlich auf die lineare Trennung der Klassen zurückgegriffen wird, wird hier insbesondere die Auswahl anhand des Normalenvektors im Vergleich zum alternativen Einsatz der Diskriminanzanalyse verwendet. Die 25 durch LDA ermittelten wichtigsten Merkmale werden in Tabelle

[24] Ein χ^2-Homogenitätstest zeigt, dass die beiden Datensätze Training und Test bei einem Signifikanzniveau von $\alpha = 0,05$ eine homogene Verteilung der Klassen aufweisen ($\chi^2 = 2,38 < 9,49 = \chi^2_{krit}$).

4.19 dargestellt. Dabei wird die Wichtigkeit eines Merkmals durch den gewichteten Diskriminanzkoeffizient bestimmt (*Decker, Temme* (2000)). Der Einsatz der schrittweisen Diskriminanzanalyse kommt zu einer ähnlichen Reihenfolge, die mittels der Rangkorrelation nach Spearman als stark korreliert zu der hier vorliegenden zu bewerten ist. Die letzte Spalte enthält den jeweils erreichten Rang. Zum Vergleich werden zusätzlich die für diese Merkmale mittels linearer SVM ($\boldsymbol{w}$) und FCS (vgl. Seite 73) ermittelten Ränge angegeben. Zusätzlich zum Rang wird bei der linearen Trennung (mit $C = 1$) mittels Normalenvektor der Scorewert angezeigt. Da hier eine OAO-Multiklassifikation durchgeführt wurde, erhalten die Merkmale einen Scorewert, der jeweils über die $\frac{5(5-1)}{2} = 10$ SVM gemittelt wurde. Abhängig vom erreichten Wert des Verfahrens ergibt sich die jeweilige Rangfolge der Merkmale[25]. Die verteilten Ränge der einzelnen Verfahren zeigen, dass die Methoden zu unterschiedlichen Ergebnissen kommen und sich in der Verteilung der Items unterscheiden. Lediglich in wenigen Merkmalen stimmen die vergebenen Ränge grob überein, wie z.B. bei Nr. 19 „Wichtigkeit der Qualität der Nahrungsmittel“ oder Nr. 51 „Vorzug von Kochen einfacher Gerichte“. Dennoch kann dieses Ergebnis so beurteilt werden, dass grob gesehen die gleichen Merkmale als wichtig (also zu den 25 ersten Merkmalen gehörig) bestimmt werden. Dies wird daran deutlich, dass nur fünf der mittels SVM für diese Merkmale vergebenen Ränge über 25 liegen (bei FCS sind es ebenfalls fünf). Würden zufällig 25 aus den 61 Variablen gezogen werden, so läge die Wahrscheinlichkeit dafür, mindestens 80% der ausgewählten Merkmale unter den 25 ersten Rängen wie bei der Auswahl mittels Diskriminanzanalyse zu finden, nur bei $2,364 \cdot 10^{-7}$. Dies zeigt, dass alle drei betrachteten Vorgehensweisen grob zu ähnlichen Ergebnissen kommen.
Sind die für eine Klassifikation relevanten Merkmale bestimmt, so können die aus einem reduzierten Datensatz resultierenden Trefferquoten bestimmt werden. Im Folgenden wurden dazu die von dem jeweils verwendeten Verfahren durch einmaliges Anwenden als unwichtig eingestuften Merkmale sukzessive aus den ursprünglichen Daten entfernt und die sich ergebende Trefferquote auf den Validierungsdaten ermittelt[26]. In Abbildung 4.10 werden die resultierenden Ergebnisse bei Reduzierung der Anzahl der Merkmale dargestellt. Die durchgezogene Linie kennzeichnet das Ergebnis der Reduzierung unter Einsatz des Normalenvektors. Die gestrichelte Kurve repräsentiert das Ergebnis auf Basis von FCS und die schraffierte Linie die Ergebnisse basierend auf der LDA.[27] Auf die Generierung und Bedeutung der vierten Linie wird auf Seite 169 Bezug genommen. Es ist deutlich zu erkennen, dass der Einsatz des Normalenvektors zur Merkmalsauswahl die anderen beiden im größten Teil dominiert, nur bei der Reduzierung der Merkmale auf bis zu 11

[25] In den Tabellen A.3 und A.4 im Anhang sind die vollständigen Items der hier nur in abgekürzter Form dargestellten Merkmale aufgelistet.

[26] Eine Alternative dazu wäre das Trainieren auf den vollständigen Daten und Validieren des Modells mittels eines Datensatzes, der ebenfalls die volle Anzahl an Merkmalen enthält, die irrelevanten jedoch durch Mittelwerte ersetzt wurden. Dies führt allerdings, wie entsprechende Analysen gezeigt haben, zu einer geringeren Trefferquote.

[27] Hierbei werden insbesondere bei SVM bei jeder Neuberechnung die zuvor ermittelten Parameterwerte verwendet, also eine lineare Trennung mit $C = 1$.

Nr.	Merkmal	$\boldsymbol{w}$-Score	Rang ($\boldsymbol{w}$)	Rang (FCS)	Rang (LDA)
46	regelmäßige Einnahme von Vitaminpräparaten	0,121	33	20	1
25	viel Zeit für Kochen	0,155	16	3	2
38	Vorliebe für ausgefallene Speisen und Gerichte	0,144	21	5	3
16	Besitz neuer Produkte vor Bekannten	0,147	19	36	4
40	Vorliebe für schnelle Gerichte	0,204	1	11	5
28	vollwertige Ernährung	0,190	3	13	6
58	Vermeidung von gesundheits-schädlicher Ernährung	0,139	22	6	7
37	Kauf von zusatzstofffreien Lebensmitteln	0,163	11	1	8
51	Vorzug von einfachen Gerichten	0,187	4	4	9
30	kein Kochen ohne Fertigprodukte	0,157	14	38	10
55	häufige Verwendung von Getreide-körnern	0,150	18	32	11
17	Festhalten an alten Gewohnheiten	0,167	9	18	12
26	Vorzug edler Speisen und Getränke	0,128	27	8	13
33	Überschätzung des Einflusses der Ernährung auf die Gesundheit	0,121	34	35	14
13	andere Interessen als nur Küche	0,178	7	25	15
44	Kauf frischer Lebensmittel	0,129	26	17	16
50	Spaß am Ausprobieren fremd-ländischer Spezialitäten	0,181	6	12	17
59	bei Wahl: Kauf von deutschen Lebensmitteln	0,161	12	15	18
48	Rücksichtnahme auf Gesundheit	0,158	13	9	19
18	Kochen von Gerichten, die garantiert gelingen	0,170	8	10	20
35	Achten auf schonende reizarme Kost	0,182	5	2	21
60	keine Ergänzungspräparate bei normaler Kost notwendig	0,052	60	59	22
27	Vorzug von Hausmannskost	0,130	25	16	23
19	Wichtigkeit der Qualität	0,135	24	22	24
53	Kochen altbewährter Gerichte	0,195	2	7	25

Tabelle 4.19: Auswahl an 25 Merkmalen, die bezüglich LDA die höchste diskriminatorische Eigenschaft besitzen

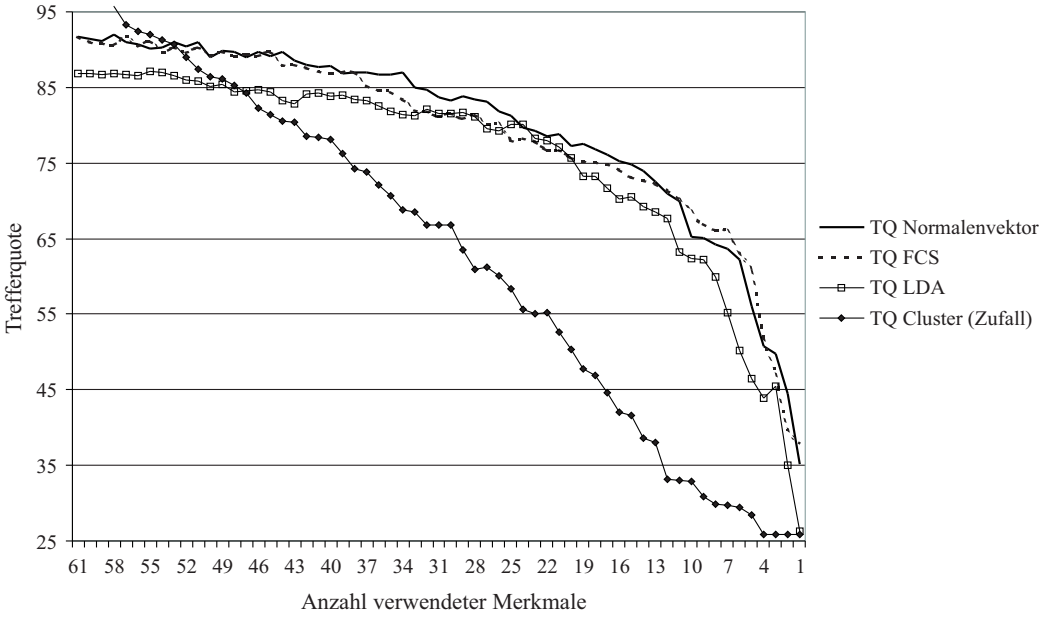

Abbildung 4.10: Trefferquote bei Reduzierung der Merkmale mittels Normalenvektor, FCS und LDA

oder weniger schwächer abschneidet[28]. Die Ergebnisse der LDA sind im Vergleich dieser drei Verfahren als die schlechtesten zu beurteilen. Dies kann u.a. darin begründet liegen, dass LDA generell bei der Trennung eine niedrigere Trefferquote erreicht. Unter Einsatz aller Merkmale erreicht die LDA lediglich eine Trefferquote von 86,9% auf den festen Testdaten, wohingegen mit SVM 91,43% erzielt werden. Werden hingegen die relativen Veränderungen zu jeweils der Trefferquote auf allen Merkmalen betrachtet, so ergeben sich für die LDA zur linearen SVM vergleichbare Ergebnisse. Bei einer Reduktion der eingehenden Merkmale auf bis zu 10 oder weniger Merkmale fällt die Trefferquote bei allen Verfahren drastisch. Dies zeigt, dass mehr als nur zehn Merkmale zur adäquaten Beschreibung der Cluster erforderlich sind.

Werden die Merkmale auf ihre inhaltliche Bedeutung untersucht, so resultieren folgende Ergebnisse, die mit der schrittweisen Diskriminanzanalyse bestätigt werden können. Die mittels aller Verfahren als eher unwichtig eingestuften[29] Variablen sind in Tabelle 4.20 zusammengefasst.

[28] Auf die ebenfalls dargestellte Kurve „TQ Cluster (Zufall)“ wird später eingegangen.

[29] Alle Variablen, die bei einer Rangfolge einen Rang unter den letzten 15 Variablen erhalten, werden hier als unwichtig bezeichnet.

Nr.	Merkmal
1	Von den Produkten, die laufend auf den Markt kommen, halte ich die meisten für überflüssig.
2	Ich liebe die Atmosphäre von kleineren Läden und Fachgeschäften.
3	Ich gehöre zu den Menschen, die Geselligkeit lieben.
10	Am wohlsten fühle ich mich zu Hause.
11	Den Aussagen der Werbung stehe ich mit sehr großem Misstrauen gegenüber.
24	Es ist mir egal, ob meine Lebensmittel aus Deutschland sind oder aus irgendeinem anderen Land.
42	Ich verwöhne mich gerne mit einem guten Essen.

Tabelle 4.20: Ausgewählte Einstellungsmerkmale, die von allen eingesetzten Verfahren als unwichtig klassifiziert werden

Bei einigen Variablen ist die Bewertung anhand der Charakterisierung der einzelnen Cluster intuitiv nachvollziehbar, wie etwa bei Item 11 „Den Aussagen der Werbung stehe ich mit sehr großem Misstrauen gegenüber“. Da nicht zu erwarten ist, dass etwa Gruppen wie „Fast-Food“-Cluster (Klasse „3“) oder „bewusste Genießer“ (Klasse „5“) durch deutliche Zustimmung oder Ablehnung dieser Aussage geprägt sind, ist diese Bewertung nachvollziehbar. Dies ist allerdings nicht für alle Variablen gegeben. Bei einer manuellen Auswahl würde das Item 10 „Am wohlsten fühle ich mich zu Hause.“ für eine Trennung der Gruppen aufgrund der ausgeprägten Zustimmung im traditionellen Segment möglicherweise hinzugezogen werden. Da hier jedoch die Trennfähigkeit der Merkmale hinsichtlich aller fünf Cluster angegeben wird, ist dies ein Grund für die Herabsetzung der Wichtigkeit dieses Merkmals. Merkmal 10 ist zwar für die Trennung des traditionellen Segmentes von den übrigen relevant, insbesondere beim Vergleich zu den „bewussten Genießern“, spielt hingegen bei der Trennung der übrigen Klassen eine untergeordnete Rolle, sodass der Gesamt-Score für die Multiklassifikation gering ausfällt.
Die jeweils als wichtigste Variablen für die Differenzierung der Klassen eingestuften Einstellungsmerkmale enthält Tabelle 4.21. Dabei ist ein ähnliches Phänomen wie bei den irrelevanten Merkmalen zu beobachten.

Nr.	Merkmal
25	Für das Kochen nehme ich mir viel Zeit.
28	Wir ernähren uns nach dem Prinzip der Vollwertküche.
40	Ich koche am liebsten Gerichte, die schnell gehen.
48	Ich achte darauf, was ich esse und trinke, denn ich muss auf meine Gesundheit Rücksicht nehmen.
50	Es macht mir Spaß, fremdländische Spezialitäten auszuprobieren.
51	Je einfacher das Kochen geht, desto lieber ist es mir.

Tabelle 4.21: Ausgewählte Merkmale, die von allen eingesetzten Verfahren als wichtig klassifiziert werden

Das Item 48 „Ich achte darauf, was ich esse und trinke, denn ich muss auf meine Gesundheit Rücksicht nehmen.“ charakterisiert keinen der angegebenen Cluster, sodass es a priori bei einer manuellen Auswahl ausgeschlossen werden würde. Die vorliegenden Analysen zeigen jedoch, dass es zur Diskriminierung der Klassen in Form einer höheren Trefferquote beiträgt. Dies zeigt, dass eine Auswahl der für eine Klassifikation wichtigen Merkmale insbesondere bei einer Multiklassifikation mittels eines computerbasierten Verfahrens durchgeführt werden sollte, um eine höhere Trefferquote als bei einem manuellen Ausschluss zu erzielen. Häufig können Zusammenhänge bei Vorliegen vieler Gruppen nicht ohne weiteres erkannt werden, sodass der Einsatz von Merkmalsreduktionsverfahren gerade hier sinnvoll ist. Ein Effekt kann auch in der Interaktion der Variablen bestehen, der bei den bisher angewendeten Verfahren nicht berücksichtigt werden kann. Darauf wird allerdings auf Seite 174 eingegangen, wenn der IRRM-Algorithmus (vgl. Abschnitt 3.5) und seine Erweiterungen angewendet werden.

Um die Gewichtung der Merkmale und damit die für die Klassifikation auszuwählenden Merkmale zu bestimmen, wurde ebenfalls der von *Guyon et al.* (2002) vorgeschlagene Algorithmus SVM RFE eingesetzt (vgl. Abschnitt 3.5). Bei dem einfacheren, bisherigen Vorgehen wird die Reihenfolge der Wichtigkeit der Merkmale lediglich durch ein einmaliges Berechnen der SVM bestimmt und die Trefferquote danach auf den jeweils entsprechend reduzierten Datensätzen berechnet. Bei RFE hingegen wird nach jedem Entfernen eines Merkmals die Reihenfolge der Wichtigkeit erneut bestimmt, was einen deutlich umfangreicheren Arbeitsaufwand darstellt. Die mittlere erzielte Trefferquote ist sehr ähnlich zu dem etwas einfacheren Vorgehen wie Abbildung 4.11 verdeutlicht. Diese Vorgehensweise wird nicht eindeutig von den Ergebnissen von RFE dominiert. Dies zeigt, dass die anfangs durch den Normalenvektor $\boldsymbol{w}$ bestimmte Reihenfolge der Merkmale sehr gut war und die wiederholte Neuordnung nach jedem Training nicht zwingend notwendig erscheint. Bei späteren Auswertungen wird auf Basis des Ellbogenkriteriums und mittleren zu erwartenden Scores ein Umfang von 27 Merkmalen ausgewählt. Damit ergeben sich 82,28% Genauigkeit der Prognose im Vergleich zu 81,86% beim trivialen Vorgehen, was einer zusätzlichen Richtigklassifikation von drei der 700 Testbeobachtungen entspricht. Aufgrund der nur geringen Verbesserung der Trefferquote, die mit einem ungleich höheren Aufwand verbunden ist, erscheint hier die einfachere Vorgehensweise sinnvoller. Spielt hingegen eine Erhöhung der Genauigkeit um wenige Prozentanteile eine entscheidende Rolle, so ist der Aufwand und die damit verbundene erhöhte Trefferquote des umfangreicheren Algorithmus SVM RFE gerechtfertigt. Dies liegt insbesondere dann vor, wenn die Fehlklassifikation einzelner Beobachtungen enorme Kosten oder negative Folgen verursacht und eine Maximierung der Trefferquote unbedingt erforderlich ist, wie etwa bei der Krebsdiagnostik. Im betriebswirtschaftlichen Kontext sei in diesem Zusammenhang die Klassifikation von Unternehmen durch Venture Capitalists zu nennen, die diejenigen Startups identifizieren möchten, in die sie investieren. Werden erfolgreiche Unternehmen nicht erkannt, so entgeht dem Venture Capitalist, der von der Erkennung Erfolg versprechender Startups lebt, ein hoher Umsatz.

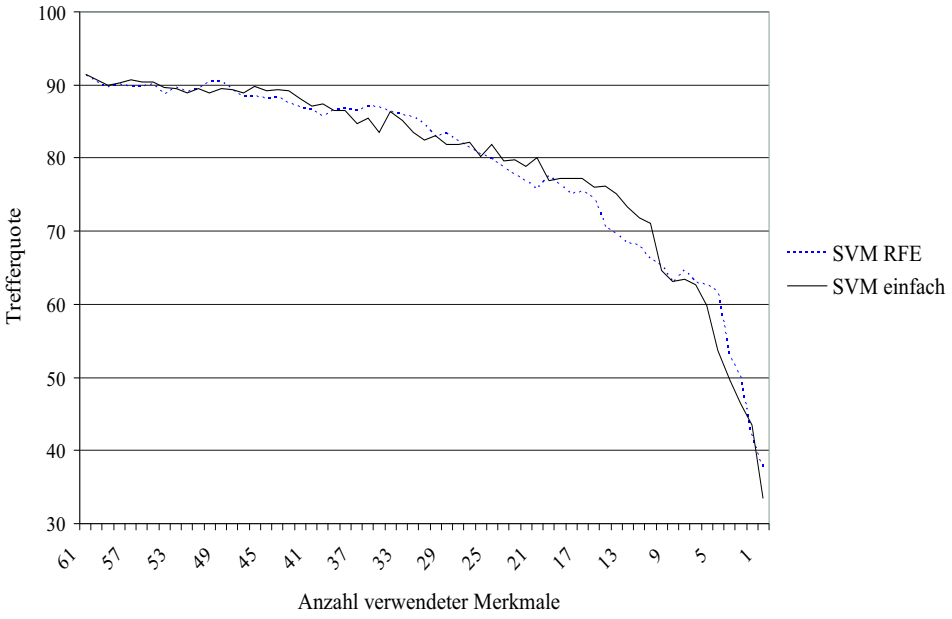

Abbildung 4.11: Trefferquote bei Reduzierung der Merkmale mittels einfachem Vorgehen und SVM RFE

Da die Gruppenbildung in diesem Fall mittels Clusteranalyse mit anschließender Klassenzuweisung durch Distanzvergleich erfolgte, ist der Vergleich der Ergebnisse einer intuitiven Zuordnung der Beobachtungen anhand einer reduzierten Menge von Merkmalen berechtigt. Darunter wird die Zuweisung der Vektoren auf Basis euklidischer Distanzen zu den fünf Clustern verstanden, ohne dabei SVM einzusetzen. Wird die Reihenfolge der Merkmale zunächst zufällig ausgewählt und die Beobachtungen dann den Clustern auf Basis der jeweils reduzierten Merkmale zugewiesen, so ergibt sich die in Abbildung 4.10 zusätzlich enthaltene Kurve „TQ Cluster (Zufall)“. Es zeigt sich, dass SVM im Vergleich zu der trivialen Vorgehensweise bei Reduktion um mindestens 10 Merkmale deutlich besser (ausgedrückt in der Trefferquote der Testdaten) abschneidet. Der Einsatz von SVM zur Reduktion der für eine Klassifikation zu verwendenden Merkmale lohnt sich demnach.

Wird die Bestimmung der Relevanz der eingehenden Merkmale auf Basis einer nicht linearen Trennung mittels Radialbasis-Kern mit $\gamma = 0,00073$ und $C = 5$ (vgl. Tabelle 4.17) vorgenommen, so resultieren ähnliche Ergebnisse. Beim Gradientenverfahren wurde zunächst eine Reihenfolge basierend auf lediglich den Margin-Vektoren, also denjenigen Support Vektoren mit $0 < \alpha_i < C$, berechnet. Dies entspricht der Festsetzung des zu wählenden Parameters ϵ auf $\epsilon = 0$ (vgl. Abschnitt 3.5). Die Erweiterung der Basis an Vektoren, die für die Gewichtung der Merkmale verantwortlich sind, um diejenigen, die in einer ϵ-Umgebung um die Hilfsebene liegen, zeigt, dass dies zu sehr unterschiedlichen Ergebnissen führen kann. Dies kann in einer möglichen Verschiebung der Wichtigkeiten begründet sein:

falls sich sehr viele Support Vektoren innerhalb der Hilfsebenen, also innerhalb der zu maximierenden Spanne befinden, wird die Wichtigkeit eines Merkmals gemäß Gleichung (3.7) auf einer deutlich vergrößerten Basis berechnet, sodass hier Veränderungen bezüglich der Gewichte der Merkmale auftreten können. Daher erscheint es sinnvoll, die Wahl des Umgebungsparameters ϵ so zu wählen, dass alle richtig klassifizierten Support Vektoren mit in die Berechnung der Relevanzen der Merkmale eingehen. Dies wird insbesondere durch $\epsilon = 1$ erreicht, da somit der Raum innerhalb der Hilfsebenen (vgl. Abbildung 2.4) abgedeckt wird. Im Folgenden werden daher die Ergebnisse des nicht linearen Vorgehens mit $\epsilon = 1$ betrachtet.
Mittels des Gradientenverfahrens werden leicht bessere Ergebnisse als bei der linearen Trennung erzielt. Im Vergleich zu den in Abbildung 4.10 gezeigten Ergebnissen der Trefferquote weicht die Kurve des nicht linearen Vorgehens in 38 der berechneten 61 Fälle verglichen zu den Trefferquoten auf Basis des Normalenvektors nach oben ab. Dabei wird im Durchschnitt eine um 2,04 Prozentpunkte höhere Trefferquote erreicht. In den übrigen Fällen ergibt sich eine geringere Trefferquote als bei der Methode des Normalenvektors. Die Abweichung beträgt hier im Durchschnitt 1,74 Prozentpunkte, sodass insgesamt von sehr ähnlichen bzw. leicht besseren Ergebnisse gesprochen werden kann.
Weiterhin ergibt eine gesonderte Analyse der erzeugten Winkel zwischen den jeweiligen Einheitsvektoren und den Gradienten der Entscheidungsfunktion, dass keinem Merkmal mehr Relevanz aufgrund eines hohen Gewichts an manchen Vektoren zugesprochen werden kann. Die erzielten Winkel zwischen den Gradienten und den jeweiligen Einheitsvektoren unterscheiden sich innerhalb der Support Vektoren lediglich um maximal 5°. Dies kann in der Form der berechneten Hyperebene begründet sein. Da die lineare Trennung der Daten zu sehr ähnlichen Ergebnissen gelangt, ist anzunehmen, dass die nicht lineare Trennung zwar von der nicht linearen Trennebene abweicht, aber dennoch eine ähnliche Struktur besitzt. Demnach können die erzielten Winkel pro Merkmal keine großen Unterschiede aufweisen.

Bei der Bestimmung der letztendlich optimalen Anzahl der Merkmale sind unterschiedliche Faktoren von Bedeutung. Neben der möglichen Reduzierung der Trefferquote spielt gerade bei der Datenerhebung die Anzahl der nicht mehr zu erfassenden Variablen eine besondere Rolle, da diese die Kosten der Erhebung deutlich vermindern kann. Aufgrund der Vielzahl an Arten von Daten, sowie unterschiedlicher Kosten ist es unmöglich, eine pauschale Regel für die Verwendung einer bestimmten Anzahl von Merkmalen aufzustellen. Dennoch kann basierend auf den Ausführungen in Abschnitt 3.5 ein Ansatzpunkt bei der Ermittlung der richtigen Anzahl gegeben werden. In diesem Fall stellt die Reduzierung der Merkmale ein entscheidendes Element dar, da die Minimierung des Umfangs eines Fragebogens zur Ermittlung des möglichen Kaufverhaltens Kern der Analyse ist. Somit ist es von Interesse, möglichst wenige Fragen in den Katalog aufzunehmen, um die Befragten nicht unnötig durch lange Interviews oder Online-Fragebögen zu ermüden. Die optimale Anzahl an Merkmalen wird somit zum einen durch die Minimierung der Fehler und durch die Minimierung der Anzahl an Fragen gekennzeichnet. Die Reduzierung der Anzahl der Merkmale zeigt jedoch lediglich

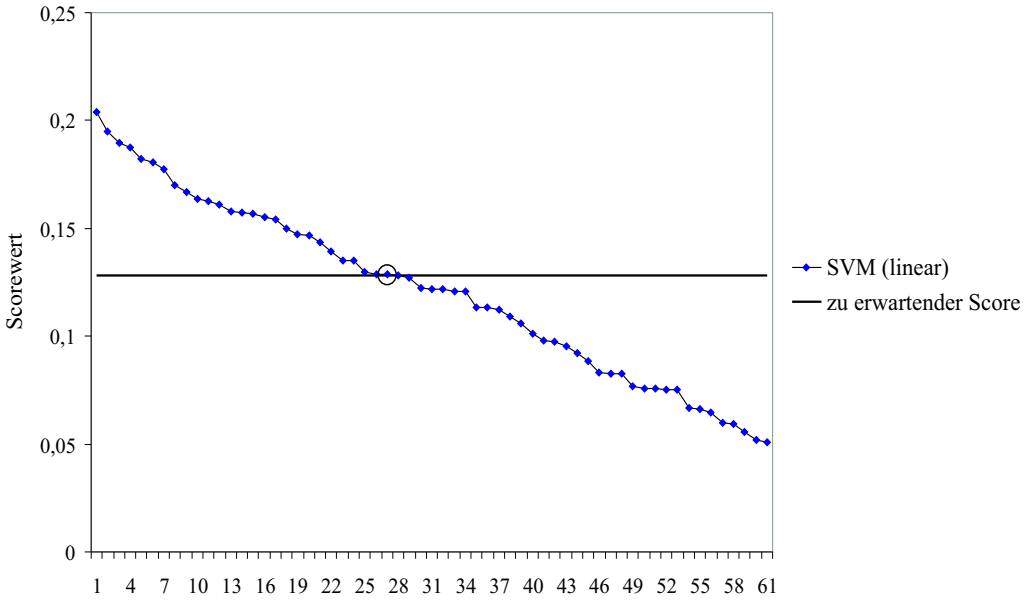

Abbildung 4.12: Nach Größe sortierte Scorewerte bei linearer Trennung mit zusätzlich zu erwartendem Score

bei der Klassifikation neuer Beobachtungen ihre Wirkung, da sich der Umfang der zu erhebenden Merkmale drastisch reduzieren kann. Zur Ermittlung der jeweiligen Merkmale und zur Bestimmung der Trennebene ist jedoch die Heranziehung aller Merkmale erforderlich. Daher kann der Einsatz von SVM hier als ein Mittel des Pretests gesehen werden. Zur Bestimmung der Anzahl der zu extrahierenden Merkmale stehen verschiedene Möglichkeiten zur Verfügung (vgl. Abschnitt 3.5). Das Ellbogenkriterium ist hier nicht anwendbar, da kein eindeutiger Ellbogen erkennbar ist (vgl. Abbildung 4.12), nach dem der Wert drastisch sinkt. Vielmehr nimmt der Score pro Merkmalsrang mehr oder weniger gleichmäßig ab. Alternativ kann der erwartete Score herangezogen werden. Wird beim Einsatz der Gewichtung durch den Normalenvektor bei der linearen Trennung danach die Anzahl der zu verwendenden Merkmale bestimmt, so resultiert in diesem Beispiel ein Datensatz im Umfang von 27 Merkmalen, was zu einer zu erwartenden Trefferquote von etwa $82,14\%$ bei der einfachen Vorgehensweise führt. Die Entscheidung ist nochmals in Abbildung 4.12 verdeutlicht. Der Scorewert beim 27. Merkmalsrang wird durch einen Kreis gekennzeichnet. Der zu erwartende Scorewert beträgt hier etwa 0,128. Die Extraktion von genau 27 Merkmalen ergibt sich aus dem Schnittpunkt der Linie, die den zu erwartenden Score repräsentiert, und der Kurve, die die Entwicklung der Wichtigkeit der Merkmale angibt. Da es sich hier nicht um die Vermarktung hochpreisiger Güter dreht, bei der die Maximierung der Trefferquote im Vordergrund steht, kann eine Reduzierung der Trefferquote um etwa 9 Prozentpunkte zugunsten der Reduzierung der zu erhebenden Merkmale um 34 Stück (55,7%) als akzeptabel bezeichnet werden. Diese Anzahl erscheint vor dem Hintergrund der Gestaltung

eines Fragebogens dennoch zu hoch, sodass dies lediglich als erster Anhaltspunkt verwendet werden kann. Wird eine weitere Reduktion der Genauigkeit des Modells akzeptiert, so kann die Anzahl der auszuwählenden Merkmale entsprechend weiter gesenkt werden. Ist eine Anzahl von 10 Fragen akzeptabel, so erreicht man durch die Verwendung der entsprechenden Merkmale immer noch eine Trefferquote von 67,86% bei der nicht linearen Trennung, bzw. 64,57% bei der linearen Trennung. Diese Trefferquote ist hinsichtlich des Vergleichs mit der zu erwartenden Trefferquote bei Verwendung des Größte-Gruppen-Kriteriums von etwa 26% immer noch als sehr gut zu bezeichnen.
Bei Betrachtung der Score-Werte, die mittels des Gradientenverfahrens ermittelt werden, ergibt sich ebenfalls kein eindeutiger Ellbogen. Dennoch kann der Umfang von 27 Merkmalen angenommen werden, da hier 28 Merkmale einen überdurchschnittlichen Score erhalten und das Ergebnis damit bestätigen.
Wird also der Umfang von 27 bzw. 10 Merkmalen festgehalten, so ergeben sich die in Tabelle 4.22 enthaltenen Ergebnisse.

Methode	10 Merkmale	27 Merkmale	61 Merkmale
SVM ($\boldsymbol{w}$)	64,57%	82,14%	91,43%
SVM RBF (Gradient)	68,00%	85,00%	91,0%
FCS	70,57%	77,86%	(91,43%)
LDA	68,57%	78,57%	86,86%

Tabelle 4.22: Ergebnisse bei fest gewählter Anzahl an Merkmalen

Angegeben werden jeweils die Trefferquoten, die mittels linearer SVM (SVM linear, FCS), nicht linearer SVM und LDA bei Reduktion der Merkmale auf einen Umfang von 10 und 27 ermittelt wurden. Die Menge der jeweils heranzuziehenden Merkmale bestimmt sich durch das jeweils eingesetzte Verfahren. Zum Vergleich sind zusätzlich die Ergebnisse angegeben, die bei Hinzuziehen aller verfügbaren Merkmale resultieren.[30] Die Reduktion der Merkmale auf die 27 wichtigsten würde somit beispielsweise bei der linearen Trennung zu einem Verlust der Trefferquote um etwa 10 Prozentpunkte führen. Bei allen Verfahren ist weiterhin ein deutlicher Verlust der Trefferquote bei Reduktion der Merkmale von 27 auf 10 zu erkennen. Hierbei erzielt das Gradientenverfahren bei 27 Merkmalen das beste Ergebnis, bei 10 Merkmalen ist die bei vorangegangenem SVM-Einsatz bei FCS ermittelte Trefferquote am höchsten, was ebenfalls in Abbildung 4.10 erkennbar ist. Der Unterschied zwischen LDA und SVM ist bei 10 Merkmalen nur sehr gering, bzw. LDA schneidet leicht besser ab, sodass hier keine eindeutige Entscheidung für oder gegen ein Verfahren getroffen werden kann.
In folgender Tabelle werden diejenigen Merkmale aufgelistet, die bei mindestens zwei

[30] FCS bezeichnet die Vorgehensweise zur Reduzierung der Merkmale. Klassifiziert wird in diesem Fall ebenfalls mittels der linearen SVM. Daher ist die eingeklammerte Trefferquote bei 61 Merkmalen unabhängig von FCS.

der vier betrachteten Verfahren unter den wichtigsten 27 Merkmalen liegen[31].

			Methode		
Nr.	Bedeutung	$\boldsymbol{w}$	Gradient	FCS	LDA
13	andere Interessen als nur Küche	X	X	X	X
17	Festhalten an alten Gewohnheiten	X	X	X	X
18	Kochen von Gerichten, die garantiert gelingen	X	X	X	X
19	Wichtigkeit der Qualität	X	X	X	X
25	viel Zeit für Kochen	X	X	X	X
26	Vorzug edler Speisen und Getränke	X	X	X	X
27	Vorzug von Hausmannskost	X	X	X	X
28	vollwertige Ernährung	X	X	X	X
35	Achten auf schonende reizarme Kost	X	X	X	X
37	Kauf von zusatzstofffreien Lebensmitteln	X	X	X	X
38	Vorliebe für ausgefallene Speisen	X	X	X	X
40	Vorliebe für schnelle Gerichte	X	X	X	X
44	Kauf frischer Lebensmittel	X	X	X	X
50	Spaß am Ausprobieren fremdländischer Spezialitäten	X	X	X	X
51	Vorzug von einfachen Gerichten	X	X	X	X
53	Kochen altbewährter Gerichte	X	X	X	X
58	Vermeidung von gesundheitsschädlicher Ernährung	X	X	X	X
59	bei Wahl Kauf von deutschen Lebensmitteln	X	X	X	X
16	Besitz neuer Produkte vor Bekannten	X	X	0	X
21	Keine Konservierungsstoffe	X	X	X	0
30	kein Kochen ohne Fertigprodukte	X	X	0	X
46	regelmäßige Vitaminpräparate	0	X	X	X
48	Rücksichtnahme auf Gesundheit	X	0	X	X
55	häufige Verwendung von Getreidekörnern	X	X	0	X
15	Genuss des Lebens in vollen Zügen	X	X	0	0
22	Achten auf Figur	X	0	0	X
33	Überschätzung des Einflusses der Ernährung auf die Gesundheit	0	X	0	X
49	kein umweltbelastete Kost	X	0	X	0
Summe		26	25	22	25

Tabelle 4.23: Inhalt und auftretende Merkmale

[31]Hier werden die Merkmale wiederum lediglich in verkürzter Form dargestellt. Ausführliche Formulierungen der Items finden sich in den Tabellen A.3 und A.4 im Anhang.

Dabei gibt der Eintrag X (bzw. 0) an, dass sich das Merkmal unter Einsatz des betreffenden Verfahrens unter den jeweils 27 wichtigsten Merkmalen befindet (bzw. sich nicht befindet). Diese Zusammenstellung umfasst 28 Merkmale, was die hohe Übereinstimmung der Verfahren aufzeigt. Lediglich mittels FCS werden leicht abweichende Merkmale (insgesamt fünf) ausgewählt, die durch kein anderes Verfahren als wichtig beurteilt werden.
Zur Bestimmung der Klassenzugehörigkeiten ist es augenscheinlich wichtig, zu erfahren, ob die Konsumenten Wert auf ihre Gesundheit beim Einkaufen legen und an alten Gewohnheiten festhalten. Dies wird anhand des größten Teils der in Tabelle 4.23 aufgezeigten Merkmale deutlich.

Ergänzend soll das Vorgehen zur Bestimmung der Anzahl der auszuwählenden Merkmalen bei Hinzunahme eines zufällig verteilten Merkmals gemäß Abschnitt 3.5 (Seite 75) herangezogen werden. Dazu wird ein Merkmal mit zufällig verteilten Ausprägungen zum Datensatz hinzugefügt und diejenigen Merkmale entfernt, deren Score unterhalb des hinzugefügten Merkmals liegt. Hierbei würden lediglich die beiden Merkmale 1 und 60 aufgrund ihres geringen Gewichts von der Untersuchung ausgeschlossen werden. Dies zeigt die hohe Relevanz der Merkmale zur Beschreibung der fünf Cluster. Da hier jedoch die Reduzierung der Merkmale um einen hohen Anteil im Vordergrund steht, ist diese Vorgehensweise an dieser Stelle nicht adäquat.

Die bisherigen Ausführungen beziehen sich alle bis auf SVM RFE auf das einmalige Sortieren und das anschließende Reduzieren der Merkmalsmenge. In Abschnitt 3.5 wurde eine Möglichkeit vorgestellt, bei Vorgabe der Anzahl der Merkmale, diejenigen auszuwählen, die gemeinsam die höchste diskriminatorische Eigenschaft besitzen. Dazu kann auf die zitierte Vorgehensweise des IRRM oder auf die erweiterte Variante unter Einbeziehung des Gradientenverfahrens (NLIRM) zurückgegriffen werden. Im Folgenden soll wiederum eine Menge von 10 und 27 Merkmalen extrahiert werden. Unter Vorgabe dieses Parameters gelangt man zu den in Tabelle 4.24 dokumentierten Ergebnissen.

Anzahl an		η			
Merkmalen	Verfahren	1	2	3	4
10	IRRM	70,29%	71,29%	71,14%	70,29%
10	NLIRM	72,0%(0)	71,29%(1)	**72,14**%(1)	71,43%(0)
27	IRRM	85,0%	86,0%	**86,29**%	85,71%
27	NLIRM	85,14%(0)	85,14%(0)	85,14%(0)	85,14%(0)

Tabelle 4.24: Ergebnisse der IRRM- und NLIRM-Verfahren bei 10 und 27 Merkmalen

Dabei bezeichnet η den für dieses Vorgehen notwendigen Parameter. In *Fröhlich, Zell* (2004) wird $\eta = 1$ als bester Wert verwendet, was hier nicht bestätigt werden kann, da $\eta = 3$ zu den bei beiden Merkmalsmengen besten Ergebnissen führt. Die Markierungen (0) und (1) geben den Wert des Parameters ϵ beim zugrunde

liegenden Gradientenverfahren an. Es zeigt sich, dass die beiden Vorgehensweisen (IRRM, NLIRM) zu sehr ähnlichen Ergebnisse gelangen. Sowohl für IRRM als auch für NLIRM gibt es eine Parameterkonstellation, für die das entsprechende Verfahren das jeweils andere bei Vorgabe von 10 oder 27 und η Merkmalen dominiert.
Werden bei allen Verfahren die wichtigsten Merkmale zusammengestellt, die sich bei der Extraktion der 27 wichtigsten Merkmale ergeben, so sind dies die Nummern 13, 18, 37, 40, 50, 51 und 53 (vgl. Tabellen in Anhang A.1). Diese sind bereits durch alle bisherigen einfacheren Methoden als wichtig eingestuft worden. Diese Merkmale besitzen im gegenseitigen Zusammenspiel die höchste Trennfähigkeit. Dies wird auch bei der näheren Betrachtung der Merkmale deutlich. Sie zeichnen sich insbesondere dadurch aus, dass sie zusammengenommen die fünf Cluster sehr gut charakterisieren, wohingegen sie einzeln gesehen nur für ein oder zwei Gruppen typische Aussagen beinhalten. So zeichnet zum Beispiel die Zustimmung zu Merkmal 53 („Kochen von altbewährten Gerichten“) die Haushalte aus der „traditionell, deutsch bürgerlichen“ Gruppe aus, wohingegen die Beobachtungen aus der Gruppe der „bewussten Genießer“ durch Ablehnung dieser Aussage gekennzeichnet sind. Die übrigen Gruppen zeigen keine typische Richtung der Einstellung, sodass die Diskriminierungsfähigkeit des Merkmals 53 isoliert betrachtet als nicht hoch bezeichnet werden kann. Zusammen mit den übrigen Merkmalen, die besonders typisch für die übrigen Gruppen sind, bildet es jedoch eine Menge von Merkmalen, die die vorliegenden Gruppen sehr gut charakterisiert und somit eine gute Trennung der Klassen bewirkt.
In der Zusammenstellung der übrigen Merkmale ergibt sich bei den betrachteten Verfahren jeweils eine Abweichung von zwei Merkmalen, die die unterschiedlichen Trefferquoten bei IRRM in Tabelle 4.24 erklären. NLIRM gelangt bei 27 Merkmalen unabhängig von η zu der gleichen Merkmalsauswahl. Bei der Betrachtung der jeweils 10 wichtigsten Merkmale ist die Übereinstimmung nicht mehr so groß. Es ergeben sich bei NLIRM bis zu 5 Abweichungen und bei IRRM bis zu 6 Abweichungen bei den ermittelten Mengen.
Der Einsatz von IRRM oder NLIRM kann eine Verbesserung der Trefferquote um 4,1 Prozentpunkte bei Vorgabe von 10 auszuwählenden Merkmalen (beim Vergleich zur mittels FCS ermittelten Trefferquote sind es lediglich 1,6 Prozentpunkte) und eine Erhöhung der Trefferquote um 1,3 Prozentpunkte bei 27 Merkmalen im Vergleich zu voran gegangenen einfacheren Vorgehen mittels Gradientenverfahren bewirken. Eine abschließende erneute Suche nach guten Parametern durch Gridsearch kann bei beispielsweise NLIRM unter Vorgabe der ausgewählten Merkmale die Trefferquote noch auf 85,86% steigern.

Zusammenfassend kann festgehalten werden, dass der Einsatz der nicht linearen Trennung mit anschließender Reduktion der Merkmale mit dem Gradientenverfahren zu einer Auswahl an Merkmalen führt, die eine respektable Trefferquote auf einer reduzierten Datenbasis liefert. Die Verbesserung in Bezug auf die lineare Trennung ist in diesem Fall sehr gering. Daher erscheint die lineare Trennung bei diesen Daten durchaus ausreichend und die aufwändige und zeitintensive Parametersuche bei der nicht linearen Trennung kann somit umgangen werden.

Die durch IRRM oder NLIRM erreichte Trefferquote erbringt lediglich eine geringe Verbesserung gegenüber dem zeitsparenderen Vorgehen, sodass diese Methode im vorliegenden Fall nicht angemessen erscheint, jedoch eine mögliche Alternative bei anderen Datensätzen darstellt.
Ziel des Abschnittes war die Reduzierung der Anzahl der Merkmale zur Generierung eines Fragenkatalogs, um darauf aufbauend Prognosen über die jeweilige Einstellung zur Ernährung zu geben. Je nach Zielsetzung muss individuell entschieden werden, welche Vorgehensweise gewählt werden soll. Diese kann durch die inhaltliche Bedeutung und die zeitliche Vorgaben bestimmt werden. Die Frage nach der Zuverlässigkeit der Ergebnisse, also der Zusammenstellung der Menge der Merkmale, kann hingegen nicht beantwortet werden. Je nach Zielsetzung der einzelnen Verfahren (SVM, LDA) werden unterschiedliche Ergebnisse und daher auch differierende Mengen an Merkmalen generiert, die für die Klassifikation herangezogen werden. Welche die vorzuziehende Menge ist, kann lediglich anhand der damit erzielten Trefferquote bestimmt werden. Dabei schneidet SVM zumindest bei den vorliegenden Daten mit unterschiedlichen Arten der Merkmalsgewichtung überwiegend besser ab.

4.4.5 Zusammenfassung

Der Einsatz von SVM auf Paneldaten hat gezeigt, dass sich diese Methodik sinnvoll im Rahmen des Direktmarketings im Lebensmitteleinzelhandel zur Identifikation von Kaufverhaltenspräferenzen einsetzen lässt. Die Daten wurden so gewählt, dass sich die Vorgehensweise problemlos auf die beispielsweise durch Kundenkarten erhobenen Daten übertragen lässt und nicht an die kostenaufwändige Erhebung von Paneldaten gebunden ist. Das Ziel des Einsatzes von SVM auf den vorliegenden Daten ist die Erzielung langfristigen Erfolgs durch die Bindung des Kunden an das Unternehmen. Dies soll derart geschehen, dass der Kunde ausschließlich Angebote, Werbung, Gutscheine etc. erhält, die für ihn von Interesse sind, weil sie gut zu seinen Gewohnheiten und seinem Lebensstil passen. Dies wird durch die Zuordnung eines Kunden zu vorher definierten Klassen erreicht, die die Ernährungstypen oder Einstellungen zur Ernährung widerspiegeln. Für bestehende Kundengruppen können somit (Direkt-)Marketingmaßnahmen durch den Einsatz von SVM geplant werden. Je nach Ausmaß der Zugehörigkeit zu Klassen, gemessen an den Entscheidungswerten, können diese weiter differenziert werden. Dabei wird durch den Einsatz von Multilabel-Klassifikation die Betrachtung der Kunden als Individuen mit mehreren Präferenzmustern und somit Klassenzugehörigkeiten ermöglicht. Bei der Ausgestaltung der Direktmarketingaktionen können die Ausprägungen der Entscheidungswerte ebenso berücksichtigt werden wie die Nichtzuweisung zu bestimmten Klassen. Je höher die Zugehörigkeit zu einer Klasse ist, desto mehr kommt der Kunde hinsichtlich dieser Klasse in den Genuss der Marketingaktivitäten. Durch die Einführung unterschiedlicher Bereiche können mehrere Angeboten realisiert werden, die sich in ihrer Intensität (z.B. Höhe des eingeräumten Rabatts) und Art (z.B. Werbemail oder Produktprobe) unterscheiden können. Somit können zahlreiche Kundenwünsche be-

dient werden. Die richtige Klassifikation von Beobachtungen einzelner Klassen kann durch ihre besondere Gewichtung vom Anwender gesteuert werden, sodass der Fokus auf die für das Unternehmen besonders interessanten Kundengruppen gelegt wird. Beim Kunden soll somit das Bewusstsein geprägt werden, dass er durch den Einsatz der Kundenkarte die Angebote erhält, die für ihn von Interesse und wichtig sind. Diese Vorgehensweise kann insbesondere bei Anwendungen mit höherpreisigen Produkten verwendet werden, bei denen die gezielte und persönliche Ansprache von Kunden ein entscheidendes Element der erfolgreichen Werbung darstellt.
Zur Reduktion der Anzahl der eingehenden Merkmale wurden unterschiedliche Methoden eingesetzt. Es hat sich gezeigt, dass die Kombination aus einmaliger SVM zur Bestimmung der Relevanz einzelner Merkmale und des darauf aufbauenden Einsatzes des Ellbogenkriteriums bzw. des Kriteriums des zu erwartenden Scores zu sehr guten Ergebnissen führt. Gerade im Vergleich zu dem deutlich zeitaufwändigerem RFE stellt dies eine lohnende Alternative dar. Ebenso wurde herausgestellt, dass der Einsatz von IRRM und NLIRM eine Verbesserung der Trefferquote bewirken kann. Bei der Bestimmung der Anzahl der zu extrahierenden Variablen gilt es den Trade-Off zwischen der Verringerung der Trefferquote einerseits und der größeren Menge an zu erhebenden Merkmalen andererseits zu bestimmen. Das Ziel der Klassifikation und die damit verbundenen Kosten von Fehlklassifikationen sowie Kosten der Merkmalserhebung sollten stets berücksichtigt werden und somit eine jeweils angemessene Variante zur Merkmalsreduktion gewählt werden. Die inhaltliche Bedeutung der ausgewählten Merkmale hat gezeigt, dass der Einsatz von SVM zu besseren Ergebnissen bei der Reduktion der Merkmale im Vergleich zur manuellen Selektion führt. Im Vergleich zur schrittweisen Auswahl der Merkmale bei der Diskriminanzanalyse schneidet SVM im Hinblick auf die jeweils erzielten Trefferquoten ebenfalls größtenteils besser ab, was den Aufwand des methodisch anspruchsvolleren Werkzeugs rechtfertigt.

4.5 Anwendung von SVM im One-to-One-Marketing

Die herkömmliche Ausrichtung der Marketingaktivitäten orientierte sich an dem Einsatz der 4Ps product, price, promotion und place (vgl. *Perreault, McCarthy* (2005)). Die modernen strategischen Überlegungen konzentrieren sich auf die Pflege, Gestaltung und Erhaltung von Geschäftsbeziehungen und zielen eher auf die Beziehung zu den Kunden ab. Die Entwicklung vom Massenmarketing zum individualisierten Relationship Marketing begann in den Neunziger Jahren (*Petrison et al.* (1997)) und wird sich aufgrund der Entwicklung technologischer Möglichkeiten weiter in diese Richtung bewegen. Dies lässt sich durch die Umkehrung der bisherigen Sichtweise von „inside-out“ in „outside-in“ beschreiben (*Bruhn* (2002)) und beinhaltet die stärkere Ausrichtung der unternehmerischen Marketingaktivitäten an den Bedürfnissen und Wünschen der Kunden.
Aus diesem Paradigmenwechsel resultiert die Tendenz zur individuellen Gestaltung

von Kundenbeziehungen, die insbesondere als One-to-One-Marketing durchgeführt wird. Dieses ist als Strategie des CRM zu verstehen, um Kunden durch eine individuelle Ansprache besser und länger an das Unternehmen zu binden (*Peppers, Rogers* (1996)). Gerade im Bereich des E-Commerce kann die Individualisierung durch die natürliche Erhebung der nötigen Daten erfolgreich genutzt werden, um Kunden direkt und ihren Bedürfnissen entsprechend anzusprechen und somit z.B. Empfehlungen für weitere Produktkäufe abgeben zu können (*Strauß, Schoder* (2000)). Dadurch wird das Massenmarketing durch eine persönlichere Form abgelöst und somit der Reiz- und Informationsüberflutung durch das Gießkannenprinzip der Werbeversendung entgegengewirkt. Das entscheidende Ziel dieser Strategie ist die Bindung des Kunden an das eigene Unternehmen durch personalisierte Ansprache. Ziel einer solchen Umsetzung von Marketingaktivitäten ist es, den Kunden auf Basis der zur Verfügung stehenden Informationen, z.B. über sein bisheriges Kaufverhalten oder seinen demografischen Daten, seinen Bedürfnissen entsprechend zu bedienen. So können etwa im Rahmen eines Vielfliegerprogrammes einzelnen Kunden für sie attraktive Angebote über spezielle Flugziele gemacht werden (*Homburg, Krohmer* (2003)). Durch diese Art der Gestaltung von Informationen kann eine große Anzahl von Kunden angesprochen und zugleich eine individuelle Kundenansprache gewährleistet werden. One-to-One-Strategien werden hauptsächlich innerhalb von Recommender-Systemen im E-Commerce umgesetzt. Dies liegt in den geringen Kosten der Datenerhebung und der zunehmenden Menge an Möglichkeiten, Personalisierung im Internet umzusetzen, begründet. Dazu werden häufig Entscheidungsbäume, Regressionsanalyse oder neuronale Netze zur Ermittlung der Empfehlungen angewendet (*Murthi, Sarkar* (2003)). Die Hauptaufgabe im E-Commerce besteht wie im vorliegenden Anwendungsbeispiel aus dem LEH auch, in der Identifikation der Wünsche der Kunden, um sie angemessen ansprechen zu können.
In den Anwendungen der Abschnitte 4.4.2 und 4.4.3 wurde bereits auf die individuelle Ausrichtung der Marketingaktivitäten eingegangen, wobei der Fokus auf der Behandlung von Kundengruppen lag, die sich durch gewisse Eigenschaften (z.B. dem Kaufverhalten) voneinander unterschieden. Hier soll verstärkt auf die Kunden als Individuen eingegangen werden, um somit speziell auf einzelne Kunden zugeschnittene Werbemittel einsetzen zu können. Ein weiterer Unterschied zu Abschnitt 4.4 liegt darin, dass hier der Focus auf eine einzelne Warengruppe gelegt wird, und somit sich die Datengrundlage und die Zusammenstellung der Daten deutlich unterscheidet. Aus diesem Grund spielt hier die inhaltliche Ausrichtung möglicher Werbemaßnahmen keine Rolle, sondern vielmehr die Entscheidung über die Adressaten der Mailings. Zusätzlich wird hier eine zeitliche Komponente mit in die Analyse aufgenommen. Ziel der folgenden Anwendung ist daher der Einsatz von SVM im One-to-One-Marketing, um so Direktmarketingmaßnahmen sinnvoll zu unterstützen und eine Individualisierung zu realisieren.

4.5.1 Problemstellung und Datenbeschreibung

Problembeschreibung

Die Methodik der SVM soll in einer individuellen One-to-one-Strategie im Rahmen der Couponverschickung im Einzelhandel eingebunden werden. Zur Demonstration der Vorgehensweise liegen demografische Kundendaten sowie Kaufhistorien von Kunden verschiedener Lebensmittelmärkte der USA bezüglich vier verschiedener Warengruppen vor. Auf dieser Grundlage soll entschieden werden, welchen Personen Rabattmarken oder Hinweise auf Sonderangebote der betreffenden Warengruppen zugestellt werden. Die hier vorliegenden Daten können durch den Einsatz von Kundenkarten generiert worden sein, mit denen die Kunden bei jedem Kauf identifiziert und das Kaufverhalten dokumentiert wird, sodass die Kunden im Rahmen des Direktmarketings kontaktiert werden können. Gerade im Lebensmitteleinzelhandel ist die Verbreitung dieser Kundenkarten, im Gegensatz zu Kaufhäusern, noch nicht weit vorangeschritten (*Wassel* (2001)). Sie werden aber bereits in manchen Ketten wie real,-, ausgewählten Edeka-Märkten oder bei Rewe eingesetzt. Couponing zählt zu den derzeit am stärksten wachsenden Marketinginstrumenten (*Ploss* (2003)) und gewinnt mit Wegfall des Rabattgesetzes und der Zugabeverordnung in 2001 zunehmend an Bedeutung. Dass sich der Einsatz von Coupons als Werbemittel durchaus lohnt, zeigt eine Studie von *Bauer et al.* (2002), bei der 44,5% der Befragten bereit sind, aufgrund eines Coupons die Einkaufsstätte zu wechseln[32]. Die Menge der eingesetzten Coupons wird derzeit hauptsächlich durch Zeitungscoupons und POS-Coupons gebildet (*Ploss* (2003)). Eine Individualisierung findet nur in den wenigsten Fällen statt. Bei der Karstadt Warenhaus AG wird eine Verknüpfung von Kundenkarteninformationen und dem Einsatz von Coupons vorgenommen (*Franz* (2003)), sodass Angebote individueller gestaltet werden können.
Genau diese Strategie soll auch in den folgenden Anwendungen verfolgt werden. Um Coupons möglichst effektiv einzusetzen und die Kunden an das eigene Unternehmen zu binden, sollen Informationen über das bisherige Kaufverhalten als Basis zur Identifikation der wichtigen Kunden dienen. Der im Anfangsstadium der Kundenkartenverbreitung befindliche Lebensmitteleinzelhandel (LEH) steht bei dieser Untersuchung im Mittelpunkt des Interesses. Da der Austauschcharakter bei der Wahl des Einzelhändlers im Lebensmittelbereich generell hoch ist, sollen denjenigen Kunden, die in naher Zukunft Bedarf an einem bestimmten Produkt haben, Rabattcoupons für die entsprechenden Produkte geschickt werden, um einen Kauf im eigenen Unternehmen sicherzustellen und eine Abwanderung an die Konkurrenz zu verhindern. Ausschlaggebend für eine solche Zuweisung ist die bisherige Kaufhistorie, die durch Kundenkarten erhoben wird. Hat ein Kunde ein bestimmtes Produkt, bzw. ein Produkt aus einer bestimmten Warengruppe länger nicht gekauft, so ist die Wahrscheinlichkeit für einen Kauf in den kommenden Perioden höher und sollte durch eine Couponversendung unterstützt werden.

[32]Im Gegensatz dazu kommt für 26% der Befragten ein Wechsel nicht in Frage. 29,4% sind unentschlossen.

Diese Muster innerhalb der Kaufhistorien werden von SVM erkannt, sodass eine Klassifikation auf Basis der bisherigen Kaufhistorien möglich wird. Der Einsatz von SVM hat hierbei eine Individualisierung der Werbung zum Ziel, wobei einem Kunden auf seine Bedürfnisse und seinem bisherigen Kaufverhalten zugeschnittene Couponbündel angeboten werden sollen. Dabei werden unter Coupons Rabattgutscheine verstanden, die dem Kunden einen Preisnachlass auf bestimmte Artikel gewähren. Diese können als Direktmailing per Post oder als e-Mail-Coupons an die Kunden versendet werden. Hierbei muss allerdings im Sinne des Permission Marketing eine Einwilligung des Kunden über den Empfang unaufgeforderter Werbung vorliegen. Andere Coupons wie Coupon-Anzeigen werden hier hinsichtlich des intendierten Ziels des One-to-One-Marketings außer Acht gelassen.

Datenbeschreibung

Grundlage eines erfolgreichen personalisierten Couponings ist eine gepflegte Kundendatenbank, die Auskunft über demografische Daten und das jeweilige Kaufverhalten gibt (*Huldi* (2003)). Diese kundenbezogenen Daten können durch Kundenkarten oder durch ein aufwändiges Haushaltspanel erhoben werden. Im vorliegenden Fall werden auf dem AC Nielsen Haushaltspanel zweier amerikanischer Städte basierende Daten herangezogen, die vom James M. Kilts Center der Universität von Chicago im Internet zur Verfügung gestellt werden[33]. Die ausgewählten Bezirke geben die typische demografische Struktur der amerikanischen Bevölkerung wieder. Zur Bestimmung und Auswahl der zu generierenden individuellen Couponbündel werden exemplarisch vier Warengruppen („Waschmittel“, „Ketchup“, „Suppen“ und „Yoghurt“) herangezogen. Die Vorgehensweise kann bei Bedarf ohne weiteres auf mehrere Warengruppen erweitert werden. Ziel ist die Versendung von Coupons für die hier ausgewählten Warengruppen auf Basis der Bedürfnisse der Kunden. Es werden hier lediglich Handelscoupons betrachtet, da sich die Anwendung auf die Umsetzung innerhalb einzelner LEH-Läden bezieht und diese bei ihren Kundenbindungsmaßnahmen unterstützen soll, wohingegen die Einflussnahme von Seiten des Handels bei Herstellercoupons sehr beschränkt ist.
Damit die Muster im Kaufverhalten durch SVM erkannt werden können, werden die auf eine einzelne Warengruppe bezogenen Kaufhistorien durch 0-1-Folgen repräsentiert. Eine 1 (bzw. 0) gibt an, dass in der betreffenden Woche[34] Produkte der entsprechenden Warengruppe gekauft (bzw. nicht gekauft) wurden. Da nur das von dem Unternehmen erhobene Kaufverhalten eines Kunden in die Analyse eingeht, und demnach die Kaufaktivitäten bei anderen Unternehmen nicht berücksichtigt werden können, deutet eine 0 nicht zwingend auf einen generellen Nichtkauf des Produktes hin. Der Kunde kann das Produkt ebenso bei der Konkurrenz gekauft haben. Je nachdem, wie lange die Kaufhistorie zurückreichen soll, um typische Kaufverhaltensmuster bezüglich einer Warengruppe abzubilden, ergeben sich unterschiedlich lange Zeitfenster. Die Klassenbildung erfolgt auf Basis der Einträge

[33]Quelle: http://gsbwww.uchicago.edu/kilts/research/db/erim/, Zugriff: 12.6.2004

[34]Die Festlegung der Periodenlänge im Wochenzyklus erscheint bei der Versendung von Coupons ein geeignetes Maß

der folgenden Wochen. Dies führt dazu, dass Coupons nach den individuellen Kaufabfolgen der einzelnen Kunden verschickt werden können. Falls in den kommenden Wochen kein Produkt der entsprechenden Warengruppe gekauft wird, erhält der Kunde einen Coupon für diese Warengruppe, um ihn zu einem Kauf im eigenen Unternehmen zu ermutigen. So sollen die Kunden genau dann animiert werden, wenn der geschätzte Bedarf vorhanden, die Wahrscheinlichkeit für einen Kauf beim eigenen Unternehmen jedoch gering ist. Wie viele Wochen für die Bildung der Klasse herangezogen werden, kann pro Warengruppe unterschiedlich sein. Eine beispielhafte Darstellung möglicher Daten ist in Abbildung 4.13 gegeben.

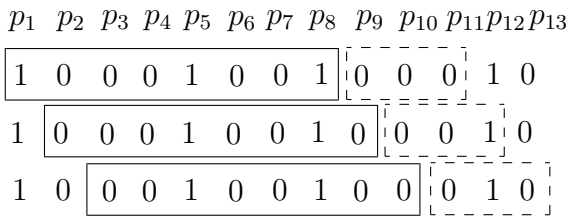

Abbildung 4.13: Mögliche Darstellung der Kaufhistorie eines Haushaltes

Dargestellt wird eine exemplarische Abfolge von Käufen einer Warengruppe eines Haushaltes, die über 13 Perioden $p_1, \ldots, p_{13}$ hinweg dokumentiert sind. Die Rahmen um die Daten eines acht Wochen Zyklus markieren die Perioden, die als Merkmalsvektor zur Beschreibung des bisherigen Kaufverhaltens dienen. Die gestrichelten Kästchen wiederum umranden die Perioden, die zur Bestimmung der Klassenzugehörigkeit beitragen. Dieser Haushalt geht aufgrund der bisher getätigten Käufe zunächst mit der ersten Kaufabfolge $p_1, \ldots, p_8$ (durchgezogener Kasten in der ersten Reihe) als Trainingsvektor der Klasse „-1“ in die Optimierung ein. Ein Haushalt wird zur Klasse „+1“ zugeordnet, wenn er in den kommenden drei Perioden $p_9,\ p_{10},\ p_{11}$ mindestens einmal die Produkte der betreffenden Warengruppe im auswertenden Unternehmen kauft. Falls kein Kauf zu erwarten ist, fällt der Kunde bzw. Haushalt in die Klasse „-1“. Sollen die Kunden in einer der folgenden Perioden klassifiziert werden, so wird das Fenster, welches 8 Perioden als Training umfasst, verschoben, sodass immer die letzten 8 Perioden als Trainingsbasis dienen und die darauf folgenden 3 Perioden ausschlaggebend für die Klassenzugehörigkeit sind. Im vorliegenden Beispiel in Abbildung 4.13 würde der betreffende Haushalt zunächst als Nichtkäufer (Klasse „-1“) klassifiziert werden, bei Verschieben des Fensters um jeweils eine Periode würde er aufgrund des Kaufes in Periode p_{12} in die Klasse der Käufer sortiert werden. Bei der Klassifikation neuer Beobachtungen liegen nur die Daten aus den acht Wochen vor. Durch die mittels SVM ermittelten Trennebene zwischen den Käufern und Nichtkäufern können so Zuweisungen für unklassifizierte Beobachtungen vorgenommen werden, sodass Aussagen darüber getroffen werden können, ob ein Kauf des betreffenden Produktes in den kommenden Perioden zu erwarten ist oder nicht. Durch die Verschiebung des Zeitfensters, d.h. durch die Betrachtung lediglich der letzten acht Perioden, wird die Dynamik innerhalb des

Kaufverhaltens mitberücksichtigt. Dies führt dazu, dass ein Haushalt nicht generell zu einer Klasse gehört wie in den übrigen Anwendungen innerhalb dieses Kapitels, sondern sich die Klassenzugehörigkeit in Abhängigkeit der jeweiligen Kaufhistorie im Zeitablauf ändern kann.
Die Länge des in Abbildung 4.13 dargestellten Zeitfensters kann und sollte je nach vorliegender Warengruppe und resultierendem Kaufverhalten variiert werden. Die vorliegende Datenbasis wird hier aus den Kaufverhaltensmustern von insgesamt 3189 Haushalten gebildet. Die Auswertungen dieser Daten beziehen sich zunächst auf die Warengruppe Waschmittel, bei der ein Zeitfenster von 15 Wochen, also 15 Perioden, gewählt wird, da diese Warengruppe nicht häufig gekauft wird und durch diese Fensterbreite ein möglicher Kauf sehr wahrscheinlich dokumentiert wird. Weiterhin ist zu erwarten, dass ein Kunde aufgrund eines Coupons für Waschmittel eher geneigt ist, die Einkaufsstätte zu wechseln. Es soll prognostiziert werden, ob in den nächsten drei Perioden ein Kauf zu erwarten ist oder nicht, um anschließend durch die Bereitstellung eines Coupons die Kunden zu einem Kauf im eigenen Hause zu animieren. Die einzelnen Muster sind aufgrund der demografischen Beschreibung der Haushalte durchaus nachvollziehbar. So kaufen große Haushalte mit mehr als 6 Mitgliedern tendenziell häufiger innerhalb der 15 betrachteten Wochen als kleine Haushalte mit nur einem Mitglied, die überwiegend gar nicht, einmal oder zweimal innerhalb der 15 Wochen Waschmittel gekauft haben. Um den Einsatz von SVM hier zu zeigen, werden zunächst 15 Wochen als Basis fixiert.

4.5.2 Auswertung

Um zunächst einen Überblick darüber zu bekommen, welche Trefferquoten bei einer oben beschriebenen Klassifikation zu erwarten sind, werden in Tabelle 4.25 die Ergebnisse zusammengefasst, die sich nach vorangegangener Parameterselektion mittels Kreuzvalidierung auf den zufällig ausgewählten 797 Testdaten[35] ergeben.

Verfahren	Parameter	TQ Kl. „+1“	TQ Kl. „-1“	Gesamt-TQ
SVM Lin.	$C = 1000, c_1 = 1,8$	63,05%	71,12%	68,13%
SVM RBF	$\gamma = 0,0002$	62,71%	71,31%	68,13%
	$C = 500, c_1 = 1,8$			
SVM RBF	$\gamma = 5 \cdot 10^{-5}$	56,27%	77,89%	69,89%
	$C = 100, c_1 = 1,8$			
LDA		61,69%	73,90%	69,38%
C4.5		38,64%	85,06%	67,88%
MLP		46,10%	79,48%	67,13%

Tabelle 4.25: Resultierende Ergebnisse einer Klassifikation auf Basis der Kaufhistorien für die Warengruppe Waschmittel

[35]Der Umfang der Testdaten ergibt sich aus dem in der 4-fachen Kreuzvalidierung verwendeten Verhältnis von Trainings- und Testdaten. Somit wird 1/4 der zur Verfügung stehenden Daten zum Testen verwendet. Die restlichen 2392 Beobachtungen dienen der Optimierung der SVM.

Die Haushalte, die in den jeweils kommenden drei Wochen Waschmittel kaufen werden, bilden einen Anteil von 37%, sodass nach dem Größte-Gruppen-Kriterium mindestens eine Gesamttrefferquote von 63% erreicht werden sollte[36]. Die eingesetzten Verfahren kommen zu sehr ähnlichen Ergebnissen, was die Gesamttrefferquote betrifft. So wird mittels lineare SVM eine Gesamttrefferquote von 68,13% erreicht, was durch den Einsatz des nicht linearen Radialbasis-Kerns auf 69,89% gesteigert werden kann. Im Mittelpunkt des Interesses liegt bei der vorliegenden Anwendung die Maximierung der Gesamttrefferquote bei gleichzeitiger Ausgewogenheit der Ergebnisse der beiden Klassen. Dies wird durch die Höhergewichtung der deutlich kleineren Klasse „+1“ erreicht. Ein Verschicken eines Coupons an Kunden, die in den kommenden Perioden Waschmittel mit großer Wahrscheinlich im eigenen Unternehmen kaufen, kann zu einem unerwünschten Umsatzverlust durch Gewährung unnötiger Rabatte führen, da diese Kunden auch ohne Coupon gekauft hätten. Somit sollen nur diejenigen Kunden durch die Coupon-Versendung erreicht werden, bei denen diese Aktivierung notwendig erscheint. Dies bedeutet, dass beide Klassen bei der Trennung möglichst hohe Trefferquoten aufweisen sollten. Unter diesem Gesichtspunkt sind die Ergebnisse von SVM, sowohl mit linearer als auch mit nicht linearer Trennung, als vergleichbar gut zu bezeichnen. Eine Veränderung der Parameter zugunsten der Gesamttrefferquote würde hierbei zu Lasten der Trefferquote der Klasse „+1“ gehen, was hinsichtlich des zu verfolgenden Ziels zu vermeiden ist.
Um eine Differenzierung bei der Gestaltung der Coupons zu erreichen, können die Entscheidungswerte eingesetzt werden. Die bei einer linearen Trennung mittels SVM ($C = 1000,\ c_1 = 1,8$) resultierende Situation wird in Abbildung 4.14 dargestellt.

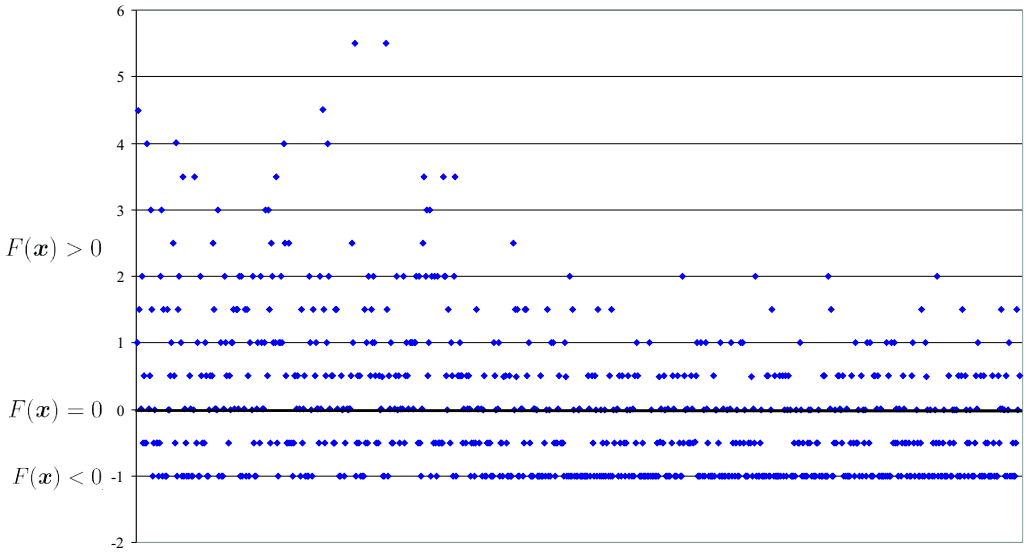

Abbildung 4.14: Differenzierung der Haushalte nach Gewährung von unterschiedlich hohen Rabattwerten

[36]Nach dem Proportional-Chance-Kriterium wird eine minimale Trefferquote von 53% erwartet.

Es wurden die 797 Beobachtungen auf Basis ihrer 15 Wochen umfassenden Kaufhistorie den beiden Klassen „Käufer“ (Klasse „+1“) und „Nichtkäufer“ (Klasse „-1“) zugewiesen. Da die Merkmalsvektoren lediglich die Einträge 0 oder 1 enthalten, kann durch die lineare Kombination dieser Vektoren in Verbindung mit einer hohen Anzahl an „bounded Support Vektoren“[37] nur eine endliche Menge an Entscheidungswerten generiert werden, was die regelmäßige Erscheinung in Abbildung 4.14 begründet. Eine Differenzierung der Couponversendung erfordert eine Strukturierung der insgesamt erreichten Entscheidungswerte durch kleinere Bereiche. Eine Einteilung des Intervalls der angenommenen Entscheidungswerte in fünf verschiedene Bereiche erscheint in der vorliegenden Situation adäquat. Dies wird in Tabelle 4.26 veranschaulicht.

Bereich	Intervall	Häufigkeit	Trefferquote	Strategie
1	$[1;6]$	127	72,44%	kein Rabatt
2	$[0;1[$	204	46,08%	5% Coupon
3	$[-0,5;0[$	169	72,78%	10% Coupon
4	$[-1;-0,5[$	284	79,58%	20% Coupon
5	$[-2;-1[$	13	61,54%	25% Coupon

Tabelle 4.26: Aus der linearen Entscheidungsfunktion resultierende mögliche Bereiche zur differenzierten Behandlung von Haushalten

Das durch die Entscheidungswerte aufgespannte Intervall [-2;6] wird in fünf Abschnitte geteilt, die die Couponversendung unterschiedlich beeinflussen. So fallen diejenigen Haushalte, die einen Entscheidungswert kleiner -1 erzielen, recht sicher in die Klasse der Nichtkäufer (Klasse „-1“). Um diese Kunden dennoch zu einem Kauf zu bewegen, werden hier Coupons versendet, die einen Rabatt in Höhe von 25% gewähren.
Da sehr viele der Beobachtungen dicht an der Ebene positioniert sind und diese Daten häufig falsch klassifiziert wurden, wird den entsprechenden Haushalten ebenfalls ein Coupon zugesandt. Der gewährte Preisnachlass fällt umso höher aus, je sicherer eine Beobachtung in die Klasse der Nichtkäufer fällt. Die Haushalte, deren Beobachtung im Intervall [-1;-0,5[liegt, werden mit einem 20%-Gutschein bedacht, die im Intervall [-0,5;0[liegenden mit 10%. Haushalte, die relativ sicher als Käufer klassifiziert wurden, erhalten keinen Rabattgutschein (Intervall [1;6]). Werden die Kunden mit einem geringen Entscheidungswert der Klasse der Käufer zugeteilt, wird ihnen nur ein geringer Rabatt von 5% eingeräumt (Intervall [0;1[). Dieses Vorgehen verhindert, dass Kunden umsatzmindernde Preisnachlässe gewährt werden, die mit sehr hoher Wahrscheinlichkeit sowieso im eigenen Unternehmen kaufen werden.
Durch die Differenzierung der Rabattwerte wird dem Umstand Rechnung getragen, dass eine Beobachtung umso sicherer zu einer Klasse gehört, je größer der

[37] Als „bounded “ werden diejenigen Support Vektoren bezeichnet, deren zugehörige Lagrange-Multiplikatoren α_i die Obergrenze C (bzw. $c_i C$) annehmen.

absolute Entscheidungswert ausfällt (vgl. dazu Abschnitt 3.7.2). Dieses Phänomen ist ebenfalls deutlich an der Trefferquote der einzelnen Bereiche zu erkennen. Diese nimmt ab je näher sich der betreffende Bereich an der Entscheidungsebene befindet. Beobachtungen im ersten Bereich werden noch mit einer Trefferquote von 72,44% richtig klassifiziert, wohingegen die Haushalte im zweiten Intervall $[0, 1[$ beispielsweise nur noch zu 46% richtig klassifiziert werden.

Im vorliegenden Beispiel soll zusätzlich untersucht werden, ob durch den Einsatz von SVM in der Praxis Umsatzsteigerungen durch die differenzierte Kundenansprache zu erwarten sind. Dazu wird die in Tabelle 4.26 vorgenommene Einteilung der Kunden mit drei alternativen Vorgehensweisen ohne den Einsatz von SVM verglichen. Die Versendung keines Coupons (1. Alternative), also die Durchführung keinerlei Marketingaktivitäten, die Versendung von Coupons in Höhe von 10% an alle registrierten Kunden (2. Alternative) und die Versendung von Coupons in Höhe von 10% an Kunden, die einen Bedarf an Waschmittel vermuten lassen (3. Alternative), sollen mit den Ergebnissen von SVM verglichen werden. Um letztendlich den resultierenden Umsatz aller Alternativen berechnen zu können, müssen einige Annahmen getroffen werden. So wird davon ausgegangen, dass die durchschnittlichen Ausgaben für den Kauf von Waschmittel mit 5 Euro angegeben werden können. Da bei SVM und den übrigen drei Alternativen diejenigen Kunden aktiviert werden sollen, die nicht als Käufer in den nächsten drei Wochen zu erwarten sind, wird eine Wechselwahrscheinlichkeit vom „Nichtkauf“zum „Kauf“in Abhängigkeit der Rabatthöhe des jeweils versendeten Gutscheins angenommen[38]. Pro Rabattprozentpunkt steige die Wahrscheinlichkeit, zu dem Unternehmen zu wechseln, bzw. hier einen Kauf zu tätigen um 1,5%. Kunden, die einen Bedarf aufweisen, und nicht als wahre Käufer in der nächsten Zeit zu zählen sind, werden der Gruppe zugeordnet, bei der diese Wechselwahrscheinlichkeit auftritt. Dieser Bedarf wird auf Basis der Kaufhistorie ermittelt. Da von einem Intervall von acht Wochen ausgegangen wird, in dem im Durchschnitt einmal Waschmittel gekauft wird, wird denjenigen Kunden ein Bedarf unterstellt, die in den vergangenen fünf Perioden keinen Kauf getätigt haben (zusammen mit den drei Prognoseperioden ergibt sich ein Intervall im Umfang von acht Wochen). Es kann natürlich nicht überprüft werden, ob die Kunden bei einem anderen Einzelhändler ihren Bedarf an Waschmittel gedeckt haben, da nur die Informationen der eigenen Kundenkarteneinsätze dokumentiert worden sind.
Bei der ersten Alternative wird keine Marketingaktivität durchgeführt. Da in den nächsten drei Wochen 295 der 797 zu betrachtenden Kunden einen Kauf tätigen werden, resultiert dies in einem Umsatz von $295 \cdot 5 = 1475$ Euro. Werden hingegen ungeachtet des Kaufverhaltens an alle Kunden ein Gutschein über 10% Rabatt geschickt, so ergibt sich eine Umsatzsteigerung auf 1528,65 Euro. Dabei wurde errechnet, dass aufgrund der aktivierten Kunden zusätzlich zu den bisherigen 295 Käufern noch 45 Käufer zu erwarten sind. Dies resultiert aus der in diesem Fall vorliegenden Wechselwahrscheinlichkeit von $10 \cdot 1,5 = 15\%$, mit der die 298 Kunden,

[38]Für die Versendung von Coupons werden hierbei keine Kosten angenommen, da diese z.B. per e-mail versendet werden können.

die zwar nicht kaufen, aber einen Bedarf aufweisen, zum betreffenden Unternehmen wechseln. Damit ergibt sich ein Umsatz von $(295 + 45) \cdot 5 \cdot 0,9 = 1528,65$ Euro für die zweite Alternative. Werden in der dritten Alternative noch zusätzlich diejenigen Kunden identifiziert, die einen Bedarf aufweisen, so können Umsatzverluste vermieden werden. Nur denjenigen Kunden, die einen Bedarf aufweisen, wird ein Coupon zugesendet. Dadurch werden Verluste, die durch Gewähren von unnötigen Rabatten entstehen, vermieden. Diese Einsparungen ergeben eine Umsatzsteigerung auf 1627,65 Euro, wobei bisher noch kein Methodeneinsatz nötig war.
Wird zur Ermittlung der richtigen Ansprache der Kunden SVM hinzugezogen, so ergibt sich eine mögliche, noch feinere Einteilung der Behandlung, die in Tabelle 4.26 dokumentiert ist. In Tabelle 4.27 werden die resultierenden Aktionen und die jeweils folgende Anzahl der zugeordneten Beobachtungen zusammengefasst.

Coupon	Käufer	Nicht-käufer	Bedarf	Wechsel-w.-keit	Anzahl Wechsler	Käufer gesamt	Umsatz
0%	92	35	3	0	0	92	460
5%	94	110	29	0,075	2	96	456
10%	46	123	68	0,15	10	56	252
20%	58	226	198	0,3	59	117	468
25%	5	8	0	0,375	0	5	18,75
Summe	295	502	298	–	71	366	**1654,75**

Tabelle 4.27: Aus linearer SVM resultierende Bereiche der Entscheidungswerte und Häufigkeiten der Kunden

Die Spalte „Bedarf“ gibt die Anzahl der Nichtkäufer an, die einen Bedarf aufweisen und somit evtl. wechseln würden. Beispielsweise fallen 204 Kunden (94 Käufer und 110 Nichtkäufer) in die Gruppe derjenigen Kunden, die einen 5%-Coupon erhalten. Von diesen Nichtkäufern weisen insgesamt 29 aufgrund ihrer Kaufhistorie einen Bedarf auf, sodass sie mit einer Wahrscheinlichkeit von 0,075 zu der Gruppe der Käufer wechseln würden. Diese Wahrscheinlichkeit ergibt sich hier aus $1,5 \cdot 5\% = 7,5\%$. Dies resultiert in zwei zusätzlichen Käufern. Die insgesamt 96 Käufer dieses Bereichs bewirken somit einen Umsatz von $96 \cdot 5 \cdot 0,95 = 456$ Euro. Alles in allem wird die ursprüngliche Anzahl an Käufern (295) um insgesamt 71 aktivierte und wechselnde Käufern erweitert, was zu einem Gesamtumsatz von 1654,75 Euro führt. Wird nun dieses Ergebnis mit den Umsätzen der drei anderen Alternativen verglichen, so ergeben sich Umsatzsteigerungen, die in Tabelle 4.28 zusammengefasst sind. Mit zunehmendem Aufwand in der Differenzierung der Kunden erhöht sich auch der zu erwartende Umsatz. Wird eine Auswertung der Kundendaten vorgenommen (Alternative 3), so resultiert dies in einem deutlich höheren zu erwartenden Umsatz. Der zusätzliche Einsatz von SVM führt hier zu einer Umsatzsteigerung von mindestens 1,66%. In diesem Beispiel sind lediglich 797 Kunden betrachtet worden, von denen 366 als Käufer in Betracht gezogen werden. Die Erhöhung des Umsatzes fällt absolut gesehen entsprechend niedrig aus.

Alternative	Umsatz	Steigerung bei SVM
1 (keine Aktivität)	1475	+12,19%
2 (10%-Gutschein für alle)	1528,65	+8,25%
3 (10%-Gutschein für Kunden, die Bedarf aufweisen)	1627,65	+1,66%
SVM (entsprechend Tabelle 4.26)	1654,75	–

Tabelle 4.28: Aus Einsatz von SVM resultierende Umsatzsteigerungen

Wird dies jedoch auf reale Daten eines Kundenbindungsprogramm hochgerechnet, so ergeben sich deutlich höhere Umsatzsteigerungen. Wird zum Beispiel das Kundenbindungsprogramm „Payback“ betrachtet, so liegen bereits 27 Millionen eingesetzte Kundenkarten vor[39]. Geht man davon aus, dass davon 8 Millionen Kunden Käufer des kooperierenden Lebensmittelmarktes sind[40], so kann mit knapp 3,6 Millionen Käufern für dieses Beispiel gerechnet werden, wenn die obigen Verhältnisse (letztendlich 366 Käufer von 797 Kunden) beibehalten werden. Unter Berücksichtigung der unterschiedlichen Rabattgewährung und einem Preis von 5 Euro für das Waschmittel entsprechen hier 1,6% Umsatzsteigerung einer Erhöhung des Umsatzes um etwa 300.000 Euro im Vergleich zur trivialen bzw. einfacheren Vorgehensweise. Umsatzsteigerungen dieser Größenordnung rechtfertigen nun die Anschaffung und den Einsatz eines Systems zur Klassifikation wie SVM. Diese Art der Prognosemethode kann bei Ausweitung auf andere Branchen oder den internationalen Kontext ebenfalls angewendet werden. Ein Einsatz des Klassifikationsinstrumentes unter den genannten Annahmen kann sich also monetär lohnen. Dabei ist nicht nur der hierbei erhöhte Umsatz zu berücksichtigen, sondern auch die durch den Einkauf induzierten Cross-Selling-Effekte, sodass sich der Einsatz von SVM zur Behandlung der Kunden sehr gut eignet.

Ein alternatives Verfahren im Rahmen der Kaufverhaltensanalyse bildet z.B. das bekannte Negativ-Binomial-Modell (*Decker, Wagner* (2002)). Dabei werden, wie in Anhang A.3 beschrieben, mit Hilfe der Kaufklassenhäufigkeiten für 15 Perioden die bedingten Wahrscheinlichkeiten für den Kauf bzw. Nichtkauf von Produkten aus der Warengruppe „Waschmittel“ berechnet, woraus dann die Trefferquoten der beiden Klassen ermittelt werden können. Diese liegen bei 76,27% (Klasse „+1“) und 56,37% (Klasse „-1“), woraus sich eine Gesamttrefferquote von 63,74% ergibt. Diese liegt deutlich unterhalb der Ergebnisse, die mit SVM erzielt werden können. Der Einsatz von Markovketten unter Berücksichtigung von Übergangswahrscheinlichkeiten bringt ebenfalls keine Verbesserung der Ergebnisse, wie dies in Anhang A.4 beschrieben wird. Dies zeigt, dass SVM im Vergleich zu den in diesem Bereich herkömmlichen Methoden vergleichbare und zum Teil sogar bessere Ergebnisse im Rahmen der Auswertung des Kaufverhaltens produzieren.

[39] Quelle: www.payback.de, Zugriff: 22.8.2005
[40] Quelle: www.metrogroup.de, Zugriff: 22.8.2005

Als Ziel dieses Abschnitts wurde die Anwendung von SVM im One-to-One-Marketing festgesetzt. Die obigen Ausführungen beinhalten die Versendung eines Coupons auf Basis individueller Kaufhistorien. Dies kann durch die Erweiterung des Ansatzes auf mehrere Warengruppen auf die bereits oben erwähnten Couponbündel ausgedehnt und somit die Gestaltung der Werbemaßnahmen individueller eingerichtet werden. Damit wird die Anwendung dem Ziel des One-to-One-Marketings durch die Gewährleistung individueller Angebote gerecht. Dazu stehen bei den vorliegenden Daten drei weitere Warengruppen („Suppen“, „Ketchup“ und „Yoghurt“) zur Verfügung, die bei anderen Anwendungen bei Bedarf um weitere ergänzt werden könnten. Wie die folgenden Ausführungen zeigen, kann diese Art der parallelen Betrachtung mehrerer Klassen als Erweiterung der Multilabel-Klassifikation gesehen werden, da die gleichzeitige Zuordnung eines Haushaltes zu mehreren Klassen untersucht wird. Der Unterschied liegt in der für jede Warengruppe unterschiedlich langen Datengrundlage und der damit verbundenen nötigen Berechnung einer SVM pro Warengruppe. Ein derartiger Einsatz von SVM stellt für jeden Haushalt auf Basis seiner bisherigen Kaufhistorie ein individuelles Couponbündel zusammen. So kann zum Beispiel der Fall auftreten, in dem ein Kunde für die Warengruppen „Suppe“ und „Yoghurt“ einen Coupon erhält, für die beiden anderen Warengruppen jedoch nicht. Die beiden Coupons können unterschiedlich hohe Rabatte beinhalten, und eine weitere Individualisierung des Angebots wäre somit möglich. Dies bildet eine nicht triviale Anwendung von SVM, die speziell auf die Art der Ergebnisdarstellung dieses Verfahrens zugeschnitten ist. Wie die Ergebnisse einer derartigen Anwendung von SVM im One-to-One-Marketing bei vorliegenden Daten aussehen, zeigt Abbildung 4.15.

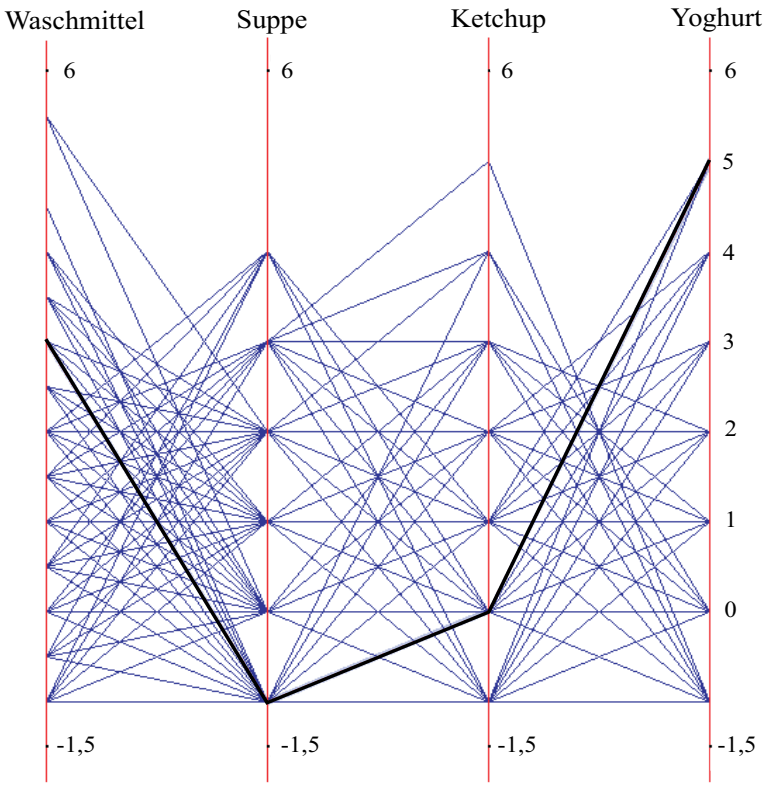

Abbildung 4.15: Visualisierung der Entscheidungswerte für die vier Warengruppen mit Hilfe von Parallelkoordinaten

Es wird davon ausgegangen, dass im Rahmen einer Couponversendung für die folgenden drei Wochen Rabatte auf die vier Warengruppen gewährt werden sollen. Die Empfänger der Coupons werden wie oben durch diejenigen Kunden gebildet, die in diesen Wochen einen Bedarf dieser Warengruppen aufweisen und diesen möglicherweise bei einem Mitbewerber decken. Der Umfang der eingehenden Kaufhistorien unterscheidet sich je nach Warengruppe. Da bei „Ketchup“ ähnlich wie bei „Waschmittel“ von einer geringen Kauffrequenz ausgegangen wird, umfassen die eingehenden Vektoren hierbei wiederum 15 Wochen. Bei den übrigen Warengruppen „Suppe“ und „Yoghurt“ wird auf eine kürzere Zeitspanne (6 bzw. 5 Wochen) zurückgegriffen. Abbildung 4.15 zeigt die resultierenden Entscheidungswerte für die Kunden, die hinsichtlich der Erstellung von Couponbündel klassifiziert werden. Für jede Warengruppe wurde zuvor mittels Kreuzvalidierung die besten Parameterkombinationen für die lineare Trennung ermittelt[41]. Wie deutlich zu erkennen ist, bewegen sich die Ausprägungen der vier Entscheidungsfunktionen in ähnlichen Bereichen. Um zu einer differenzierten Zusammenstellung von Couponbündeln für die einzelnen Haushalte zu gelangen, kann daher die Einteilung der Intervalle in Tabelle 4.26 für jede Warengruppe als Vorlage dienen. Für den Haushalt, der in Abbildung 4.15 fett markiert ist, ergeben sich die folgenden Werte bezüglich der Entscheidungsfunktionen

[41] Für die Warengruppe „Suppe“ wurde $C = 500,\ c_1 = 1,55$, für die Warengruppe „Ketchup“ $C = 500,\ c_1 = 5,9$ und für die Warengruppe „Yoghurt“ $C = 1000,\ c_1 = 3,9$ gewählt.

der vier Warengruppen[42]:

$$\begin{aligned} F_W(\boldsymbol{x}) &= 2,99 \\ F_S(\boldsymbol{x}) &= -0,99 \\ F_K(\boldsymbol{x}) &= 0,00 \\ F_Y(\boldsymbol{x}) &= 4,99 \end{aligned}$$

Durch den Vektor $\boldsymbol{x}$ wird die bei den Warengruppen jeweils eingehende Kaufhistorie berücksichtigt, sodass $\boldsymbol{x}$ je nach Warengruppe unterschiedliche Länge aufweist. Unter Berücksichtigung der Einteilung in Tabelle 4.26 wird diesem Haushalt aufgrund der ermittelten Werte für die Warengruppen „Suppe“ und „Ketchup“ jeweils ein Coupon zukommen. Für die Warengruppen „Waschmittel“ und „Yoghurt“ erhält der Haushalt keinen Coupon, da die positiven Werte auf einen sicheren Kauf schließen lassen und eine zusätzliche Aktivierung dieses Kunden nicht nötig erscheint. Bei der Bestimmung der Gewährung der Rabatte ist die Höhe der Entscheidungswerte relevant. Im vorliegenden Fall ergibt sich die Zusammenstellung

Kein Coupon für die Warengruppe „Waschmittel“
Coupon für die Warengruppe „Suppe“ in Höhe von 20%
Coupon für die Warengruppe „Ketchup“ in Höhe von 5%
Kein Coupon für die Warengruppe „Yoghurt“.

Bei den in Abbildung 4.15 visualisierten Entscheidungswerten ist für die drei weiteren Warengruppen eine zu der Warengruppe „Waschmittel“ unterschiedliche Einteilung der Bereiche denkbar. So wäre die Einführung von lediglich vier Bereichen mit keinem Coupon, Coupons mit 5%, 10% und 20% Rabatt ebenfalls denkbar, um die Haushalte zu differenzieren. Die Wahl einer geeigneten Einteilung und folgender Implikationen für die entsprechenden Beobachtungen kann nicht pauschalisiert werden, sondern hängt immer sowohl von dem zugrunde liegenden Datenmaterial, als auch insbesondere von der vorgenommenen Trennung und deren Generalisierungsfähigkeit ab.

Die dynamische Variante der SVM, das Online-Training, kann hier im Gegensatz zu anderen Situationen nicht sinnvoll angewendet werden. Dies liegt insbesondere daran, dass das Training des zu verwendenden Modells zu einem bestimmten Zeitpunkt stattfindet, um die Couponversendung für die kommenden Perioden zu steuern. In diesem Fall liegen die erforderlichen Daten, also die Haushalte mit ihren entsprechenden Kaufhistorien bereits vollständig vor[43]. Es findet keine sukzessive Erweiterung der Datenbasis statt, sodass die neu eingetroffenen Kunden in das Modell integriert werden können. Da auf das Kaufverhalten der jeweils vergangenen Perioden zurückgegriffen wird, muss bei einer Planung der Couponversendung jeweils ein neues Modell berechnet werden, sodass sich die dynamische Komponente

[42] Die Indizes der Entscheidungsfunktionen geben die Warengruppe an (W: „Waschmittel“; S: „Suppe“; K: „Ketchup“ und Y: „Yoghurt“)

[43] Selbstverständlich kann der Umfang des Kundenstammes um weitere Neukunden ergänzt werden, wenn die erforderlichen Daten vorliegen, sodass durchaus eine Erweiterung der Datenbasis durchgeführt wird.

von SVM hier nicht geeignet ausnutzen lässt. Allerdings liegt bei dieser Anwendung eine andere Art der Dynamik vor. Einem Kunden wird je nach Zeitpunkt der Klassifikation ein neues Couponbündel auf Basis des jeweils bisherigen Verhaltens zugewiesen, im Gegensatz zur herkömmlichen Anwendung, in der die Zuweisung eines Kunden vorwiegend statischer Natur ist. Dazu wird das oben verwendete Zeitfenster mit der Zeit verschoben und aktualisiert. Da ein Haushalt je nach Zeitpunkt also unterschiedliche Klassenzugehörigkeiten aufweisen kann, sind diese sowie die Merkmalsvektoren der Datenbasis als dynamisch zu bezeichnen. Dies wird durch die Interpretation der Merkmalsvektoren als das „Kaufverhalten der letzten 15 Perioden“ (im Fall der Warengruppe „Waschmittel“) ermöglicht. Eine Überlappung der jeweiligen Kaufhistorien ist nicht ausgeschlossen. Eine erneute Optimierung des Modells zu Beginn einer neuen Planungsperiode berücksichtigt das sich evtl. im Zeitablauf ändernde Kaufverhalten der Kunden, sodass das Modell ständig aktualisiert wird. Wird das oben verwendete Zeitfenster um drei Wochen verschoben, so erhält der in Abbildung 4.15 fett markierte Haushalt folgende Entscheidungswerte:

$$\begin{aligned} F_W(\boldsymbol{x}) &= 2,00 \\ F_S(\boldsymbol{x}) &= 0,00 \\ F_K(\boldsymbol{x}) &= 0,00 \\ F_Y(\boldsymbol{x}) &= 4,99 \end{aligned}$$

Wird wiederum die in Tabelle 4.26 vorgeschlagene Einteilung der Bereiche verwendet, so erhält dieser Haushalt für die hier folgenden Perioden keine Coupons für die Warengruppen „Waschmittel“ und „Yoghurt“. Die hohe Zugehörigkeit zur Klasse der „Yoghurt“-Käufer ist sehr gut an seinem bisherigen Kaufverhalten zu erkennen, welches für jede der verwendeten fünf Wochen einen Eintrag (Kauf) aufweist. Für den bereits oben betrachteten Zeitraum erhält der Kunde zusätzlich einen Coupon für die Warengruppe „Suppe“ in Höhe von 5%, da sich der entsprechende Entscheidungswert deutlich erhöht hat. Der Wert für die Warengruppe „Ketchup“ ist unverändert geblieben, sodass hier zur Herstellung eines Kaufinteresses wiederum ein Gutschein über einen 5%-Rabatt vergeben wird.

4.5.3 Zusammenfassung

Die Differenzierung der Kunden nach den individuellen Kaufwahrscheinlichkeiten in den kommenden Perioden und der darauf basierenden Segmentierung des Kundenstammes dient der Erhöhung der Kundenzufriedenheit. Dem Kunden soll durch die differenzierte Ausgestaltung der Mailings der Eindruck vermittelt werden, dass das klassifizierende Unternehmen sich auf ihn als Individuum konzentriert und seinen Bedürfnissen entgegen kommt. Es wird von der Annahme ausgegangen, dass die Wahrscheinlichkeit, auf Mailings zu reagieren, also in diesem Fall einen Kauf zu tätigen, höher ist, wenn das Angebot individuell zugeschnitten ist, als wenn es sich um ein nicht erkennbar differenziertes Massenmailing handelt. Eine Einschränkung der Rabattgewährung auf die Gewinn versprechenden Kunden hat ebenfalls eine Kostensenkung derartiger Aussendungen zur Folge. Dabei ist nicht

die Individualisierung im Sinne des One-to-One-Marketings, sondern die Auswahl der Zielgruppe mittels SVM als neues Element innerhalb des Direktmarketings zu betrachten. Die Anwendung hat gezeigt, dass sich eine Vielzahl unterschiedlicher Zuweisungsmuster ergeben und sich nur wenige Couponbündel gleichen.
Neben der reinen Zuordnung der Kunden zu den einzelnen Segmenten auf Basis der bisherigen Kaufhistorie kann eine Überprüfung der rechtmäßigen Versendung der Coupons mit Hilfe der traditionellen ABC-Analyse (vgl. *Krafft, Albers* (2000)) durchgeführt werden. Es könnte zusätzlich überprüft und differenziert werden, ob ein Kunde tatsächlich in den Genuss eines höherwertigen Coupons gelangen sollte, oder nicht. Da beide Verfahren unterschiedliche Ziele verfolgen[44], kann das eingesetzte Konzept der SVM durch herkömmliche Verfahren zur Kundenklassifikation sinnvoll erweitert werden.
Wendet man analog zu Abschnitt 4.4.2 die Verfahren zur Ermittlung des Einflusses der eingehenden Merkmale an, so zeigt dies hier, dass etwa bei der Warengruppe „Ketchup“ die weiter in der Vergangenheit liegenden Perioden einen höheren Einfluss auf die Lage der Trennebene und die Klassifikation haben, als die jüngeren Daten. Demnach wird die These gestützt, dass die Muster bei der Untersuchung des Kaufverhaltens innerhalb der Warengruppe „Ketchup“ von einer längeren Kaufhistorie (in diesem Fall 15 Wochen) abhängt.

Beim Einsatz von SVM im Rahmen der Ausgestaltung von One-to-One-Marketingstrategien können Probleme unterschiedlicher Art auftreten, auf die im Folgenden kurz eingegangen werden soll.

Zunächst sei auf die zugrunde liegenden Daten und deren Interpretation verwiesen. Aus den hier verwendeten Merkmalsvektoren geht neben dem Indikator für Kauf nicht die Menge der gekauften Artikel hervor. Dies impliziert insbesondere, dass bei länger andauerndem Nichtkauf nicht zwangsläufig auf Bedarf geschlossen werden muss, da durchaus Vorratskäufe vorliegen können, die durch die Struktur der Daten nicht erkannt werden können. Zur Berücksichtigung des Problems müssten die entsprechenden Beobachtungen bei Auftreten eines Kaufs von überdurchschnittlichem Umfang aus dem Datensatz entfernt werden, um die Struktur der Daten beibehalten zu können. Eine weitere Alternative besteht in der Verwendung von metrischen Daten, die den Umfang der gekauften Artikel aus der entsprechenden Warengruppe dokumentieren. Da sich die Muster bei metrischen Daten allerdings deutlich voneinander unterscheiden können, ist nicht zu erwarten, dass ein derartiges Vorgehen zu guten Klassifikationsergebnissen führt.
Die Gutscheine werden versendet, wenn in den kommenden Wochen kein Kauf getätigt wird. Eine alternative Marketingstrategie liegt darin, gerade den treuen Käufern einen Gutschein als „Belohnung“ zukommen zu lassen oder gerade diejenigen Kunden zu aktivieren, für die ein Kauf in den folgenden Perioden zu erwarten ist. Dies könnte ein eventuelles Abwandern zur Konkurrenz aufgrund von entsprechenden Werbemaßnahmen der Wettbewerber verhindern. Somit ergibt sich

[44]Bei der Klassifikation mittels SVM steht nicht die Bewertung von Kunden wie in Abschnitt 4.4.2 im Vordergrund, sondern lediglich die Einordnung bezüglich des Kaufverhaltensmusters.

die inhaltliche Bedeutung der verschiedenen erhaltenen Bereiche nicht automatisch, sondern ist immer ein Teil der zu verfolgenden Marketingstrategie.
Ein weiterer, nicht zu vernachlässigender Punkt ist die Auswahl der zu bewerbenden Produkte. Die vier hier ausgewählten Warengruppen sind in unterschiedlichem Maße für diese Analyse geeignet. Häufig gekaufte Warengruppen, die bei nahezu jedem Einkauf vertreten sein können, wie hier etwa die Warengruppe „Yoghurt“, sollten evtl. von derartigen Analysen ausgeschlossen werden, da hier keine starke Differenzierung zwischen den Käufern und Nichtkäufern zu erwarten ist, bzw. diese auf regelmäßige Käufer, die nicht unbedingt zu bewerben sind, und unregelmäßige Käufer beschränkt ist. Letztere bilden die eigentliche Zielgruppe derartiger Analysen. Es wurde hier dennoch die Warengruppe „Yoghurt“ hinzugezogen, da aufgrund mangelnder Datengrundlage keine Daten weiterer Warengruppen zur Verfügung standen und eine Präsentation der Vorgehensweise bei mehreren Warengruppen durchgeführt werden sollte.
Eine weitere Problemquelle kann die Inhomogenität der Zusammenstellung der Warengruppen bilden. Diese tritt bei obigem Beispiel insbesondere bei der Warengruppe „Waschmittel“ auf, bei dem neben dem 10kg Waschpulver-Eimer ebenfalls kleine Tuben mit Spezialreiniger oder Flüssigwaschmittel in kleinen Portionen auftreten können. Dies führt dazu, dass auf Basis der Kaufhistorien nicht unbedingt auf den Bedarf an Waschmittel geschlossen werden kann, da auch der Kauf ähnlicher Produkte dokumentiert wird und somit die Mustererkennung innerhalb der Daten nur eingeschränkt möglich ist. Dies ist u.a. an den eher mittelmäßigen Trefferquoten zu erkennen. Eine vorherige Säuberung der Daten, bei der die Homogenität der betrachteten Warengruppen gewährleistet wird, ist daher unbedingt empfehlenswert[45]. Bei der sehr geringen Gewinnspanne gerade bei der Warengruppe „Waschmittel“ muss hier auf Cross-Selling-Effekte gesetzt werden. Das vorrangige Ziel sollte bei der Betrachtung des Lebensmitteleinzelhandels in der Bindung des Kunden an das eigene Unternehmen liegen, zunächst unabhängig von dem bei einer Warengruppe erwarteten Gewinn.
Bei Analysen dieser Art können meist nur Daten des eigenen Unternehmens aus Kundenkarteninformationen verwendet werden, sodass das Kaufverhalten nur begrenzt dokumentiert werden kann und bei Einsatz von Klassifikationsverfahren immer Einschränkungen bei der Genauigkeit der Prognosen gemacht werden müssen. Alternative Datengewinnung wäre mittels Haushaltspanels möglich, die einerseits vollständige Daten liefern würden, aber andererseits sehr aufwändig und kostenintensiv sind. Der Einsatz von Fragebögen kommt aufgrund der Notwendigkeit des regelmäßigen Einsatzes hierbei nicht in Frage.
Im Gegensatz dazu liefert diese Art der Vorgehensweise aber auch Vorteile, die nicht unterschätzt werden sollten. So wird durch die spezielle Art der Datengrundlage eine dynamische Behandlung und Klassifikation der Kunden realisiert. Es werden individuelle Kaufzyklen und Kaufhistorien berücksichtigt, sodass die Klassenzugehörigkeiten sich verändern und nicht starr vorgegeben sind. Weiterhin

[45] In der vorliegenden Anwendung ist eine derartige Säuberung nicht durchführbar, da die Differenzierung der gekauften Produkte aufgrund fehlender Informationen über amerikanische Produkte nicht möglich ist.

können viele Warengruppen berücksichtigt werden und somit eine Vielzahl von Angeboten bzw. individuelle Couponversendungen realisiert werden. Dies wird dadurch gewährleistet, dass für jede einzelne Warengruppe ein Modell berechnet und Muster identifiziert werden. Außerdem kann der Umfang der Couponbündel variieren. Der Einsatz von SVM kann somit insbesondere zur Ergänzung existierender Werbestrategien eingesetzt werden, um ein individuelles One-to-One-Marketing umzusetzen und traditionelle Ziele beizubehalten.
Weiterhin können die Kunden klassifiziert werden und Angebote erstellt werden, ohne dass persönliche Daten dieser vorliegen müssen. In die Modellberechnungen gehen lediglich die Kaufdaten der Kunden ein. Voraussetzung dafür ist der Einsatz von Kundenkarten, um die Kunden bei jedem Kauf eindeutig identifizieren zu können. Den Einsatz dieser Karten vorausgesetzt werden somit kostenaufwändige und zeitintensive Erhebungen soziodemografischer Kundendaten und das Auftreten fehlender Werte vermieden.
Die hier verwendeten Daten basieren auf dem Einsatz dieser Kundenkarten im Lebensmitteleinzelhandel, der in der gegenwärtigen Situation noch sehr wenig verbreitet ist und starkes Entwicklungspotenzial besitzt. SVM können hierbei als ein Hilfsmittel der Entscheidungsunterstützung bei der Auswertung dieser Daten dienen. Eine derartige Klassifikation der Kunden kann in regelmäßigen Abständen wiederholt werden, wobei die ermittelten SVM je nach Warengruppe und deren Saisonzyklen neu trainiert werden müssen. Durch die Datenauswahl und die Möglichkeiten der Ergebnisinterpretation bieten SVM hier einen viel versprechenden Ansatz, um als ergänzendes Verfahren eingesetzt zu werden.

Kapitel 5

Zusammenfassung und Ausblick

Ziel dieser Arbeit war es, durch die anwendungsbezogene Betrachtung der noch relativ jungen SVM eine neue Methode zur Analyse a priori eingeteilter Gruppen innerhalb des quantitativen Marketings einzubringen. Neben der Erweiterung und Anpassung des bestehenden Instrumentariums an marketingspezifische Anforderungen sollten anhand empirischer Untersuchungen die Potenziale von SVM aufgezeigt werden. Dazu wurden zunächst in Kapitel 2 die methodischen Grundlagen erörtert. Neben den Basisansätzen zur linearen und nicht linearen Trennung zweier Klassen wurde auf unterschiedliche Arten der Multiklassifikation eingegangen. Anschließend wurden in Kapitel 3 Ergänzungen vorgestellt und diskutiert, die nützliche Eigenschaften aufweisen, welche beim Einsatz von SVM im Marketing zum Tragen kommen. Die Eignung von SVM und der vorgestellten Erweiterungen für den empirischen Einsatz wurde anhand der Anwendung auf unterschiedlich strukturierte Datensätze aus dem Bereich der Kundenklassifikation demonstriert. Die erzielten Ergebnisse wurden denen gegenübergestellt, die mit herkömmlichen, im Marketing etablierten Methoden erreicht wurden. Es hat sich gezeigt, dass die SVM in allen Anwendungsbeispielen ein leistungsstärkeres Verfahren zur Klassifikation, gemessen an den Trefferquoten, darstellt. Daneben bieten SVM vielfältige Möglichkeiten zur Einflussnahme von Seiten des Nutzers, um die gewünschte Anwendung realisieren zu können.

Die Ausrichtung der Arbeit zielte insbesondere auf die Überprüfung der anwendungsbezogenen Bereiche der SVM. Diese umfassen unterschiedliche Aspekte. Für einen erfolgreichen Einsatz eines Klassifikationsinstrumentes in der Praxis muss die Benutzerfreundlichkeit gewährleistet sein. Dies beinhaltet insbesondere die Festlegung des Modells durch die Wahl geeigneter Parameter. Die Auswertungen haben gezeigt, dass sehr unterschiedliche Kombinationen der zu wählenden Parameter zu guten Ergebnissen führen können, sodass keine generelle Empfehlung hinsichtlich guter Werte gegeben werden kann. Somit ist übereinstimmend mit verschiedenen Literaturquellen in diesem Bereich ein weiterer Forschungsbedarf bestätigt worden. Andererseits kann dies aber auch als ein Vorteil für den Benutzer gewertet werden, da die Suche nach Parametern durch mehrere gute Kombinationen im Gegensatz zu nur einer optimalen Lösung vereinfacht

bzw. verkürzt wird. Bei Festlegung des Kerns hat sich der Radialbasis-Kern als eine gute Wahl für die nicht lineare Trennung herausgestellt. Dabei hat sich die im Vergleich zu etwa dem Einsatz des polynomiellen Kerns deutlich geringere Rechenzeit bei Verwendung der LIBSVM-Software als vorteilhaft erwiesen, was einen entscheidenden Punkt bei der betriebswirtschaftlichen Anwendung bildet. Hierbei ist hinzuzufügen, dass eine zukünftige intensive Auseinandersetzung mit der Anwendung von SVM im Marketing zur Entwicklung neuer Kerne führen kann, um spezielle betriebswirtschaftliche Situationen besser abbilden und bearbeiten zu können, wie dies in anderen Bereichen bereits der Fall ist. Die Verfügbarkeit eines Kerns, der zum Beispiel zusätzlich fehlende Werte verarbeiten kann oder den Einsatz von kategorialen Daten ermöglicht, würde zu einem vermehrten Einsatz von SVM in marketingbezogenen Anwendungen beitragen können. Da diese beiden Eigenschaften bei Daten des Marketings häufig auftreten, bieten SVM diesbezüglich bislang noch kein adäquates Analysewerkzeug für Daten dieser Art. Somit bildet die Entwicklung von Kernen, die sich an die Probleme der betriebswirtschaftlichen Anwendungen anpassen, also insbesondere fehlende Werte explizit berücksichtigen, eine interessante Forschungsaufgabe für die Zukunft.
Ferner ist auf die Möglichkeit eingegangen worden, durch Gewichtungen der Klassen und Merkmale Einfluss auf die Trennung der Gruppen zu nehmen. Häufig liegen im Marketing Daten vor, die nicht balanciert sind, sodass die Zielgruppe die Minderheit innerhalb der Daten bildet. Für derartige Daten ist es schwierig, Modelle zu finden, die gute Vorhersagen für die kleinere Klasse liefern. SVM bietet durch Vergabe unterschiedlicher Kosten für fehlklassifizierte Beobachtungen die Möglichkeit, diesem Problem entgegen zu treten und somit bessere Prognosen liefern zu können als herkömmliche Methoden. Auch bei der Betrachtung von Merkmalen kann eine Gewichtung nützlich sein. In der Literatur sind derartige Modifikationen bei SVM-Anwendungen allerdings nicht verbreitet. Die empirischen Analysen haben in dieser Arbeit hingegen gezeigt, dass eine unterschiedliche Gewichtung von Merkmalen zu einer Verbesserung der Ergebnisse führen kann. Bei der Wahl des Merkmalseinflusses könnte die Einbeziehung der Ergebnisse einer Hauptkomponentenanalyse hilfreich sein. So können die dadurch erzielten Ladungen einen Hinweis auf mögliche Gewichte und deren Verhältnisse zueinander geben. Dieser Ansatz kann die Basis weiterer Forschungsarbeiten bilden, wobei die hohe Sensibilität von SVM im Bezug auf die Reaktion auf Veränderungen von Parametern zu berücksichtigen ist.
Bezüglich der Anwenderfreundlichkeit spielt die Nützlichkeit und Verständlichkeit der resultierenden Ergebnisse eine entscheidende Rolle. Es wurde insbesondere auf die Interpretation der Ergebnisse Wert gelegt, die hier über die in der Regel in der Literatur lediglich verwendeten Trefferquoten hinausgeht und damit die Basis für weitere Anwendungen schafft. Die Werte der Klassifikationsfunktionen dienten als Entscheidungsbasis zur Entwicklung differenzierter Strategien in der Behandlung des Marktes und des Kundenkreises, um den individuellen Bedürfnissen der Kunden nachkommen zu können und sie somit an das Unternehmen zu binden. Sowohl bei der Binärklassifikation als auch im Mehrklassen-Fall unter Zuhilfenahme der Fuzzy-SVM ist es möglich, aus den Entscheidungswerten die Intensität der

Zugehörigkeit eines Objektes zu einer Klasse zu bestimmen. Somit können die im Finanzbereich gängigen Bewertungen von Unternehmen für die Interpretation übernommen werden. Im Fall der nicht linearen Trennung, die oftmals bessere Generalisierungsergebnisse aufweist, wurde dieses Konzept in der vorliegenden Arbeit um die Einbeziehung von Distanzen im Eingaberaum erweitert, um so eine Interpretation im Sinne des im Marketing intendierten Klassifikationsziels zu ermöglichen. Die empirischen Analysen von Daten des One-to-One-Marketings haben gezeigt, dass sich der Einsatz von SVM auch monetär lohnen kann. Die differenzierte Betrachtung und Behandlung des Kundenstammes basierend auf den Ergebnissen der SVM lieferte in dieser Arbeit neben dem Eingehen auf die Bedürfnisse des Kunden Umsatzsteigerungen im Vergleich zu einer homogenen Betrachtung der gesamten Kunden. Diese Eigenschaft der SVM ermöglicht damit einen effizienten und effektiven Einsatz im Rahmen des Marketings, insbesondere beim Direktmarketing, wo eine derartige Zuordnungsintensität von besonderer Bedeutung für die erfolgreiche Kundenbindung sein kann.

Die Weiterführung dieses Ansatzes liegt in der Heranziehung von Klassifikationswahrscheinlichkeiten, was in dieser Arbeit nicht betrachtet wurde. Diese können auf unterschiedliche Weise aus den Entscheidungswerten der Klassifikationsfunktion generiert werden. Neben einem Überblick über bestehende Methoden bietet *Platt* (2000) einen weiteren Ansatz, um Klassifikationswahrscheinlicheiten zu berechnen. Dieser wird in *Rüping* (2004) mit einer einfacheren Version verglichen. *Rüping* kommt zu dem Schluss, dass die Version von *Platt* die besten Ergebnisse liefert, die einfachere Methode jedoch ebenfalls gute Resultate erzeugt. So ist auch hier, wie bei der Reduktion der Merkmale, der Trade-Off zwischen einem guten Ergebnis und dem dazu nötigen Aufwand zu finden. In *Wu et al.* (2004) wird eine Erweiterung dieses Ansatzes auf die Multiklassifikation vorgenommen, wobei die OAO-Trennung zugrunde gelegt wird. Die Verwendung von Wahrscheinlichkeiten kann bei der Anwendung von SVM im Marketing eine sinnvolle Erweiterung der in dieser Arbeit betrachteten Interpretation der Entscheidungswerte darstellen. Diese könnten die Aussagen über die Verlässlichkeit der Prognose konkretisieren.

Eine weitere Ausgabe des berechneten Modells bilden die Lagrange-Multiplikatoren. Diese geben den Einfluss eines Vektors auf die Lage der Ebene wieder, sodass neben der Bestimmung von Support Vektoren ebenfalls ein Vergleich der Bedeutsamkeiten einzelner Objekte vorgenommen werden kann. Eine weiterführende Interpretation, die die Höhe der Lagrange-Multiplikatoren berücksichtigt, oder eine Weiterverwendung dieser Werte ist bisher allerdings nicht möglich.

Der Flexibilität der SVM kommt hinsichtlich der Akzeptanz im praktischen Einsatz eine entscheidende Bedeutung zu. Bei SVM bieten sich unterschiedliche Möglichkeiten der Erweiterungen an, die die verschiedenen Gegebenheiten innerhalb der zu analysierenden Daten berücksichtigen. Neben dem klassischen Einsatz ist in dieser Arbeit die Ausdehnung auf das Online-Learning betrachtet worden, womit ein entscheidender Bereich, nämlich die POS-Scannerdaten, adäquat analysiert werden kann. Durch die Online-Variante könnte somit eine Empfehlung in Echtzeit realisiert werden, wobei sich die Trennung der vorgegebenen Klassen mit den hinzukommenden Daten ebenfalls verändert. Dies wäre insbesondere auch für

Anwendungen im eCommerce interessant. Der Algorithmus wurde in dieser Arbeit auf die Multiklassifikation ausgeweitet, um den häufig eintretenden Fall mehrerer Klassen betrachten zu können.
Ein ebenfalls mögliches Szenarium bildet die Zugehörigkeit zu mehreren Klassen, die hier als Multilabel-Klassifikation bei SVM vorgestellt wurde. Dabei ist in Zukunft eine intensivere Auseinandersetzung mit der Bewertung der Resultate erforderlich, um zu einem aussagekräftigen Maß zu gelangen. Hier kann die Einbeziehung von Koeffizienten zur Bewertung von Ähnlichkeiten von Objekten (vgl. *Decker, Wagner* (2002)) hilfreich sein, um die Gleichartigkeit von zugewiesenen Mengen und tatsächlich vorliegenden Klassenzugehörigkeiten an verschiedenen Eigenschaften festzumachen. Denkbar ist beispielsweise eine Einbeziehung des gleichzeitigen Nichtauftretens einer prognostizierten und wahren Klasse, um den in dieser Arbeit betrachteten Scorewert zu ergänzen.
Falls eine große Anzahl beschreibender Variablen vorliegt, ist es sinnvoll, Merkmalsreduktionsverfahren anzuwenden, um den Aufwand bei der Erhebung der Merkmale zur Prognose der Klassenzugehörigkeiten neuer Objekte zu minimieren. Wie in der Arbeit gezeigt wurde, bieten SVM dazu mehrere Möglichkeiten, die im praktischen Einsatz einen verschieden hohen Aufwand erfordern. Im empirischen Teil der Arbeit wurden diese hinsichtlich ihrer Leistungsstärke verglichen. Die bestehenden Verfahren wurden in das IRRM-Verfahren integriert, was mit NLIRM bezeichnet wurde. In den empirischen Analysen hat sich gezeigt, dass NLIRM gute Ergebnisse im Vergleich zu bestehenden Verfahren liefert. Die teilweise auch besseren Trefferquoten des neuen, modifizierten Algorithmus rechtfertigen einen Einsatz dieses Verfahrens in zukünftigen Anwendungen. Der Einsatz des Ellbogenkriteriums und des zu erwartenden Scores liefert eine einfache Möglichkeit, die Anzahl der Merkmale im Vorfeld einer Reduktion zu bestimmen, wobei dies lediglich als eine Alternative verstanden werden soll, die nicht immer die optimale Lösung liefern muss.
Bei der Prognose von Klassenzugehörigkeiten, die auf Basis reellwertiger Daten ermittelt wurden, bietet sich der Einsatz regressionsanalytischer Verfahren an Stelle der Klassifikation an, um die funktionale Abhängigkeit innerhalb der Daten zu beschreiben. In Abschnitt 4.3.1 wurde z.B. die durch den Umsatz ermittelte Klassenzugehörigkeit einer Apotheke ermittelt. Hierbei kann ebenfalls die regressionsanalytische Variante von SVM zum Einsatz kommen, die durch den Einsatz von Kernen die Schätzung nicht linearer Zusammenhänge ermöglicht. Dieser Ansatz ist neben der Klassifikation ein noch nicht weit verbreiteter Bereich von SVM. Künftige Forschungsarbeiten im Bereich des Marketings sollten sich gerade mit diesem Aspekt befassen, da SVM durch ihre Flexibilität die bisherige Analyse mittels Regression bereichern können. Im Gegensatz zur traditionellen Regressionsanalyse können somit insbesondere hochgradig nicht lineare Abhängigkeiten metrisch skalierter Merkmale identifiziert werden und somit die Prognosewerte bei Vorliegen entsprechender Daten verbessert werden.
In Tabelle 5.1 werden die in dieser Arbeit behandelten Aspekte von SVM sowie deren Vor- und Nachteile dargestellt, um die Eigenschaften des Verfahrens kompakt zusammenzufassen. Ein großer Vorteil von SVM liegt darin, dass bei SVM der

Bereich	Vor- und Nachteile
Gewichtung	+ Hervorhebung einzelner Elemente + Einflussnahme auf die Trennung – zu hohe Sensibilität der SVM
Parameter-bestimmung	+ große Auswahl an Kernen + mehrere Parameterkombinationen führen zu guten Ergebnissen – aufwändiges Gridsearch notwendig
Multiklassifikation	+ einfache Vorgehensweise + viele Methoden zur Auswahl
Multilabel-Klassifikation	+ einfache Vorgehensweise – schwierige Beurteilung der Prognosegüte
Online-Lernen	+ gut durchführbar – keine Erfahrungswerte aus Veröffentlichungen
Prognosegüte	+ hohe Genauigkeit + bedingt steuerbar durch Gewichtungen
Merkmalsauswahl	+ mehrere Algorithmen zur Auswahl + gute, bisherige Ergebnisse – teilweise sehr aufwändige Verfahren
Entscheidungswerte	+ zusätzliche wichtige Informationen + gerade bei Multiklassifikation sinnvoll + gut im Marketing einsetzbar – Festlegung von Bereichsgrenzen erfordert Analyseerfahrung
Ergebnis-visualisierung	+ Entscheidungswerte bilden die Basis + durch Parallelkoordinaten durchführbar – Trennung im hochdimensionalen Raum nicht visualisierbar

Tabelle 5.1: Überblick über die die SVM aus Marketingsicht auszeichnenden Eigenschaften.

Anwender in gewissem Maße Einfluss auf die Ausrichtung des Modells nehmen kann (z.B. durch Gewichtung der Merkmale oder Klassen), was mit alternativen Verfahren nur schwer oder gar nicht realisierbar ist.
Wie bereits in der Einleitung erwähnt ist als Ergänzung dieser Arbeit der Einsatz von kontrollierten Experimenten denkbar. Durch die systematischen Veränderungen von simulierten Daten könnte noch die Wirkung von strukturellen Variationen auf die Ergebnisse von SVM untersucht werden.

Insgesamt hat sich gezeigt, dass SVM eine leistungsstarke Alternative zu herkömmlichen Methoden zur Klassifikation im Marketing darstellen. Dabei ist allerdings bisher ein großer Erfahrungsschatz bei der Einstellung der Parameter notwendig, um zu guten Ergebnissen bei der nicht linearen Trennung zu gelangen. Ist eine lineare Abhängigkeit zu vermuten und eine geringfügig schlechtere Trefferquote akzeptabel, so sollten herkömmliche Verfahren herangezogen werden. Ist jedoch eine verbesserte Trefferquote und damit eine Optimierung der aus der Klassifikation resultierenden Handlungsanweisungen wünschenswert, so lassen SVM insbesondere unter Einsatz der nicht linearen Trennung gute Ergebnisse erwarten. Also sollten SVM insbesondere immer dann eingesetzt werden, wenn bei der Anwendung eine Maximierung der Trefferquote nötig ist. Die SVM auszeichnende Flexibilität wird durch die Einführung von Kernen erreicht. Durch die Möglichkeit, eine Vielzahl von Kernen einzusetzen, umfassen SVM, wie *Joachims* (1999) ausführt, „eine weite Klasse von neuronalen Netzen und auch viele andere Klassifikatoren“. Dadurch erhalten SVM einen großen Vorteil gegenüber anderen Klassifikationsverfahren, die auf eine Methode beschränkt sind. Ein zusätzlicher Nutzen von SVM liegt in der Rückführung des Optimierungsproblems auf die entscheidenden Beobachtungen, die Support Vektoren. Neben der Reduzierung der Komplexität des Problems kann dies zur Bildung eines Gütemaßes der Trennung mittels SVM bei zukünftigen Forschungsarbeiten genutzt werden. Je weniger Support Vektoren bei einem Modell vorliegen, desto eher ist von einer guten Trennung der Trainingsdaten (insbesondere bei einem nicht linearen Kern) auszugehen. In Verbindung mit den Entscheidungswerten könnte diese Tatsache zur Entwicklung neuer Gütemaße zur Bewertung der Prognosegüte eines Modells beitragen. Ein weiterer Vorteil liegt darin, dass SVM im Gegensatz zu z.B. neuronalen Netzen, die von zufälligen Startkonfigurationen abhängig sind, bei gleicher Parameterwahl bei mehreren Optimierungen zum gleichen Ergebnis führen. Diese Stabilität der Ergebnisse ist darauf zurückzuführen, dass keine zufälligen Werte in die Optimierung eingehen und es keine lokalen Optima gibt. Weiterhin sind SVM unabhängig von der Verteilung der in die Optimierung eingehenden Daten. Es muss lediglich die Repräsentativität der in das Training eingehenden Daten für aussagekräftige Prognosen sichergestellt sein. Die Klassifikation beruht häufig auf Vergangenheitsdaten. Wenn angenommen werden muss, dass sich die Daten im Laufe der Anwendungsperiode ändern, so ist das wiederholte Training der SVM nötig, um die gute Generalisierungsfähigkeit des Modells zu gewährleisten. Eng damit verbunden ist die Erfolgskontrolle, um zu bestimmen, ob die eingesetzten SVM und die darauf aufbauende Einteilung der Bereiche und resultierende Marketingstrategien erfolgreich waren. Ist dies nicht

gegeben, so ist die Optimierung erneut durchzuführen, oder die resultierenden Implikationen zu überdenken.

Die Auswahl eines Klassifikationsinstruments beruht nicht allein auf diskriminatorischen Eigenschaften, sondern z.B. auch auf Kosten der Implementation, Einfachheit der Benutzung oder Interpretation der Ergebnisse. Diese Arbeit hat gezeigt, dass SVM einerseits wegen der zum Teil schwierigen Einstellungen noch nicht immer den Ansprüchen der Anwender an ein Klassifikationsinstrument im Marketing genügen. Andererseits lassen SVM durch das viel versprechende Potenzial sehr gute Ergebnisse bei Klassifikationsproblemen im Marketing erwarten. An dieser Stelle sei angemerkt, dass der bisher zurückhaltende Einsatz von SVM im Marketing auch auf die fehlende Verfügbarkeit von SVM in der gängigen Standard-Software in der Marketingforschung (wie SAS® oder SPSS®) zurückzuführen ist. Die Anwendung von SVM wird allerdings (wie beispielsweise bei Clementine®) durch Zugriff auf die Oracle Data Mining®-Suite als einem optionalen Element der Oracle 10g® Datenbank ermöglicht. SVM bilden hier einen bereits integrierten Bestandteil des Softwarepakets (*Mann, Monien* (2004)).
Die gesamtheitliche Betrachtung der SVM als Klassifikationswerkzeug in dieser Arbeit liefert eine über die in der Literatur vorliegenden Ansätze hinausgehende Vorstellung des Einsatzes von SVM im Marketing und gibt einen ausführlichen Überblick über die Methodik. Die vorgenommenen Erweiterung von Algorithmen und die Interpretation der Ausgabe bilden neue Ansätze, die im empirischen Teil erfolgreich auf ihre Eignung getestet wurden. Die neu aufkommenden Publikationen in marketingspezifischen Zeitschriften zur Anwendung von SVM im Marketing zeigen die Aktualität der in dieser Arbeit behandelten Thematik. Dies belegt auch die mehrfach erwähnte, zunehmende Relevanz von SVM in diesem Bereich. Es lässt sich festhalten, dass in dieser Arbeit viele Aspekte von SVM aufgedeckt wurden, die mit anderen Verfahren häufig nicht oder nur eingeschränkt umgesetzt werden können, sodass SVM dadurch zu einem viel versprechenden Instrumentarium innerhalb des Marketings werden und die bestehenden Alternativen sinnvoll ergänzen. Die Autorin hofft, mit dieser Arbeit einen Beitrag zur Etablierung von SVM als Analyseinstrument im Marketing geleistet und weiteren Forschungsbedarf aufgezeigt zu haben.

Literaturverzeichnis

Aaker, D.; V. Kumar; G. Day (2004): *Marketing Research*, 8. Aufl., New York, Wiley & Sons.

Abe, S. (2003): Analysis of Multiclass Support Vector Machines, *Proceedings of the International Conference on Computational Intelligence for Modelling, Control, and Automation* (CIMCA 2003), Wien, 385–396.

Abe, S.; T. Inoue (2002): Fuzzy Support Vector Machines for Multiclass Problems, *Proceedings of European Symposium on Artificial Neural Networks*, 113–118.

Albers, S. (2002): Besuchsplanung, in: Albers, S. (Hrsg.): *Verkaufsaußendienst: Planung, Steuerung, Kontrolle*, Düsseldorf, Symposium, 173–195.

Ankerst, M. (2000): *Visual Data Mining*, dissertation.de, Berlin, Verlag im Internet.

Ayat, N.E.; M. Cheriet; C. Y. Suen (2002): KMOD - A Two-parameter SVM Kernel for Pattern Recognition, *Proceedings of the International Conference on Pattern Recognition*, Canada, IEEE Computer Society.

Backhaus, K.; B. Erichson; W. Plinke; R. Weiber (2003): *Multivariate Analysemethoden*, 10. Auflage, Berlin, Springer.

Bahlmann, C.; B. Haasdonk; H. Burkhard (2002): Online Handwriting Recognition with Support Vector Machines - A kernel approach, *Proceedings of the 8^{th} International Workshop on Frontiers in Handwriting Recognition* (IWFHR), 49–54.

Baier, D.; M. Queitsch; F. Wartenberg (2004): eDetailing - Neue Anforderungen und innovative Konzepte, in: Decker, R.; F. Wartenberg (Hrsg.): *Vertriebs- und Kundenmanagement - Moderne Marketingmethoden im Einsatz*, Lohmar, Eul, 117–133.

Bauer, H.; G. Görtz; L. Dünnhaupt (2002): Der Einzug von Coupons in Deutschland, *Arbeitspapier M 70*, Institut für Marktorientierte Unternehmensführung, Universität Mannheim.

Bazzani, A.; A. Bevilacqua; D. Bollini; R. Brancaccio; R. Campanini; N. Lanconelli; A. Riccardi; D. Romani (2001): An SVM classifier to seperate false signals from

microcalcifications in digital mammograms, *Physics in Medicine and Biology*, Vol. 46, Nr. 6, 1651–1663.

Bennett, K. P.; C. Campbell (2000): Support Vector Machines: Hype or Hallelujah?, *SIGKDD Explorations*, Vol. 2, No. 2, 1–13.

Bennett, K. P.; D. Wu; L. Auslender (1998): On Support Vector Decision Trees for Database Marketing, *R.P.I. Math Report No. 98–100*, Rensselaer Polytechnic Institute, Troy.

Berekoven, L.; Eckert, W.; Ellenrieder, P. (2004): *Marktforschung*, 10. Aufl., Wiesbaden, Gabler.

Berry, M.; G. Linoff (2000): *Mastering Data Mining - The Art and Science of Customer Relationship Management*, New York, Wiley.

Bodapati, A.; S. Gupta (2004): A Direct Approach to Predicting Discretized Response in Target Marketing, *Journal of Marketing Research*, Vol. 41, Nr. 1, 73–85.

Bomhardt, C. (2004): NewsRec, a SVM-driven Personal Recommendation System for News Websites, IEEE/WIC/ACM International Conference on Web Intelligence, Bejing, 545–548.

Boser, B. E.; I. M. Guyon; V. Vapnik (1992): A Training Algorithm for Optimal Margin Classifiers, in: Haussler, D. (Hrsg.): *Proceedings of the 5^{th} Annual ACM Workshop on Computational Learning Theory*, Pittsburgh, ACM Press, 144–152.

Boutell, M.; X. Shen; J. Luo; C. Brown (2003): Multi-label Semantic Scene Classification, *Technical Report 813*, Department of Computer Science, University of Rochester, Rochester.

Bruhn, M. (2002): *Marketing, Grundlagen für Studium und Praxis*, 6. Aufl., Wiesbaden, Gabler.

Burges, C. (1998): A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery*, Vol. 2, Nr. 2, 121–167.

Chang, C.-C.; C.-J. Lin (2001): LIBSVM : a library for Support Vector Machines, http://www.csie.ntu.edu.tw/∼cjlin/libsvm, Zugriff: 3.8.2001

Chapelle, O.; V. Vapnik; O. Bousquet; S. Murkherjee (2002): Choosing Multiple Parameters for Support Vector Machines, *Machine Learning* Vol. 46, Nr. 1, 131–159.

Cheung, K.-W.; J. T. Kwok; M. H. Law; K.-C. Tsui (2003): Mining Customer Product Ratings for Personalized Marketing, *Decision Support Systems – Special Issue on Web Data Mining*, Vol. 35, Nr. 2, 231–243.

Cortes, C.; V. Vapnik (1995): Support Vector Networks, *Machine Learning*, Vol. 20, Nr. 3, 273–297.

Crammer, K.; Y. Singer (2001): On the algorithmic implementation of multiclass Kernel-Based Vector Machines, *Journal of Machine Learning Research*, Vol. 2, No. 2, 265–292.

Cristianini, N.; J. Shawe-Taylor (2000): *An introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge, Cambridge University Press.

Crone, S. F.; S. Lessmann; R. Stahlbock (2004): Empirical Comparison and Evaluation of Classifier Performance for Data Mining in Customer Relationship Management, *Proceedings of the IEEE International Joint Conference on Neural Networks*, New York, Vol. 1, 442–448.

Cui, D.; D. Curry (2005): Prediction in Marketing using the Support Vector Machine, erscheint in *Marketing Science*.

Curry, D.; D. Cui (2005): Applications of Support Vector Machines in Marketing Engineering: An Exposition, erscheint in *Decision Science*.

Decker, R. (2005): Market Basket Analysis by Means of a Growing Neural Network, *The International Review of Retail, Distribution, and Consumer Research*, Vol. 15, Nr. 2, 151–169.

Decker, R. (2005a): A Growing Self-Organizing Neural Network for Lifestyle Segmentation, erscheint in: *Journal of Data Science*.

Decker, R.; K. Monien (2003): Support-Vektor-Maschinen als Analyseinstrument im Marketing am Beispiel der Neukundenklassifikation, *Der Markt*, Vol. 42, Nr. 1, 3–13.

Decker, R.; T. Temme (2000): Diskriminanzanalyse, in: Herrmann, A.; C. Homburg (Hrsg.): *Marktforschung: Methoden, Anwendungen, Praxisbeispiele*, 2. Aufl., 295–335.

Decker, R.; R. Wagner (2002): *Marketingforschung, Methoden und Modelle zur Bestimmung des Käuferverhaltens*, München, Moderne Industrie.

Dialego (2004): Selbstmedikation, Dialego Deutschland NetJet. April 2004, www.dialego.de, Zugriff: 15.3.2005.

Dietterich, T.; G. Bakiri (1995): Solving Multiclass Learning Problems via Error-Correcting Output Codes, *Journal of Artificial Intelligence Research*, Vol. 2, 263–286.

Diller, H. (1995): Kundenmanagement, in: Tietz, B.; R. Köhler; J. Zentes (Hrsg.): *Handwörterbuch des Marketing*, 2. Aufl., Stuttgart, Schäffer-Poeschel, 1363–1376.

Elisseeff, A.; J. Weston (2002): A kernel method for multi-labelled classification, in: Dietterich, T.; S. Becker; Z. Ghahramani (Hrsg.): *Advances in Neural Information Processing Systems*, 14, Cambridge, MIT Press, 681–687.

Evgeniou, T.; C. Boussios; G. Zacharia (2005): Generalized Robust Conjoint estimation, *Marketing Science*, Vol. 24, Nr. 3, 415–429.

Fayyad, U.; G. Piatetsky-Shapiro; R. Uthurusamy (Hrsg.) (1996): *Advances in Knowledge Discovery and Data Mining*, Menlo Park, AAAI Press.

Fisher, R.A. (1936): The Use of Multiple Measurements in Taxonomic Problems, *Annals of Eugenics*, Vol. 7, 179–188.

Franz, T. (2003): Karstadt Warenhaus AG: Couponing - Ein neues Instrument für das Marketing, in: Hartmann, W.; R.T. Kreutzer; H. Kuhfuß (Hrsg): *Handbuch Couponing*, Wiesbaden, Gabler, 539–561.

Fröhlich, H.; A. Zell (2004): Feature Subset Selection for Support Vector Machines by Incremental Regularized Risk Minimization, *IEEE International Joint Conference on Neural Networks*, Vol. 3, 2041–2046.

Furey, T.; N. Cristianini; N. Duffy; D. Bednarski; M. Schummer; D. Haussler (2000): Support Vector Machine Classification and Validation of Cancer Tissue Samples using Microarray Expression Data, *Bioinformatics*, Vol. 16, Nr. 10, 906–914.

Gehrig, W. (1992): *Pharma-Marketing*, 2. Aufl., Zürich, Moderne Industrie.

Goodhardt, G.J.; A.S.C Ehrenberg (1967): Conditional Trend Analysis: A Breakdown by Initial Purchasing Level, *Journal of Marketing Research*, Vol. 4, 155–161.

Gündling, C. (1997): *Maximale Kundenorientierung*, 2. Aufl., Stuttgart, Schäffer-Poeschel.

Guyon, I.; A. Elisseeff (2003): An Introduction to Variable and Feature Selection, *Journal of Machine Learning Research*, Vol. 3, 1157–1182.

Guyon, I.; J. Weston; S. Barnhill; V. Vapnik (2002): Gene Selection for Cancer Classification using Support Vector Machines, *Machine Learning*, 46, 389–422.

Hanley, J.; B. McNeil (1982): The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve, *Radiology*, Vol. 143, 29–36.

Hennig-Thurau, T.; U. Hansen (2001): Kundenzufriedenheit, in: H. Diller (Hrsg.): *Vahlens Großes Marketing Lexikon*, 2. Aufl., München, Beck/Vahlen, 878–881.

Hermes, L.; J. Buhmann (2000): Feature Selection for Support Vector Machines, *Proceedings of the International Conference on Pattern Recognition* (ICPR 2000), Vol. 2, 716–719.

Heutschi, R.; R. Alt (2003): eDetailing - elektronisches Marketing in der Pharmaindustrie, Bericht BE HSG/CC BN2/6, Institut für Wirtschaftsinformatik, Universität St. Gallen, St. Gallen.

Heutschi, R.; C. Legner; A. Schiesser; V. Barak; H. Österle (2003): Potential benefits and challenges of e-detailing in Europe, *International Journal of Medical Marketing*, Vol. 3, Nr. 4, 263–273.

Hippner, H.; K. Wilde (2001): Der Prozess des Data Mining im Marketing, in: Hippner, H.; U. Küsters; M. Meyer; K. Wilde (Hrsg.): *Handbuch Data Mining im Marketing – Knowledge Discovery in Marketing Databases*, Braunschweig/Wiesbaden, Vieweg/Gabler, 2–91.

Holland, H. (2004): *Direktmarketing*, 2. Aufl., München, Vahlen.

Homburg, H.; H. Krohmer (2003): *Marketingmanagement*, Wiesbaden, Gabler.

Homburg, C.; F. Sieben (2005): Customer Relationship Management (CRM) - Strategische Ausrichtung statt IT-getriebenem Aktivismus, in: Bruhn, M.; Homburg, C. (Hrsg.): *Handbuch Kundenbindungsmanagement*, 5. Aufl., Wiesbaden, Gabler, 435–461.

Hosmer, D.; S. Lemeshow (2000): *Applied Logistic Regression*, 2. Aufl., New York, Wiley.

Hsu, C.-W.; C.-C. Chang; C.-J. Lin (2003): A Practical Guide for Support Vector Classification, http://www.csie.ntu.edu.tw/∼cjlin/papers/guide/guide.pdf, Zugriff: 6.6.2005.

Hsu, C.-W.; C.-J. Lin (2002): A Comparison of Methods for Multi-class Support Vector Machines, *IEEE Transactions on Neural Networks*, Vol. 13, No. 2, 415–425.

Huang, H.-P.; Y.-H. Liu (2002): Fuzzy Support Vector Machines for Pattern Recognition and Data Mining, *International Journal of Fuzzy Systems*, Vol. 4, No. 3, 826–835.

Huang, H.-S.; K.-L. Lin; J. Hsu; C.-N. Hsu (2005): Item-Triggered Recommendation for Identifying Potential Customers of Cold Sellers in Supermarkets, *Beyond Personalization 2005*, Workshop on the Next Stage of Recommender Systems Research, International Conference on Intelligent User Interfaces (IUI 2005), San Diego, USA.

Huldi, C. (2003): Couponing als Bestandteil von Database-Marketing-Strategien, in: Hartmann, W.; R.T. Kreutzer; H. Kuhfuß (Hrsg): *Handbuch Couponing*, Wiesbaden, Gabler, 261–277.

Inoue, T.; S. Abe (2001): Fuzzy Support Vector Machines for Pattern Classification, *Proceedings of International Conference on Neural Networks*, Vol. 2, 1449–1454.

Inselberg, A. (1985): The Plane with parallel coordinates, *The Visual Computer*, Vol. 1, 69–91.

Joachims, T. (1999): Support Vector Machines, *Künstliche Intelligenz*, Heft 4, 54–55.

Joachims, T. (2002): *Learning to classify text using support vector machines*, Boston, Kluwer.

Keerthi, S. (2002): Efficient Tuning of SVM Hyperparameters Using Radius/Margin Bound and Iterative Algorithms, *IEEE Transactions on Neural Networks*, Vol. 13, Nr. 5, 1225–1229.

Kikuchi, T.; S. Abe (2003): Error Correcting Output Codes vs. Fuzzy Support Vector Machines, *Proceedings of Artificial Neural Networks in Pattern Recognition*, ANNPR, Florenz, Italien, 192–196.

Klemz, B.; P. Dunne (2000): Exploratory Analysis using Parallel Coordinate Systems: Data Visualization in N Dimensions, *Marketing Letters*, Vol. 11, Nr. 4, 323–333.

Kotler, P.; F. Bliemel (2001): *Marketing Management*, 10. Aufl., Schäffer-Poeschel, Stuttgart.

Krafft, M.; S. Albers (2000): Ansätze zur Segmentierung von Kunden – Wie geeignet sind herkömmliche Konzepte?, *Zeitschrift für betriebswirtschaftliche Forschung*, Vol. 52, Nr. 6, 515–536.

Krafft, M.; O. Götz (2004): Der Zusammenhang zwischen Kundennähe, Kundenzufriedenheit und Kundenbindung sowie deren Erfolgswirkungen, in: Hippner, H.; Wilde, K. D. (Hrsg.): *Grundlagen des CRM - Konzepte und Gestaltung*, Gabler, Wiesbaden, 265–296.

Kreßel, U. (1999): Pairwise Classification and Support Vector Machines, in: Schölkopf, B.; C. Burges; A. Smola (Hrsg.): *Advances in Kernel Methods: Support Vector Learning*, Cambridge, MIT Press, 255–268.

Lau, K. W.; Q. H. Wu (2003): Online Training of Support Vector Classifier, *Pattern Recognition*, Vol. 36, Nr. 8, 1913–1920.

Lawrence, R.D.; G.S Almasi; V. Kotlyar; M.S. Viveros; S. Duri (2001): Personalization of Supermarket Product Recommendations, *Data Mining and Knowledge Discovery*, Vol. 11, Nr. 5, 11–32.

Lim, T.-S.; W.-Y. Loh; Y.-S. Shih (2000): A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms, *Machine Learning*, Vol. 40, Nr. 3, 203–228.

Lin, C.-F.; S.-D. Wang (2002): Fuzzy Support Vector Machines, *IEEE Transactions on Neural Networks*, Vol. 13, Nr. 2, 464–471.

Lin, Y.; Y. Lee; G. Wahba (2002): Support Vector Machines for Classification in Nonstandard Situations, *Machine Learning*, Vol. 46, Nr. 1-3, 191–202.

Link, J. (2001): Direktmarketing, in: Diller, H.: *Vahlens großes Marketing Lexikon*, 2. Aufl., München, Beck/Vahlen, 308–310.

Link, J.; V.G. Hildebrand (1997): Ausgewählte Konzepte der Kundenbewertung im Rahmen des Database Marketing, in: Link, J.; D. Brändli; C. Schleuning; R. Kehl (Hrsg.): *Handbuch Database Marketing*, Ettlingen, IM Fachverlag Marketing-Forum.

Lüdtke, H.; J. Schneider (2001): Can patterns of everyday consumption indicate Lifestyles? A secondary analysis of expenditures for fast moving goods and their social contexts, in: Papastefanou, G.; P. Schmidt; H. Lüdtke; A. Börsch-Supan; U. Oltersdorf (Hrsg.): *Social Research with Consumer Panel-Data*, ZUMA-Nachrichten Spezial, Band 7, Mannheim, 26–54.

Ma, Y.; X. Ding (2002): Face Detection Based on Cost-sensitive Support Vector Machines, *Proceedings of the First International Workshop on Pattern Recognition with Support Vector Machines*, Lecture Notes in Computer Science, Vol. 2388, London, Springer, 260–267.

Mann, M.; K. Monien (2004): Klassifikation von Unternehmensdaten mit Oracle Data Mining, *Vortragsband zur 17. Deutschen Oracle-Anwenderkonferenz*, Berlin, DOAG - Deutsche Oracle Anwendergruppe, 1048–1063.

Männche, S. (2004): Allokation von Besuchszeiten und Festlegung der Außendienststärke, in: Decker, R.; F. Wartenberg (Hrsg.): *Vertriebs- und Kundenmanagement - Moderne Marketingmethoden im Einsatz*, Lohmar, Eul, 19–35.

Mercer, J. (1909): Functions of positive and negative type, and their connection with the theory of integral equations, *Philosophical Transactions of the Royal Society*, London, A 209, 415–446.

Metz, C. (1978): Basic Principles of ROC Analysis, *Seminars in Nuclear Medicine*, Vol. 8, Nr. 4, 283–298.

Meyer, D.; F. Leisch; K. Hornik (2003): The Support Vector Machine under Test, *Neurocomputing*, Vol. 55, Nr. 1-2, 169–186.

Meyer, M. (1987): *Die Beurteilung von Länderrisiken der internationalen Unternehmung*, Berlin, Duncker & Humblot.

Mika, S. (2002): *Kernel Fisher Discriminant Analysis*, University of Technology, Berlin.

Mika, S.; G. Rätsch; J. Weston; B. Schölkopf; K.-R. Müller (1999): Fisher Discriminant Analysis with Kernels, in: Hu, Y.-H.; J. Larsen; E. Wilson; S. Douglas (Hrsg.): *Neural Networks for Signal Processing IX*, IEEE, 41–48.

Monien, K.; R. Decker (2004): Strengths and Weaknesses of Support Vector Machines within Marketing Data Analysis, in: Baier, D.; K.-D. Wernecke (Hrsg.): *Innovations in Classification, Data Science, and Information Systems*, Heidelberg, Springer, 355–362.

Monien, K.; R. Decker (2004a): Segmentierung und Klassifikation von Kunden, in: Decker, R.; F. Wartenberg (Hrsg.): *Vertriebs- und Kundenmanagement - Moderne Marketingmethoden im Einsatz*, Lohmar, Eul, 155–166.

Morik, K.; P. Brockhausen; T. Joachims (1999): Combining statistical learning with a knowledge-based approach - A case study in intensive care monitoring, *Proceedings of the 16th International Conference on Machine Learning (ICML-99)*, San Francisco, Morgan Kaufman, 268–277.

Morrison, D. (1969): On the Interpretation of Discriminant Analysis, *Journal of Marketing Research*, Vol. 6, 156–163.

Müller, K.-R.; S. Mika; G. Rätsch; K. Tsuda; B. Schölkopf (2001): An Introduction to Kernel-based Learning Algorithms, *IEEE Transactions on Neural Networks*, Vol. 12, Nr. 2, 181–202.

Murthi, B.; S. Sarkar (2003): The Role of the Management Sciences in Research on Personalization, *Management Science*, Vol. 49, Nr. 10, 1344–1362.

Orsenigo, C.; C. Vercellis (2003): Multivariate classification trees based on minimum features discrete support vector machines, *IMA Journal of Management Mathematics*, Vol. 14, Nr. 3, 221–234.

Ou, Y.-Y.; C.-Y. Chen; S.-C. Hwang; Y.-J. Oyang (2003): Expediting Model Selection for Support Vector Machines Based on Data Reduction, *IEEE International Conference on Systems, Man and Cybernetics*, 786–791.

Oukhellou, L; P. Aknin; H. Stoppiglia; G. Dreyfus (1998): A New Decision Criterion for Feature selection, *European Signal Processing Conference*, EUSIPCO 1998, Vol. 1, 411–414.

Papastefanou, G.; P. Schmidt; H. Lüdtke; A. Börsch-Supan; U. Oltersdorf (Hrsg.) (2001): Social Research with Consumer Panel-Data, *ZUMA-Nachrichten Spezial*, Band 7, Mannheim.

Pavlidis, P.; J. Weston; J. Cai; W. N. Grundy (2001): Gene functional classification from heterogeneous data, *Proceedings of the 5th International Conference on Computational Molecular Biology*, New York, ACM Press, 249–255.

Peppers, D.; M. Rogers (1996): *Strategien für ein individuelles Kundenmarketing - Die 1:1 Zukunft*, Knaur, München.

Perreault,W. D.; E. J. McCarthy (2005): *Basic Marketing: A Global Approach*, 15. Aufl., Homewood, Irwin.

Petrison, L.; R. Blattberg; P. Wang (1997): Database Marketing: Past, Present, and Future, *Journal of Direct Marketing* Vol. 11, Nr. 4, 109–125.

Platt, J. (2000): Probabilities for SV Machines, in: Schölkopf, B.; C. Burges; A. Smola (Hrsg.): *Advances in Large Margin Classifiers*, Cambridge, MIT Press, 61–73.

Platt J.; N. Cristianini; J. Shawe-Taylor (2000): Large Margin DAGs for Multiclass Classification, *Advances in Neural Information Processing Systems*, Vol. 12, 547–553.

Ploss, D. (2003): Couponing in der Praxis - Bestandsaufnahme und Entwicklungsperspektiven, in: Hartmann, W.; R.T. Kreutzer; H. Kuhfuß (Hrsg.): *Handbuch Couponing*, Wiesbaden, Gabler.

Quinlan, J. R. (1993): *C4.5: Programs for Machine Learning*, San Mateo, Morgan Kaufman.

Rüping, S. (2001): Incremental Learning with Support Vector Machines, *Proceedings of the IEEE International Conference on Data Mining*, 641–642.

Rüping, S. (2004): A Simple Method for Estimating Conditional Probabilities in SVMs, in: Abecker, A.; S. Bickel; U. Brefeld; I. Drost; N. Henze; O. Herde; M. Minor; T. Scheffer; L. Stojanovic; S. Weibelzahl (Hrsg.): *LWA 2004 - Lernen - Wissensentdeckung - Adaptivität*, Berlin, Humboldt-Universität Berlin.

Saunders, C.; M. O. Stitson; J. Weston; L. Bottou; B. Schölkopf; A. Smola (1998): *Support Vector Machine - Reference Manual*, Royal Holloway Technical Report CSD-TR-98-03, Royal Holloway, University of London.

Schafer, J.B.; J. Konstan; J. Riedl (2001): E-Commerce Recommendation Applications, *Data Mining and Knowledge Discovery*, Vol. 5, Nr. 1-2, 115–153.

Schölkopf, B. (1997): *Support Vector Learning*, München, Oldenbourg.

Schölkopf, B.; C. Burges; A. Smola (1999): Introduction to Support Vector Learning, in: Schölkopf, B.; C. Burges; A. Smola (Hrsg.): *Advances in Kernel Methods: Support Vector Learning*, Cambridge, MIT Press, 1–15.

Schölkopf, B.; K.-R. Müller; A. Smola (1999b): Lernen mit Kernen, *Informatik-Forschung und Entwicklung*, Vol. 14, 154–163.

Schölkopf, B.; A. Smola (2002): *Learning with Kernels*, Cambridge, MIT Press.

Schölkopf, B.; A. Smola; K.-R. Müller (1999c): Kernel Principal Component Analysis, in: Schölkopf, B.; C. Burges; A. Smola (Hrsg.): *Advances in Kernel Methods: Support Vector Learning*, Cambridge, MIT Press, 327–352.

Standard & Poor's (2004): *Ratings im Versicherungssektor: Definitionen für Insurer Financial Strength Ratings*, www.standardandpoors.de, Zugriff: 17.3.2004.

Strauß, R.; D. Schoder (2000): Wie werden die Produkte den Kundenwünschen angepasst? - Massenhafte Individualisierung, in: S. Albers; M. Clement; K. Peters; B. Skiera (Hrsg.): *E-Commerce*, Frankfurt, F.A.Z.-Institut, 109–121.

Syed, N. A.; H. Liu; K. K. Sung (1999): Incremental Learning with Support Vector Machines, *Proceedings of the Workshop on Support Vector Machines at the International Joint Conference on Artificial Intelligence*, Stockholm, Schweden.

Temme, T. (2002): *Integrierte Entscheidungsfindung in der Marktforschung*, Lohmar, Eul.

Teow, L.-N.; K.-F. Loe (2000): Selection of Support Vector Kernel Parameters for Improved Generalization, *Proceedings of the 17^{th} International Conference on Machine Learning*, 967–974.

Trommsdorff, V.; M. Drüner (2001): Kundenorientierung, in: Diller, H. (Hrsg.): *Vahlens Großes Marketing Lexikon*, 2. Aufl., München, Beck/Vahlen, 870–871.

Valentini, G.; M. Muselli; F. Ruffino (2004): Cancer recognition with bagged ensembles of Support Vector Machines, *Neurocomputing*, Vol. 56, 461–466.

Vapnik, V. (1995): *The Nature of Statistical Learning Theory*, New York, Springer.

Vapnik, V. (1998): *Statistical Learning Theory*, New York, Wiley.

Veropoulos, K.; C. Campbell; N. Cristianini (1999): Controlling the Sensitivity of Support Vector Machines, *Proceedings of the International Joint Conference on Artificial Intelligence*, Stockholm, Schweden, 55–60.

Viaene S.; B. Baesens; T. Van Gestel; J. Suykens; D. Van den Poel; B. De Moor; J. Vanthienen; G. Dedene (2001): Knowledge Discovery in a Direct Marketing Case Using Least Squares Support Vector Machines, *International Journal of Intelligent Systems*, Vol. 16, Nr. 9, 1023–1036.

Wagner, R.; T. Temme; R. Decker (1998): Die Behandlung fehlender Werte in der angewandten Markforschung, *Jahrbuch der Absatz- und Verbrauchsforschung*, Jg. 44, Nr. 4, 395–417.

Wassel, P. (2001): Deutschland - ein Kundenkarten-Entwicklungsland, *Absatzwirtschaft Online*, 28. November 2001, Zugriff: 5.1.2005

Weston, J.; S. Mukherjee; O. Chapelle; M. Pontil; T. Poggio; V. Vapnik (2000): Feature Selection for SVMs, in: Solla, S.; T. Leen; K.-R. Müller (Hrsg.): *Advances in Neural Information Processing Systems*, 13, Cambridge, MIT Press.

Weston, J.; C. Watkins (1999): Support Vector Machines for Multi-Class Pattern Recognition, *Proceedings of the 7^{th} European Symposium on Artificial Neural Networks* (ESANN '99), 219–224.

Wilde, K. (2001): Data Warehouse, OLAP und Data Mining im Marketing - Moderne Informationstechnologien im Zusammenspiel, in Hippner, H.; Küsters, U.; Meyer, M.; Wilde, K.D. (Hrsg.): *Handbuch Data Mining im Marketing - Knowledge Discovery in Marketing Databases*, Vieweg, Wiesbaden, 33–51.

Witten, I.; E. Frank (2000): *Data Mining*, San Francisco, Morgan Kaufmann Publishers.

Wrobel, S.; K. Morik; T. Joachims (2000): Maschinelles Lernen und Data Mining, in: Görz, G.; C.-R. Rollinger; J. Schneeberger (Hrsg.): *Handbuch der Künstlichen Intelligenz*, 3. Aufl., München, Oldenbourg, 517–597.

Wu, T.-F.; C.-J. Lin; R. Wang (2004): Probability Estimates for Multi-class Classification by Pairwise Coupling, *Journal of Machine Learning Research*, Vol. 5, 975–1005.

Yang, Q. (2002): Towards Statistical Planning for Marketing Strategies, in: Ghallab, M.; J. Hertzberg; P. Traverso (Hrsg.): *Proceedings of the Artificial Intelligence Planning Conference*, AAAI Press.

Zadrozny, B.; C. Elkan (2002): Transforming Classifier Scores into Accurate Multiclass Probability Estimates, *Proceedings of the* 8^{th} *International Conference on Knowledge Discovery and Data Mining*, 694–699.

Zhang, T.; V. Iyengar (2002): Recommender Systems Using Linear Classifiers, *Journal of Machine Learning Research*, Vol. 2, 313–334.

Anhang A

A.1 Datengrundlagen

Nr.	Warengruppe	Nr.	Warengruppe
1	Fenster-/Teppich-/WC-Reiniger	31	Knäckebrot
2	Nudelsoßen, Ketchup	32	Spezialsalz für Geschirrspüler
3	Salatsoßen	33	Kaffeemittel
4	Feinwaschmittel	34	Feinkostsalate
5	Geschirrspülmittel	35	Sherry/Portwein
6	Haushaltsreiniger/Scheuermittel	36	Alkoholfreie Getränke
7	Zahnbürsten/Mundwasser		mit Kohlensäure
8	Gemüsekonserven	37	Stärken/Steifen/Gardinenpflege
9	Weichspüler/Wasserenthärter	38	Dosenmilch/Kaffeesahne
10	Zahnpflege (Zahncreme etc.)	39	Weinhaltige Getränke
11	Bohnenkaffee-Röstware	40	Herbstartikel (Lebkuchen etc.)
12	Bohnenkaffee-Extrakt	41	Essig
13	Speisesalz	42	Süßgebäck
14	Geflügel	43	Tiernahrung/Katzenstreu
15	Tiefkühlkost	44	Zwieback
16	Tee	45	Trockenfertiggerichte
17	Kakao	46	Produkte für Zimmerpflanzen
18	Spirituosen	47	Knabbergebäck
19	Universalwaschmittel	48	Instantfertiggerichte/-suppen
20	Nahrungsfette	49	Teilfertiggerichte in Dosen
21	Fertigkuchen	50	Mehl
22	Sekt	51	Backofen-/Grill-/Spezialreiniger
23	Apfel-Cidre	52	Backpulver/Hefe/Vanillinzucker
24	Bodenpflege	53	Mineralwasser
25	Badezusätze	54	Eiscreme
26	Kartoffelfertigprodukte	55	Backmischungen
27	Bier	56	Reis
28	Wermut/Aperitif	57	Teigwaren (Nudeln)
29	Alkoholfreie Getränke	58	Komplettmenues in Schalen
	ohne Kohlensäure	59	Geröstete Cocktailartikel
30	Sauer-/Krautkonserven		

Tabelle A.1: In Panel 6 vorliegende und verwendete Warengruppen

Nr.	Warengruppe, Merkmal	Beschreibung
1	Warengruppe 8	Gemüsekonserven
2	Warengruppe 11	Bohnenkaffee-Röstware
3	Warengruppe 12	Bohnenkaffe-Extrakt
4	Warengruppe 15	Tiefkühlkost
5	Warengruppe 18	Spirituosen
6	Warengruppe 20	Nahrungsfette
7	Warengruppe 22	Sekt
8	Warengruppe 25	Badezusätze
9	Warengruppe 26	Kartoffelfertigprodukte
10	Warengruppe 28	Wermut/Aperitif
11	Warengruppe 31	Knäckebrot
12	Warengruppe 33	Kaffeemittel
13	Warengruppe 34	Feinkostsalate
14	Warengruppe 35	Sherry/Portwein
15	Warengruppe 42	Süßgebäck
16	Warengruppe 44	Zwieback
17	Warengruppe 45	Trockenfertiggerichte
18	Warengruppe 47	Knabbergebäck
19	Warengruppe 48	Instantfertiggerichte/-suppen
20	Warengruppe 49	Teilfertiggerichte in Dosen
21	Warengruppe 54	Eiscreme
22	Warengruppe 55	Backmischungen
23	Warengruppe 58	Komplettmenus in Schalen
24	Warengruppe 59	Geröstete Cocktailartikel
25	Junk-TK-Anteil	Anteil an Fastfood in WG 15
26	Health-TK-Anteil	Anteil an Gemüse/Obst in WG 15
27	Diet-Fett-Anteil	Anteil Diätprodukte in WG 20
28	Diet-Salate-Anteil	Anteil an kalorienreduzierten Artikeln in WG 34
29	Lose-Salate-Anteil	Anteil an losen Salaten in WG 34
30	Fabrik-Salate-Anteil	Anteil an fabrikverpackten Salaten in WG 34
31	Diet-Back-Anteil	Anteil kalorienreduzierten Artikeln in WG 55
32	Schalen-Voll-Anteil	Anteil an Vollwertkost in WG 58
33	Diet-Schalen-Anteil	Anteil an kalorienreduzierten Artikeln in WG 58
34	Anteil-LEH-Einkauf	Anteil an Einkäufen in kleinen LEH

Tabelle A.2: Merkmale bzw. Warengruppen (WG), die bei der Multilabel-Zuweisung verwendet werden

Nummer	Einstellungsmerkmal
1	Von den Produkten, die laufend auf den Markt kommen, halte ich die meisten für überflüssig.
2	Ich liebe die Atmosphäre von kleineren Läden und Fachgeschäften.
3	Ich gehöre zu den Menschen, die Geselligkeit lieben.
4	Ich probiere gerne neue Produkte aus.
5	Viele Artikel, die ich schon kaufe, kennen andere Hausfrauen noch gar nicht.
6	In meiner Freizeit unternehme ich viel.
7	Wenn man ganz neue Produkte kauft, fällt man oft herein.
8	Man sollte sich mit seinem Geld lieber ein schönes Leben machen, als es zu sparen.
9	Ich bin immer auf der Suche nach neuen Produkten, die meinen Bedürfnissen eher entsprechen.
10	Am wohlsten fühle ich mich zu Hause.
11	Den Aussagen der Werbung stehe ich mit sehr großem Misstrauen gegenüber.
12	Neue Produkte sind oft teurer als die alten, aber nicht besser.
13	Ich habe ganz andere Interessen, als lange in der Küche zu stehen.
14	Ich möchte beim Einkaufen auf die persönliche Bedienung nicht verzichten.
15	Ich will mein Leben in vollen Zügen genießen.
16	Produkte, die neu herauskommen, habe ich oft früher als meine Bekannten.
17	In meiner Lebensführung mag ich keine Veränderungen, ich halte mich lieber an meine alten Gewohnheiten.
18	Ich koche nur Gerichte, von denen ich auch weiß, dass sie mir gelingen.
19	Beim Einkauf von Nahrungsmitteln achte ich grundsätzlich auf Qualität, auch wenn es deutlich teurer ist.
20	Markenartikel sind besser als Produkte mit unbekannten Namen.
21	Produkte, in denen Konservierungsstoffe enthalten sind, lehne ich ab.
22	Ich achte beim Essen und Trinken auf meine Figur.
23	Für Spezialitäten aus anderen Ländern kann ich mich begeistern.
24	Es ist mir egal, ob meine Lebensmittel aus Deutschland sind oder aus irgendeinem anderen Land.
25	Für das Kochen nehme ich mir viel Zeit.
26	Edle Speisen und Getränke gehören zu meinem Lebensstil.
27	Ich esse am liebsten Hausmannskost.
28	Wir ernähren uns nach dem Prinzip der Vollwertküche.
29	Ich leiste mir öfter mal Delikatessen.
30	Ohne Fertigprodukte kann ich mir das Kochen kaum noch vorstellen.
31	Bei Nahrungsmitteln achte ich mehr auf den Preis als auf Marken.

Tabelle A.3: 61 Items bzgl. des täglichen Leben und der Essgewohnheiten (Teil I)

Nummer	Einstellungsmerkmal
32	Multivitaminsäfte sind eine wichtige Ergänzung der Ernährung.
33	Der Einfluss der Ernährung auf die Gesundheit wird oft überschätzt.
34	Ich achte streng darauf, möglichst wenig Fett zu essen.
35	In meinem Haushalt achte ich sehr auf schonende, reizarme Kost.
36	Ich würde mich als sehr schlankheitsbewußt bezeichnen.
37	Beim Einkaufen achte ich sehr darauf, Lebensmittel ohne jegliche Zusatzstoffe zu wählen.
38	Ich koche gerne ausgefallene Speisen und Gerichte.
39	Nahrungsmittel aus Deutschland sind für mich qualitativ am besten.
40	Ich koche am liebsten Gerichte, die schnell gehen.
41	Wir ernähren uns vegetarisch (ohne Fleisch, ohne Wurst).
42	Ich verwöhne mich gerne mit einem guten Essen.
43	Ich esse gerne herzhafte, deftige Mahlzeiten.
44	Bei Lebensmitteln kaufe ich ausschließlich frische Produkte anstelle von z.B. Konserven oder Tiefkühlkost.
45	Zu Nahrungsmitteln ohne Markenbezeichnung habe ich kein echtes Vertrauen.
46	Ich verwende regelmäßig Vitamin- und Mineralstoffpräparate, um mich körperlich fit zu halten.
47	Es wird zuviel Wirbel um die Ernährung gemacht.
48	Ich achte darauf, was ich esse und trinke, denn ich muss auf meine Gesundheit Rücksicht nehmen.
49	Ich informiere mich darüber, welche Lebensmittel umweltbelastet sind und kaufe sie nicht mehr.
50	Es macht mir Spaß, fremdländische Spezialitäten auszuprobieren.
51	Je einfacher das Kochen geht, desto lieber ist es mir.
52	Aus Schlankheitsgründen achte ich darauf, dass ich pro Tag eine bestimmte Kalorienzahl nicht überschreite.
53	Ich halte mich beim Kochen am liebsten an altbewährte Rezepte.
54	Beim Essen und Trinken bin ich sehr anspruchsvoll.
55	Gerichte aus Getreidekörnern sind immer häufiger Bestandteile meines Speiseplans.
56	Lebensmittel bekannter Marken sind besser als Produkte mit unbekannten Namen.
57	Heutzutage schmecken mir Konserven genauso gut wie Frisches.
58	Bei der Ernährung vermeide ich alles, was der Gesundheit schadet.
59	Wenn ich die Wahl habe, kaufe ich Nahrungsmittel aus Deutschland.
60	Eine normale Kost enthält alle lebenswichtigen Nährstoffe, dazu braucht man nichts zusätzlich einzunehmen.
61	Würde man alles glauben, was heute über die Ernährung geredet wird, dürfte man gar nichts mehr essen oder trinken.

Tabelle A.4: 61 Items bzgl. des täglichen Leben und der Essgewohnheiten (Teil II)

A.2 ECOC-Codematrizen

Hier werden die in Abschnitt 4.4.2 eingesetzten Codematrizen zur Optimierung mittels ECOC (vgl. Abschnitt 2.2.5) aufgelistet. Dabei bezeichnet ein +, respektive −, die Klassen, die in der betreffenden SVM (a_i) zu den positiven respektive negativen Beobachtungen zählen. Eine 0 gibt an, dass die Klasse (Kl.) bei der entsprechenden SVM nicht berücksichtigt wird. Die Multiklassifikation wird dementsprechend auf binäre Trennungen zurückgeführt, bei der im Falle des Einsatzes einer dünn besetzten (sparse) Matrix nicht alle Klassen in jeder SVM involviert sind.

Kl.	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}	a_{15}
1	-	0	+	0	-	+	0	-	+	-	0	+	+	0	+
2	0	-	0	+	0	0	-	0	-	+	0	-	-	0	0
3	-	0	-	-	0	-	+	-	+	0	+	-	+	-	0
4	+	0	-	0	+	+	+	-	0	-	0	0	-	+	0
5	0	+	+	0	-	+	0	+	0	0	-	+	+	-	-

Kl.	a_{16}	a_{17}	a_{18}	a_{19}	a_{20}	a_{21}	a_{22}	a_{23}	a_{24}	a_{25}	a_{26}	a_{27}	a_{28}	a_{29}	a_{30}
1	+	0	0	+	0	+	+	+	0	-	+	+	0	-	+
2	-	0	+	+	+	+	0	0	+	+	-	0	0	-	+
3	-	-	0	+	0	-	0	-	0	-	-	0	-	-	0
4	+	+	+	0	-	+	+	+	-	0	0	-	+	+	-
5	-	+	-	-	+	+	-	0	0	-	-	0	0	0	-

Tabelle A.5: Darstellung der dünn besetzten Codematrix (5×30)

Klasse	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
1	-	+	+	-	+	-	-	-	-	+
2	+	-	+	-	-	-	-	+	+	-
3	-	+	-	+	-	-	+	+	-	-
4	+	-	-	-	-	+	+	-	-	+
5	-	-	-	+	+	+	-	-	+	-

Tabelle A.6: Darstellung der dicht besetzten Codematrix (5×10)

A.3 Berechnungen zum Negativ-Binomial-Modell

Die in Abschnitt 4.5.2 mittels SVM vorgenommenen Berechnungen zum One-to-One-Marketing werden durch Ergebnisse des Negativ-Binomial-Modells (NB-Modell)

ergänzt, um die Resultate mit denen traditioneller Verfahren vergleichen zu können. Dazu sind die folgenden Daten gegeben. Tabelle A.7 enthält die Anzahl der für das NB-Modell benötigten Käufe der Haushalte pro Periode für den 15 Wochenzyklus für die Warengruppe Waschmittel.

$\tilde{n}=0$	$\tilde{n}=1$	$\tilde{n}=2$	$\tilde{n}=3$	$\tilde{n}=4$	$\tilde{n}=5$	$\tilde{n}=6$	$\tilde{n}=7$
205	148	153	110	58	54	25	17

$\tilde{n}=8$	$\tilde{n}=9$	$\tilde{n}=10$	$\tilde{n}=11$	$\tilde{n}=12$	$\tilde{n}=13$	$\tilde{n}=14$	$\tilde{n}=15$
8	11	2	3	1	2	0	0

Tabelle A.7: Anzahl der Haushalte, die $\tilde{n}$-mal innerhalb von 15 Wochen gekauft haben

Pro Woche kann höchstens ein Kauf getätigt werden. Mehrfachkäufe pro Periode werden demnach nicht berücksichtigt. Daraus ergeben sich nach dem NB-Modell (vgl. z.B. *Decker, Wagner* (2002)) folgende Parameter

$$\begin{aligned}\gamma_1 &= 9,4220\\ \gamma_2 &= 1,4257.\end{aligned}$$

Mittels

$$P(\tilde{n}\text{ Käufe})_T = \begin{cases} (\frac{\gamma_1}{\gamma_1+T})^{\gamma_2} & \text{für } \tilde{n}=0 \\ P(\tilde{n}-1\text{ Käufe})_T \frac{T(\tilde{n}+\gamma_2-1)}{\tilde{n}(\gamma_1+T)} & \text{für } \tilde{n}=1,2,\ldots,T \end{cases}$$

ergeben sich die Schätzer für die Kaufklassenhäufigkeiten wie folgt (Tabelle A.8):

$\tilde{n}=0$	$\tilde{n}=1$	$\tilde{n}=2$	$\tilde{n}=3$	$\tilde{n}=4$	$\tilde{n}=5$	$\tilde{n}=6$	$\tilde{n}=7$
204,99	179,50	133,72	93,78	63,73	42,48	27,94	18,21

$\tilde{n}=8$	$\tilde{n}=9$	$\tilde{n}=10$	$\tilde{n}=11$	$\tilde{n}=12$	$\tilde{n}=13$	$\tilde{n}=14$	$\tilde{n}=15$
13,46	8,66	5,54	3,54	2,25	1,43	0,90	0,57

Tabelle A.8: Geschätzte Kaufklassenhäufigkeiten

Ein χ^2-Anpassungstest zeigt, dass die beobachteten Kaufklassenhäufigkeiten sich durch das NB-Modell beschreiben lassen.[1]

Um Trefferquoten für die beiden Klassen berechnen zu können, bietet sich

[1] Bei einem Signifikanzniveau von $\alpha = 0,01$ ergibt sich $\chi^2_{emp} = 19,77 < 27,69 = \chi^2_{krit}$.

der Einsatz von bedingten Wahrscheinlichkeiten an, mit denen die Wahrscheinlichkeit eines Kaufes in der nächsten Periode unter der Bedingung mehrerer Käufe in der Vorperiode geschätzt werden kann. Die Herleitung von derartigen bedingten Wahrscheinlichkeiten auf Basis des NB-Modells sind bereits bei *Goodhardt, Ehrenberg* (1967) zu finden. Es wird gezeigt, dass die Verteilung der Käufe in einer Folgeperiode unter der Bedingung des Vorliegens von $\tilde{r}$ Käufen in der Startperiode ebenfalls einer Negativ-Binomial-Verteilung genügt mit den Parametern $\gamma_1 + \tilde{r}$ und $\gamma_2 + \tilde{r}$. Damit ergibt sich für die bedingten Wahrscheinlichkeiten eines Nichtkaufes in der Folgeperiode folgendes[2]:

$$
\begin{aligned}
P(Y=0|\tilde{r}=0) &= \left(\tfrac{\gamma_1+\tilde{r}}{\gamma_1+\tilde{r}+T}\right)^{\gamma_2+\tilde{r}} = \left(\tfrac{9{,}422+0}{9{,}422+0+3}\right)^{1{,}4257+0} = 0{,}6743\\
P(Y=0|\tilde{r}=1) &= 0{,}5413\\
P(Y=0|\tilde{r}=2) &= 0{,}4498\\
P(Y=0|\tilde{r}=3) &= 0{,}3839\\
P(Y=0|\tilde{r}=4) &= 0{,}3347\\
P(Y=0|\tilde{r}=5) &= 0{,}2969\\
P(Y=0|\tilde{r}=6) &= 0{,}2672\\
P(Y=0|\tilde{r}=7) &= 0{,}2432\\
P(Y=0|\tilde{r}=8) &= 0{,}2237\\
P(Y=0|\tilde{r}=9) &= 0{,}2074\\
P(Y=0|\tilde{r}=10) &= 0{,}1938\\
P(Y=0|\tilde{r}=11) &= 0{,}1821\\
P(Y=0|\tilde{r}=12) &= 0{,}1721\\
P(Y=0|\tilde{r}=13) &= 0{,}1634\\
P(Y=0|\tilde{r}=14) &= 0{,}1558\\
P(Y=0|\tilde{r}=15) &= 0{,}1491.
\end{aligned}
$$

Ist diese Wahrscheinlichkeit größer als $0,5$, so wird die Kundengruppe derjenigen, die bereits $\tilde{r}$ mal gekauft haben komplett den Nichtkäufern zugeschrieben. Als Käufer in der Folgeperiode gelten die Beobachtungen, falls die berechneten Wahrscheinlichkeiten der entsprechenden Kundengruppe für einen Nichtkauf kleiner als $0,5$ sind. Eine derartige Zuordnung ergibt eine Trefferquote von 76,27% in der Klasse der Käufer und eine Trefferquote von nur 56,37% in der Klasse der Nichtkäufer. Es ergibt sich eine Gesamttrefferquote von 63,74%, was vergleichbar ist mit den Ergebnissen, die mit anderen Verfahren erzielt werden, aber dennoch um ein paar Prozentpunkte schlechter abschneidet.

A.4 Berechnungen zum Markovmodell

Als weiteres herkömmliches Modell wird im Folgenden das Markovmodell herangezogen. Die Datengrundlage wird von den gleichen Beobachtungen gebildet wie in

[2]Mit Y wird die Anzahl der Käufe in der Folgeperiode bezeichnet und $\tilde{r}$ gibt die Menge der Käufe in der vorangegangenen Periode an. Die Folgeperiode sie T Einheiten lang. Im vorliegenden Beispiel gilt $T = 3$.

4.5 auch, also durch binäre Kaufhistorien, die Aufschluss über Kauf oder Nichtkauf eines bestimmten Produktes (hier wird die Warengruppe „Waschmittel“ herangezogen) gibt. Wenn diese Daten aufgefasst werden als der Kauf einer Warengruppe und dem Nichtkauf der Warengruppe, bzw. dem Kauf anderer alternativer Artikel so bieten sich Markovketten zur Bestimmung von Kaufwahrscheinlichkeiten an (vgl. *Decker, Wagner* (2002)). Diese werden im herkömmlichen Einsatz zur Bestimmung von Markenwahlwahrscheinlichkeiten eingesetzt, dies kann aber problemlos auf den vorliegenden Fall übertragen werden.
Ziel ist die Prognose von Kaufwahrscheinlichkeiten für die im Testdatensatz aufgeführten Kunden, um somit Vergleiche zu den mittels SVM ermittelten Trefferquoten ermöglichen zu können. Dazu werden zunächst für jeden Kunden auf Basis seines bisherigen, 15 Wochen umfassenden Kaufverhaltens Übergangsmatrizen geschätzt, die die Wahrscheinlichkeiten enthalten, von einem Zustand in den anderen zu wechseln. Die Matrix

$$\boldsymbol{U}^{[\tilde{k}]} = \begin{pmatrix} u_{00}^{[\tilde{k}]} & u_{01}^{[\tilde{k}]} \\ u_{10}^{[\tilde{k}]} & u_{11}^{[\tilde{k}]} \end{pmatrix}$$

enthält somit beispielsweise mit $u_{01}^{[\tilde{k}]}$ die Wahrscheinlichkeit, dass der betreffende Kunde $\tilde{k}$ von einem Nichtkauf zu einem Kauf wechselt. Die Kaufwahrscheinlichkeit in einer Periode hängt weiterhin gemäß einer Markovkette von den Kaufwahrscheinlichkeiten in der vorangegangenen Periode ab. Somit ist die Wahrscheinlichkeit ($p_{\tau,\tilde{k}}^{[K]}$ bzw. $p_{\tau,\tilde{k}}^{[NK]}$) für einen Kauf (K) bzw. Nichtkauf (NK) von Kunde $\tilde{k}$ in einer Periode τ durch

$$(p_{\tau,\tilde{k}}^{[K]}, p_{\tau,\tilde{k}}^{[NK]}) = (p_{\tau-1,\tilde{k}}^{[K]}, p_{\tau-1,\tilde{k}}^{[NK]})\boldsymbol{U}^{[\tilde{k}]}$$

gegeben. Die Startwahrscheinlichkeiten werden ebenfalls auf Basis der 15 Wochen umfassenden Kaufhistorie gebildet. Gilt $p_{\tau,\tilde{k}}^{[K]} > 0$ für eine der drei zu prognostizierenden Perioden (vgl. S. 181) so ist es wahrscheinlich, dass Kunde $\tilde{k}$ in diesen Perioden mindestens einen Kauf tätigt. Daher wird er der Klasse der Käufer zugeordnet. Anderenfalls zählt er zu den Nichtkäufern. Liegen in den Daten Nullvektoren vor, also wurde in dem Betrachtungszeitraum für einen Kunden kein Kauf registriert, so wird er bei diesem Vorgehen direkt den Nichtkäufern zugeordnet, um Berechnungsprobleme bei der Bestimmung der Übergangsmatrizen zu umgehen.
Ein derartiges Vorgehen führt bei den vorliegenden 797 Testdaten zu einer Trefferquote von 54,45%, was von den Ergebnissen mittels SVM und des NB-Modells deutlich übertroffen wird. Daher scheint dieses Verfahren bei den vorliegenden Daten nicht adäquat zu sein.