

Inhaltsverzeichnis

Teil I	Knowledge Discovery in Databases und Data Mining	
Kapitel 1	Einleitung	3
Kapitel 2	Eine Einordnung der Begriffe KDD und Data Mining.....	7
2.1	Abgrenzung der Begriffe KDD und Data Mining.....	9
2.1.1	Knowledge Discovery in Databases (KDD).....	9
2.1.2	Data Mining.....	10
2.2	Problemfelder.....	12
2.3	Verwandte Gebiete.....	15
2.3.1	Maschinelles Lernen.....	16
2.3.2	Statistik.....	18
2.3.3	Datenbanksysteme.....	18
2.3.4	Computerbasierte Entscheidungsunterstützung.....	19
2.3.5	Visualisierung.....	21
2.4	KDD-Prozeßmodelle.....	22
2.4.1	Das Prozeßmodell von Brachman, Anand (1996).....	22
2.4.2	Das Prozeßmodell von Chapman et al. (1998).....	25
2.4.3	Das Prozeßmodell von Fayyad et al. (1996a).....	28
2.4.4	Das Prozeßmodell von John (1997).....	29
2.4.5	Das Prozeßmodell von Wirth, Reinartz (1996).....	31
2.4.6	Ein neuer Ansatz basierend auf den Hauptschritten des KDD- Prozesses	33
2.4.6.1	Die Hauptschritte des KDD-Prozesses.....	34
2.4.6.2	Das verallgemeinerte KDD-Prozeßmodell.....	36
2.5	Im Rahmen des KDD und Data Mining diskutierte Verfahren.....	40
2.5.1	Data Mining Aufgaben.....	40
2.5.2	Data Mining Verfahren.....	42

Kapitel 3	Eigene Ansätze zur Charakterisierung des Forschungsgebietes KDD und Data Mining.....	47
3.1	Fallbeispiele und Anwendungsberichte.....	48
3.1.1	Ausgewählte Anwendungsberichte	48
3.1.2	Studie über Fallbeispiele und Anwendungsberichte	50
3.1.2.1	Anwendungsbereiche	51
3.1.2.2	Eingesetzte Data Mining Verfahren.....	52
3.2	Kommerzielle Softwareprodukte	56
3.2.1	Ansätze und Arbeiten zum Vergleich von Data Mining Softwareprodukten	57
3.2.1.1	Abbott et al. (1998)	58
3.2.1.2	Collier et al. (1998)	60
3.2.1.3	Elder, Abbott (1998)	62
3.2.2	Positionierung und Segmentierung ausgewählter Data Mining Softwareprodukte	64
Teil II	Ausgewählte Verfahrensklassen und ihre Erweiterungsmöglichkeiten	
Kapitel 4	Entscheidungsbaumverfahren	79
4.1	Grundlagen und Begriffe.....	80
4.1.1	Mögliche Unterteilungen.....	83
4.1.1.1	Nominale Merkmale.....	83
4.1.1.2	Quantitative Merkmale.....	84
4.1.2	Divisive Induktion von Entscheidungsbäumen	85
4.2	Auswahlmaße	88
4.2.1	Informationsgewinn.....	89
4.2.2	Informationsgewinnverhältnis	91
4.2.3	Gini-Index	91
4.2.4	Twoing-Wert	93
4.2.5	χ^2 -Maß.....	93
4.3	Pruning-Techniken	96
4.3.1	Error-Complexity-Pruning (ECP)	97
4.3.2	Pessimistic-Error-Pruning (PEP).....	100
4.3.3	Error-Based-Pruning (EBP).....	102
4.4	Vergleich von Entscheidungsbaumverfahren.....	104
4.4.1	ID3	105

4.4.2	C4.5	105
4.4.3	CART	106
4.4.4	ChAID	106
4.5	Entscheidungsbaum-Softwareprodukte.....	107
Kapitel 5 Assoziationsregeln		109
5.1	Grundlagen und Begriffe.....	110
5.2	Algorithmen zur Generierung von Assoziationsregeln.....	112
5.2.1	Der AIS-Algorithmus	113
5.2.2	Der SETM-Algorithmus.....	115
5.2.3	Der Apriori-Algorithmus.....	116
5.2.4	Der PARTITION-Algorithmus.....	121
5.3	Assoziationen in hierarchischen Datenstrukturen.....	124
5.3.1	Der Ansatz von Srikant, Agrawal (1995) für hierarchische Datenstrukturen	125
5.4	Behandlung quantitativer Merkmale.....	129
5.5	Weitere Bewertungsmaße für Assoziationsregeln	133
5.5.1	Interest	135
5.5.2	Conviction.....	137
5.5.3	Factor.....	137
Kapitel 6 Ein neuer Ansatz zur Kombination von Entscheidungs- baumverfahren mit Assoziationsregelalgorithmen		139
6.1	Multivariate Entscheidungsbaumverfahren.....	140
6.2	Ein neuer Ansatz eines multivariaten Entscheidungsbaumverfahrens mit Hilfe eines modifizierten Assoziationsregelalgorithmus	145
6.2.1	Allgemeine Vorgehensweise	146
6.2.2	Bestimmen der häufigen Itemmengen.....	149
6.2.3	Generieren von Regeln.....	151
6.2.4	Berechnen der Werte des Auswahlmaßes mit Hilfe von Support und Konfidenz	151
6.2.5	Behandlung quantitativer Merkmale	155
6.2.5.1	Auswahl der univariaten Unterteilungen.....	155
6.2.5.2	Wahl der optimalen Support-Schranke	159
6.2.5.3	Zusätzliche Parameter zur Einschränkung des Suchraumes ..	161
6.3	Anwendung des neuen Ansatzes und Vergleich mit existierenden Entscheidungsbaumverfahren	163
6.3.1	Verfahrensvergleich	163
6.3.2	Simulationsstudie zur Parameterwahl	166

Teil III Analyse des Markenwechselverhaltens als ausgewählte
Data Mining Anwendung

Kapitel 7 Analyse des Markenwechselverhaltens von Konsumenten.. 171

- 7.1 Einführung und Problemspezifikation 171
- 7.2 Eigene Ansätze unter Verwendung von Data Mining Verfahren..... 174
 - 7.2.1 Generalisierte Analyse des Markenwechselverhaltens mit Hilfe
eines modifizierten Assoziationsregelalgorithmus 174
 - 7.2.2 Ein Loyalitätsmaß für Switching-Matrizen höherer Ordnung..... 180
- 7.3 Auswertung eines realen Paneldatensatzes 184

Kapitel 8 Abschließende Betrachtungen..... 193

Anhang 197

Literaturverzeichnis 221

Abbildungsverzeichnis

Abbildung 2.1:	Begriffsabgrenzung Data Mining im engeren und im weiteren Sinne	11
Abbildung 2.2:	KDD als interdisziplinärer Bereich	16
Abbildung 2.3:	Das Prozeßmodell von Brachman, Anand (1996)	22
Abbildung 2.4:	Das Prozeßmodell von Chapman et al. (1998)	25
Abbildung 2.5:	Generic Tasks und Outputs des Prozeßmodells von Chapman et al. (1998)	27
Abbildung 2.6:	Das Prozeßmodell von Fayyad et al. (1996a)	28
Abbildung 2.7:	Das Prozeßmodell von John (1997)	30
Abbildung 2.8:	Das Prozeßmodell von Wirth, Reinartz (1996)	32
Abbildung 2.9:	Entscheidungsfindungsprozeß im interdisziplinären Forschungsumfeld Datenanalyse, Expertenwissen und Entscheidungsunterstützung	34
Abbildung 2.10:	Die fünf Hauptschritte des KDD-Prozesses und ihre Zuordnung in ausgewählten Prozeßmodellen	35
Abbildung 2.11:	Eigener Ansatz eines KDD-Prozeßmodells mit den fünf Hauptschritten und den wichtigsten Unteraufgaben	39
Abbildung 2.12:	Mögliche Data Mining Verfahrensklassen unterteilt nach mustererkennenden und musterbeschreibenden Verfahren	43
Abbildung 2.13:	Eine Auswahl wichtiger Data Mining Verfahren, unterteilt nach der Analyse-Aufgabe	46
Abbildung 3.1:	Jahr der Veröffentlichung der betrachteten Berichte	51
Abbildung 3.2:	Anwendungsbereiche der betrachteten Berichte	52
Abbildung 3.3:	In den betrachteten Berichten eingesetzte Verfahren	53
Abbildung 3.4:	In den betrachteten Berichten bearbeitete Data Mining Aufgaben ...	55
Abbildung 3.5:	Zahl der „False Alarms“	59
Abbildung 3.6:	Zahl der richtig erkannten betrügerischen Transaktionen	59

Abbildung 3.7:	Top Two Vendors Seriously Under Evaluation for Data Mining Tools	65
Abbildung 3.8:	Data Mining Verfahren der betrachteten Softwareprodukte	68
Abbildung 3.9:	Merkmale zur Untersuchung der Unterstützungsmöglichkeiten der Hauptschritte des KDD-Prozeßmodells	70
Abbildung 3.10:	Segmentierung und Positionierung der ausgewählten Data Mining Softwareprodukte	73
Abbildung 3.11:	Dendrogramm des Average Linkage Verfahrens	74
Abbildung 4.1:	Algorithmus zur divisiven Induktion eines Entscheidungsbaumes ...	86
Abbildung 4.2:	Entscheidungsbaum zum Datensatz aus Tabelle 4.2, für den nur Unterteilungen der Form (4.1) zugelassen sind	87
Abbildung 4.3:	Entscheidungsbaum zum Datensatz aus Tabelle 4.2, für den nur Unterteilungen der Form (4.3) zugelassen sind	87
Abbildung 4.4:	Mit Hilfe des Informationsgewinnes generierter Baum.....	90
Abbildung 4.5:	Mit Hilfe des Gini-Index generierter Baum.....	95
Abbildung 4.6:	Großer Entscheidungsbaum aus den Daten der Tabelle 4.4	99
Abbildung 5.1:	Der AIS-Algorithmus	114
Abbildung 5.2:	Der SETM-Algorithmus	115
Abbildung 5.3:	Der Apriori-Algorithmus	117
Abbildung 5.4:	Der AprioriTID-Algorithmus	120
Abbildung 5.5:	Der PARTITION-Algorithmus.....	123
Abbildung 5.6:	Graphische Darstellung der Hierarchie \mathcal{K}	124
Abbildung 5.7:	Der Algorithmus Cumulate.....	127
Abbildung 5.8:	Die häufigen Mengen für $s_{\min} = 1/3$	129
Abbildung 6.1:	Übersicht über das Vorgehen des CART-LC-Algorithmus zur Bestimmung eines multivariaten Splits im aktuellen Knoten des Baumes.....	141
Abbildung 6.2:	Übersicht über das Vorgehen des OC1-Algorithmus zur Bestimmung eines multivariaten Splits im aktuellen Knoten des Baumes.....	144
Abbildung 6.3:	Schematische Darstellung des DTAR-Verfahrens	148
Abbildung 6.4:	Allgemeine Vorgehensweise zur Generierung der häufigen Itemmengen im DTAR-Verfahren	150
Abbildung 6.5:	Univariater Entscheidungsbaum zum Datensatz aus Tabelle 6.3	153
Abbildung 6.6:	Multivariater Entscheidungsbaum des DTAR-Verfahrens zum Datensatz aus Tabelle 6.3	155
Abbildung 6.7:	Vorgehensweise für quantitative Merkmale zur Generierung der häufigen Itemmengen im DTAR-Verfahren	157
Abbildung 6.8:	Berechnung der Unterschranke des minimalen Supports am Beispiel des Gini-Index	160

Abbildung 6.9:	Durchschnittliche Zahl der Endknoten innerhalb der Simulationsstudie	167
Abbildung 6.10:	Durchschnittliche Klassifikationsgüte innerhalb der Simulationsstudie	168
Abbildung 7.1:	Modifizierter Apriori-Algorithmus für Kauffolgen	178
Abbildung 7.2:	Mit Hilfe des modifizierten Apriori-Algorithmus für Kauffolgen generierte Mengen	179
Abbildung 7.3:	Mit Hilfe des modifizierten Apriori-Algorithmus für Kauffolgen generierte Switching-Regeln	180
Abbildung 7.4:	Loyalitäts-Werte der Marke 2 (Ariel) im Zeitverlauf	191
Abbildung 7.5:	Loyalitäts-Werte der Marke 5 (Sunil) im Zeitverlauf	191

Tabellenverzeichnis

Tabelle 1.1:	Übersicht über Definitionen des Begriffes Data Mining in der Literatur	4
Tabelle 2.1:	Verschiedene Beschreibungen der wichtigsten Data Mining Aufgaben	41
Tabelle 3.1:	Ausgewählte Data Mining Anwendungsberichte von 1996-1998	49
Tabelle 3.2:	Ausführlichere Betrachtung der Verfahrensklassen Neuronale Netze und Entscheidungsbaumverfahren	54
Tabelle 3.3:	Data Mining Verfahren aus Abbildung 2.13 und ihr Einsatz in der vorliegenden Studie	54
Tabelle 3.4:	Studien zum Vergleich von Data Mining Softwareprodukten	57
Tabelle 3.5:	Ease of Use Comparison	58
Tabelle 3.6:	Bewertung von fünf Data Mining Tools	61
Tabelle 3.7:	Bedienbarkeit ausgewählter Data Mining Tools	62
Tabelle 3.8:	Automatisierungsmöglichkeiten ausgewählter Data Mining Tools	63
Tabelle 3.9:	Stärken und Schwächen ausgewählter Data Mining Software Tools	63
Tabelle 3.10:	Ausgewählte Data Mining Tools	66
Tabelle 3.11:	Data Mining Aufgaben und Verfahren der vorgestellten Software- produkte	67
Tabelle 3.12:	Gegenüberstellung der am meisten eingesetzten Data Mining Verfahren der Studien aus Kapitel 3	69
Tabelle 3.13:	Unterstützungsmöglichkeiten der ausgewählten Data Mining Softwareprodukte für drei der Hauptschritte des KDD-Prozesses sowie zusätzliche Eigenschaften	72
Tabelle 3.14:	Durchschnittliche Ausprägungswerte der 4-Klassenlösung des Average Linkage-Verfahrens	74
Tabelle 4.1:	Kontingenztafel für die Unterteilung $S_k(L^p)$	85
Tabelle 4.2:	Beispiel eines Kundendatensatzes	87
Tabelle 4.3:	Kontingenztafel für die Unterteilung $S_k(L)$	89
Tabelle 4.4:	Kundendatensatz	94

Tabelle 4.5:	Werte der Unreinheitsmaße für die wichtigsten Unterteilungen der Merkmale Einkommen und Alter für die gesamten Trainingsdaten in der Wurzel des Baumes	95
Tabelle 4.6:	Werte des Error-Complexity-Pruning	100
Tabelle 4.7:	Werte für Kriterium (4.12) für die Knoten des Baumes aus Abbildung 4.6	102
Tabelle 4.8:	Schätzer für die falschen Vorhersagen des Baumes aus Abbildung 4.2	103
Tabelle 4.9:	Eigenschaften der vorgestellten Post-Pruning-Techniken	104
Tabelle 4.10:	Eigenschaften der betrachteten Entscheidungsbaumverfahren	104
Tabelle 4.11:	Entscheidungsbaum-Softwareprodukte	107
Tabelle 5.1:	Zwischenergebnisse des Apriori-Algorithmus	118
Tabelle 5.2:	Zwischenergebnisse des AIS-Algorithmus	119
Tabelle 5.3:	Mögliche Datenbasis aus Kundentransaktionen	125
Tabelle 5.4:	Mögliche Assoziationsregeln	125
Tabelle 5.5:	Einfache Datenbasis aus Kundeninformationen	130
Tabelle 5.6:	Binärdarstellung der Datenbasis aus Tabelle 5.5	131
Tabelle 5.7:	Mögliche quantitative Assoziationsregeln	131
Tabelle 5.8:	Vierfeldertafel für paarweise Verbundbeziehungen	134
Tabelle 5.9:	Vierfeldertafel für Assoziationsregeln $X \rightarrow Y$ und $Y \rightarrow X$	135
Tabelle 5.10:	Kontingenztafel für die Ereignisse Tee und Kaffee	136
Tabelle 5.11:	Eigenschaften der vorgestellten Bewertungsmaße	138
Tabelle 6.1:	Kontingenztafel zur Unterteilung $S_2(L,W)$	151
Tabelle 6.2:	Darstellung der Kontingenztafel aus Tabelle 6.1 mit Hilfe der Support-Werte der zugehörigen Assoziationsregeln	152
Tabelle 6.3:	Beispieldatensatz bestehend aus drei Merkmalen und zwei Klassenbezeichnungen	153
Tabelle 6.4:	Support-Werte der einelementigen Itemmengen	154
Tabelle 6.5:	Support-Werte der dreielementigen Itemmengen	154
Tabelle 6.6:	Gesamtzahl der Teilintervalle des Merkmals „Anzahl der Kinder“ des Kundendatensatzes aus Tabelle 4.4	158
Tabelle 6.7:	Elemente aus I	158
Tabelle 6.8:	Elemente aus \hat{I}	158
Tabelle 6.9:	Zur Analyse verwendete Datensätze und ihre Eigenschaften	163
Tabelle 6.10:	Durchschnittliche Klassifikationsgüte der angewendeten Verfahren bei 10-facher Kreuzvalidierung	164
Tabelle 6.11:	Zahl der Endknoten und maximale Tiefe der generierten Bäume bei 10-facher Kreuzvalidierung	165
Tabelle 7.1:	Ansätze zur Marktstrukturanalyse basierend auf aggregierten Markenwechseldaten	172

Tabelle 7.2:	Hauptkategorien von Loyalitätsmaßen	173
Tabelle 7.3:	Untersuchte Marken und zugehörige Marktanteile	184
Tabelle 7.4:	Zahl der Regeln für ausgewählte Support-Schranken	185
Tabelle 7.5:	Switching-Regeln (X, Y) mit $\bar{s}_0(X \cup, Y) \geq 180$	186
Tabelle 7.6:	Zahl der Haushalte bezüglich der Mindestanzahl von Kaufakten	187
Tabelle 7.7:	Loyalitäts-Werte des verallgemeinerten Modells für Switching- Matrizen höherer Ordnung	188
Tabelle 7.8:	Anpassungsgüte der geschätzten $\hat{p}_{q_1; q_2 \dots q_l}$ Werte an die beobachteten Größen $p_{q_1; q_2 \dots q_l}$	188
Tabelle 7.9:	Eigenschaften der untersuchten Teilintervalle des Betrachtungs- zeitraumes	189
Tabelle 7.10:	Werte der verallgemeinerten Loyalitätsmaße $\alpha'_q(\kappa)$ für Marke q , Switching-Regeln der Länge l und Teilintervall k	189
Tabelle 7.11:	Anpassungsgüte der gefundenen Lösungen für die betrachteten Teilintervalle	190