

Inhaltsverzeichnis

1	Einführung	1
1.1	Auswertung von Massendaten	1
1.2	Data Mining und Business Intelligence	3
1.3	Ablauf einer Datenanalyse	4
1.4	Interdisziplinarität	11
1.5	Erfolgreiche Beispiele	14
1.6	Werkzeuge	16
1.6.1	KNIME	17
1.6.2	WEKA	26
1.6.3	JavaNNS	31
2	Grundlagen des Data Mining	37
2.1	Grundbegriffe	37
2.2	Datentypen	39
2.3	Abstands- und Ähnlichkeitsmaße	43
2.4	Grundlagen Künstlicher Neuronaler Netze	47
2.5	Logik	52
2.6	Überwachtes und unüberwachtes Lernen	55
3	Anwendungsklassen	57
3.1	Cluster-Analyse	57
3.2	Klassifikation	59
3.3	Numerische Vorhersage	61
3.4	Assoziationsanalyse	63
3.5	Text Mining	65
3.6	Web Mining	66

4	Wissensrepräsentation	69
4.1	Entscheidungstabelle	69
4.2	Entscheidungsbäume	71
4.3	Regeln	72
4.4	Assoziationsregeln	73
4.5	Instanzenbasierte Darstellung	79
4.6	Repräsentation von Clustern	79
4.7	Neuronale Netze als Wissensspeicher	80
5	Klassifikation	83
5.1	K-Nearest Neighbour	83
5.1.1	K-Nearest-Neighbour-Algorithmus	85
5.1.2	Ein verfeinerter Algorithmus	89
5.2	Entscheidungsbaumlernen	92
5.2.1	Erzeugen eines Entscheidungsbaums	92
5.2.2	Auswahl eines Attributs	94
5.2.3	Der ID3-Algorithmus zur Erzeugung eines Entscheidungsbaums	96
5.2.4	Entropie	104
5.2.5	Der Gini-Index	106
5.2.6	Der C4.5-Algorithmus	106
5.2.7	Probleme beim Entscheidungsbaumlernen	108
5.2.8	Entscheidungsbaum und Regeln	109
5.3	Naive Bayes	111
5.3.1	Bayessche Formel	111
5.3.2	Der Naive-Bayes-Algorithmus	112
5.4	Vorwärtsgerichtete Neuronale Netze	117
5.4.1	Architektur	117
5.4.2	Das Backpropagation-of-Error-Lernverfahren	119
5.4.3	Modifikationen des Backpropagation-Algorithmus	123
5.4.4	Ein Beispiel	125
5.5	Support Vector Machines	128
5.5.1	Grundprinzip	128
5.5.2	Formale Darstellung von Support Vector Machines	130
5.5.3	Ein Beispiel	132
6	Cluster-Analyse	135
6.1	Arten der Cluster-Analyse	135
6.2	Der k-Means-Algorithmus	139
6.3	Der k-Medoid-Algorithmus	148

6.4	Erwartungsmaximierung	153
6.5	Agglomeratives Clustern	155
6.6	Dichtebasiertes Clustern	160
6.7	Clusterbildung mittels selbstorganisierender Karten	163
6.7.1	Aufbau	163
6.7.2	Lernen	164
6.7.3	Visualisierung einer SOM	167
6.7.4	Ein Beispiel	168
6.8	Clusterbildung mittels neuronaler Gase	170
6.9	Clusterbildung mittels ART	172
7	Assoziationsanalyse	175
7.1	Der A-Priori-Algorithmus	175
7.1.1	Generierung der Kandidaten	177
7.1.2	Erzeugen der Regeln	179
7.2	Frequent Pattern Growth	185
7.3	Assoziationsregeln für spezielle Aufgaben	189
7.3.1	Hierarchische Assoziationsregeln	189
7.3.2	Quantitative Assoziationsregeln	190
7.3.3	Erzeugung von temporalen Assoziationsregeln	192
8	Datenvorbereitung	195
8.1	Motivation	195
8.2	Arten der Datenvorbereitung	197
8.2.1	Datenselektion und -integration	198
8.2.2	Datensäuberung	199
8.2.3	Datenreduktion	206
8.2.4	Datentransformation	209
8.3	Ein Beispiel	215
9	Bewertung	221
9.1	Prinzip der minimalen Beschreibungslängen	222
9.2	Interessantheitsmaße für Assoziationsregeln	222
9.2.1	Support	223
9.2.2	Konfidenz	223
9.2.3	Gain-Funktion	225
9.2.4	p - s -Funktion	226
9.2.5	Lift	227
9.3	Gütemaße und Fehlerkosten	227
9.3.1	Fehlerraten	227

9.3.2	Weitere Gütemaße für Klassifikatoren	228
9.3.3	Fehlerkosten	230
9.4	Testmengen	231
9.5	Qualität von Clustern	233
9.6	Visualisierung	235
10	Eine Data-Mining-Aufgabe	245
10.1	Die Aufgabe	245
10.2	Das Problem	246
10.3	Die Daten	248
10.4	Datenvorbereitung	253
10.5	Experimente	256
10.5.1	K-Nearest Neighbour	258
10.5.2	Naive Bayes	260
10.5.3	Entscheidungsbaumverfahren	262
10.5.4	Neuronale Netze	265
10.6	Auswertung der Ergebnisse	272
A	Anhang	275
A.1	Iris-Daten	275
A.2	Sojabohnen	277
A.3	Wetter-Daten	279
A.4	Kontaktlinsen-Daten	281
	Abbildungsverzeichnis	283
	Tabellenverzeichnis	291
	Verzeichnis der Symbole	293
	Verzeichnis der Abkürzungen	295
	Literaturverzeichnis	297
	Index	303