

Jürgen Cleve, Uwe Lämmel

Data Mining

2. Auflage

**DE GRUYTER
OLDENBOURG**

Inhaltsverzeichnis

1	Einführung	1
1.1	Auswertung von Massendaten	1
1.2	Data Mining und Business Intelligence	3
1.3	Ablauf einer Datenanalyse	5
1.4	Interdisziplinarität	12
1.5	Erfolgreiche Beispiele	15
1.6	Werkzeuge	17
1.6.1	KNIME	18
1.6.2	WEKA	27
1.6.3	JavaJVNS	32
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	107
5.2.7	Probleme beim Entscheidungsbaumlernen	108
5.2.8	Entscheidungsbaum und Regeln	109
5.3	Naive Bayes	111
5.3.1	Bayessche Formel	112
5.3.2	Der Naive-Bayes-Algorithmus	112
5.4	Vorwärtsgerichtete Neuronale Netze	117
5.4.1	Architektur	118
5.4.2	Das Backpropagation-of-Error-Lernverfahren	120
5.4.3	Modifikationen des Backpropagation-Algorithmus	125
5.4.4	Ein Beispiel	126
5.5	Support Vector Machines	130
5.5.1	Grundprinzip	130
5.5.2	Formale Darstellung von Support Vector Machines	131
5.5.3	Ein Beispiel	133
6	Cluster-Analyse	137
6.1	Arten der Cluster-Analyse	137
6.2	Der k-Means-Algorithmus	141
6.3	Der k-Medoid-Algorithmus	151

6.4	Erwartungsmaximierung	156
6.5	Agglomeratives Clustern	158
6.6	Dichtebasiertes Clustern	163
6.7	Clusterbildung mittels selbstorganisierender Karten	166
6.7.1	Aufbau	166
6.7.2	Lernen	167
6.7.3	Visualisierung einer SOM	170
6.7.4	Ein Beispiel	170
6.8	Clusterbildung mittels neuronaler Gase	173
6.9	Clusterbildung mittels ART	175
6.10	Der Fuzzy-c-Means-Algorithmus	177
7	Assoziationsanalyse	181
7.1	Der A-Priori-Algorithmus	181
7.1.1	Generierung der Kandidaten	183
7.1.2	Erzeugen der Regeln	185
7.2	Frequent Pattern Growth	191
7.3	Assoziationsregeln für spezielle Aufgaben	195
7.3.1	Hierarchische Assoziationsregeln	195
7.3.2	Quantitative Assoziationsregeln	196
7.3.3	Erzeugung von temporalen Assoziationsregeln	198
8	Datenvorbereitung	201
8.1	Motivation	201
8.2	Arten der Datenvorbereitung	205
8.2.1	Datenselektion und-integration	205
8.2.2	Datensäuberung	207
8.2.3	Datenreduktion	212
8.2.4	Datentransformation	216
8.3	Ein Beispiel	224
9	Bewertung	229
9.1	Prinzip der minimalen Beschreibungslängen	230
9.2	Interessantheitsmaße für Assoziationsregeln	230
9.2.1	Support	231
9.2.2	Konfidenz	231
9.2.3	Gain-Funktion	233
9.2.4	p-s-Funktion	234
9.2.5	Lift	235

9.3	Gütemaße und Fehlerkosten	235
9.3.1	Fehlerraten	235
9.3.2	Weitere Gütemaße für Klassifikatoren	236
9.3.3	Fehlerkosten	239
9.4	Testmengen	240
9.5	Qualität von Clustern	242
9.6	Visualisierung	244
10	Eine Data-Mining-Aufgabe	255
10.1	Die Aufgabe	255
10.2	Das Problem	256
10.3	Die Daten	258
10.4	Datenvorbereitung	263
10.5	Experimente	266
10.5.1	K-Nearest Neighbour	268
10.5.2	Naive Bayes	270
10.5.3	Entscheidungsbaumverfahren	272
10.5.4	Neuronale Netze	275
10.6	Auswertung der Ergebnisse	282
A	Anhang	285
A.1	Iris-Daten	285
A.2	Sojabohnen	287
A.3	Wetter-Daten	289
A.4	Kontaktlinsen-Daten	291
	Abbildungsverzeichnis	293
	Tabellenverzeichnis	301
	Verzeichnis der Symbole	303
	Verzeichnis der Abkürzungen	305
	Literaturverzeichnis	307
	Index	313