

Inhalt

1	Einführung — 1
1.1	Auswertung von Massendaten — 1
1.2	Ablauf einer Datenanalyse — 2
1.3	Das Vorgehensmodell von Fayyad — 8
1.4	Interdisziplinarität von Data Mining — 11
1.5	Wozu Data Mining? — 17
1.6	Werkzeuge — 20
1.6.1	KNIME — 21
1.6.2	WEKA — 29
1.6.3	JavaNNS — 35
2	Grundlagen des Data Mining — 41
2.1	Grundbegriffe — 41
2.2	Datentypen — 43
2.3	Abstands- und Ähnlichkeitsmaße — 47
2.4	Grundlagen Künstlicher Neuronaler Netze — 51
2.5	Logik — 56
2.6	Überwachtes und unüberwachtes Lernen — 59
3	Anwendungsklassen — 61
3.1	Cluster-Analyse — 61
3.2	Klassifikation — 64
3.3	Numerische Vorhersage — 66
3.4	Assoziationsanalyse — 67
3.5	Text Mining — 69
3.6	Web Mining — 70
4	Wissensrepräsentation — 73
4.1	Entscheidungstabelle — 73
4.2	Entscheidungsbäume — 75
4.3	Regeln — 76
4.4	Assoziationsregeln — 77
4.5	Instanzenbasierte Darstellung — 83
4.6	Repräsentation von Clustern — 83
4.7	Neuronale Netze als Wissensspeicher — 85
5	Klassifikation — 87
5.1	K-Nearest Neighbour — 87
5.1.1	K-Nearest-Neighbour-Algorithmus — 89

5.1.2	Ein verfeinerter Algorithmus — 93
5.2	Entscheidungsbaumlernen — 96
5.2.1	Erzeugen eines Entscheidungsbaums — 96
5.2.2	Auswahl eines Attributs — 98
5.2.3	Der ID3-Algorithmus zur Erzeugung eines Entscheidungsbaums — 101
5.2.4	Entropie — 108
5.2.5	Der Gini-Index — 110
5.2.6	Der C4.5-Algorithmus — 111
5.2.7	Probleme beim Entscheidungsbaumlernen — 113
5.2.8	Entscheidungsbaum und Regeln — 114
5.3	Naive Bayes — 116
5.3.1	Bayessche Formel — 117
5.3.2	Der Naive-Bayes-Algorithmus — 117
5.4	Vorwärtsgerichtete Neuronale Netze — 125
5.4.1	Architektur — 125
5.4.2	Das Backpropagation-of-Error-Lernverfahren — 128
5.4.3	Modifikationen des Backpropagation-Algorithmus — 132
5.4.4	Ein Beispiel — 134
5.4.5	Convolutional Neural Networks — 137
5.5	Support Vector Machines — 138
5.5.1	Grundprinzip — 138
5.5.2	Formale Darstellung von Support Vector Machines — 140
5.6	Ensemble Learning — 144
5.6.1	Bagging — 145
5.6.2	Boosting — 145
5.6.3	Random Forest — 146
6	Cluster-Analyse — 147
6.1	Arten der Cluster-Analyse — 147
6.1.1	Partitionierende Cluster-Bildung — 147
6.1.2	Hierarchische Cluster-Bildung — 148
6.1.3	Dichtebasierte Cluster-Bildung — 150
6.1.4	Cluster-Analyse mit Neuronalen Netzen — 150
6.2	Der k-Means-Algorithmus — 151
6.3	Der k-Medoid-Algorithmus — 161
6.4	Erwartungsmaximierung — 167
6.5	Agglomeratives Clustern — 168
6.6	Dichtebasiertes Clustern — 172
6.7	Cluster-Bildung mittels selbstorganisierender Karten — 176
6.7.1	Aufbau — 176
6.7.2	Lernen — 177
6.7.3	Visualisierung einer SOM — 179

6.7.4	Ein Beispiel —	181
6.8	Cluster-Bildung mittels neuronaler Gase —	183
6.9	Cluster-Bildung mittels ART —	185
6.10	Der Fuzzy-c-Means-Algorithmus —	187
7	Assoziationsanalyse —	191
7.1	Der A-Priori-Algorithmus —	191
7.1.1	Generierung der Kandidaten —	193
7.1.2	Erzeugen der Regeln —	196
7.2	Frequent Pattern Growth —	201
7.3	Assoziationsregeln für spezielle Aufgaben —	206
7.3.1	Hierarchische Assoziationsregeln —	206
7.3.2	Quantitative Assoziationsregeln —	207
7.3.3	Erzeugung von temporalen Assoziationsregeln —	209
8	Datenvorbereitung —	211
8.1	Motivation —	211
8.2	Arten der Datenvorbereitung —	215
8.2.1	Datenselektion und -integration —	216
8.2.2	Datensäuberung —	217
8.2.3	Datenreduktion —	223
8.2.4	Ungleichverteilung des Zielattributs —	226
8.2.5	Datentransformation —	227
8.3	Ein Beispiel —	238
9	Bewertung —	245
9.1	Prinzip der minimalen Beschreibungslängen —	246
9.2	Interessantheitsmaße für Assoziationsregeln —	246
9.2.1	Support —	247
9.2.2	Konfidenz —	247
9.2.3	Completeness —	248
9.2.4	Gain-Funktion —	249
9.2.5	p -s-Funktion —	250
9.2.6	Lift —	251
9.2.7	Einordnung der Interessantheitsmaße —	252
9.3	Gütemaße und Fehlerkosten —	252
9.3.1	Fehlerraten —	252
9.3.2	Weitere Gütemaße für Klassifikatoren —	253
9.3.3	Fehlerkosten —	257
9.4	Testmengen —	258
9.5	Qualität von Clustern —	260
9.6	Visualisierung —	263

10 Eine Data-Mining-Aufgabe — 273

- 10.1 Die Aufgabe — 273
- 10.2 Das Problem — 274
- 10.3 Die Daten — 276
- 10.4 Datenvorbereitung — 282
- 10.5 Experimente — 284
 - 10.5.1 K-Nearest Neighbour — 286
 - 10.5.2 Naive Bayes — 289
 - 10.5.3 Entscheidungsbaumverfahren — 291
 - 10.5.4 Neuronale Netze — 295
- 10.6 Auswertung der Ergebnisse — 302

A Anhang – Beispieldaten — 305

- A.1 Iris-Daten — 305
- A.2 Sojabohnen — 307
- A.3 Wetter-Daten — 308
- A.4 Kontaktlinsen-Daten — 310

Literatur — 313

Stichwortverzeichnis — 317