

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 — 30
 - 1.6.3 JavaNNS — 34
 - 1.6.4 Python — 39

2 Grundlagen des Data Mining — 45

- 2.1 Grundbegriffe — 45
- 2.2 Datentypen — 47
- 2.3 Abstands- und Ähnlichkeitsmaße — 51
- 2.4 Grundlagen Künstlicher Neuronaler Netze — 55
- 2.5 Logik — 59
- 2.6 Überwachtes und unüberwachtes Lernen — 63

3 Anwendungsklassen — 65

- 3.1 Cluster-Analyse — 66
- 3.2 Klassifikation — 68
- 3.3 Numerische Vorhersage — 70
- 3.4 Assoziationsanalyse — 72
- 3.5 Text Mining — 74
- 3.6 Web Mining — 75

4 Wissensrepräsentation — 77

- 4.1 Entscheidungstabelle — 77
- 4.2 Entscheidungsbäume — 79
- 4.3 Regeln — 81
- 4.4 Assoziationsregeln — 82
- 4.5 Instanzenbasierte Darstellung — 87
- 4.6 Repräsentation von Clustern — 88
- 4.7 Neuronale Netze als Wissensspeicher — 89

5 Klassifikation — 91

- 5.1 K-Nearest Neighbour — 91
- 5.1.1 K-Nearest-Neighbour-Algorithmus — 93
- 5.1.2 Ein verfeinerter Algorithmus — 97
- 5.2 Entscheidungsbaumlernen — 100
- 5.2.1 Erzeugen eines Entscheidungsbaums — 100
- 5.2.2 Auswahl eines Attributs — 102
- 5.2.3 Der ID3-Algorithmus zur Erzeugung eines Entscheidungsbaums — 105
- 5.2.4 Entropie — 112
- 5.2.5 Der Gini-Index — 114
- 5.2.6 Der C4.5-Algorithmus — 115
- 5.2.7 Probleme beim Entscheidungsbaumlernen — 117
- 5.2.8 Entscheidungsbaum und Regeln — 118
- 5.3 Naive Bayes — 121
- 5.3.1 Bayessche Formel — 121
- 5.3.2 Der Naive-Bayes-Algorithmus — 122
- 5.4 Vorwärtsgerichtete Neuronale Netze — 130
- 5.4.1 Architektur — 130
- 5.4.2 Das Backpropagation-of-Error-Lernverfahren — 133
- 5.4.3 Modifikationen des Backpropagation-Algorithmus — 137
- 5.4.4 Ein Beispiel — 138
- 5.4.5 Convolutional Neural Networks — 142
- 5.5 Support Vector Machines — 143
- 5.5.1 Grundprinzip — 143
- 5.5.2 Formale Darstellung von Support Vector Machines — 145
- 5.6 Ensemble Learning — 149
- 5.6.1 Bagging — 150
- 5.6.2 Boosting — 151
- 5.6.3 Random Forest — 152

6 Cluster-Analyse — 153

- 6.1 Arten der Cluster-Analyse — 153
- 6.1.1 Partitionierende Cluster-Bildung — 153
- 6.1.2 Hierarchische Cluster-Bildung — 154
- 6.1.3 Dichtebasierte Cluster-Bildung — 156
- 6.1.4 Cluster-Analyse mit Neuronalen Netzen — 156
- 6.2 Der k-Means-Algorithmus — 157
- 6.3 Der k-Medoid-Algorithmus — 167
- 6.4 Erwartungsmaximierung — 172
- 6.5 Agglomeratives Clustern — 174
- 6.6 Dichtebasiertes Clustern — 178

6.7	Cluster-Bildung mittels selbstorganisierender Karten —	181
6.7.1	Aufbau —	182
6.7.2	Lernen —	183
6.7.3	Visualisierung einer SOM —	185
6.7.4	Ein Beispiel —	187
6.8	Cluster-Bildung mittels neuronaler Gase —	189
6.9	Cluster-Bildung mittels ART —	191
6.10	Der Fuzzy-c-Means-Algorithmus —	193
7	Assoziationsanalyse —	197
7.1	Der A-Priori-Algorithmus —	197
7.1.1	Generierung der Kandidaten —	199
7.1.2	Erzeugen der Regeln —	201
7.2	Frequent Pattern Growth —	208
7.3	Assoziationsregeln für spezielle Aufgaben —	212
7.3.1	Hierarchische Assoziationsregeln —	212
7.3.2	Quantitative Assoziationsregeln —	213
7.3.3	Erzeugung von temporalen Assoziationsregeln —	215
8	Datenvorbereitung —	217
8.1	Motivation —	217
8.2	Arten der Datenvorbereitung —	220
8.2.1	Datenselektion und -integration —	221
8.2.2	Datensäuberung —	222
8.2.3	Datenreduktion —	228
8.2.4	Ungleichverteilung des Zielattributs —	231
8.2.5	Datentransformation —	232
8.3	Ein Beispiel —	243
9	Bewertung —	249
9.1	Prinzip der minimalen Beschreibungslängen —	250
9.2	Interessantheitsmaße für Assoziationsregeln —	250
9.2.1	Support —	251
9.2.2	Konfidenz —	251
9.2.3	Completeness —	252
9.2.4	Gain-Funktion —	253
9.2.5	p -s-Funktion —	254
9.2.6	Lift —	255
9.2.7	Einordnung der Interessantheitsmaße —	256

XII — Inhalt

9.3	Gütemaße und Fehlerkosten —	256
9.3.1	Fehlerrate —	256
9.3.2	Weitere Gütemaße für Klassifikatoren —	257
9.3.3	Fehlerkosten —	261
9.4	Testmengen —	262
9.5	Qualität von Clustern —	264
9.6	Visualisierung —	267
10	Eine Data-Mining-Aufgabe —	277
10.1	Die Aufgabe —	277
10.2	Das Problem —	278
10.3	Die Daten —	280
10.4	Datenvorbereitung —	286
10.5	Experimente —	288
10.5.1	K-Nearest Neighbour —	290
10.5.2	Naive Bayes —	293
10.5.3	Entscheidungsbaumverfahren —	295
10.5.4	Neuronale Netze —	298
10.6	Python-Programm —	305
10.6.1	Datenvorverarbeitung —	306
10.6.2	Modellentwicklung —	311
10.6.3	Anwendung des Modells – die Vorhersage —	315
10.7	Auswertung der Ergebnisse —	320
A	Anhang – Beispieldaten —	323
A.1	Iris-Daten —	323
A.2	Sojabohnen —	324
A.3	Wetter-Daten —	326
A.4	Kontaktlinsen-Daten —	327
Literatur —		329
Stichwortverzeichnis —		333