

Wilhelm Nüsser (Hrsg.)  
Carsten Weigand (Hrsg.)  
Raphael Fockel (Autor)

## **Methoden des Data Mining im praktischen Einsatz**

FHDW Paderborn / Bielefeld  
Fachbericht Nr. 1/2009



FHDW-Fachbericht

Band 1/2009

**Raphael Fockel**  
**Wilhelm Nüsser (Hrsg.)**  
**Carsten Weigand (Hrsg.)**

**Methoden des Data Mining im praktischen Einsatz**

Shaker Verlag  
Aachen 2009

**Bibliografische Information der Deutschen Nationalbibliothek**

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Copyright Shaker Verlag 2009

Alle Rechte, auch das des auszugsweisen Nachdruckes, der auszugsweisen oder vollständigen Wiedergabe, der Speicherung in Datenverarbeitungsanlagen und der Übersetzung, vorbehalten.

Printed in Germany.

ISBN 978-3-8322-8775-7

ISSN 1861-3292

Shaker Verlag GmbH • Postfach 101818 • 52018 Aachen

Telefon: 02407 / 95 96 - 0 • Telefax: 02407 / 95 96 - 9

Internet: [www.shaker.de](http://www.shaker.de) • E-Mail: [info@shaker.de](mailto:info@shaker.de)

# Vorwort

Die Bewältigung drastisch steigender Datenmengen auf der einen und die Notwendigkeit auf sich schnell verändernde Umgebungen angemessen zu reagieren – vor diesen Herausforderungen stehen Wissenschaft, Unternehmen und Einzelpersonen heute. Data Mining ist die „Kunst“ nützliche Informationen aus diesen großen Datenmengen zu extrahieren, um dann angemessen agieren zu können. Dabei stehen neben der Analyse quantitativer Daten auch zunehmend Techniken zur Untersuchung von Textdaten im Blickpunkt.

Wegen der großen Datenmengen ist der Wunsch nach Verfahren entstanden, automatisiert interessante Muster zu extrahieren. Die stetige Verbesserung der Leistungsfähigkeit der Informationstechnik ist ein zusätzlicher Faktor, der die Entwicklung solcher Verfahren unterstützt.

Herr Raphael Fockel gibt in der vorliegenden Arbeit einen umfangreichen Überblick über gängige Methoden des Data Mining, wobei auch entsprechende Software berücksichtigt wird. Ferner wird aufgezeigt, wie diese Methoden im praktischen, insbesondere betriebswirtschaftlichen Einsatz angewendet werden können.

Nach einem Überblick über das Thema Data Mining werden spezielle, häufig genutzte Data-Mining-Methoden im Detail vorgestellt. Weiterhin werden Software-Produkte, sowohl frei verfügbare als auch kommerzielle, vorgestellt. Hier nahm der Autor die mühselige aber verdienstvolle Arbeit auf sich, nach der jeweiligen Beschaffung und Installation die Funktionsweise und die Bedienung ausgewählter Softwareprodukte vorzustellen.

Schließlich wird ein Überblick über Anwendungsbereiche von Data-Mining-Methoden gegeben. Der Autor analysiert hier eine beachtliche Anzahl unterschiedlicher Quellen im Hinblick auf die Anwendung der Methoden, vor allem im betriebswirtschaftlichen Bereich.

In der Arbeit wird in systematischer Weise ein Überblick über die wichtigsten Facetten des Data Mining gegeben, wobei vorhandene Software wesentlich in die Betrachtung mit einbezogen wird. Eine solche Zusammenstellung suchte man in der Literatur bisher vergeblich. Es ist ein Werk entstanden, das nicht nur als Einführung in die Thematik des Data Mining dienen kann, sondern dem Leser einen fundierten Einblick in den derzeitigen Stand dieses Gebietes erlaubt. Es zeigt auch, dass die Kenntnis unterschiedlicher Verfahren des Data Mining von zentraler Bedeutung für die Gewinnung relevanter Informationen ist. Das umfangreiche Quellenverzeichnis erlaubt dem Interessierten schließlich, weiter in das Thema einzusteigen.

Paderborn, im Oktober 2009

Die Herausgeber:  
W. Nüßer  
C. Weigand

# Inhalt

<b>Abbildungsverzeichnis.....</b>	<b>IV</b>
<b>Tabellenverzeichnis.....</b>	<b>VI</b>
<b>Einleitung.....</b>	<b>1</b>
<b>1. Data Mining und Knowledge Discovery in Databases.....</b>	<b>3</b>
<b>2. Methoden des Data Mining.....</b>	<b>9</b>
2.1 Überblick.....	9
2.2 Clusteranalyse.....	12
2.3 Künstliche Neuronale Netze.....	19
2.4 Entscheidungsbäume.....	28
2.5 Assoziationsanalyse.....	32
2.6 Naive-Bayes-Klassifikation.....	35
2.7 k-Nächste-Nachbarn-Klassifikation.....	38
<b>3. Data-Mining-Software.....</b>	<b>41</b>
3.1 Überblick.....	41
3.2 Kommerzielle Produkte.....	42
3.2.1. SPSS.....	42
3.2.2 SAS.....	44
3.2.3 Statistica.....	45
3.2.4 XLMiner.....	46
3.3 Frei verfügbare Produkte.....	47
3.3.1 R.....	47
3.3.2 RapidMiner.....	48
3.3.3 Weka.....	49
3.3.4 KNIME.....	50
3.4 Fallbeispiele: Ablaufschritte ausgewählter Data-Mining-Methoden am Beispiel R und XLMiner.....	51
3.4.1 Überblick.....	51
3.4.2 „k-Nächste-Nachbarn-Verfahren“.....	51
3.4.3 Naive-Bayes-Verfahren.....	55
3.4.4 Assoziationsanalyse.....	58
<b>4. Bewertungskriterien.....</b>	<b>61</b>

4.1 Bewertung der Qualität der Daten.....	61
4.2 Bewertung der Methoden.....	63
4.3 Bewertung der Software.....	65
4.4 Bewertung der Ergebnisse.....	68
<b>5. Funktionale und branchenspezifische Anwendungsbereiche der Data-Mining-Methoden .....</b>	<b>74</b>
5.1 Überblick.....	74
5.2 Segmentierung.....	80
5.3 Fallbeispiel: Partitionierende Clusteranalyse bei der Kundensegmentierung am Beispiel der Produkte Statistica, XLMiner, R und RapidMiner .....	90
5.4 Klassifikation .....	101
5.5 Prognose .....	109
5.6 Assoziation.....	111
<b>6. Weitere Anwendungsgebiete und Ausprägungen des Data Mining.....</b>	<b>114</b>
<b>Fazit und Ausblick .....</b>	<b>117</b>
<b>Anhang... ..</b>	<b>119</b>
<b>Glossar.....</b>	<b>119</b>
<b>Produktübersicht .....</b>	<b>122</b>
<b>Literatur.....</b>	<b>124</b>

## Abbildungsverzeichnis

Abbildung 1: Prozess des Knowledge Discovery in Databases (nach Fayyad (1996), S. 10) .....	5
Abbildung 2: Eingesetzte Methoden in Veröffentlichungen der Jahre 1994-1998 (eigene Darstellung, nach Säuberlich (2000), S. 52) .....	9
Abbildung 3: Eingesetzte Data-Mining-Methoden bei Web-Mining-Projekten (eigene Darstellung, nach Hippner u. a. (2002), S. 330) .....	10
Abbildung 4: Systematisierung wichtiger Data Mining Methoden (eigene Darstellung, in Anlehnung an Pietsch (2003), S. 80; Knobloch (2000), S. 23).....	11
Abbildung 5: Gegenüberstellung von Ähnlichkeit und Distanz (eigene Darstellung) ...	14
Abbildung 6: Beispiel eines Dendrogramms (Weiß (2007), S. 125).....	17
Abbildung 7: Das Modell eines Neurons (eigene Darstellung) .....	20
Abbildung 8: Grundstruktur eines zweischichtigen neuronalen Netzes (Wiedmann, Buckler (2001, S. 57)).....	21
Abbildung 9: Ausgewählte Typen Künstlicher-Neuronaler-Netze-Verfahren (eigene Darstellung, nach Backhaus u. a. (1998), S. 755).....	23
Abbildung 10: Aufbau eines aktiven Neurons in der verdeckten Schicht (Backhaus u. a. (2008), S. 769) .....	24
Abbildung 11: Alternative Aktivierungsfunktion bei Künstlichen Neuronalen Netzen (Backhaus u. a. (2008), S. 771).....	26
Abbildung 12: Entscheidungsbaum mit 6 splits (Rudolph (2007), S. 214) .....	28
Abbildung 13: Datensatz "Gartentechnik".....	40
Abbildung 14: Screenshot eines Prozessflussdiagramms bei KNIME (www.knime.org) .....	41
Abbildung 15: Screenshot der Benutzeroberfläche von SPSS 17 (eigene Darstellung).....	43
Abbildung 16: Screenshot der Benutzeroberfläche von SAS (eigene Darstellung) .....	44
Abbildung 17: Screenshot der Benutzeroberfläche von Statistica (eigene Darstellung) .....	45
Abbildung 18: Screenshot der Benutzeroberfläche von XLMiner (eigene Darstellung).....	46
Abbildung 19: Screenshot der Benutzeroberfläche von R. Das Beispiel zeigt ein k-Means Clusterverfahren (eigene Darstellung) .....	47
Abbildung 20: Screenshot der Benutzeroberfläche von RapidMiner. Das Beispiel zeigt eine Visualisierung durch einen Entscheidungsbaum (eigene Darstellung).....	49
Abbildung 21: Screenshot von WEKA (eigene Darstellung) .....	50
Abbildung 22: Screenshot der Eingabemaske k-Nearest-Neighbor-Verfahren bei XLMiner (eigene Darstellung).....	53
Abbildung 23: Screenshot der Ausgabe Validierung bei XLMiner (eigene Darstellung) .....	54
Abbildung 24: Screenshot der Klassifikation bei XLMiner (eigene Darstellung).....	54
Abbildung 25: Screenshot der Klassifikationsmatrix bei XLMiner (eigene Darstellung) .....	54
Abbildung 26: Screenshot des Outputs bei XLMiner beim Naive-Bayes-Verfahren (eigene Darstellung).....	57
Abbildung 27: Screenshot des Outputs bei XLMiner beim Naive-Bayes-Verfahren (eigene Darstellung).....	58
Abbildung 28: Screenshot der Ausgabe bei einer Assoziationsanalyse bei XLMiner (eigene Darstellung).....	60
Abbildung 29: Beispiel einer ROC-Kurve. Dargestellt ist der Kurvenverlauf sowie vergleichsweise die Mittelsenkrechte der beiden Achsen (Eigene Darstellung)....	70

Abbildung 30: Anwendungsbereiche des Data Mining (eigene Darstellung, nach Säuberlich (2000), S. 52) .....	75
Abbildung 31: Anwendungsgebiete nach Funktionen (eigene Darstellung, nach Küppers (2000), S. 124) .....	76
Abbildung 32: Anwendungsgebiete nach Branchen (eigene Darstellung, nach Küppers (2000), S. 124) .....	77
Abbildung 33: Ausgewählte Data Mining Anwendungen in der Industrie (eigene Darstellung, nach Otte u. a. (2004), S. 32).....	78
Abbildung 34: Web-Mining-Taxonomie (eigene Darstellung, nach Säuberlich (2001), S. 106) .....	79
Abbildung 35: Beispiel Clusterbildung von Bankkunden .....	84
Abbildung 36: Screenshot eines Dendrogramms bei Statistica (eigene Darstellung).....	93
Abbildung 37: Screenshot der Mittelwerte der Cluster bei Statistica (eigene Darstellung) .....	94
Abbildung 38: Screenshot der Ausgaben bei Statistica (Varianzanalyse, Cluster und Distanzen) (eigene Darstellung).....	96
Abbildung 39: Screenshot der Eingabemaske beim K-Means-Verfahren bei XLMiner (eigene Darstellung) .....	98
Abbildung 40: Screenshot der Ausgabe beim k-Means-Verfahren bei XLMiner (eigene Darstellung).....	98
Abbildung 41: Screenshot der Grafik der Mittelwerte der drei Cluster bei RapidMiner (eigene Darstellung).....	99
Abbildung 42: Clusteranzahl und Mittelwerte bei RapidMiner (eigene Darstellung)..	100

## Tabellenverzeichnis

Tabelle 1: Aufgabenstellungen und eine Auswahl an Methoden des Data Mining (eigene Darstellung, in Anlehnung an Bankhofer (2004), S. 397) .....	10
Tabelle 2: Beispiel Einkaufstransaktionen.....	34
Tabelle 3: Tabelle mit Assoziationsregeln.....	35
Tabelle 4: Beispiel Elektronikfachgeschäft .....	37
Tabelle 5: Prozessschritte und Beispiele für statistische Verfahren im Rahmen der Bewertung von Datensätzen (eigene Darstellung, in Anlehnung an Otte u.a. (2004), S. 61-86; Hippner, Wilde (2001), S. 37-63).....	62
Tabelle 6: Bewertungsgrundlagen für Data-Mining-Methoden (eigene Darstellung, in Anlehnung an Küppers (1999), S. 87) .....	63
Tabelle 7: Beispiel für den Ansatz einer Bewertungsmatrix bei der Auswahl von Data-Mining-Produkten (eigene Darstellung, in Anlehnung an Collier u.a. (1998), S. 6) .....	68
Tabelle 8: Formale Darstellung einer Klassifikationstabelle (eigene Darstellung, in Anlehnung an Bonne, Armingier (2001), S. 229, Shmueli u. a. (2007), S. 56).....	69
Tabelle 9: Gütemaße für Prognosen (eigene Darstellung, nach Otte u. a., S. 141; Petersohn (2005), S. 170-172) .....	72
Tabelle 10: Beispiel Clusterlösung bei der Kundensegmentierung eines Kreditinstituts .....	83
Tabelle 11: Screenshot des Datensatzes "Einkauf" (eigene Darstellung) .....	92