

**Jürgen Sembdner**

**Data Mining**

Ein Überblick über Verfahren und Anwendungsfelder

**Diplomarbeit**

## **Bibliografische Information der Deutschen Nationalbibliothek:**

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

Dieses Werk sowie alle darin enthaltenen einzelnen Beiträge und Abbildungen sind urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsschutz zugelassen ist, bedarf der vorherigen Zustimmung des Verlanges. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen, Auswertungen durch Datenbanken und für die Einspeicherung und Verarbeitung in elektronische Systeme. Alle Rechte, auch die des auszugsweisen Nachdrucks, der fotomechanischen Wiedergabe (einschließlich Mikrokopie) sowie der Auswertung durch Datenbanken oder ähnliche Einrichtungen, vorbehalten.

Copyright © 2000 Diplom.de  
ISBN: 9783832428440

**Jürgen Sembdner**

## **Data Mining**

**Ein Überblick über Verfahren und Anwendungsfelder**



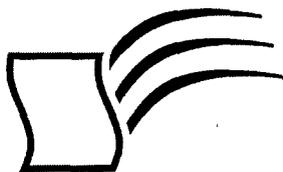
---

Jürgen Sembdner

# Data Mining

*Ein Überblick über Verfahren und Anwendungsfelder*

**Diplomarbeit**  
**an der FernUniversität - Gesamthochschule Hagen**  
**Fachbereich Wirtschaftswissenschaft**  
**April 2000 Abgabe**



***Diplomarbeiten Agentur***  
**Dipl. Kfm. Dipl. Hdl. Björn Bedey**  
**Dipl. Wi.-Ing. Martin Haschke**  
**und Guido Meyer GbR**

**Hermannstal 119 k**  
**22119 Hamburg**

**agentur@diplom.de**  
**www.diplom.de**

ID 2844

Sembdner, Jürgen: Data Mining: Ein Überblick über Verfahren und Anwendungsfelder /  
Jürgen Sembdner - Hamburg: Diplomarbeiten Agentur, 2000  
Zugl.: Hagen, Universität - Gesamthochschule, Diplom, 2000

---

Dieses Werk ist urheberrechtlich geschützt. Die dadurch begründeten Rechte, insbesondere die der Übersetzung, des Nachdrucks, des Vortrags, der Entnahme von Abbildungen und Tabellen, der Funksendung, der Mikroverfilmung oder der Vervielfältigung auf anderen Wegen und der Speicherung in Datenverarbeitungsanlagen, bleiben, auch bei nur auszugsweiser Verwertung, vorbehalten. Eine Vervielfältigung dieses Werkes oder von Teilen dieses Werkes ist auch im Einzelfall nur in den Grenzen der gesetzlichen Bestimmungen des Urheberrechtsgesetzes der Bundesrepublik Deutschland in der jeweils geltenden Fassung zulässig. Sie ist grundsätzlich vergütungspflichtig. Zuwiderhandlungen unterliegen den Strafbestimmungen des Urheberrechtes.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, daß solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Die Informationen in diesem Werk wurden mit Sorgfalt erarbeitet. Dennoch können Fehler nicht vollständig ausgeschlossen werden, und die Diplomarbeiten Agentur, die Autoren oder Übersetzer übernehmen keine juristische Verantwortung oder irgendeine Haftung für evtl. verbliebene fehlerhafte Angaben und deren Folgen.

Dipl. Kfm. Dipl. Hdl. Björn Bedey, Dipl. Wi.-Ing. Martin Haschke & Guido Meyer GbR  
Diplomarbeiten Agentur, <http://www.diplom.de>, Hamburg 2000  
Printed in Germany



## Wissensquellen gewinnbringend nutzen

**Qualität, Praxisrelevanz und Aktualität** zeichnen unsere Studien aus. Wir bieten Ihnen im Auftrag unserer Autorinnen und Autoren Wirtschaftsstudien und wissenschaftliche Abschlussarbeiten – Dissertationen, Diplomarbeiten, Masterarbeiten, Staatsexamensarbeiten und Studienarbeiten zum Kauf. Sie wurden an deutschen Universitäten, Fachhochschulen, Akademien oder vergleichbaren Institutionen der Europäischen Union geschrieben. Der Notendurchschnitt liegt bei 1,5.

**Wettbewerbsvorteile verschaffen** – Vergleichen Sie den Preis unserer Studien mit den Honoraren externer Berater. Um dieses Wissen selbst zusammenzutragen, müssten Sie viel Zeit und Geld aufbringen.

<http://www.diplom.de> bietet Ihnen unser vollständiges Lieferprogramm mit mehreren tausend Studien im Internet. Neben dem Online-Katalog und der Online-Suchmaschine für Ihre Recherche steht Ihnen auch eine Online-Bestellfunktion zur Verfügung. Inhaltliche Zusammenfassungen und Inhaltsverzeichnisse zu jeder Studie sind im Internet einsehbar.

**Individueller Service** – Gerne senden wir Ihnen auch unseren Papierkatalog zu. Bitte fordern Sie Ihr individuelles Exemplar bei uns an. Für Fragen, Anregungen und individuelle Anfragen stehen wir Ihnen gerne zur Verfügung. Wir freuen uns auf eine gute Zusammenarbeit.

### Ihr Team der *Diplomarbeiten* Agentur

#### ***Diplomarbeiten* Agentur**

Dipl. Kfm. Dipl. Hdl. Björn Bedey —  
Dipl. Wi.-Ing. Martin Haschke —  
und Guido Meyer GbR —

Hermannstal 119 k —  
22119 Hamburg —

Fon: 040 / 655 99 20 —  
Fax: 040 / 655 99 222 —

agentur@diplom.com —  
www.diplom.com —

## Inhaltsverzeichnis

1	Einleitung.....	6
1.1	Motivation zur Anwendung von „Data Mining“.....	6
1.2	Zielsetzung der Arbeit.....	7
1.3	Aufbau und Schwerpunktsetzung.....	8
2	Einordnung und Begriffsbestimmung.....	8
2.1	Der Gesamtprozess „Knowledge Discovery in Databases (KDD)“.....	8
2.2	Definition „Data Mining“.....	10
2.3	Abgrenzung zu anderen Disziplinen.....	12
2.3.1	Data Warehouse.....	12
2.3.2	Visualisierungstechniken.....	13
2.3.3	Statistik.....	14
2.3.4	Maschinelles Lernen.....	15
2.3.5	Expertensysteme.....	16
3	Eigenschaften von Data-Mining-Verfahren.....	17
3.1	Clusteranalyse.....	18
3.1.1	Hierarchische Clusterung.....	18
3.1.1.1	Agglomerative Methoden.....	23
3.1.1.2	Divisive Methoden.....	26
3.1.1.3	Eigenschaften hierarchischer Methoden.....	28
3.1.2	Partitionierende Clusterung.....	29
3.1.2.1	K-Means-Algorithmus.....	29
3.1.2.2	FKM-Algorithmus.....	33
3.1.2.3	Eigenschaften partitionierender Methoden.....	39
3.2	Statistische Verfahren .....	40
3.2.1	Bayes-Klassifikation.....	40
3.2.1.1	Beschreibung des Verfahrens .....	40
3.2.1.2	Eigenschaften der Bayes-Klassifikation.....	44
3.2.2	Assoziationsanalyse.....	45
3.2.2.1	Beschreibung des Verfahrens.....	45
3.2.2.2	Eigenschaften der Assoziationsanalyse.....	49

3.3	Induktives Lernen.....	49
3.3.1	Konzeptionelles Clustern.....	49
3.3.1.1	Beschreibung des Verfahrens .....	50
3.3.1.2	Eigenschaften des konzeptionellen Clusters.....	55
3.3.2	Entscheidungsbäume und –regeln.....	56
3.3.2.1	ID-3-Algorithmus.....	56
3.3.2.2	Eigenschaften des ID3-Algorithmus.....	61
3.4	Neuronale Netze.....	61
3.4.1	Multilayer-Perceptron.....	62
3.4.1.1	Beschreibung des Verfahrens.....	62
3.4.1.2	Eigenschaften des Multilayer-Perceptrons.....	66
3.4.2	Kohonen-Algorithmus.....	67
3.4.2.1	Beschreibung des Verfahrens .....	67
3.4.2.2	Eigenschaften des Kohonen-Algorithmus.....	71
3.5	Genetische Algorithmen.....	72
3.5.1	Basisalgorithmus.....	72
3.5.2	Anwendung auf „Data Mining“.....	74
3.5.3	Eigenschaften von Genetischen Algorithmen.....	78
4	Ausgewählte Anwendungsfelder für „Data Mining“.....	79
4.1	Warenkorbanalysen.....	80
4.2	Kundensegmentierung.....	81
4.3	Betrugserkennung.....	83
4.4	Datenreinigung.....	84
4.5	Prognose von Aktienkursen.....	85
5	Zusammenfassung und Ausblick.....	86
5.1	Data Mining zur Segmentierung.....	87
5.2	Data Mining zur Klassifizierung und Prognose.....	88
5.3	Data Mining zur Assoziationsanalyse.....	89
5.4	Genetische Algorithmen und hybride Verfahren.....	89
5.5	Eigenschaften der dargestellten Verfahren.....	90
5.6	Auswahl von Data-Mining-Werkzeugen.....	91
5.7	Anwendung von Data Mining auf andere Datentypen.....	91
5.8	Data Mining und strategische Wettbewerbsvorteile.....	92

## Abbildungsverzeichnis

Abbildung 1: Der KDD-Prozeß.....	9
Abbildung 2: Begriffsabgrenzung Data Mining.....	11
Abbildung 3: Data Mining: Verwandte Disziplinen.....	12
Abbildung 4: Parallele Koordinaten .....	14
Abbildung 5: Methodenüberblick Data Mining.....	17
Abbildung 6: Dendrogramm, agglomerative und divisive Clusterung.....	19
Abbildung 7: Verschiedenheitsmatrix nach Gower.....	22
Abbildung 8: Dendrogramm für das durchgängige Beispiel mit agglomerativer hierarchischer Clusterung.....	25
Abbildung 9: Dendrogramm für das durchgängige Beispiel mit divisiver hierarchischer Clusterung.....	27
Abbildung 10: ID3-Entscheidungsbaum für das durchgängige Beispiel .....	59
Abbildung 11: Multilayer-Perceptron für das durchgängige Beispiel .....	63
Abbildung 12: Kohonen-Karte.....	67
Abbildung 13: Selektion, Kreuzung und Mutation bei genetischen Algorithmen.....	73

## Tabellenverzeichnis

Tabelle 1: Kundendatensätze.....	21
Tabelle 2: Verschlüsselung des Berufsstatus.....	21
Tabelle 3: Ergebnis der agglomerativen hierarchischen Clusterung.....	25
Tabelle 4: Ergebnis der divisiven hierarchischen Clusterung.....	28
Tabelle 5: Normierte Daten des durchgängigen Beispiels.....	31
Tabelle 6: Clusterung nach dem ersten Durchlauf k-Means-Clusterung.....	32
Tabelle 7: Ergebnisse des k-Means-Verfahrens für das durchgängige Beispiel.....	33
Tabelle 8: Startpartition für das FKM-Verfahren.....	35
Tabelle 9: Clusterzentren im ersten Schritt des FKM-Verfahrens.....	35
Tabelle 10: Zugehörigkeitsmatrix nach der ersten Iteration des FKM-Verfahrens.....	36
Tabelle 11: Clusterzentren im zweiten Schritt des FKM-Verfahrens.....	37
Tabelle 12: Zugehörigkeitsmatrix nach der zweiten Iteration des FKM-Verfahrens.....	38

Tabelle 13: Umgerechnete Clusterzentren des FKM-Verfahrens.....	38
Tabelle 14: Beispieldaten für das Bayes-Verfahren.....	41
Tabelle 15: Schlüsselungen für das Bayes-Verfahren.....	42
Tabelle 16: Ergebnis der Wahrscheinlichkeitsschätzung.....	43
Tabelle 17: Testdaten für das Bayes-Verfahren.....	44
Tabelle 18: Wahrscheinlichkeiten für die Testdaten.....	44
Tabelle 19: Beispieldaten für die Assoziationsanalyse.....	46
Tabelle 20: Kombinationsmatrix zur Assoziationsanalyse.....	47
Tabelle 21: Konfidenz der Regeln mit Mindestsupport von 30 Prozent.....	48
Tabelle 22: Eingangsdaten für Cluster/2.....	50
Tabelle 23: Stars der Startcluster.....	51
Tabelle 24: Kombinationen der Starbeschreibungen.....	51
Tabelle 25: Zuordnung der Kunden zu den Clustern.....	52
Tabelle 26: Ausnahmelisten für Cluster/2.....	52
Tabelle 27: Zwischenergebnis Cluster/2.....	53
Tabelle 28: Mögliche Zuordnungen der Ausnahmeobjekte bei Kombination 2.....	54
Tabelle 29: Modifizierte Eingangsdaten für ID3.....	58
Tabelle 30: Klassifikation von Testdaten durch ID3.....	60
Tabelle 31: Input und Output der Neuronen für Kunde 1.....	64
Tabelle 32: Input und Output der Neuronen nach der ersten Gewichtskorrektur.....	65
Tabelle 33: Ergebnis der Kohonen-Clusterung für das durchgängige Beispiel .....	71
Tabelle 34: Startpopulation für den Genetischen Algorithmus.....	75
Tabelle 35: Fitneßwerte der Startpopulation.....	76
Tabelle 36: Neue Genome nach Selektion und Kreuzung.....	77
Tabelle 37: Neue Generation nach Selektion und Kreuzung.....	77