

ilmedia

 TECHNISCHE  
UNIVERSITÄT  
ILMENAU

---

*Bankhofer, Udo:*

## **Data Mining und seine betriebswirtschaftliche Relevanz**

---

In veränderter Fassung erschienen als:  
" Data Mining und seine betriebswirtschaftliche Relevanz ", in:  
Betriebswirtschaftliche Forschung und Praxis : BFuP. - Herne, Westf.  
: Verl. Neue Wirtschaftsbriefe, ISSN 0340-5370, 56. Jg. (2004), S.  
395-412

# Data Mining und seine betriebswirtschaftliche Relevanz

Von Udo Bankhofer

Data Mining, Knowledge Discovery in Databases, Clusteranalyse, Neuronale Netze, Entscheidungsbaumverfahren, Assoziationsanalyse

*Mit dem Begriff „Data Mining“ wird die Anwendung geeigneter Methoden zur Entdeckung von Strukturen und Beziehungen in großen Datenmengen umschrieben. Im Gegensatz zum Ansatz der induktiven Statistik stellen nicht im Vorfeld formulierte und zu überprüfende Hypothesen den Ausgangspunkt der Betrachtung dar; vielmehr sollen die in den Daten enthaltenen Informationen zur Theorieentwicklung genutzt werden. Damit wird mit dem Data Mining der eher modellgetriebenen Sichtweise der klassischen Statistik ein datengetriebener Ansatz vorangestellt, da auf Basis der in den Daten gefundenen Muster Hypothesen aufgestellt werden können, die dann mittels neuer Datensätze und konfirmatorischer Verfahren überprüft werden können. Darüber hinaus kann bereits das Entdecken von Strukturen ökonomisch wertvoll sein, was sich vor allem in betrieblichen Data-Mining-Anwendungen widerspiegelt.*

## 1 Gegenstand des Data Mining

### 1.1 Data Mining im Prozeß des Knowledge Discovery in Databases

Der Begriff „Knowledge Discovery in Databases“ (KDD) umschreibt den gesamten Prozeß der interaktiven und iterativen Entdeckung und Interpretation von nützlichem Wissen aus Daten.<sup>1</sup> Dazu müssen nach der Festlegung des Untersuchungsgegenstands zunächst geeignete Daten gesichtet und ausgewählt werden. Beispielsweise kann in einer Verkaufsdatenbank nur ein bestimmter Zeitraum von Interesse sein. Im Anschluß daran werden die Daten aufbereitet, um die Qualität des Datenbestandes zu verbessern. Dies bedeutet, daß vor allem fehlende Daten und Ausreißer berücksichtigt und adäquat behandelt werden müssen. Im Rahmen der Datenvorverarbeitung können dann gegebenenfalls noch Transformationen der Daten zur Vorbereitung auf die nachfolgende Analyse durchgeführt werden. Damit wird in erster Linie das Ziel verfolgt, die ausgewählten Daten auf das gemäß der Zielsetzung der Untersuchung notwendige Maß zu reduzieren. Durch den Einsatz von Data Warehouses oder Data Marts für einzelne Abteilungen eines Unternehmens können die eben skizzierten Schritte vereinfacht und beschleunigt werden, da mit diesen Systemen regelmäßig und systematisch Daten zusammengeführt werden.

Das Herzstück des KDD-Prozesses stellt schließlich das Data Mining dar. Dabei geht es um die Auswahl und Anwendung geeigneter Methoden zur Entdeckung von Mustern und Beziehungen in den Daten im Hinblick auf den betrachteten Untersuchungsgegenstand. Die gewonnenen Aussagen müssen jedoch noch auf ihre Plausibilität hin untersucht werden und letztendlich – und dies stellt einen weiteren wichtigen Schritt dar – in Erkenntnisse bzw. Wis-

---

<sup>1</sup> vgl. z.B. Düsing (1999, S. 346-347)

sen umgesetzt werden. Wie im letzten Teil dieser Arbeit noch ausführlich aufgezeigt wird, stellen die in den Daten entdeckten Muster eine wesentliche Grundlage zum Aufstellen betriebswirtschaftlich relevanter Hypothesen dar, wobei allein das Entdecken von Strukturen bereits ökonomisch wertvoll sein kann.

In der nachfolgenden Abbildung 1 ist der Prozeß des Knowledge Discovery in Databases noch einmal zusammengefaßt.<sup>2</sup> Dabei bleibt festzuhalten, daß zwischen den einzelnen Schritten jederzeit Rückkopplungen möglich sind. So kann beispielsweise bei der Anwendung von Data Mining Methoden festgestellt werden, daß die vorliegenden Daten erneut aufbereitet oder gegebenenfalls durch die Auswahl weiterer Daten ergänzt werden müssen. Dies verdeutlicht nochmals, daß das Data Mining in einen iterativen und interaktiven Gesamtprozeß der Wissensentdeckung eingebunden ist, bei dem vor allem auch die Vor- und Nachverarbeitung der Daten von zentraler Bedeutung sind.

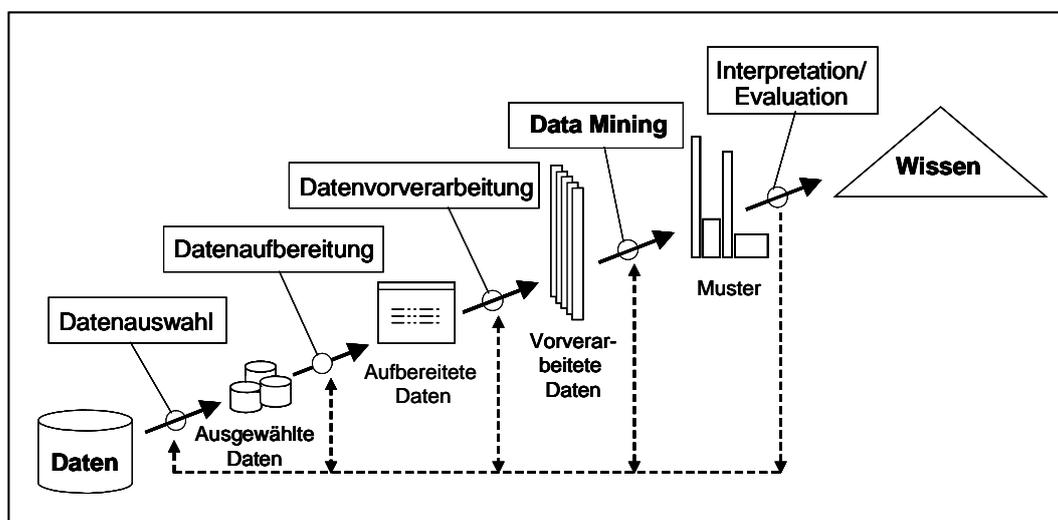


Abbildung 1: Prozeß des Knowledge Discovery in Databases

Überträgt man diesen KDD-Prozeß auf ein Unternehmen, dann stellt das Verständnis einer betrieblichen Problemstellung den Ausgangspunkt dar, der zur Auswahl entsprechender Daten führt. Aus dem resultierenden Wissen können dann Erkenntnisse zur Lösung der Problemstellung gewonnen werden, die zur Ableitung von Handlungsempfehlungen bzw. Strategien führen und damit zur Weiterentwicklung des Unternehmens beitragen. Daraus können schließlich neue und weiterführende betriebliche Problemstellungen resultieren, die den Prozess erneut auslösen.

## 1.2 Anwendungsbereiche und Methoden im Überblick

Es haben sich im Wesentlichen vier Anwendungsbereiche des Data Mining herausgebildet, und zwar die Segmentierung, die Klassifikation, die Vorhersage sowie der Ansatz der Assoziation.<sup>3</sup> Die Aufgabe der Segmentierung besteht in der Bildung von Gruppen ähnlicher Objek-

<sup>2</sup> vgl. *Fayyad et al.* (1996, S. 9-11)

<sup>3</sup> vgl. *Lusti* (2002, S. 261-262)

te. Die klassische betriebswirtschaftliche Anwendung stellt die Kundensegmentierung bzw. die Bildung entsprechender Typologien dar. Damit wird das Ziel verfolgt, Produkte, Dienste und Kommunikationsmaßnahmen auf die Bedürfnisse der gefundenen homogenen Zielgruppen abstimmen zu können. Ein weiterer Anwendungsbereich des Data Mining ist die Klassifikation. Im Gegensatz zur Segmentierung soll hier keine neue Klassifikation entdeckt, sondern die Klassenzugehörigkeit der Objekte anhand gegebener Merkmale bestmöglich identifiziert werden. Das Ergebnis einer Klassifikation kann und wird zwar zur Einteilung neuer Objekte in bereits vorliegende Klassen herangezogen, jedoch ist die Vorhersage als weiterer Anwendungsbereich des Data Mining davon abzugrenzen. Dies liegt daran, daß hier auf Basis einer geschätzten funktionalen Beziehung die Werte einer abhängigen quantitativen Variablen prognostiziert werden. Während beispielsweise eine klassifikatorische Bonitätsbeurteilung einen Kunden in die Klassen der kreditwürdigen und nicht kreditwürdigen Personen einordnet, könnte bei einer vorhersagenden Bonitätsbeurteilung die Bonität z.B. als maximal einräumbares Forderungsvolumen definiert werden. Den vierten Anwendungsbereich des Data Mining stellt schließlich die so genannte Assoziation dar. Als grundlegende Aufgabenstellung kann die Entdeckung struktureller Zusammenhänge in Datenbasen mit Hilfe von Assoziationsregeln genannt werden.

<b>Anwendungsbereich</b>	<b>Aufgabenstellung</b>	<b>Wesentliche Methoden</b>
Segmentierung	Bildung von Klassen aufgrund von Ähnlichkeiten der Objekte	<ul style="list-style-type: none"> <li>■ Clusteranalyse</li> <li>■ Neuronale Netze</li> </ul>
Klassifikation	Identifikation der Klassenzugehörigkeit von Objekten auf der Basis gegebener Merkmale	<ul style="list-style-type: none"> <li>■ Diskriminanzanalyse</li> <li>■ Neuronale Netze</li> <li>■ Entscheidungsbäume</li> </ul>
Vorhersage	Prognose der Werte einer abhängigen kontinuierlichen Variable auf Basis einer funktionalen Beziehung	<ul style="list-style-type: none"> <li>■ Regressionsanalyse</li> <li>■ Neuronale Netze</li> <li>■ Entscheidungsbäume</li> </ul>
Assoziation	Aufdeckung von strukturellen Zusammenhängen in Datenbasen mit Hilfe von Regeln	<ul style="list-style-type: none"> <li>■ Assoziationsanalyse</li> </ul>

Tabelle 1: Anwendungsbereiche und Methoden des Data Mining

Abhängig vom jeweiligen Anwendungsbereich des Data Mining kann eine Vielzahl verschiedener Methoden zum Einsatz kommen, die in der Tabelle 1 zusammenfassend dargestellt sind. Neben den bereits seit langem aus der Multivariaten Datenanalyse bekannten Verfahren der Cluster-, Diskriminanz- und Regressionsanalyse sind für das Data Mining vor allem die Neuronalen Netze, die Entscheidungsbäume sowie die Assoziationsanalyse von Bedeutung.<sup>4</sup> Bei diesen letztgenannten Methoden steht der für das Data Mining typische experimentelle Charakter im Vordergrund.

<sup>4</sup> vgl. *Schinzer et al. (1999, S. 107-108)*

Die Neuronale Netze und die Entscheidungsbäume werden in der einschlägigen Literatur zwar schon seit längerem behandelt, haben mit dem Data Mining aber ein prädestiniertes Einsatzfeld gefunden und sind im Rahmen dieses Ansatzes entsprechend weiterentwickelt worden. Auch die Clusteranalyse hat mit dem Data Mining eine Weiterentwicklung erfahren. Neben den klassischen Verfahren gibt es mittlerweile eine Reihe von Varianten und neueren Algorithmen, die speziell auf den Ansatz des Data Mining zugeschnitten sind. Im Folgenden soll auf diese drei Methodenbereiche sowie auf die ebenfalls speziell für das Data Mining konzipierte Assoziationsanalyse näher eingegangen werden.

## 2 Ausgewählte Methoden des Data Mining

### 2.1 Clusteranalyse

Das Ziel einer Clusteranalyse besteht darin, die Elemente einer Menge derart zu gruppieren, daß die Elemente innerhalb der gebildeten Klassen möglichst ähnlich und zwischen den Klassen möglichst verschieden sind. Ausgehend von einer Datenmatrix, in der zeilenweise die Objekte und spaltenweise die Merkmale enthalten sind, kommen als zu klassifizierende Elemente die Objekte und/oder die Merkmale in Betracht. Sollen lediglich Objekte oder Merkmale in Klassen zusammengefaßt werden, spricht man von einmodaler Clusteranalyse. Im Fall einer gleichzeitigen Segmentierung von Objekten und Merkmalen liegt eine zweimodale Clusteranalyse vor. Da für entsprechende Fragestellungen das Data Mining vor allem einmodale Segmentierungsverfahren, mit denen die Objekte einer Datenbasis gruppiert werden sollen, von Bedeutung sind, wird im Folgenden auf diese Ansätze ausführlicher eingegangen.<sup>5</sup>

Die traditionellen Clusteranalyseverfahren lassen sich im Wesentlichen in hierarchische und partitionierende Verfahren unterscheiden. Während bei den hierarchischen Ansätzen durch die Vereinigung von zwei Clustern oder die Aufspaltung eines Clusters schrittweise immer größere bzw. feinere Segmentierungen erzeugt werden, wird bei den partitionierenden Verfahren eine vorliegende Zerlegung der Objektmenge durch den Austausch von Objekten iterativ verbessert. Für das Data Mining eignen sich die hierarchischen Clusteranalyseverfahren weniger, da Hierarchien mit zunehmender Objektzahl exponentiell anwachsen.<sup>6</sup> Des Weiteren weisen die in Hierarchien enthaltenen detaillierten Informationen über den Fusionsprozeß und die Entwicklung von Klassenstrukturen einen vergleichsweise geringen Nutzen für die Segmentierungen großer Objektmengen auf, da sie nicht mehr sinnvoll interpretiert werden können und dies im Rahmen des Data Mining Ansatzes auch nicht von primärem Interesse ist. Letztendlich soll am Ende eine interpretierbare und zur Lösung der Problemstellung geeignete Segmentierung der Objekte vorliegen, ohne die in Hierarchien zwangsläufig enthaltene Entstehung dieser Segmente schrittweise nachvollziehen zu wollen. Aus diesen Gründen sind für das Data Mining vor allem die partitionierenden Verfahren von Bedeutung. Dazu zählen beispielsweise das K-Means-Verfahren für quantitative Daten sowie das auf Distanzen basierende CLUDIA-Verfahren für beliebige Daten.<sup>7</sup> Die mittlerweile zahlreich existierenden Vari-

---

<sup>5</sup> Darstellungen zu zweimodalen Verfahren der Clusteranalyse können beispielsweise *Eckes und Orlik* (1991) oder *Espejo und Gaul* (1986) entnommen werden.

<sup>6</sup> vgl. *Küsters* (2001, S. 113)

<sup>7</sup> vgl. z.B. *Bausch und Opitz* (1993, S. 57-58)

ten partitionierender Clusteranalysealgorithmen unterscheiden sich vor allem hinsichtlich der gewählten Bewertungskriterien für die resultierende Segmentierung, der Bestimmung einer initialen Clustereinteilung sowie der Art und Weise des Objekttauschs im Rahmen der einzelnen Iterationen.<sup>8</sup>

Bei den bislang behandelten Ansätzen der Clusteranalyse wird ein Objekt eindeutig einer Klasse zugeordnet. In realen Anwendungen ist dies jedoch durchaus problematisch, da vor allem Ausreißer eine Segmentierung erschweren und zu relativ inhomogenen Clustern führen können. Zu Lösung dieses Problems sind Verfahren entwickelt worden, die unter dem Begriff der unscharfen Clusteranalyse zusammengefaßt werden und für die einzelnen Objekte Zugehörigkeitsgrade zu den Klassen bestimmen. Auf Basis dieser Zugehörigkeitsgrade können dann Kern-, Rand- oder Zwischenobjekte von Klassen entdeckt werden, so daß im Ergebnis eine deutliche differenziertere Segmentierung der Objekte vorliegt.<sup>9</sup>

Abschließend sollen noch kurz zwei weitere Ansätze angesprochen werden, die in der Literatur zwar oft im Zusammenhang mit der Clusteranalyse im Rahmen des Data Mining erwähnt werden, jedoch gemäß der Systematisierung in dieser Arbeit als eigenständige Methoden zur Segmentierung zu betrachten sind. Zum einen sind dies Neuronale Netze in Form der so genannten Self-Organizing Feature Maps (SOM), auf die im nachfolgenden Abschnitt noch eingegangen wird. Zum anderen finden sich in der neueren Literatur vermehrt Beiträge zur Anwendung evolutionärer Algorithmen in der Clusteranalyse. Auf diesen Methodenbereich wird im Rahmen dieser Arbeit nicht weiter eingegangen.<sup>10</sup>

## 2.1 Neuronale Netze

Zur grundlegenden Darstellung des Ansatzes der Neuronalen Netze wird im Folgenden vom Beispiel einer Kreditwürdigkeitsprüfung ausgegangen. Besteht das Ziel einer derartigen Analyse in der Identifizierung der kreditwürdigen und nicht kreditwürdigen Kunden, dann liegt in diesem Fall die Problemstellung der Klassifikation vor. Dabei stellen alle als grundsätzlich relevant erachteten und gegebenenfalls codierten Kundendaten die Eingangsinformationen für das Neuronale Netz dar, auf deren Basis eine Einteilung hinsichtlich der Kreditwürdigkeit erfolgen soll. Wie der nachfolgenden Abbildung 2 entnommen werden kann, ist das Grundelement eines Neuronalen Netzes ein Verarbeitungselement, das mehrere gewichtete Eingänge, eine Transformationsfunktion sowie einen Ausgang besitzt.

Bei der Informationsverarbeitung in einem Neuron  $j$  werden die Ausgabewerte der vorgelagerten Neuronen zunächst zu einem Nettoeingangssignal  $net_j$  verarbeitet, wobei hier meist eine gewichtete Summe verwendet wird. Im zweiten Schritt erzeugt eine Aktivierungsfunktion  $f_{akt}$  aus dem Nettoeingangssignal den Aktivierungszustand  $a_j$  des Neurons, der diskrete oder in einem Intervall definierte stetige Werte annehmen kann. Als typische Vertreter von Aktivierungsfunktionen können Schwellenwert-, Rampen- oder logistische Funktionen genannt wer-

---

<sup>8</sup> Eine umfassende Darstellung dazu kann z.B. *Rudolph* (1999) entnommen werden.

<sup>9</sup> Hinsichtlich einer ausführlichen Darstellung zur Fuzzy-Clusteranalyse sei exemplarisch auf *Höppner et al.* (1997) verwiesen.

<sup>10</sup> Eine Darstellung zu dieser Thematik kann beispielsweise *Freitas* (2002, S. 165-178) entnommen werden.

den. Im letzten Schritt wird dann der Aktivierungszustand in einen Ausgabewert  $z_j$  überführt, wobei als Ausgabefunktion  $f_{out}$  meist die Identitätsfunktion herangezogen wird.

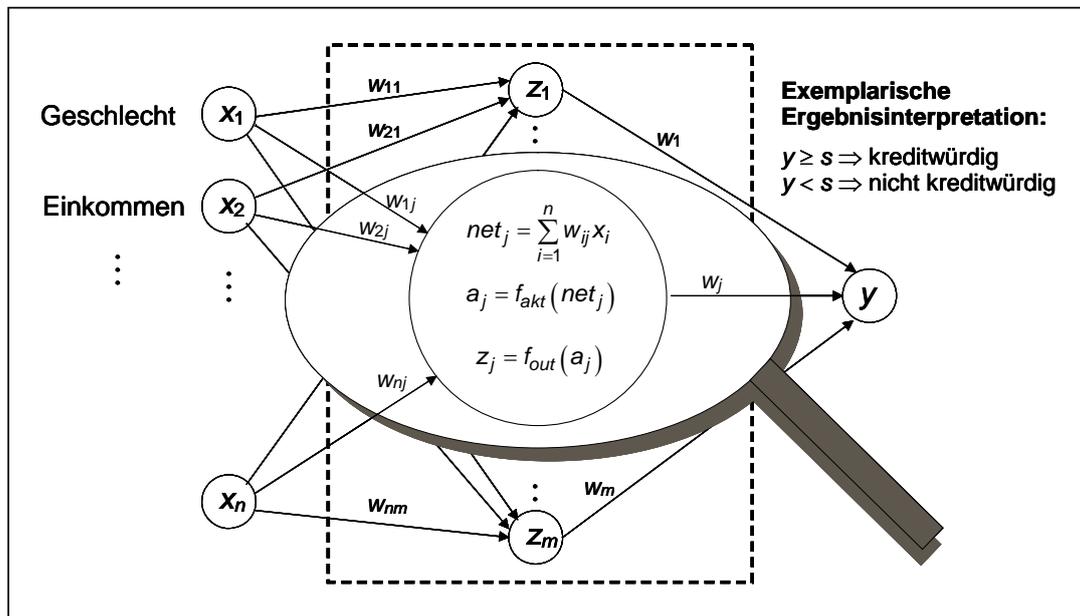


Abbildung 2: Grundstruktur eines Neuronen Netzes

Durch die Verbindung mehrerer Verarbeitungselemente auf verschiedenen Ebenen entsteht dann ein Netz. Ein derartiges Netz besteht zumindest aus einer Eingabeschicht sowie einer Ausgabeschicht. Darüber hinaus sind beliebig viele Zwischenschichten denkbar, mit denen die Repräsentationsfähigkeit des Netzes enorm erhöht werden kann. Für reale betriebswirtschaftliche Anwendungen reichen jedoch im Allgemeinen Netze mit höchstens ein bis zwei Zwischenschichten aus. Neben dem in der Abbildung 2 dargestellten Typ eines Feedforward-Netzes sind auch Feedback-Netze bzw. rekurrente Netztypen denkbar, bei denen Rückkopplungen zu vorgelagerten Schichten zugelassen sind.<sup>11</sup>

Ausgehend von einer festgelegten Netzstruktur müssen schließlich in der Lernphase die Verknüpfungen geeignet angepaßt werden. Dabei unterscheidet man im Wesentlichen zwischen überwachtem und unüberwachtem Lernen. Beim überwachten Lernen werden die Gewichte so angepaßt, daß bei einem vorgegebenen Eingangssignal ein bekanntes Ergebnis eintritt. Diese Vorgehensweise kann im Rahmen des Data Mining für die Problemstellungen der Klassifikation sowie der Prognose zur Anwendung kommen. Zur Schätzung der Modellparameter in Abhängigkeit vom vorliegenden Netztyp existieren zahlreiche Techniken, die meist spezielle Varianten von Gradientenverfahren darstellen.<sup>12</sup>

<sup>11</sup> vgl. z.B. *Rehkugler und Kerling* (1995, S. 314); ausführliche Darstellungen zu der Vielzahl möglicher Netzwerkarchitekturen können beispielsweise den Arbeiten von *Bigus* (1996, S. 65-67) sowie *Podding und Sidrovitch* (2001, S. 372-375) entnommen werden.

<sup>12</sup> Eine umfassende Behandlung dieser Ansätze sowie neuere Entwicklungen wie die Radialen Basisfunktionen (RBF) finden sich in den Arbeiten von *Neuneier und Zimmermann* (1998) sowie *Neuneier und Tresp* (1994).

Im Gegensatz zum überwachten Lernen liegen beim unüberwachten Lernen keine Soll-Ausgabewerte vor, so daß die Strukturen selbständig entdeckt werden müssen. Für derartige Problemstellungen, wie sie im Data Mining im Fall einer Segmentierung gegeben sind, können dann zum Beispiel die von Kohonen entwickelten Self-Organizing Feature Maps herangezogen werden.<sup>13</sup> Es bleibt noch festzuhalten, daß im Anschluß an die Trainingsphase des Netzes eine Testphase mit neuen Daten durchgeführt werden sollte, um die Leistungsfähigkeit des Netzes überprüfen zu können.<sup>14</sup>

Für das eingangs angesprochene Beispiel einer Kreditwürdigkeitsprüfung wird mit einem auf Basis von Kundendaten trainierten und auch getesteten Netz ein Klassifikator gebildet, mit dem eine Einteilung in die kreditwürdigen und nicht kreditwürdigen Kunden möglich ist. Für einen zu klassifizierenden Kunden wird dabei ein Ausgabewert  $y$  ermittelt, mit dem durch einen Vergleich mit dem Schwellenwert  $s$  eine entsprechende Einteilung erfolgen kann (vgl. Abbildung 2).

### 2.3 Entscheidungsbäume

Neben den im vorherigen Abschnitt vorgestellten Neuronalen Netzen könnte für die Problemstellung einer klassifikatorischen Kreditwürdigkeitsprüfung auch der Ansatz der Entscheidungsbäume herangezogen werden. Darauf soll im Folgenden näher eingegangen werden, wobei sich die folgenden Darstellungen zu den Entscheidungsbäumen aus Gründen der Übersichtlichkeit zunächst auf diskrete Bäume mit ausschließlich binären Variablen beschränken.

Im Rahmen der Konstruktion eines Entscheidungsbaums erfolgt im Prinzip eine automatische Auswahl von geeigneten Merkmalen zur Aufteilung der Datensätze im Hinblick auf die vorliegende Zielgröße. Für das in der Abbildung 3 betrachtete Beispiel der Kreditwürdigkeitsprüfung stellt damit die Kreditwürdigkeit mit den beiden Ausprägungen „kreditwürdig“ und „nicht kreditwürdig“ die zu identifizierende Zielvariable  $y$  dar. Des Weiteren sind mit den Kundenmerkmalen die unabhängigen Variablen  $x_1, \dots, x_n$  gegeben, die zusammen mit der Zielvariablen die Ausgangs- bzw. Trainingsdaten zur Konstruktion des Entscheidungsbaums bilden.

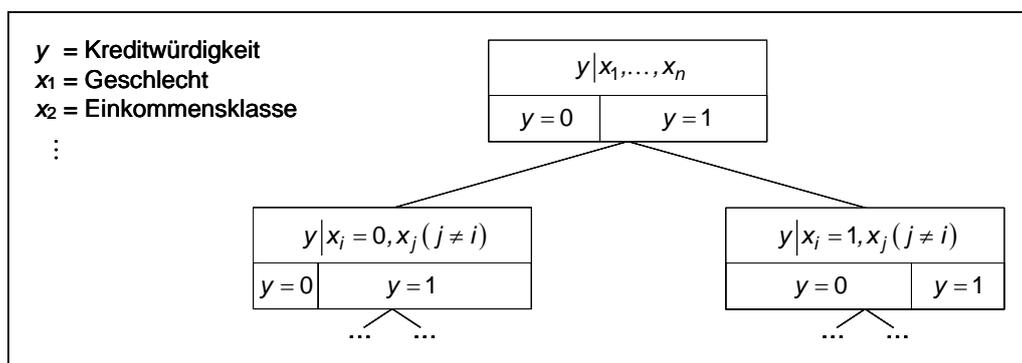


Abbildung 3: Binärer Entscheidungsbaum

<sup>13</sup> vgl. z.B. Kohonen (1995)

<sup>14</sup> vgl. z.B. Lackes und Mack (2000, S. 138-157)

Ausgehend von der zunächst vorliegenden Gruppierung der Datensätze in die beiden Klassen der kreditwürdigen ( $y = 1$ ) und nicht kreditwürdigen ( $y = 0$ ) Personen wird im ersten Schritt eine Variable  $x_i$  gesucht, so daß die resultierende Unterteilung der Daten bestmöglich nach der Klassenzugehörigkeit diskriminiert. Dies bedeutet, daß die resultierende Unterteilung bezüglich der Zielvariablen maximal homogen sein soll. Als Homogenitätsmaße sind eine Reihe unterschiedlicher Maße denkbar, wobei vor allem das Chi-Quadrat-Maß, die Entropie, der Gini-Index sowie der Twoing-Wert zu nennen sind.<sup>15</sup> Das Chi-Quadrat-Maß basiert auf einer entsprechenden Kontingenztafel der Ziel- und der Unterteilungsvariablen und vergleicht die gemeinsame Verteilung mit der Randverteilung. Im Fall einer maximalen Abweichung zwischen der tatsächlichen und der unter der Unabhängigkeitsprämisse erwarteten Verteilung liegt eine maximale Abhängigkeit zwischen den beiden Variablen und damit eine maximale Homogenität innerhalb der aufgeteilten Daten vor.

Nachdem mit der ersten Unterteilungsvariablen eine bestmögliche Diskriminierung nach der Klassenzugehörigkeit durchgeführt worden ist, wird im nächsten Schritt eine darauf aufbauende weitere Unterteilung der Daten gesucht, die wiederum zu einer maximalen Homogenität führt. Dies ist in der Abbildung 3 mit den Punkten angedeutet. Auf eine weitere Unterteilung kann auch verzichtet werden, wenn zum Beispiel die Werte des Homogenitätsmaßes zu gering werden. Des Weiteren ist unter Umständen gar keine bessere Aufteilung mehr möglich. Dies ist beispielsweise der Fall, wenn alle Datensätze eines Knotens bereits derselben Klassen angehören oder denselben Merkmalsvektor aufweisen. Alle diese Abbruchkriterien dienen dazu, ein unnötiges Anwachsen des Baumes bereits bei der Konstruktion zu verhindern. Aus diesem Grund wird in diesem Zusammenhang auch von Pre-Pruning-Techniken gesprochen. Demgegenüber können auch Post-Pruning-Techniken zur Anwendung kommen, die einen bereits fertig konstruierten Baum durch ein nachträgliches Zurückschneiden wieder verkleinern. Damit soll vor allem erreicht werden, daß die Fehlklassifikationsrate bei neuen Daten, die aufgrund der extremen Anpassung des Baumes an die Trainingsdaten unter Umständen sehr hoch sein kann (Problem des Overfitting), wieder reduziert wird.<sup>16</sup>

Neben den hier kurz vorgestellten diskreten Entscheidungsbäumen sind auch stetige Entscheidungsbäume denkbar, bei denen anstelle einer nominalen eine quantitative Variable identifiziert wird, um auf Basis der Ergebnisse dann vor allem eine Prognose durchführen zu können. Dieser Ansatz basiert auf dem Grundgedanken, daß innerhalb der Klassen, die durch eine Unterteilungsvariable vorgegeben werden, eine möglichst geringe Variation der Zielvariablen vorliegen sollte.<sup>17</sup> Stetige Entscheidungsbäume weisen allerdings im Vergleich zum diskreten Ansatz eine deutlich geringere Anwendungsrelevanz in entsprechenden betriebswirtschaftlichen Problemstellungen auf.<sup>18</sup> Dies mag auch daran liegen, daß stetige Entscheidungsbäume im Wesentlichen für Intervallprognosen geeignet sind und mit der Regressions-

---

<sup>15</sup> vgl. z.B. *Säuberlich* (2000, S. 88-94)

<sup>16</sup> Ausführliche Darstellungen dazu können beispielsweise den Arbeiten von *Quinlan* (1986, 1993), *Breiman et al.* (1984) sowie *Säuberlich* (2000, S. 96-103) entnommen werden.

<sup>17</sup> vgl. ausführlich z.B. *Schlittgen* (1998)

<sup>18</sup> vgl. *Küsters* (2001, S. 110)

analyse sowie den Neuronalen Netzen weitere Verfahren existieren, die für Prognosezwecke die unter Umständen besseren Alternativen darstellen.

## 2.4 Assoziationsanalyse

Ausgangspunkt für den Ansatz der Assoziationsanalyse ist eine Menge  $I = \{i_1, \dots, i_m\}$ , die diskrete Größen oder so genannte Items enthält. Dies können beispielsweise die Artikel eines Supermarkts sein. Des Weiteren bezeichnet  $T$  eine Transaktion, die eine Teilmenge der Itemmenge  $I$  darstellt. Eine Transaktion kann somit beispielsweise als Kaufaktion in Form einer speziellen Warenzusammenstellung eines Kunden gesehen werden. Werden alle vorliegenden Transaktionen, d.h. in dem hier betrachteten Beispiel alle Kaufaktionen eines gegebenen Zeitraums zusammengefaßt, ergibt dies die Datenbasis  $D = (T_1, \dots, T_n)$  mit  $T_j \subseteq I$ . Eine Assoziationsregel stellt dann eine Regel der Form „wenn  $X$ , dann  $Y$ “ ( $X \rightarrow Y$ ) dar, wobei die Menge  $X$  im Regelrumpf und die Menge  $Y$  im Regelkopf disjunkt und echte Teilmengen der Itemmenge  $I$  sind.<sup>19</sup>

Zur Bewertung einer Assoziationsregel werden zumindest zwei Maße in Form des so genannten Supports sowie der Confidence benötigt.<sup>20</sup> Der Support einer Itemmenge  $X \subset I$  ist definiert als die relative Häufigkeit dieser Itemmenge in der Datenbasis, d.h.

$$\text{Support}(X) = \frac{|T \in D : X \subseteq T|}{|D|}.$$

Damit wird zum Ausdruck gebracht, ob eine Kombination von Items zur Bildung einer Assoziationsregel überhaupt von Bedeutung ist. Das Ausmaß der Gültigkeit einer Regel  $X \rightarrow Y$  wird dann durch die Confidence angegeben. Dieser Wert ergibt sich gemäß

$$\text{Confidence}(X, Y) = \frac{|T \in D : (X \cup Y) \subseteq T|}{|T \in D : X \subseteq T|}.$$

Dabei wird ausgehend von den Transaktionen, die eine Itemmenge  $X$  enthalten, der relative Anteil an diesen Transaktionen bestimmt, die auch die Menge  $Y$  enthalten. Im Prinzip wird damit der Support der Vereinigungsmenge der Mengen  $X$  und  $Y$  ins Verhältnis zum Support der Menge  $X$  gesetzt.

Der Support und die Confidence einer Assoziationsregel sollen im Folgenden anhand eines sehr einfachen Zahlenbeispiels erläutert werden (vgl. Abbildung 4). Ausgehend von insgesamt 1.000.000 Einkaufstransaktionen wurden 200.000-mal Chips und 50.000-mal Bier gekauft, wobei 20.000-mal beide Produkte gemeinsam gekauft worden sind. Anhand des Venn-Diagramms rechts oben in der Abbildung 4 wird deutlich, wie sich der Support für die Kombination der beiden Artikel ergibt. Der ganz dunkel schraffierte Bereich, der den Einkäufen entspricht, in denen Chips und Bier gleichzeitig enthalten waren, wird dabei ins Verhältnis zur Gesamtzahl aller vorliegenden Einkäufe gesetzt. Dabei ergibt sich in diesem Beispiel ein prozentualer Anteil von 2 Prozent. Betrachtet man nun als mögliche Assoziationsregel die Regel

<sup>19</sup> vgl. z.B. *Säuberlich* (2000, S. 109-110)

<sup>20</sup> Eine Darstellung weiterer Bewertungsmaße (Lift bzw. Interest, Factor, etc.) kann beispielsweise der Arbeit von *Hettich* und *Hippner* (2001, S. 445-449) entnommen werden.

„Wenn Chips gekauft werden, dann wird auch Bier gekauft“, so ergibt sich eine Confidence von 10 Prozent. Anhand des entsprechenden Venndiagramms in der Abbildung 4 wird ersichtlich, daß die Anzahl der Kombinationseinkäufe von Chips und Bier jetzt ins Verhältnis zur Gesamtanzahl der Chips-Einkäufe gesetzt wird. Im Fall einer umgekehrten Assoziationsregel resultiert entsprechend eine Confidence von 40 Prozent, da jetzt die Gesamtanzahl der Biereinkäufe die Bezugsgröße darstellt. Damit würde in diesem Beispiel die zweite dargestellte Regel ein Muster in den Daten aufzeigen, das der Supermarkt exemplarisch wie folgt nutzen könnte: Durch die Anordnung der Chips in Sichtweite des Biers könnte gegebenenfalls eine Zunahme dieser Verbundkäufe erreicht werden. Des Weiteren sollte nicht bei beiden Produkten gleichzeitig ein Sonderangebot gemacht werden, sondern – sofern derartige Angebote überhaupt geplant sind – idealerweise nur beim Bier. Diese Angebote könnten dann auch zu einer absoluten Zunahme der Chips-Verkäufe führen.

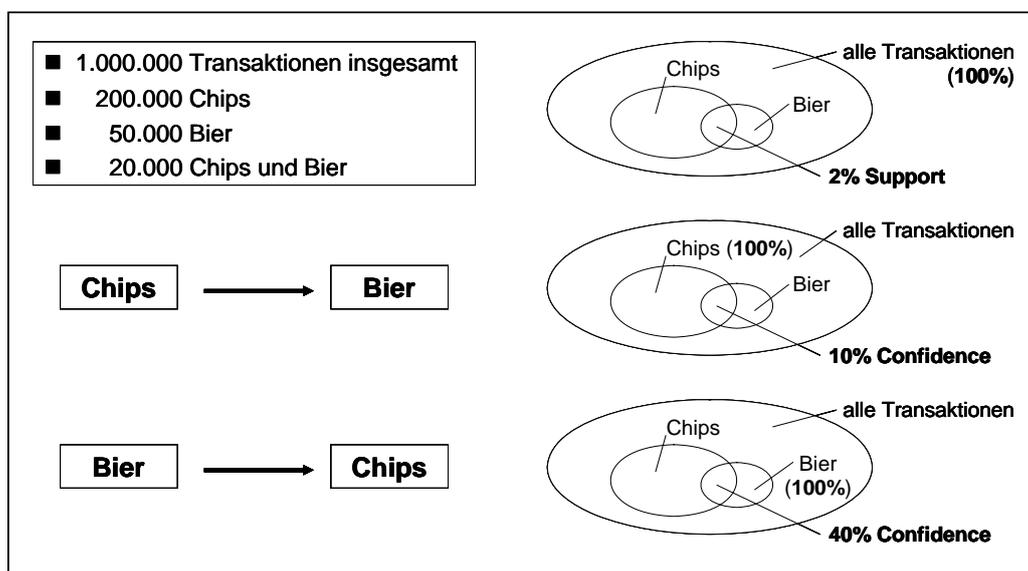


Abbildung 4: Beispiel zur Assoziationsanalyse

Um ausgehend von einer gegebenen Datenbasis alle bedeutsamen Assoziationsregeln generieren zu können, werden zunächst die so genannten häufigen Itemmengen einer Datenbasis bestimmt. Dies sind alle Itemmengen, deren Support größer oder gleich dem vorgegebenen Mindest-Support ist. Da die Anzahl aller möglichen Kombinationen von Items der Ausgangsitemmenge  $I$  im Allgemeinen sehr groß ist, scheidet eine enumerative Problemlösung aus. Die Aufgabe der Assoziationsregelalgorithmen besteht daher darin, möglichst effizient alle häufigen Itemmengen zu finden, ohne jedoch alle grundsätzlich möglichen Itemmengen auf ihren Support hin untersuchen zu müssen. Sind schließlich alle häufigen Itemmengen gefunden, werden im zweiten Schritt aus jeder häufigen Itemmenge alle möglichen Regeln generiert, deren Confidence größer gleich dem geforderten Mindestwert ist.

Der AIS-Algorithmus stellt den ersten in der Literatur beschriebenen Algorithmus zur Generierung häufiger Itemmengen dar.<sup>21</sup> Ausgehend von allen einelementigen häufigen Itemmengen

<sup>21</sup> vgl. Agrawal et al. (1993)

werden mehrelementige Kandidaten für häufige Itemmengen dadurch schrittweise gewonnen, daß auf Basis der vorliegenden Transaktion zu den bereits erkannten häufigen Itemmengen einzelne Items hinzugefügt werden. Aus diesen Kandidaten können dann durch einen Vergleich mit dem Mindest-Support diejenigen Itemmengen bestimmt werden, die auch häufig sind. Einen im Prinzip ähnlichen Verfahrensablauf weist der Apriori-Algorithmus auf, der aber eine Weiterentwicklung des AIS-Algorithmus darstellt. Der Apriori-Algorithmus verwendet zur Gewinnung von Kandidaten häufiger Itemmengen eine spezielle vorgeschaltete Operation, mit der die häufigen Itemmengen im Allgemeinen deutlich schneller gefunden werden.<sup>22</sup> Die grundlegende Idee dabei ist, daß eine Itemmenge nur dann häufig sein kann, wenn auch alle möglichen Teilmengen dieser Itemmenge, die jeweils ein Element weniger enthalten, häufig sind. Dadurch sinkt der Arbeitsaufwand zur Gewinnung potentieller häufiger Itemmengen deutlich, was in realen Problemstellungen zu einer erheblichen Reduzierung des Rechenaufwands und damit der Laufzeit führen kann.<sup>23</sup>

Während Assoziationsregeln normalerweise die Abhängigkeiten zwischen dem Auftreten von Items untersuchen, werden mit den so genannten „dissociation rules“ oder auch „negative association rules“ Beziehungen zwischen dem Fehlen und dem Vorhandensein von Items analysiert.<sup>24</sup> Für reale Problemstellungen bedeutet dies, daß neben Verbundkäufen damit auch Substitutionskäufe aufgedeckt werden können. Abschließend bleibt noch festzuhalten, daß neben der effizienten Generierung häufiger Itemmengen auch die effiziente Regelkonstruktion einen entscheidenden Einfluß auf die Effizienz bei der Generierung von Assoziationsregeln hat.<sup>25</sup>

### 3 Betriebswirtschaftliche Relevanz des Data Mining

#### 3.1 Wissenschaftstheoretische Ansatzpunkte

Wissenschaftstheoretische Veröffentlichungen gehen üblicherweise der Frage nach, wie man von einer Theorie zu wissenschaftlichen Erkenntnissen gelangt. Für die formalen Wissenschaften stellt dieser Schritt im Allgemeinen kein Problem dar, da aus Axiomen und Gesetzen entsprechende Aussagen mittels Deduktion, d.h. logischem Schließen abgeleitet werden können. In den Realwissenschaften, zu denen auch die Betriebswirtschaftslehre zählt, ist diese deduktive Vorgehensweise jedoch meist nicht möglich. Daher sind in der wissenschaftstheoretischen Literatur eine Vielzahl entsprechender Denkansätze entwickelt und diskutiert worden, wovon drei der nachfolgenden Abbildung 5 entnommen werden können.

Auf eine ausführlichere Darstellung der einzelnen wissenschaftstheoretischen Denkansätze wird an dieser Stelle verzichtet, da lediglich verdeutlicht werden soll, daß sich das Augenmerk in der wissenschaftstheoretischen Diskussion im Wesentlichen auf den Weg von der Theorie zur wissenschaftlichen Erkenntnis beschränkt. Damit bleibt aber die Frage, wie man

---

<sup>22</sup> vgl. *Agrawal und Srikant* (1994)

<sup>23</sup> Eine ausführliche Darstellung dieser und weiterer Assoziationsregelalgorithmen kann beispielsweise *Hettich und Hippner* (2001, S. 430-438) sowie *Säuberlich* (2000, S. 112-123) entnommen werden. Ansätze, bei denen genetische Algorithmen im Rahmen der Assoziationsanalyse zur Anwendung kommen, finden sich z.B. in der Arbeit von *Freitas* (2002, S. 107-163)

<sup>24</sup> vgl. *Berry und Linoff* (1997, S. 148-149) sowie *Zhang und Zhang* (2002, S. 47-84)

<sup>25</sup> Einige Ansatzpunkte dazu können z.B. der Arbeit von *Bankhofer* (2002, S. 46) entnommen werden.

überhaupt zu einer Theorie gelangt, unbeantwortet. Neben Inspiration, Wissenstransfer und zufälligen Ereignissen sind in diesem Zusammenhang vor allem Daten und Beobachtungsaussagen zu nennen, aus denen mittels Induktion Gesetzmäßigkeiten abgeleitet werden können (vgl. Abbildung 5). Hier leistet die explorative Datenanalyse und im Zeitalter zunehmender Datenflut vor allem das Data Mining einen wertvollen Beitrag, Informationen aus vorliegenden Daten zur Theorieentwicklung zu gewinnen. Bei Entdeckung bestimmter Muster oder Strukturen in den Daten können Hypothesen aufgestellt werden, die dann mit Hilfe neuer Datensätze und konfirmatorischer Verfahren zu überprüfen sind. Darüber hinaus kann auch allein das Entdecken von Strukturen bereits ökonomisch wertvoll sein.

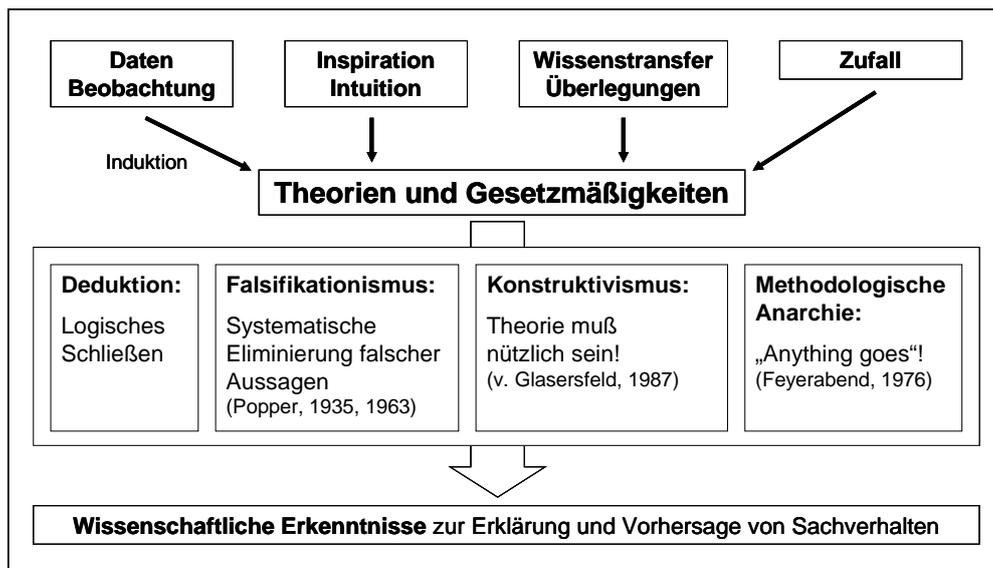


Abbildung 5: Wissenschaftstheoretischer Bezugsrahmen

### 3.2 Betrieblicher Bezugsrahmen

Im Folgenden soll auf die wesentlichen Einflußfaktoren eingegangen werden, die in den Unternehmen in verstärktem Maße zu einem Bedarf an einer weitgehend automatischen Informations- und Wissensgenerierung führen. Als erster zentraler Faktor sind in diesem Zusammenhang vor allem die drastisch wachsenden Datenbestände im betrieblichen Umfeld zu nennen. Beispielsweise speichert die amerikanische Warenhauskette Wal-Mart mit mehr als 2000 Geschäften weltweit ca. 20 Millionen Einkaufstransaktionen täglich. Da Wal-Mart zu einem der weltweit größten Einzelhandelsunternehmen zählt, mögen diese Dimensionen nicht verwundern. Dennoch können bereits bei mittelständischen Unternehmen Datenbestände vorliegen, die ein beachtliches Volumen aufweisen und damit den Wunsch nach einer automatischen Entdeckung von Mustern und Beziehungen in diesen Daten mit sich bringen.<sup>26</sup>

Die im betrieblichen Umfeld vorliegenden Daten stellen für ein Unternehmen im Prinzip Kapital dar, da die darin enthaltenen Informationen zur Steigerung des Profits oder auch anderer Zielgrößen genutzt werden können. Darüber hinaus zwingt der zunehmende Wettbewerbs-

<sup>26</sup> vgl. Küppers (1999, S. 13)

druck die Unternehmen, alle verfügbaren Informationsquellen zu nutzen. Neben den unternehmensinternen Daten sind in diesem Zusammenhang vor allem auch unternehmensexterne Daten zu nennen, auf die ein Unternehmen zurückgreifen muß. Aufgrund der zunehmenden Umweltkomplexität sind in der heutigen Zeit jedoch immer mehr Faktoren zu berücksichtigen, die für das Unternehmen entscheidungsrelevant sind.

Als weiterer Faktor, der die Entwicklung und Anwendung von Data Mining Verfahren im Unternehmen unterstützt, kann die Unzufriedenheit mit den vorhandenen Auswertungsmethoden genannt werden, da oft allein die Menge an Daten die Anwendung dieser traditionellen Ansätze verhindert. Schließlich muß in diesem Zusammenhang auch auf die kontinuierliche Verbesserung der Leistungsfähigkeit der Informationstechnik sowie die Entwicklung neuer Informationstechnologien hingewiesen werden, durch die eine Verarbeitung großer Datenmengen überhaupt erst möglich wird.

In der nachfolgenden Abbildung 6 sind die angesprochenen Faktoren, die im betrieblichen Umfeld als Promotoren für eine weitgehend automatische Informations- und Wissensgenerierung angesehen werden können, noch einmal zusammengefaßt.

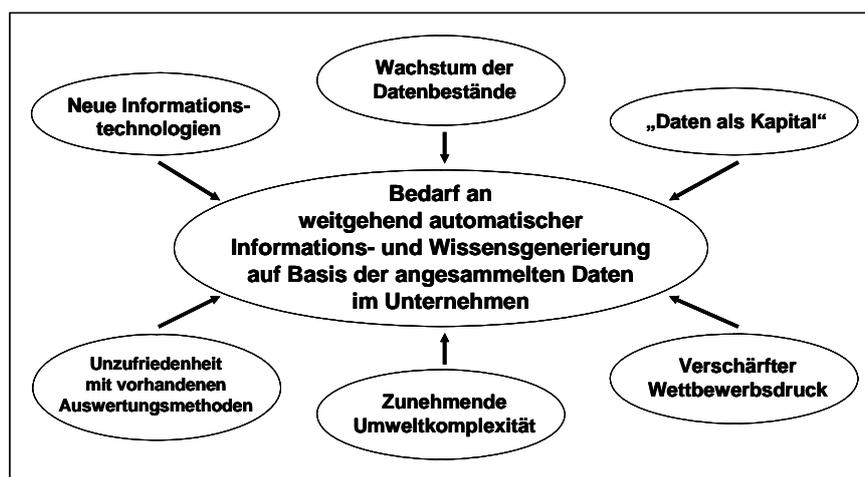


Abbildung 6: Promotoren des Data Mining im betrieblichen Umfeld

Trotz des dargestellten Bedarfs eines Einsatzes von Data Mining Verfahren im Unternehmen stellt sich für die betriebliche Praxis die Problematik, daß im Unternehmen die Fähigkeit zur sinnvollen Auswertung der Daten vorhanden sein muß. Ohne diese Data Mining Kompetenz sind die im Unternehmen vorhandenen Daten wertlos. Der Aufbau entsprechender Kompetenzen stellt mit Sicherheit eine zentrale Aufgabe dar, der sich die Unternehmen heute und in der Zukunft stellen müssen.

### 3.3 Betriebliche Anwendungen im Überblick

Wie in den beiden vorhergehenden Abschnitten aufgezeigt wurde, ergibt sich für die Unternehmen in der heutigen Zeit die Notwendigkeit, möglichst umfassend alle aus betrieblicher Sicht relevanten Beziehungen und Strukturen in ihren im Allgemeinen sehr großen Datenbeständen zu entdecken, um daraus Erkenntnisse zur Ableitung des richtigen unternehmerischen Handeln zu gewinnen. Aus dieser Notwendigkeit heraus haben sich mittlerweile zahlreiche

betriebliche Anwendungen und Einsatzgebiete des Data Mining herausgebildet, auf die im Folgenden kurz eingegangen werden soll.

Das dominierende betriebswirtschaftliche Einsatzgebiet des Data Mining stellt mit Sicherheit das Marketing dar. Neben der Kundensegmentierung und der damit verbundenen Individualisierung der Kundenansprache sind vor allem Anwendungen aus den Bereichen Preisfindung, Warenkorbanalysen sowie Storno- bzw. Kündigungsanalysen zu finden.<sup>27</sup> Aber auch neben dem Marketing sind in der Literatur zahlreiche betriebswirtschaftliche Anwendungen anzutreffen, die vor allem den Bereichen Beschaffung/Produktion, Controlling sowie Finanzdienstleistungen zugeordnet werden können.<sup>28</sup> Aufgrund der Vielzahl einzelner Anwendungen muß an dieser Stelle auf eine umfassende und detaillierte Beschreibung verzichtet werden, zumal dies auch den Rahmen der vorliegenden Arbeit sprengen würde. Statt dessen wird mit der nachfolgenden Abbildung 7 lediglich ein kurzer Überblick über einige grundlegende Anwendungen aus den systematisierten Bereichen mit entsprechenden Literaturverweisen gegeben, der allerdings keinerlei Anspruch auf Vollständigkeit erhebt. Um dennoch die möglichen Erfolgspotentiale des Data Mining in entsprechenden betrieblichen Problemstellungen zumindest ansatzweise anzudeuten, sollen im Folgenden exemplarisch drei dieser Anwendungen kurz vorgestellt werden, und zwar die Responseanalyse von Werbemitteln, die Qualitätssicherung und -kontrolle sowie die Entdeckung von Kreditkartenmißbrauchsversuchen.

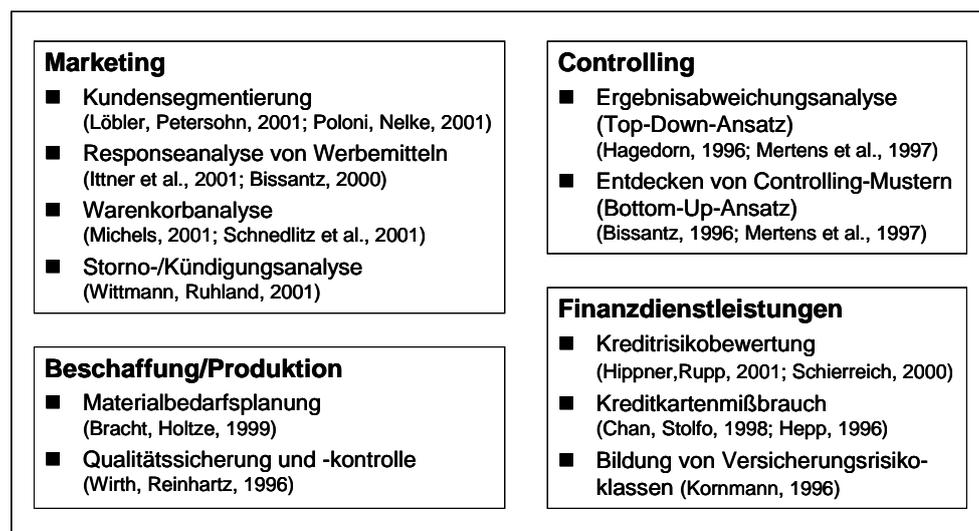


Abbildung 7: Betriebliche Anwendungen des Data Mining im Überblick

Die gezielte Versendung von Werbemitteln stellt vor allem für den Versandhandel ein wesentliches Instrumentarium dar, mit dem Kunden in Kontakt zu treten. Bei einer nicht individualisierten Kundenansprache würden bei Direktmailingaktionen alle Kunden das gleiche Mailing erhalten, wobei in diesem Fall von Responseraten von ca. 1 bis 2 Prozent ausgegangen werden kann. Aufgrund dieser geringen Effizienz besteht das Ziel einer Responseanalyse

<sup>27</sup> vgl. dazu ausführlich Hippner et al. (2001, S. 591-970) sowie Schinzer et al. (1999, S. 126-130)

<sup>28</sup> vgl. z.B. Berry und Linoff (2000, S. 253-483), Cabena et al. (1998, S. 25-38) sowie Küppers (1999, S. 134-149)

darin, durch eine Individualisierung der Mailings eine deutliche Verbesserung der Response-rate zu erreichen. Dazu werden auf der Basis einer Kundendatenbank diejenigen Kundengruppen identifiziert, die ein bestimmtes Konsumverhalten oder auf Basis eines Testmailings eine hohe Responserate aufweisen und damit für eine geplante Mailingaktion das größte Erfolgspotential besitzen. Durch die Beschränkung auf diese Kunden können dann bei weitgehender Ausschöpfung des gesamten Responsepotentials die Kosten für derartige Aktionen erheblich reduziert werden. Damit läßt sich der Gesamterfolg einer Mailingaktion deutlich steigern.

Das nächste Anwendungsbeispiel betrifft die Qualitätssicherung und -kontrolle eines namhaften Automobilherstellers. Das Unternehmen verfügt über eine Qualitätsdatenbank, in der die Daten zu den Ausstattungsmerkmalen sowie den in der Garantiezeit eingetretenen Schäden vergangener Produktionsperioden enthalten sind. Das Ziel der Analyse besteht in der Aufdeckung von Ausstattungsmerkmalen der produzierten Fahrzeuge, die in der Vergangenheit häufiger zu Garantiefällen geführt haben. Dazu wird die entsprechende Klassifikationsvariable „Schäden in der Garantiezeit“ mit Hilfe von Neuronalen Netzen oder Entscheidungsbäumen anhand der Ausstattungsmerkmale identifiziert. Auf Basis der Analyseresultate ist dann eine frühzeitige Problem- und Fehlererkennung möglich, die im Ergebnis zu einer Senkung der Kosten aus Garantiefällen führt. Des Weiteren können Ansätze zur Produktverbesserung abgeleitet und gegebenenfalls teure Rückrufaktionen vermieden werden, indem erwartete Störung im Rahmen der normalen Wartungsarbeiten rechtzeitig behoben werden. In diesem Fall würden die Kunden die eventuell vorliegenden Produktfehler erst gar nicht bemerken.

Als letztes Anwendungsbeispiel soll noch kurz auf die Entdeckung von Kreditkartenmißbrauchsversuchen eingegangen werden. Bei der Gesellschaft für Zahlungssysteme wurde beispielsweise ein System auf der Basis eines Neuronalen Netzes installiert, das bei der Autorisierung der Kreditkarte automatisch Mißbrauchsversuche aufdeckt, ohne die rechtmäßigen Transaktionen nennenswert zu beeinträchtigen. Dazu wird auf Basis des Karteninhaberprofils in Kombination mit weiteren Attributen wie Tageszeit oder typische Geschäfte ein Risikowert ermittelt, der der Wahrscheinlichkeit eines Mißbrauchsversuchs entspricht. Falls der Risikowert hoch ist, wird die Autorisierung erst nach telefonischer Rücksprache erteilt. In diesem Anwendungsbeispiel konnten die Kosten für die Systementwicklung und Installation in Höhe von mehreren Millionen Mark bereits nach neun Monaten eingespart werden.

#### **4. Zusammenfassung**

Das Data Mining ist ein wesentlicher Bestandteil im Prozeß des Knowledge Discovery in Data Bases. Durch die Anwendung geeigneter Methoden sollen dabei Muster und Beziehungen in großen Datenmengen entdeckt werden. Wie im Rahmen der vorliegenden Arbeit aufgezeigt wurde, besitzt das Data Mining ein breites Anwendungsspektrum und zum einsetzbaren Methodeninstrumentarium zählen neben den klassischen Verfahren der Multivariaten Datenanalyse vor allem die Neuronalen Netze und Entscheidungsbäume sowie der neu konzipierte Ansatz der Assoziationsanalyse.

Was die betriebswirtschaftliche Relevanz des Data Mining angeht, so erscheint diese offensichtlich zu sein. Zum einen kann das Data Mining zur Generierung von Hypothesen und Entwicklung von Theorien beitragen, obwohl gerade in betrieblichen Anwendungsfällen oft

allein das Entdecken von Strukturen bereits ökonomisch wertvoll ist. Zum anderen können die Methoden des Data Mining in einer Reihe betrieblicher Anwendungen zum Einsatz kommen und sind damit im Sinne einer Querschnittstechnologie über alle betrieblichen Funktionsbereiche, in denen große Datenmengen vorkommen, nutzbar.

## Literaturverzeichnis

- Agrawal, R.; Imielinski, T.; Swami, A.* (1993): Mining Associations between Sets of Items in Massive Data Bases, in: Proceedings of the ACM SIGMOD Int. Conference on Management of Data, Washington, S. 207-216
- Agrawal, R.; Srikant, R.* (1994): Fast Algorithms for Mining Association Rules, in: Proceedings of the 20th VLDB Conference, Santiago
- Bankhofer, U.* (2002): Data Mining mit Assoziationsregeln, in: Wirtschaftswissenschaftliches Studium, Heft 12, S. 43-47
- Bausch, T.; Opitz, O.* (1993): PC-gestützte Datenanalyse mit Fallstudien aus der Marktforschung, Vahlen, München
- Berry, M.J.A.; Linoff, G.* (1997): Data Mining Techniques, Wiley, New York
- Berry, M.J.A.; Linoff, G.* (2000): Mastering Data Mining, Wiley, New York
- Bigus, J.P.* (1996): Data Mining with Neural Networks, McGraw-Hill, New York
- Bissantz, N.* (1996): Clusmin – Ein Beitrag zur Analyse von Daten des Ergebniscontrollings mit Datenmustererkennung (Data Mining), in: *Bissantz, N.; Hagedorn, J.*, Data Mining im Controlling, Arbeitsbericht des Instituts für mathematische Maschinen und Datenverarbeitung, Band 29, Nr. 7, Universität Erlangen, Teil A
- Bracht, U.; Holtze T.* (1999): Data Mining für eine bessere Teilebedarfsprognose: Neue Ansätze zur Planung einer variantenreichen Serienfertigung hochkomplexer Produkte, in: ZWF (1999), Hanser, Heft 3, S. 119-122
- Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J.* (1984): Classification and Regression Trees, Wadsworth International, Belmont
- Cabena, P.; Hadjinian, P.; Stadler, R.; Verhees, J.; Zanasi, A.* (1998): Discovering Data Mining – From Concept to Implementation, Prentice Hall, Upper Saddle River
- Düsing, R.* (1999): Knowledge Discovery in Databases und Data Mining, in: Chamoni, P.; Gluchowski, P. (Hrsg.), Analytische Informationssysteme, 2. Auflage, Springer, Berlin, S. 345-353
- Chan, P.K.; Stolfo, S.J.* (1998): Toward Scalable Learning with Non-uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection, in: Proceedings of the Forth International Conference on Knowledge Discovery and Data Mining, AAAI Press, Menlo Park
- Eckes, T.; Orlik, P.* (1991): An Agglomerative Method for Two-Mode Hierarchical Clustering, in: *Bock, H.H.; Ihm, P.* (Hrsg.), Classification, Data Analysis and Knowledge Organisation, North-Holland, Amsterdam, S. 3-8
- Espejo, O.E.; Gaul, W.* (1986): Two-Mode Hierarchical Clustering as an Instrument for Marketing Research, in: *Gaul, W.; Schader, M.* (Hrsg.), Classification as a Tool of Research, North-Holland, Amsterdam, S. 121-128
- Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.* (1996): From Data Mining to Knowledge Discovery: An Overview, in: *Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R.* (Hrsg.), Advances in Knowledge Discovery and Data Mining, AAAI Press/MIT Press, Menlo Park, S. 1-34
- Feyerabend, P.* (1976): Wider den Methodenzwang, Frankfurt/Main
- Freitas, A.A.* (2002): Data Mining and Knowledge Discovery with Evolutionary Algorithms, Springer, Berlin
- Glasersfeld, E.v.* (1987): Wissen, Sprache und Wirklichkeit. Arbeiten zum radikalen Konstruktivismus, Braunschweig
- Hagedorn, J.* (1996): Die automatische Filterung von Controlling-Daten unter besonderer Berücksichtigung der Top-Down-Navigation, in: *Bissantz, N.; Hagedorn, J.*, Data Mining im Controlling, Arbeitsbericht des Instituts für mathematische Maschinen und Datenverarbeitung, Band 29, Nr. 7, Universität Erlangen
- Hepp, H.-M.* (1996): Vom Neuronalen Netz zum Data Mining, Data Mining Konferenz, Frankfurt/Main
- Hettich, S.; Hippner, H.* (2001): Assoziationsanalyse, in: *Hippner, H.; Ulrich, K.; Meyer, M.; Wilde, K.* (Hrsg.), Handbuch Data Mining im Marketing, Vieweg, S. 427-463
- Hippner, H.; Ulrich, K.; Meyer, M.; Wilde, K.* (Hrsg.) (2001): Handbuch Data Mining im Marketing, Vieweg
- Hippner, H.; Rupp, A.* (2001): Kreditwürdigkeitsprüfung im Versandhandel, in: *Hippner, H.; Ulrich, K.; Meyer, M.; Wilde, K.* (Hrsg.), Handbuch Data Mining im Marketing, Vieweg, S. 685-706

- Höppner, F.; Klawonn, F.; Kruse, R. (1997): Fuzzy-Clusteranalyse: Verfahren für die Bildererkennung, Klassifizierung und Datenanalyse, Vieweg, Braunschweig
- Ittner, A.; Sieber, H.; Trautzsch, S. (2001): Nichtlineare Entscheidungsbäume zur Optimierung von Direktmailingaktionen, in: Hippner, H.; Ulrich, K.; Meyer, M.; Wilde, K. (Hrsg.), Handbuch Data Mining im Marketing, Vieweg, S. 707-723
- Kohonen, T. (1995): Self-Organizing Maps, Springer, Berlin
- Kornmann, K. (1996): Entwicklung eines Frühwarnsystems, Data Mining Konferenz, Frankfurt/Main
- Küppers, B. (1999): Data Mining in der Praxis, Lang, Frankfurt
- Küsters, U. (2001): Data Mining Methoden: Einordnung und Überblick, in: Hippner, H.; Ulrich, K.; Meyer, M.; Wilde, K. (Hrsg.), Handbuch Data Mining im Marketing, Vieweg, S. 95-130
- Lackes, R., Mack, D. (2000): Neuronale Netze in der Unternehmensplanung, Vahlen, München
- Löbler, H.; Petersohn, H. (2001): Kundensegmentierung im Automobilhandel zur Verbesserung der Marktbearbeitung, in: Hippner, H.; Ulrich, K.; Meyer, M.; Wilde, K. (Hrsg.), Handbuch Data Mining im Marketing, Vieweg, S. 623-641
- Lusti, M. (2002): Data Warehousing und Data Mining, 2. Auflage, Springer, Berlin
- Mertens, P.; Bissantz, N.; Hagedorn, J. (1997): Data Mining im Controlling, in: Zeitschrift für Betriebswirtschaft, Jg. 67, Heft 2, S. 179-201
- Michels, E. (2001): Data Mining Analyse im Handel – konkrete Einsatzmöglichkeiten und Erfolgspotentiale, in: Hippner, H.; Ulrich, K.; Meyer, M.; Wilde, K. (Hrsg.), Handbuch Data Mining im Marketing, Vieweg, S. 933-950
- Neuneier, R.; Tresp, V. (1994): Radiale Basisfunktionen, Dichteschätzungen und Neuro-Fuzzy, in: Rehkugler, H.; Zimmermann, H.G. (Hrsg.), Neuronale Netze in der Ökonomie, München, S. 89-130
- Neuneier, R.; Zimmermann, H.G. (1998): How to Train Neural Networks, in: Orr, G.B.; Müller, K.-R. (Eds.), Neural Networks: Tricks of the Trade, Berlin, S. 373-423
- Podding, T.; Sidorovitch, I. (2001): Künstliche neuronale Netze: Überblick, Einsatzmöglichkeiten und Anwendungsprobleme, in: Hippner, H.; Ulrich, K.; Meyer, M.; Wilde, K. (Hrsg.), Handbuch Data Mining im Marketing, Vieweg, S. 363-402
- Poloni, M.; Nelke, M. (2001): Kundensegmentierung und Zielgruppendefinition im Database Marketing am Beispiel von Direktvertriebsprodukten, in: Hippner, H.; Ulrich, K.; Meyer, M.; Wilde, K. (Hrsg.), Handbuch Data Mining im Marketing, Vieweg, S. 643-650
- Popper, K.R. (1935): Die Logik der Forschung, Wien
- Popper, K.R. (1963): Conjectures and Refutations, London
- Quinlan, J.R. (1986): Induction of Decision Trees, in: Machine Learning, Vol. 1, S. 81-106
- Quinlan, J.R. (1993): C4.5: Programs for Machine Learning, Morgan, Kaufmann, San Mateo
- Rehkugler, H.; Kerling, M. (1995): Einsatz neuronaler Netze für Analyse- und Prognosezwecke, in: Betriebswirtschaftliche Forschung und Praxis, Heft 3, S. 306-324
- Rudolph, A. (1999): Data Mining in action: Statistische Verfahren der Klassifikation, Shaker, Aachen
- Säuberlich, F. (2000): KDD und Data Mining als Hilfsmittel zur Entscheidungsunterstützung, Lang, Frankfurt
- Schedlitz, P.; Reutterer, T.; Joos, W. (2001): Data-Mining und Sortimentsverbundanalyse im Einzelhandel, in: Hippner, H.; Ulrich, K.; Meyer, M.; Wilde, K. (Hrsg.), Handbuch Data Mining im Marketing, Vieweg, S. 951-970
- Schierreich, T. (2000): Bonitätsprüfung im Versandhandel - Über die Konstruktion von Entscheidungsbäumen, in: Alpar, P.; Niedereichholz, J. (Hrsg.), Data Mining im praktischen Einsatz, Vieweg
- Schinzer, H.; Bange, C.; Mertens, H. (1999): Data Warehouse und Data Mining, 2. Auflage, Vahlen, München
- Schlittgen, R. (1998): Regressionsbäume, in: Allgemeines Statistisches Archiv, 82, S. 291-311
- Wirth, R.; Reinhardt, T.P. (1996): Detecting Early Indicator Cars in an Automotive Data Base: A Multi-Strategy Approach, in: Simoudis, E.; Han, J.; Fayyad, U. (Hrsg.), Proceedings – The Second International Conference of Knowledge Discovery and Data Mining, Menlo Park, S. 76 ff.
- Wittmann, T.; Ruhland, J. (2001): Neuro-Fuzzy Data Mining zur Zielgruppenselektion im Bankbereich, in: Hippner, H.; Ulrich, K.; Meyer, M.; Wilde, K. (Hrsg.), Handbuch Data Mining im Marketing, Vieweg, S. 787-804
- Zhang, C.; Zhang, S. (2002): Association Rule Mining, Springer, Berlin