

# Data Mining: eine Einführung

Vincent OBERLÉ (vincent@oberle.com)

13. März 2000

## 1 Einführung

Seit Data Mining ein heißes Thema geworden ist und sich nicht nur die Wissenschaft, aber auch die Industrie dafür interessiert, gibt es eine Menge von Definitionen. Im allgemeinen wird Data Mining als Teil eines komplexen Prozesses definiert, der als KDD, „Knowledge Discovery in Databases“ (Entdeckung von Wissen in Datenbanken), bekannt ist. Motivation für die Entwicklung von KDD-Technologien ist die einfache Feststellung von mehr und mehr Organisationen: Die Menge von Daten wird immer größer (man spricht von einem exponentialen Wachstum), aber die Bedeutung, die Erkenntnisse, die man aus diesen Daten zieht, steigen nicht entsprechend. Neue Techniken und Software sind notwendig, um den Menschen zu helfen, mit diesen Bergen von Daten einsichtsvoll und selbständig fertig zu werden.

Eine Definition für KDD stellt [2] vor:

*Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially usefull, and ultimately understandable patterns in data.*

Daten sind konkreten Tatsachen, die z.B. in einer Datenbank gespeichert sind. *Pattern* kann man als Regeln sehen, die, wenn sie gültig (*valid*), neu (*novel*), nützlich (*potentially useful*) und begreiflich für Menschen (*ultimately understandable*) sind, zu Wissen werden.

Dieser Beitrag ist wie folgt aufgebaut: Abschnitt 2 erklärt, warum Data Mining gebraucht wird. Abschnitt 3 stellt Anwendungsbeispiele vor. Der 4. Abschnitt erklärt die verschiedenen Stufen eines KDD Prozeß, und im 5. Abschnitt werden drei Data Mining Techniken vorgestellt. Der Beitrag schließt mit einem Ausblick auf zukünftige Data Mining Arbeitsgebiete (Abschnitt 6) und einer Zusammenfassung (Abschnitt 7).

## 2 Warum Data Mining?

### 2.1 Aktion und Wissen

Abbildung 1 zeigt die Verbindungen zwischen Aktion und Wissen in Abhängigkeit von der Veränderung der Daten [1]. Das Wissen beeinflusst mehrere Aktionsstufen:

1. Zuerst beeinflusst es die Datenerfassung. Wenn Daten nicht bedeutsam scheinen, werden sie nicht im Entscheidungsprozeß integriert.

Zum Beispiel kann die Nachricht „Luxusbier einkaufen“ kann unwichtig scheinen. Aber wenn man klarstellt, daß 80 % der Luxusbier-Käufer auch ein CD-Rom Lesegerät in den nächsten sechs Monaten kaufen, dann wird dieses Luxusbierkaufen viel wichtiger und bekommt eine andere Bedeutung.

2. Das Wissen führt auch dazu, daß eine Veränderung in den gespeicherten Daten passiert und sie zu Informationen werden. Der Verantwortliche der Computerabteilung wird sich vielleicht an den neuen Information über die Luxusbier-Verkäufe interessieren und die Luxusbier-Käufer über CD-Rom-Angebote informieren.

### Abbildung 1: Aktion und Wissen

3. Schließlich ist das Wissen wichtig, um die Entscheidungen zu treffen. Wann sollte man dem Luxusbier-Käufer Werbung zuschicken?

Vergleicht man die ersten zwei Fragestellungen mit der letzten, so kann Data Mining am meisten in der letzten Stufe helfen. Data Mining trifft keine Entscheidung, es bringt nur die Information zu Tage, um sie einfacher treffen zu können.

## 2.2 Zuviele Daten töten Information

WalMart, das größte Verkaufsunternehmen in den USA, sammelt täglich 20 Millionen Transaktionen, um den Trend von jedem Produkt zu berechnen [1]. Die neuen Datenbanken ermöglichen die Speicherung von so vielen Daten, aber es gibt keine Analyseprodukte für diese Datenmengen.

Computer haben Wissen versprochen. Im allgemeinen geben sie nur eine Menge von unbeherrschten Daten aus. Wenn man diese Daten mit traditionellen statistischen Methoden analysieren würde, wären die Statistiker die Berufsgruppe Nr. 1 im Jahre 2015!

Außerdem sind die aktuellen Datenbankanfragetechnologien ziemlich primitiv. Man ist noch weit von der naturellen Sprache entfernt, wie sie im Film „2001: A Space Odyssey“ mit dem HAL Computer möglich ist. Im Gegenteil, um ein Problem zu lösen, muß man schon wissen, wie die Daten organisiert sind. Zum Beispiel bei der Frage „Kauft dieser Mann regelmäßig in diesem Geschäft?“ kann keine Datenbank das Wort „regelmäßig“ verstehen.

## 2.3 Das Datawarehouse

Ein Datawarehouse (oder Datenlager) ist eine Ansammlung von strukturierten Daten, die von verschiedenen Betriebssystemen kommen. Ein Datawarehouse ist der Entscheidungshilfe gewidmet. Das Ziel ist es, einen Überblick über das gesamte Produktionssystem zu bekommen.

Die Datawarehouse-Technologie bildet die Basis für die Entwicklung einer neuen Generation von Produkten zur Analyse der Daten. Hier findet man die OLAP-Produkte, und natürlich auch die KDD-Lösungen, die man allgemein Data Mining Produkte nennt.

Ein Datawarehouse ist dennoch nicht unbedingt notwendig, um Data Mining zu machen.

## 2.4 Von Produktionssystemen zum Data Mining

Die Informatikanwendungen innerhalb eines Unternehmens kann man in zwei Kategorien einteilen: Anwendungen innerhalb der Produktion und Anwendungen zur Entscheidungshilfe.

Die Produktionssysteme sind die älteren und sind auch noch immer sehr stark und wichtig. Man braucht nur den Erfolg von den ERP-Systemen<sup>1</sup> wie SAP oder Baan zu sehen. Aber diese Systeme unterscheiden sich kaum. Die Firmen sind am Ende ungefähr mit denselben Systemen ausgerüstet, und deshalb ist es nicht so schwer, zum Beispiel eine gute Lagerungsführung zu haben.

Die richtige Entscheidung zu treffen, ist aber eine andere Sache. Die Entscheidungsinformatik umfaßt alle Anwendungen, die Teil eines Entscheidungsprozesses innerhalb eines Unternehmens führen können [1]. Diese Anwendungen können in vier Bereiche eingeteilt werden. Jede Anwendung wird von den anderen beeinflusst (siehe auch Abbildung 2).

- Datenbanken (Oracle, IBM DB2, Informix. . .) für die Datenlagerung.
- Reporting Produkte (Business Objects, Impromptu, Brio Query. . .) für die statistischen Berichte über Datenanfragen.
- OLAP-Produkte (SAS MDDDB, Oracle Express, Cognos Powerplay. . .) für die mehrdimensionale Analyse.
- Data Mining Produkte für das Entdecken von verstecktem Wissen in den Daten.

Abbildung 2: Anwendungen der Entscheidungsinformatik

---

<sup>1</sup>ERP: *Enterprise Resource Planning*

## 3 Der Aufschwung des Data Mining

### 3.1 Wesentliche Anwendungen des Data Mining

Im allgemeinen ist Data Mining überall da sinnvoll, wo es viele Information gibt und wo Prozesse verbessert werden können, das heißt in fast allen Wirtschaftsbereichen. Praktisch wird Data Mining besonders bei der Kundenanalyse oder bei der Betrugsentdeckung benutzt.

Die folgende Liste zeigt bekannte, wichtige Anwendungen [1].

#### **Einkaufscenter und Versandverkauf**

- Suche von Gleichartigkeit zwischen Kunden in Abhängigkeit von geographischen oder soziologischen Kriterien
- Vorhersehung von Antwortquoten
- Optimale Gestaltung der Belieferung

#### **Pharmazeutische Labore**

- Findung der besten Therapie für verschiedene Krankheiten

#### **Banken**

- Suche nach der Kartenbenutzung, die typisch für einen Betrug ist
- Vorhersehung von Eintreibungskapazitäten

#### **Versicherungen**

- Kundenauswahlmodelle
- Schadensanalyse

#### **Industrie**

- Qualitätskontrolle, Vorhersehung von Defekten

#### **Telekommunikation**

- Preissimulation
- Betrugsentdeckung

### 3.2 Der Aufschwung der Data Mining Produkte

Die Konjunkturforschung gehen davon aus, daß der Entscheidungshilfemarkt ungefähr 30 Milliarden DM im Jahr 1995 groß war und daß der Data Mining-Teil 100 Millionen DM umfaßte. Es wird vorausgesagt, daß der Data Mining-Teil auf 1,5 Milliarden DM im Jahr 2000 anwachsen (40 % Aufschwung pro Jahr) wird.

### 3.3 *Return of Investment*

Data Mining kann ein richtige Goldmine für ein Unternehmen sein. Die Investition kann sich in sehr kurzer Zeit bezahlt machen. Dies wird im folgenden anhand eines Beispiel erklärt.

Im Versandverkauf habe eine Firma ein Hauptkatalog, der sich nicht mehr vergrößern läßt. Deswegen werden zusätzlich spezielle Kataloge für kleinere Märkte, wie für die Jugend, oder für Video, Innenarchitektur, usw. erstellt.

Das Problem ist dann: Welchem Kunden wird einen Hauptkatalog und welchem ein Spezialkatalog zugeschickt, wobei die Gesamtkosten zu optimieren sind.

Mit Data Mining Techniken kann der Verkäufer ein Modell entwickeln, der einen potentiellen Käufer in einem speziellen Katalog von einem Käufer in dem Hauptkatalog unterscheiden kann. Mit diesem Modell erzielt er eine Einkaufsquote von 8,5 %, gegen 7,7 % mit dem alten System (Prozentzahlen geben an, wieviele Personen, denen ein Katalog zugeschickt wurde, etwas aus diesem Katalog kaufen).

Das bringt bei 10 000 mehr Verkäufen pro Jahr und einem Gewinn von ungefähr 30 DM pro Verkauf, 300 000 DM insgesamt. Der Data Mining Prozeß hat 25 000 DM gekostet, das macht also eine Return-of-Investment-Zeit von ungefähr einem Monat. Die Rentabilität für dieses Bereich ist enorm hoch.

## 4 Der KDD Prozeß

Data Mining Software oder Data Mining Algorithmen dürfen nicht mit dem Data Mining oder KDD Prozeß verwechselt werden. Die Data Mining Produkte sind nur ein Teil eines Prozesses, der aus acht Stufen besteht. [1] und [2] geben ein komplettes Beispiel für diesen Prozeß. Diese acht Stufen werden nun einzeln erläutert.

### 4.1 Die Problemstellung

Diese Stufe ist wichtig, weil man hier die Grundlagen legt, besonders um das Thema eines Problems richtig zu verstehen. Das Problem wird in ein Format gebracht, so daß KDD-Techniken angewandt werden können. Die erwarteten Ziele werden definiert. In dieser Stufe könnten auch die Erfahrungen von Experten benutzt werden.

### 4.2 Suche nach Daten

Hier fragt man sich, welche Daten benutzt werden und wie man sie bekommt. Der Anzahl der Variablen wird auch reduziert, so daß die Leistung der Data Mining Anwendung akzeptabel bleibt und daß das Modell allgemein bleibt und sich nicht in Einzelheiten verliert.

Das stellt natürlich die Frage, welche Variablen am treffendsten sind.

### 4.3 Die Auswahl der wesentlichen Daten

Die Auswahl von Daten, die Grundlagen des Data Mining, ist mehr oder weniger einfach in Abhängigkeit von den Technologien in der Firma. Ein bestehendes Datawarehouse oder etablierte Datenbanken können diese Stufe ziemlich vereinfachen. Dennoch kann sie bis zu 80 % der Arbeit des ganzen Prozesses einnehmen.

Hier wird auch unterschieden, ob man alle Daten oder nur die eines Musters benutzt.

### 4.4 Die Reinigung der Daten

Diese Stufe hat als Ziel, sich mit den schlechten Daten zu befassen. Es gibt verschiedene Typen zur Prüfung, die mit dem Ursprung der Daten verbunden sind. Die abweichenden und die fehlenden Daten müssen verwaltet werden.

## 4.5 Die Behandlung der Variablen

In dieser Stufe werden die Variablen bearbeitet, so daß sie besser von der Data Mining Software verwertbar sind.

Innerhalb der Variable kann man die Maßeinheit anpassen, das Datum in die Dauer und die geographische Daten in Koordinaten umrechnen. Mehrere Variablen können auch zusammengebracht werden (Frequenz, Kennziffer, usw.)

## 4.6 Die Wahl des Modells

Diese Stufe wird im allgemeinen im Vergleich zum gesamten Prozeß als die Data Mining Stufe bezeichnet. Bevor es Data Mining Software gab, wurde diese Stufe mit statistischen Methoden durchgeführt: Die Forscher wollten eine Hypothese bestätigen.

Die Suche eines Modells passiert auf einer Lerndatenbank, die verschieden von die Testdatenbank ist. Mehr und mehr Data Mining Software ist nicht mehr vollautomatisch, sondern agiert in Wechselbeziehung mit dem Benutzer. Dieser kann die Forschung führen und die Software erstellt die Modelle.

Die verschieden Data Mining Algorithmen werden in Abschnitt 5 beschrieben.

## 4.7 Der Überschlag des Ergebnisses

Die Ergebnisse des KDD-Prozeß müssen geprüft werden. Dies kann mit der Testdatenbank gemacht werden, das heißt man prüft, ob das Modell mit neuen Beispielen umgehen kann, die es noch nie gesehen hat.

## 4.8 Die Festigung des Wissens

Wissen hat keine Bedeutung, wenn es nicht Aktionen und Entscheidungen verändert. Das kann bedeuten, daß es in das Informationssystem der Firma einbezogen wird, oder nur einfach beurkundet wird und dem interessierten Leser zur Verfügung steht.

Eine mögliche Veränderung kann sein, wenn die Daten schlecht sind, die Fütterung des Datawarehouse anzupassen.

# 5 Data Mining Techniken

Data Mining ist ganz sinnlos ohne die Erfahrungen und das Wissen von Experten, das ist klar. Aber was man unter dem Wort Data Mining versteht, sind die Techniken und die Algorithmen. Es werden nun die drei wichtigen Techniken vorgestellt und nach den Problemen, die sie lösen müssen, eingeordnet. Die drei Techniken sind: Assoziationen finden, Gruppieren in *Cluster* und Entdeckung von Klassifikationen.

## 5.1 Assoziationsregeln (*Association*)

### 5.1.1 Definition

Diese Data Mining Technik besteht in der Suche von Verbindungen, sprich konditionalen Regeln. Eine konditionale Regel ist eine Folge von „Wenn Bedingungen, dann Ergebnis“. Die Suche kann sich auf alle mögliche Ergebnisse beziehen, oder nur auf ein vom Benutzer gewähltes Ergebnis [4].

Die hauptsächtliche Nutzung der Assoziationsregeln liegt heutzutage im Bereich der Kreditentscheidungen und in der Analyse von Registrierkassenbons. Dieses letzte Beispiel wird jetzt besonders erklärt.

### 5.1.2 Anwendungsbeispiel

Die selbstverständliche Anwendung von Assoziationsregel-Techniken finden sich in der Analyse des Einkaufs von Supermarktkunden. Der Ziel ist es, Assoziation zwischen den verschiedenen gekauften Produkten zu finden. Zum Beispiel, 80 % der Käufer von Windeln kaufen auch Bier. Eine andere Möglichkeit besteht darin, Einkäufe über die Zeit zu analysieren, um zu sehen, wie sich die Gewohnheiten eines Käufers entwickeln. Das alles ist jetzt möglich, dank Datawarehouse und Höchstleistungsrechner.

Es gibt mehrere Gründe, um Assoziationsregeln zu suchen. Der Lagerbestand kann optimisiert werden, wenn man voraussehen kann, was sich mehr verkaufen lassen wird. Das Finden von Assoziationsregeln zwischen Produkten kann auch zur Reorganisation der Verkaufsfläche führen.

### 5.1.3 Aufstellung der Assoziationsregeln

Im folgenden wird anhand eines Einkaufsbons-Beispiels das Aufstellen der Assoziationsregeln erklärt [1]. Die Datenbank eines Supermarkt umfaßt die vier Einkaufsbons, die in der Tabelle 1 beschrieben sind.

1	2	3	4
Mehl	Eier	Mehl	Eier
Zucker	Zucker	Eier	Schokolade
Milch	Schokolade	Zucker	Tee
		Schokolade	

Tabelle 1: Einkaufsbons

Eine Assoziationsregel ist eine Ausprägung vom Typ  $X \rightarrow Y$  und zeigt, daß die Präsenz von  $X$  mit der Präsenz von  $Y$  korreliert. Gesucht sind die Regeln, die ein großes Vertrauen und eine große Unterstützung genießen. Die Maße Vertrauen und Unterstützung werden weiter unten im Text erklärt.

Es wird definiert, daß eine Transaktion die Assoziationsregeln ( $Mehl \rightarrow Zucker$ ) enthält, wenn man beide Produkte auf einem Einkaufsbon findet.

Der erste Bon hat die Paare:

$$\begin{array}{lll}
 Mehl \rightarrow Zucker & Zucker \rightarrow Mehl & Mehl \rightarrow Milch \\
 Milch \rightarrow Mehl & Zucker \rightarrow Milch & Milch \rightarrow Zucker
 \end{array}$$

Man sieht, daß das Paar  $Mehl \rightarrow Zucker$  auf den Bons 1 und 3 ist.

Nun die Definition und Anwendung der oben bereits eingeführten Maße:

**Das Vertrauen** gibt an, wie oft die Assoziationsregel  $Mehl \rightarrow Zucker$  vorhanden ist, dividiert durch die Anzahl des Vorkommens der Bedingung Mehl. Mehl kommt zweimal vor, so daß das Vertrauen für die Regel  $Mehl \rightarrow Zucker$  100 % ist. Das Vertrauen gibt an, wie stark eine Assoziation ist.

**Die Unterstützung** gibt an, wie oft die Assoziationsregel  $Mehl \rightarrow Zucker$  vorhanden ist, dividiert durch die Zahl aller Bons. Hier ist die Unterstützung für  $Mehl \rightarrow Zucker$  50 %. Die Unterstützung gibt die gibt die prozentuale Häufigkeit des gemeinsamen Auftretens von den beiden Elementen einer Assoziation bezogen auf alle Mengen.

Das Ziel ist es, die Artikel zu finden, die ein hohes Vertrauen und eine hohe Unterstützung haben. Zuerst werden die Produkte ausgesucht, die eine größere Unterstützung als ein bestimmter Zahl haben (Tabelle 2). Zum Beispiel kann man eine Unterstützung von 30 % nehmen (hier 30 % von 4 Transaktionen). So kann man schon die Produkte Milch und Tee eliminieren, weil sie nur eine Häufigkeit von 25 % haben.

Produkt	Häufigkeit
Mehl	2 = 50 %
Zucker	3 = 75 %
Milch	1 = 25 %
Eier	3 = 75 %
Schokolade	3 = 75 %
Tee	1 = 25 %

Tabelle 2: Frequenz der verschiedenen Produkte

Dann nimmt man die anderen Artikel und erstellt alle mögliche Paare, wie in der Tabelle 3.

Assoziation	Unterstützung
{ Mehl, Zucker }	2 = 50 %
{ Mehl, Eier }	1 = 25 %
{ Mehl, Schokolade }	1 = 25 %
{ Zucker, Eier }	2 = 50 %
{ Zucker, Schokolade }	2 = 50 %
{ Eier, Schokolade }	3 = 75 %

Tabelle 3: Frequenz der verschiedenen Assoziationen

Dann eliminiert man auch die Paare, die eine Unterstützung kleiner als beispielweise 30 % haben. So bleibt noch { Mehl, Zucker }, { Zucker, Eier }, { Zucker, Schokolade }, und { Eier, Schokolade } übrig. Endlich erzeugt man die mögliche Assoziationen und berechnet das Vertrauen (Tabelle 4).

Assoziation	Vertrauen
<i>Mehl</i> → <i>Zucker</i>	2/2 = 100 %
<i>Zucker</i> → <i>Mehl</i>	2/3 = 66 %
<i>Zucker</i> → <i>Eier</i>	2/3 = 66 %
<i>Zucker</i> → <i>Schokolade</i>	2/3 = 66 %
<i>Eier</i> → <i>Zucker</i>	2/3 = 66 %
<i>Eier</i> → <i>Schokolade</i>	3/3 = 100 %
<i>Schokolade</i> → <i>Zucker</i>	2/3 = 66 %
<i>Schokolade</i> → <i>Eier</i>	3/3 = 100 %

Tabelle 4: Frequenz der verschiedenen Assoziationen

Nun versucht man Gruppen von Produkten zu finden, die gemeinsam den Kauf eines anderen Produkts erzwingen. Die einzige Gruppe besteht aus Zucker – Eier – Mehl. Man kann es zwei mal finden, so hat es eine Unterstützung von 50 %.

Das Algorithmus endet hier, weil es keine Gruppe von 4 Artikeln gibt.

Aus dem gewonnen Wissen können folgende Aktionen resultieren:

- *Mehl* → *Zucker* hat ein Vertrauen von 100 % und eine Unterstützung von 50 %. Das bedeutet, daß wenn ein Kunde Mehl kauft, kauft er auch Zucker.
- *Zucker* → *Mehl* hat ein Vertrauen von 66 % und eine Unterstützung von 50 %. So kann man zum Beispiel den Preis für Zucker-Käufer senken.
- *Eier* → *Schokolade* und *Schokolade* → *Eier* haben beide 100 % Vertrauen und 100 % Unterstützung. Die beide Produkte sollten dann vielleicht nebeneinander verkauft werden.

## 5.2 Klassifikation

### 5.2.1 Definition

Das Klassifikationsproblem kann so erklärt werden: Es gibt, Daten anhand einer Beispielmenge zu analysieren. Jeder Datensatz besitzt mehrere Kennzeichen, und ist einer Klasse zugeordnet. Die Beispielmenge wird benutzt, um einen Modell zu erstellen. Dieses Modell soll es erlauben, neue Daten aufgrund ihrer Kennzeichen in die Klassen einzuordnen.

### 5.2.2 Klassifikationstechniken

Um das Klassifikationsproblem zu lösen, gibt es eine Menge von bekannten Techniken. Einige sind [3]:

- **Entscheidungsbäume (*decision tree*).**

Ein Entscheidungsbaum ist eine Aneinanderkettung von logischen Regeln, die automatisch aufgebaut werden aufgrund einer Beispieldatenbank. Jede Regel besitzt die *if...then* Struktur.

- **Genetische Algorithmen.**

Sie sind ziemlich neu<sup>2</sup> und basieren auf denselben Regeln wie die natürliche Auslese. Sie beschreiben die Entwicklung einer Testgruppe in Abhängigkeit von seiner Umwelt.

Genetische Algorithmen sind einfach zu benutzen, aber sehr wirksam. Ein Beispiel wird in 5.2.3 gegeben.

- **Bayesche Netze (oder *Bayesian classifiers*).**

Bayesche Netze sind eine klassische Technik. Sie wird benutzt, um die Wahrscheinlichkeit eines Ereignis zu finden, wenn andere Ereignisse bekannt sind.

Obwohl das Modell ziemlich einfach ist, ist er ziemlich gut bei vielen Problemen.

- **Neurale Netze.**

Neurale Netze sind wieder in den letzten Jahren ein heißes Thema geworden. Ein Neuronales Netz ist ein Computerprogramm, welches die prinzipielle Wirkungsweise eines Gehirns nachbildet. Es lernt selbständig Zusammenhänge aufgrund von Beispielen.

Zum Beispiel, wenn wir eine Herdplatte sehen und die Luft darüber warm ist, so haben wir einmal gelernt, daß es sehr schmerzhaft sein kann, auf die Herdplatte zu fassen. Unser menschliches Neuronales Netz hat nun die Regel gelernt, daß „wenn Herdplatte & Luft heiß“ (Inputs), dann „nicht anfassen“ (Output). Der Vorteil des Computerprogramms ist es jedoch, daß es weitaus kompliziertere Regeln bilden kann, und das in weitaus kürzerer Zeit.

### 5.2.3 Prinzip und Beispiel eines genetischen Algorithmus

Genetische Algorithmen werden viel benutzt, um die Leistung von Data Mining Software zu optimieren. Zum Beispiel können Entscheidungsbäume optimiert werden: Der genetische Algorithmus würde die Variablen finden, die am bedeutsamsten sind.

Die Theorie von Darwin stellt das folgenden Prinzip fest: eine Bevölkerung der Erde verändert sich in jeder Generation, weil sich jede Generation an eine andere Umwelt adaptieren muß.

Dieser Entwicklungsprozeß wird von den Genen gesteuert. Die Gene werden in Chromosomen organisiert. In der Natur überleben wegen des Faustrechts nur die Kreaturen, die am besten an die Umwelt adaptiert sind. Die Reproduktion einer Spezies bringt auch Diversität. Sie basiert auf der Zusammenstellung der Gene der beiden Eltern zu einem neuen Wesen [1].

Genetische Algorithmen basieren auch auf einem solchen Entwicklungsprozeß. Sie verändern eine Datenmenge, bis ein Optimum erreicht wird. So könnte man zum Beispiel die Daten einer Datenbank über Kunden nach folgenden Regeln binär kodieren:

1. Alter einer Bestellung (1 für weniger als 6 Monate, 0 sonst)
2. Umsatz pro Jahr (1 für weniger als 300 DM, 0 sonst)
3. Zahl der Bestellung pro Jahr (1 für mehr als 2, 0 sonst)
4. Alter des Kund (1 für weniger als 45 Jahre, 0 sonst)
5. Hat er Kinder? (1 für ja, 0 für nein)

So ist 10110 der Code für die Kunden, die seit weniger als 6 Monaten etwas bestellt haben, die einen Umsatz höher als 300 DM haben, die mehr als zwei mal bestellt haben, die jünger als 45 sind und die keine Kinder haben. Andere Codierungen sind natürlich auch möglich, wie für die Nummer 1.43, der 143 oder 10001111 in binärer Schreibweise.

---

<sup>2</sup>John Holland, 1975

Danach braucht man noch eine Schätzungsfunktion  $F(n)$ , um die interessanten Kunden zu identifizieren. In dem Beispiel würde  $F(n)$  eine Einkaufsquote sein, die mit traditionellen statistischen Methoden berechnet wurde, wie in der Tabelle 5. In Tabelle 5 ist auch noch angegeben, wieviele Kunden es von jedem Typ gibt (Stärke).

Typ	$F(n)$	Stärke
01000	1.75 %	5000
00010	0.25 %	2500
10110	3.28 %	1500
00111	2.35 %	1000

Tabelle 5: Chromosomen und ihre Schätzungsfunktion

Der Durchschnittswert von  $F(n)$  ist 2.0 %.

**Der Auswahlprozeß** Die Auswahl der besten Chromosomen basiert auf der Schätzungsfunktion. Die Chromosomen, die überleben dürfen, werden mit einer Zufallsmethode gewählt. Die Funktion ist die folgende:  $(2 * \pi) * (f_i/f)$  mit  $f_i$  als Schätzungsfunktion für den Typ und  $f$  als Schätzungsfunktion für die Bevölkerung (Durschnitt von  $F(n)$ ).

Für das Beispiel sind die Auswahlwerte in der Tabelle 6 zu finden.

Typ	$F(n)$	
01000	1.75 %	$2 * \pi * (1.75/2.00) = 5.49$
00010	0.25 %	$2 * \pi * (0.25/2.00) = 0.78$
10110	3.28 %	$2 * \pi * (3.28/2.00) = 10.30$
00111	2.35 %	$2 * \pi * (2.35/2.00) = 7.38$
		<i>Summe = 23.95</i>

Tabelle 6: Auswahl von den Chromosomen

Damit läßt sich für die Gruppe 10110 berechnen, daß der Anteil der Gruppe 10110 in der nächsten Generation  $10.30/23.95 = 43$  % betragen wird. Am Anfang war dieser Gruppe nur 15 % stark, so darf jeder Chromosom in dieser Gruppe 2.8 Söhne haben.

**Genetische Behandlungen** Für jede Generation erzeugt der Algorithmus neue Chromosomen mit den drei folgenden Techniken.

- Die Vermischung (*cross-over*: ein Teil von zwei Chromosomen wird vertauscht.

Beispiel:

**0 1 0 1 0** → **1 0 0 1 0**

**1 0 0 1 1** → **0 1 0 1 1**

- Die Mutation ist die Veränderung eines Elements in einem Chromosom. So wird die Situation nicht blockiert, weil es nicht genügend Verschiedenartigkeiten gibt.

Beispiel:

**0 1 0 1 0** → **0 1 1 1 0**

- Die Inversion.

Beispiel:

**0 1 0 1 0** → **1 0 0 1 0**

#### 5.2.4 Klassifikationsanwendungen

Das Klassifikationsproblem hat mehrere direkte Anwendungen im Marketing und E-commerce, einfach um vorherzusehen, welche Benutzergruppe eine bestimmte Werbung erreicht. Ein anderer Bereich ist die Diagnosehilfe. Algorithmen zur Texteinordnung sind auch sehr wertvoll, um automatisch Bibliotheken zu erstellen.

### 5.3 Clustering

*Clustering* ist eine Technik, um Gleichartigkeiten zu finden, Kunden in Kategorien einzuteilen, die Erkennung von Motiven oder die Analyse von Trends. *Clustering* Techniken werden von den Statistikern und Datenbankern schon sehr lange studiert [3].

#### 5.3.1 Definition

Zuerst muß man die Zusammenstellung von Daten (*clustering data points*) definieren: Man hat Daten in einer mehrdimensional Fläche. Das Ziel ist es, eine Teilung dieser Daten in *Cluster* zu finden, so daß die Punkte in jedem *Cluster* sehr nahe sind [3].

#### 5.3.2 Das Dimensionenauswahlproblem

Die meisten clustering Algorithmen sind nicht sehr wirksam, wenn es viele Dimensionen gibt. Bei vielen Dimensionen gibt es allgemein nur einige, wo die Punkte bedeutsam im *Cluster* zusammenliegen. Deswegen müssen *clustering* Algorithmen zuerst sinnvolle Dimensionen auswählen. Das Ziel ist es, besondere Dimensionen zu finden, wo die Daten in Wechselbeziehung zueinander stehen.

Dieses Prozeß reduziert das Datengeräusch, leider können auch wichtige Nachrichten verloren gehen.

Die Abbildung 3 erläutert dieses Problem. In zwei verschiedenen Dimensionen sind unterschiedliche Gruppe gewählt worden, aber in jedem Fall gehen Informationen verloren. Es gibt zwei Cluster, je nachdem welche Dimensionen man sieht, die Informationen aus der anderen Dimension gehen verloren.

#### 5.3.3 Clustering-Techniken

Zuerst definiert man ein projiziertes *cluster*. Dazu wählt man zunächst Dimensionen aus, so daß es in der Projektion *Cluster* (eng zusammenliegende Daten) gibt. In der Abbildung 3 gibt es zwei verschiedene projizierte *Cluster*. Der erste gibt es nur in x-y und das andere in x-z.

Sonst sind *Clustering* Algorithmen vom selben Typ wie Klassifikationstechniken (siehe 5.2.2).

#### 5.3.4 Vorteile und Anwendungen

Ein Vorteil der *clustering* Techniken ist, daß sie nicht nur *Cluster* produzieren, aber auch eine Beschreibung über die *Cluster*, über die Dimensionen, und das gibt den *Clustern* einen Sinn.

*Clustering* kann sehr gut mit sehr großen Mengen von Daten durchführen, wie mit Video-Datenbanken oder Marketing Anwendungen. Daten werden in Gruppe organisiert. Die Gruppen sind über die Dimensionen definiert, und das ist die wichtige Information über die Daten in jeder Gruppe.

## 6 Perspektiven

### 6.1 Zugänglichkeit und Leistung

Der allgemeine Trend von Data Mining Software geht in zwei gegensätzliche Richtungen: die Zugänglichkeit und die Leistung [1].

**Die Zugänglichkeit:** Software vereinfacht immer mehr die Benutzung eines KDD-Prozesses und verdeckt die Komplexität der Modelle. Data Mining wird allgemeinverständlich.

**Die Leistung:** Die Algorithmen werden immer besser und schneller. Die Voraussagen werden präziser.

Das hat zwei wichtige Folgen. Die Benutzer haben jetzt Techniken, die früher für Spezialisten bestimmt waren, und es gibt jetzt neue Spezialisten, um diese komplexen Algorithmen zu parametrisieren.

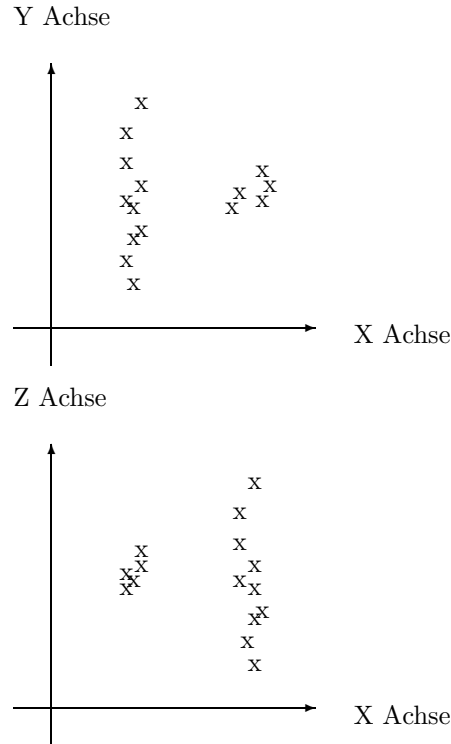


Abbildung 3: Schwierigkeiten mit Dimensionen Auswahl

## 6.2 Die Vereinigung von Data Mining und Datenbanken

Wie gerade gesagt, strebt die Entwicklung von Data Mining nach einer Vereinfachung der Benutzung. So ist normalerweise die beste Data Mining Lösung, die Lösung, die man ganz vergißt. Die Einbeziehung der KDD Techniken direkt in Datenbanken geht in dieser Richtung.

Dieser Trend steht nur am Anfang, aber die großen Hersteller von Datenbanken wie Oracle oder IBM mit DB2 haben schon die Türen geöffnet, so daß die Data Mining Algorithmen im Kern der Datenbank integriert werden können.

## 6.3 Die Vereinigung von Data Mining und OLAP

OLAP<sup>3</sup> bezeichnet eine Kategorie von Datenuntersuchungssoftware, um Daten in mehreren Dimensionen sichtbar zu machen. Ein anderer Trend der Data Mining Techniken ist sie nicht in die Datenbanken einzubeziehen, aber lieber in die Software, die die Daten anfragen oder sichtbar machen. OLAP Software, wie die Produkte von Business Objects, illustrieren diese Trend.

Data Mining Algorithmen werden benutzt, um dem Benutzer vorzustellen, welche Dimension er als erstes studieren sollte. Es ist ein Ratgeber. Um sich von den Konkurrenzprodukten zu unterscheiden, ist es klar, daß mehr und mehr OLAP Produkte Data Mining Funktionen realisieren werden.

## 6.4 Data Mining und Multimedia

Data Mining wird am ehesten für strukturierte Daten benutzt. Aber neurale Netze zum Beispiel werden schon lange in der Schrift- und Bildererkennung benutzt.

**Text Mining** analysiert Dokumente, um Wörter- und Konzeptverbindungen zu finden. Es kann zum Beispiel für die Analyse von Kundenkommentaren benutzt werden.

<sup>3</sup>On Line Analytical Processing

**Image Mining** sucht Verbindungen zwischen Bildern, zum Beispiel zwischen medizinischen Bildern nach derselben Krankheit.

**Video Mining** ist dasselbe wie Image Mining, aber zwischen Videos. Bisher ist es aber nur Theorie, weil diese Technik sehr viel Leistung braucht.

Wenn die Aussage von Moore<sup>4</sup> stimmt, dann wird es in Zukunft noch mehr Anwendungen für diese Techniken geben.

## 6.5 Data Mining und Internet

Die Wechselwirkungen zwischen Data Mining und Internet lassen sich in drei Kategorien einteilen:

- Das Internet verändert stark die Datensammlung, besonders die Sammlung von Daten über Kunden. Für diese sehr großen Datenbanken sind KDD-Techniken unentbehrlich.
- Data Mining bringt dem Benutzer neue Lösungen, um seine Suche und Navigation zu vereinfachen.
- Das Internet bietet für Data Mining Produkte eine Benutzeroberfläche mit allen Vorteilen von Internet-Benutzeroberflächen (im Vergleich zu Client/Server).

Wenn man sieht, wie die Wörter Internet und Data Mining das Silicon Valley anregt haben, so ist es klar, daß bald neue Ideen in diesem Bereich herauskommen werden.

### 6.5.1 Internet für die Datensammlung

Das Web ist schon lange mehr als nur Darstellungsmittel von Multimedia-Daten. Jetzt ist es eine Entwicklungsplattform für neue Dienste. Der Preis der Datensammlung ist auch reduziert, weil die Datenerfassung selbst von dem Benutzer gemacht wird.

### 6.5.2 Knowbots und intelligente Agenten

Knowbot kommt von Knowledge und Robot, das englische Wort für intelligenter Agent. Agenten sind Entitäten, die autonom in einer heterogenen Welt agieren können [1]. Knowbots können Probleme lösen, aufgrund von Wechselwirkungen untereinander, und sind eine Alternative zu komplexen, zentralen Systemen. Ein Agent hat sein eigenes Ziel, und das ist ein wichtiger Unterschied zu traditionellen Programmen. Die Struktur eines Agenten ist in der Abbildung 4 zusammengefaßt.

Diese intelligenten Agenten finden natürlich ihre Nützlichkeit im Internet. Sie können zum Beispiel die günstigsten, beste Preise suchen, oder auch ein Portal in Abhängigkeit von den persönlichen Interessen aufbauen.

Ein interessantes Beispiel ist die Web Site der Firma Firefly<sup>5</sup>. Es funktioniert ungefähr so: Der Benutzer muß Fragen über seine Interessen beantworten. Dann stellt ihm der Server eine Liste von Artikeln vor, die Leute mit denselben Interessen gelesen und gemocht haben. Man kann seine eigenen Vorstellungen angeben und entsprechend werden die Vorstellungen der Sites immer präziser.

### 6.5.3 Internet als Zugangsmittel zum Data Mining

Im Data Mining haben Produkte wie SAS, Information Discovery oder DSS Agent schon Zugangsmöglichkeiten von Internet Browsern. Das ist eine normale Entwicklung, wodurch das Data Mining mehr Benutzern zugänglich wird.

## 6.6 Data Mining und Recht

Anwendungen von Data Mining sind vielschichtig, aber eine der ersten beschäftigte sich mit den Erkenntnissen über Kunden und Anwendungen im Marketing.

Frankreich hat schon lange ein Gesetz genannt *Informatique et liberté* (Informatik und Freiheit), das den Menschen gegen die ungewünschte Verwendung von seinen persönlichen Daten schützt. Es gibt auch eine Organisation, die CNIL. Alle Firmen müssen ihre Datenbanken über ihre Kunden beim CNIL melden.

<sup>4</sup>Moore, ein Gründer von Intel, verspricht uns ein exponentielles Wachstum der Prozessorleistung.

<sup>5</sup><http://www.firefly.com>

Abbildung 4: Struktur eines Agentes

Deutschland, obwohl es nicht so weit geht, hat auch gleichartige Gesetze.

Data Mining bleibt rechtlich erlaubt, aber die Prozesse dürfen natürlich keine Kriterien wie die politische Meinung oder das Geschlecht berücksichtigen, und auch nicht herabsetzende Bezeichnungen enthalten. Man kann sich vorstellen, daß jemand eine Werbung bekommt, die viele persönliche Informationen hat. Das Gesetz sagt, daß dieser Mensch einen Zugang zum Ursprung dieser Informationen haben muß. Was wird man ihm denn geben? Ein neurales Netz, ganz unverständlich für den Nichtkenner.

## 7 Zusammenfassung

Data Mining ist weniger eine Revolution, als eine normale Entwicklung der bestehenden Produkte. Sie werden einfacher und zugänglicher. Die Preissenkung der Prozessorleistung hilft auch, daß der normale Benutzer Zugang zu Produkten findet, die früher nur für Spezialisten waren.

Data Mining wird sehr stark bleiben und noch wichtiger werden. Was aber nicht so sicher ist, ist, ob die Technologie autonom bleiben wird oder ob es ganz in andere Produkte (Datenbanken, OLAP Produkte...) integriert wird.

Data Mining darf kein exklusives Verfahren sein. Nicht nur Informatiker, aber auch andere müssen im Prozeß einbezogen werden. Die Ergebnisse haben nur Wert im Vergleich zu den Erkenntnissen der Firma.

Data Mining braucht keine Datawarehouse, um anzufangen, und obwohl der äußere Schein manchmal täuscht, die Techniken funktionieren wirklich. Investitionen in Data Mining Techniken sind schnell rentabel.

## Literatur

- [1] R. Lefébure, G. Venturi, *Le Data Mining*, Eyrolles, 1998.
- [2] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, *From Data Mining to Knowledge Discovery: An Overview*, *Advances in Knowledge Discovery and Data Mining*, MIT Press, 1996.

- [3] C. C. Aggarwal, P. S. Yu, *Data Mining Techniques for Associations, Clustering and Classification, Methodologies for Knowledge Discovery and Data Mining*, Lecture Notes in Artificial Intelligence 1574, April 1999.
- [4] I. Cengiz, *Mining Association Rules*,  
<http://www.cs.bilkent.edu.tr/~icengiz/datamine/mining.htm>.
- [5] R. Rastogi, K. Shim, *Recent Advances in Data Mining Algorithms on Large Databases*,  
<http://www.bell-labs.com/project/serendip>.