**Zeitschrift:** Arbido
**Herausgeber:** Verein Schweizerischer Archivarinnen und Archivare; Bibliothek Information Schweiz
**Band:** 14 (1999)
**Heft:** 5

**Artikel:** Data mining - wie finde ich in Datensammlungen entscheidungsrelevante Muster?
**Autor:** Schmid, Josef
**DOI:** https://doi.org/10.5169/seals-769097

# DATA MINING – WIE FINDE ICH IN DATENSAMMLUNGEN ENTSCHEIDUNGSRELEVANTE MUSTER?

*von Josef Schmid, SPSS (Schweiz) AG*

## WAS IST DATA MINING?

***Data Mining – ein neuer Begriff taucht auf***

In Zusammenhang mit Datenbanken und Unternehmensdaten wird immer häufiger ein relativ neuer Begriff ins Spiel gebracht, nämlich *Data Mining*. Neuerdings wird über Data Mining sogar in der Tagespresse geschrieben. So berichtete beispielsweise die *Sonntags Zeitung* vom 14.3.1999 unter dem Titel «Migros hat mit Cumulus die Nase vorn», dass die Migros offenbar aus der Cumulus-Kundenbindungsaktion gewonnene Kundendaten für einen Versand benütze – und ein solches Vorgehen hat sehr viel mit Data Mining zu tun.

***Lernen aus Erfahrung***

Was ist nun aber unter Data Mining konkret zu verstehen? Im anglo-amerikanischen Raum ist der Begriff wesentlich bekannter als in Europa, und entsprechend finden sich wesentlich mehr Publikationen aus diesem Raum zum Thema. Was sicherlich gesagt werden kann ist, dass der Begriff Data Mining über fast so viele unterschiedliche Definitionen verfügt, wie es Publikationen gibt. Ein wesentliches gemeinsames Element aller Definitionen ist aber, dass Data Mining mit *Lernen* zu tun hat. Data Mining heisst im Kern nichts anderes, als dass aus Erfahrungen für zukünftiges Handeln gelernt wird. Das ist nun aber wirklich nichts Neues – es ist sogar ein menschliches Grundmuster. Jeder erfolgreiche Geschäftsmann beispielsweise wendet *«Lernen durch Erfahrung»* an: seine Erfahrung aus früheren Geschäftsprozessen zeigt ihm, wie er sich in zukünftigen Geschäften am erfolgversprechendsten verhalten soll. Hier zeigt sich ein weiteres Merkmal von Data Mining: das Gelernte soll auch *in neue Entscheidungen umgesetzt* werden.

***Datenmengen und moderne Algorithmen***

Neu im heutigen Umfeld ist aber, dass einem Geschäftsmann beispielsweise wesentlich mehr und differenziertere Informationen zur Verfügung stehen als früher. Ein Unternehmen ohne Datenbanken ist im heutigen Umfeld gar nicht mehr vorstellbar: Transaktionsdatenbanken wie Auftragsbearbeitung, Buchhaltung, Lagerverwaltung, Rechnungswesen, Personalverwaltung etc. speichern enorme Informationsmengen über gegenwärtige wie auch über vergangene Vorgänge. Grössere Unternehmen verfügen in einem zunehmenden Masse über sogenannte *Data Warehouses*, in denen die wichtigsten Informationen zusammenhängend und bereinigt zur Verfügung gestellt werden und über *Data Marts*, einen fokussierten *Subset* des Warehouses.

Diese Informationsmenge ist für eine Einzelperson – eben unseren Geschäftsmann – nicht mehr so einfach zu verarbeiten. Angefangen schon bei der enormen Fülle, die kaum überblickt werden kann, bilden das Herstellen von Zusammenhängen oder auch nur schon die Interpretation (Was ist in diesem Feld überhaupt gespeichert?) ohne Hilfsmittel unüberwindbare Hürden. Auf der anderen Seite stellt das rasche Lernen aus den Daten, sprich das schnelle Fällen von optimalen Entscheidungen, je länger je mehr einen erfolgskritischen Faktor für Unternehmen dar. Hier setzen nun die modernen Data-Mining-Instrumente an: sie ermöglichen den Umgang mit den vorhandenen Datenmengen und die schnelle Extraktion der wichtigen Entscheidungsgrundlagen.

***Muster in den Daten entdecken***

Aus den alten Daten lernen, heisst Muster in den vorhandenen Daten zu entdecken. Solche Muster lassen sich in folgende Typen einteilen:

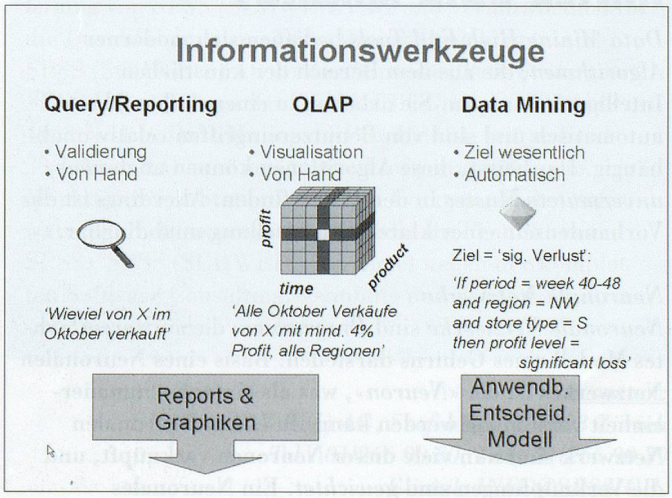
- *Klassifizierung:* z.B. Käufer von Nichtkäufern unterscheiden
- *Assoziation/Sequenz:* z.B. kaufen Bierkäufer häufig auch Nüsse
- *Wahrscheinlichkeit:* mit einer Wahrscheinlichkeit von 74% wird Herr Y dieses Jahr eine Versicherung abschliessen
- *Gruppierung:* z.B. Konsumenten von Luxusgütern oder Gesunde charakterisieren

*Data Mining formt Daten in Entscheidungsgrundlagen um.*

## WERKZEUGE ZUR INFORMATIONSGEWINNUNG

***Queryabfragen/OLAP-Tools***

Um Muster in den Daten zu entdecken, stehen heute verschiedenste moderne EDV-gestützte Werkzeuge für die



unterschiedlichsten Anwendungsbereiche zur Verfügung. So können *Datenbankabfragen* (relativ komplex und umständlich) getätigt oder *EIS/OLAP- und Visualisierungstools verwendet werden, um komprimierte Informationen aus der Datenfülle zu gewinnen*. Gemeinsam ist diesen Werkzeugen, dass sie eine genaue Vorstellung dessen erfordern, was als Information gewünscht wird und nur mit relativ hohem Aufwand an gewandelte Anforderungen angepasst werden kann.

***Statistik***

*Statistische Methoden* gehen über die reine Beobachtung hinaus: mit ihnen werden Beobachtungen in den Daten beispielsweise daraufhin geprüft, ob beobachtete Unterschiede als aussagekräftig oder lediglich zufällig betrachtet werden können. Wichtig ist auch die Bildung von Modellen, die aus Basisinformationen die Ableitung oder Prognose anderer Informationen erlaubt. Zentrale Voraussetzung für die Anwendung statistischer Methoden ist das Vorhandensein von *Hypothesen* oder *Vermutungen*, die mit den statistischen Methoden geprüft werden können. So eignen sich statistische Methoden auch, Resultate aus einem Data-Mining-Projekt zu validieren.

***Data Mining***

Den wesentlichen Schritt zum schnellen Entdecken von Mustern stellen aber die heutigen *AI-Technologien (Artificial Intelligence)* dar. Sie erlauben es, weitgehend *automatisch* die eigenen Daten auf das Vorhandensein von Mustern zu untersuchen. Zu erwähnen sind im wesentlichen *Regelinduktion* und *Neuronale Netzwerke*. Eine wichtige Eigenschaft dieser Technologien ist es, dass sie oft *direkt umsetzbare Resultate* erbringen. Deshalb wird unter Data Mining oft vor allem die Anwendung dieser Technologien verstanden. Diese Technologien erfordern nicht, dass eine Hypothese formuliert wird, sie erfordern aber, dass eine *klare Fragestellung* vorliegt.

**BERICHT ZUR SVD-ARBEITSTAGUNG DATA MINING:**
**vgl. Seiten 19 und 20 in diesem Heft**

## WELCHE TECHNOLOGIEN STEHEN HINTER DATA MINING?

*Data Mining High End Tools* bedienen sich moderner *Algorithmen*, die aus dem Bereich der künstlichen Intelligenz stammen. Sie arbeiten in einem hohen Masse automatisch und sind von Benutzereingriffen relativ unabhängig. Das heisst, diese Algorithmen können auch *neue, unvermutete Muster* in den Daten finden. Allerdings ist das Vorhandensein einer klaren Fragestellung unabdingbar.

***Neuronale Netzwerke***

*Neuronale Netzwerke* sind Programme, die ein vereinfachtes Modell eines Gehirns darstellen. Basis eines Neuronalen Netzwerkes ist ein *«Neuron»*, was als einfache Summiereinheit verstanden werden kann. In einem Neuronalen Netzwerk sind nun viele dieser Neuronen verknüpft, und die Verknüpfungen sind *gewichtet*. Ein Neuronales Netzwerk *«lernt»* die Struktur eines Datensatzes durch Anpassung der Gewichte und kann das gelernte Wissen dann an einem neuen Datensatz anwenden. Man kann sich die Wirkungsweise eines Neuronalen Netzwerkes an einem Beispiel vor Augen führen: Sie resp. das Neuronale Netzwerk probiert einen Wein. Die Sinnesorgane stellen verschiedene Informationen zur Verfügung: Farbe, Geruch, Geschmack etc. Aufgrund der gespeicherten Kenntnisse entscheidet das Neuronale Netz nun, was das für ein Wein ist – z.B. ein Madeira.

Neuronale Netzwerke sind sehr mächtig und können komplexe Muster «lernen», z.B. das Erkennen eines Gesichtes oder die Identifikation von Unterschriften. Auf der anderen Seite sind sie undurchsichtig: was wirklich in einem Neuronalen Netz passiert und welche Einflussstärke die verschiedenen Faktoren haben, bleibt verborgen.

***Regelinduktion***

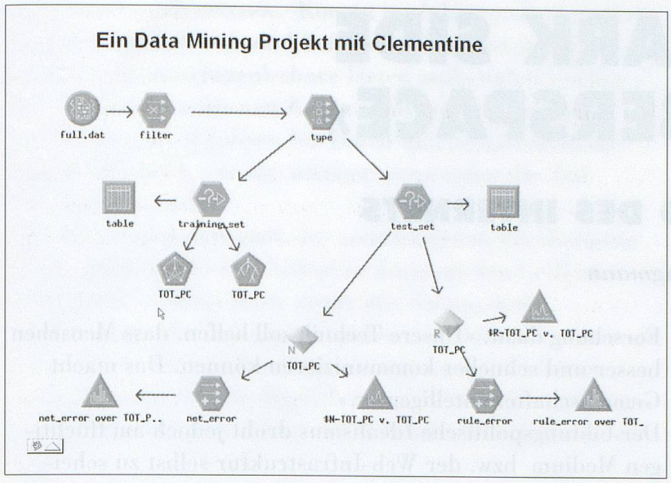
Die *Regelinduktion* stellt einen anderen weitverbreiteten Data-Mining-Algorithmus dar. In der Form eines Baumes (*Decision Tree*) werden Regeln dargestellt, die möglichst präzise zwischen den verschiedenen Fällen unterscheiden sollen. Der Entscheid, welche Informationen in welcher Reihenfolge zum Entscheid benützt werden, erfolgt auf der Basis von Statistik und/oder Informationstheorie. Die Reihenfolge der Auswahl der Informationen im Entscheidungsbaum gibt Auskunft über deren Wichtigkeit für das Modell: Wenn eine Information am Anfang eines Entscheidungsbaumes benützt wird, so heisst dies, dass sie auch entsprechend wichtig ist.

***Welcher Algorithmus ist der wirksamste?***

Neuronale Netzwerke und Regelinduktion benützen unterschiedliche Technologien und haben unterschiedliche Vor- und Nachteile. Demzufolge benutzen Data Mining High End Tools die verschiedenen Ansätze gemeinsam vergleichend oder kombiniert, analog einem Expertenpanel, um zu einem optimierten Resultat zu kommen.

## DATA MINING IN DER PRAXIS

Wer setzt nun aber diese Methoden wirklich ein? Während im anglo-amerikanischen Raum eine eindrucksvolle Liste von Referenzen aus den Bereichen Finanzen, Produktion, Dienstleistung, Telekommunikation, Pharma, Verkauf, Verwaltung, Medien, Consulting etc. bereits zur Verfügung steht (z.B. Reuters, Caterpillar, AirTouch Cellular, Unilever, Daimler Benz, American Century, BBC), scheint sich Data Mining im deutschsprachigen Raum noch am Anfang einer vielversprechenden Entwicklung zu befinden. Welchen Nutzen Data Mining erbringen kann, mag ein Beispiel zeigen: Winterthur Versicherungen Spanien hatte das Ziel, unter den Kunden diejenigen identifizieren zu können, die einerseits mit grosser Wahrscheinlichkeit kündigen würden, auf der anderen Seite aber auch als gute Kunden galten. Mit Hilfe von Regelinduktion konnte aufgrund der vorliegenden Daten über Kündiger und Nichtkündiger ein Modell gebildet werden, das in einem weiteren unabhängigen Datensatz eine Prognosegenauigkeit von mehr als 91% erreichte! Das heisst, dass innerhalb von 12% der Kunden 75% der potentiellen Kündiger identifiziert werden konn-




ten. So konnten potentielle Kündiger gezielt angesprochen und damit ein Teil der Kündigungen verhindert werden.

## EINIGE DATA-MINING-MYTHEN

Begriffe wie Neuronale Netzwerke, Data Mining etc. mögen sehr komplex klingen und den Eindruck erwecken, dass hochspezialisiertes Fachwissen und Hochleistungsrechner nötig seien. Stimmt dies nun wirklich?

***Mythos Nr. 1: Data Mining ist extrem komplex und braucht spezialisierte Experten***
Moderne Softwaretechnologie macht den ganzen Data-Mining-Prozess überschaubar und stellt extrem einfach bedienbare Tools zur Verfügung. So ist nicht mehr IT-Spezialwissen wichtig, sondern viel wichtiger ist es, *die richtigen Fragen zu stellen und über Fachwissen zur Fragestellung zu verfügen*. Auf den Business-Bereich übertragen heisst dies, dass die besten Resultate zustande kommen, wenn Business-Leute Business-Daten «minen». Nur so kann entschieden werden, welche Resultate wirklich von Bedeutung sind – Algorithmen können keine Relevanz einschätzen. So hat beispielsweise das High End Data Mining Tool *Clementine* immer wieder gezeigt, dass erfolgreiches Data Mining nicht nur für Spezialisten möglich ist. Der Data-Mining-Prozess wird visuell mit der Maus durchgeführt und erfordert keine Programmierung. Die Resultate können ebenfalls direkt mit der Maus (*drag and drop*) angewendet werden.

***Mythos Nr. 2: Es gibt den Algorithmus, der überall am besten funktioniert***
In der Praxis ist immer schwer zu entscheiden, welcher Algorithmus konkret die besten Resultate erbringt. Oft erbringt eine Kombination von Algorithmen die besten Resultate – es ist in jedem Falle vorteilhaft, verschiedene Algorithmen zur Verfügung zu haben und auf einfache Art und Weise anwenden zu können. Wichtiger als der gewählte Algorithmus ist, dass Data Mining als *Prozess* verstanden wird, der systematisch durchgeführt werden muss, und zwar nicht nur technisch, sondern auch und vor allem nach *inhaltlichen* Kriterien. Wesentliche Schritte sind dabei:

- *Das Problem verstehen*
- *Die Daten verstehen*
- *Die Daten vorbereiten (Cleaning, Umformung, Integration)*
- *Modelle bilden*
- *Modelle beurteilen*
- *Modell umsetzen*

In diesem Prozess können die verschiedenen Schritte immer wieder von neuem aufgenommen werden, bis befriedigende Resultate erreicht werden.

***Mythos Nr. 3: Der Schlüssel zum Erfolg liegt in der Analyse einer möglichst grossen Datenmenge***
In der Praxis wird Data Mining fast immer an relativ kleinen Stichproben erfolgreich durchgeführt. Auch hier lohnt es sich, *gezielt und intelligent* vorzugehen und nicht einfach brutale Rechengewalt einzusetzen (auch wenn diese vorhanden wäre):

- *Zuerst entscheiden, wonach gesucht wird*
- *Dann vorhandenes Wissen anwenden*
- *Mit einer Stichprobe ein Modell bilden*
- *Das Modell an einer zweiten Stichprobe verifizieren*
- *Schliesslich das verifizierte Modell umsetzen*

Ein grosser Teil erfolgreicher Data-Mining-Produkte ist auf handelsüblichen Personalcomputern durchgeführt worden – obwohl die eigentlich zur Verfügung stehende Datenbasis die Kapazitäten eines solchen Computers gesprengt hätte.

## DATA MINING – AUCH IN DER SCHWEIZ?

Auch in der Schweiz wächst das Interesse an Data Mining zusehends. Verschiedene konkrete Data-Mining-Projekte – vor allem aus dem Marketing-Bereich – wurden schon durchgeführt oder sind am Anlaufen. Data Mining kann überall da erfolgreich eingesetzt werden, wo folgende Voraussetzungen gegeben sind:

- *Eine formulierte Fragestellung*
- *Vorhandene aussagekräftige Daten*
- *Wissen über Fragestellung und Daten und*
- *Wille und Mittel zur Umsetzung der Resultate*

### DER AUTOR

*Josef Schmid*, lic. phil. I, ist Mitglied der Geschäftsleitung von *SPSS (SCHWEIZ) AG* und verantwortlich für die Durchführung von Data-Mining-Projekten.
SPSS (SCHWEIZ) AG ist die Schweizer Vertretung von *SPSS Inc., Chicago*, laut der *META Group* «1997 und 1998 Nr.1 in Data Mining». SPSS Inc. ist seit mehr als 30 Jahren Anbieter von Statistiksoftware und heute Marktführer in Statistik auf dem Desktop mit der Software SPSS. *Clementine* ist ebenfalls ein Produkt von SPSS. SPSS (SCHWEIZ) AG bietet neben der kompletten Software Consulting, Schulung und die Durchführung von Data-Mining-Projekten an.

*contact:*
*SPSS (SCHWEIZ) AG, Seefeldstr. 9, 8008 Zürich*
*Tel.: 01/266 90 30, Fax: 01/266 90 39*
*E-mail: INFO@SPSS.CH.*