Intelligente Systeme zur Gewinnung führungsrelevanter Informationen aus großen Datenmengen – Systematisierung und Bewertung von Data Mining Verfahren

Claudia Heidsieck
Technische Universität Dresden (Claudia.Heidsieck@mailbox.tu-dresden.de)
Betreuer: Prof. Dr. W. Uhr, Technische Universität Dresden

Inhalt

1 Zusammenfassung

In der Arbeit wird der Stand des Zugangs zu führungsrelevanten Informationen in großen Datenmengen erfaßt und mit dem Akzent auf Intelligenz bewertet. Data Mining wird als Methode der Datenanalyse vorgestellt, die über die Fähigkeit verfügt, neues, bisher verborgenes Wissen in den vorhandenen Datenbeständen zu entdecken. Auf eine kurze Darstellung und Einordnung der Verfahren für Data Mining folgt der Aufbau eines Kriterienkatalogs, der die Grundlage für ihre Klassifikation bildet. Die verschiedenen Algorithmen zur Umsetzung der Data Mining Verfahren werden nicht näher betrachtet. Abschließend erfolgt eine beispielhafte Systematisierung von der Analysephase der Data Mining Verfahren.

2 Motivation und Einordnung

Zunehmend sehen sich Unternehmen mit wechselnden Umweltbedingungen konfrontiert, deren Wandel durch die Globalisierung der Märkte und die Verschärfung des Wettbewerbs sowie den Fortschritt in der Informationstechnologie verursacht wird. Information wird zum kritischen Erfolgsfaktor für fundierte Führungsentscheidungen, die ein zielgerichtetes Agieren am Markt ermöglichen und damit den Bestand des Unternehmens absichern. Gleichzeitig wächst die Menge der unternehmensintern sowie der weltweit verfügbaren Daten, was zur Folge hat, daß die Anzahl und Größe von Datenbanken ansteigt. Damit wird die Situation von dem scheinbaren Widerspruch zwischen der steigenden Datenflut und einem gleichzeitigen Informationsdefizit geprägt. Abhilfe schaffen können intelligente Systeme, die aus den großen Datenmengen nur die relevanten Informationen gewinnen, aus denen Nutzen geschaffen werden kann.

Diese Arbeit betrachtet intelligente Systeme zur Gewinnung führungsrelevanter Informationen aus großen Datenmengen. Eine Abgrenzung findet insbesondere von schwach strukturierten und unstrukturierten Daten sowie methodisch von Expertensystemen statt. Mit diesem Hintergrund läßt sich das Thema in das Data-Warehouse Umfeld als Organisationsform für stark strukturierte Daten einordnen.

3 Begriffliche Einordnung

Von intelligenten Systemen wird die Fähigkeit, neue und unerwartete Beziehungen zwischen Variablen und Datenmuster erkennen zu können, gefordert. Die so generierten Informationen sind Entscheidungsinformationen und verringern qualitative und Effizienz-Defizite in der Informationsversorgung.

Knowledge Discovery in Databases (KDD) ist der nicht-triviale Prozeß der Identifikation gültiger, neuer, potentiell nützlicher und schlußendlich verständlicher Muster in (großen) Datenbeständen (FAYYAD u.a. 1996). Über den Begriff des Musters wird deutlich gemacht, daß das Wissen unterschiedliche Formen annehmen kann, wie z. B. Regeln, Assoziationen, Objektgruppierungen oder Entscheidungsbäume. Die Nichttrivialität des Prozesses grenzt ihn gegenüber einfachen Datenbankabfragen ab, die zwar Wissen produzieren, aber keine Entdeckungen im eigentlichen Sinne liefern und hebt den Gesichtspunkt der automatischen Suche nach gültigen Mustern hervor. Durch die Betonung des Prozeßaspekts wird verdeutlicht, daß KDD alle Schritte von der Planung der Analyseaufgabe bis hin zu der Verwendung der Analyseergebnisse in Berichten umfaßt. Im wissenschaftlichen Bereich wird der eigentliche Analyseschritt, in dem Hypothesen gesucht und bewertet werden, als Data Mining bezeichnet. Ebenfalls gebräuchlich ist die Verwendung von Data Mining als Synonym für KDD, die sich überwiegend im kommerziellen Bereich findet.

Der KDD Prozeß besteht aus folgenden Phasen:

- In der Planungsphase findet die Aufgabenfestlegung statt und werden die notwendigen Ressourcen und Ziele bestimmt.
- Die Vorbereitungsphase umfaßt die Selektion der für die Auswertung wichtigen Daten, um zu verhindern, daß durch eine zu große Datenbasis die Effizienz des Analysealgorithmus verringert wird, das Preprocessing mit der Replikation der

Daten aus verschiedenen Systemen z. B. in ein Data Warehouse sowie die Transformation der Daten zur Sicherung der Datenqualität.

- In der Data Mining Phase findet die eigentliche Analyse statt und interessante Muster oder Beziehungen zwischen den Daten werden generiert.
- Die Ergebnisse werden in der Auswertungsphase nach Interessantheit gefiltert.

Data Mining benötigt Tools zur Datenmustererkennung und zum Aufzeigen von Querverbindungen zwischen Variablen und deren Visualisierung. Forschungsschwerpunkte sind momentan die Preprocessing Phase des KDD Prozesses sowie Data Mining Methoden und Verfahren.[1]

4 Wissenschaftliche Zielstellung und Methodik

Das Wissenschaftsziel dieser Arbeit ist die Entwicklung eines Konzeptes für die Systematisierung intelligenter Systeme zur Gewinnung führungsrelevanter Informationen und die Beschreibung und Bewertung der Data Mining Methoden und Verfahren. Parallel zu dem Ansatz aus Verfahrenssicht erfolgt eine Systematisierung, formale Beschreibung und Bewertung der Problemstellungen im betriebswirtschaftlichen Kontext, die sich für den Einsatz von KDD eignen. Darauf aufbauend werden beide Ansätze aufeinander abgebildet, so daß Referenzen aus der Problem- in die Verfahrenssicht und umgekehrt möglich werden. Damit soll das allgemeingültige Kriterium für den Einsatz von KDD, das große Datenmengen und umfangreiche Datenstrukturen vorschreibt, verfeinert werden. Das praktische Ergebnis der Arbeit ist eine Empfehlung für einen differenzierten Einsatz von intelligenten Systemen zur Gewinnung führungsrelevanter Informationen sowie für den Einsatz spezieller Data Mining Verfahren in konkreten Situationen, die eine bestimmte Problemstruktur aufweisen. Dieses Ziel wird über eine Zuordnung von den bewerteten Verfahren und den Problemstrukturen erreicht.

5 Data Mining Verfahren

Im folgenden wird die Data Mining Phase, die das Ziel hat, aussagekräftige Muster in den Daten zu identifizieren, näher betrachtet. Zur Informationsgewinnung findet eine Auswertung über stufenweises Eingrenzen statt. Die Data Mining Verfahren für die Entdeckung und Modellierung neuen Wissens lassen sich über ihre Ziele voneinander abgrenzen.[2]

---

1 vgl. CABENA u.a. 1997, CHAMONI / GLUCHOWSKI u.a. 1998, NAKHAEIZADEH u. a. 1998

2 eine ausführlichere Beschreibung findet sich in CHEN u. a. 1996 , NAKHAEIZADEH u. a. 1998

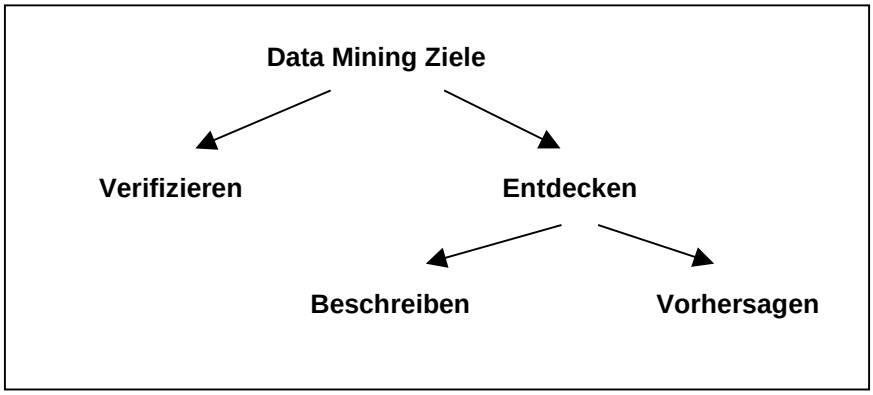

Abbildung 1: Data Mining Ziele

5.1 Datenbeschreibung und -zusammenfassung

Die Datenbeschreibung ermittelt die generellen Eigenschaften einer Teilmenge von Daten, die durch den Benutzer ausgewählt wurde, um ihre Struktur zu verdeutlichen. Bei der mit dieser Methode verwandten Datenzusammenfassung wird das Ziel verfolgt, die Daten in unterschiedlicher Granularität zu präsentieren.[3] Beide Verfahren bewegen sich an der Grenze von Data Mining und streben die Entdeckung einer treffenden Beschreibung der Daten an. Sie werden besonders vor der Anwendung anderer Data Mining Verfahren eingesetzt, um Teilmengen von Daten zu identifizieren, deren genauere Analyse Erfolg verspricht. Außerdem eignen sie sich zur Beschreibung und Zusammenfassung der Data Mining Ergebnisse vor der Präsentation.

5.2 Segmentierung

Die Aufgabe der Segmentierung, die häufig auch als Clusterung bezeichnet wird, ist die Unterteilung der Daten in sinnvolle und interessante Klassen. Verwendet wird sie z. B. für die Kundensegmentierung. Wenn die Anwendung dieses Verfahrens nicht der eigentliche Data Mining Endzweck ist, kann es auch eingesetzt werden, um Datenmengen zu reduzieren oder homogenere Teilmengen zu identifizieren, damit die Auswertung der Daten wie z. B. bei der Warenkorbanalyse vereinfacht wird. Das Ziel der Segmentierung ist eine Beschreibung der Daten über die Bildung von Clustern.

5.3 Klassifikation

Die Klassifikation basiert auf der Segmentierung und verfolgt eines der zentralen Data Mining Ziele. Diese Methode untersucht die Objektbeziehungen innerhalb einer bereits bekannten Klasse, lernt aus den vorgegebenen Beispielen die gewünschten Werte einer Zielfunktion und bildet daraus eine allgemeine Funktionsbeschreibung. Diese Funktionsbeschreibung wird bei neuen Objekten zur Vorhersage des Zielfunktions-

[3] vgl. auch OLAP (On-line Analytical Processing)

werts genutzt. Eine Beispielanwendung dieser Methode ist die Bestimmung von Risiken z. B. bei Krediten.

Mit der Klassifikation eng verwandt ist die Prognose. Diese beiden Methoden unterscheiden sich in dem Wertebereich der vorherzusagenden Daten. Bei der Klassifikation hat die Zielvariable einen symbolischen Wert, wogegen bei der Prognose ein numerischer Wert vorhergesagt wird.

Ebenfalls einen starken Bezug zu der Klassifikation hat die Konzeptbeschreibung. Sie basiert auf der Segmentierung und bei ihr steht das Ziel, eine Beschreibung für vorhandene Klassen zu finden, im Vordergrund und nicht die Einordnung neuer Datenobjekte. Die erzielte Konzeptbeschreibung kann dann wieder zur Klassifikation verwendet werden. Genauso kann die Funktionsbeschreibung, die bei der Klassifikation erzeugt wird, als Konzeptbeschreibung verwendet werden, wenn sie allgemein verständlich ist.

Die Data Mining Intention von Klassifikation und Prognose ist die Vorhersage, wogegen die Konzeptbeschreibung auf die bessere Beschreibung der Analysedaten zielt.

5.4 Abweichungsentdeckung

Die Abweichungsentdeckung findet Muster mit statistisch unerwarteten Ausprägungen und verfolgt damit den Zweck, die Daten zu beschreiben. Die entdeckte Abweichung der Merkmale von einem Erwartungswert kann dabei über der Zeitachse oder über der Gesamtheit der Daten sein.

Bei der Abweichungsanalyse werden wird im allgemeinen das Vorgehen eines menschlichen Analytikers bei der Navigation durch Geschäftsdaten nachgeahmt, indem die Hauptursache für eine Abweichung auf der jeweils nächsten niedrigeren Ebene gesucht wird. Diese Methode eignet sich auch für die Datenbereinigung in der Vorbereitungsphase des KDD Prozesses, da die extrahierten Muster Anomalien z. B. im Konsumerverhalten aber auch in der Datenbasis sein können und da Fehler besonders auffällige Muster produzieren. So führt z. B. das Fehlen von Plandaten zu einer besonders hohen Plan-Ist-Abweichung.[4]

5.5 Abhängigkeitsentdeckung

Die Abhängigkeitsentdeckung modelliert die statistischen Abhängigkeiten zwischen den Merkmalswerten von Objekten und ermöglicht die Vorhersage für die Wahrscheinlichkeit des Auftretens eines bestimmten Wertes. Die Abhängigkeiten werden in Form von Assoziationsregeln dargestellt. Die Ableitung des Risikos einer Lebensversicherung aus Alter und Beruf ist eine mögliche Anwendung der Abhängigkeitsentdeckung.

---

[4] Vgl. BISSANTZ u. a. 1996

5.6 Grafisches Data Mining

Dieses Verfahren nutzt die beim Menschen besonders ausgeprägte Fähigkeit, Muster visuell zu erkennen, indem verschiedene Sichten auf die Daten gegeben werden, aus denen sich der Nutzer das Wissen selbst ableiten muß. Angestrebt wird die Beschreibung der Analysedaten. Es handelt sich nur dann um Grafisches Data Mining, wenn diesem Verfahren keine andere Datenanalyse vorausgeht.

6 Kriterienkatalog

Für die Systematisierung aus Verfahrenssicht wurde ein Kriterienkatalog mit den Aspekten Analyseansatz, Ausgangsdaten, Analyseablauf und Analyseergebnis aufgestellt. Darüber hinaus wird die Kombinierbarkeit der Data Mining Verfahren betrachtet.

6.1 Analyseansatz

Beim Analyseansatz kann es sich um einen benutzergetriebenen handeln, bei dem der Anwender die zu analysierende Datengruppe auswählt und eine scharfe oder unscharfe Anfrage an das System stellt. Neben benutzergetrieben wird in daten- und problemgetrieben unterschieden. Während bei dem problemgetriebenen Ansatz das Analyseziel vorgegeben wird und dementsprechend die Anwendung und das zum Einsatz kommende Verfahren ausgesucht wird, ermittelt das System bei der datengetriebenen Analyse automatisch Auffälligkeiten und generiert Regeln.

6.2 Analysedaten

Die Analysedaten können nach ihren Inhalten in betriebswirtschaftliche und technische Daten unterteilt werden. Unter technischen Daten werden reale oder simulierte Meßwerte von physikalischen Eigenschaften verstanden (z. B. Simulationen des globalen Klimas oder medizinische Untersuchungsergebnisse) wogegen betriebswirtschaftliche Datenbanken Objekte wie Namen, Adressen und Zahlen enthalten. [5] Aus dieser Unterscheidung lassen sich dann die Eigenschaften von Datenvolumen, -haltung und –format ableiten.

Außerdem können die Analysedaten in der Menge, dem Wertebereich der Attribute und in der Verteilung differieren. Der letzte Aspekt beschreibt, ob die Daten in der Ausprägung ihrer Attribute bestimmte Kriterien wie z. B. Gleich- oder Normalverteilungsannahmen erfüllen. Hauptuntersuchungsgegenstand des Data Minings sind sogenannte Ausreißerwerte. Die attributorientierte Induktion zur Generalisation und Zusammenfassung von Daten anhand von Konzepthierarchien ist z.B. eine Methode, die wenig Anforderungen an die Ausgangsdaten stellt und besonders auch für große Datenmengen geeignet ist, einen beliebigen Wertebereich der Attribute unterstützt und relativ robust gegenüber dem Einfluß der Verteilung innerhalb der Daten ist.

---

[5] vgl. zur Problematik von technischen Daten FORTNER 1998

6.3 Analyseablauf

Charakteristika des Analyseablaufs sind die Nutzung von Hintergrundwissen, das notwendig oder nicht notwendig ist oder eingebracht werden kann, sowie die Autonomie der Verfahren, die überwacht oder unüberwacht ablaufen, die intuitive Benutzbarkeit und die Abhängigkeit des Wachstums der Rechenzeit von der Vergrößerung der Datenbasis.

6.4 Analyseergebnis

Wichtigster Aspekt bei der Systematisierung intelligenter Verfahren der Datenanalyse ist das erzielbare Ergebnis. Hier ist entscheidend, ob Vorhersagemöglichkeiten bestehen oder nur eine Musterbeschreibung stattfindet. Außerdem kann ein Zeitbezug vorhanden sein und das Aufzeigen von Entwicklungstendenzen erlauben. Starke Abweichungen ergeben sich aus den Anforderungen der Analyseergebnisse an die Visualisierung. Für den Endanwender sicherlich die bedeutsamste Eigenschaft des Analyseergebnisses ist die Interessantheit, deren Facetten Validität, Neuheit, Nützlichkeit und Verständlichkeit auf objektive Maße abgebildet werden müssen.[6] Eine stärkere Anforderung als die Sicherheit, an der das Resultat von Data Mining gemessen werden muß, ist die Robustheit des extrahierten Wissens. Ergebnisse gelten als robust, wenn es unwahrscheinlich ist, daß sie nach Änderungen in der Datenbasis inkonsistent werden, sie also nicht nur für den aktuellen Zustand der Datenbasis gelten, sondern mit einiger Sicherheit auch noch nach dem Einfügen, Löschen oder Ändern eines Datensatzes.[7]

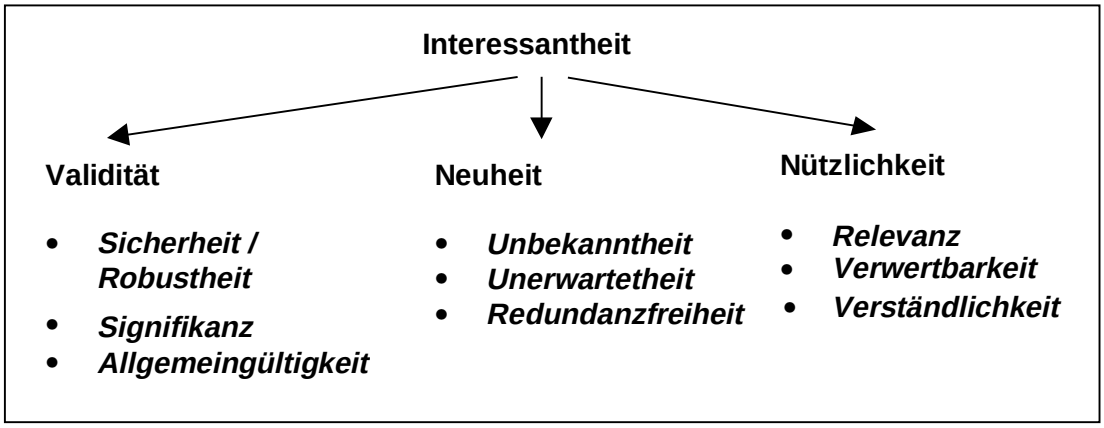

Abbildung 2: Interessantheitsfacetten[6]

[6] vgl. MÜLLER u.a. 1998, SILBERSCHATZ / TUZHILIN 1995

[7] vgl. zur Schätzung von Robustheit HSU / KNOBLOCK 1995

7 Anwendung des Kriterienkatalogs

Zu der Systematisierung der Data Mining Verfahren liegen erste Ergebnisse für die Analysephase vor, die im folgenden in Tabelle 1 dargestellt werden.

Die Autonomie der Segmentierung ist als sehr hoch zu bewerten, weil das Verfahren die Daten vollkommen automatisch einteilt und der Benutzer keinen Einfluß auf die Bildung der Cluster hat. Deshalb spielt die intuitive Benutzbarkeit in diesem Fall keine Rolle. Bei der Segmentierung ohne Hintergrundwissen werden die Cluster anhand statistischer Abstandsmaße eingeteilt. Algorithmen, welche die Nutzung von Hintergrundwissen unterstützen, berücksichtigen Erwartungswerte für die Häufigkeit des Auftretens bestimmter Werte und bilden die Cluster gemäß statistischer Abstandsmaße kombiniert mit Wahrscheinlichkeiten.

Die Klassifikation ist im Gegensatz zu der unüberwachten Segmentierung eine überwachte Data Mining Methode, die aus zwei Phasen besteht. In einer ersten Phase werden die Regeln für die Zuordnung neuer Objekte mit Hilfe von Trainingsdaten gelernt, die dann in der zweiten Phase automatisch auf neue Daten angewendet werden.

	Intuitive Benutzbarkeit	**Autonomie**	**Nutzung von Hintergrundwissen**
Datenbeschreibung und -zusammenfassung	Gut	gering	hoch
Segmentierung	-	hoch	je nach Algorithmus unterschiedlich
Klassifikation	-	mittel, da Trainingsdaten notwendig	gering
Abweichungsentdeckung	-	je nach Algorithmus unterschiedlich	gering
Abhängigkeitsentdeckung	-	hoch	gering
Grafisches Data Mining	sehr gut	sehr gering	gering

Tabelle 1: Analyseablauf

Um die Abweichungsanalyse zu beschleunigen und gezielte Analysen über bestimmte Teilmengen zu ermöglichen, gibt es Algorithmen wie z. B. EXPLORA von Hoschka und Klösgen, die dem Benutzer erlauben, Hypothesen einzugeben, die dann vom System getestet werden. (BISSANTZ u.a. 1996)

8 Weitere Schritte

Nachdem die Zusammenstellung der zu betrachtenden Kriterien beendet ist, findet nun die Systematisierung Data Mining Verfahren statt. Dabei soll zur feineren Klassifikation über die Ebene der Verfahren hinaus auch deren exemplarische Realisierung durch Algorithmen betrachtet werden. Die Systematisierung und Bewertung der Problemstellungen, die sich für den Einsatz von KDD eignen, muß noch erfolgen.

Literaturverzeichnis

BISSANTZ u.a. 1996 Bissantz, N. / Hagedorn, J. / Mertens, P.: Data-Mining als Komponente eines Data-Warehouses; in: Mucksch, H. / Behme, W. (Hrsg.): Das Data-Warehouse-Konzept; Gabler Verlag, Wiesbaden 1996

CABENA u.a. 1997 Cabena, P. / Hadjinian, P. / Stadler, R. / Verhees, J. / Zanasi, A.: Discovering Data Mining: From Concept to Implementation; Prentice Hall PTR, Upper Saddle River 1997

CHAMONI / GLUCHOWSKI u.a. 1998 Chamoni, P. / Gluchowski, P. (Hrsg.): Analytische Informationssysteme: Data Warehouse, On-Line Analytical Processing, Data Mining; Springer Verlag, Heidelberg 1998

CHEN u. a. 1996 Chen, M.-S. / Han, J. / Yu, P. S.: Data Mining: An Overview from Database Perspective; IBM Research Division, New York 1996

FAYYAD u.a. 1996 Fayyad, U. M. / Piatetsky-Shapiro, G. / Smyth. P.: 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 / MIT Press, Menlo Park 1996

FORTNER 1998 Fortner, B.: The technical data crisis: The untold story of write-only data; in: scientific data management, September 1998

HSU / KNOBLOCK 1995 Hsu, C.-N. / Knoblock, C. A.: Estimating the Robustness of Discovered Knowledge; in: Fayyad, U. M. / Uthurusamy, R. (Hrsg.): Proceedings: The First International Conference on Knowledge Discovery & Data Mining; AAAI Press, Menlo Park 1995

MÜLLER u. a. 1998 Müller, M. / Hausdorf, C. / Schneeberger, J.: Zur Interessantheit bei der Entdeckung von Wissen in Datenbanken; in: NAKHAEIZADEH u. a. 1998

NAKHAEIZADEH u. a. 1998 Nakhaeizadeh, G. (Hrsg.): Data Mining: Theoretische Aspekte und Anwendungen; Physica-Verlag, Heidelberg 1998

SILBERSCHATZ / TUZHILIN 1995 Silberschatz, A. / Tuzhilin, A.: On Subjective Measures of Interestingness in Knowledge Discovery; in: Fayyad, U. M. / Uthurusamy, R. (Hrsg.): Proceedings: The First International Conference on Knowledge Discovery & Data Mining; AAAI Press, Menlo Park 1995