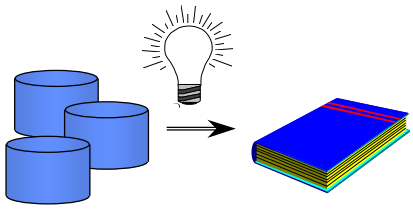
Data Mining
Anwendungen und Techniken

Knut Hinkelmann
DFKI GmbH

Entdecken von Wissen in Datenbanken

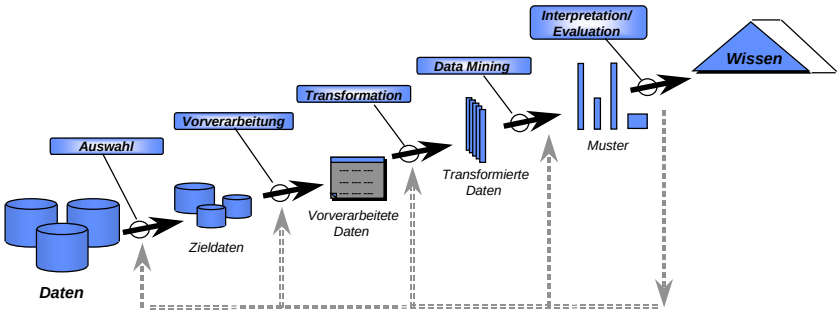


- Unternehmen sammeln ungeheure Datenmengen
- Daten enthalten wettbewerbs-relevantes Wissen
- Ziel: Entdecken dieses verborgenen Wissens
- Allgemeine Kriterien an das zu findende Wissen
  - nicht-trivial
  - bisher unbekannt
  - potentiell nützlich

Beispiel aus der Automobilindustrie

- 7 - 10 Jahre Historie für ca. 7.000.000 Fahrzeuge
  - Fahrzeugdaten (Produktionsdaten; Daten über Motor, Getriebe, ...)
  - Beanstandungen (Schadensteil, Schadensart, ...)
  - Werkstattaufenthalte
- Frage: Wie kann man das Auto zuverlässiger machen?
- Suche in Datenbank nach möglichen Gründen für Ausfälle
- mögliche Umsetzung des Wissens:
  - Änderung in Konstruktion
  - Wechsel des Zulieferers
  - Kundendienst: vorbeugende Wartung
  - usw.

Data Mining ist eine Phase im Prozess der Wissensentdeckung aus Datenbanken



Forschungsgruppe Wissensmanagement
DFKI GmbH, 8/97
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

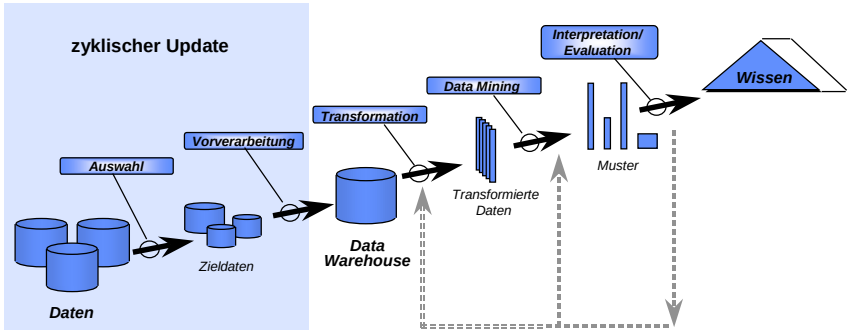
Prozeßschritte der Wissensentdeckung

- Anforderungsanalyse
  - Ziele, Kriterien festlegen
- Auswahl
  - Fokussierung, Auswahl relevanter Daten
- Vorverarbeitung
  - Bereinigung der Daten
- Transformation
  - Anzahl der Variablen reduzieren, Datenformat vereinheitlichen
- Data Mining
  - Auswahl von Techniken und Methoden
  - evtl. viele Testläufe mit verschiedenen Parametern
- Interpretation/Evaluierung
  - Beurteilung der Ergebnisse bzgl. festgelegter Kriterien
  - Dokumentation, Visualisierung der Ergebnisse
  - Überführung in die Anwendung

Probleme beim Data Mining

- Datenbank muß gut organisiert sein
- Ungeeignete Daten:
  - inkonsistente Einträge
  - redundante Einträge
  - Lücken in den Daten
  - Freitextfelder mit relevanten Informationen
- Daten müssen zugreifbar sein
- Einheitliches Format der Daten
- Lösungen:
  - Bereinigung der Daten, Fokussierung
  - evtl. Vereinheitlichung der Daten in einem Data Warehouse

Data Warehouse und Wissensentdeckung in Datenbanken



Mögliche Unterscheidung von Data-Mining-Verfahren

- Einsatz des Systems
  - Überprüfung einer Hypotheses des Benutzers
  - Autonomes Entdecken von Regeln bzw. Mustern
- Techniken, z.B.
  - Maschinelles Lernen
  - Statistik
  - Datenvisualisierung
- Art des zu erkennenden Wissens

Erkennung verschiedener Arten von Wissen

- Abhängigkeiten
  - Lernen von Abhängigkeiten zwischen Variablen
- Klassifikation
  - Lernen einer Menge von Regeln, die Objekte aufgrund ihrer Attribute vorgegebenen Klassen zuordnen
- Clustering
  - Zusammenfassung ähnlicher Objekte
- Abweichungen
  - Entdecken der signifikantesten Abweichungen von vorherigen oder normalen Werten
- Zeitreihenanalyse
  - Lernen von Regeln zeitlicher Veränderungen

Entdecken von Abhängigkeiten

- Gegeben ist eine Menge von Datensätzen
- Lernen von Regeln der Form $A \Rightarrow B$:
  - A und B sind Mengen von Attributwerten
  - Bedeutung der Regel:
    "Wenn ein Datensatz die Attributwerte A_1 und ... und A_n enthält, dann enthält er auch die Werte B_1 und ... und B_m"
- Beispiel:
  - Gegeben eine Kunden-Datenbasis, z.B. Versicherungen
  - Gesucht sind Regeln der Art: Leben ⇒ Unfall
    - Wenn ein Kunde eine Lebensversicherung abschließt, dann schließt er oft auch eine Unfallversicherung ab
  - Anwendung der Regeln:
    Entwicklung gezielter Verkaufsmethoden (Cross-Selling)

Beurteilung der Güte einer Abhängigkeitsregel

- Angenommen man hat eine Regel $X \Rightarrow Y$ gelernt.
- *Konfidenz* bezeichnet die Stärke einer Regel:
  - Eine Regel $X \Rightarrow Y$ hat Konfidenz c, wenn c% der Datensätze in D, die X enthalten, auch Y enthalten.
- *Unterstützung* bezeichnet die Häufigkeit der Muster:
  - Eine Regel $X \Rightarrow Y$ hat Unterstützung s, wenn für s% der Datensätze in D sowohl X als auch Y gilt
- Interessant sind Regeln mit *hoher* Konfidenz und *großer* Unterstützung
- Wichtig: Unabhängigkeit der Attribute:
  - Beispiel: 75% aller Studenten essen Cerealien,
    60% aller Studenten spielen Basketball,
    40% essen Cerealien und spielen Basketball
  - Obwohl die Regel *Basketball* $\Rightarrow$ *Cerealien* hohe Konfidenz (66%) hat, ist sie irreführend: der Anteil der Cerealien-Esser an der Gesamtheit aller Studenten ist höher als der an Basketballspielern.

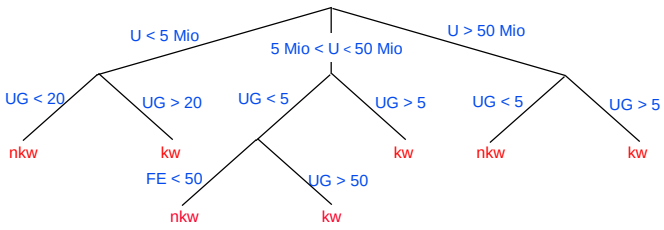
Lernen von Klassifikationsregeln

- Gegeben ist eine Menge von Datensätzen
- Jedem Datensatz ist eine Klasse zugeordnet
- Welche Attributwerte bestimmen die Klassenzugehörigkeit?
- Vorgehensweise: Aufteilung der Daten in zwei Teilmengen
  - *Trainingsmenge* zum Lernen der Regeln
  - *Testmenge* zum Überprüfen der gelernten Regeln:
    - Eine Menge von Klassifikationsregeln ist gut, wenn ein gewisser Anteil der Datensatze korrekt klassifiziert wird

Beispiel für Klassifikation: Kreditwürdigkeitsprüfung

- Gegeben ist eine Menge von Daten über eine Firma:
  - Bilanzdaten, z.B.
    - U: Umsatz
    - G: Gewinn
    - E: Eigenkapital
    - F: Fremdkapital
  - Kennzahlen, z.B.
    - UG: Gewinnanteil am Umsatz
    - FE: Fremdkapital/Eigenkapital
- Jeder Datensatz ist einer von zwei Klassen zugeordnet:
  - kw: "kreditwürdig"
  - nkw: "nicht-kreditwürdig"
- Welche Attributwerte bestimmen, daß eine Firma kreditwürdig ist?

Ergebnisdarstellung: Entscheidungsbaum

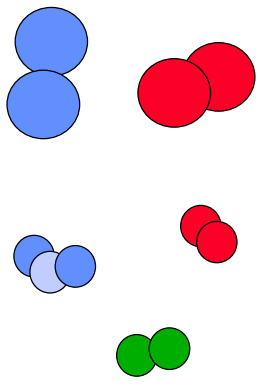


- Ein Entscheidungsbaum ist eine graphische Darstellung einer Menge von Regeln:
  - die Knoten im Baum entsprechen Entscheidungen
  - an den Wurzeln sind Klassen
- Beispiel: *U < 5 Mio und UG < 5 ⇒ nkw*
  - "Wenn der Umsatz kleiner als 5 Mio DM ist und der Gewinn weniger als 20% des Umsatzes beträgt, dann ist die Firma nicht kreditwürdig"

Techniken für Klassifikation

- Entscheidungsbaumverfahren
- Maschinelles Lernen, Regelinduktion
- Neuronale Netze
- Fuzzy-Regeln
- statistische Verfahren
- Fallbasiertes Schließen

Clustering: Lernen von Klassen



- Clustering teilt Datensätze/Objekte aufgrund von Ähnlichkeiten in Klassen (Cluster) ein:
  - ähnliche Datensätze bilden eine Klasse
    - möglichst große Ähnlichkeit mit anderen Objekten der Klasse
    - möglichst geringe Ähnlichkeit mit Objekten außerhalb der Klasse
- Voraussetzung: Definition von Ähnlichkeit zwischen Datensätzen
- Beispiel:
  - Gegeben: Datenbank mit Verkaufszahlen
  - Gesucht: Typische Kundenprofile
  - Verwendung: Entwicklung gezielter Marketingstrategien

Entdecken von Abweichungen

- Erkennen von Abweichungen von normalen Werten
- Typische Muster von Abweichungen sind bekannt
- Gesucht sind Vorkommen dieser Muster in einer Menge von Datensätzen
- Die Muster müssen nicht exakt vorkommen, so daß ein Ähnlichkeitsmaß definiert werden muß
- Beispiel:
  - Erkennen von Kreditkartenbetrug

Zeitreihenanalyse

- Lernen zeitlicher Verläufe aus bekannten Entwicklungen
- Anwendung: Prognose zeitlicher Veränderungen
- Beispiele:
  - Kontenentwicklung, Liquiditätsverlauf, Aktienkurse
  - Analyse von Messdaten zur Vorbeugung von Systemausfällen

Zusammenfassung

- Data Mining ist Teil eines komplexeren Prozesses: Wissensentdeckung aus Datenbanken
- Es gibt keine universellen Data Mining Algorithmus. Die Wahl der Techniken hängt ab von
  - dem zu lernenden Wissen (Klassifikator, Abhängigkeiten, ...)
  - der Beschaffenheit der Daten
- Daten müssen gut organisiert sein
- Data Mining setzt Wissen über die Anwendung voraus. Ein großer Aufwand besteht in der richtigen Formulierung des Problems