Ludwig-Maximilians-Universität München
Institut für Informatik
Lehr- und Forschungseinheit für Datenbanksysteme

Skript

Knowledge Discovery in Databases II

Winter Semester 2013/2014

Kapitel 1: Einleitung und Überblick

Vorlesung: PD Dr Matthias Schubert
Übungen: Erich Schubert

http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_II_(KDD_II)

Organisatorisches

- **Veranstaltungstermine**
  - Vorlesung: Dienstag, 14:00-17:00, Raum A120 (Hauptgebäude)
  - Übungen: Donnerstag, 16:00-18:00, Raum U127 (Oettingenstr. 67)
    Freitag, 12:00-14:00, Raum U151 (Oettingenstr. 67)
  - Alle Informationen zur Vorlesung finden Sie auf:
    http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_II_(KDD_II)

- **Klausur**
  - Anmeldung zur Klausur:
    https://uniworx.ifi.lmu.de/?action=uniworxCourseWelcome&id=91
  - Die Klausur dauert 90 min und zählt für 6 ECTS Punkte.
  - In der Klausur wird der Stoff geprüft, der in der Vorlesung und den Übungen besprochen wurde:
    Das zur Vorlesung erhältliche Skript ist lediglich als Lernhilfe zu verstehen. Der Besuch der Übung ist nicht hinreichend zur Vorbereitung.

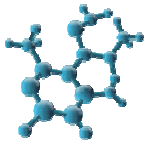
Inhalt des Kapitels

- Knowledge Discovery in Databases und Data Mining
- Standard Data Mining Techniken (Wiederholung der Inhalte von KDD I)
- Überblick über die Inhalte von KDD II
- Lerninhalte und weiterführende Literatur

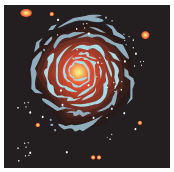
Knowledge Discovery in Datenbanken

Große Mengen an komplexen Datenobjekten

Verbindungsdaten/ Orte/

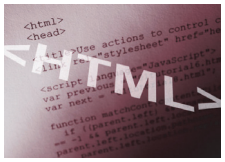


Moleküle/ Prozeßdaten



Telekop-Daten

Zahlungsdaten



Webdaten/ Clickstreams

Manuelle Analyse zu aufwendig !!

→ **Knowledge Discovery in Datenbanken und Data Mining**

Ziel: - Deskriptive Muster: Warum verhalten sich die Daten so ?
(Explizite Beschreibung von Beobachtungen in Regeln)
- Praeskriptive Muster: Wie werden sich Daten verhalten ?
(Vorhersage der Objektklasse und des Objektverhaltens)

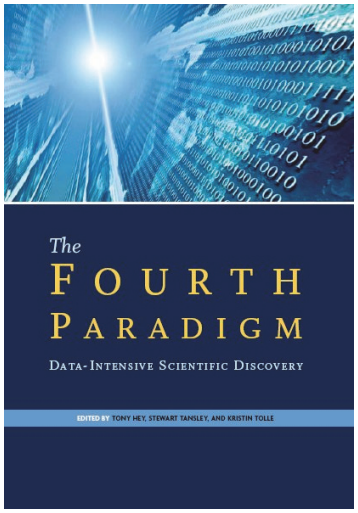
WICHTIG: Gefundene Muster sind selten allgemeingültig sondern meist nur häufig gültig.

The Fourth Paradigm

1. Vor Eintausend Jahren: Experimentelle Wissenschaft
   - Beschreibung natürlicher Phänomene
2. In den letzten Jahrhunderten: Theoretische Wissenschaft
   - Newton'sche Gesetze, Maxwell-Gleichungen, …
3. Die letzten Jahrzehnte: Computergestützte Wissenschaft
   - Simulation komplexer Phänomene
4. Heute: Daten-Intensive Wissenschaft
   - Vereinigt Theorie, Experimente und Simulation

"Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets."

-The Fourth Paradigm - Microsoft



Big Data

Neue Herausforderung für Wissenschaftler aller Bereiche: Der Umgang mit riesigen Datenmengen

Auszug aus dem McKinsey Report *"Big data: The next frontier for innovation, competition, and productivity"*, Juni 2011:

Capturing the value of Big data:

- 300Mrd USD potentieller Wert für das amerikanische Gesundheitswesen, jährlich.
- 250Mrd Euro potentieller Wert für den öffentlichen Dienst in Europa, jährlich.
- 600Mrd USD potentieller Wert durch die Verwendung von location based services.

"The United States alone faces a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts to analyze big data and make decisions based on their findings."

-

www.mckinsey.com/mgi/publications/big_data/

Definition KDD

[Fayyad, Piatetsky-Shapiro & Smyth 1996]

Knowledge Discovery in Databases (*KDD*) ist der Prozess der (semi-) automatischen Extraktion von Wissen aus Datenbanken, das

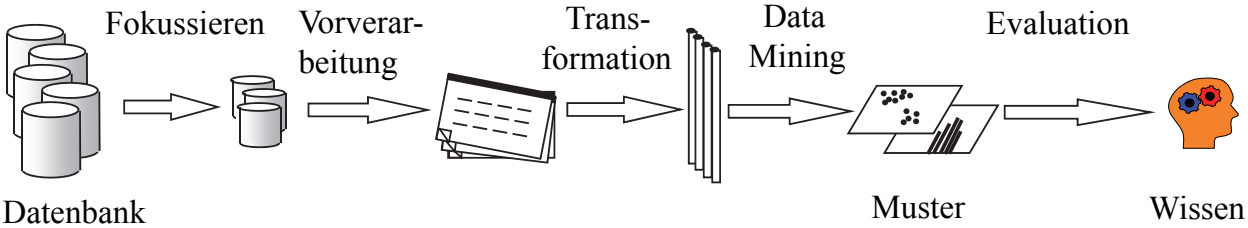
- *gültig*
- *bisher unbekannt*
- und *potentiell nützlich* ist.

Bemerkungen:

- (*semi-*) *automatisch*: im Unterschied zu manueller Analyse. Häufig ist trotzdem Interaktion mit dem Benutzer nötig.
- *gültig*: im statistischen Sinn.
- *bisher unbekannt*: bisher nicht explizit, kein „Allgemeinwissen".
- *potentiell nützlich*: für eine gegebene Anwendung.

Das KDD-Prozessmodell

Prozessmodell nach Fayyad, Piatetsky-Shapiro & Smyth



Fokussieren:
- Beschaffung der Daten
- Verwaltung (File/DB)
- Selektion relevanter Daten

Vorverarbeitung:
- Integration von Daten aus unterschiedlichen Quellen
- Vervollständigung
- Konsistenzprüfung

Transformation
- Diskretisierung numerischer Merkmale
- Ableitung neuer Merkmale
- Selektion relevanter Merkm.

Data Mining
- Generierung der Muster bzw. Modelle

Evaluation
- Bewertung der Interessantheit durch den Benutzer
- Validierung: Statistische Prüfung der Modelle

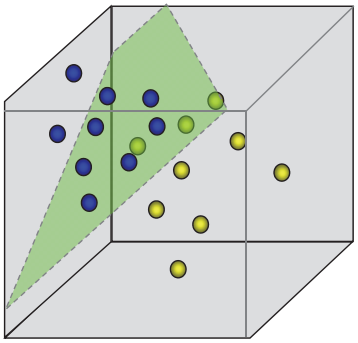
Inhalt des Kapitels

- Knowledge Discovery in Databases und Data Mining
- Standard Data Mining Techniken (Wiederholung der Inhalte von KDD I)
- Überblick über die Inhalte von KDD II
- Lerninhalte und weiterführende Literatur

Inhalte KDD I

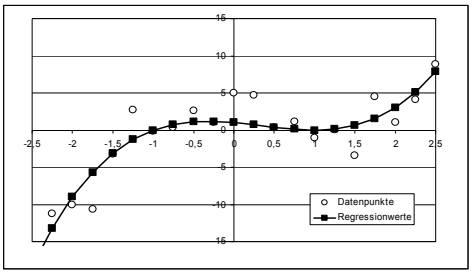
- Clustering
  partitionierendes, agglomeratives, dichte-basiertes Clustering usw.
- Outlier Detection
- Klassifikation
  NN-Klassifikation, Bayes-Verfahren, SVM, Entscheidungsbäume
- Assoziationsregeln
  (Pattern Mining)
- Regression

Klassifikation und Regression



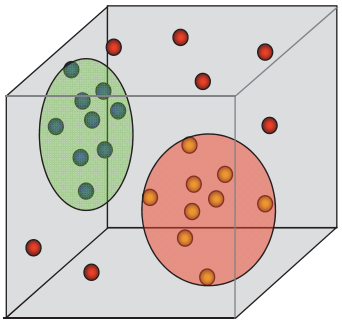
Klassifikation: (supervised learning)
Lerne anhand von Beispielen, wie sich verschiedene Klassen von Objekten unterscheiden.
⇒ Bestimmung der Klasse neuer Objekte
⇒ Erkenne Charakteristika der einzelnen Klassen

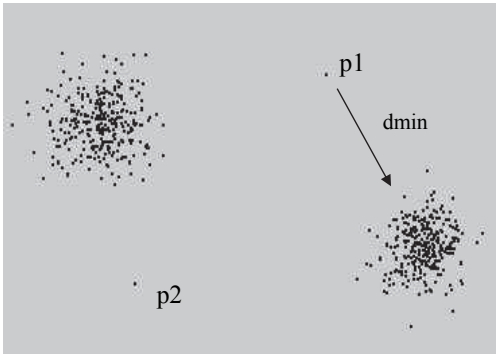
Regression: (supervised learning)
Lerne anhand von Beispielen, wie sich verschiedene Objekte auf eine reelle Zielvariable abbilden lassen:
⇒ Bestimmung der Zielvariable für neue Objekte
⇒ Erkenne Zusammenhang zwischen Objekten und Zielvariable



Clustering und Outlier Detection

Clustering: (Unsupervised Learning)
Finde Objektgruppen (Cluster) in der Datenbank, so dass:
⇒ Objekte des gleichen Clusters ähnlich
⇒ Objekte unterschiedlicher Cluster unähnlich





Outlier Detection:
Finde Objekte, die nicht durch die in einer Datenmenge bekannte Mechanismen erklärbar sind:
⇒ Lerne Outlier (Supervised)
⇒ Finde Outlier (z. B. über Distanzen) (Unsupervised)

Assoziationregeln

TransaktionsID	Items
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Assoziationsregeln:
Finde Regeln über die Elemente in großen Transaktionsdatenbanken.
(Transaktion = Teilmenge aller möglichen Mengenelemente)

⇒ finde häufig zusammen auftretende Elemente

⇒ finde Regeln der Form:
*Wenn **A** in Transaktion **T** enthalten ist, dann ist auch **B** mit Wahrscheinlichkeit **x%** in **T** enthalten.*

Inhalt des Kapitels

- Knowledge Discovery in Databases und Data Mining
- Standard Data Mining Techniken (Wiederholung der Inhalte von KDD I)
- Überblick über die Inhalte von KDD II
- Lerninhalte und weiterführende Literatur

Inhalte von KDD II

- Einleitung

Teil1: Hochdimensionale Daten

- Feature-Selektion
- Feature Reduction und Distance Learning
- Subspace Clustering

Teil 2: Mining in Strukturierten Daten

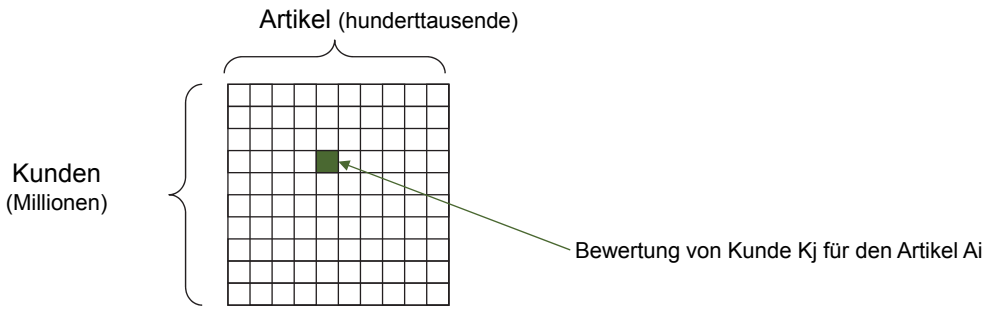
- Ensemble learning und Multi-View Daten
- Multi-Instance Data Mining
- Link Mining und Graph Mining

Teil 3: Big Data

- Sampling, Microclustering und Approximative Suche
- Distributed Data Mining & Privacy
- Data Mining in Streams

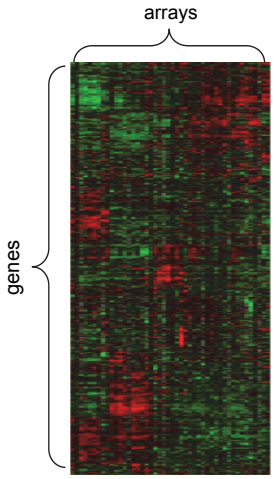
Teil 1: Hochdimensionale Daten - I

- Reale Datenmengen werden häufig sind häufig über sehr viel Attribute beschreibbar.
- Beispiel: Collaborative Filtering
  - Benutzerbewertungen für Artikel (Filme, Songs, …)
  - Logisch bilden diese Daten eine Kunden-Artikel Matrix



Part 1: Hochdimensionale Daten - II

- Beispiel: Micro Array Daten
  - Messung der Genexpression
  - Häufig tausende Gene (Attribute)
  - Aber nur 10-100 Patienten



- Beispiel: Textdaten
  - Einzelne Wörter (Unigramme) oder Wordkombinationen (n-Gramme) als Features
    → Sehr hohe Anzahl potentieller Attribute
  - Ein Text enthält in der Regel sehr viele Worte
  - Darstellung als Vektor der Worthäufigkeiten

Teil 1: Hochdimensionale Daten - III

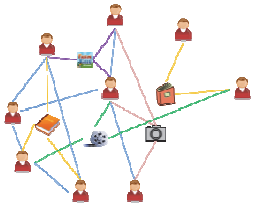
- Herausforderungen für die Analyse hochdimensionaler Daten
  - Distanzfunktionen (für das Clustering, Outlier Detection, …) verlieren ihre diskriminative Aussage in hohen Dimensionen.
    (Curse of Dimensionality)
  - Muster in den Daten können sich auf unterschiedliche Attribute beziehen.
    (Jedes Muster muss in einem eigenen Unterraum betrachtet werden)
- Wir besprechen hier die folgenden Ansätze
  - Featureselektion
  - Dimensionsreduktion
  - Distance Learning
  - Subspace Clustering

Teil 2: Strukturierte Daten I

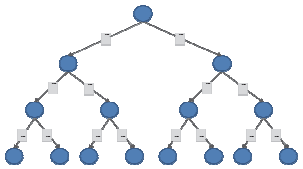
- Standardverfahren betrachten Daten häufig als Vektoren von unabhängigen Merkmalen

  Aber: Daten haben häufig komplexe Strukturen

- **Beispiele**:
  - Graph Daten: Objekte (Knoten) stehen in Beziehung (Kanten) miteinander
    - Soziale Netzwerke (Twitter graph, Facebook graph)
    - Co-Authoren-Netzwerk (DPLP)
    - Protein-Interaktions-Netzwerke



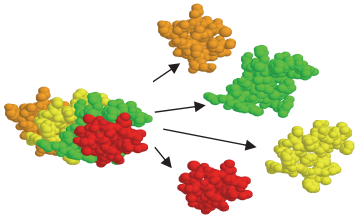
  - Baumstrukturierte Objekte
    - XML Dokumente
    - Sensor Netzwerke



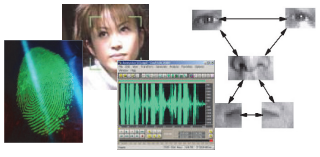
Teil 2: Strukturierte Daten II

- Weitere Arten Strukturierte Objekte
  - Sequenzen:
    - Videos, Audiodaten, Zeitreihen
    - Verhaltensweisen und Bewegunestrajektorien

  - Multi-Instanz Objekte:
    - Teams, Lokale Bilddeskriptoren(SIFT Darstellung)
    - Molekül-Konformationen, Multiple Messdaten



  - Multirepräsentierte Objekte:
    - Bildrepräsentation über Kombination von Form, Farb und Kantenvektoren
    - Proteine über Primär-, Sekundär- und Tertiär-Strukturedeskriptoren



Teil 2: Strukturierte Daten III

- Herausforderungen durch Strukturierte Objekte
  - Kombination von Aussagen aus unterschiedlichen Feature-Räumen
  - Definition von Struktureller Ähnlichkeit
  - Neue Arten von Mustern und Aussagen

- Wir befassen uns in der Vorlesung mit
  - Multi-Instance Data Mining
  - Multi-View Data Mining
  - Link-mining
  - Graph-mining

Teil 3: Big Data - I

- Durch die rasante Entwicklung von Hardware, Infrastruktur und die Entstehung neuer Dienste, können riesige Datenmengen gesammelt werden

- Beispiel: Telekommunikationsanbieter

  - Verbindungsdaten
  - Positionsdaten (über verbundene Sendemasten)
  - IP Addressen/Datenverkehr

- Beispiel: WWW
  - Neue Posts/ Tweets/ Videos ..

- Beispiel : Facebook
  - Neue Benutzer/ Verbindung zwischen Benutzern/ Gruppen
  - Neue Posts/Likes (z.B. Videos, Bilder) / Links auf externe Inhalte

Teil 3: Big Data - II

- Beispiel: Linkedin / XING
  - Neue Benutzer/ Firmen/ Seiten
  - Neue Verbindungen zwischen den Benutzern/ Firmen/ Seiten
  - Neue Interaktionen (e.g., Empfehlungen, Einladungen, Likes)
- Beispiel: Umwelt Monitoring Projekte
  - Sensoren messen Temperatur, Luftfeuchtigkeit, Verschmutzung..
  - Kameras zeigen Bilder von vielen Bereichen der Welt

- Beispiel: Wissenschaftliche Großversuche im CERN
  - Experimente generieren jede Sekunde ein Petabyte Daten
  - "We don't store all the data as that would be impractical. Instead, from the collisions we run, we only keep the few pieces that are of interest, the rare events that occur, which our filters spot and send on over the network," he said.
  - Das CERN speichert jedes Jahr immer noch 25PB an Daten, – das Äquivalent zu 1,000 Jahren Videodaten in DVD Qualität – die von Wissenschaflter untersucht und analysiert warden können um Hinweise für die Struktur des Universums zu finden.
  - http://www.v3.co.uk/v3-uk/news/2081263/cern-experiments-generating-petabyte

Teil 3: Big data - III

- Herausforderungen im Bereich Big Data
  - Lösungen die mit großen Datenmengen flexibel umgehen können.
  - Parallele Berechnung von multiplen Problemen zur gleichen Zeit (Berechnung aller Kundenpräferenzen)
  - Einsatz moderner Hardware (Cloud Computing)
  - Berechnung von Mustern, die nur auf sehr großen Datenmengen signifikant nachgewiesen werden können. (Komplexe Muster)

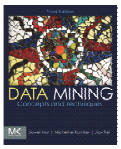
- Wir befassen uns in der Vorlesung mit
  - Approximativen Lösungen und Datenkompression
  - Verteilten und Parallelen Data Mining
  - Privacy Preserving Data Mining
  - Stream Mining

Inhalt des Kapitels

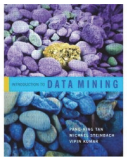
- Knowledge Discovery in Databases und Data Mining
- Standard Data Mining Techniken (Wiederholung der Inhalte von KDD I)
- Überblick über die Inhalte von KDD II
- Lerninhalte und weiterführende Literatur

Empfohlene Begleitliteratur

- Han J., Kamber M., Pei J. (English)
  Data Mining: Concepts and Techniques
  3rd ed., Morgan Kaufmann, 2011



- Tan P.-N., Steinbach M., Kumar V. (English)
  Introduction to Data Mining
  Addison-Wesley, 2006



- Mitchell T. M. (English)
  Machine Learning
  McGraw-Hill, 1997

- Witten I. H., Frank E. (English)
  Data Mining: Practical Machine Learning Tools and Techniques
  Morgan Kaufmann Publishers, 2005

- Ester M., Sander J. (German)
  Knowledge Discovery in Databases: Techniken und Anwendungen
  Springer Verlag, September 2000

Weitere Bücher

- C. M. Bishop, „*Pattern Recognition and Machine Learning*“, Springer 2007.
- S. Chakrabarti, „ *Mining the Web: Statistical Analysis of Hypertext and Semi-Structured Data*”, Morgan Kaufmann, 2002.
- R. O. Duda, P. E. Hart, and D. G. Stork, „*Pattern Classification*“, 2ed., Wiley-Inter-science, 2001.
- D. J. Hand, H. Mannila, and P. Smyth, „*Principles of Data Mining*“, MIT Press, 2001.
- U. Fayyad, G. Piatetsky-Shapiro, P. Smyth: ``*Knowledge discovery and data mining: Towards a unifying framework''*, in: Proc. 2nd ACM Int. Conf. on Knowledge Discovery and Data Mining (KDD), Portland, OR, 1996

Online Resourcen

- *Mining of Massive Datasets* book by Anand Rajaraman and Jeffrey D. Ullman
  - http://infolab.stanford.edu/~ullman/mmds.html
- *Machine Learning* class by Andrew Ng, Stanford
  - http://ml-class.org/
- *Introduction to Databases* class by Jennifer Widom, Stanford
  - http://www.db-class.org/course/auth/welcome
- Kdnuggets: Data Mining and Analytics resources
  - http://www.kdnuggets.com/

KDD /Data Mining tools

- Several options for either commercial or free/ open source tools
  - Check an up to date list at: http://www.kdnuggets.com/software/suites.html
- Commercial tools offered by major vendors
  - e.g., IBM, Microsoft, Oracle …
- Free/ open source tools

Weka

R

SciPy + NumPy

Elki

Orange

Rapid Miner (free, commercial versions)

Lernziele

- Definition KDD und Data Mining
- Der KDD Prozess
- Supervised vs. Unsupervised Learning
- Hauptaufgaben im Data Mining Schritt
  - Clustering
  - Classification
  - Association rules mining
  - Outlier detection
  - Regression