

Skript zur Vorlesung
Knowledge Discovery in Databases II
im Sommersemester 2008

Kapitel 1: Einleitung und Überblick

Skript © 2008 Matthias Schubert

<http://www.dbs.ifi.lmu.de/Lehre/KDD2>

Organisatorisches

Aktuelles

- Vorlesung: Mittwoch, 8:30-11 Uhr, 1.14 (Oettingenstr. 67)
- Übung: Montag, 14-16 Uhr, 0.33 (Oettingenstr. 67)
Montag, 16-18 Uhr, 0.33 (Oettingenstr. 67)
Freitag, 12-14 Uhr, 1.43 (Oettingenstr. 67)

Alle wichtigen Informationen (z.B. Skript) sind im WWW unter:

http://www.dbs.ifi.lmu.de/Lehre/KDD_II/

Scheinerwerb:

In der Abschlussprüfung wird der Stoff geprüft, der in der Vorlesung besprochen wurde.

Das zur Vorlesung erhältliche Skript ist lediglich als Lernhilfe zu verstehen.



Dr. Matthias Schubert

<http://www.dbs.ifi.lmu.de/Mitarbeiter/schubert.html>

Arthur Zimek

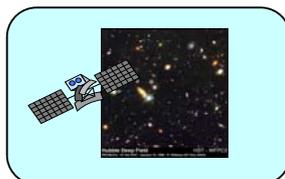
<http://www.dbs.ifi.lmu.de/Mitarbeiter/zimek.html>



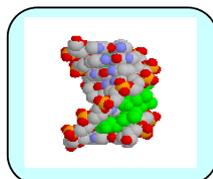
3

Knowledge Discovery in Datenbanken

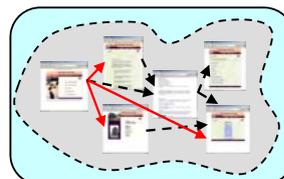
Große Mengen an komplexen Datenobjekten



Space Telescopes



Biomolekül-
Datenbanken



WWW



CAD-Kataloge

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.

4

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.

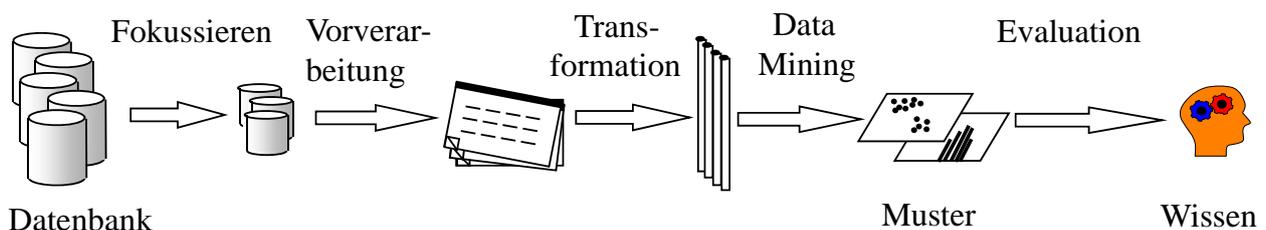
Bemerkungen:

- *(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.

5

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 Merkmale.

Data Mining

- Generierung der Muster bzw. Modelle

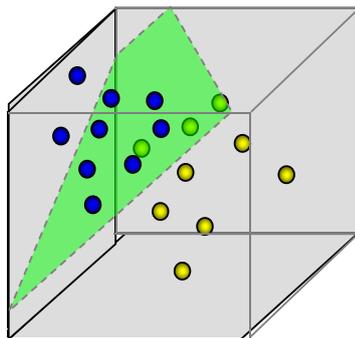
Evaluation

- Bewertung der Interessanzheit durch den Benutzer
- Validierung: Statistische Prüfung der Modelle

6

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

Klassifikation und Regression



Klassifikation: (supervised learning)

Lerne anhand von Beispielen, wie sich verschiedene Klassen von Objekten unterscheiden.

⇒ Bestimmung der Klasse neuer Objekte

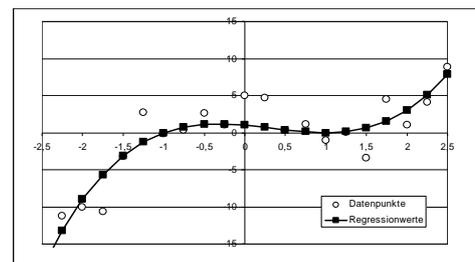
⇒ Erkenne Charakteristika der einzelnen Klassen

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

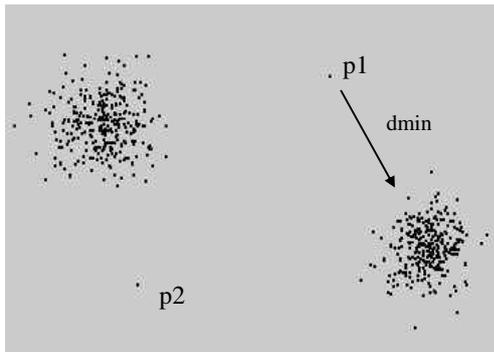
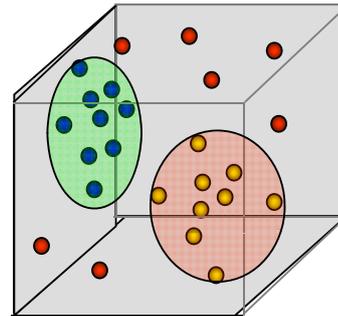


Clustering und Outlier Detection

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 (Supervise)
- ⇒ Finde Outlier über Distanzen (Unsupervised)

9

Assoziationsregeln

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.

Weitere Themen:

- Effizienzsteigerungen bei großen Datenbanken
- Data Warehouses als Systemarchitektur für Data Mining Datenbestände

10

Inhalte KDD II

Weiterführende Methoden des Data Mining:

- Data Mining und Knowledge Discovery ist ein noch junges Gebiet (ca. 12 Jahre)
- in den letzten Jahren, Einzug in die Praxis
 - ⇒ neue Probleme aus der Praxis
- neue Lösungsansätze und Themengebiete
- Grundannahmen der Basisalgorithmen sind häufig verletzt

Basisannahmen:

- Daten alle in einer Datenbank
- alle Features sind potentiell nützlich
- Daten werden durch einen Vektorraum beschrieben (eine DB-Tabelle)
- Datenobjekte unabhängig voneinander

⇒ Neue Lösungsansätze,
die ohne diese Annahmen auskommen

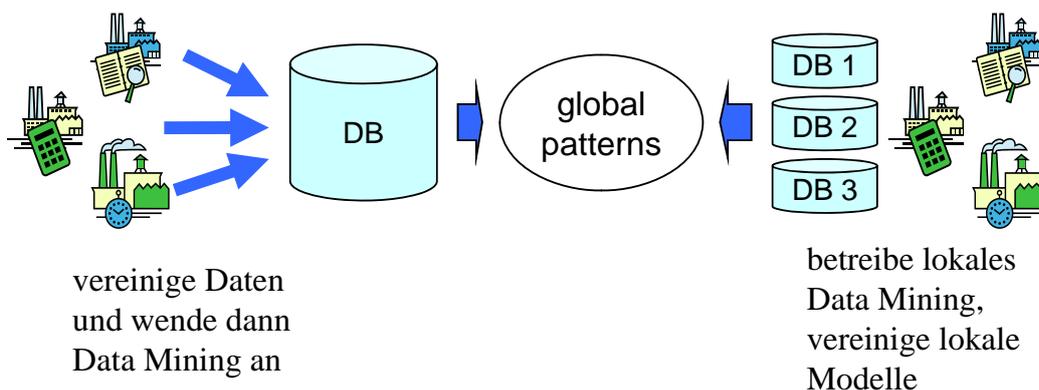
11

Verteilte Datenbestände

Bisher:

Alle Daten sind vollständig in einer Datenbank integriert.

Aber: Verteilte Systeme, Parallele Verarbeitung, Schutz der Privatsphäre bei Data Mining über mehrere Datenquellen hinweg.



12

Verteilte Datenbestände

Problemstellung:

Erst Daten integrieren und dann Data Mining anwenden, ist häufig problematisch:

- Datentransfer teuer
- Update Problematik
- Datenschutz für lokale Informationen

Lösungsansatz:

Suche lokale Muster und kombiniere sie zu globalen Mustern.

Anforderungen:

- Datenschutz für lokale Daten (Privacy Preservation)
- manche globale Muster lassen sich nicht aus lokalen ableiten.
- mit zentraler Instanz (alle Muster werden bei einem Server kombiniert)
- ohne zentral Instanz (lokale Muster werden unter gleich berechtigten Partnern ausgetauscht)

13

Strukturierte Datenobjekte

Bisher:

Daten als Featurevektoren mit moderater Dimensionalität dargestellt

Aber:

In vielen Anwendungen sind Daten durch sehr hochdimensionale Featurevektoren oder durch Mengen, Tupel oder Graphen anderer Repräsentationen beschrieben.



14

Strukturierte Datenobjekte

Problemstellung:

Eigenschaften haben unterschiedlichen Informationsgehalt, beschreiben unterschiedliche Aspekte und haben eine Struktur.

Lösungsansatz:

Finde Data Mining Algorithmen, die mit strukturierten Objekten Arbeiten (Sequenzen, Tupel, Mengen, Graphen..)

Anforderungen:

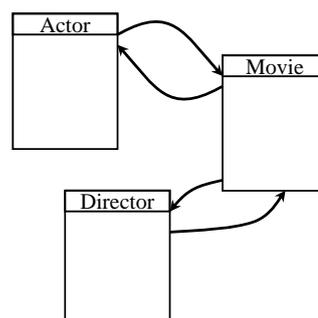
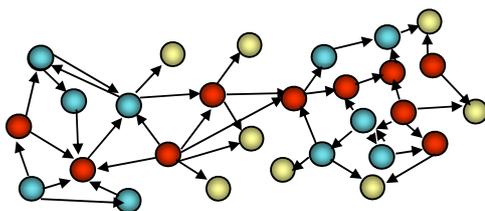
- mehr Information => mehr Vergleichsmöglichkeiten
Welche Information hilft und welche ist überflüssig ?
- Effizienz: Vergleiche von Mengen und Graphen deutlich teurer als Vergleich zwischen Arrays

15

Verlinkte und Relationale Daten

Bisher: „identical and independent distributed Objects“ (IID)
Daten sind unabhängig und können sich nicht gegenseitig beeinflussen.

Aber: Häufig interagieren Objekte in einem Netzwerk.
=> Man muss das System verstehen und nicht nur die Eigenschaften der Einzelobjekte.



16

Verlinkte und Relationale Daten

Problemstellung:

Verhalten der Daten ist nicht unter „iid“-Hypothese erklärbar.
Muster sind nur unter der Berücksichtigung von
Objektinteraktion erklärbar.

Lösung:

Berücksichtigung von interagierenden Datenbeständen.

Aspekte:

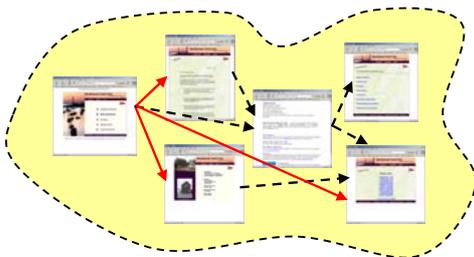
- Modellierung der Datenbank als Interaktionsgraph
- Welche Objekte sind wichtig und welche nicht ?
- Wie wirkt sich Änderung eines Objekts auf alle anderen aus
- Finde Art der Beziehung
(Ursache-Wirkung-Co-Occurence Problematik)

17

Spezielle Anwendungen

Bisher: Allgemeine Verfahren zum Data Mining in Vektoren

Aber: bestimmte Daten z.B. Text- und Web-Daten haben
spezielle Charakteristiken, die für optimale Ergebnisse
berücksichtigt werden sollten.



...	...
Schnee	1
Eis	1
Fahrzeug	1
Politik	0
...	...

18

Spezielle Anwendungen

Problemstellung:

Spezielle Charakteristika der Daten werden nicht ausgenutzt.
Allgemeinheit der Verfahren kann Zeit und Qualität kosten.

Lösungsansatz:

Entwickle spezielle Verfahren, die Eigenschaften der Applikation berücksichtigen.

Aspekte:

- Lösung so allgemein wie möglich halten
- Untersuchung der Unterschiede: Warum sind Textvektoren keine herkömmlichen Feature-Vektoren.
- Welche Aspekte lassen sich ausnutzen und welche nicht?
- Spezielle Systemarchitekturen, die Data Mining verwenden (Focused Crawler und Search Engines)

19

Überblick über die Vorlesung (Momentaner Stand der Planung)

- | | |
|--|---|
| 1. Einleitung | 5. Multirepräsentiertes Data Mining und Ensemble Methoden |
| 2. Feature-Selektion und Feature-Reduktion | 6. Multi-Instanz Data Mining |
| 3. Clustering in hochdimensionalen Räumen | 7. Graphmodelierte Daten |
| 4. Paralleles, Verteiltes und Privacy Preserving Data Mining | 8. Link Mining und Relationales Data Mining |

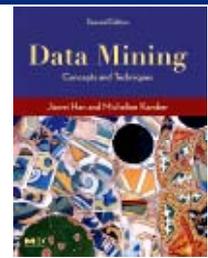
20

Begleitliteratur zur Vorlesung:

Han J., Kamber M.

„*Data Mining: Concepts and Techniques*“

ISBN: 1558609016, Morgan Kaufmann Publishers, March 2006, €54,95



Allgemein Literatur zum Thema Data Mining und KDD:

1. C. M. Bishop, „*Pattern Recognition and Machine Learning*“, Springer 2007.
2. S. Chakrabarti, „*Mining the Web: Statistical Analysis of Hypertext and Semi-Structured Data*“, Morgan Kaufmann, 2002.
3. R. O. Duda, P. E. Hart, and D. G. Stork, „*Pattern Classification*“, 2ed., Wiley-Interscience, 2001.
4. D. J. Hand, H. Mannila, and P. Smyth, „*Principles of Data Mining*“, MIT Press, 2001.
5. T. M. Mitchell, „*Machine Learning*“, McGraw Hill, 1997.
6. P.-N.Tan, M. Steinbach, and V. Kumar, „*Introduction to Data Mining*“, Addison-Wesley, 2006. ISBN: 0-321-32136-7
7. I. H. Witten and E. Frank, „*Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*“, Morgan Kaufmann, 2nd ed., 2005, ISBN 0-12-088407-0
8. 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