Ludwig-Maximilians-Universität München
Lehrstuhl für Datenbanksysteme und Data Mining
Prof. Dr. Thomas Seidl

Knowledge Discovery and Data Mining I

Winter Semester 2018/19

People

Lecturer

- Prof. Dr. Thomas Seidl

Assistants

- Max Berrendorf
- Julian Busch

Student Assistants

- Maximilian Hünemörder
- Florentin Schwarzer

Schedule

Lecture (begins: 16.10.2018)

Tu.	09:15-11:45	B U101 (Oettingenstraße. 67)

Tutorials (begin: 25.10.2018)

Th.	12:15-13:45	Lehrturm-VU107 (Prof.-Huber-Pl. 2)
Th.	14:15-15:45	Lehrturm-VU107 (Prof.-Huber-Pl. 2)
Fr.	12:15-13:45	Lehrturm-V005 (Prof.-Huber-Pl. 2)
Fr.	14:15-15:45	C 111 (Theresienstr. 41)

Exam

1. Hauptklausur:
   Mo., 25.02.19, 14:00-16:00, B 101 B 201 (Hauptgebäude)
2. Nachholklausur:
   tba

Material, Tutorials & Exam

Material (Slides, Exercises, etc.)

Available on course webpage:

http://www.dbs.ifi.lmu.de/cms/studium_lehre/lehre_master/kdd1819/index.html

Tutorial

- Python Introduction now available on website
- First exercise sheet available for download around 18.10.2018
- Prepare at home
- Presentation and discussion one week after

Exam

- Written exam at the end of semester
- All material discussed in the lecture and tutorials
- Registration via UniWorX

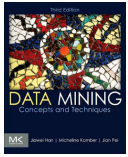
Content of this Course

1. Introduction
   - 1.1 Organisation
   - 1.2 Motivation
   - 1.3 Knowledge Discovery Process
2. Basics
   - 2.1 Data Representation
   - 2.2 Data Reduction
   - 2.3 Visualization
   - 2.4 Privacy
3. Unsupervised Methods
   - 3.1 Frequent Pattern Mining
   - 3.2 Clustering
   - 3.3 Outlier Detection
4. Supervised Methods
   - 4.1 Classification
   - 4.2 Regression
5. Advanced Topics
   - 5.1 Process Mining
   - 5.2 Outlook

Textbook / Acknowledgements

The slides used in this course are modified versions of the copyrighted original slides provided by the authors of the adopted textbooks:

- © Jiawei Han, Micheline Kamber, Jian Pei: *Data Mining – Concepts and Techniques*, 3rd ed., Morgan Kaufmann Publishers, 2011.
  http://www.cs.uiuc.edu/~hanj/bk3
- © Martin Ester and Jörg Sander: *Knowledge Discovery in Databases – Techniken und Anwendungen* Springer Verlag, 2000 (in German).



Motivation

- Data Mining = extraction of patterns from data
- Patterns
  - Regularities – examples: frequent itemsets, clusters
  - Irregularities – examples: outliers
- Not all patterns are useful
  - *"all mothers in our database are female"* $\rightsquigarrow$ trivial/known
  - *"bread, butter is frequent"* given *"bread, butter, salt is frequent"* $\rightsquigarrow$ redundant
- Aggregation of data may help: Basic statistics

What is Data Mining?

Knowledge Discovery in Databases (Data Mining)

Extraction of interesting *(non-trivial, implicit, previously unknown and potentially useful)* information or patterns from data in *large databases*

Roots of Data Mining

- Statistics
- Machine Learning
- Database Systems
- Information Visualization

Data Mining: Motivation

"Necessity is the mother of invention"

Data Explosion Problem

Tremendous amounts of data caused by

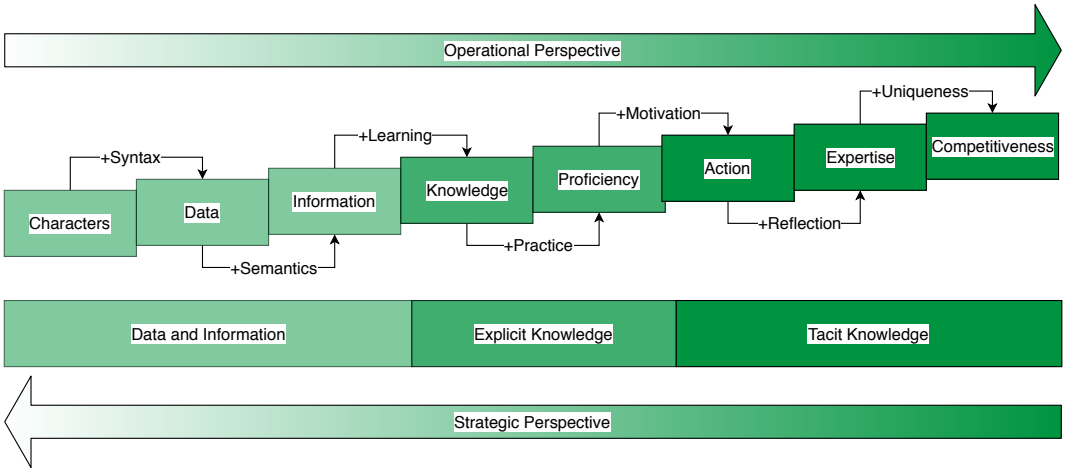
- Automated data collection
- Mature database technology

"We are drowning in data, but starving for knowledge!"

Solution

- Data Warehousing and on-line analytical processing (OLAP)
- Data Mining: Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases

Data Mining: Motivation



Stairs of Knowledge (K. North) [1]

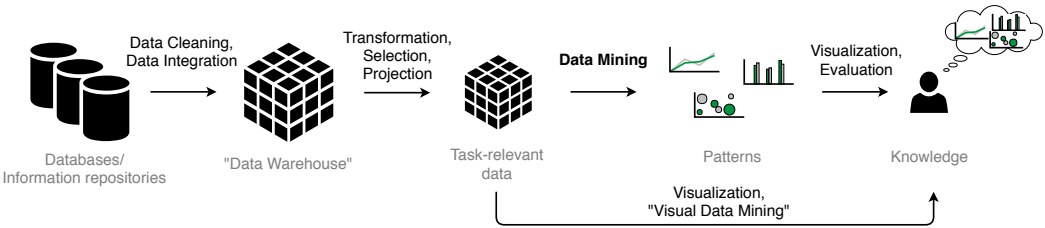
[1] Stairs of Knowledge: North, K.: Wissensorientierte Unternehmensführung - Wertschöpfung durch Wissen. Gabler, Wiesbaden 1998.

Data Mining: Potential Applications

- Database analysis and decision support
  - *Market analysis and management*:
    target marketing, customer relation management, market basket analysis, cross selling, market segmentation
  - *Risk analysis and management*:
    Forecasting, customer retention ("Kundenbindung"), improved underwriting, quality control, competitive analysis
  - *Fraud detection and management*
- Other Applications:
  - Text mining (news group, email, documents) and Web analysis.
  - Intelligent query answering

The Knowledge Discovery Process

- The KDD-Process (Knowledge Discovery in Databases)

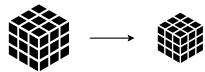


- Data Mining:
  - Frequent Pattern Mining
  - Clustering
  - Classification
  - Regression
  - Process Mining
  - ...

KDD Process: Data Cleaning & Integration

- . . . may take 60% of effort
- Integration of data from different sources
  - Mapping of attribute names, e.g. `C_Nr` $\rightarrow$ `O_Id`
  - Joining different tables, e.g. `Table1 = [C_Nr, Info1]` and `Table2 = [O_Id, Info2]`
    $\rightsquigarrow$ `JoinedTable = [O_Id, Info1, Info2]`
- Elimination of inconsistencies
- Elimination of noise
- Computation of missing values (if necessary and possible): Possible strategies e.g. default value, average value, or application specific computations

KDD Process: Focusing on Task-Relevant Data



Task

- Find useful features, dimensionality/variable reduction, invariant representation
- Creating a target data set

Selections

Select the relevant tuples/rows from the database tables, e.g., sales data for the year 2001

KDD Process: Focusing on Task-Relevant Data

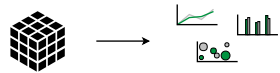
Projections

Select the relevant attributes/columns from the database tables, e.g., (`id`, `name`, `date`, `location`, `amount`) $\rightsquigarrow$ (`id`, `date`, `amount`)

Transformations, e.g.:

- Discretization of numerical attributes, e.g.,
  `amount`: [0, 100] $\rightsquigarrow$ `d_amount`: {low, medium, high}
- Computation of derived tuples/rows and derived attributes:
  - aggregation of sets of tuples, e.g., total amount per months
  - new attributes, e.g., `diff = sales current month - sales previous month`

KDD Process: Basic Data Mining Tasks



Goal

Find patterns of interest

Tasks

- Identify task: Are there labels (in the training data)?
  - *Many* $\rightsquigarrow$ Supervised learning (focus on given concepts)
  - *Some few* $\rightsquigarrow$ Semi-supervised learning (focus on few hidden concepts)
  - *None* $\rightsquigarrow$ Unsupervised learning (many hidden concepts)
- Choose fitting mining algorithm(s)

Basic Mining Tasks: Frequent Itemset Mining

Setting

Given a database of transactions, e.g.

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Motivation

Frequently co-occurring items in the set of transactions indicate correlations or causalities

Examples

- `buys(x, "diapers")`$\Rightarrow$`buys(x, "beers")` [supp: 0.5%, conf: 60%]
- `major(x, "CS")`$\wedge$`takes(x, "DB")`$\Rightarrow$`grade(x,"A")` [supp: 1.0%, conf: 75%]

Basic Mining Tasks: Frequent Itemset Mining

Applications

- Market-basket analysis
- Cross-marketing
- Catalogue design
- Also used as a basis for clustering, classification
- Association rule mining: Determine correlations between different itemsets

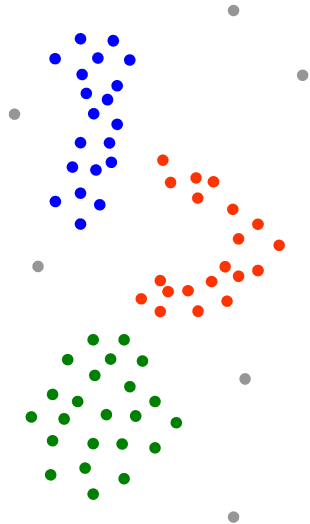
Basic Mining Tasks: Clustering

Setting

- Database of objects
- (Dis-)Similarity function between objects
- Unknown class labels

Task

Group objects into sub-groups (clusters)
"maximizing" intra-class similarity and
"minimizing" interclass similarity



Basic Mining Tasks: Clustering

Applications

- Customer profiling/segmentation
- Document or image collections
- Web access patterns
- ...

Basic Mining Tasks: Classification

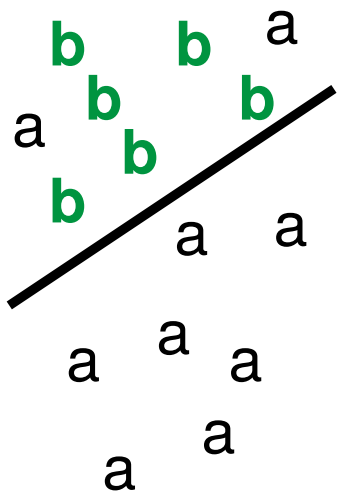
Setting

Class labels are known for a small set of "training data"

Task

Find models/functions/rules (based on attribute values of the training examples) that

- describe and distinguish classes
- predict class membership for "new" objects



Basic Mining Tasks: Classification

Applications

- Classify disease type for tissue samples from gene expression values
- Automatic assignment of categories to large sets of newly observed celestial objects
- Predict unknown or missing values (cf. KDD data cleaning & integration)
- ...

Basic Mining Tasks: Regression

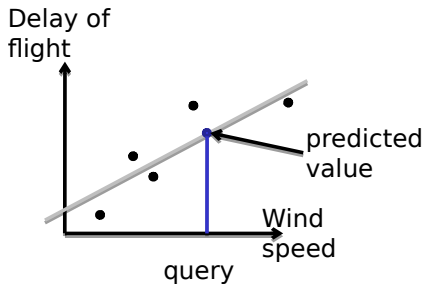
Setting

Numerical output values are known for a small set of "training data"

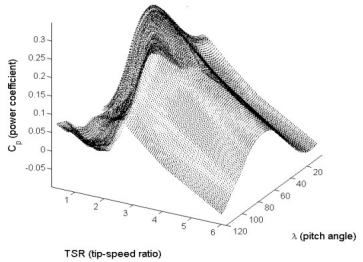
Task

Find models/functions/rules (based on attribute values of the training examples) that

- describe the numerical output values of the training data
- predict the numerical value for "new" objects



Basic Mining Tasks: Regression

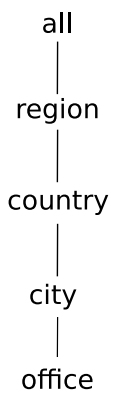


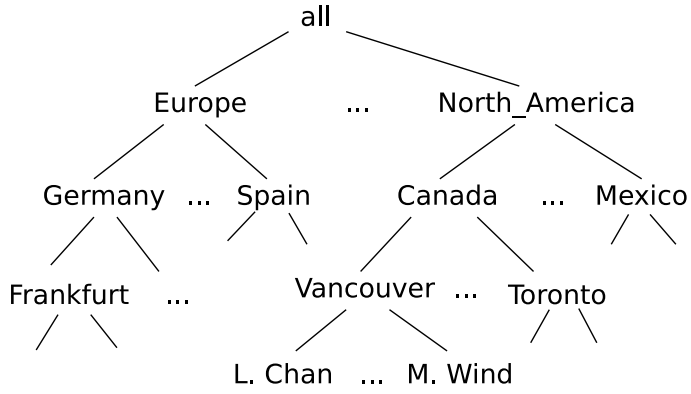
Applications

- Build a model of the housing values, which can be used to predict the price for a house in a certain area
- Build a model of an engineering process as a basis to control a technical system
- ...

Basic Mining Tasks: Generalization Levels

- Generalize, summarize, and contrast data characteristics
- Based on attribute aggregation along concept hierarchies
  - Data cube approach (OLAP)
  - Attribute-oriented induction approach





Basic Mining Tasks: Other Methods

Outlier Detection

Find objects that do not comply with the general behaviour of the data (fraud detection, rare events analysis)

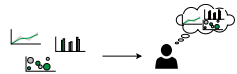
Trends and Evolution Analysis

Sequential patterns (find re-occurring sequences of events)

Methods for special data types, and applications

- Process Mining
- Spatial Data Mining
- Graphs
- ...

KDD Process: Evaluation and Visualization



- Pattern evaluation and knowledge presentation: Visualization, transformation, removing redundant patterns, etc.
- Integration of visualization and data mining:
  - *data* visualization
  - data mining *result* visualization
  - data mining *process* visualization
  - *interactive* visual data mining
- Different types of 2D/3D plots, charts and diagrams are used, e.g. box-plots, trees, scatterplots, parallel coordinates
- Use of discovered knowledge

Summary

- Data mining = Discovering interesting patterns from large amounts of data
- A natural evolution of database technology, machine learning, statistics, visualization, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.

References

	Conference	Journal
Data Mining and KDD	KDD, PKDD, SDM, PAKDD, ICDM, ...	Data Mining and Knowledge Discovery, ...
Database Field	ACM-SIGMOD, ACM-PODS, VLDB, ICDE, EDBT, CIKM, ...	ACM-TODS, J. ACM, IEEE-TKDE, JIIS, VLDBJ, ...
AI and Machine Learning	Machine learning, AAAI, IJCAI, ICLR, ...	Machine Learning, Artificial Intelligence, ...
Statistics	Joint Stat. Meeting, ...	Annals of Statistics, ...
Visualization	CHI (Comp. Human Interaction), ...	IEEE Trans. Visualization and Computer Graphics, ...