Data mining

Extraction of interesting high-level knowledge from large amounts of data
or simply
modern data analysis

- Motivation: data generation grows faster than data understanding
- What could existing data tell us, if we were able to ask the right questions?

Examples of discovered knowledge

Example domain: supermarket customer analysis

- **Regularities in purchases:** "Beer and chips tend to be purchased together."
- **Clusters of customers:** "Singles (small purchases containing half-ready meals), families, families at weekends, ..."
- **Exceptions within customers:** "Mary, although the mother of a large family, makes small purchases."
- **Time series:** "Annual/weekly cycles in the consumption of ice cream."

Application examples

- Customer analysis, direct marketing (What are our customers like? What makes customers happy? How can we better serve our customers?)
- Risk and fraud analysis (Which long distance calls probably are fraudulent? Which loan applications have a high risk?)
- Trend analysis (How are health care costs changing?)
- Text analysis (Information retrieval, web search robots)
- Science (Cataloging stars, climate reconstruction)

More about applications

- Most data mining techniques are domain independent
- There are countless potential applications in countless domains
- Reported commercial applications are few: you don't want to tell your competitors how you improved your own competence
- Successful applications often are based on a small improvement in a large number of cases (example: direct marketing)

Typical data mining: Association rules

Association rules indicate correlation between observed items

- Retail database ("market basket analysis"):
  chips ⇒ beer (confidence=52%, frequency=3.2%)
- Risk estimation for vehicle insurance:
  sex = male, age < 25 ⇒ insurance claim (21%, 1.2%)
- Natural language sentences:
  "WWW", "Netscape" ⇒ "browser", "internet" (89%, 0.12%)

Task: find **all** association rules that have a frequency of at least c

Association rules—so what?

- The task is to find *all* frequent rules, not just those with 2 or 3 elements
- There are thousands of products, millions of customer transactions
- The objective is to find *unexpected* relationships
- Emphasis is on *description*, not on prediction
- The idea of associations is domain independent

Frequent episodes indicate associations in sequential data

- Telecommunication alarm database:
  battery low in X, auxiliary power failure in $X \Rightarrow$
  high error rate in X (within 1 min)
- Course enrollment database:
  Data Communications $\Rightarrow$ Programming in C
  (within previous 3 years)
- WWW log, accesses to WWW pages:
  Dept. of CS Home Page, Research Groups $\Rightarrow$ Data Mining
- Running text:
  "innovative" ... "IBM" $\Rightarrow$ <period> in between

A "library" of approaches

One set of approaches/techniques for data mining, from simple to complex patterns

- Describing data
- Frequent patterns, associations
- ...
- (Complex) stochastic models
- Exceptions

Frequent patterns

Q: Which similar situations (patterns) occur often in the data?

Example: telecommunication alarm analysis

Input: raw data (sequences of alarms from logs)

Output: repeating (=frequent) patterns

- Parameters: what type of patterns are considered, frequency threshold
- Algorithms and tools: few (tools for specific patterns types in new data mining software)
- Note: how can the pattern types be varied?

Association rules

Q: Which items imply the occurrence of some other items?

Example: market basket analysis

Input: collection of sets (shopping basket contents of customers)

Output: association rules "if X then Y, with confidence c and frequency f"

- Parameters: confidence threshold, frequency threshold
- Algorithms and tools: many (in new data mining software)
- Note: how do you find the most useful association rules from 10 000 rules that are discovered?

Clustering

Q: What natural groups are there in the data?

Example: customer segmentation for direct marketing

Input: set of cases (customers)

Output: partitioning of the set into clusters (sets of different customer types)

- Cases within a cluster are similar to each other
- Cases in different clusters are different

- Parameters: relative weights for attributes, sometimes the number of clusters
- Algorithms and tools: numerous
- Note: does the number of clusters depend on data, or is it a user-defined parameter; scalability

Classification

Q: How to predict the type of a new case?

Example: learning to recognize fraudulent long distance calls

Input: a set of known cases (telephone calls) and their classifications (non-fraudulent/fraudulent)

Output: a rule that predicts the class of an unknown case

- Unlike in clustering, here the classes/groups are given by the user

- Parameters: attributes used in classification
- Algorithms and tools: many (decision trees, rule learners, neural networks)
- Note: comprehensibility, accuracy, scalability

Prediction

Q: What is the value of an unknown attribute likely to be?

Example: How many cars will be sold during the following month?

Input: a set of known cases (previous months)

Output: a rule that predicts the value of the attribute for an unknown case

- Parameters: attributes used in prediction
- Algorithms and tools: several (regression, rule learners, neural networks)
- Note: scalability, comprehensibility, accuracy

Trends

Q: What systematic changes are there in a time series?

Example: How does the consumption of ice cream vary?

Input: a sequence of values (ice cream sold per day)

Output: trends (sales grow 10% each year), cycles (summer sales are larger by 60% than winter sales)

- Parameters: possible cyclic periods (year, week, month?); functions to fit
- Algorithms and tools: some (statistical packages)
- Note: quality of fit

(Complex) statistical models

Q: What are appropriate parameter values for my model, given the data?

Example: What is the response of certain aquatic micro organisms to the lake temperature?

Input: a (detailed) model with some free parameters (a formula for the response curve, with unspecified optimum temperature and tolerance), and data (lakes with the temperature and the micro organisms measured)

Output: parameter values with which the model fits the data (optimum temperatures and tolerances that reflect the data)

- Parameters: the model; it can be very simple or very complex
- Works best when the model is designed to answer specific questions, the model can, in principle, be designed to explore a wide range of possibilities
- Algorithms and tools: almost non-existent (statistical packages, new MCMC software)
- Note: modeling often requires mathematical expertise

Exceptions

Q: Which cases seem exceptional?

Example: Where are health care costs higher than the norm?

Input: data (expenditures on different areas of health care) + a way of knowing the norm

- The norm can be set by discovered patterns, by past data, or by normative statistics

Output: exceptions or deviations (areas of health care where expenditures could be cut down significantly by reaching the norm)

- Parameters: where to find the norm
- Algorithms and tools: hardly any specific ones; see other approaches

Describing data

Large parts of data mining are about finding good descriptions

Q:	In short, what is the data like?
Example:	overviews of students in Canada
Input:	Collection of data (students)
Output:	Description or summary (typical students, counts of different generalizations of students such as science students in Alberta)

- Giving a good overview of *essential* characteristics of data is useful, and often not trivial
- Algorithms and tools: mostly simple statistics
- Note: what kind of questions can be answered reasonably accurately from the summaries and what not?
- OLAP is a way of getting certain types of descriptions of data

Visualization

- Use the vast human capacity of visual processing
- How to visualize a large amount of multi-dimensional data?
- Conventional types of graphs are easy to read
- Dynamic exploration $\Rightarrow$ visual data mining
- Algorithms and tools: some

An example application problem

Problem: Identification of fraudulent uses of credit card

- Stolen cards as deviations or exceptions to a norm?
- Trend detection in shopping patterns, prediction of card use?
- Classification of cards to stolen and not stolen?
- Clustering of users and their shopping habits?
- Analysis of frequent usage patterns of stolen and not stolen cards?

How to evaluate techniques?
When to use which technique?

- What is the problem?
  Finding any unexpected associations or identifying fraud?
- What could be useful questions to ask?
  Which cases belong to class “fraud”, or which cases are exceptions?

- What kind of knowledge about the problem is already available?
  Which one is more important: to identify fraudulent cases, or to identify honest cases?

In the following we list some properties where there are fundamental differences between approaches

The properties correlate strongly

Focus: Do the discovered patterns describe relationships between any attributes (no focus) or between specific attributes (strong focus)?

- No focus: associations, frequent patterns
- Some focusing possible: clustering (weights of attributes say which attributes are important)
- Focused: classification, prediction (fixed target attribute, sometimes also predictive attributes)
- Strong focus: statistical model fitting in some cases (all attributes fixed, only parameters estimated)

Pattern complexity/expressive power: How complex are the patterns, or much information do they carry?

- Simple: association rules (conditional properties of sets of items)
- Fairly simple: frequent patterns; clustering (cluster descriptions usually are not complex)
- More complex: classification, prediction (can be fairly simple, such as linear regression coefficients, or fairly complex but regular, such as decision tree or neural network)
- Arbitrary degree of complexity: statistical models

Number of patterns found: Does the method find a number of patterns or one pattern?

- Many patterns: association rules, frequent patterns (a lot!)
- Some patterns: clustering (many methods output alternative or hierarchical segmentations)
- One pattern: classification, regression; model fitting (sometimes with probability distributions of parameter values)

Use of existing knowledge: Can the method take advantage of existing knowledge?

- Not really: association rules, classification, prediction
- Little: clustering (user-defined weights)
- Yes: statistical models (the model can, in principle, contain any exact knowledge available)

Ability to deal with structured data: Can the method use, e.g., relationships between rows in different tables?

- Not really: most methods
- Yes: statistical models (anything goes, in principle)
- For most practical purposes, however, there are simple fixes

When to use which technique

Note the strong correlation between properties

- Simple patterns are cheap (fast to find and evaluate), complex ones expensive
- Practical methods are in balance: they either look for a particular but expensive pattern in a specific place, or for a number of cheap patterns all over the database

⇒ Simple patterns (e.g., association rules) are useful for finding “something interesting”

⇒ Complex patterns (e.g., decision trees) are useful for well focused problems

What do you need for data mining?

- Data, possibly large masses (data warehouse is useful)
- Some expertise on the data
- Suspicion that important knowledge is hidden in the data
- Questions you cannot express as statistical or OLAP queries
- Commitment: time and money
- Realistic expectations

You need to know what you want to mine.

- What sort of data are you analyzing?
  For what purposes?
- What sort of patterns can at all be discovered from your data, and for your problem?
- Are association rules useful, or episodes, clusters, trends, deviations, classification, regression, or something else?

You need to know how to mine your data.

- What is an abstract description of your task?
- Which methods are applicable in the task?
- Which methods and variations are useful in your case?
- Are there tools that can help you?
- A probable outcome is that there are no methods and tools in the market that perfectly fit the problem

To understand the findings, you need to know the mining methods and the data.

- Why did the tool produce the result it did?
  E.g., how did the tool rank different outcomes?
- What are the limitations of the tool?
  E.g., is it guaranteed to find the best answers?
- What do the findings mean for your domain?
- Do the findings make sense?
- Are the findings useful? Can you apply some of them?

How can you mine your data better?