Course overview

T-122.103 Algorithmic methods of data mining (3 cr) P

- T-122.103 Tiedon louhinnan algoritmiset menetelmät (3 ov) L
- Data mining, also called knowledge discovery in databases (KDD)
- In Finnish: tiedon louhinta, tietämyksen muodostaminen
- Goal of the course: an overview of pattern discovery
- Biased overview
- Theory and examples
- Course home page:
  http://www.cis.hut.fi/Opinnot/T-122.103/
- Email: t122103@james.hut.fi

Contents of the course

- Introduction: what is knowledge discovery, types of tasks, etc.
- Discovery of association rules
- Frequent episodes
- Case study: discovery of episodes from telecommunications alarm databases
- Discovery of all frequent patterns
- Basics of cluster analysis
- Link analysis: basic ideas
- (Search for integrity constraints in databases)

Course organization

- Lectures Friday 10-12 in T4, Heikki Mannila
- Exercises: Friday 12.30-14 in T4, Kai Puolamäki, starting September 19
- Language of instruction: Finnish
- A small project assignment, themes will be given in early October.
- Exam on December 9, 2003 at 9-12

Course material

- H. Mannila, H. Toivonen: Knowledge discovery in databases: the search for frequent patterns; available from the web page of the course
- copies of slides available on the web during the course
- Background material: D. Hand, H. Mannila, P. Smyth: Principles of Data Mining, MIT Press 2001.

Data mining activities at HUT/CIS

- basic research: algorithms, theory
- applied research: genetics, ecology, ubiquitous computing, documents, natural language, ...
- FDK "From Data to Knowledge": Academy of Finland Center of Excellence
- HIIT Basic Research Unit
- Neural Networks Research Center: Academy of Finland Center of Excellence

Chapter 1: Introduction

Chapter 1. Introduction

- What is KDD?
- Some examples
- Data mining vs. statistics and machine learning

What is data mining?

Goal: obtain useful knowledge
from large masses of data.

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data analyst

- "Tell something interesting about this data."
- "Describe this data."
- exploratory data analysis on large data sets

Examples of discovered knowledge

- association rules:
  "80 % of customers who buy beer and sausage buy also mustard"
- rules: "if Age $<$ 40 then Income $<$ 10"
- functional dependencies:
  $A \to B$, i.e., "if $t[A] = u[A]$, then $t[B] = u[B]$"
- models: $Y = aX + b$
- clusterings

Example: sales data

- 0/1 matrix D
- rows: customer transactions
- columns: products
- $t(A) = 1$ iff transaction t contained product A
- thousands of columns, millions of rows
- easy to collect
- find something interesting from this?

Association rules

- mustard, sausage, beer $\Rightarrow$ chips
- conditional probability (*accuracy*) of the rule: 0.8
- *frequency* of the rule: 0.2
- arbitrary number of conjuncts on the left-hand side

Find all association rules with frequency at least *min_fr*.

Example: student/course data

- matrix D with $D(t, c) = 1$ iff student t has taken course c

```
1 1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
1 1 0 0 0 0 0 0 1 0 0 0 1 0 1 0 0 0 1 0 0 0 1 0 0 0 0
0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0
0 1 0 0 1 1 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1
1 1 0 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Example: student/course data

- example rule:

  `if {Data Communications, Unix, Networks}`

  `then (graduation) (0.3)`

Example: words in documents

- Given a collection of documents
- Find out which words define topics, i.e., some sort of clusters of words that tend to occur together
- Topic 1: "mining", "data", "association", "rule", etc.
- Topic 2: "lecture", "exercise", "exam", "material", "assignment", etc.
- Some documents talk about topic 1, some about topic 2, some about both
- How to find the topics from data?

Example: SKICAT sky survey

approximately 1000 photographs of 13000x13000 pixels

$\Rightarrow 3 \cdot 10^9$ smudges of light, each with 40 attributes

- task 1: classify the objects to
  - stars
  - galaxies
  - others
- task 2: find some interesting clusters of objects
  - quasars with high redshift
  - ...

machine learning? pattern recognition?

Why data mining?

- raw data is easy to collect, expensive to analyze
- more data than can be analyzed using traditional methods
- suspicion that important knowledge could be there
- successful applications
- methods: machine learning, statistics, databases
- a tool for exploratory data analysis
- can and has to be used in combination with traditional methods
- strong interest from 1989–, hype from 1995–

The KDD process

- understanding the domain,
- preparing the data set,
- discovering patterns, clusters, models, etc
- postprocessing of discovered regularities, and
- putting the results into use

Where is the effort?

iteration, interaction

Phases of finding patterns, models, clusters, etc.

- What is the task? What do we want to find out?
- What is the model, or pattern class?
- What is the score function, i.e., how do we know which model fits the data well?
- What is the algorithm? How do we find the model?

Data mining tasks

- Pattern discovery: find interesting patterns that occur frequently in the data
- Prediction: predict the value of a variable in the data
- Clustering: group the observations (or variables) into groups of similar objects
- Outlier detection: find observations that are unusual
- ...

Example: words in documents

- 0-1 matrix representation of documents
- Rows: document
- Columns: words
- Entry M_{ij} has a 1 if document i contains word j, 0 otherwise
- "Bag-of-words" representation
- Loses a lot of information, but a lot is retained

1	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	1	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	1	0	0	1	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
1	1	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Data mining tasks in document data

- Find collections of word that occur frequently together in the same documents
- Find topics from data
- Cluster similar documents together
- Will term x occur in the document?
- Which are unusual documents?

Another view of data mining tasks

- Exploratory data analysis
- Descriptive modeling
- Predictive modeling
- Discovering patterns and rules
- Retrieval by content

Data mining and related areas

- How does data mining relate to statistics?
- How does data mining relate to machine learning?
- Other related areas?

Data mining vs. machine learning

- machine learning methods are used for data mining
  - classification, clustering, ...
- amount of data makes a difference
  - accessing examples can be a problem
- data mining has more modest goals:
  - automating tedious discovery tasks, not aiming at human performance in real discovery
  - helping user, not replacing them
- this course: data mining $\setminus$ machine learning

Data mining vs. statistics

"tell me something interesting about this data" – what else is this than statistics?

- the goal is similar
- different types of methods
- in data mining one investigates a lot of possible hypotheses
- amount of data
- data mining as a preliminary stage for statistical analysis
- challenge to data mining: better contact with statistics

Data mining and databases

- ordinary database usage: deductive
- knowledge discovery: “inductive”
- new requirements for database management systems
- novel data structures, algorithms, and architectures are needed
- SQL queries; OLAP queries; data mining

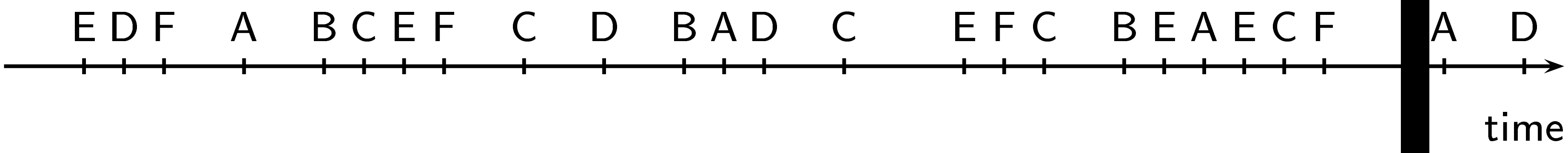
Data mining and algorithms

- Lots of nice connections
- A wealth of interesting research questions
- Some will be treated later in the course

Example of pattern discovery: episodes in sequences

- Data: events in time, e.g.,
  - alarms in telecommunications networks
  - user actions in an user interface
  - database transactions
  - biostatistical events
- Goals: understanding the structure of the process producing the events, prediction of future events.
- How to discover which combinations of events occur frequently?

Example sequence



Observations:

- whenever E occurs, F occurs soon
- whenever A and B occur (in either order), C occurs soon

Telecommunications alarm log

Event	Time
KE82K02-31	780560888
KE82K10-16	780560892
H-M-K09-57	780560917
SOT-K01-03	780560926
MUU-K03-04	780561011
KE82K10-16	780561015
PAK-K14-27	780561119
KE82K10-16	780561138

100–400 different events

events occur recurrently

1 month = 70 000 alarms

Sequences and episodes

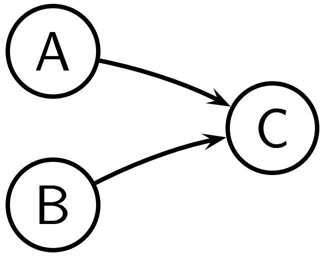
Data a sequence of (event, time) pairs

$$(A, 123), (B, 125), (D, 140), (A, 150), (C, 151), (D, 201), (A, 220)$$

Patterns episodes: a set of events

occurring close to each other in time

in an order extending a given partial order



Pattern class All serial episodes / all parallel episodes / all episodes / ...

Occurrence pattern occurs frequently in data if there are sufficiently many windows of size W in the data such that the pattern occurs in the window

Discovering frequent patterns

- find all frequent episodes of size 1
- build candidate episodes of size 2
- check which episodes of size 2 occur frequently
- continue
- incremental recognition, using previous rounds of computation, ...

A general model for pattern discovery

- given a class of patterns
- an occurrence criterion
- find all patterns from the class that occur frequently enough

Why is KDD an interesting research area?

- practically relevant
- good theoretical questions
- the whole spectrum: from theoretical issues to systems issues to concrete data cleaning to novel discoveries
- easy to cooperate with other areas and with industry