Knowledge Discovery Process and Data Mining - Final remarks

Lecturer: JERZY STEFANOWSKI
Institute of Computing Sciences
Poznan University of Technology
Poznan, Poland
Lecture 14
SE Master Course
2008/2009

Growth Trends

- Moore’s law
  - Computer Speed doubles every 18 months
- Storage law
  - Total storage doubles every 9 months
- Consequence
  - Very little data will ever be looked at by a human
- Knowledge Discovery is **NEEDED** to make sense and use of data.
Data Mining a step in A KDD Process

- Data mining: the core step of knowledge discovery process.

Steps of a KDD Process

- Learning the application domain:
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing
- Data reduction and projection:
  - Find useful features, dimensionality/variable reduction, invariant representation.
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest
- Interpretation: analysis of results.
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge
Interacting with a user / expert in KDD

- KDD is not a fully automatically way of analysis.
- The user is an important element in KDD process.
- Should decide about, e.g.
  - Choosing task and algorithms, selection in preprocessing.
  - Interpretation and evaluation of patterns
    - Objective interestingness measures,…
    - Subjective,…
- By definition, KDD may have several iterations.

Data Preparation for Knowledge Discovery

A crucial issue: The majority of time / effort is put there.
Data Understanding: Relevance

• What data is available for the task?
• Is this data relevant?
• Is additional relevant data available?
• How much historical data is available?
• Who is the data expert?

Data Mining: On What Kinds of Data?

• Relational database
• Data warehouse
• Transactional database
• Advanced database and information repository
  • Object-relational database
  • Spatial and temporal data
  • Time-series data
  • Stream data
  • Multimedia database
  • Heterogeneous and legacy database
  • Text databases & WWW
Are All the “Discovered” Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting
  - Suggested approach: Human-centered, query-based, focused mining

- Interestingness measures
  - A pattern is interesting if it is easily understood by humans, valid on new or test data with some degree of certainty, potentially useful, novel, or validates some hypothesis that a user seeks to confirm

- Objective vs. subjective interestingness measures
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

Can We Find All and Only Interesting Patterns?

- Find all the interesting patterns: Completeness
  - Can a data mining system find all the interesting patterns?
  - Heuristic vs. exhaustive search
  - Association vs. classification vs. clustering

- Search for only interesting patterns: An optimization problem
  - Can a data mining system find only the interesting patterns?
  - Approaches
    - First general all the patterns and then filter out the uninteresting ones.
    - Generate only the interesting patterns—mining query optimization
Examples of Systems for Data Mining

- IBM: QUEST and Intelligent Miner
- Silicon Graphics: MineSet
- SAS Institute: Enterprise Miner
- Statistica Data Miner
- SPSS / Integral Solutions Ltd.: Clementine
- Oracle 9i Miner
- Rapid Miner (YALE)
- Orange
- Other systems
  - Information Discovery Inc.: Data Mining Suite
  - SFU: DBMiner, GeoMiner, MultiMediaMiner

WEKA – Machine Learning and Data Mining

Java implementation of many algorithms

No ideal solutions → but …
RapidMiner (YALE)

Some Rapidminer screenshots
KNIME

- KNIME was developed (and will continue to be expanded) by the Chair for Bioinformatics and Information Mining at the University of Konstanz, Germany.
- It integrates all analysis modules of the well known Weka data mining environment and additional plugins allow R-scripts to be run, offering access to a vast library of statistical routines.
KNIME - An Example of Data Analysis Workflow

• More http://www.knime.org/

Developing Trees

• Node flows
KNIME working with decision trees

Cell Miner

- KNIME has been used to analyze cell images.
Orange (Slovenia)

Following are some screenshots of Orange Widgets and Orange’s visual programming interface for data mining:

Classification tree viewer with a navigator.

Orange - clustering

Data mining © by J.Stefanowski 2005
R project – statistical data exploration

R project
2008 Pool on the popular free software

- Internet users - www.eruditionhome.com

IBM Intelligent Miner: Major Features

- Highly scalable, large database-oriented data mining algorithms
- Multiple data mining functions:
  - Association
  - Classification
  - Sequencing analysis
  - Clustering.
- Visual graphical display
- Influential in database and data mining research communities.
IBM Miner – example of visualisation

Statistica – Statsoft (www.statsoft.pl / *.com)

- User friendly for MS Windows; mainly based on statistical approaches.
- It contains numerous data analysis methods.
- Efficient calculations, good managing results and reports.
- Excellent graphical visualisation.
- Comprehensive help, documentations, supporting books and teaching materials.
- Drivers to data bases and other data sources

Main systems:
- Statistica 6.0 – mainly statistical software
- Statistica Data Miner – specific for DM / user friendly
- Specialized systems – Statistica Neural Networks.
- Quality and Control Cards
- Corporation Tools
- …
DataMiner – main panel

Data Miner – loading data and selecting attributes
Data Miner – choosing methods

Extra tools for defining projects
Using several methods on the same data

SAS Enterprise Miner

data organized in libraries
database-specific parameters
supported databases
Enterprise miner project

SEMMA nodes to choose from

form training, test, and validation sets

impute missing values

linear or logistic

statistical exploration and analysis

various algorithms

assess and compare models

data table

Data Mining and Business Intelligence

Increasing potential to support business decisions

Making Decisions

Data Presentation Visualization Techniques

Data Mining Information Discovery

Data Exploration Statistical Analysis, Querying and Reporting

Data Warehouses / Data Marts OLAP, MDA

Data Sources Paper, Files, Information Providers, Database Systems, OLTP

End User

Business Analyst

Data Analyst

DBA
Industries/fields where you currently apply data mining [KDD Pool - 216 votes total]

<table>
<thead>
<tr>
<th>Industry</th>
<th>Votes</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Banking</td>
<td>29</td>
<td>13%</td>
</tr>
<tr>
<td>Bioinformatics/Biotech</td>
<td>18</td>
<td>8%</td>
</tr>
<tr>
<td>Direct Marketing/Fundraising</td>
<td>19</td>
<td>9%</td>
</tr>
<tr>
<td>eCommerce/Web</td>
<td>12</td>
<td>6%</td>
</tr>
<tr>
<td>Entertainment/News</td>
<td>1</td>
<td>0%</td>
</tr>
<tr>
<td>Fraud Detection</td>
<td>19</td>
<td>9%</td>
</tr>
<tr>
<td>Insurance</td>
<td>15</td>
<td>7%</td>
</tr>
<tr>
<td>Investment/Stocks</td>
<td>9</td>
<td>4%</td>
</tr>
<tr>
<td>Manufacturing</td>
<td>9</td>
<td>4%</td>
</tr>
<tr>
<td>Medical/Pharma</td>
<td>15</td>
<td>7%</td>
</tr>
<tr>
<td>Retail</td>
<td>9</td>
<td>4%</td>
</tr>
<tr>
<td>Scientific data</td>
<td>20</td>
<td>9%</td>
</tr>
<tr>
<td>Security</td>
<td>8</td>
<td>4%</td>
</tr>
<tr>
<td>Telecommunications</td>
<td>12</td>
<td>6%</td>
</tr>
<tr>
<td>Travel</td>
<td>2</td>
<td>1%</td>
</tr>
<tr>
<td>Other</td>
<td>19</td>
<td>9%</td>
</tr>
</tbody>
</table>

Market Analysis and Management

- Where does the data come from?
  - Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies

- Target marketing
  - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
  - Determine customer purchasing patterns over time

- Cross-market analysis
  - Associations/co-relations between product sales, & prediction based on such association

- Customer profiling
  - What types of customers buy what products (clustering or classification)

- Customer requirement analysis
  - identifying the best products for different customers
  - predict what factors will attract new customers

- Provision of summary information
  - multidimensional summary reports
  - statistical summary information (data central tendency and variation)
Corporate Analysis & Risk Management

- Finance planning and asset evaluation
  - cash flow analysis and prediction
  - contingent claim analysis to evaluate assets
  - cross-sectional and time series analysis (financial-ratio, trend analysis, etc.)
- Resource planning
  - summarize and compare the resources and spending
- Competition
  - monitor competitors and market directions
  - group customers into classes and a class-based pricing procedure
  - set pricing strategy in a highly competitive market

Fraud Detection & Mining Unusual Patterns

- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications: Health care, retail, credit card service, telecomm.
  - Auto insurance: ring of collisions
  - Money laundering: suspicious monetary transactions
  - Medical insurance
    - Professional patients, ring of doctors, and ring of references
    - Unnecessary or correlated screening tests
  - Telecommunications: phone-call fraud
    - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
- Retail industry
  - Analysts estimate that 38% of retail shrink is due to dishonest employees
- Anti-terrorism
Other Applications

- **Sports**
  - IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat

- **Astronomy**
  - JPL and the Palomar Observatory discovered 22 quasars with the help of data mining

- **Internet Web Surf-Aid**
  - IBM Surf-Aid applies data mining algorithms to Web access logs for market-related pages to discover customer preference and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.

Controversial Issues: Society and Privacy

- Data mining (or simple analysis) on people may come with a profile that would raise controversial issues of
  - Discrimination
  - Privacy
  - Security
  - Examples:
    - Should males between 18 and 35 from countries that produced terrorists be singled out for search before flight?
    - Can people be denied mortgage based on age, sex, race?
    - Women live longer. Should they pay less for life insurance?
    - Can discrimination be based on features like sex, age, national origin?
    - In some areas (e.g. mortgages, employment), some features cannot be used for decision making
Data Mining and Privacy

• Can information collected for one purpose be used for mining data for another purpose
  • In Europe, generally no, without explicit consent!
  • In US, generally yes,…
• Companies routinely collect information about customers and use it for marketing, etc.
• People may be willing to give up some of their privacy in exchange for some benefits

Data Mining Future Directions

• Currently, most data mining is on flat tables
• Richer data sources
  • text, links, web, images, multimedia, knowledge bases
• Advanced methods
  • Link mining, Stream mining, …
• Applications
  • Web, Bioinformatics, Customer modeling, …
Challenges for Data Mining

- Technical
  - tera-bytes and peta-bytes
  - complex, multi-media, structured data
  - integration with domain knowledge
- Business
  - finding good application areas
- Societal
  - Privacy issues

The Hype Curve for Data Mining and Knowledge Discovery

- Over-inflated expectations
- Disappointment
- Rising expectations
- Performance
- Growing acceptance and mainstreaming
Data Mining Central Quest

Find true patterns and avoid **overfitting** (false patterns due to randomness).

So, be lucky in using this course!

Background literature

- Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press 1996.
Any questions and remarks

- I prefer other questions than
- What about the final exam?

Thank you!