

Welcome

Data Mining: Updates in Technologies

Xindong Wu

Dept of Math and Computer Science

Colorado School of Mines

Golden, Colorado 80401, USA



Email: xwu@mines.edu

Home Page: <http://kais.mines.edu/~xwu/>

Outline

- Data mining: Techniques and applications
- How to select the proper techniques for various kinds of projects (supervised, unsupervised, and hybrid)
- Problems with data - cleaning, transforming, structuring, and organizing data
- Mining at large: Knowledge management
Intelligent learning database systems
- Incremental data mining
- Leading data mining tools

Data mining: Techniques and applications (with an insurance context)

- **Cluster analysis:** Marketing to identify potential clients and to better serve policyholders.
- **Classification** of existing products, to determine the most appropriate policy for a new customer
- **Association analysis** to prevent customer loss (based on cancellations and unemployment for specific regions)
- **Pattern recognition** and **deviation detection** in claims analysis for detecting areas of potential fraud
- **Simulation of Changes and Potential Policies** (interpreting data mining results)

How to select the proper techniques for various kinds of projects

- Data characteristics
- Intended use of the mined knowledge
- Typical techniques
 - Classification
 - Rule induction
 - Decision tree construction
 - Association analysis
 - Statistical patterns
 - Neural networks
 - Genetic algorithms
 - Inductive logic programming

Problems with data

- Data cleaning
 - Contradiction and redundancy elimination
 - Missing and stale data
- Data transforming
 - Feature enhancement and dimension reduction
 - Mapping time-series data
- Data structuring and organization
 - Common terminology
 - Sampling strategy

Mining at large: Knowledge management

- How to **handle massive bodies** of knowledge discovered from large, real-world data?
- When the data comes from different sources (e.g., from different hospitals), how to maintain the **consistency** of the knowledge that is extracted from each of these data sources? Would bagging and boosting work for different data sources?
- If the extracted knowledge **overfits** or **underfits** the data, how to resolve its incompleteness and contradiction?

Knowledge management (2)

- No match: No knowledge extracted is applicable to a new problem.
- Multiple match: Knowledge from data mining gives conflicting advice to a new problem.
- Single match: a test example matches with the description of a specific class

Intelligent learning database systems

- An ILDB system provides mechanisms for
 - Preparation and translation of standard (e.g. relational) **database information** into a form suitable for use by its induction engines,
 - Using **induction** techniques to extract knowledge from databases, and
 - Interpreting the extracted knowledge to solve users' problems by **deduction** (or KBS technology)

Incremental data mining

- Quinlan's **windowing** technique: costs vs benefits of windowing
- **Incremental learning**: training data items one by one
- **Multi-layer incremental induction** (Wu and Lo 1998): Subsets (or samples) of a database are processed one by one
 - **Data Partitioning**
 - **Generalization**: A set of rules is learned.
 - **Reduction**: Behavioral examples are derived.
 - From behavioral examples, generalization can extract new rules, which are expected to correct defects and inconsistencies of previous rules.
 - Successive applications of the above generalization-reduction process allow more accurate and more complex (because of disjunctive) rules to be discovered, by sequentially handling the subsets of examples

Incremental data mining (2)

- **Incremental batch learning** [Clearwater et al 1989]:
 - Rules are generated on each batch of examples, "almost-good-rules" are collected, an additional filter (such as Gen-Spec method and the n -Best) is applied, and then a RL technique (derived from the Meta-DENDRAL) is used to refine the remaining rules on the next batch of examples.
- **Meta-learning** for scalable inductive learning [Chan 1997]:
 - Partition a large database into subsets, run a learning algorithm on each of the subsets, and combine the results in some principled fashion. Since the learning processes are independent, they can be run in parallel for speed up over a sequential system.
 - Arbiters and combiners

Incremental data mining (3)

- Learning “ensembles of classifiers” (such as **bagging** and **boosting**): Generates multiple versions of a predictor by running an existing learning algorithm many times on a set of re-sampled data. A data item in the original database can be used in many samples for generating different versions of the predictor.
- **Stacked generalization** [Wolpert 1992, Ting & Witten 1997] uses a high-level model to combine lower-level models to achieve greater predictive accuracy. These lower-level models are generated also by sampling the original database, and cross-validation is used to generate higher-level data for the high-level model.

Leading data mining tools

- **C5.0/See5**, RuleQuest Research (Australia)
 - decision tree construction and rule induction
 - based on best-known data mining
- **Intelligent Miner** family, IBM (USA)
 - customer relationship marketing; fraud and abuse detection
 - database support
- **Clementine**, Integral Solutions (UK)
 - neural networks and rule induction
 - data visualization in tables, histograms, plots and “webs”
- **Enterprise Miner**, SAS Institute (USA)
 - clustering, decision trees, & neural networks
 - GUI designed for business users

Leading data mining tools (2)

- **DBMiner**, DBMiner Technology (Canada)
 - Data warehousing support
 - A comprehensive set of mining facilities
- **Darwin**, Thinking Machines (US, distributors in Japan)
 - parallel, scalable, data mining architecture
 - NN, CART, and GAs
- **Pattern Recognition Workbench**, Unica (USA)
 - pattern classification and function estimation with an experiment manager
 - statistical regression, non-parametric, and NN algorithms

Leading data mining tools (3)

- Other data mining tools
(<http://www.datamininglab.com/resources.html>)
 - A Comparison of Leading Data Mining Tools
 - Data Mining Tools for Fraud Detection
 - Fourteen Desktop Data Mining Tools