Illinois Institute of Technology

Introduction to Data Mining (CS 422)

Illinois Institute of Technology

Course Outline

- Introduction
- Data Pre-processing
- Data Mining Algorithms
  - » Supervised
    - – Naïve Bayes
    - – Neural Network
    - – Decision Tree
    - – K-Nearest Neighbor,…
    - – Rule based
    - – Ensemble methods
  - » Unsupervised
    - – Association mining
    - – Clustering…
- Evaluation

Introduction

Illinois Institute of Technology

- Mining potential useful knowledge from a large amount of data.
- Other terminologies:
  - » Knowledge Discovery in Databases (KDD)- although DM is a step in KDD.
  - » Knowledge Extraction
  - » Data Analysis
  - » Information Harvesting
  - » and more….

Data Mining vs. KDD

Illinois Institute of Technology

- Knowledge Discovery in Databases (KDD) is the process of finding useful information and patterns in the data.
- Data Mining is the use of algorithms to find the useful information in the KDD process.
- KDD process is:
  - » Data cleaning & integration (Data Pre-processing)
  - » Creating a common data repository for all sources, such as data warehouse.
  - » Data mining
  - » Visualization for the generated results

Data Mining vs. Database

Illinois Institute of Technology

- DB's user knows what is looking for.
- DM's user might/might not know what is looking for.
- DB's answer to query is 100% accurate, if data correct.
- DM's effort is to get the answer as accurate as possible.
- DB's data are retrieved as stored.
- DM's data need to be cleaned (some what) before producing results.
- DB's results are subset of data.
- DM's results are the analysis of the data.
- The meaningfulness of the results is not the concern of Database as it is the main issue in Data Mining.

Data Mining Applications

Illinois Institute of Technology

- Fraud Detection and Risk Analysis
  - » Credit card fraud detection
  - » Money laundry detection
  - » Risk of loan payment
  - » etc….
- Retail
  - » Sale and Promotion
  - » Coupon
  - » etc…
- Stock Market Analysis

Data Mining Applications

Illinois Institute of Technology

- Identifying Criminals & Profiling
- Flood Prediction
- Telecommunications
- Medical Diagnosis & Treatment
- Biomedical & DNA Data Analysis
  - » Which genes co-occur with other genes?
  - » What are the sequence of genetic activities in stages of a disease?
- Web Mining
  - » What are associations among different pages?
  - » What are web page characteristics?
  - » What is the distribution of information on web?

Privacy Issues

Illinois Institute of Technology

- DM applications derive demographics about customers via
  - – Credit card use
  - – Store card
  - – Subscription
  - – Book, video, etc rental
  - – and via more sources…
- As the DM results are deemed to be a good estimate or prediction, one has to be sensitive to the results not to violate privacy.

DM Commercial Tools

Illinois Institute of Technology

- Problem: Not a common model/ architecture.
  - » Accessing different but not necessarily all type of data repositories.
  - » Supporting one or more of the DM algorithms.
  - » May/may not supporting all data types.
  - » Supporting different but not all functionalities.
  - » platform dependant.
  - » => Each application might work with one commercial tool and not with the other tool.

Some of the DM Commercial Tools

Illinois Institute of Technology

- Darwin (Oracle Corp.)
- MineSet (Silicon Graphics Inc. - SGI)
- Intelligent Miner (IBM Corp)
- Enterprise Miner (SAS Institute Inc.)
- Clementine (SPSS Inc – Integral Solutions)
- BrainMaker (California Scientific Software)
- CART (Salford Systems)
- MARS (Salford Systems)
- Scenario (Cognos Inc.)
- Web Analyst (Megaputer Intelligence Inc.)
- SurfAid Analysis (IBM corp)
- etc….

Different Data Sources

Illinois Institute of Technology

- Relational Database
- Data Warehouse
- Flat Files
- Web
- Object Oriented database
- Multi Media

Data Warehouse

Illinois Institute of Technology

- Many enterprises consolidate data from their different homogeneous and heterogeneous data repositories into one common data source called Data Warehouse (DW).
- Data Warehouse contains current and historical data to be used for planning and forecasting in Decision Support Systems (DSS).
- Traditional Databases are operational databases that are day-to-day data.
- Star-schema, Snow-Flakes, Galaxy are modeling schemes in DW.

Data Warehouse (Cont'd)

Illinois Institute of Technology

- To improve the performance in DW different techniques such as Summarization and Denormalization are used.
- Usually but not always DW is accessed by On-Line Analytical Processing (OLAP).
- SQL gives a precise answer to a user query.
- OLAP gives a multi-dimensional view of data and is as extension of some aggregate functions in SQL.
- OLAP Operations are Slice, Dice, Roll-up, and Drill-down.

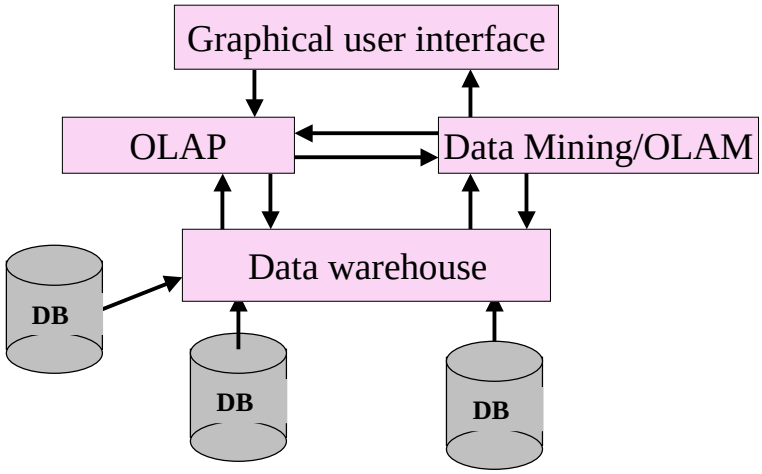
OLAP vs. Data Mining

Illinois Institute of Technology

- OLAP is a data summarization/ aggregation tool that facilitates the data analysis for the user by providing a multi-dimensional view of the data.
- Data Mining Tool provides an automated discovery of knowledge and gives more in-depth knowledge about data and hidden information.
- OLAM (OLAP Mining) is the integration of OLAP with Data Mining.

OLAM Architecture

Illinois Institute of Technology



DM and Other Disciplines

Illinois Institute of Technology

- Statistical Concepts
  - » Bayes Theorem
  - » Regression
  - » etc…
- Machine Learning
  - » Neural Network
  - » Genetic Algorithm
  - » Clustering
  - » Association Rule

Scalability

Illinois Institute of Technology

- Statistical approach deal with small data sets.
  - » Believe that all data must be cleaned and reduced.
- Machine Learning deal with small data sets.
  - » Goal is to make machine learn.
  - » Applications such as Chess Playing rather than applications that deal market analysis.
- Real life data to be mined is huge, thus need scalable algorithms.

DM Algorithms

Illinois Institute of Technology

- Supervised (Classification)
  - » Bayesian
  - » Neural Network
  - » Decision Tree
  - » Others: Genetic Algorithms, Fuzzy Set, K-Nearest Neighbor
- Unsupervised
  - » Association Rules
  - » Clustering

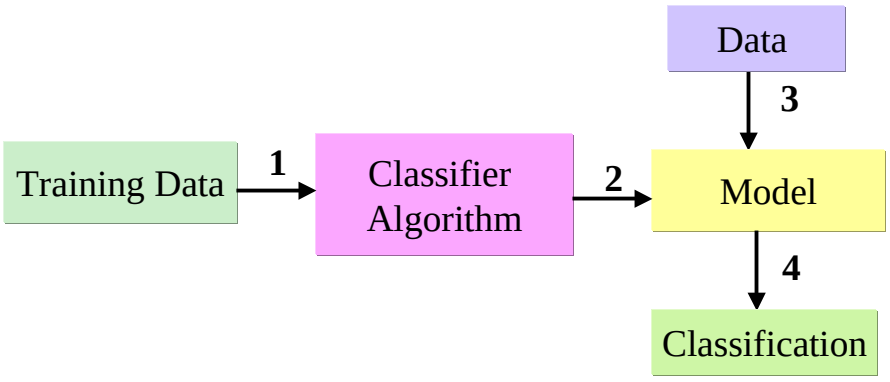
Supervised vs. Unsupervised

Illinois Institute of Technology

- Supervised algorithms
  - » Learning by example:
    - – Use training data which has correct answers (class label attribute)
    - – Create a model by running the algorithm on the training data
    - – Identify a class label for the incoming new data
- Unsupervised algorithms
  - » Do not use training data.
  - » Classes may not be known in advance.

Supervised Algorithms

Illinois Institute of Technology



DM Evaluation Metrics

Illinois Institute of Technology

- Not always straightforward.
- ROI (Return on Investment) used in business to measure benefit of using Data Mining.
- Lift Chart used to visualize and measure response modeling performance.
- Traditional Computer Science evaluation metrics are space requirement and time complexity to compare the algorithms.
- Measuring accuracy of DM results:
  - » Use of Cross-Validation in Supervised algorithms.
  - » Information Retrieval measures of Precision & recall.
  - » Various accuracy measures based on each algorithm.

Summary

Illinois Institute of Technology

- Data Mining algorithms are used to detect the information that we did not know.
- There are various data sources, types,formats and applications for Data Mining.
- Usually Data Mining is used on a Data Warehouse.
- There are many Data Mining algorithms.
- Scalability of Data Mining differentiates it from statistical and Machine Learning approach, as in Data Mining we deal with huge amount of data.