Introduction to KDD and data mining

Nguyen Hung Son

This presentation was prepared on the basis of the following public materials:

1. Jiawei Han and Micheline Kamber, „Data mining, concept and techniques” [http://www.cs.sfu.ca](http://www.cs.sfu.ca)
Lecture plan

- Motivations: why data mining?
- Definitions of data mining?
- Examples of applications
- Data mining systems and functionality
- Methods in data mining
- Data mining: a KDD process
- Data mining issues
Motivation: large scale databases

- Advanced methods in data extraction and data storing techniques
- Growth of many application areas

- More generated data:
  - Bank, telecom, other business transactions ... 
  - Scientific data: astronomy, biology, etc
  - Web, text, and e-commerce
Massive data sources

- Huge number of records
  $10^6 - 10^{12}$ in case of databases about celestial objects (astronomy)

- Huge number of attributes (features, measurements, columns)
  Hundreds of variables in patient records corresponding to results of medical examinations
Motivation

- „We are melting in a ocean of data, but we need a knowledge”
- PROBLEM: How to get a useful information/knowledge from large databases?
- SOLUTION: Data wherehouse + data mining
Lecture plan

- Motivations: why data mining?
- Definitions of data mining?
- Examples of applications
- Data mining systems and functionality
- Methods in data mining
- Data mining: a KDD process
- Data mining issues
What Is Data Mining?

An iterative and interactive process of discovering
- novel,
- valid,
- useful,
- comprehensive and
- understandable
patterns and models in
MASSIVE data sources (databases).

- Novel: something we are not aware of
- Valid: generalise to the future
- Useful: some reaction is possible
- Understandable: leading to insight
- Iterative: many steps and many passes
- Interactive: human is a part of the system
What is Data Mining

- Alternative names and their “inside stories”:
  - Data mining: a misnomer?
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

- What is not data mining?
  - (Deductive) query processing.
  - Expert systems or small ML/statistical programs
Evolution of Database Technology

- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.) and application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s—2000s:
  - Data mining and data warehousing, multimedia databases, and Web databases
Big Data Examples

- Europe's Very Long Baseline Interferometry (VLBI) has 16 telescopes, each of which produces 1 Gigabit/second of astronomical data over a 25-day observation session
  - storage and analysis a big problem
- AT&T handles billions of calls per day
  - so much data, it cannot be all stored -- analysis has to be done “on the fly”, on streaming data
Largest databases in 2003

- Commercial databases:
  - Winter Corp. 2003 Survey: France Telecom has largest decision-support DB, ~30TB; AT&T ~ 26 TB

- Web
  - Alexa internet archive: 7 years of data, 500 TB
  - Google searches 4+ Billion pages, many hundreds TB
  - Internet Archive (www.archive.org), ~ 300 TB
5 million terabytes created in 2002

- UC Berkeley 2003 estimate: 5 exabytes (5 million terabytes) of new data was created in 2002.
- US produces ~40% of new stored data worldwide
Data Growth Rate

- Twice as much information was created in 2002 as in 1999 (~30% growth rate)
- Other growth rate estimates even higher
- Very little data will ever be looked at by a human
- Knowledge Discovery is **NEEDED** to make sense and use of data.
Lecture plan

- Motivations: why data mining?
- Definitions of data mining?
- Examples of applications
- Data mining systems and functionality
- Methods in data mining
- Data mining: a KDD process
- Data mining issues
Data Mining

Application areas

- Science
  - astronomy, bioinformatics, drug discovery, …

- Business
  - advertising, CRM (Customer Relationship management), investments, manufacturing, sports/entertainment, telecom, e-Commerce, targeted marketing, health care, …

- Web:
  - search engines, bots, …

- Government
  - law enforcement, profiling tax cheaters, anti-terror(?)

KDD and DM 15
Data Mining for Customer Modeling

- Customer Tasks:
  - attrition prediction
  - targeted marketing:
    - cross-sell, customer acquisition
  - credit-risk
  - fraud detection

- Industries
  - banking, telecom, retail sales, …
Customer Attrition: Case Study

- Situation: Attrition rate at for mobile phone customers is around 25-30% a year!

Task:

- Given customer information for the past N months, predict who is likely to attrite next month.
- Also, estimate customer value and what is the cost-effective offer to be made to this customer.
Customer Attrition Results

- Verizon Wireless built a customer data warehouse
- Identified potential attriters
- Developed multiple, regional models
- Targeted customers with high propensity to accept the offer
- Reduced attrition rate from over 2%/month to under 1.5%/month (huge impact, with >30 M subscribers)
  (Reported in 2003)
Assessing Credit Risk: Case Study

- **Situation:** Person applies for a loan
- **Task:** Should a bank approve the loan?
- **Note:** People who have the best credit don’t need the loans, and people with worst credit are not likely to repay. Bank’s best customers are in the middle
Credit Risk - Results

- Banks develop credit models using variety of machine learning methods.
- Mortgage and credit card proliferation are the results of being able to successfully predict if a person is likely to default on a loan.
- Widely deployed in many countries.
Successful e-commerce – Case Study

- A person buys a book (product) at Amazon.com.
- Task: Recommend other books (products) this person is likely to buy
- Amazon does clustering based on books bought:
  - customers who bought “Advances in Knowledge Discovery and Data Mining”, also bought “Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations”
- Recommendation program is quite successful
Unsuccessful e-commerce case study (KDD-Cup 2000)

- Data: clickstream and purchase data from Gazelle.com, legwear and legcare e-tailer
- Q: Characterize visitors who spend more than $12 on an average order at the site
- Dataset of 3,465 purchases, 1,831 customers
- Very interesting analysis by Cup participants
  - thousands of hours - $X,000,000 (Millions) of consulting
- Total sales -- $Y,000
- Obituary: Gazelle.com out of business, Aug 2000
Genomic Microarrays – Case Study

Given microarray data for a number of samples (patients), can we

■ Accurately diagnose the disease?
■ Predict outcome for given treatment?
■ Recommend best treatment?
Example: ALL/AML data

- 38 training cases, 34 test, ~ 7,000 genes
- 2 Classes: Acute Lymphoblastic Leukemia (ALL) vs Acute Myeloid Leukemia (AML)
- Use train data to build diagnostic model

Results on test data:
33/34 correct, 1 error may be mislabeled
Security and Fraud Detection - Case Study

- Credit Card Fraud Detection
- Detection of Money laundering
  - FAIS (US Treasury)
- Securities Fraud
  - NASDAQ KDD system
- Phone fraud
  - AT&T, Bell Atlantic, British Telecom/MCI
- Bio-terrorism detection at Salt Lake Olympics 2002
Lecture plan

- Motivations: why data mining?
- Definitions of data mining?
- Examples of applications
- Data mining systems and functionality
- Methods in data mining
- Data mining: a KDD process
- Data mining issues
Architecture of a Typical Data Mining System

- Graphical user interface
- Pattern evaluation
- Data mining engine
- Database or data warehouse server
- Knowledge-base
- Databases
- Data Warehouse
- Data cleaning & data integration
- Filtering
Data Mining: On What Kind of Data?

- Relational databases
- Data warehouses
- Transactional databases
- Advanced DB and information repositories
  - Object-oriented and object-relational databases
  - Spatial databases
  - Time-series data and temporal data
  - Text databases and multimedia databases
  - Heterogeneous and legacy databases
  - WWW
Data Mining Functionalities (1)

- **Concept description: Characterization and discrimination**
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions

- **Association** (correlation and causality)
  - Multi-dimensional vs. single-dimensional association
  - \( \text{age}(X, "20..29") \land \text{income}(X, "20..29K") \Rightarrow \text{buys}(X, "PC") \) [support = 2%, confidence = 60%]
  - \( \text{contains}(T, "computer") \Rightarrow \text{contains}(T, "software") \) [1%, 75%]
Data Mining Functionalities (2)

- **Classification and Prediction**
  - Finding models (functions) that describe and distinguish classes or concepts for future prediction
  - E.g., classify countries based on climate, or classify cars based on gas mileage
  - Presentation: decision-tree, classification rule, neural network
  - Prediction: Predict some unknown or missing numerical values

- **Cluster analysis**
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Clustering based on the principle: maximizing the intra-class similarity and minimizing the interclass similarity
Data Mining Functionalities (3)

- **Outlier analysis**
  - Outlier: a data object that does not comply with the general behavior of the data
  - It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis

- **Trend and evolution analysis**
  - Trend and deviation: regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis

- **Other pattern-directed or statistical analyses**
Lecture plan

- Motivations: why data mining?
- Definitions of data mining?
- Examples of applications
- Data mining systems and functionality
- Methods in data mining
- Data mining: a KDD process
- Data mining issues
Component of a Data Mining algorithm

- Knowledge representation model
- Evaluation criteria
- Search strategy
Knowledge representation

- Using logical language to describe mined patterns. E.g.,
  - Logical formulas
  - Decision tree
  - Neural networks (!)
Search strategy

- Parameter searching
- Model searching
Are All the “Discovered” Patterns Interesting?

- A data mining system/query 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?
  - Association vs. classification vs. clustering

- **Search for only interesting patterns: Optimization**
  - 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
Major Data Mining Methods

- **Classification:** predicting an item class
- **Clustering:** finding clusters in data
- **Associations:** e.g. A & B & C occur frequently
- **Visualization:** to facilitate human discovery
- **Summarization:** describing a group
- **Deviation Detection:** finding changes
- **Estimation:** predicting a continuous value
- **Link Analysis:** finding relationships
- …
Related techniques

- Neural Networks
- Fuzzy Sets
- Rough Sets
- Time series analysis
- Bayesian Networks
- Decision trees
- Evolutionary programming and GA
- Markov modelling
Example

Income

Debt

KDD and DM 40
Linear classification

![Diagram showing linear classification with axes for Income and Debt, and symbols representing loan and no loan.]

KDD and DM 41
Linear regression

Regression line

Debt

Income

X X O O X X O O O O O O
X X X X X X O O O O O O
X X

KDD and DM 42
Clustering

income

Debt
Single threshold (cut)
Nonlinear classifier
Nearest neighbour

- Income
- Debt

- No Loan
- Loan

KDD and DM 46
Lecture plan

- Motivations: why data mining?
- Definitions of data mining?
- Examples of applications
- Data mining systems and functionality
- Methods in data mining
- Data mining: a KDD process
- Data mining issues
KDD

Artificial Intelligence

Database

High Performance Computing

Statistics

Soft Computing

KDD
Data Mining: a KDD process
Steps of a KDD Process

1. Learning the application domain:
   - relevant prior knowledge and goals of application
2. Creating a target data set: data selection
3. Data cleaning and preprocessing: (may take 60% of effort!)
4. Data reduction and transformation:
   - Find useful features, dimensionality/variable reduction, invariant representation.
5. Choosing functions of data mining
   - summarization, classification, regression, association, clustering.
6. Choosing the mining algorithm(s)
7. Data mining: search for patterns of interest
8. Pattern evaluation and knowledge presentation
   - visualization, transformation, removing redundant patterns, etc.
9. Use of discovered knowledge
The goals of Data Mining

- **Prediction**: To foresee the possible future situation on the basis of previous events.
  
  *Given sales recordings from previous years can we predict what amount of goods we need to have in stock for the forthcoming season?*

- **Description**: What is the reason that some events occur?
  
  *What are the reasons for the cars of one producer to sell better than equal products of other producers?*

- **Verification**: We think that some relationship between entities occur.
  
  *Can we check if (and how) the thread of cancer is related to environmental conditions?*

- **Exception detection**: There may be situations (records) in our databases that correspond to something unusual.
  
  *Is it possible to identify credit card transactions that are in fact frauds?*
Classification of Data Mining systems

- General functionality
  - Descriptive data mining
  - Predictive data mining

- Different views, different classifications
  - Kinds of databases to be mined
  - Kinds of knowledge to be discovered
  - Kinds of techniques utilized
  - Kinds of applications adapted
A Multi-Dimensional View of Data Mining Classification

- **Databases to be mined**
  - Relational, transactional, object-oriented, object-relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW, etc.

- **Knowledge to be mined**
  - Characterization, discrimination, association, classification, clustering, trend, deviation and outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels

- **Techniques utilized**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, neural network, etc.

- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, etc.
Lecture plan

- Motivations: why data mining?
- Definitions of data mining?
- Examples of applications
- Data mining systems and functionality
- Methods in data mining
- Data mining: a KDD process
- Data mining issues
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
Major Issues in Data Mining (1)

- **Mining methodology and user interaction**
  - Mining different kinds of knowledge in databases
  - Interactive mining of knowledge at multiple levels of abstraction
  - Incorporation of background knowledge
  - Data mining query languages and ad-hoc data mining
  - Expression and visualization of data mining results
  - Handling noise and incomplete data
  - Pattern evaluation: the interestingness problem

- **Performance and scalability**
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed and incremental mining methods
Major Issues in Data Mining (2)

- **Issues relating to the diversity of data types**
  - Handling relational and complex types of data
  - Mining information from heterogeneous databases and global information systems (WWW)

- **Issues related to applications and social impacts**
  - Application of discovered knowledge
    - Domain-specific data mining tools
    - Intelligent query answering
    - Process control and decision making
  - Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem
  - Protection of data security, integrity, and privacy
References


- *Data Mining Techniques: for Marketing, Sales and Customer Support*. M. Berry, G. Linoff (Wiley)


- *Rough Sets in Knowledge Discovery I & II*. L. Polkowski, A. Skowron (Springer)