Lecture 1

Introduction to Data Mining

Zhou Shuigeng (周水庚)

Feb. 19, 2006
Motivation: Why data mining?
What is data mining?
Data Mining: On what kind of data?
Data mining functionality
Are all the patterns interesting?
Classification of data mining systems
Major issues in data mining
Open topics
Necessity Is the Mother of Invention

- Data explosion problem
  - Automated data collection tools and mature database technology lead to tremendous amounts of data accumulated and/or to be analyzed in databases, data warehouses, and other information repositories

- We are drowning in data, but starving for knowledge! (我们湮没于数据之中，却因知识匮乏而饥肠辘辘！)

- Solution: Data warehousing and data mining
  - Data warehousing and on-line analytical processing
  - Mining interesting knowledge (rules, regularities, patterns, constraints) from data in large databases
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.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)

- **1990s:**
  - Data mining, data warehousing, multimedia databases, and Web databases

- **2000s**
  - Data mining with a variety of applications
  - Web technology and global information systems
  - Stream data management and mining
  - Biological data mining
  - Privacy preservation data mining, etc.
What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?

- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

- Watch out: Is everything “data mining”?
  - (Deductive) query processing.
  - Expert systems or small ML/statistical programs
Why Data Mining?—Potential Applications

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)

- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - DNA and bio-data analysis
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**
Web Analysis

- Web content analysis
  - Web pages clustering: finding similar pages in the web
  - Web pages classification to organize web pages by category

- Web structure analysis
  - search: the web search engine Google is developed through data mining research in Stanford
  - Re-structure web sites

- Web usage analysis (Web log analysis)
  - if customer visit page A and page B, then he/she is likely to go to page C and then buy product E
  - sequences clustering, finding group of customers who have very similar page visit sequences
Security

- Network Security
  - used to detect network intrusion
  - analyze the commands issued
  - analyze the flow of network traffic
- Anti-terrorism
  - Money-flow analysis
  - Communication analysis
- Security Camera
  - analyze abnormal movement in a room or in an airplane
  - fast detection of weapons during X-ray scan
Geographical Data Mining

- Data that are presented on a map
- Weather prediction
- Finding pollution sources
- Analyze crime patterns
- Location planning
- Traffic analysis
Performance Optimization

- cache pre-fetching
- semantic compression using data mining
- better indexes through clustering
- supply-chain management
Bioinformatics

- DNA sequences analysis:
  - indexing
  - clustering
  - Compression

- Gene expression analysis
  - function prediction
  - visualization
  - clustering
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.
Data Mining: A KDD Process

- Data mining—core of knowledge discovery process

Data Cleaning → Data Integration → Task-relevant Data → Data Mining → Knowledge

Pattern Evaluation

Data Warehouse

Databases
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: (may take 60% of effort!)
- Data reduction and transformation
  - Find useful features, dimensionality/variable reduction, invariant representation.
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering.
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest
- Pattern evaluation and knowledge presentation
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge
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
Architecture: Typical Data Mining System

- **Graphical user interface**
- **Pattern evaluation**
- **Data mining engine**
- **Database or data warehouse server**
- **Knowledge-base**

- **Data cleaning & data integration**
- **Filtering**

- **Databases**
- **Data Warehouse**
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, Bio-database, and WWW
Data Mining Functionalities

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

- **Association (correlation and causality)**
  - Diaper -> Beer [0.5%, 75%]

- **Classification and Prediction**
  - Construct 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
  - Predict some unknown or missing numerical values
Data Mining Functionalities (2)

- **Cluster analysis**
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing interclass similarity

- **Outlier analysis**
  - Outlier: a data object that does not comply with the general behavior of the data
  - Noise or exception? No! 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**
Association Rule Discovery

- **Association rule mining:**
  - Finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories.

- **Applications:**
  - Basket data analysis, cross-marketing, catalog design, loss-leader analysis, clustering, classification, etc.

- **Examples.**
  - Rule form: “Body ® Head [support, confidence]”.
  - `b(, “diapers”) ® b(, “beers”) [0.5%, 60%]
  - `m(x, “CS”) ^ t(x, “DB”) ® g(x, “A”) [1%, 75%]`
Cluster Analysis (1)

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Cluster analysis
  - Grouping a set of data objects into clusters
Cluster Analysis (2)

- Issues to be consider in clustering include
  - Types of distance functions
  - Objective measures
  - Handling high dimensionalities
  - Scalability
  - Selecting relevant dimensions ....

- Clustering had been studied in established field like statistic but continue to be an important research topic in data mining as new applications and data types emerge
Outlier Detection

- What are outliers?
  - The set of objects are considerably dissimilar from the remainder of the data
  - Example: Sports (Michael Jordon), ...
- Naïve way to find outliers: cluster objects and objects that are very far from the clusters are outliers
- Challenge is to find outliers without clustering
Classifier Building

- **Classification:**
  - Classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data.

- **Model construction:** Describing a set of predetermined classes.
  - Each tuple is assumed to have a class label attribute.
  - The set of tuples used for model construction: *training set*.
  - The model is represented as classification rules.

- **Model usage:** For classifying future or unknown objects.
  - Estimate accuracy of the model.
    - The known label of test sample is compared with the classified result from the model.
    - Accuracy rate is the percentage of test set samples that are correctly classified by the model.
Classification: Training Dataset

This follows an example from Quinlan’s ID3

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunny</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>N</td>
</tr>
<tr>
<td>sunny</td>
<td>hot</td>
<td>high</td>
<td>true</td>
<td>N</td>
</tr>
<tr>
<td>overcast</td>
<td>hot</td>
<td>high</td>
<td>false</td>
<td>P</td>
</tr>
<tr>
<td>rain</td>
<td>mild</td>
<td>high</td>
<td>false</td>
<td>P</td>
</tr>
<tr>
<td>rain</td>
<td>cool</td>
<td>normal</td>
<td>false</td>
<td>P</td>
</tr>
<tr>
<td>rain</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>N</td>
</tr>
<tr>
<td>overcast</td>
<td>cool</td>
<td>normal</td>
<td>true</td>
<td>P</td>
</tr>
<tr>
<td>sunny</td>
<td>mild</td>
<td>high</td>
<td>false</td>
<td>N</td>
</tr>
<tr>
<td>sunny</td>
<td>cool</td>
<td>normal</td>
<td>false</td>
<td>P</td>
</tr>
<tr>
<td>rain</td>
<td>mild</td>
<td>normal</td>
<td>false</td>
<td>P</td>
</tr>
<tr>
<td>sunny</td>
<td>mild</td>
<td>normal</td>
<td>true</td>
<td>P</td>
</tr>
<tr>
<td>overcast</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>P</td>
</tr>
<tr>
<td>overcast</td>
<td>hot</td>
<td>normal</td>
<td>false</td>
<td>P</td>
</tr>
<tr>
<td>rain</td>
<td>mild</td>
<td>high</td>
<td>true</td>
<td>N</td>
</tr>
</tbody>
</table>
Classification: Decision Tree Model

outlook?

sunny

Humidity?

high

N

normal

P

overcast

P

wind?

strong

N

weak

P
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 generate all the patterns and then filter out the uninteresting ones.
  - Generate only the interesting patterns—mining query optimization
Data Mining: Classification Schemes

- General functionality
  - Descriptive data mining
  - Predictive data mining

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

- **Data to be mined**
  - Relational, data warehouse, transactional, stream, object-oriented/relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW

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

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

- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, Web mining, etc.
Data Mining: A Generalized Framework

**Techniques**
- Association rules discovery
- Sequential Pattern Discovery
- Cluster analysis
- Outlier Detection
- Classifier Building
- Data Cube/Data Warehouse Construction
- Visualization ...

**Database Technology:**
- Indexing, Compression, Data Structure
- AI/ Machine Learning
- Statistics
- Information Theory
- Theoretical CS:
  - Approximate, Random, Online Algorithms
- Mathematical Programming
- Computational Geometry ...

**Applications**
- Customer Relationship Management (CRM)
- Web pages Searches and Analysis
- Network Security
- Geographical Data Analysis
- Genomic Database ...

**Principles**
Data Mining: Confluence of Multiple Disciplines

- Database Systems
- Statistics
- Machine Learning
- Algorithm
- Visualization
- Other Disciplines
Techniques like indexing, compression and query optimization help to provide quick access to relevant data for the mining algorithm.

**Examples**

- Many clustering algorithms require k-nearest neighbors searches which is support by index structure like R* tree etc. With an index, time complexity is $O(n \log n)$, otherwise $O(n^2)$.

- After being compressed, databases can be stored in the main memory for fast mining operation.
Many techniques from data mining are in fact machine learning algorithms which are made scalable using database technologies, approximation algorithms etc. However, data mining is NOT only machine learning.

Example:

Classification is well studied in machine learning. Decision tree induction algorithms like C4.5, ID3 are first developed by the machine learning researchers and make efficiency by the database community.
Machine Learning

- How to construct computer programs that automatically improve with experience
- Machine learning theory
  - How does learning performance vary with the number of training examples presented
  - What learning algorithms are most appropriate for various types of learning tasks
ML Formal Definition

A computer program is said to learn from experience $E$ w.r.t some classes of tasks $T$ and performance $P$, if its performance at tasks in $T$, as measured by $P$, improves with experience $E$. 
Machine Learning vs. Data Mining(I)

- Since data mining is essentially the use of historical data to improve decisions, we can see this as trying to learn from previous experience. Machine learning can provide many useful tools and techniques for this purpose.
- Machine learning on the other hand does not need to worry about the interpretability of the knowledge being learned or discovered.
Machine Learning vs. Data Mining (II)

- Data mining also need to deal with the tasks of handling massive datasets which mean techniques from database research must be brought in.
- Generally, we can say that machine learning deal with the effectiveness aspect of data mining while database research deal with the efficiency aspect.
A Generalized View of ML(or DM)

1. The *task* the algorithm is used to address (e.g. classification, clustering, etc.)

2. The *structure of the model or pattern* we are fitting to the data (e.g. a linear regression model)

3. The *score function* used to judge the quality of the fitted models or patterns (e.g. accuracy, BIC, etc.)

4. The *search or optimization method* used to search over parameters and structures (e.g. steepest descent, MCMC, etc.)

5. The *data management technique* used for storing, indexing, and retrieving data (critical when data too large to reside in memory)
Models vs. Patterns

- **Models**
  - Global summary of the dataset
  - Example: Fitting the line equation $Y=aX+c$ to all the data points

- **Patterns**
  - Local feature of the dataset. Limited to a subset of rows and attributes. Can be caused by concept drift.
  - Example: A small portion of the data above does not conform to $Y=aX^2+c$ but instead conform to $Y=aX+c$

- Boundary between models and patterns is not always clear
Types of Models

- **Prediction Model**
  - In a predictive model, one of the variable $Y$ should be predicted from the other variables $X_1, \ldots, X_p$
  - Also called *supervised learning*
  - If $Y$ is numerical, we call it *regression*
  - If $Y$ is categorical, we call it *classification*

- **Descriptive Model**
  - Aim is to produce a summary or description of the data
  - Also called *unsupervised learning*
  - Example: Clustering, data cubes

- **Models for Structure Data**
  - To model situations in which data items affect one another
  - Example: time series, sequences, spatial data
Principles of DM: Statistics

- **Statistics**
  - A discipline dedicated to data analysis
  - Try to fit the data to a mathematical model with various form of statistical testing
  - Often used in data mining for rules verification etc

- **Example**
  - Association rules are found based on **support** and **confidence**. However these two quantities tend to be high for items that occur frequently. For example, if two items A and B occur in the database 80% and 90% of the time, then the rules A→B will have expected support of 72% and confidence of 90% even if A and B occur together by random. To filter off such rules, we used the **chi-squared test** which test the independency between items.
Principles of DM: Information Theory

- **Information Theory**
  - is originally developed by *Claude Shannon* and applied in the area of communication
  - quantitative measurement of information using entropy or the minimum number of bits needed to encode a dataset
  - dataset is separated into a model and noise
  - minimize the total number of bits that is needed to represent the model and noise

- **Example**
  - Often used in decision tree building to avoid building over-complex decision tree
  - Often used to find deviants (which are noise)
Principles of DM: Others

- Theoretical CS
  - Many rules discovery problem in data mining are NP-hard
  - Solution: Used Approximate, Randomized, Online Algorithms

- Mathematical Programming
  - Lots of optimization in data mining too
  - Algorithms like neural net training, k-means clustering, SVMs are in fact mathematical programming

- Computational Geometry ...
OLAP Mining: Integration of Data Mining and Data Warehousing

- Data mining systems, DBMS, Data warehouse systems coupling
  - No coupling, loose-coupling, semi-tight-coupling, tight-coupling
- On-line analytical mining data
  - Integration of mining and OLAP technologies
- Interactive mining multi-level knowledge
  - Necessity of mining knowledge and patterns at different levels of abstraction by drilling/rolling, pivoting, slicing/dicing, etc.
- Integration of multiple mining functions
  - Characterized classification, first clustering and then association
An OLAM Architecture

Layer 4: User Interface
- Mining query
- Mining result

Layer 3: OLAP/OLAM
- User GUI API
- Data Cube API

Layer 2: MDDB
- OLAM Engine
- OLAP Engine
- MDDB
- Meta Data

Layer 1: Data Repository
- Databases
- Data Warehouse
- Data cleaning
- Data integration
- Filtering
- Filtering & Integration
- Database API
Major Issues in Data Mining (1)

- **Mining methodology**
  - Mining different kinds of knowledge from diverse data types, e.g., bio, stream, Web
  - Performance: efficiency, effectiveness, and scalability
  - Pattern evaluation: the interestingness problem
  - Incorporation of background knowledge
  - Handling noise and incomplete data
  - Parallel, distributed and incremental mining methods
  - Integration of the discovered knowledge with existing one: knowledge fusion
Major Issues in Data Mining (2)

- **User interaction**
  - Data mining query languages and ad-hoc mining
  - Expression and visualization of data mining results
  - Interactive mining of knowledge at multiple levels of abstraction

- **Applications and social impacts**
  - Domain-specific data mining & invisible data mining
  - Protection of data security, integrity, and privacy
Open Topics in DM & KDD

- Streaming data mining
- Biological data mining
- Privacy-preservation mining
- Data mining in distributed environments (e.g., P2P and sensor network environment - mining with incomplete data in dynamic situation)
- Web mining to detect rare events, emerging events, community
- ......
Summary

- Data mining: discovering interesting patterns from large amounts of data
- A natural evolution of database technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Mining can be performed in a variety of information repositories
- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.
- Data mining systems and architectures
- Major issues in data mining
A Brief History of Data Mining Society

- **1989 IJCAI Workshop on Knowledge Discovery in Databases (Piatetsky-Shapiro)**
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991)

- **1991-1994 Workshops on Knowledge Discovery in Databases**
  - Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1996)

- **1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98)**
  - Journal of Data Mining and Knowledge Discovery (1997)

- **1998 ACM SIGKDD, SIGKDD‘1999-2001 conferences, and SIGKDD Explorations**

- **More conferences on data mining**
Where to Find References?

- **Data mining and KDD (SIGKDD: CDROM)**
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations

- **Database systems (SIGMOD: CDROM)**
  - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
  - Journals: ACM-TODS, IEEE-TKDE, JIIS, J. ACM, etc.

- **AI & Machine Learning**
  - Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), etc.
  - Journals: Machine Learning, Artificial Intelligence, etc.

- **Statistics**
  - Conferences: Joint Stat. Meeting, etc.
  - Journals: Annals of statistics, etc.

- **Visualization**
  - Conference proceedings: CHI, ACM-SIGGraph, etc.
  - Journals: IEEE Trans. visualization and computer graphics, etc.
Recommended Reference Books

- R. Agrawal, J. Han, and H. Mannila, Readings in Data Mining: A Database Perspective, Morgan Kaufmann (in preparation)
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2001
- S. M. Weiss and N. Indurkhya, Predictive Data Mining, Morgan Kaufmann, 1998
References Addresses

- Searching references from/via
  - http://www.informatik.uni-trier.de/~ley/db/indices/t-form.html
  - http://www.google.com/
  - http://citeseer.nj.nec.com/
Homework

- Read Chapter 1 of the textbook
- Access Weka Website to familiarize with the Weka project and its source codes package