Kaynaklar

- Data Mining: Concepts and Techniques, Third Edition, Jiawei Han, Micheline Kamber, Jian Pei
- Data Mining: Practical Machine Learning Tools and Techniques, Third, Ian H. Witten, Eibe Frank, Mark A. Hall

Necessity Is the Mother of Invention

- Data explosion problem
  - Automated data collection tools, widely used database systems, computerized society, and the Internet lead to tremendous amounts of data accumulated and/or to be analyzed in databases, data warehouses, WWW, and other information repositories

- We are drowning in data, but starving for knowledge!

- Solution: Data warehousing and data mining
  - Data warehousing and on-line analytical processing (OLAP)
  - 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
  - Edgar F. Codd (1923-2003)
  - Structured English Query Language (SEQUEL), SQL

- 1980s:
  - Advanced data models (extended-relational, OQ, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)

- 1990s:
  - Data mining, data warehousing, multimedia databases
  - Web databases (Amazon, etc.)

- 2000s
  - Stream data management and mining
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems
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 (interesting patterns?)
- Data mining - a misnomer? (erro de nome)
- 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
  - Medical data mining
  - Bioinformatics and bio-data analysis

Example 1: 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

Example 2: Corporate Analysis & Risk Management

- Finance planning and asset evaluation
  - Cash flow analysis and prediction (feature development)
  - Contingent claim analysis to evaluate assets (componente do ativo)
  - Cross-sectional and time series analysis (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

Example 3: Fraud Detection & Mining Unusual Patterns

- Approaches:
  - Unsupervised Learning: Clustering
  - Supervised Learning: Neuronal Networks
  - Model construction for frauds
  - Outlier analysis

Example 3: Fraud Detection & Mining Unusual Patterns
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 (Vendor a varejo)
  - Analysts estimate that 38% of retail shrink is due to dishonest employees
- Anti-terrorism

Data Mining and Knowledge Discovery (KDD) Process

- Data mining—core of knowledge discovery process
- Data Cleaning
- Data Integration
- Databases
- Data Warehouse
- Task-relevant Data
- Data Mart
- Data Cleaning

Steps of a KDD Process (1)

- 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!)
- Understand data (statistics)
- Data reduction and transformation
  - Find useful features, dimensionality/variable reduction, invariant representation

Steps of a KDD Process (2)

- 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

KDD Sample

Overview of Data Mining Methods

- Automated Exploration/Discovery
  - E.g., discovering new market segments
  - Distance and probabilistic clustering algorithms
- Prediction/Classification
  - E.g., forecasting gross sales given current factors
  - Regression, neural networks, genetic algorithms
- Explanation/Description
  - E.g., characterizing customers by demographics and purchase history
  - Inductive decision trees, association rule systems
  - Focus is on induction of a model from specific examples
Data Mining Methods

Automated Exploration and Discovery
- Distance-based numerical clustering
  - metric grouping of examples (KNN)
  - graphical visualization can be used
- Bayesian clustering
  - search for the number of classes which result in best fit of a probability distribution to the data
- Unsupervised Learning

Prediction and Classification
- Function approximation (curve fitting)
- Classification (concept learning, pattern recognition)
- Methods:
  - Statistical regression
  - Artificial neural networks
  - Genetic algorithms
  - Nearest neighbour algorithms
- Supervised Learning

Unsupervised Learning

Income
Age

Bayesian clustering

Generalization
- The objective of learning is to achieve good generalization to new cases, otherwise just use a look-up table.
- Generalization can be defined as a mathematical interpolation or regression over a set of training points:

\[ f(x) \]

Classification
- Find ways to separate data items into pre-defined groups
- If we know X and Y belong together, find other things in same category
- Requires “training data”. Data items where group is known
- Profiling
- Technologies:
  - Generate decision trees (results are human understandable)
- Neural Nets

Association Rules
- Identify dependencies in the data
- “Find groups of items commonly purchased together”
- People who purchase fish are extraordinarily likely to purchase wine
- People who purchase Turkey are extraordinarily likely to purchase cranberries
- Uses:
  - Targeted marketing
- Technologies:
  - AIS, SETM, Hugin, TETRAD II

Clustering
- Find groups of similar data items
- Statistical techniques require some definition of “distance” (e.g., between travel profiles) within the background concepts and logical descriptions
- Uses:
  - Demographic analysis
- Technologies:
  - Self-Organizing Maps
  - Probability Densities
  - Conceptual Clustering

Generalization can be defined as a mathematical interpolation or regression over a set of training points:

\[ f(x) \]

“Group people with similar travel profiles”
- George, Patricia
- Jeff, Evelyn, Chris
- Rob
Sequential Associations

+ Find event sequences that are unusually likely
+ Requires “training” event list, known “interesting” events
+ Must be robust in the face of additional “noise” events

Uses:
+ Failure analysis and prediction

Technologies:
+ Dynamic programming (Dynamic time warping)
+ “Custom” algorithms

Deviation Detection

+ Find unexpected values, outliers
+ “Find unusual occurrences in IBM stock prices”

Uses:
+ Failure analysis
+ Anomaly discovery for analysis

Technologies:
+ Clustering/classification methods
+ Statistical techniques
+ Visualization

Data Mining and Business Intelligence

- Increasing potential to support business decisions
- End User
- Business Analyst
- Data Analyst
- Data Miner
- OLAP, MDX
- Data Sources
- Paper, Files, Information Providers, Database Systems, OLTP
- Making Decisions
- Data Presentation
- Visualization Techniques
- Data Mining
- Information Discovery
- Data Exploration
- Statistical Analysis, Querying and Reporting
- Data Warehouses / Data Marts
- OLAP, MDX

Data Mining Functionalities (1)

+ Multidimensional concept description: Characterization and discrimination
+ Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
+ Frequent patterns, association, correlation and causality
+ Smoking → Cancer (Correlation or causality?)
+ 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
+ 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: Data object that does not comply with the general behavior of the data
+ Noise or exception?

+ Trend and evolution analysis
  + Trend and deviation: e.g., regression analysis
+ Sequential pattern mining, periodicity analysis
+ Similarity-based analysis

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, and validated some hypotheses 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: Confluence of Multiple Disciplines

- Database Technology
- Statistics
- Machine Learning
- Data Mining
- Visualization
- Algorithm
- Other Disciplines

Data Mining: Classification Schemes

- General functionality
  - Descriptive data mining
  - Predictive data mining
- Different views lead to different classifications
  - Kinds of data to be mined
  - Kinds of knowledge to be discovered
  - Kinds of techniques utilized
  - Kinds of applications adapted

Data Mining from different perspectives

- Data to be mined
  - Object-oriented(relational), 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 warehouses, machine learning, statistics, visualization, etc.
- Applications adapted
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web-mining, etc.

Primitives that Define a Data Mining Task

- Task-relevant data
- Type of knowledge to be mined
- Background knowledge
- Pattern interestingness measurements (?)
- Visualization/presentation of discovered patterns

Primitive 1: Task-Relevant Data

- Database or data warehouse name
- Database tables or data warehouse cubes
- Condition for data selection
- Relevant attributes or dimensions
- Data grouping criteria
Primitive 2: Types of Knowledge to Be Mined

- Characterization (Categories)
- Discrimination
- Association
- Classification/Prediction
- Clustering
- Outlier analysis
- Other data mining tasks

Primitive 3: Background Knowledge

- Schema hierarchy (taxonomy)
  - E.g., street < city < province_or_state < country
- Set-grouping hierarchy
  - E.g., [20-39] = young, [40-59] = middle aged
- Operation-derived hierarchy
  - Email address: hagonzal@cs.uw.edu
    - login-name < department < university < country
- Rule-based hierarchy
  - low_profit_margin(X) <= price(X, P_1) and cost(X, P_2) and (P_1 - P_2) < $50

Primitive 4: Measurements of Pattern Interestingness

- Simplicity
  - E.g., (association) rule length, (decision) tree size
- Certainty
  - E.g., confidence, classification reliability or accuracy, certainty factor, rule strength, rule quality, discriminating weight, etc.
- Utility
  - Potential usefulness, e.g., support (association), noise threshold (description)
- Novelty
  - Not previously known, surprising (used to remove redundant rules)

Primitive 5: Presentation of Discovered Patterns

- Different backgrounds/uses may require different forms of representation
  - E.g., rules, tables, crosstabs, pie/bar chart, etc.
- Concept hierarchy is also important
  - Discovered knowledge might be more understandable when represented at high level of abstraction
- Interactive drill up/down, pivoting, slicing and dicing provide different perspectives to data
- Different kinds of knowledge require different representation: association, classification, clustering, etc.

Why Data Mining Query Language?

- Automated vs. query-driven?
  - Finding all the patterns autonomously in a database?—unrealistic because the patterns could be too many but uninteresting
- Data mining should be an interactive process
- User directs what to be mined
- Users must be provided with a set of primitives to be used to communicate with the data mining system
- Incorporating these primitives in a data mining query language
- More flexible user interaction
- Foundation for design of graphical user interface
- Standardization of data mining industry and practice

Architecture: Typical Data Mining System
**State of Commercial/Research Practice**

- Increasing use of data mining systems in financial community, marketing sectors, retailing
- Still have major problems with large, dynamic sets of data (need better integration with the databases)
- Most research emphasizes machine learning; little emphasis on database side (especially text)
- People achieving results are not likely to share knowledge

**Related Techniques: OLAP**

- On-Line Analytical Processing tools provide the ability to pose statistical and summary queries interactively (traditional On-Line Transaction Processing (OLTP) databases may take minutes or even hours to answer these queries)
- Advantages relative to data mining
  - Can obtain a wider variety of results
  - Generally faster to obtain results
- Disadvantages relative to data mining
  - User must "ask the right question"
  - Generally used to determine high-level statistical summaries, rather than specific relationships among instances

**OLAP: On-Line Analytical Processing**

- OLAP Functionality
  - Dimension selection
  - Rotation
  - Filtration
  - Hierarchies
  - Drill-downs to lower levels
  - Roll-ups to higher levels

**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/pivoting, slicing/dicing, etc.
- Integration of multiple mining functions
  - Characterized classification, first clustering and then association

**An OLAM Architecture**

**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/pivoting, slicing/dicing, etc.
- Integration of multiple mining functions
  - Characterized classification, first clustering and then association
Coupling Data Mining with DB/DW Systems

- No coupling—flat file processing, not recommended
- Loose coupling—fetching data from DB/DW
- Semi-tight coupling—enhanced DM performance
  - Provide efficient implementation of data mining primitives in a DB/DW system, e.g., sorting, indexing, aggregation, histogram analysis, multiway join, precomputation of some data functions
- Tight coupling—a uniform information processing environment
  - DM is smoothly integrated into a DB/DW system, mining query is optimized based on mining query, indexing, etc.

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
  - (constraints, taxonomy)
- Handling noise and incomplete data (preprocessing)
- Parallel, distributed and incremental mining methods
- Integration of the discovered knowledge with existing one

Summary

- Data mining: discovering interesting patterns from large amounts of data (DB)
- A natural evolution of database technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration (Data Warehouse), data selection (Data Mart), 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.
  - Subjective, requires expert knowledge
- Data mining systems and architectures

A Brief History of Data Mining Society

- 1989 IJCAI Workshop on Knowledge Discovery in Databases
  - Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1992)
- 1993–1994, Workshops on Knowledge Discovery in Databases
  - Advances in Knowledge Discovery and Data Mining (C. 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 (1995)
- ACM SIGKDD conferences since 1998 and SIGKDD Explorations
  - More conferences on data mining

Conferences and Journals on Data Mining

- ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD)
- SIAM Data Mining Conf. (SDM)
- IEEE Int. Conf. on Data Mining (ICDM)
- Conf. on Principles and Practice of Knowledge Discovery and Data Mining (PKDD)
- Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)

Journals:
- Data Mining and Knowledge Discovery (DAMI or DMKD)
- IEEE Trans. On Knowledge and Data Eng. (TKDE)
- KDD Explorations
- ACM Trans. on KDD
Where to Find References?—DBLP, CiteSeer, Google

- Data mining and KDD (SIGKDD: CD-ROM)
  - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-Da, PKDD, PAKDD, etc.
  - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM-TKDD
- Database systems (SIGMOD: ACM SIGMOD Anthology—CD-ROM)
  - Conferences: ACM-SIGMOD, ACM-PADS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
- AI & Machine Learning
  - Conferences: Machine Learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
  - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.