Knowledge Discovery and Data Mining

Spring 2021

Chap 13. Data Mining Trends and Research Frontiers

Jiawei Han, Micheline Kamber and Jian Pei, Data Mining: Concepts and Techniques, 3rd ed., The Morgan Kaufmann Series in Data Management Systems Morgan Kaufmann Publishers, July 2011. ISBN 978-0123814791

Tuong Le, PhD

Outline

- ❑ Mining Complex Types of Data
- ❑ Other Methodologies of Data Mining
- ❑ Data Mining Applications
- ❑ Data Mining and Society
- ❑ Data Mining Trends
- ❑ Summary

- ❑ Mining Sequence Data
  - ▪ Mining Time Series
  - ▪ Mining Symbolic Sequences
  - ▪ Mining Biological Sequences
- ❑ Mining Graphs and Networks
- ❑ Mining Other Kinds of Data

Mining Sequence Data

- Similarity Search in Time Series Data
  - Subsequence match, dimensionality reduction, query-based similarity search, motif-based similarity search
- Regression and Trend Analysis in Time-Series Data
  - long term + cyclic + seasonal variation + random movements
- Sequential Pattern Mining in Symbolic Sequences
  - GSP, PrefixSpan, constraint-based sequential pattern mining
- Sequence Classification
  - Feature-based vs. sequence-distance-based vs. model-based
- Alignment of Biological Sequences
  - Pair-wise vs. multi-sequence alignment, substitution matirces, BLAST
- Hidden Markov Model for Biological Sequence Analysis
  - Markov chain vs. hidden Markov models, forward vs. Viterbi vs. Baum-Welch algorithms

Mining Graphs and Networks

- Graph Pattern Mining
  - Frequent subgraph patterns, closed graph patterns, gSpan vs. CloseGraph
- Statistical Modeling of Networks
  - Small world phenomenon, power law (log-tail) distribution, densification
- Clustering and Classification of Graphs and Homogeneous Networks
  - Clustering: Fast Modularity vs. SCAN
  - Classification: model vs. pattern-based mining
- Clustering, Ranking and Classification of Heterogeneous Networks
  - RankClus, RankClass, and meta path-based, user-guided methodology
- Role Discovery and Link Prediction in Information Networks
  - PathPredict
- Similarity Search and OLAP in Information Networks: PathSim, GraphCube
- Evolution of Social and Information Networks: EvoNetClus

Mining Other Kinds of Data

- Mining Spatial Data
  - Spatial frequent/co-located patterns, spatial clustering and classification
- Mining Spatiotemporal and Moving Object Data
  - Spatiotemporal data mining, trajectory mining, periodica, swarm, …
- Mining Cyber-Physical System Data
  - Applications: healthcare, air-traffic control, flood simulation
- Mining Multimedia Data
  - Social media data, geo-tagged spatial clustering, periodicity analysis, …
- Mining Text Data
  - Topic modeling, i-topic model, integration with geo- and networked data
- Mining Web Data
  - Web content, web structure, and web usage mining
- Mining Data Streams
  - Dynamics, one-pass, patterns, clustering, classification, outlier detection

Outline

- Mining Complex Types of Data
- Other Methodologies of Data Mining
  - Statistical Data Mining
  - Views on Data Mining Foundations
  - Visual and Audio Data Mining
- Data Mining Applications
- Data Mining and Society
- Data Mining Trends
- Summary

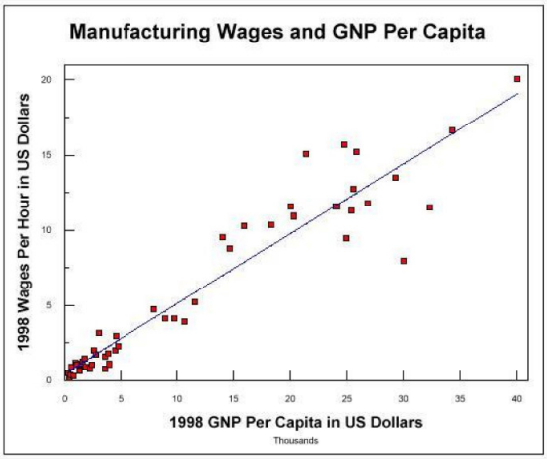
Major Statistical Data Mining Methods

- ❑ Regression
- ❑ Generalized Linear Model
- ❑ Analysis of Variance
- ❑ Mixed-Effect Models
- ❑ Factor Analysis
- ❑ Discriminant Analysis
- ❑ Survival Analysis

Statistical Data Mining (1)

- ❑ There are many well-established statistical techniques for data analysis, particularly for numeric data
  - applied extensively to data from scientific experiments and data from economics and the social sciences

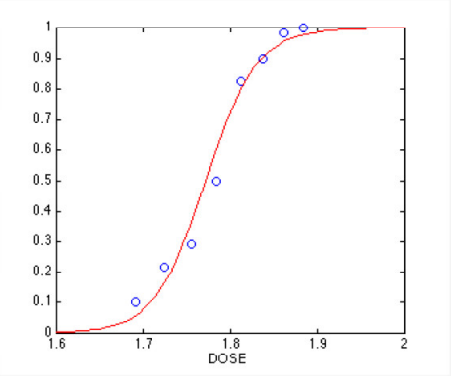
- ❑ **Regression**
  - predict the value of a response (dependent) variable from one or more predictor (independent) variables where the variables are numeric
  - forms of regression: linear, multiple, weighted, polynomial, nonparametric, and robust



Scientific and Statistical Data Mining (2)

❑ **Generalized linear models**

- allow a categorical response variable (or some transformation of it) to be related to a set of predictor variables
- similar to the modeling of a numeric response variable using linear regression
- include logistic regression and Poisson regression



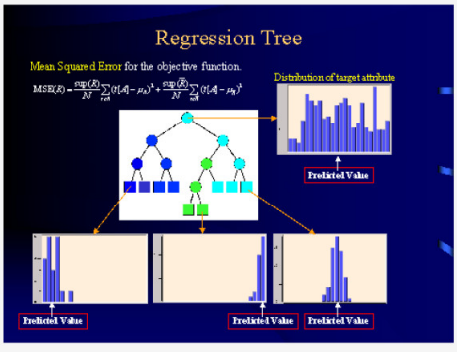
❑ **Mixed-effect models**

- For analyzing grouped data, i.e. data that can be classified according to one or more grouping variables
- Typically describe relationships between a response variable and some covariates in data grouped according to one or more factors

Scientific and Statistical Data Mining (3)

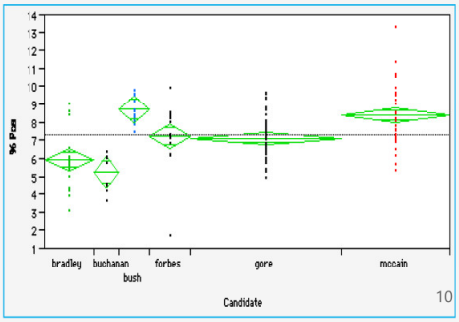
❑ **Regression trees**

- Binary trees used for classification and prediction
- Similar to decision trees:Tests are performed at the internal nodes
- In a regression tree the mean of the objective attribute is computed and used as the predicted value



❑ **Analysis of variance**

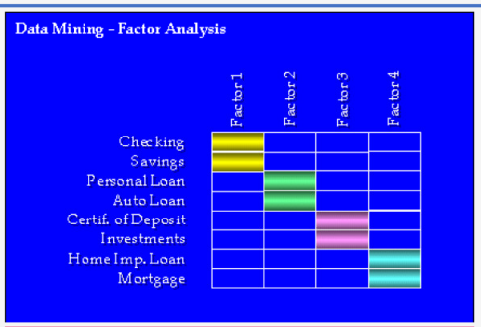
- Analyze experimental data for two or more populations described by a numeric response variable and one or more categorical variables (factors)

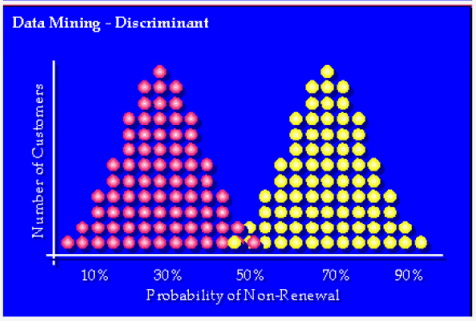


Statistical Data Mining (4)

- **Factor analysis**
  - determine which variables are combined to generate a given factor
  - e.g., for many psychiatric data, one can indirectly measure other quantities (such as test scores) that reflect the factor of interest

- **Discriminant analysis**
  - predict a categorical response variable, commonly used in social science
  - Attempts to determine several discriminant functions (linear combinations of the independent variables) that discriminate among the groups defined by the response variable

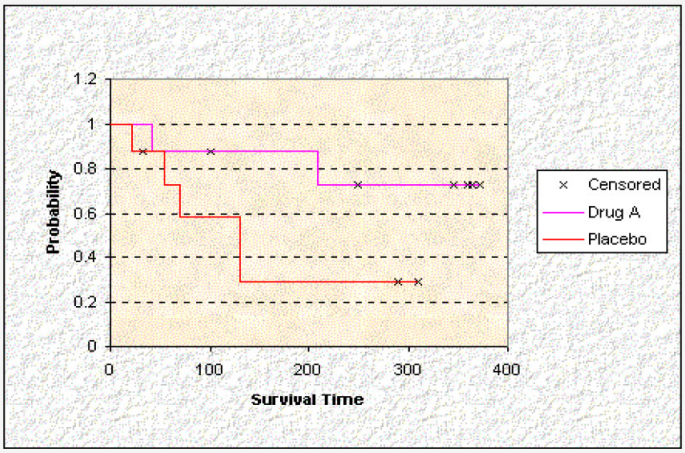




www.spss.com/datamine/factor.htm

Statistical Data Mining (5)

- **Time series**: many methods such as autoregression, ARIMA (Autoregressive integrated moving-average modeling), long memory time-series modeling
- **Quality control:** displays group summary charts
- **Survival analysis**
  - Predicts the probability that a patient undergoing a medical treatment would survive at least to time *t* (life span prediction)

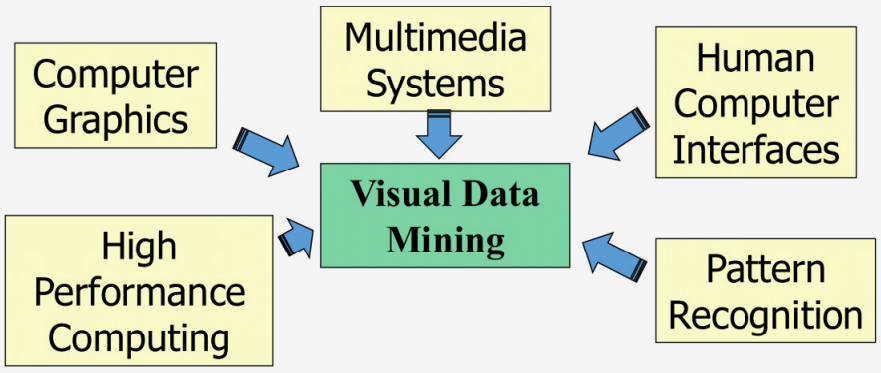


Views on Data Mining Foundations

- Data reduction
  - Basis of data mining: Reduce data representation
  - Trades accuracy for speed in response
- Data compression
  - Basis of data mining: Compress the given data by encoding in terms of bits, association rules, decision trees, clusters, etc.
- Probability and statistical theory
  - Basis of data mining: Discover joint probability distributions of random variables
- Microeconomic view
  - A view of utility: Finding patterns that are interesting only to the extent in that they can be used in the decision-making process of some enterprise
- Pattern Discovery and Inductive databases
  - Basis of data mining: Discover patterns occurring in the database, such as associations, classification models, sequential patterns, etc.
  - Data mining is the problem of performing inductive logic on databases
  - The task is to query the data and the theory (i.e., patterns) of the database
  - Popular among many researchers in database systems

Visual Data Mining

- Visualization: Use of computer graphics to create visual images which aid in the understanding of complex, often massive representations of data
- Visual Data Mining: discovering implicit but useful knowledge from large data sets using visualization techniques



Audio Data Mining

- Uses audio signals to indicate the patterns of data or the features of data mining results
- An interesting alternative to visual mining
- An inverse task of mining audio (such as music) databases which is to find patterns from audio data
- Visual data mining may disclose interesting patterns using graphical displays, but requires users to concentrate on watching patterns
- Instead, transform patterns into sound and music and listen to pitches, rhythms, tune, and melody in order to identify anything interesting or unusual

Outline

- Mining Complex Types of Data
- Other Methodologies of Data Mining
- **Data Mining Applications**
- Data Mining and Society
- Data Mining Trends
- Summary

Data Mining Applications

- Data mining: A young discipline with broad and diverse applications
  - There still exists a nontrivial gap between generic data mining methods and effective and scalable data mining tools for domain-specific applications
- Some application domains (briefly discussed here)
  - Data Mining for Financial data analysis
  - Data Mining for Retail and Telecommunication Industries
  - Data Mining in Science and Engineering
  - Data Mining for Intrusion Detection and Prevention
  - Data Mining and Recommender Systems

Data Mining for Financial Data Analysis

- Financial data collected in banks and financial institutions are often relatively complete, reliable, and of high quality
- Design and construction of data warehouses for multidimensional data analysis and data mining
  - View the debt and revenue changes by month, by region, by sector, and by other factors
  - Access statistical information such as max, min, total, average, trend, etc.
- Loan payment prediction/consumer credit policy analysis
  - feature selection and attribute relevance ranking
  - Loan payment performance
  - Consumer credit rating
- Classification and clustering of customers for targeted marketing
  - Multidimensional segmentation by nearest-neighbor, classification, decision trees, etc. to identify customer groups or associate a new customer to an appropriate customer group
- Detection of money laundering and other financial crimes
  - integration of from multiple DBs (e.g., bank transactions, federal/state crime history DBs)
  - Tools: data visualization, linkage analysis, classification, clustering tools, outlier analysis, and sequential pattern analysis tools (find unusual access sequences)

Data Mining for Retail & Telcomm. Industries

- ❑ Retail industry: huge amounts of data on sales, customer shopping history, e-commerce, etc.
- ❑ Applications of retail data mining
  - Identify customer buying behaviors
  - Discover customer shopping patterns and trends
  - Improve the quality of customer service
  - Achieve better customer retention and satisfaction
  - Enhance goods consumption ratios
  - Design more effective goods transportation and distribution policies
- ❑ Telcomm. and many other industries: Share many similar goals and expectations of retail data mining

Data Mining Practice for Retail Industry

- ❑ Design and construction of data warehouses
- ❑ Multidimensional analysis of sales, customers, products, time, and region
- ❑ Analysis of the effectiveness of sales campaigns
- ❑ Customer retention: Analysis of customer loyalty
  - Use customer loyalty card information to register sequences of purchases of particular customers
  - Use sequential pattern mining to investigate changes in customer consumption or loyalty
  - Suggest adjustments on the pricing and variety of goods
- ❑ Product recommendation and cross-reference of items
- ❑ Fraudulent analysis and the identification of usual patterns
- ❑ Use of visualization tools in data analysis

Data Mining in Science and Engineering

- ❑ Data warehouses and data preprocessing
  - Resolving inconsistencies or incompatible data collected in diverse environments and different periods (e.g. eco-system studies)
- ❑ Mining complex data types
  - Spatiotemporal, biological, diverse semantics and relationships
- ❑ Graph-based and network-based mining
  - Links, relationships, data flow, etc.
- ❑ Visualization tools and domain-specific knowledge
- ❑ Other issues
  - Data mining in social sciences and social studies: text and social media
  - Data mining in computer science: monitoring systems, software bugs, network intrusion

Data Mining for Intrusion Detection and Prevention

- ❑ Majority of intrusion detection and prevention systems use
  - Signature-based detection: use signatures, attack patterns that are preconfigured and predetermined by domain experts
  - Anomaly-based detection: build profiles (models of normal behavior) and detect those that are substantially deviate from the profiles
- ❑ What data mining can help
  - New data mining algorithms for intrusion detection
  - Association, correlation, and discriminative pattern analysis help select and build discriminative classifiers
  - Analysis of stream data: outlier detection, clustering, model shifting
  - Distributed data mining
  - Visualization and querying tools

Data Mining and Recommender Systems

- ❑ Recommender systems: Personalization, making product recommendations that are likely to be of interest to a user
- ❑ Approaches: Content-based, collaborative, or their hybrid
  - Content-based: Recommends items that are similar to items the user preferred or queried in the past
  - Collaborative filtering: Consider a user's social environment, opinions of other customers who have similar tastes or preferences
- ❑ Data mining and recommender systems
  - Users C × items S: extract from known to unknown ratings to predict user-item combinations
  - Memory-based method often uses k-nearest neighbor approach
  - Model-based method uses a collection of ratings to learn a model (e.g., probabilistic models, clustering, Bayesian networks, etc.)
  - Hybrid approaches integrate both to improve performance (e.g., using ensemble)

Outline

- ❑ Mining Complex Types of Data
- ❑ Other Methodologies of Data Mining
- ❑ Data Mining Applications
- ❑ **Data Mining and Society**
- ❑ **Data Mining Trends**
- ❑ Summary

Ubiquitous and Invisible Data Mining

- Ubiquitous Data Mining
  - Data mining is used everywhere, e.g., online shopping
  - Ex. Customer relationship management (CRM)
- Invisible Data Mining
  - Invisible: Data mining functions are built in daily life operations
  - Ex. Google search: Users may be unaware that they are examining results returned by data
  - Invisible data mining is highly desirable
  - Invisible mining needs to consider efficiency and scalability, user interaction, incorporation of background knowledge and visualization techniques, finding interesting patterns, real-time, …
  - Further work: Integration of data mining into existing business and scientific technologies to provide domain-specific data mining tools

Privacy, Security and Social Impacts of Data Mining

- Many data mining applications do not touch personal data
  - E.g., meteorology, astronomy, geography, geology, biology, and other scientific and engineering data
- Many DM studies are on developing scalable algorithms to find general or statistically significant patterns, not touching individuals
- The real privacy concern: unconstrained access of individual records, especially privacy-sensitive information
- Method 1: Removing sensitive IDs associated with the data
- Method 2: Data security-enhancing methods
  - Multi-level security model: permit to access to only authorized level
  - Encryption: e.g., *blind signatures*, *biometric encryption*, and *anonymous databases* (personal information is encrypted and stored at different locations)
- Method 3: Privacy-preserving data mining methods

Privacy-Preserving Data Mining

- Privacy-preserving (privacy-enhanced or privacy-sensitive) mining:
  - Obtaining valid mining results without disclosing the underlying sensitive data values
  - Often needs trade-off between information loss and privacy
- Privacy-preserving data mining methods:
  - Randomization (e.g., perturbation): Add noise to the data in order to mask some attribute values of records
  - K-anonymity and l-diversity: Alter individual records so that they cannot be uniquely identified
    - k-anonymity: Any given record maps onto at least k other records
    - l-diversity: enforcing intra-group diversity of sensitive values
  - Distributed privacy preservation: Data partitioned and distributed either horizontally, vertically, or a combination of both
  - Downgrading the effectiveness of data mining: The output of data mining may violate privacy
    - Modify data or mining results, e.g., hiding some association rules or slightly distorting some classification models

Trends of Data Mining

- Application exploration: Dealing with application-specific problems
- Scalable and interactive data mining methods
- Integration of data mining with Web search engines, database systems, data warehouse systems and cloud computing systems
- Mining social and information networks
- Mining spatiotemporal, moving objects and cyber-physical systems
- Mining multimedia, text and web data
- Mining biological and biomedical data
- Data mining with software engineering and system engineering
- Visual and audio data mining
- Distributed data mining and real-time data stream mining
- Privacy protection and information security in data mining

Summary

- We present a high-level overview of mining complex data types
- Statistical data mining methods, such as regression, generalized linear models, analysis of variance, etc., are popularly adopted
- Researchers also try to build theoretical foundations for data mining
- Visual/audio data mining has been popular and effective
- Application-based mining integrates domain-specific knowledge with data analysis techniques and provide mission-specific solutions
- Ubiquitous data mining and invisible data mining are penetrating our data lives
- Privacy and data security are importance issues in data mining, and privacy-preserving data mining has been developed recently
- Our discussion on trends in data mining shows that data mining is a promising, young field, with great, strategic importance