

Data Mining: Concepts and Techniques

(3rd ed.)

— Chapter 13 —

Jiawei Han, Micheline Kamber, and Jian Pei
University of Illinois at Urbana-Champaign &
Simon Fraser University


©2011 Han, Kamber & Pei. All rights reserved.



Chapter 13: Data Mining Trends and Research Frontiers

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

Mining Complex Types of Data

- 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 matrices, 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

Chapter 13: Data Mining Trends and Research Frontiers

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

Other Methodologies of Data Mining

- Statistical Data Mining 
- Views on Data Mining Foundations
- Visual and Audio Data Mining

Major Statistical Data Mining Methods

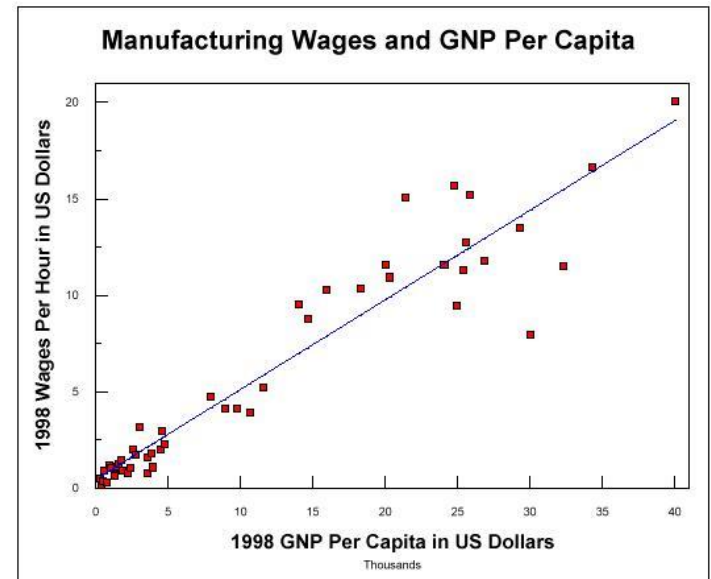
- Regression
- Generalized Linear Model
- Analysis of Variance
- Mixed-Effect Models
- Factor Analysis
- Discriminant Analysis
- Survival Analysis

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



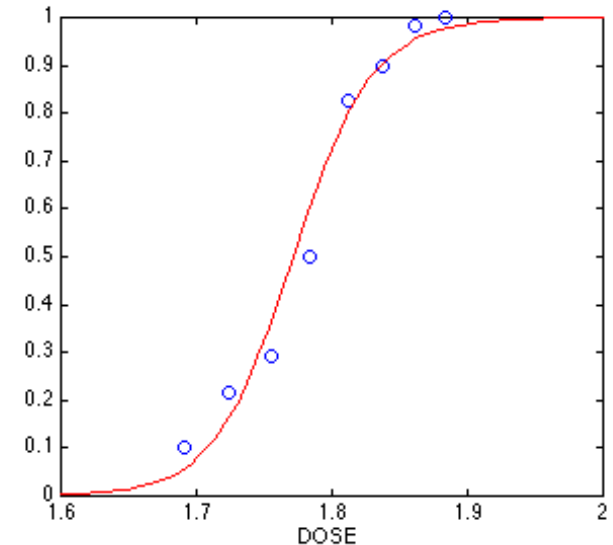
Scientific and Statistical Data Mining (2)

■ 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



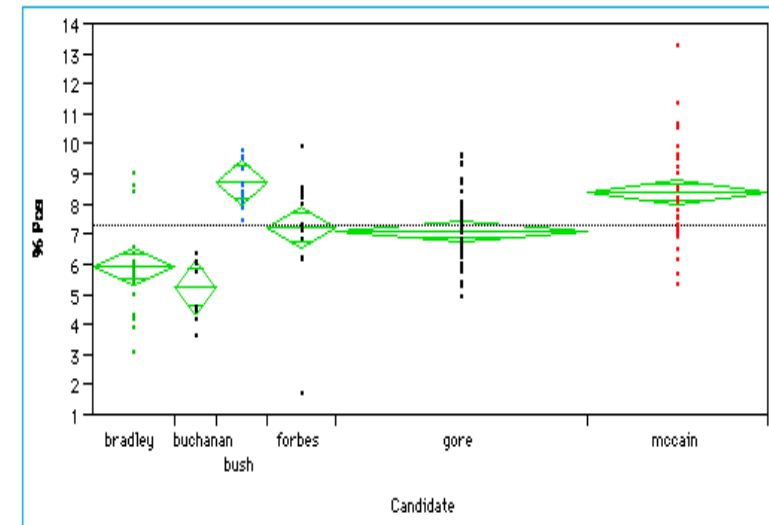
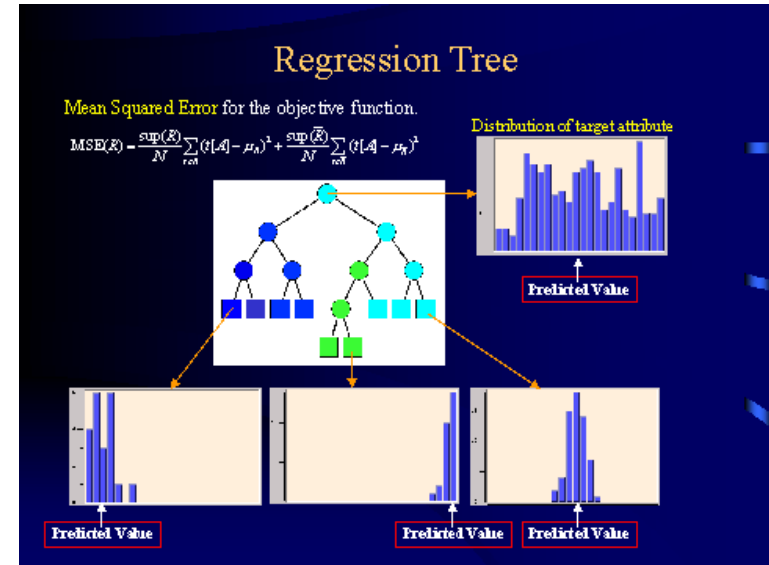
Scientific and Statistical Data Mining (3)

■ 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)

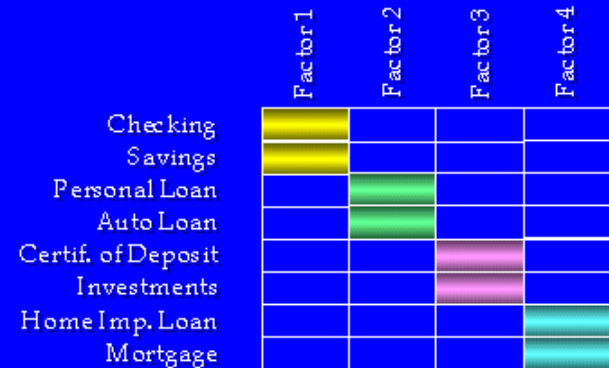
■ Factor analysis

- 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

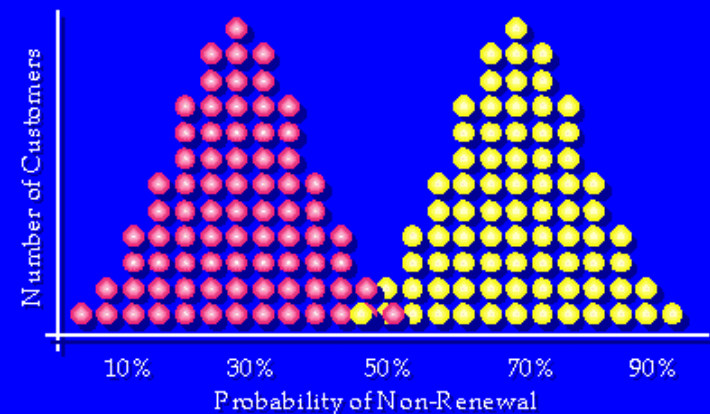
■ Discriminant analysis

- 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

Data Mining - Factor Analysis



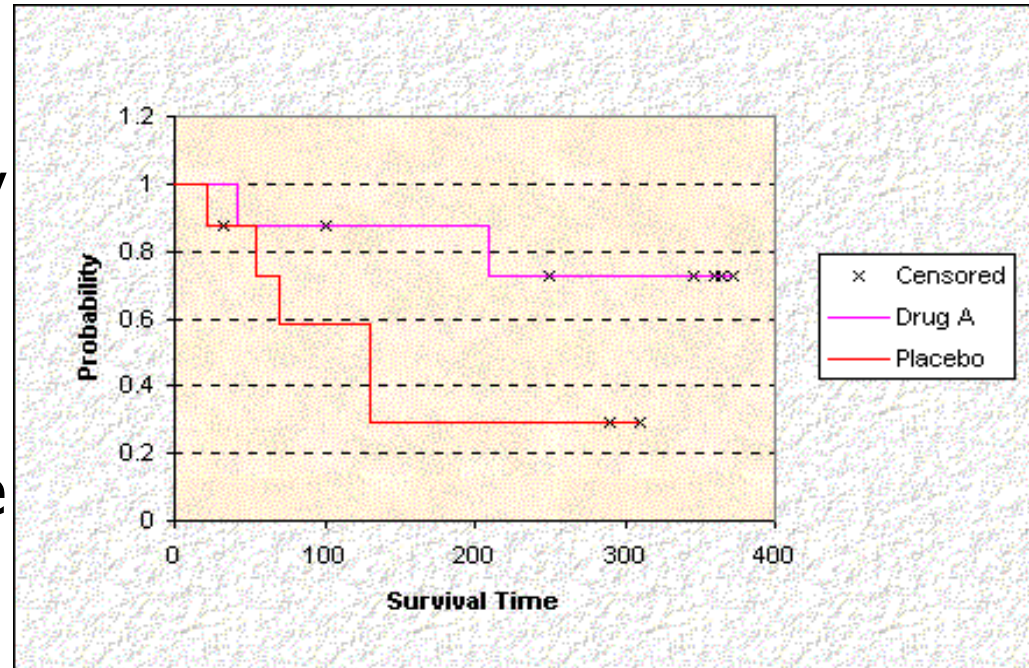
Data Mining - Discriminant



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)



Other Methodologies of Data Mining

- Statistical Data Mining
- Views on Data Mining Foundations
- Visual and Audio Data Mining




Views on Data Mining Foundations (I)

- 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

Views on Data Mining Foundations (II)

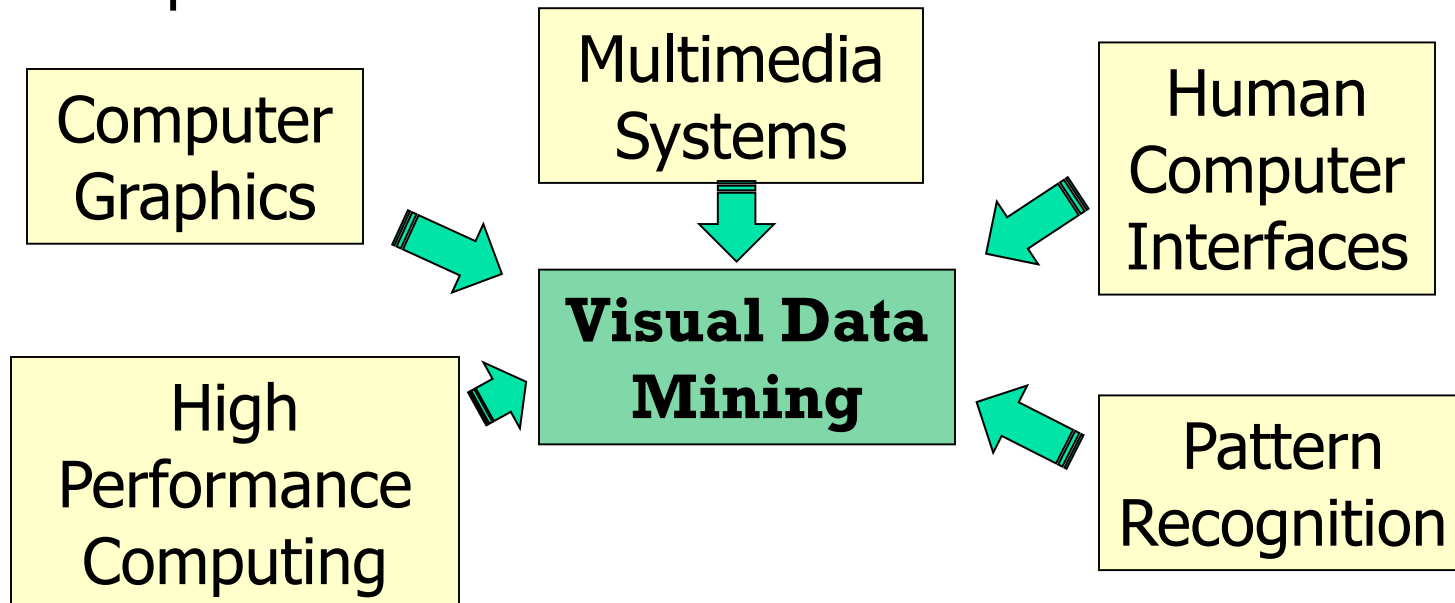
- Microeconomic view
 - A view of utility: Finding patterns that are interesting only to the extent in that they can be used in the decision-making process of some enterprise
- Pattern Discovery and Inductive databases
 - Basis of data mining: Discover patterns occurring in the database, such as associations, 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

Other Methodologies of Data Mining

- Statistical Data Mining
- Views on Data Mining Foundations
- Visual and Audio Data Mining 

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



Visualization

- Purpose of Visualization
 - Gain insight into an information space by mapping data onto graphical primitives
 - Provide qualitative overview of large data sets
 - Search for patterns, trends, structure, irregularities, relationships among data.
 - Help find interesting regions and suitable parameters for further quantitative analysis.
 - Provide a visual proof of computer representations derived

Visual Data Mining & Data Visualization

- Integration of visualization and data mining
 - data visualization
 - data mining result visualization
 - data mining process visualization
 - interactive visual data mining
- Data visualization
 - Data in a database or data warehouse can be viewed
 - at different levels of abstraction
 - as different combinations of attributes or dimensions
 - Data can be presented in various visual forms

Data Mining Result Visualization

- Presentation of the results or knowledge obtained from data mining in visual forms
- Examples
 - Scatter plots and boxplots (obtained from descriptive data mining)
 - Decision trees
 - Association rules
 - Clusters
 - Outliers
 - Generalized rules

The screenshot displays the SIMULINK 4.0 software interface. The main window shows a 3D bar chart titled "AVERAGE RAINFALL (by REGION and COUNTRY)". The chart has three axes: "Region" (North, South, East, West), "Country" (USA, UK, France, Germany, Italy), and "Rainfall" (mm/year). The bars are colored blue and red, representing different rainfall levels. A legend on the right indicates that blue bars represent "Rainfall (mm/year)" and red bars represent "Rainfall (mm/year)".

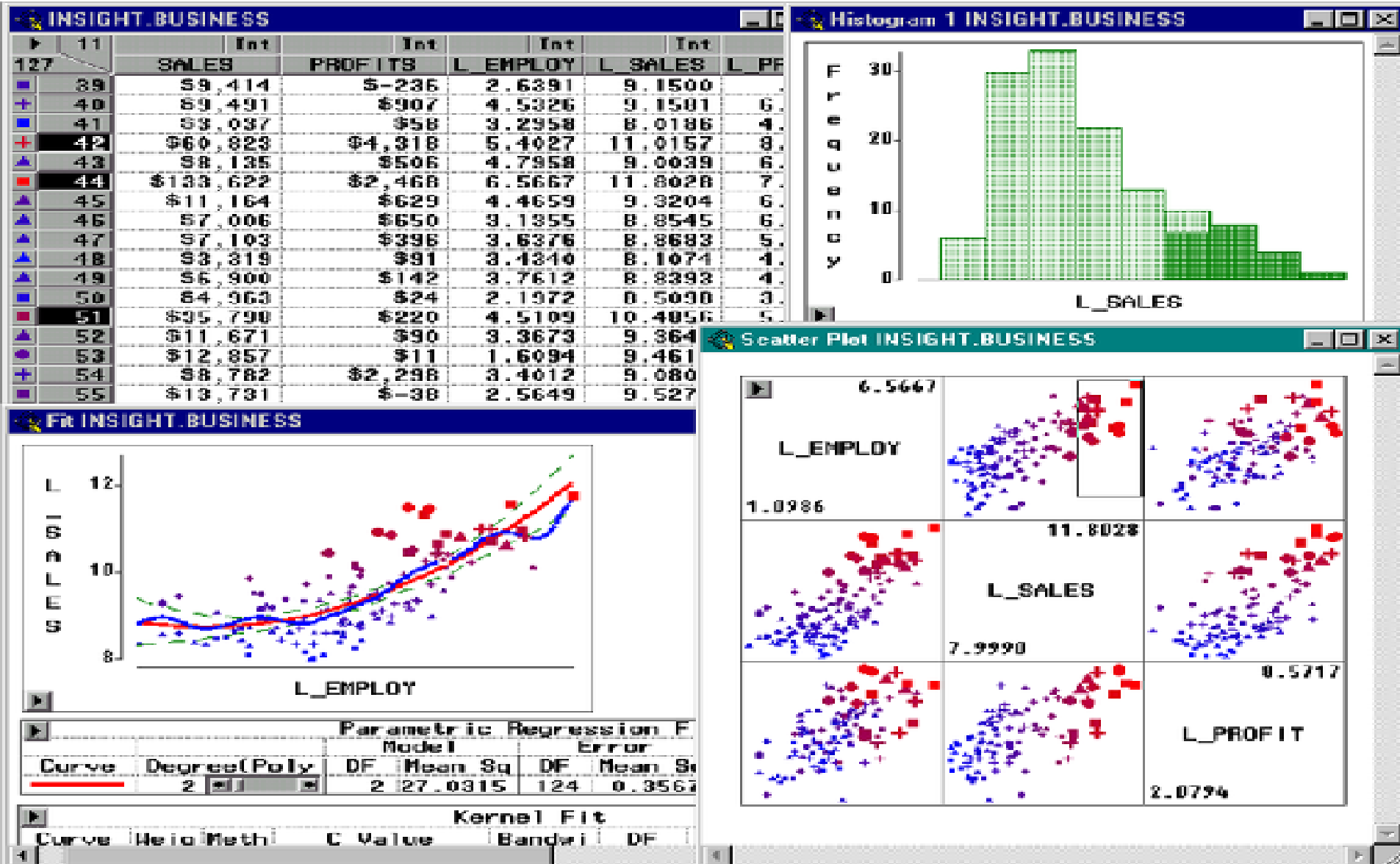
Below the main chart, there is a section titled "AVERAGE RAINFALL CHANGE (by REGION and COUNTRY)". This section contains a 3D bar chart with a legend and a 3D bar chart. The legend indicates that blue bars represent "Rainfall (mm/year)" and red bars represent "Rainfall (mm/year)".

On the left side of the interface, there is a data table titled "The Rainfall (mm/year) in 1990". The table has three columns: "COUNTRY", "REGION", and "RAINFALL". The data is as follows:

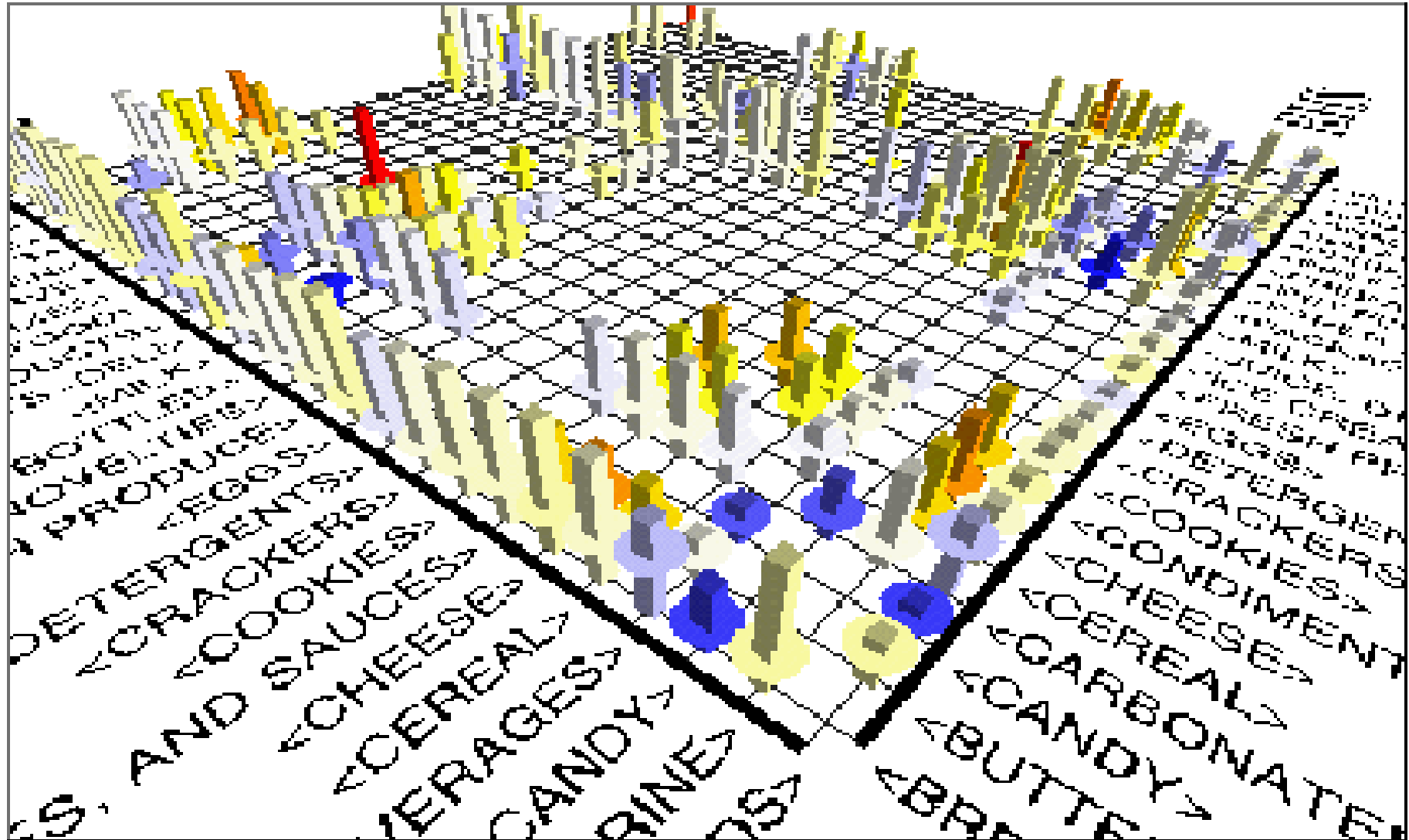
| COUNTRY | REGION | RAINFALL |
|---------|--------|----------|
| USA | NORTH | 1000 |
| USA | SOUTH | 1000 |
| USA | EAST | 1000 |
| USA | WEST | 1000 |
| UK | NORTH | 1000 |
| UK | SOUTH | 1000 |
| UK | EAST | 1000 |
| UK | WEST | 1000 |
| France | NORTH | 1000 |
| France | SOUTH | 1000 |
| France | EAST | 1000 |
| France | WEST | 1000 |
| Germany | NORTH | 1000 |
| Germany | SOUTH | 1000 |
| Germany | EAST | 1000 |
| Germany | WEST | 1000 |
| Italy | NORTH | 1000 |
| Italy | SOUTH | 1000 |
| Italy | EAST | 1000 |
| Italy | WEST | 1000 |

At the bottom of the interface, there is a section titled "AVERAGE RAINFALL CHANGE (by REGION and COUNTRY)". This section contains a 3D bar chart with a legend and a 3D bar chart. The legend indicates that blue bars represent "Rainfall (mm/year)" and red bars represent "Rainfall (mm/year)".

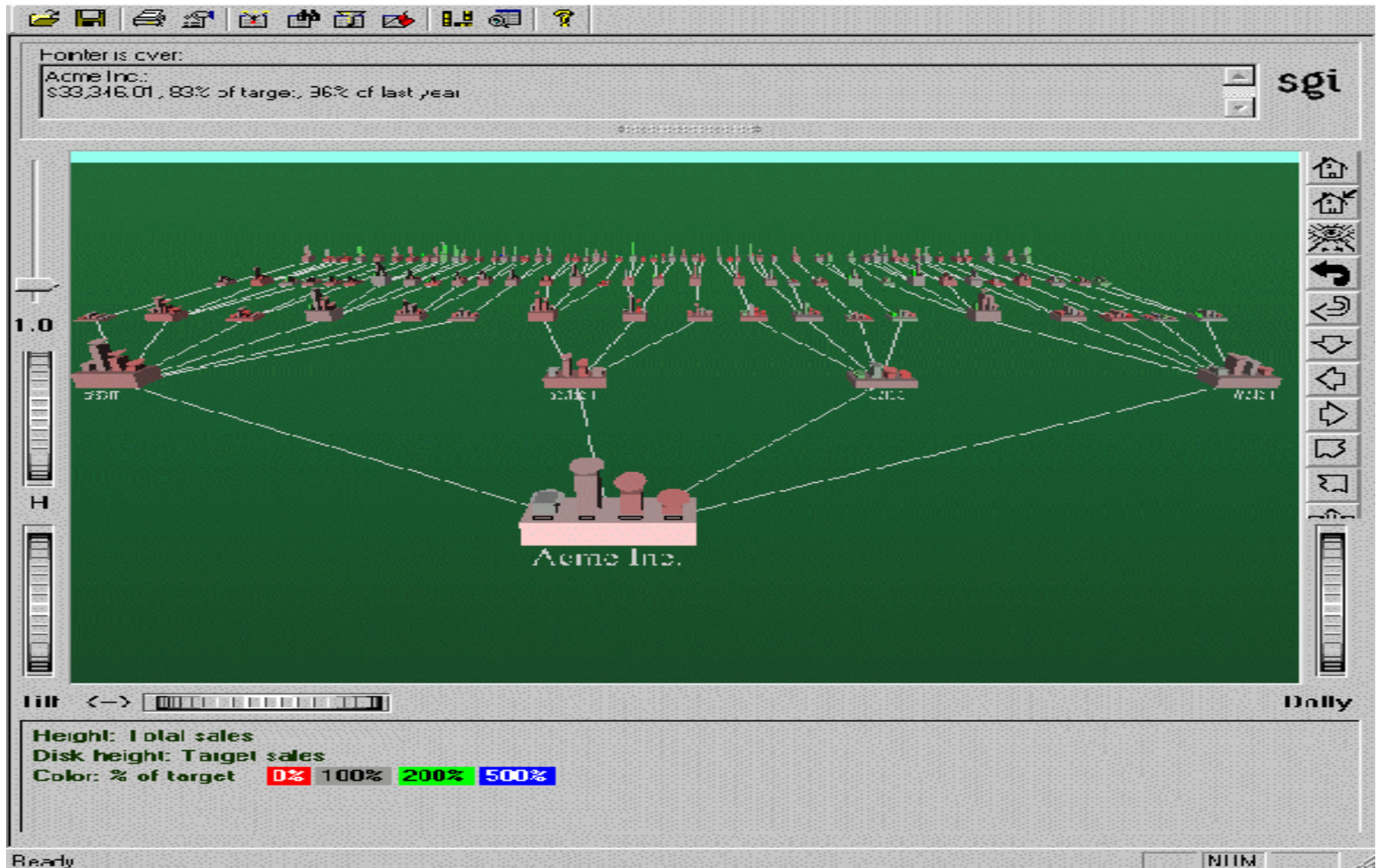
Visualization of Data Mining Results in SAS Enterprise Miner: Scatter Plots



Visualization of Association Rules in SGI/MineSet 3.0



Visualization of a **Decision Tree** in SGI/MineSet 3.0



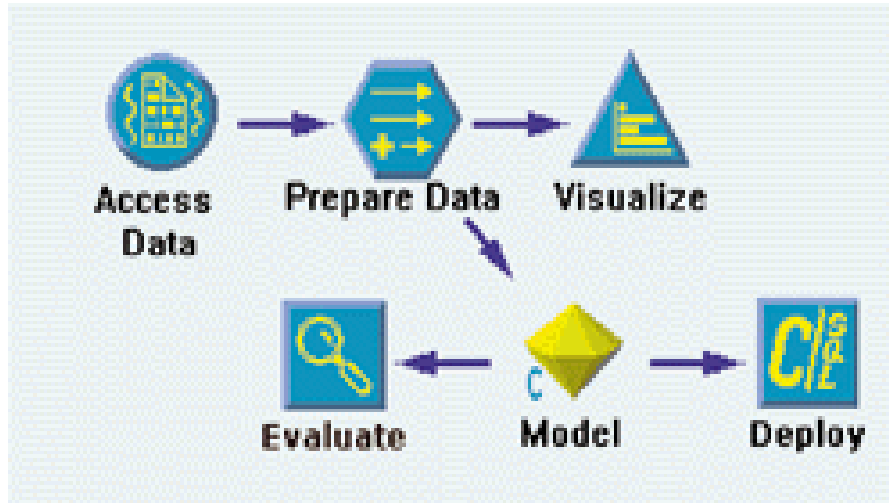
Visualization of **Cluster Grouping** in IBM Intelligent Miner



Data Mining Process Visualization

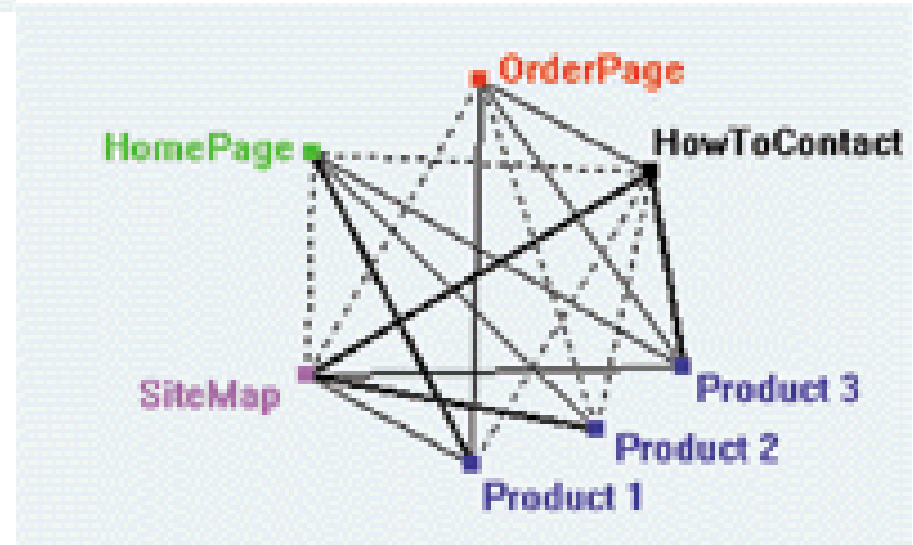
- Presentation of the various processes of data mining in visual forms so that users can see
 - Data extraction process
 - Where the data is extracted
 - How the data is cleaned, integrated, preprocessed, and mined
 - Method selected for data mining
 - Where the results are stored
 - How they may be viewed

Visualization of **Data Mining Processes** by Clementine



See your solution
discovery
process clearly

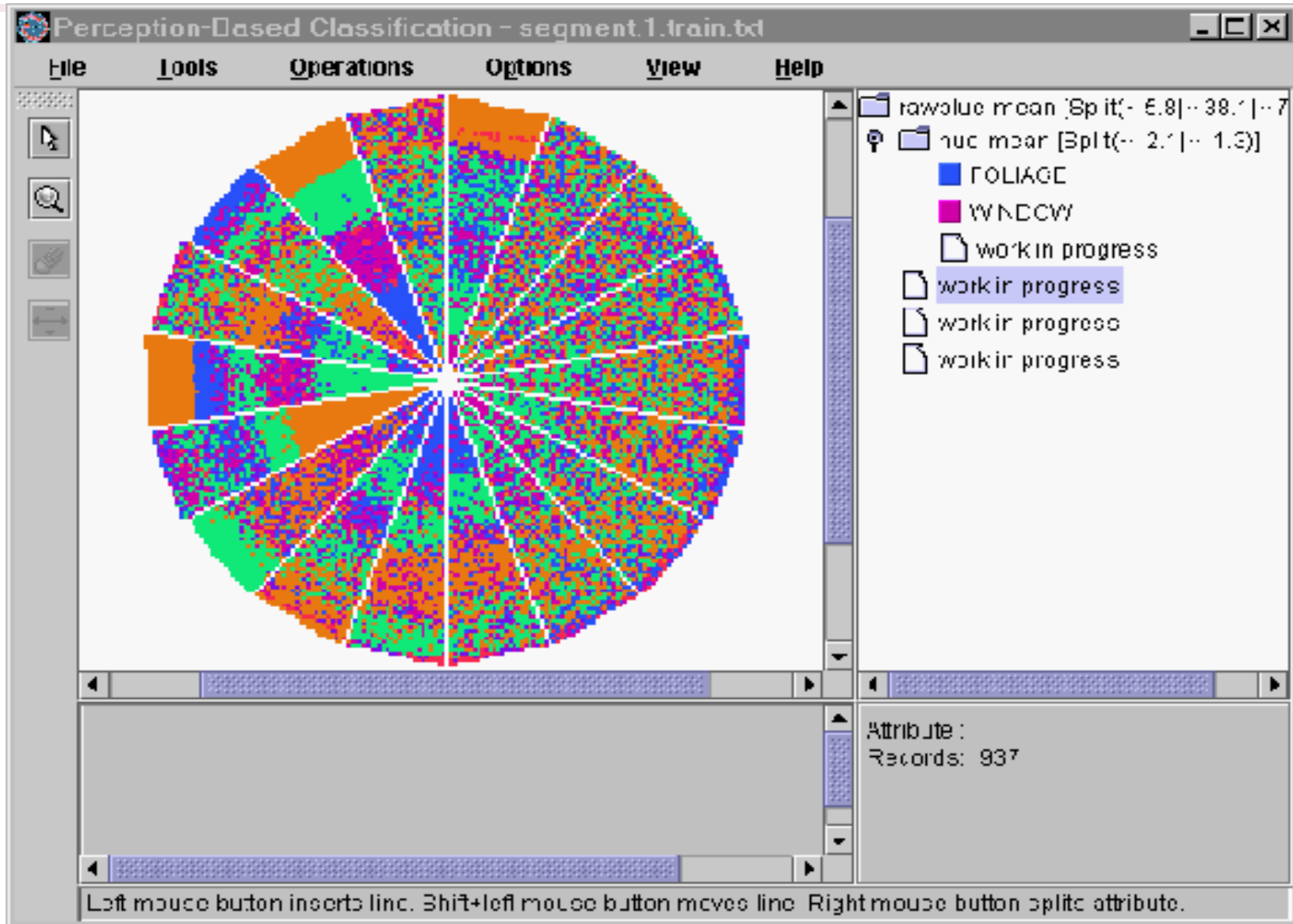
Understand
variations with
visualized data



Interactive Visual Data Mining

- Using visualization tools in the data mining process to help users make smart data mining decisions
- Example
 - Display the data distribution in a set of attributes using colored sectors or columns (depending on whether the whole space is represented by either a circle or a set of columns)
 - Use the display to which sector should first be selected for classification and where a good split point for this sector may be

Interactive Visual Mining by Perception-Based Classification (PBC)



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

Chapter 13: Data Mining Trends and Research Frontiers

- 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 (I)

- 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

Data Mining for Financial Data Analysis (II)

- 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 (I)

- 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 $C \times$ 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)

Chapter 13: Data Mining Trends and Research Frontiers

- 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

Chapter 13: Data Mining Trends and Research Frontiers

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

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

Chapter 13: Data Mining Trends and Research Frontiers

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

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

References and Further Reading

- ❖ The books lists a lot of references for further reading. Here we only list a few books
- E. Alpaydin. *Introduction to Machine Learning*, 2nd ed., MIT Press, 2011
- S. Chakrabarti. *Mining the Web: Statistical Analysis of Hypertext and Semi-Structured Data*. Morgan Kaufmann, 2002
- R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*, 2ed., Wiley-Interscience, 2000
- D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010.
- U. Fayyad, G. Grinstein, and A. Wierse (eds.), *Information Visualization in Data Mining and Knowledge Discovery*, Morgan Kaufmann, 2001
- J. Han, M. Kamber, J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 3rd ed. 2011
- T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer-Verlag, 2009
- D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.
- B. Liu. *Web Data Mining*, Springer 2006.
- T. M. Mitchell. *Machine Learning*, McGraw Hill, 1997
- M. Newman. *Networks: An Introduction*. Oxford University Press, 2010.
- P.-N. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining*, Wiley, 2005
- I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, 2nd ed. 2005

