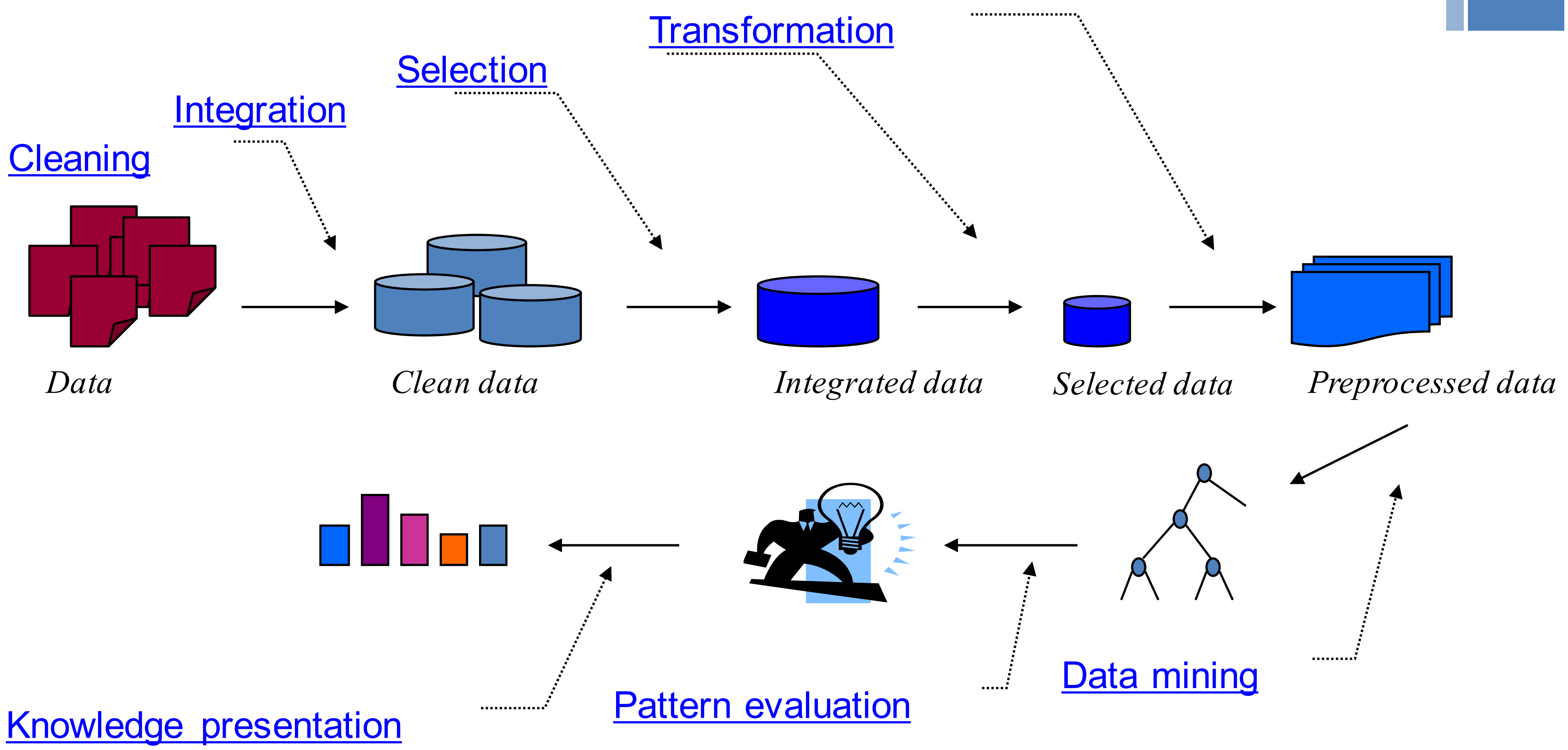
Knowledge discovery

Preliminary definitions

- Data mining why?
  - Huge amounts of data in electronic forms
  - Turn data into useful information and knowledge for broad applications
    - Market analysis, business management, decision support
- Data mining the process of discovering interesting knowledge from huge amounts of data
  - Patterns, associations, changes, anomalies and significant structures
  - Databases, data warehouses, information repositories

Discovery knowledge process



Simplified KDD discovery process

- Database/ data warehouse construction
  - Cleaning, integration, selection and transformation

- Iterative process: data mining
  - Data mining, pattern evaluation, knowledge representation

Major tasks in data mining

- Description
  - Describes the data set in a concise and summarized manner
  - Presents interesting general properties of data
- Prediction
  - Constructs one or a set of models
  - Performs inference on the data set
  - Attempts to predict the behavior of new data sets

Data mining system tasks

- Class description
  - Concise and succinct summarization of a collection of data → characterization
  - Distinguishes it from others → comparison or discrimination
  - Aggregation: count, sum, avg
  - Data dispersion: variance, quartiles, etc.
  - Example: compare European and Asian sales of a company, identify important factors which discriminate the two classes
- Association
- Classification
- Prediction
- Clustering
- Time series analysis

Data mining system tasks

- Class description
- Association
  - Discovery of correlations or association relationships among a set of items
  - Expressed in the form a rule: attribute value conditions that occur frequently together in a given set of data
  - X→Y: database tuples that satisfy X are likely to satisfy Y
  - Transaction data analysis for directed marketing, catalog design, etc.
- Classification
- Prediction
- Clustering
- Time series analysis

Data mining system tasks

- Class description
- Association
- Classification
  - Analyze a set of training data (i.e., a set of objects whose class label is known)
  - Construct a model for each class based on the data features
  - A decision tree or classification rules are generated
    - Better understanding of each class
    - Classification of future data
    - Diseases classification to help to predict the kind of diseases based on the symptoms of patients
  - Classification methods proposed in machine learning, statistics, database, neural networks, rough sets.
  - Customer segmentation, business modelling and credit analysis
- Prediction
- Clustering
- Time series analysis

Data mining system tasks

- Class description
- Association
- Classification
- Prediction
  - Predict possible values of some missing data or the value distribution of certain attributes in a set of objects
    - Find a set of relevant attributes to the attribute of interest (e.g., by some statistical analysis)
    - Predict the value distribution based on the set of data similar to the selected objects
    - An employees potential salary can be predicted based on the salary distribution of similar employees in a company
  - Regression analysis, generalized linear models, correlation analysis, decision trees used in quality prediction
  - Genetic algorithms and neural network models also popular
- Clustering
- Time series analysis

Data mining system tasks

- Class description
- Association
- Classification
- Prediction
- Clustering
  - Identify clusters embedded in the data
    - Cluster is a collection of data objects similar to one another
    - Similarity expressed by distance functions specified by experts
  - Good cluster method produces high quality clusters to ensure that
    - inter cluster similarity is low
    - intra cluster similarity is high
  - Cluster the houses of Cholula according to their house category, floor area and geographical locations
- Time series analysis

Data mining system tasks

- Class description
- Association
- Classification
- Prediction
- Clustering
- Time series analysis
  - Analyze large set of time-series data to find regularities and interesting characteristics
    - Search for similar sequences, sequential patterns, periodicities, trends and derivations
  - Predict the trend of the stock values for a company based on its stock history, business situation, competitors' performance and current market

Data mining challenges

- Handling of different types of data
  - Knowledge discovery system should perform efficient and effective data mining on different kinds of data
  - Relational data, complex data types (e.g. structured data, complex data objects, hypertext, multimedia, spatial and temporal, transaction, legacy data)
  - Unrealistic for one single system
- Efficiency and scalability of data mining algorithms
- Usefulness, certainty and expressiveness of data mining results
- Expression of various kinds of data mining results
- Interactive mining knowledge at multiples abstraction levels
- Mining information from different sources of data
- Prediction of privacy and data security

Data mining challenges

- Handling of different types of data
- Efficiency and scalability of data mining algorithms
  - Running times predictable and acceptable in large databases
  - Algorithms with exponential or medium order polynomial complexity are not practical
- Usefulness, certainty and expressiveness of data mining results
- Expression of various kinds of data mining results
- Interactive mining knowledge at multiples abstraction levels
- Mining information from different sources of data
- Prediction of privacy and data security

Data mining challenges

- Handling of different types of data
- Efficiency and scalability of data mining algorithms

- Usefulness, certainty and expressiveness of data mining results
  - Discovered knowledge must
    - Portray the contents of a database accurately:
    - Useful for certain applications
  - Uncertainty measures (approximate or quantitative rules)
  - Noise and exceptional data: statistical, analytical and simulative models and tools

- Expression of various kinds of data mining results
- Interactive mining knowledge at multiples abstraction levels
- Mining information from different sources of data
- Prediction of privacy and data security

Data mining challenges

- Handling of different types of data
- Efficiency and scalability of data mining algorithms
- Usefulness, certainty and expressiveness of data mining results

- Expression of various kinds of data mining results

  - Different kinds of knowledge can be discovered
  - Examine from different views and present in different forms
    - Express data mining requests and discovered knowledge in high level languages or graphical interfaces
    - Knowledge representation techniques

- Interactive mining knowledge at multiples abstraction levels
- Mining information from different sources of data
- Prediction of privacy and data security

Data mining challenges

- Handling of different types of data
- Efficiency and scalability of data mining algorithms
- Usefulness, certainty and expressiveness of data mining results
- Expression of various kinds of data mining results
- Interactive mining knowledge at multiples abstraction levels
  - Difficult to predict what can be discovered
  - High level data mining query should be treated as a probe disclosing interesting traces to be further explored
    - Interactive discovery: refine queries, dynamically change data focusing, progressively deepen a data mining process, flexibly view data and data mining results at multiple abstraction levels and different angles
- Mining information from different sources of data
- Prediction of privacy and data security

Data mining challenges

- Handling of different types of data
- Efficiency and scalability of data mining algorithms
- Usefulness, certainty and expressiveness of data mining results
- Expression of various kinds of data mining results
- Interactive mining knowledge at multiples abstraction levels
- Mining information from different sources of data
  - Mine distributed and heterogeneous (structure, format, semantic)
  - Disclose high level data regularities in heterogeneous databases hardly discovered by query systems
  - Huge size, wide distribution and computational complexity of data mining methods → parallel and distributed algorithms
- Prediction of privacy and data security

Data mining challenges

- Handling of different types of data
- Efficiency and scalability of data mining algorithms
- Usefulness, certainty and expressiveness of data mining results
- Expression of various kinds of data mining results
- Interactive mining knowledge at multiples abstraction levels
- Mining information from different sources of data
- Prediction of privacy and data security
  - Data viewed from different angles and abstraction levels → threaten security and privacy
  - When is it invasive and how to solve it?
    - Conflicting goals
    - Data security protection vs. Interactive data mining of multiple level knowledge from different angles

Data mining approaches

- Needs the integration of approaches from multiple disciplines
  - Database systems & data warehousing
  - Statistics, machine learning, data visualization, information retrieval, high performance computing
  - Neural networks, pattern recognition, spatial data analysis, image databases, spatial processing, probabilistic graph theory and inductive logic programming
- Large set of data mining methods
  - Machine learning: classification and induction problems
  - Neural networks: classification, prediction, clustering analysis tasks

  → Scalability and efficiency
- Data structures, indexing, data accessing techniques

Data analysis vs. data mining

- Data analysis
  - Assumption driven
    - Hypothesis is formed and validated against data
- Data mining
  - Discovery-driven
    - Patterns are automatically extracted from data
    - Substantial search efforts
  - High performance computing
    - Parallel, distributed and incremental data mining methods
    - Parallel computer architectures

Classifying data mining techniques

- What kinds of databases to work on
  - DMS classified according to the kinds of database on which data mining is performed
  - Relational, transactional, OO, deductive, spatial, temporal, multimedia, heterogeneous, active, legacy, Internet-information base
- What kind of knowledge to be mined
  - Kind of knowledge
    - Association, characteristic, classification, discriminant rules, clustering evolution, deviation analysis
  - Abstraction level of the discovered knowledge
    - Generalized knowledge, primitive-level knowledge, multiple-level knowledge
- What kind of techniques to be utilized
  - Driven methods
    - Autonomous, data-driven, query-driven, interactive
  - Data mining approach
    - Generalization-based, pattern-based, statistics and mathematical theories, integrated approaches

Data mining algorithms

- Description
  - Discover knowledge contained in a data collection → decision making
  - Algorithms
    - Clustering
    - Association rules
  - Business and scientific areas

- Prediction
  - Forecast the value of a variable based on the previous knowledge of that variable
  - Algorithms
    - Classification
    - Prediction
    - Trend detection
  - Disasters prediction like floods, earthquake, volcanoes eruptions

Mining Different Kinds of Knowledge in Large Databases

- **Characterization**: Generalize, summarize, and possibly contrast data characteristics, e.g., grads/undergrads in CS.
- **Association**: Rules like "buys(x, milk) à buys(x, bread)".
- **Classification**: Classify data based on the values in a classifying attribute, e.g., classify cars based on gas mileage.
- **Clustering**: data to form new classes, e.g., cluster houses to find distribution patterns.
- **Trend and deviation analysis**: Find and characterize evolution trend, sequential patterns, similar sequences, and deviation data, e.g., stock analysis.
- **Pattern-directed analysis**: Find and characterize user-specified patterns in large databases.

Conclusions

- **Data mining**: A rich, promising, young field with broad applications and many challenging research issues.
- **Recent progress**: Database-oriented, efficient data mining methods in relational and transaction DBs.
- **Tasks**: Characterization, association, classification, clustering, sequence and pattern analysis, prediction, and many other tasks.
- **Domains**: Data mining in extended-relational, transaction, object-oriented, spatial, temporal, document, multimedia, heterogeneous, and legacy databases, and WWW.
- **Technology integration**:
  - Database, data mining, & data warehousing technologies.
  - Other fields: machine learning, statistics, neural network, information theory, knowledge representation, etc.

Knowledge discovery phases

- Data cleaning
  - handle noisy, erroneous, missing or irrelevant data (e.g., AJAX)
- Data integration
- Data selection
- Data transformation
- Data mining
- Pattern evaluation
- Knowledge presentation

Knowledge discovery phases

- Data cleaning
- Data integration
  - multiple, heterogeneous data sources may be integrated into one
- Data selection
- Data transformation
- Data mining
- Pattern evaluation
- Knowledge presentation

Knowledge discovery phases

- Data cleaning
- Data integration
- Data selection
  - relevant data for the analysis task retrieved from the database
- Data transformation
- Data mining
- Pattern evaluation
- Knowledge presentation

Knowledge discovery phases

- Data cleaning
- Data integration
- Data selection
- Data transformation
  - data transformed or consolidated into forms appropriate for mining (i.e., aggregation)
- Data mining
- Pattern evaluation
- Knowledge presentation

Knowledge discovery phases

- Data cleaning
- Data integration
- Data selection
- Data transformation
- Data mining:
  - intelligent methods are applied in order to extract data patterns
- Pattern evaluation
- Knowledge presentation

Knowledge discovery phases

- Data cleaning
- Data integration
- Data selection
- Data transformation
- Data mining
- Pattern evaluation:
  - identify the truly interesting patterns representing knowledge ← interestingness measures
- Knowledge presentation

Knowledge discovery phases

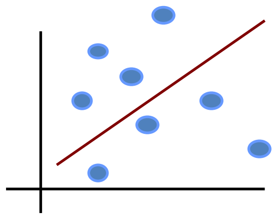
- Data cleaning
- Data integration
- Data selection
- Data transformation
- Data mining
- Pattern evaluation
- Knowledge presentation
  - visualization and knowledge representation techniques
  - present knowledge to the “decision makerPattern evaluation”

Prediction

- Compute a future value for a variable

A1	A2	...	An
4°	800 lts	...	??

A1	A2	...	An
23°	30 lts	...	Normal
34°	0 lts	...	Normal
5°	500 lts	...	Flood
...	...	...	...
9°	600 lts	...	Flood



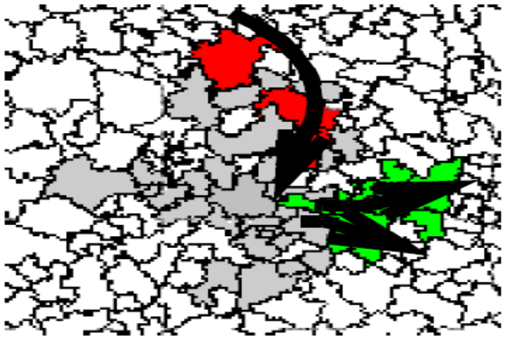
Linear Regression

A1	A2	...	An
4°	800 lts	...	flood

?

Trend detection

- Discover information, given an object and its neighbors



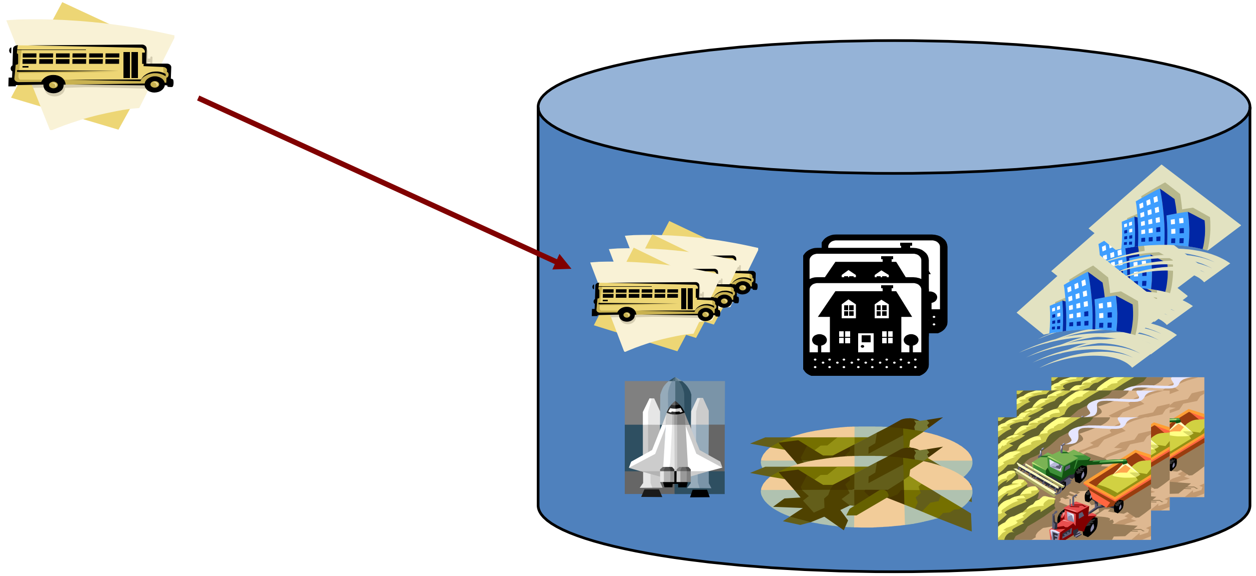
?

Classification

- General principle and definitions
- Classification based on decision trees
- Methods for performance improvement

General principle

- Given a set of classes identify whether a new object belongs to one of them



Definitions

- Process which
  - Finds the common properties among a set of objects in a database
  - Classifies them into different classes according to a classification model

- Classification model
  - Sample database E treated as a training set where each tuple
    - Consists of the same set of multiple attributes (or features)
    - Has a known class identity associated with it

- Objective
  - First analyze the training data and develop an accurate description of model for each class using the features
  - Class descriptions used to classify future test data or develop a better description (classification rules)

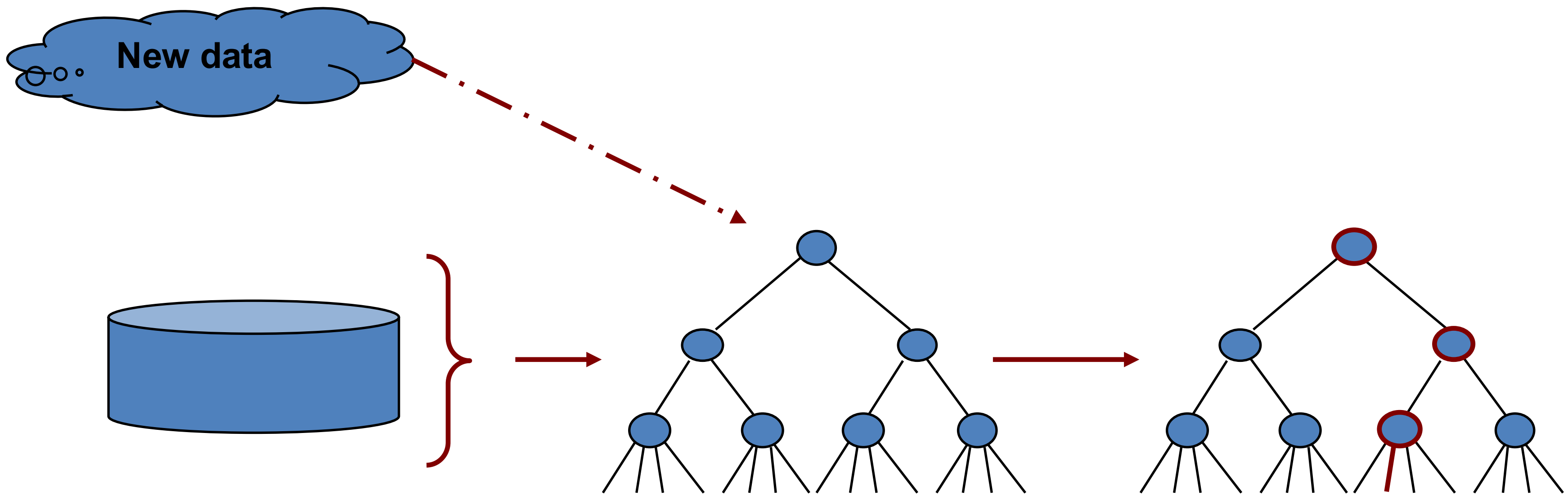
- U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996
- S.M. Weiss, C.A. Kulikowski, *Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning and Expert Systems*, Morgan Kaufman, 1991

Classification

- ✓ General principle and definitions
- Classification based on decision trees
- Methods for performance improvement

Decision trees

- Organized data with respect to variable class
- Algorithms: ID3, C4.5, C5, CART, SLIQ, SPRINT

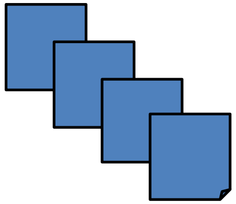


Decision trees

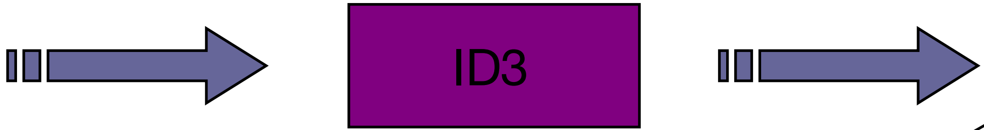
- A decision tree based classification method is a supervised learning method
  - Constructs a decision tree from a set of examples
  - The quality function of a tree depends on the classification accuracy and the size of the tree
- Choose a subset of training examples (a window) to form the decision tree
  - If the tree does not give the correct answer for all the objects
    - a selection of exceptions is added to the window
    - The process continues until de right decision tree is found
  - A tree in which
    - Each leaf carries a class name
    - Interior node specifies and attribute with a branch corresponding to each possible value of that attribute

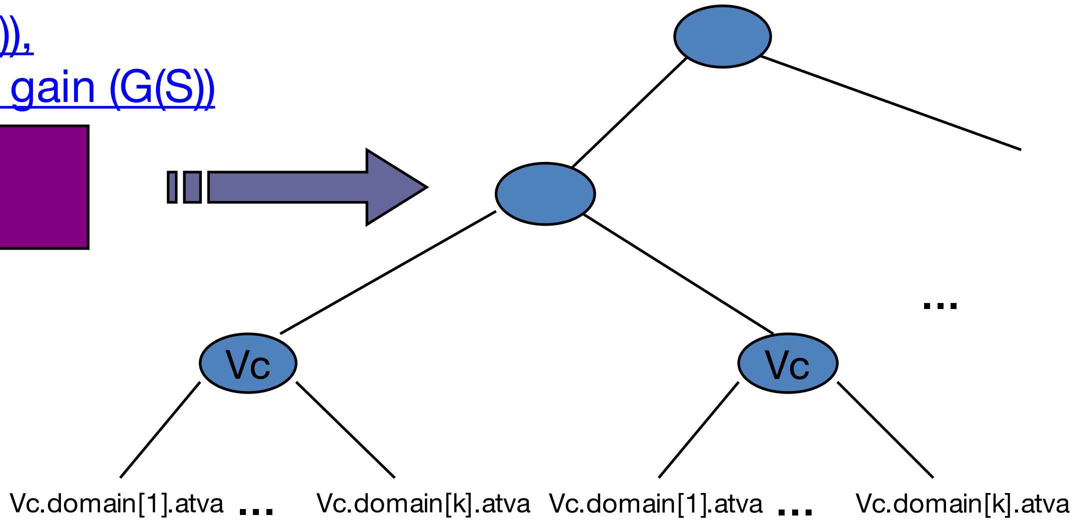
Prediction algorithm: Interactive Dichotomizer (ID3)

A_1	...	A_i	Vc



- Top down
- Greedy
- [Entropy (I(S)),]()
- [Information gain (G(S))]()





Data collection:

- **Set of attributes**
  - $A_i \rightarrow$ {<value$_1$, occurrence number>,
    …,
    <value$_j$, occurrence number>}
- **Class variable denotes values characterizing the represented model**
  Vc $\rightarrow$ {<value$_1$, occurrence number>,

```
dat{
  Tuple[] domain;
}
```

```
Tuple{
  String atVa;
  Number nOc;
```

Information gain of an attribute

$$G(A_i) = I - I(A_i)$$

- $G(A_i)$ = Information gain for attribute A_i
- I= Entropy of the class variable
- $I(A_i)$ = Entropy of attribute A_i

Attribute entropy

$$I(A_i) = \sum_{j=1}^{nv(A_i)} \frac{n_{ij}}{n} I_{ij}$$

- nv(Ai) = The different values number that the attribute Ai can take.
- nij/n = The probability that the attribute Ai appears in the collection
- n = The number of the rows in the data collection
- Iij = Entropy of the attribute Ai with value j

Entropy of the values of an attribute Ai

- Given the value j of attribute Ai :

$$I_{ij} = -\sum_{k=1}^{nc} \frac{n_{ijk}}{n_{ij}} \log_2 \frac{n_{ijk}}{n_{ij}}$$

- nc = class variable domain cardinality
- n_{ijk}/n_{ij}
  - Given a value j of attribute Ai and a value k of the class variable
  - Probability of the occurrence of tuples in the collection containing j and k
- $\log_2 n_{ijk}/n_{ij}$ = number of digits need for representing the probability n_{ijk}/n_{ij} in binary system

Gain table

- For each attribute of the data collection
  - Compute information gain
- Order the table

Attribute	Information gain
A_1	0.6
A_2	0.5
. . .	
A_i	0.1

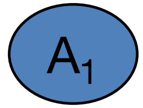
Decision tree construction: first step

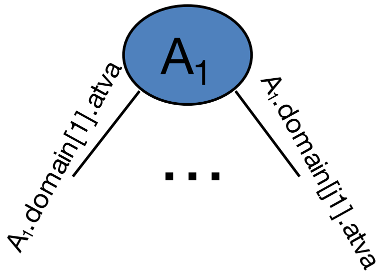
- Identify the class variable
- Compute variable base
- Compute gain table
- Root: the attribute with the highest gain
- Edges
  - Number: Vc domain cardinality
  - Label: vc in Vc.domain

A_1	...	A_i	Vc

Attribute	Information gain
A_1	0.6
A_2	0.5
. . .	
A_i	0.1

$A_1 \rightarrow$ {<value$_1$, occurrence number>,
…
<value$_{j1}$, occurrence number>}





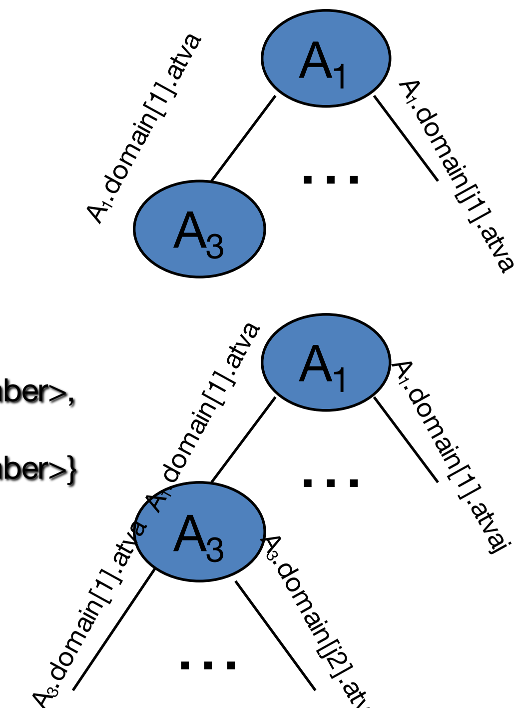
Decision tree construction: step2..n

- Class variable: root
- Select one of the root's edges
- Compute the gain table
- Node$_i$:
  - Attribute with the highest gain in the new table
- Edges
  - Number: Vc domain cardinality
  - Lable: vc in Vc.domain

A_1	...	A_i	Vc

Attribute	Information gain
A_2	0.5
A3	0.7
. . .	
A_i	0.3

$A_3 \rightarrow$ {<value$_1$, occurrence number>,
…
<value$_{j2}$, occurrence number>}



→Recursively compute nodes n_{i+1} until each root's edges have

- The original class variable is the last node and

Classification

- ✓ General principle and definitions
- ✓ Classification based on decision trees
- Methods for performance improvement

Performance improvement

- Scaling up problems
  - Realatively well performance in small databases
  - Poor performance or accuracy reduction with large training sets
- Databases indices to improve on data retrieval but not in classification efficiency
  - R. Agrawal, S. Ghosh, T. Imielinsky, B. Iyer, A. Swami, An interval classifier for database mining applications, Proceedings of the 18th International Conference on Very Large Databases, August, 1992
- DBMiner improve classification accuracy: multi-level classification technique
  - Classification accuracy in large databases with attribute oriented induction and classification methods
  - J. Han, Y. Fu, Exploration of the power of attribute-oriented induction in data mining, In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, eds., Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996
  - J. Han, Y. Fu, W. Wang, J. Chiang, W. Gong, K. Koperski, D. Li, Y. Lu, A. Rajan, N. Stefanovic, B. Xia, O.R. Zaiane, DBMiner: A system for mining knowledge in large relational databases, In Proceedings of the International Conference on Datamining and knowledge discovery, August, 1996
- SLIQ (Supervised Learning in QUEST)
  - Mining classification rules in large databases
  - Decistion tree classifier for numerical and categorical attributes
  - Pre-sorting technique, tree pruning
  - P.K. Chan, S.J. Stolfo, Learning arbiter and combiner trees from partitioned data for scaling machine learning, Proceedings of the 1st International Conference On Knowledge discovery and Data mining, August, 1995

?

Clustering

- General principle and definitions
- Randomized search for clustering large applications
- Focusing methods
- Clustering feature and CF trees

Clustering

- Discover a set of classes given a data collection

Definitions

- Process of grouping physical or abstract objects into classes of similar objects
  → Clustering or unsupervised classification
- Helps to construct meaningful partitioning of a large set of objects
  - Divide and conquer methodology
  - Decompose a large scale system into smaller components to simplify design and implementation
- Identifies clusters or densely populated regions
  - According to some distance measurement
  - In a large multidimensional data set
  - Given a set of multidimensional data points
    - The data space is usually not uniformly occupied
    - Data clustering identifies the sparse and the crowded places
    - Discovers the overall distribution patterns of the data set

Approaches

- As a branch of statistics, clustering analysis extensively studied focused on distance-based clustering analysis
  - AutoClass with Bayesian networks
    - P. Cheeseman, J. Stutz, Bayesian classification (AutoClass): Theory and results, In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, eds. Advances in Knowledge Discovery and Data mining, AAAI/MIT Press, 1996
  - Assume that all data point are given in advance and can be scanned frequently
- In machine learning clustering analysis
- Clustering analysis

Approaches

- As a branch of statistics
- In machine learning clustering analysis
  - Refers to unsupervised learning
    - Classes to which an object belongs to are not pre-specified
  - Conceptual clustering
    - Distance measurement may not be based on geometric distance but on that a group of objects represents a certain conceptual class
    - One needs to define a similarity between the objects and then apply it to determine the classes
    - Classes are collections of objects low interclass similarity and high intra class similarity
- Clustering analysis

Approaches

- As a branch of statistics
- In machine learning clustering analysis
- Clustering analysis
  - Probability analysis
    - Assumption that probability distributions on separate attributes are statistically independent one another (not always true)
    - The probability distribution representation of clusters → expensive clusters' updates and storage
  - Probability-based tree built to identify clusters is not height balanced
    - Increase of time and space complexity

- D. Fisher, Improving inference through conceptual clustering, *In Proceedings of the AAAI Conference*, July, 1987
- D. Fisher, Optimization and simplification of hierarchical clusterings, *In Proceedings of the 1st Conference on Knowledge Discovery and Data mining*, August, 1985

Clustering

- ✓ General principle and definitions
- ■ Randomized search for clustering large applications
- ■ Focusing methods
- ■ Clustering feature and CF trees

Clustering Large applications based upon randomized Search[62]

- PAM (Partitioning Around Medoids)
  - Finds k clusters in n objects
    - First finding a representation object for each cluster
      - The most centrally located point in a cluster: medoid
    - After selecting k medoids,
      - Tries to make a better choice
      - Analyzing all possible pairs of objects such that one object is a medoid and the other is not
      - The measure of clustering quality is calculated for each such combination
    - Cost of a single iteration $O(k(n-k)^2)$ → inefficient if k is big
- CLARA (Clustering Large Applications)

Clustering Large applications based upon randomized Search

- PAM (Partitioning Around Medoids)
- CLARA (Clustering Large Applications)
  - Uses sampling techniques
  - A small portion of the real data is chosen as a representative of the data
  - Medoids are chosen from this sample using PAM
    - If the sample is selected in a fairly random manner
    - Correctly represents the whole data set
    - The representative objects (medoids) will be similar to those chosen for the whole data set
- CLARANS integrate PAM and CLARA
    - Searching only the subset of the data set not confining it to any sample at any given time
    - Draw a sample randomly in each step
    - Clustering process as searching a graph where every node is a potential solution

Clustering

- ✓ General principle and definitions
- ✓ Randomized search for clustering large applications
- ■ Focusing methods
- ■ Clustering feature and CF trees

Focusing methods*

- CLARANS assumes that the objects are all stored in main memory
  - Not valid for large databases →
    - Disk based methods required
    - R*-trees[11] tackle the most expensive step (i.e., calculating the distances between two clusters)
- Reduce the number of considered objects: *focusing on representative objects*
  - A centroid query retunrs the most cetrnal object of a leaf node of the R*-tree where neighboring points are stored
  - Only these objects used to compute medoids of the clusters
  - ☺ The number of objects is reduced
  - ☹ Objects that could have been better medoids are not considered
- Restrict the access to certain objects that do not actually contribute to the computation: computation performed only on pairs of objects that can improve the clustering quality
  - Focus on relevant clusters
  - Focus on a cluster

*M. Ester, H.P. Kriegel and X. Xu, Knowledge discovery in large spatial databases: Focusing techniques for efficient class identification, *In Proceedings of the 4th Symposium on Large Spatial Databases,* August, 1995

Clustering

- ✓ General principle and definitions
- ✓ Randomized search for clustering large applications
- ✓ Focusing methods
- ■ Clustering feature and CF trees

Clustering feature and CF trees

- R-trees not always available and time consuming construction

- BIRCH (Balancing Iterative Reducing and Clustering)
  - Clustering large sets of points
  - Incremental method
  - Adjustment of memory requirements according to available size

Clustering feature and CF trees: concepts

- Clustering feature
  - CF is the triplet summarizing information about subclusters of points. Given n-dimensional points in a subcluster {Xi}

$$CF = \left(N, \overrightarrow{LS}, SS\right)$$

  - N is the number of points in the subcluster
  - LS is the linear sum on N points $\sum_{i=1}^{N} \overrightarrow{X_i}$
  - SS is the squares sum of data points $\sum_{i=1}^{N} \overrightarrow{X_i}^2$
- Clustering features
  - Are sufficient for computing clusters
  - Summarize information about subclusters of points instead of storing all points
    - Constitute an efficient storage information method since they

Clustering feature and CF trees: concepts

- Clustering feature tree
  - Branching factor B specifies the maximum number of children
  - Threshold T specifies the maximum diameter of subclusters stored at the leaf nodes
    - Changing the T we can change the size of the tree
  - Non leaf nodes are storing sums of their children CF's → summarize information about their children
  - Incremental method: built dynamically as data points are inserted
    - A point is inserted in the closes leaf entry
      - If the diameter of the cluster stored in the leaf node after insertion is larger than T
        - Split it and eventually other nodes
    - After insertion the information about the new point is transmitted to the root

Clustering feature and CF trees: concepts

- Clustering feature tree
  - The size of CF tree can be changed by changing T
  - If the size of the memory needed for storing the CF tree is larger than the size of the main memory
    - Then a larger T is specified and the tree is rebuilt
  - Rebuild process is done by building a new tree from the leaf nodes of the old tree
    - Reading all the points is not necessary

- CPU and I/O costs of BIRCH $O(N)$
  - Linear scalability of the algorithm with respect to the number of points
  - Insensibility of the input order
  - Good quality of clustering of the data

- T. Zhang, R. Ramakrishman, M. Livy, BIRCH: an efficient data clustering method for very large databases, In Proceedings of the ACM SIGMOD International Conference on Management of Data, June 1996

?

Association rules

- General principle
- A priori algorithm: an example
- Mining generalized rules
- Improving efficiency of mining association rules

Association rules

- Given a data collection determine possible relationships among the variables describing such data

- Relationships are expressed as association rules X →Y

Association rules

- ✓ General principle
- A priori algorithm: an example
- Mining generalized rules
- Other issues on mining association rules
  - Interestingness of discovered association rules
  - Improving efficiency of mining association rules

Mathematical model

- Let $I = \{i_1, \ldots, i_n\}$ be a set of literals called items
- D a set of transaction where each t in T is a set of items such that $T \subseteq I$
- Each transaction has in TID
- Let X be a set of items, T is said to contain X iff X in T
- An association rule X→Y where X in I, Y in I and X does not intersect Y
  - Holds in the transaction set D with confidence c if c% of the transactions in D that contain X also contain Y
  - Has support s in the transaction set D if s% of transactions in D contain the intersection of X and Y

Mathematical model

- Confidence denotes the strength of implication
- Support indicates the frequencies of the occurring patterns in the rule
  - Reasonable to pay attention to rules with reasonably large support: strong rules
  - Discover strong rules in large data bases
    - Discover large item sets
      - the sets of itemsets that have transaction support above a predetermined minimum support s
    - Use large itemsets to generate association rules for the database

Algorithm a priori*

Database D

TID	Items
100	A C D
200	B C E
300	A B C E
400	B E

Scan D →

C_1

Itemset	s
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	s
{A}	2
{B}	3
{C}	3
{E}	3

- In each iteration
  - Construct a candidate set of large itemsets
  - Count the number of occurrences in of each candidate itemset
  - Determine large itemsets based on a pre-determined minimum support
- In the first iteration
  - Scan all transactions to count the number of occurrences for each item

*R. Agrawal, R. Srikant, Mining Sequential Patterns, Proceedings of the 11th International Conference on Data Engineering, March, 1995

Algorithm a priori*

C_2

Itemset
{A,B}
{A,C}
{A,E}
{B,C}
{B,E}
{C,E}

Scan D →

Itemset	s
{A,B}	1
{A,C}	2
{A,E}	1
{B,C}	2
{B,E}	3
{C,E}	2

L_1

Itemset	s
{A,C}	2
{B,C}	2
{B,E}	3
{C,E}	2

- Second iteration
  - Discover 2-itemsets
  - Candidate set C_2: $L_1 * L_1$ $$L_k * L_k = \{X \cup Y \mid X, Y \in L_K, |X \cap Y| = k-1\}$$
  - C_2 consists of $\binom{|L_1|}{2}$ 2-itemsets

*R. Agrawal, R. Srikant, Mining Sequential Patterns, Proceedings of the 11th International Conference on Data Engineering, March, 1995

Algorithm a priori*

C_3

Itemset
{B,C,E}

Scan D →

Itemset	s
{B,C,E}	2

L_3

Itemset	s
{B,C, E}	2

- From L_2
  - two large 2-itemsets are identified with the same first item: {B,C} and {B,E}
  - {C,E} is a two large 2-itemset? YES!
- No candidate 4-itemset → END
- HOMEWORK: Analyze DHP in

  J.-S. Park, P.S. Yu, An effective hash based algorithm for mining association rules, Proceedings of the ACM SIGMOD, May, 1995

*R. Agrawal, R. Srikant, Mining Sequential Patterns, Proceedings of the 11th International Conference on Data Engineering, March, 1995

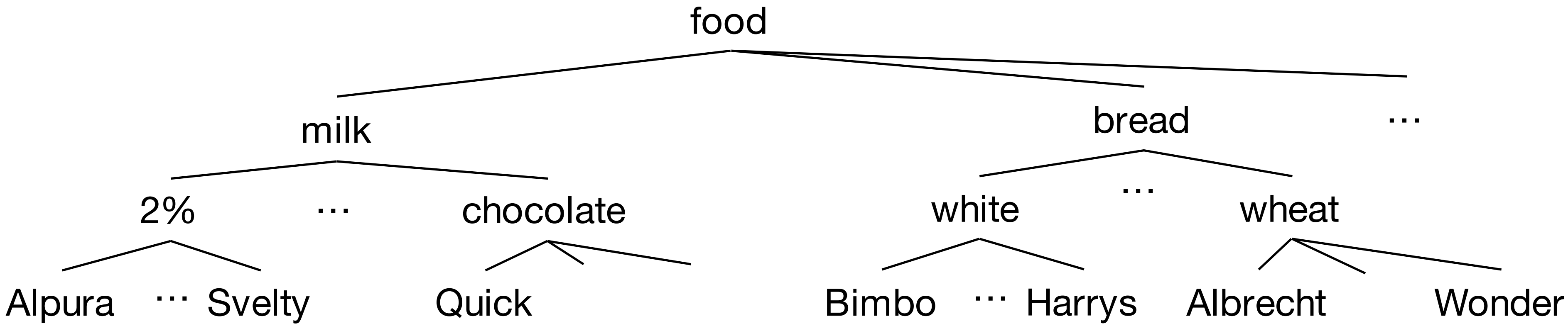
Association rules

- ✓ General principle
- ✓ A priori algorithm: an example
- Mining generalized rules
- Other issues on mining association rules
  - Interestingness of discovered association rules
  - Improving efficiency of mining association rules

Mining generalized and multiple-level association rules

- Interesting associations among data items occur at a relatively high concept level
  - Purchase patterns in a transaction database many not show substantial regularities at a primitive data level (e.g., bar code level)
  - Interesting regularities at some high concept level such as milk and bread
- Study association rules at a generalized abstraction level or at multiple levels

Mining generalized and multiple-level association rules*



- The bar codes of 1 gallon of Alpura 2% milk and 1lb of Wonder wheat bread: what for?
- 80% of the customers that purchase milk also purchase bread
- 70% of people buy wheat bread if they buy 2% milk

* J. Han, Y. Fu, Discovery of Multiple-Level Association Rules from Large Databases, *Proceedings of the 21th International Conference of Very Large Databases*, September 1995

Mining generalized and multiple-level association rules*

- Low level associations may be examined only when
  - High level parents are large at their corresponding levels
  - Different levels may adopt different minimum support thresholds

- Four algorithms developed for efficient mining of association rules
  - Based on different ways of sharing multiple level mining processes and reduction of encoded transaction tables

- Mining of quantitative association rules
  - R. Srikant, R. Agrawal, Mining Generalized Association Rules, *Proceedings of the 21st International Conference on Very Large Databases*, September, 1995

- Meta rule guided mining of association rules in relational databases
  - Y. Fu, J. Han, Meta rule guided mining of association rules in relational databases, *Proceedings of the 1st International Workshop on Integration of Knowledge with Deductive and Object Oriented Databases (KDOOD),* Singapore, December, 1995
  - W. Shen, K. Ong, B. Mitbander, C. Zaniolo, Metaqueries for data mining, In U.M. Fayard, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, eds., *Advances in Knowledge Discovery and Data mining*, AAAI/MIT Press, 1996

* J. Han, Y. Fu, Discovery of Multiple-Level Association Rules from Large Databases, *Proceedings of the 21th International Conference of Very Large Databases*, September 1995

Association rules

- ✓ General principle
- ✓ A priori algorithm: an example
- ✓ Mining generalized rules
- Other issues on mining association rules
  - Interestingness of discovered association rules
  - Improving efficiency of mining association rules

Interestingness of discovered association rules

- Not all discovered association rules are strong (i.e., passing the minimum support and minimum confidence thresholds)
- Consider a survey done in a university of 5000 students: students and activities they engage in the morning
  - 60% of the students play basket ball, 75% eat cereal, 40% both play basket ball and eat cereal
  - Suppose that a mining program runs
    - Minimal student support s = 2000
    - Minimal confidence is 60%
    - Play basket ball → eat cereal
      - 2000/3000=0,66
      - Pb!!! The overall percentage of students eating cereal is 75% > 66%
      - Playing basket ball and eating cereal are negatively associated: being involved in one decreases the likelihood of being involved in the other

Interestingness of discovered association rules

- Filter out misleading associations
  - A → B is interesting if its confidence exceeds a certain measure
  - Test of statistical independence

$$\frac{P(A \cap B)}{P(A)} - P(B) > d \qquad P(A \cap B) - P(A) * P(B) > k$$

- Interestingness studies

- G. Piatetsky-Shapiro, Discovery analysis and presentation of strong rules, In G. Piatetsky-Shapiro and W.J. Frawley, eds. *Knowledge Discovery in Databases*, AAAI/MIT press, 1991
- A. Silberschatz, M. Stonebraker, J.D. Ullman, Database research: Achievements and opportunities into the 21st century, *In Report of an NSF Workshop on the Future of Database Systems Research*, May, 1995
- R. Srikant, R. Agrawal, Mining generalized association rules, *Proceedings of the 21st Internation Conference on Very Large Databases*, September, 1995

Association rules

- ✓ General principle
- ✓ A priori algorithm: an example
- ✓ Mining generalized rules
- Other issues on mining association rules
  - ✓ Interestingness of discovered association rules
  - Improving efficiency of mining association rules

Improving the efficiency of mining association rules

- Database scan reduction:
  - Profit from database scans C_i in order to compute in advance L_i and L_{i+1}
    - M.S. Chen, J.S. Park, P.S. Yu, Data mining for path traversal patterns in a Web Environment, *Proceedings of the 16ty International Conference on Distributed Computing Systems*, May, 1996
- Sampling: mining with the adjustable accuracy
- Incremental updating of discovered association rules
- Parallel data mining

Improving the efficiency of mining association rules

- Database scan reduction:
- Sampling: mining with the adjustable accuracy
  - Frequent basis for mining transaction data to capture behavior
  - Efficiency more important than accuracy
  - Attractive due to the increasing size of databases

  - H. Mannila, H. Toivonen, A. Inkeri Verkamo, Efficient algorithms for discovering association rules, Proceedings of the AAAI Workshop on Knowledge Discovery in Databases, July, 1994
  - J.-S. Park, M.S. Chen, P.S. Yu, Mining association rules with adjustable accuracy, IBM research report, 1995
  - R. Srikant, R. Agrawal, 1995, *ibidem*.
- Incremental updating of discovered association rules
- Parallel data mining

Improving the efficiency of mining association rules

- Database scan reduction:

- Sampling: mining with the adjustable accuracy
  - Frequent basis for mining transaction data to capture behavior
  - Efficiency more important than accuracy
  - Attractive due to the increasing size of databases

  - H. Mannila, H. Toivonen, A. Inkeri Verkamo, Efficient algorithms for discovering association rules, Proceedings of the AAAI Workshop on Knowledge Discovery in Databases, July, 1994
  - J.-S. Park, M.S. Chen, P.S. Yu, Mining association rules with adjustable accuracy, IBM research report, 1995
  - R. Srikant, R. Agrawal, 1995, *ibidem*.

- Incremental updating of discovered association rules

- Parallel data mining

Improving the efficiency of mining association rules

- Database scan reduction:
- Sampling: mining with the adjustable accuracy
- Incremental updating of discovered association rules
  - On data base updates →
    - Maintenance of discovered association rules required
    - Avoid redoing data mining on the whole updated database
    - Rules can be invalidated and weak rules become strong
  - Reuse information of the large itemsets and integrate the support information of new ones
    - Reduce the pool of candidate sets to be examined
  - D.W. Cheung, J. Han, V. Ng, C.Y. Wong, Maintenance of discovered association rules in large databases: an incremental updating technique, *In Proceedings of the International Conference on Data Engineering*, February, 1996
- Parallel data mining

Improving the efficiency of mining association rules

- Database scan reduction:
- Sampling: mining with the adjustable accuracy
- Incremental updating of discovered association rules
- Parallel data mining
  - Progressive knowledge collection and revision based on huge transaction databases
  - DB partitioned → inter-node data transmission for making decisions can be prohibitively large

- IBM *Scalable POWERparallel Systems*, Technical report GA23-2475-02, February, 1995
- J.S. Park, M.S. Chen, P.S. Yu, Efficient parallel data mining for association rules, *Proceedings of the 4th International Conference on Information and Knowledge Management*, November, 1995

?