




Data Mining for Studying Large Databases

Mitsunori Ogihara
Department of Computer Science
University of Rochester
ogihara@cs.rochester.edu

April 8, 2004

Mitsunori Ogihara, CSC200 Talk

1



Data Mining (and Knowledge Discovery)

- Use computation to learn from large databases of routinely collected data

April 8, 2004

Mitsunori Ogihara, CSC200 Talk

2



Characteristics of Data Mining

- Practical implications, multidisciplinary nature
 - The extracted information applied to business, medicine, science, and engineering.
- Large databases of routinely collected data
 - Data are routinely collected, otherwise trashed.
- Intensive computation
 - Compute-intensive, ad hoc, approaches are used.




Interdisciplinary Nature of Data Mining

- Extracting information → Information retrieval
- Computational modeling of data → Machine learning
- Database interfaces → Database Theory
- Hardware issues → Systems, Networks
- Statistical underpinnings of results → Statistics
- Application dependent analysis of results → Medicine, Economics, Management, etc.
- **Heuristic algorithm design → Algorithms**



Data Mining Tasks Are Selected Depending on Attribute Types

- Numeric (the set of real numbers)
- Discrete (the set of integers)
- Categorical (finite sets)
- Binary
- Plain Text
- Semi-Structured Text Data (HTML)
- Image
- Sequential Data



Data Mining Tasks Are Selected Depending on Data Layout

- Temporal data: Attributes include the time stamp. The temporal changes are studied in the attributes of the data entries having the same ID



Major Data Mining Tasks

- ***Association Mining & Sequence Mining***
 - Finding significant patterns in the database
- ***Classification***
 - Learning to assign labels to data points
- **Clustering**
 - Grouping similar data points together



Talk Outline

- └ Introduction
- Association mining
- Sequence mining
- Classification
- Conclusions



Association Mining

- A classical example of data mining
 - The very first use of “data mining” on this topic
- Attributes are binary, called *items*
 - A collection of items is called an *itemset*
- Each data point is a subset of the set of all attributes, called a *transaction*
- L , $0 < L < 1$, an input parameter called *minimum support*
- An itemset is *frequent* if it appears in at least L of the transactions
- **Goal: Identify all frequent itemsets**



Association Rule Mining

- Association rule mining
 - Another parameter C , $0 < C < 1$
 - Goal: Find all itemset pairs (S, T) , such that
 - Both S and T are frequent
 - At least C of the data points containing S contain T
 - Interpretation:
 - If you see S in a transaction you are likely to see T as well

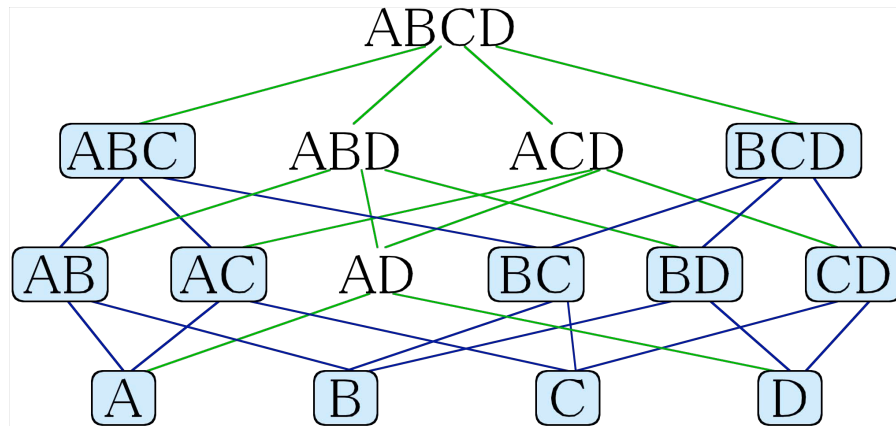
■ Basket Database: A Popular Example of Target Databases of Association Rule Mining

- Items purchased at a store
- Quantities are reduced to 0/1
- Customer ID removed, multiple transactions of same customers
- Association Mining
 - “At least 1.5% of the customers buy shampoo, toothpaste, and milk at one time.”
- Association Rule Mining
 - “If a customer buys shampoo and toothpaste then he/she is likely to buy milk with 90% of the time.”

■ The Subset Property of Frequent Attribute Sets

- Every subset of frequent attribute subset is frequent
- The frequent attribute sets form a set lattice ...
The mining problem is equivalent to the problem finding maximal elements in a set lattice
- The mining has exponential complexity ... Need to design heuristic search algorithms

Subset Property of Frequent Itemsets – Itemset Lattice



Major Issues in Association and Association Rule Mining

- Selection of the parameters is very important
 - The higher L or C, the fewer the number of itemsets or rules discovered
- Once the combinations have been identified their significance can be statistically measured (Chi-squared test) ... but can't use the measure to prune search
- Numeric attributes must be converted to binary ones
 - Solution: divide the range into overlapping/non-overlapping intervals ... difficult if the underlying distribution is unknown
- Missing data are often seen in real datasets
 - How should missing binary attributes be modeled?



The Apriori Algorithm (Agrawal & Srikant '93)

- Level-wise search in the itemset lattice:
 - Combine frequent itemsets of size k to generate candidates for frequent itemsets of size $(k+1)$
 - A size- $(k+1)$ itemset is a candidate if and only if all of its size- k subsets are frequent
- The proof-of-concept algorithm ... correct, but slow
 - Many heuristics have been designed



The CLIQUE Algorithm (Zaki, Ogihara, Parthasarathy, Li '98)

- Discover all frequent item pairs
- View the pairs as edges to draw a graph whose nodes are the items
 - Each frequent itemset is a subset of a clique in the graph
 - Use a maximal clique enumeration algorithm the pairs to prune the search space
- Can be efficiently run on parallel processor machines
- Twice to thrice faster than Apriori on single-processor machines on synthetic datasets



Rochester Perinatal Database

- Database of babies given birth in the Rochester area
- A very small number of vital statistics (height, weight, etc.)
- Mostly health conditions of the mothers collected by questionnaire
 - Before labor
 - During labor
- Analysis not successful
 - Too many missing data entries
 - 30% of data points had some missing data
 - 50% of attributes had missing entries
 - Too many high confidence rules
 - Expert knowledge must be incorporated (but never happened)



Talk Outline

- ┌ Introduction
- ┌ Association mining
- Sequence mining
- Classification
- Conclusions



Sequence Mining

- The time dimension is added to association mining
 - Each data point is a sequence of attribute sets, each labeled with a time point
 - One more parameter $W > 0$
 - Goal: Find all sequences of attribute sets that
 - Appear in at least L of the data points
 - The time interval between each neighboring attribute set pair is at most W



Sequence Mining of Basket Databases

- Items purchased; 0/1 instead of quantity
- Each data point has a unique customer ID, it is a sequence of transactions of the customer
- Sequence Mining
 - “At least 2.5% of the customers buy shampoo, toothpaste, and milk at one time, and then buy soap and peanut butter at one time in a month.”



Major Issues in Sequence Mining

- Again, parameter selection is critical
 - ... actually, more critical than association mining since the search space size is much more sensitive to parameter choice




Data Classification Using Frequent Patterns

- Artificial planning domain for fire extinction
 - 10x10 area of base, water, ground
 - Fire starts off in a point
 - Randomly fire spreads out
 - Ground and base are flammable, but water is not
 - Base must be protected from burning ... Failure if any one of base area is burnt
 - Bulldozers can be deployed to run over an area to extinguish fire



Data Classification Using Frequent Patterns

- 700 execution traces, where each entry of a trace has:
 - Time value
 - Wind direction & speed
 - Locations of bulldozers
 - Areas where fire started
- Goal: Identify patterns that predict ultimate failure



The FeatureMine Algorithm (Lesh, Zaki, Ogihara '01)

- Find sequences that are frequent in unsuccessful traces and are not frequent in successful traces
- Prune as much as possible early on
 - Collapse 100% implications
 - If B occurs after A occurs with probability 1 remove B from consideration
 - Remove partial sequences whose frequency is the same between successful traces and unsuccessful traces



Results of FeatureMine

- With FeatureMine, prediction accuracy is improved from 70% to 80% using Bayesian inference
- With the pruning strategy, CPU time for training was reduced from 5.8 hours to 560



Talk Outline

- ┌ Introduction
- ┌ Association mining
- ┌ Sequence mining
- Classification
- Conclusions



Classification

- Data points are divided into classes
- Goal: Develop an accurate, efficient algorithm for inferring membership of a new data point with unknown class membership
- Major Issues:
 - Extracting / selecting features that are useful for classification
 - Selection / development of classification algorithms



Popular Classification Algorithms

- Decision Trees
 - Data points are hierarchically divided into groups; each division is based on the value of a particular attribute
 - Can deal with more than two classes
- Gaussian Mixture Models
 - Numeric attributes
 - In each class, each attribute is subject to a linear combination of Gaussian distributions; obtain maximum-likelihood estimations of the parameters



Classification Algorithms

- Support Vector Machines
 - Find a linear separation with a wide gap between two classes
 - If linear separation is impossible in the given space, transform the data points into a higher dimensional space to separate linearly



Multi-class Extensions of Binary Classifiers

- One-versus-all: For each class, train a classifier that distinguishes the class from the rest. Assemble predictions of the classifiers.
- Pair-wise: For each class pair, train a classifier. Assemble predictions of them.
- Error-Correcting-Output-Coding (ECOC): Assemble many binary classifiers



Eigenvectors as Features

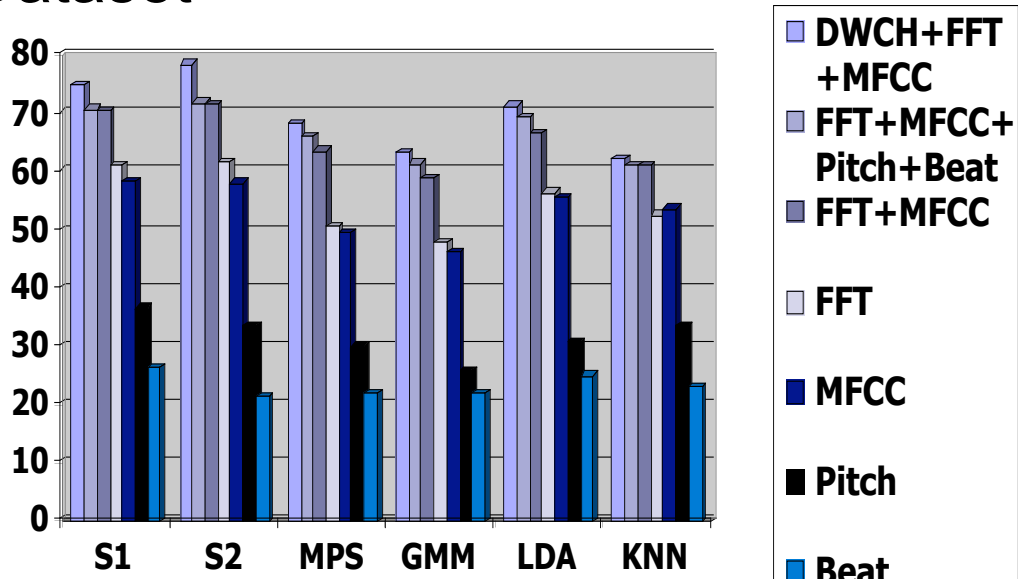
- Data points are viewed as points in a high dimensional space
- Use covariance matrix to view relations between coordinates
- Calculate largest eigenvectors, which represent the covariance
- Project data points on the eigenvectors



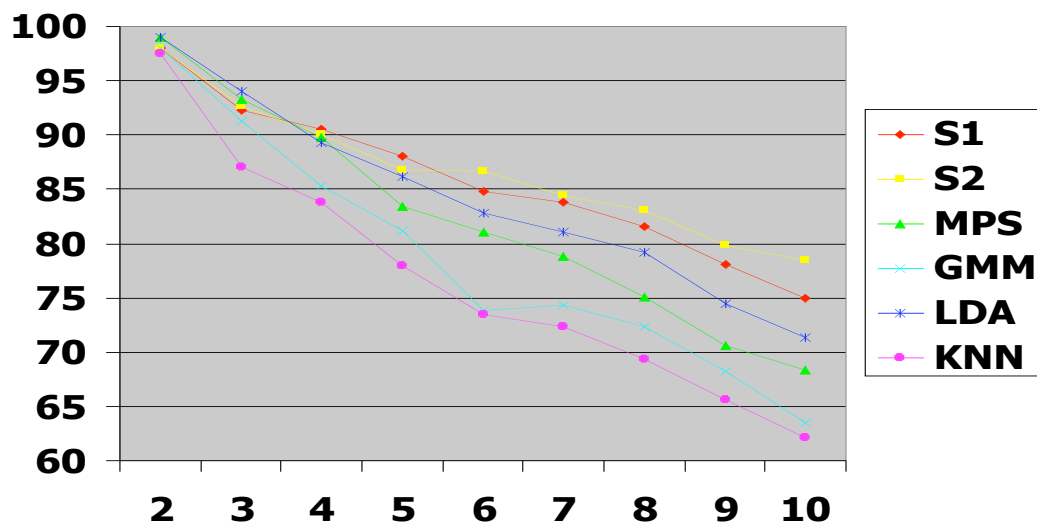
Music Information Retrieval

- Growing on-line music information (text & acoustic data)
 - Efficient tools for classifying and storing music data
- Classification of genre, style, artist, emotion, etc. is a fundamental problem
- Extracting features
 - Acoustic Data
 - Fast Fourier Transform
 - Wavelet histograms (Li, Ogihara, Li '03)
 - Text Data (Lyrics, in particular)
 - Bag-Of-Words, Part-of-Speech Statistics, Lexical Features, Orthographic Features

Genre Classification: 10 Genre Dataset



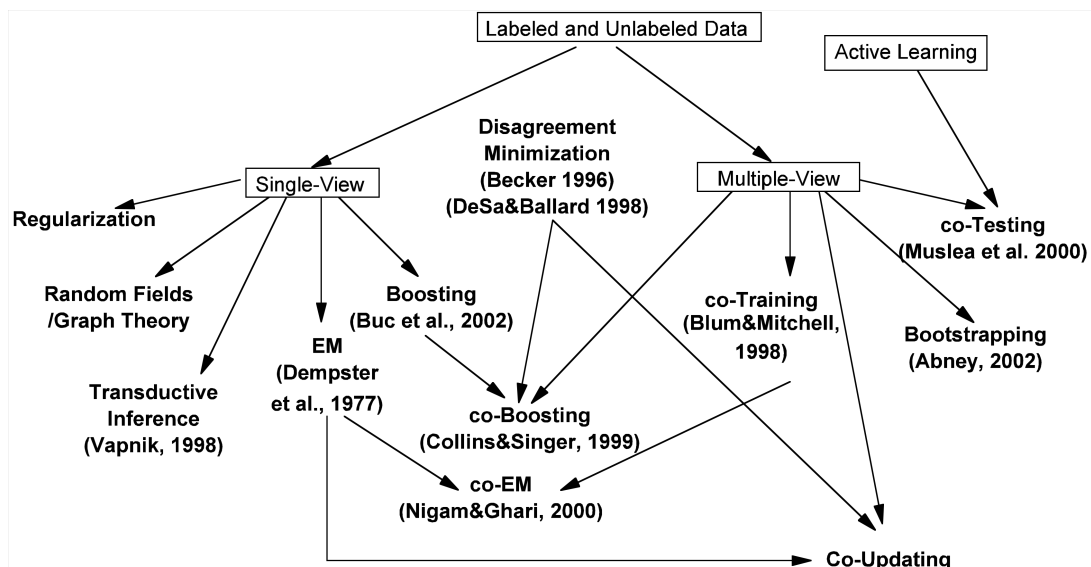
With an Increasing Number of Genres



Co-Updating (Li, Zhu, Ogiwara, Li '03)

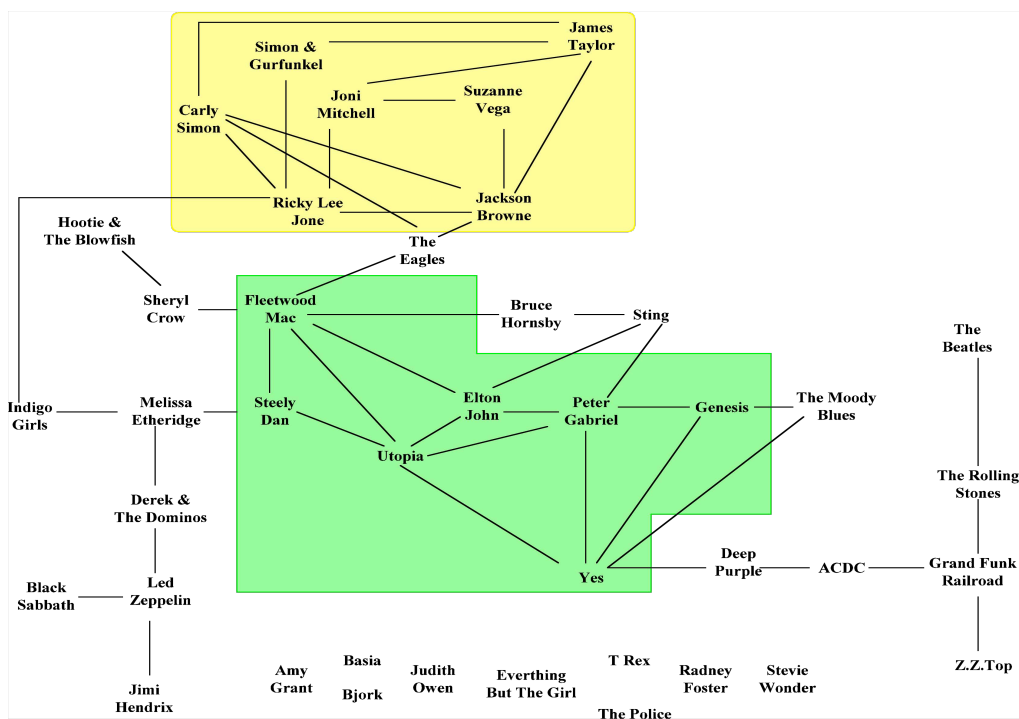
- Classifiers are built on more than one feature set
- Existence of unlabeled data
- Each classifier advises the others on the class membership of each unlabeled data point
 - Randomized process of removing disagreement among classifiers

Related Concepts



Artist Cluster Identification (Using All Music Guide as the Ground Truth)

- 45 artists, 55 albums
- Cluster 1: Fleetwood Mac, Yes, Utopia, Elton John, Genesis, Steely Dan, Peter Gabriel
- Cluster 2: Carly Simon, Joni Mitchell, James Taylor, Suzanne Vega, Ricky Lee Jones, Simon & Garfunkel



Classification Results: Cluster 1

Classifier	Accuracy	Precision	Recall
Lyrics	0.5065	0.5000	0.3844
Sound	0.5122	0.5098	0.3377
Combined	0.5309	0.5571	0.4675
Co-Upd./Lyrics	0.6358	0.5726	0.6221
Co-Upd./Sound	0.6853	0.6548	0.7143
Co-Upd./Combined	0.6875	0.6944	0.6494

Classification Results: Cluster 2

Classifier	Accuracy	Precision	Recall
Lyrics	0.5068	0.3824	0.4643
Sound	0.6027	0.4262	0.4677
Combined	0.6301	0.5167	0.5536
Co-Upd./Lyrics	0.6644	0.5636	0.5536
Co-Upd./Sound	0.6853	0.5833	0.6250
Co-Upd./Combined	0.6986	0.6111	0.5893



Talk Outline

- └ Introduction
- └ Association mining
- └ Sequence mining
- └ Classification
- **Conclusions**



Conclusions

- Data mining is exploration for knowledge in large databases
- Various techniques exist for mining
 - Choice of the technique is crucial for successful mining



Acknowledgements

- Yin-He Cheng (UR, CS Grad Student)
- Neal Lesh (MERL)
- Qi Li (U. Delaware, CIS Grad Student)
- Tao Li (UR, CS Grad Student)
- Srinivasan Parthasarathy (Ohio State U., CIS Faculty)
- Mohammed Zaki (RPI, CS Faculty)