8. New Research and Application Fields

Main Contents

- Data warehouse
- OLAP
- Data mining
- Information retrieval
- Semistructured data and XML

8.2 Data Mining

Definition:

Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data.

Example pattern (Census Bureau Data):
If (relationship = husband), then (gender = male). 99.6%

Definition (Cont.)

Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data.

- Valid: The patterns hold in general.
- Novel: We did not know the pattern beforehand.
- Useful: We can devise actions from the patterns.
- Understandable: We can interpret and comprehend the patterns.

Why Use Data Mining Today?

Human analysis skills are inadequate:

- Volume and dimensionality of the data
- High data growth rate

Availability of:

- Data
- Storage
- Computational power
- Off-the-shelf software
- Expertise

An Abundance of Data

- Supermarket scanners, POS data
- Preferred customer cards
- Credit card transactions
- Direct mail response
- Call center records
- ATM machines
- Demographic data
- Sensor networks
- Cameras
- Web server logs
- Customer web site trails

Much Commercial Support

- Many data mining tools
  - http://www.kdnuggets.com/software
- Database systems with data mining support
- Visualization tools
- Data mining process support
- Consultants

Why Use Data Mining Today?

Competitive pressure!

"The secret of success is to know something that nobody else knows."

Aristotle Onassis

- Competition on service, not only on price (Banks, phone companies, hotel chains, rental car companies)
- Personalization, CRM
- The real-time enterprise
- "Systemic listening"
- Security, homeland defense

Data Mining is Supported by Three Sufficiently Mature Technologies

- **Massive data collections**
  Commercial databases (using high performance engines) are growing at exceptional rates
- **Powerful multiprocessor computers**
  cost-effective parallel multiprocessor computer technology
- **Data mining algorithms**
  under development for decades, in research areas such as statistics, artificial intelligence, and machine learning, but now implemented as mature, reliable, understandable tools that consistently outperform older statistical methods

Applications, Operations, and Techniques

Applications	Database Marketing Customer Segmentation Customer Retention Fraud Detection Credit Checking Web Site Analysis, etc.
Operations	Classification and Prediction Clustering Association Analysis Forecasting
Techniques	Cluster Analysis Nearest Neighbour Neural Networks Naïve-Bayes Decision Trees

The Knowledge Discovery Process

Steps:

1. Identify business problem
2. Data mining
3. Action
4. Evaluation and measurement
5. Deployment and integration into businesses processes

Data Mining Step in Detail

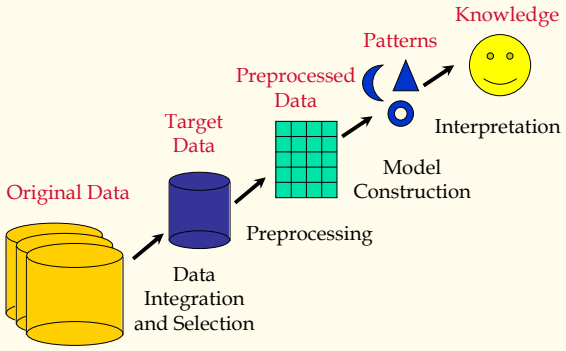
2.1 Data preprocessing

- Data selection: Identify target datasets and relevant fields
- Data cleaning
  - Remove noise and outliers
  - Data transformation
  - Create common units
  - Generate new fields

2.2 Data mining model construction

2.3 Model evaluation

Preprocessing and Mining



Example Application: Sky Survey

- Input data: 3 TB of image data with 2 billion sky objects, took more than six years to complete
- Goal: Generate a catalog with all objects and their type
- Method: Use decision trees as data mining model
- Results:
  - 94% accuracy in predicting sky object classes
  - Increased number of faint objects classified by 300%
  - Helped team of astronomers to discover 16 new high red-shift quasars in one order of magnitude less observation time

Example Application: Market Baskets

- An important form of mining involves *market baskets* = sets of "items" that are purchased together as a customer leaves a store.
- Summary of basket data is *frequent itemsets* = sets of items that often appear together in baskets.
- If people often buy hamburger and ketchup together, the store can:
  1. Put hamburger and ketchup near each other and put potato chips between.
  2. Run a sale on hamburger and raise the price of ketchup.

Data Mining: Types of Data

- Relational data and transactional data
- Spatial and temporal data, spatial-temporal observations
- Time-series data
- Text
- Images, video
- Mixtures of data
- Sequence data

- Features from processing other data sources

Types of Variables

- *Numerical*: Domain is ordered and can be represented on the real line (e.g., age, income)
- *Nominal* or *categorical*: Domain is a finite set without any natural ordering (e.g., occupation, marital status, race)
- *Ordinal*: Domain is ordered, but absolute differences between values is unknown (e.g., preference scale, severity of an injury)

Data Mining Operations

Each operation (or task) reflects a different way of distinguishing patterns or trends in complex datasets

- Classification and prediction
- Clustering
- Association analysis and sequential analysis
- Forecasting

Classification and Prediction

Classification is the operation most commonly supported by commercial data mining tools

- It is the process of sub-dividing a data set with regard to a number of specific outcomes.
  - For example, classifying customers into 'high' and 'low' categories with regard to credit risk.
- The category or 'class' into which each customer is placed is the 'outcome' of the classification.

Techniques for Classification and Prediction

- **Decision trees**
- **Neural networks**
- **Nearest neighbour algorithms**

Understanding v Prediction

- Sophisticated classification techniques enable us to discover new patterns in large and complex data sets.
- Classification is a powerful aid to understanding a particular problem.
- In some cases, improved understanding is sufficient. It may suggest new initiatives and provide information that improves future decision making.
- Often the reason for developing an accurate classification model is to improve our capability for prediction.

Training

- A classification model is said to be 'trained' on historical data, for which the outcome is known for each record.
- But beware overfitting: 100 per cent of customers called Smith who live at 28 Arcadia Street responded to our offer.
- One would then use a separate test dataset of historical data to validate the model.
- The model could then be applied to a new, unclassified data set in order to predict the outcome for each record.

Clustering

- Clustering is an unsupervised operation - no training.
- It is used to find groupings of similar records in a data set without any preconditions as to what that similarity may involve.
- Clustering is used to identify interesting groups in a customer base that may not have been recognised before.
- Often undertaken as an exploratory exercise before doing further data mining using a classification technique.

Techniques for Clustering

- **Cluster analysis**
- **Neural networks**
- **+ Good visualization support for "playing" with clusters, to see if they make sense and are useful in a business context**

Association Analysis

- **Association Rule** looks for links between records in a data set.

  Sometimes referred to as 'market basket analysis', its most common aim is to discover which items are generally purchased at the same time.

- **Time Sequential** looks for temporal links between purchases, rather than relationships between items in a single transaction.

Example of Association Rule

- Consider the following beer and nappy example:
  - 500,000 transactions
  - 20,000 transactions contain nappies (4%)
  - 30,000 transactions contain beer (6%)
  - 10,000 transactions contain both nappies and beer (2%)

Support (or prevalence)

- Measures how often items occur together, as a percentage of the total transactions.
- In this example, beer and nappies occur together 2% of the time (10,000/500,000).

Confidence (or predictability)

- Measures how much a particular item is dependent on another.
- Because 20,000 transactions contain nappies and 10,000 of these transactions contain beer, when people buy nappies, they also buy beer 50% of the time.
- The confidence for the rule:

  When people buy nappies they also buy beer 50% of the time.

  is 50%.

Confidence (cont)

- Because 30,000 transactions contain beer and 10,000 of these transactions contain nappies, when people buy beer, they also buy nappies 33.33% of the time.
- The confidence for the rule:

  When people buy beer they also buy nappies 1/3 of the time.

  is 33.33%.
- Both these rules have the same support: 2%

Expected Confidence

- In the absence of any knowledge about what else was bought, we can also make the following assertions from the available data:

  People buy nappies 4% of the time.

  People buy beer 6% of the time.
- These numbers - 4% and 6% - are called the **expected confidence** of buying nappies or beer, regardless of what else is purchased.

Lift

- Measures the ratio between the confidence of a rule and the expected confidence that the second product will be purchased. Lift is measures of the strength of an effect.
- In our example, the confidence of the nappies-beer buying rule is 50%, whilst the expected confidence is 6% that an arbitrary customer will buy beer.
- So, the lift provided by the nappies-beer rule is: 8.33 (= 50%/6%).
- A key goal of an association analysis is to find rules that have a substantial lift like this.

Forecasting

- Forecasting (unlike prediction based on classification models) concerns the prediction of continuous values, such a person's income based on various personal details, or the level of the stock market.
- Simpler forecasting problems involve a single continuous value based on a series of unordered examples.
- More complex problem is to predict one or more values based on a sequential pattern.
- Techniques include statistical time-series analysis as well as neural networks.

Data Mining Techniques

Each technique is used to support a particular data mining operation

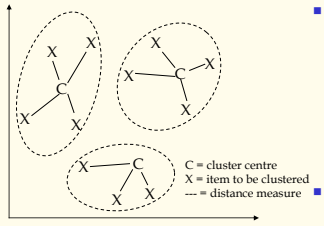
- Cluster Analysis
- Association Rules Algorithm
- Time Sequence Algorithm
- Decision Trees

Cluster Analysis

- Identifies the relationships that exist between items on the basis of their similarity and dissimilarity.
- Clusters are typically based around a "centre" value.
- How centres are initially defined and adjusted varies between algorithms.
- One method is to start with a random set of centres, which are then adjusted, added to and removed as the analysis progresses.

Cluster Analysis (cont)

- To identify items that belong to a cluster, some measure must be used that gauges the similarity between items within a cluster and their dissimilarity to items in other clusters.



- This is typically measured as their distance from each other and from the cluster centres within a multi-dimensional space, where each dimension represents one of the variables being compared.
- The key is the design and implementation of distance function.

Association Rule Algorithm

- Apriori algorithm (R.Agrawal, etc.)
  1. Find out all frequent itemsets
  2. Generate candidate association rules based on frequent itemsets and verify their cinfidence.

```
Procedure AprioriAlg()
Begin
    L1:={frequent 1-itemsets};      /* 1-itemsets is itemsets with single item */
    for(k:=2; L_{k-1}≠Φ; k++) {
    C_k := AprioriGen(L_{k-1});     /* candidate itemset generated by AprioriGen */
    forall transactions in the dataset do
      {forall candidates c ∈ C_k contained in t do
       c.count++;
      }
    L_k := {c ∈ C_k | c.count >= min-support}
    }
    Answer := ∪_k L_k;
end
```

Time Sequence Algorithm

- AprioriALL algorithm

1. Sort (according to customer ID and then transaction time
2. Find out frequent itemset (similar as Apriori, but count according to customer)
3. Transform customer transaction sequence
4. Find out frequent sequences (similar as Apriori, but AprioriGen() is some different)
5. Select maximum frequent sequences, they are the answer.

Classification Algorithm

- The generation and maintenance algorithm of decision tree

```
partition(S)
    if(all tuples in S are of the same class ) then
      return;          /*don't need to classify */
    foreach value of attribute A do
      evaluate the split on that value;
    use best split found to partition S into S1 and S2;
    partition S1;      /* call partition() recursively */
    partition S2;
```

- Evaluate each split:

Gini(S) = 1 - $\sum_{j=1}^{n} p_j^2$ /* n is class number in S; pj is frequency of every class in S */

$Gini_{split}$ (S) = (N1/N)*gini(S1) + (N2/N)*gini(S2)

Select smallest $Gini_{split}$ (S) as split condition.