Data Mining Methods and Applications

S Prathibha Sai, B Shalini, JS Anand Kumar, Assit. Professor

Abstract : Data mining is a process to store large amount of data. The paper discuss data mining methods and applications. This data mining methods and applications improve our business and found extradinary results .. Data Mining is a broad term for a variety of data analysis techniques applied to the problem of extracting meaningful knowledge from large, noisy databases. An important feature present in most of these techniques is an ability to adapt to the local characteristics of the data. Such techniques are applied to electric load profiling tasks; load profiling consists of modelling the way in which daily load shape (load profile) relates to various factors such as weather, time and customer characteristics. An implementation of an adaptive load profiling methodology is presented.

1.INTRODUCTION

For this recent interest in the data mining area arises from its applicability to a wide variety of problems, including not only databases containing consumer and transaction information, but also advanced databases on multimedia, spatial and temporal information. In this paper, we will concentrate on discussing a few of the important problems in the topic of data mining. These problems include those of finding associations, clustering and classification. In this section, we will provide a brief introduction to each.

2.DATA MINING OVERVIEW :

The development of information technology has generated large amount of database and huge amount of data collect to different areas.

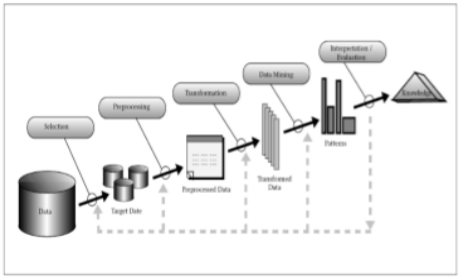


Figure 1 : Knowlwdge discovery process.

The process of data mining is a logical process and order to search data is useful data.

3. DATA MINING ALGORITHMS AND TECHNIQUES

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from database.

DATA MINING TYPES:

Predicative: It produces the model of the system described by the given data. It uses some variables or fields in the data set to predict unknown or future values of other variables of interest.

Descriptive: It produces new, non trivial information based on the available data set. It focuses on finding patterns describing the data that can be interpreted by humans.

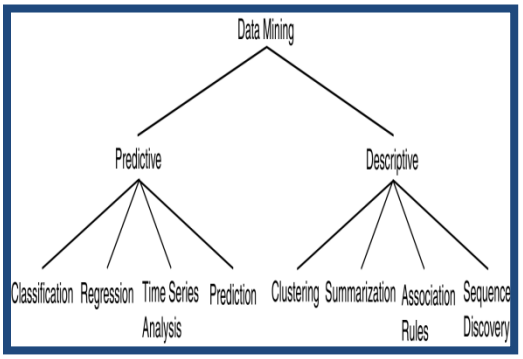


Figure 2: Data Mining techniques types

There are several major data mining techniques have been developing and using in data mining projects recently
including association, classification, clustering, predicti on, sequential patterns and decision tree. We will briefly examine those data mining techniques in the following sections.

3.1 ASSOCIATION:

This problem often occurs in the process of finding relationships between different attributes in large customer databases. These attributes may either be 0-1 literals, or they may be quantitative. The idea in the association rule problem is to find the nature of the causalities between the values of the different attributes. Consider a supermarket example in which the information maintained for the different transactions is the sets of items bought by each consumer. In this case, it may be desirable to find how the purchase behavior of one item affects the purchase behavior of another. Association Rules help in finding such relationships accurately. Such information may be used in order to make target marketing decisions. It can also be generalized to do classification of high dimensional data.

Types of association rules:

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

S. Prathibha Sai, Dept. of MCA 2nd year, KMMIPS, Tirupati, mail id:prathibharaji63@gamil.com.

B.Shalini, Dept. of MCA 2nd year, KMMIPS, Tirupati, Mail id:bobbushalini2@gamil.com.

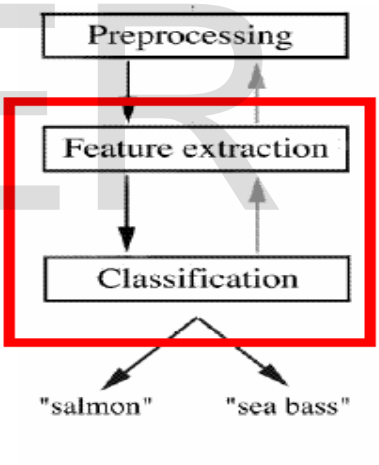
J S Ananda Kumar, Dept .of MCA, KMMIPS, Tirupati, mail id: jsanandkumar@gmail.com.

3.2 CLASSIFICATION :

Following are the examples of cases where the data analysis task is Classification −

- A bank loan officer wants to analyze the data in order to know which customer (loan applicant) are risky or which are safe.
- A marketing manager at a company needs to analyze a customer with a given profile, who will buy a new computer.

In both of the above examples, a model or classifier is constructed to predict the categorical labels. These labels are risky or safe for loan application data and yes or no for marketing data.



- **Preprocessing:** Images of different fishes are isolated from one another and from background;
- **Feature extraction:** The information of a single fish is then sent to a feature extractor, that measure certain “features” or “properties”;
- **Classification:** The values of these features are passed to a classifier that evaluates the evidence presented, and build a model to discriminate between the two species

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and creditrisk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

Types of classification models:

- Classification by decision tree induction
- Bayesian Classification
- Back propagation
- Support Vector Machines (SVM)
- Classification Based on Associations
- Genetic Algorithm
- Rough Set Approach
- Fuzzy Set Approach

3.3 CLUSTERING :

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as apreprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality.

TYPES OF CLUSTERING METHODS

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

3.4 PREDICTION :

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models. Types of regression methods

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

Following are the examples of cases where the data analysis task is Prediction –

Suppose the marketing manager needs to predict how much a given customer will spend during a sale at his company. In this example we are bothered to predict a numeric value. Therefore the data analysis task is an example of numeric prediction. In this case, a model or a predictor will be constructed that predicts a continuous-valued-function or ordered value.

HOW DOES IT DIFFER FROM CLASSIFICATION?

- A classification problem could be seen as a predictor of classes.
- Predicted values are usually continuous whereas classifications are discreet.
- Predictions are often (but not always) about the future whereas classifications are about the present.
- Classification is more concerned with the input than the output

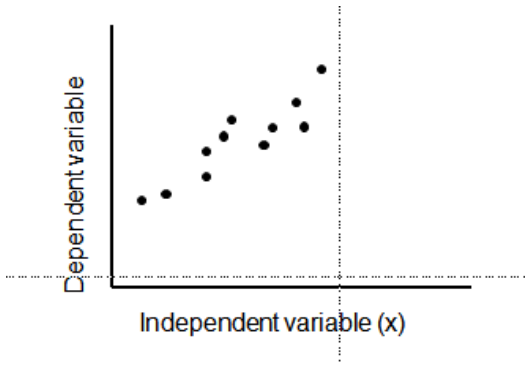
Usual Examples

- Predicting levels of sales that will result from a price change or advert.
- Predicting whether or not it will rain based on current humidity
- Predicting the colour of a pottery glaze based on a mixture of base pigments
- Predicting how far up the charts a single will go
- Predicting how much revenue a book of debt will bring

Techniques

- Most prediction techniques are based on mathematical models:
  - Simple statistical models such as regression
  - Non-linear statistics such as power series
  - Neural networks, RBFs, etc
- All based on fitting a curve through the data, that is, finding a relationship from the predictors to the predicted.

3.5 REGRESSION:



Regression is the attempt to explain the variation in a dependent variable using the variation in independent variables.

TYPES OF REGRESSION METHODS

- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

Regression is thus an explanation of causation.

If the independent variable(s) sufficiently explain the variation in the dependent variable, the model can be used for prediction.

- Regression testing is applied to code immediately after changes are made.
- The goal is to assure that the changes have not had unintended consequences on the behaviour of the test object.
- We can apply regression testing during development and in the field after the system has been upgraded or maintained in some other way.
- Good regression tests give us confidence that we can change the object of test while maintaining its intended behaviour.
- So, for example, we can change to a new version of some piece of infrastructure in the environment, make changes to the system to take

account of that and then ensure the system behaves as it should.

- Regression testing is an important way of monitoring the effects of change.
- There are many issues but the balance of confidence against cost is critical.

3.6 TIME SERIES ANALYSIS:

One definition of a time series is that of a collection of quantitative observations that are evenly spaced in time and measured successively. Examples of time series include the continuous monitoring of a person's heart rate , hourly readings of air temperature, daily closing price of a company stock, monthly rainfall data, and yearly sales figures. Time series analysis is generally used when there are 50 or more data points in a series. If the time series exhibits seasonality, there should be 4 to 5 cycles of observations in order to fit a seasonal model to the data.

GOALS OF TIME SERIES ANALYSIS:

1. **Descriptive:** Identify patterns in correlated data trends and seasonal variation
2. **Explanation:** Understanding and modeling the data.
3. **Forecasting:** Prediction of short-term trends from previous patterns.
4. **Intervention analysis:** How does a single event change the time series?
5. **Quality control:** Deviations of a specified size indicate a problem.

3.7 SUMMARIZATION:

Summarization is taking a large amount of information and condensing it so that the main points are covered but there is a reduced amount of statements. A summary should be comprehensive and state the important and pertinent information in a brief and concise format. Summarization is used in many ways in our day to day lives. Books usually contain summaries on the back or dust jacket which briefly explain the plot to help a person decide if they are interested in reading it. Sections at the end of chapters in textbooks frequently summarize the contents of the chapter. It will list important terms and concepts introduced so that the take away message contains the most relevant information. Study guides are also forms of summarization. For finals instead of reading through the entirety of the notes from the class it is much more helpful to summarize and condense the information into a few pages for a more manageable tool for studying.

4. DATA MINING APPLICATIONS:

Data mining is a relatively new technology that has not fully matured. Despite this, there are a number of industries that are already using it on a regular basis. Some of these organizations include retail stores, hospitals, banks, and insurance companies. Many of these organizations are combining data mining with such things as statistics, pattern recognition, and other important tools. Data mining can be used to find patterns and connections that would otherwise be difficult to find. This technology is popular with many businesses because it allows them to learn more about their customers and make smart marketing decisions. Here is overview of business problems and solutions found using data mining technology.

A. Financial Data Analysis:

The financial data in banking and financial industry is generally reliable and of high quality which facilitates the systematic data analysis and data mining. Here are the few typical cases: Design and construction of data warehouses for multidimensional data analysis and data mining. Loan payment prediction and customer credit policy analysis. Classification and clustering of customers for targeted marketing. Detection of money laundering and other financial crimes.

B. Retail Industry:

Data Mining has its great application in Retail Industry because it collects large amount data from

on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of increasing ease, availability and popularity of web. The Data Mining in Retail Industry helps in identifying customer buying patterns and trends. That leads to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in retail industry:

- Design and Construction of data warehouses based on benefits of data mining.
- Multidimensional analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.
- Product recommendation and cross-referencing of items.

C. Telecommunication Industry:

Today the Telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, Internet messenger, images, email, web data transmission etc. Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business. Data Mining in Telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list examples for which data mining improve telecommunication services as

- Multidimensional Analysis of Telecommunication data.
- Fraudulent pattern analysis.
- Identification of unusual patterns.
- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.
- Use of visualization tools in telecommunication data analysis.

D. Biological Data Analysis:

Now a days we see that there is vast growth in field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is very important part of Bioinformatics. Following are the aspects in which Data mining contribute for biological data analysis:

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.

CONCLUSION: Data mining has importance regarding finding the patterns, forecasting, discovery of knowledge etc., in different business domains. Data mining techniques and algorithms such as classification, clustering etc., helps in finding the patterns to decide upon the future trends in businesses to grow. Data mining has wide application domain almost in every industry where the data is generated that's why data mining is considered one of the most important frontiers in database and information systems and one of the most promising interdisciplinary developments in Information Technology.

REFERENCES:

1. https://www.allbusiness.com/Technology /computer-software-data-management/ 633425-1.html, last retrieved on 15th Aug 2010.
2. http://www.kdnuggets.com/.
3. Paulraj Ponnian, .Data Warehousing Fundamentals. John Wiley.
4. M. Holsheimer and A. Siebes, Data Mining - The Search For Knowledge in Databases, Report CS-R9406, CWI, Amsterdam, ISSN 0169-118-X.
5. G.Piatetsky-Shapiro,W.J.Frawley(Eds.) Knowledge Discovery in Databases,AAAI Press, 1991.