ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 2, Issue 2, pp: (258-264), Month: April-June 2014, Available at: www.researchpublish.com

A Study of Predictive Data Mining Techniques

M.K.Saranya[1], R.Rathnavathy[2] and Dr.G.N.K.Suresh Babu[3]

[1,2]Research Scholars, Dept of Computer Applications, GKM College of Engineering and Technology, Chennai-63
[3]Professor and Head, Dept of Computer Applications, GKM College of Engineering and Technology, Chennai-63

Abstract: Large numbers of data are generated everyday in many organizations. To extract hidden predictive information from large volumes of data, data mining (DM) techniques are needed. Organizations are starting to realize the importance of data mining in their strategic planning and successful application of DM techniques can be an enormous payoff for the organizations. This paper discusses the requirements and challenges of DM, and describes major DM techniques such as statistics, artificial intelligence, decision tree approach, genetic algorithm, and visualization. DM is the search for valuable information in large volumes of data. It is the process of nontrivial extraction of implicit, previously unknown and potentially useful information such as knowledge rules, constraints, and regularities from data stored in repositories using pattern recognition technologies as well as statistical and mathematical techniques. Many companies have recognized DM as an important technique that will have an impact on the performance of the companies.

Keywords: Statistics, Machine Learning, Genetic Algorithms.

1. INTRODUCTION

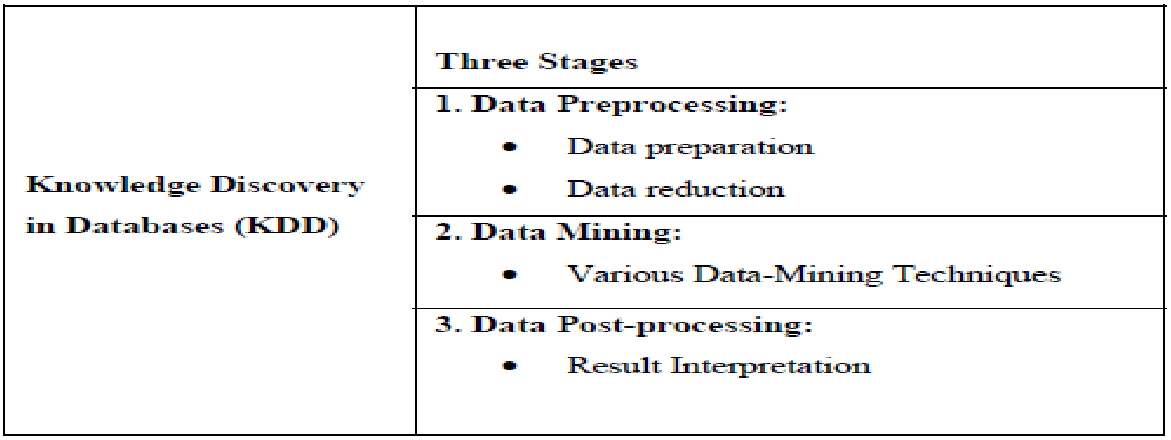
In recent years, data-mining (DM) has become one of the most valuable tools for extracting and manipulating data and for establishing patterns in order to produce useful information for decision-making. The failures of structures, metals, or materials (e.g.buildings, oil, water or sewage pipes) in an environment are often either a result of ignorance or the inability of people to take note of past problems or study the patterns of past incidents in order to make informed decisions that can forecast future occurrences. Nearly all areas of life activities demonstrate a similar pattern. Whether the activity is finance, banking, marketing, retail sales, production, population study, employment, human migration, health sector, monitoring of human or machines, science or education, all have ways to record known information but are handicapped by not having the right tools to use this known information to tackle the uncertainties of the future. Planning for the future is very important in business. Estimates of future values of business variables are needed. The commodities industry needs prediction or forecasting of supply, sales, and demand for production planning, sales, marketing and financial decisions. In a production or manufacturing environment, we battle with the issues of process optimization, job-shop scheduling, sequencing, cell organization, quality control, human factors, material requirements planning, and enterprise resource planning in lean environments, supply-chain management, and future-worth analysis of cost estimations, but the knowledge of data-mining tools that could reduce the common nightmares in these areas is not widely available.

Predictive data mining (PDM) works the same way as does a human handling data analysis for a small data set; however, PDM can be used for a large data set without the constraints that a human analyst has. PDM "learns" from past experience and applies this knowledge to present or future situations. Predictive data-mining tools are designed to help us understand what the “gold,” or useful information looks like and what has happened during past “gold-mining” procedures. Therefore, the tools can use the description of the “gold” to find similar examples of hidden information in the database and use the information learned from the past to develop a predictive model of what will happen in the future.

2. KNOWLEDGE DISCOVERY IN DATABASES (KDD)

KDD is an umbrella name for all those methods that aim to discover relationships and regularity among the observed data. KDD includes various stages, from the identification of initial business aims to the application decision rules. It is, therefore, the name for all the stages of finding and discovering knowledge from data, with data-mining being one of the stages. "data mining is the process of selection, exploration, and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database."

Table 1. Represents three stages of KDD



Knowledge Discovery in Databases (KDD)	Three Stages
	1. Data Preprocessing: • Data preparation • Data reduction
	2. Data Mining: • Various Data-Mining Techniques
	3. Data Post-processing: • Result Interpretation

3. CLASSIFYING DM TECHNIQUES

Many DM techniques and systems have been developed and designed. These techniques can be classified based on the database, the knowledge to be discovered, and the techniques to be utilized. Based on the database There are many database systems that are used in organizations, such as relational database, transaction database, object oriented database, spatial database, multimedia database, legacy database, and Web database. A DM system can be classified based on the type of database it is designed for. For example, it is a relational DM system if the system discovers knowledge from relational database and it is an object-oriented DM system if the system finds knowledge from object-oriented database. Based on the knowledge DM systems can discover various types of knowledge, including association rules, characteristic rules, classification rules, clustering, evolution, and deviation analysis. DM systems can also be classified according to the abstraction level of the discovered knowledge. The knowledge may be classified into general knowledge, primitive-level knowledge, and multiple level knowledge. Based on the techniques DM systems can also be categorized by DM techniques. For example, a DM system can be categorized according to the driven method, such as autonomous knowledge mining, data driven mining, query-driven mining, and interactive DM techniques. Alternatively, it can be classified according to its underlying mining approach, such as generalization based mining, pattern-based mining, statistical- or mathematical-based mining and integrated approaches.

4. MAJOR DM TECHNIQUES

In this section, we review and discuss the major DM techniques:

4.1 Statistics

Statistics is an indispensable component in data selection, sampling, DM, and extracted knowledge evaluation. It is used to evaluate the results of DM to separate the good from the bad. In data cleaning process, statistics offer the techniques to detect ``outliers'', to smooth data when necessary, and to estimate noise. Statistics can also deal with missing data using estimation techniques. Techniques in clustering and designing of experiments come into play for exploratory data analysis. Work in statistics, however, has emphasized generally on theoretical aspects of techniques and models. As a result, search, which is crucial in DM, has received little attention. In addition, interface to database, techniques to deal with massive data sets, and techniques for efficient data management are very important issues in DM.

4.2 Artificial intelligence (AI)

AI techniques are widely used in DM. Techniques such as pattern recognition, machine learning, and neural networks have received much attention. Other techniques in AI such as knowledge acquisition, knowledge representation, and search, are relevant to the various process steps in DM. Classification is one of the major DM problems. Classification is the process of dividing a data set into mutually exclusive groups such that the members of each group are as ``close'' as possible to one another, and the members of different groups are as ``far'' as possible from one another. For example, a typical classification problem is to divide a database of customers into groups that are as homogeneous as possible with respect to a variable such as creditworthiness. One solution to the classification problem is to use neural network. Neural network-based DM approach consists of three major phases:

4.2.1 Network construction and training: in this phase, a layered neural network based on the number of attributes, number of classes, and chosen input coding method are trained and constructed.

4.2.2 Network pruning: in this phase, redundant links and units are removed without increasing the classification error rate of the network.

4.2.3 Rule extraction: classification rules are extracted in this phase.

4.3 Decision tree

Decision trees are tree-shaped structures that represent sets of decisions. The decision tree approach can generate rules for the classification of a data set. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a data set. They provide a set of rules that can be applied to a new (unclassified) data set to predict which records will have a given outcome. CART typically requires less data preparation than CHAID.

4.4 Genetic algorithm

Genetic algorithm is a relatively new software paradigm inspired by Darwin's theory of evolution. A population of rules, each representing a possible solution to a problem, is initially created at random. Then pairs of rules (usually the strongest rules are selected as parents) are combined to produce offspring for the next generation. A mutation process is used to randomly modify the genetic structures of some members of each new generation. The system runs for dozens or hundreds of generations. The process is terminated when an acceptable or optimum solution is found, or after some fixed time limit. Genetic algorithms are appropriate for problems that require optimization with respect to some computable criterion. This paradigm can be applied to DM problems. The quantity to be minimized is often the number of classification errors on a training set. Large and complex problems require a fast computer in order to obtain appropriate solutions in a reasonable amount of time. Mining large data sets by genetic algorithms has become practical only recently due to the availability of affordable high-speed computers.

4.5 Visualization

A picture is worth thousands of numbers! Visual DM techniques have proven the value in exploratory data analysis, and they also have a good potential for mining large database. This approach requires the integration of human in the DM process. Visualization techniques have been extended to work on large data sets and produce interactive displays. There are several well-known techniques for visualizing multidimensional data sets: scatterplot matrices, coplots, prosection matrices, parallel coordinates, projection matrices, and other geometric projection techniques such as icon-based techniques, hierarchical techniques, graph-based techniques, and dynamic techniques.

5. PREDICTIVE DATA MINING

Data mining is the exploration of historical data (usually large in size) in search of a consistent pattern and/or a systematic relationship between variables; it is then used to validate the findings by applying the detected patterns to new subsets of data. The roots of data mining originate in three areas: classical statistics, artificial intelligence (AI) and machine learning. Data mining as a blend of statistics, artificial intelligence, and database research, and noted that it was not a field of interest to many until recently. Data mining can be divided into two tasks: predictive tasks and descriptive tasks. The ultimate aim of data mining is prediction; therefore, predictive data mining is the most common type of data mining and is the one that has the most application to businesses or life concerns.

DM starts with the collection and storage of data in the data warehouse. Data collection and warehousing is a whole topic of its own, consisting of identifying relevant features in a business and setting a storage file to document them. It also involves cleaning and securing the data to avoid its corruption. A data ware house is a copy of transactional or non-transactional data specifically structured for querying, analyzing, and reporting. Data exploration, which follows, may include the preliminary analysis done to data to get it prepared for mining. The next step involves feature selection and or reduction. Mining or model building for prediction is the third main stage, and finally come the data post-processing, interpretation, and/or deployment.

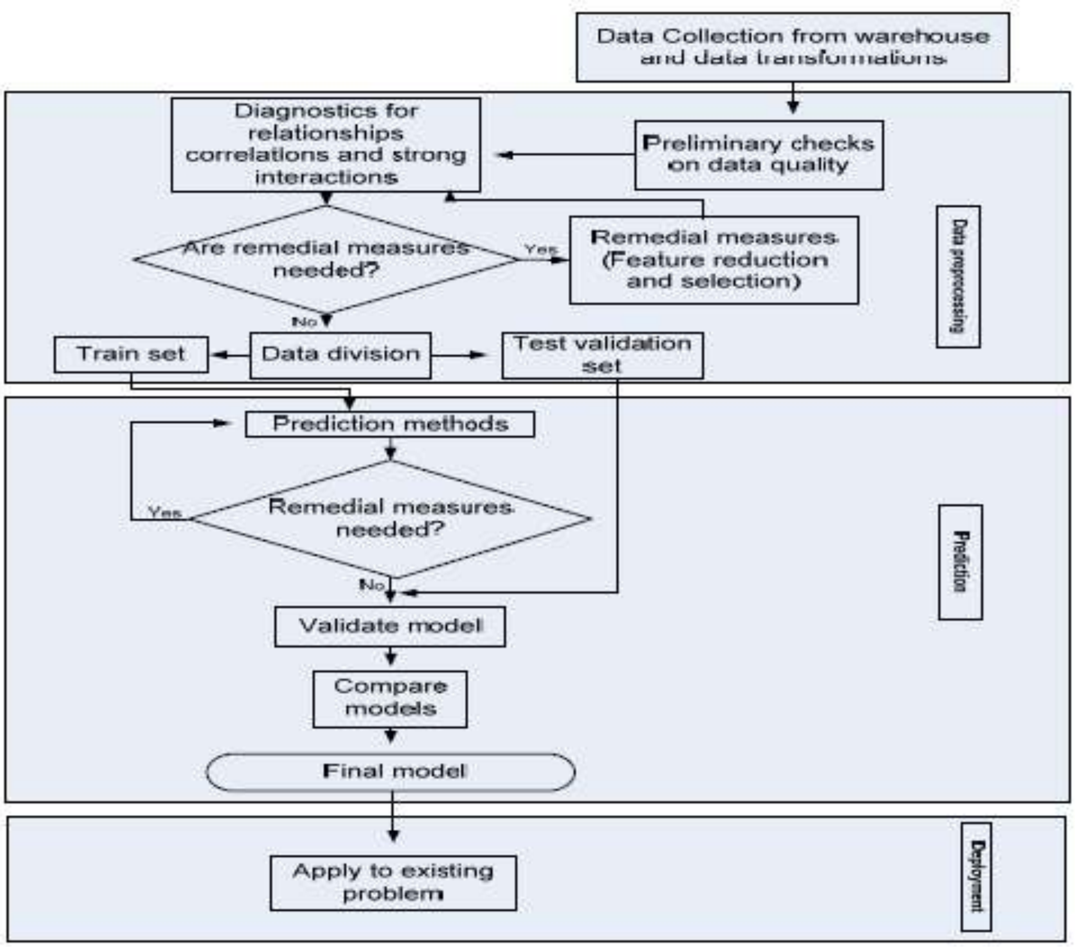

Data Collection from warehouse and data transformations
Diagnostics for relationships correlations and strong interactions
Preliminary checks on data quality
Are remedial measures needed?
Yes
Remedial measures (Feature reduction and selection)
No
Train set
Data division
Test validation set
Data preprocessing
Prediction methods
Yes
Remedial measures needed?
No
Validate model
Compare models
Final model
Prediction
Apply to existing problem
Deployment

Figure 1 represents the stages of predictive data mining.

6. PREDICTIVE ANALYTICS TECHNIQUES

6.1 Linear regression is the first kind of regression analysis to be studied and applied in various practical applications like epidemiology, environmental science and finance. In a very lucid term, linear regression otherwise called as straight line regression analysis is a regression to estimate the unknown effect of changing one variable over another. Specifically, it models *Y* as a linear function of *X* i.e. how much *Y* changes when *X* changes one unit. So it is expressed as a straight line equation:

$Y = b + w_X$ (1)

Here *b* and w are the regression coefficients where *b* is the *Y* intercept and *w* is the slope of the line. In cases, the coefficients can be assumed to be weights, where:

$Y=w_0+w1X$ *(2)*

This can be solved for the coefficients by method of least squares so as to minimize the error between the actual data and the estimated data. A training data set *D*, consisting of several predictor variable *X* and response variable *Y*,

$|D| = \{ (X1 , Y1), (X2 , Y2)....(X|D| , Y|D|) \}$ (3)

The estimated regression coefficient is given as:

$w_1 = \Sigma_i (X_i – MeanX) (Yi – MeanY) / \Sigma i (Xi – Mean_X)^2$ *where i = 1 to |D|,* (4)

$w_0=Mean_Y–w_1\text{-}Mean_x$ *(5)*

Linear regression model identifies the relationship between a single predictor variable *Xi* and the response variable *Y* when all other predictor variables in the model are "held fixed". This is called as the *unique effect Xi* on *Y*.

6.2 Multiple linear regression (MLR) is a mathematical technique that uses a number of variables to predict some unknown variable. It is a study on the relationship between a single dependent variable and one or more independent variables. This model describes a dependent variable *Y* by independent variable $X_1, X_2...X_p$ *(p>1)* is expressed by the equation as ,

$Y = \alpha + \Sigma k\ \beta k\ Xk + €$ (6)

Where *α, βk (k = 1,2...p)* are the parameters and *€* is the error term. MLR combines the idea of correlation and linear regression.

6.3 Logistic regression is a type of predictive model that can be used when the target variable is categorical variable that has exactly two categories like, win game/doesn't win, live/die. Technically it can be said as logistic regression is used for binomial regression. Simultaneously it also applies to continuous target variable that models the probability of some event occurring as a linear function of a set of predictor variable. Due to this it has extensively applied in the field of medical sciences, marketing application and social sciences. Mathematically,

$f(Z) = eZ / (eZ + 1) = 1 / (1+e –Z)$ (7)

Z is called as the *logit*, exposure to some set of independent variable *f(Z)* is probability of a particular outcome. Logistic regression takes *Z* as input and outputs *f(Z)* i.e. it can take input as any value from negative to positive infinity and give output between 0 and 1.

7. MORE ADVANCED PREDICTIVE ANALYTICS TECHNIQUES

7.1 Time series forecasting predicts the future value of a measure based on past values.Time series forecasting uses a model to forecast future events based on known past events. Examples include stock pries and sales revenue.

7.2 Data profiling and transformation uses functions that analyze row and column attributes and dependencies, change data formats, merge fields, aggregate records, and join rows and columns.

7.3 Bayesian analytics capture the concepts used in probability forecasting. It is a statistical procedure which estimate parameters of an underlying distribution based on the observed distribution.

7.4 Regression analysis is a statistical tool for the investigation of relationships between variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another-the effect of a price increase upon demand, for example, or the effect of changes in the money supply upon the inflation rate.

7.5 Classification used attributes in data to assign an object to a predefined class or predict the value of a numeric variable of interest. Examples include credit risk analysis, likelihood to purchase. Examples include acquisition, cross-sell, attrition, credit scoring and collections.

7.6 Clustering or segmentation separates data into homogeneous subgroups based on attributes. Clustering assigns a set of observations into subsets (clusters) so that observations in the same cluster are similar. An example is customer demographic segmentation.

8. APPLICATIONS OF PREDICTIVE ANALYTICS

Generally Predictive Analytics can be put to use in many applications, some of them are as follows:

• **Analytical customer relationship management (CRM)** in which Predictive Analysis is applied to customer data to pursue CRM objectives.

• **Clinical decision support systems** most frequently use Predictive Analysis in health care to determine patient severity in certain conditions like heart disease, cancer, diabates, and other life time illnesses.

• **Cross-sell ,** collect and maintain abundant data (e.g. customer records, sale transactions) and exploiting hidden relationships in the data can provide a competitive advantage to the organization. For an organization that offers multiple products, an analysis of existing customer behavior can lead to efficient cross sell of products. This directly leads to higher profitability per customer and strengthening of the customer relationship. Predictive analytics can help analyze customers' spending, usage and other behavior, and help cross-sell the right product at the right time.

• **Fraud detection,** fraudulent insurance claims and credit card transactions alone cost tens of billions of dollars a year. In the case of credit card fraud, artificial neural-networks have been widely-used by many banks. The pattern of fraudulent transactions varies with time, requiring relatively frequent and rapid generation of new models..

• **Direct marketing** are the amount of competing services available, businesses need to focus efforts on maintaining continuous consumer satisfaction. In such a competitive scenario, consumer loyalty needs to be rewarded and customer attrition needs to be minimized Predictive analytics can also predict this behavior accurately and before it occurs, so that the company can take proper actions to increase customer activity.

9. CONCLUSION AND FUTURE WORK

Having the right information at the right time is crucial for making the right decision. The problem of collecting data, which used to be a major concern for most organizations, is almost resolved. In the millennium, organizations will be competing in generating information from data and not in collecting data. Industry surveys indicated that over 80 percent of Fortune 500 companies believe that data mining would be a critical factor for business success. This study mainly intends to focus on the mining techniques using predictive analytics. Since predictive analytics is a major area of interest to almost all communities and organization, the application of it has provided a very high level of predictive performance. At the same time the widespread availability of several new computational methods and tools for predictive modeling assists the researchers and the practitioners to select the most appropriate strategy. We have presented an overview of some of the notable techniques for prediction. All analytical tools enable greater transparency and can find and analyze past trends to predict the probable future outcome of an event or its likelihood to occur, as well as to discover the hidden nature of data. We can use the above techniques hybridized with few soft computing techniques to predict the future trends. DM will be one of the main competitive focuses of organizations. Although progresses are continuously been made in the DM field, many issues remain to be resolved and much research has to be done.

REFERENCES

[1] Agresti, Alan. (2002). Categorical Data Analysis. New York: Wiley-Interscience. ISBN 0-471-36093-7.

[2] Banerjee and J. Langford. An objective evaluation criterion for clustering. In Proceedings of KDD-2004. ACM, New York, 2004.

[3] Enders, Walter (2004). Applied Time Series Econometrics. Hoboken: John Wiley and Sons. ISBN 052183919X.

[4] Greene, William H. (2003). Econometric Analysis, fifth edition. Prentice Hall. ISBN 0-13-066189-9.

[5] Jonathan T, (2009), Introduction to Applied Statistics, Introduction to applied statistics, Department of statistics, Stanford University, Statistics 191, pp. 1-38.

[6] Patricia E.N. Lutu , Andries P. Engelbrecht , "A decision rule-based method for feature selection in predictive data mining",Expert Systems with Applications 37 (2010) 602-609.

[7] Riccardo Bellazzi , Blaz Zupan , "Predictive data mining in clinical medicine: Current issues and guidelines", International Journal of Medical Informatics 77 (2008) 81-97.

[8] Se June Hong , Sholom M. Weiss, "Advances in predictive models for data mining" Pattern Recognition Letters 22 (2001) 55 – 61.

[9] Smyth G. K., (2002), Non Linear Regression, Journal of American Statistical association, Encyclopaedia of Environ metrics, ISBN 0471 899976, Vol. 3, pp. 1405-1411.

[10] Weiss, S.H. and Indurkhya, N. (1998), Predictive Data Mining: A Practical Guide, Morgan Kaufmann Publishers, San Francisco, CA.