International Journal of Computers & Technology
Volume 4 No. 2, March-April, 2013, ISSN 2277-3061

CONCEPTUAL THREE PHASE KDD MODEL AND FINANCIAL RESEARCH

Dr. Simmi Bagga
Assistant Professor,
Sant Hira Dass Kanya Maha Vidyalaya,
Kala Sanghian, Distt. Kpt
simmibagga12@gmail.com

ABSTRACT

KDD model becomes used in financial process. Data Mining tools can be used to improve the efficiency of the professionals. The integration of Data Mining tools with the traditional financial research methods is relatively a new concept. If Data Mining tools for Financial are developed then it make the process fast, cheaper and relatively much more efficient. In this paper we have discussed the three phase model of KDD on Financial Research

General Terms

Philosophical Layer, Technique Layer and Application Layer

INTRODUCTION

Financial researcher analyzes financial statements of the organization and other documents to estimate the financial conditions and other important details regarding customers. Researcher uses some software to analyze the available financial data to forecast various trends. If the Financial researcher fails to extract hidden patterns from such business data then a lots of opportunities may lose in this competitive environment. The lacks of sufficient information is the main challenge to the organizations to survive in the competitive world. Search for useful patterns among the data are very difficult to find. With the development of information technology the challenge is now to obtain useful information from a large pool of collected data and searching involves the relevant connection from the available data set.

In the recent year there has been widespread change in the adoption and utilization of new technologies in business. These days even small business has large number of financial transaction. Due to change in business trends, its very difficult and complicated to examine or predict financial transactions by traditional or manual methods. The limitations of these manual methods can overcome by using Data Mining.

KDD helps us to discover hidden knowledge about the organizations like problem of variation and behavior of customers etc from the available data. The searched information must be specific and refined. To perform successful search data mining is the best way to obtain information. By the successful search financial researcher can predict the behavior of customers and market trends easily. Finding the useful patters in the data set are not easy to extract. Managers of the organization do not understand the technical details of all the data available in the data base hence he do not analyze the various relations available in the data.

However, many managers do not notice the importance of data and the information for data analysis. Also, most managers do not understand the relationship between the data due to the lack of technical background. For example, a financial marketing manager does not understand the relationship between the hidden patterns and the customer's portfolio. Therefore, it is important to use sophisticated tool that help companies to find out the relationship between different data.

The data mining is used to identify and find knowledge from collected data set. There are various types of data mining techniques based on supervised and unsupervised learning. Statistics plays a common role with KDD. Statistics provide a mechanism that helps to find a general pattern from all available data. According to Croxton and Cowden, "Statistics is defined as a science of collection, presentation, analysis and interpretation of numerical data." Statistics is a method of taking decisions on the basis of numerical data properly collected, organized, presented, analysed and interpreted.

Data Mining helps to making better decision regarding the use of data. Data Mining plays an important role in financial research like in forecasting stock market, understanding and managing financial risk, understanding trading futures etc. In stock market we use Data Mining to forecast market trends, identifying the best time to purchase the stocks and also the decision regarding the investment strategies. Data mining in finance follows the set of Data Mining steps such as problem understanding, data collection and cleaning, selecting the appropriate data set, building a model by applying various techniques like classification, clustering and association rules, model evaluation and deployment .The ultimate goal of data mining is the prediction of unknown patterns of business.

Data mining techniques have been used to discover hidden patterns and predict future trends and behaviors in financial markets. In stock market, Data Mining uses quality of the rules for obtaining knowledge. To discover knowledge using Data Mining the relationships obtained are too complex to understand. Data Mining task applied to a particular application is successful if results are tested properly. Financial Research provides an environment where efficiency of the methods can be tested instantly, not only by testing data. This process can be repeated daily for several months collecting quality estimates.

There are main three steps in the conceptual three phase model of KDD. In the three phase model of KDD process the whole process of KDD is divided in to following three layers:

- Philosophy Layer

International Journal of Computers & Technology
Volume 4 No. 2, March-April, 2013, ISSN 2277-3061

- Technique Layer
- Application Layer

The Philosophy Layer involves task discovery, data selection, data cleaning and data transformation. The result of this stage is pattern selection to which we performs next layer i.e. Technique Layer. The outcome of this layer is to identify the patterns and relationships between data. The last step of this model is the Application Layer. This layer mainly focuses on the usefulness and relevance of discovered knowledge for the particular domain. In this stage results have to be interpreted and evaluated to discover knowledge from the patterns. This extracted pattern helps us to take various important decisions regarding business activities.

1) PHILOSOPHICAL LAYER:

The first step of conceptual three phase model is the Philosophy Layer. The Philosophy layer mainly involves Problem description. In this step we select the target data because KDD process is never done over the entire database. We select and prepare a data set for process from the large database that data is usually refers to as a target data. Selection is an appropriate procedure for generating the target data set from the database. If data base do not have sufficient data to fulfill the requirement of the target data then we can also integrate data from the external sources. Philosophical study deals with the technology that helps in understanding of our world and also establishes the operational boundaries of knowledge.

Problem description mainly involves the formulating of overall plan of the Financial Research i.e. means how to proceed the forecasting of the financial decision in the organization. Financial researcher must have deep knowledge of the business or the organization. Its mainly involve how researcher perform the process effectively & efficiently in timely manner. He also plans about his important areas of business on which decision have to perform.

The important issues related to this layer involve the representation of knowledge, the communication of knowledge in languages and the relationship between knowledge in the mind and with the real world and also involves organization of knowledge. This is the main step of the process because data plays an important role. In short we can say this step deals to find the prior knowledge and the setting the goal of the application process.

PROBLEM DESCRIPTION

One of the important task involved in Philosophy layer is Problem Description because if problem is not understandable we cannot reach to the exact required solution of that problem. For describing problem properly information must be gathered from the various sources. Information gathering mainly involves: background information of the organization, understanding of business structure that involves the operating and control structure of the business and understand all internal controls. This information must be accurate, complete, up-to-date, and relevant to the goals. Otherwise, financial plans based on the information will be erroneous for the decision making. Other core financial task involved are managing financial risk, loan management ,uncovering market trends, planning investment strategies, identifying the best time to purchase the stocks and what stocks to purchase etc.

TECHNOLOGY CURRENTLY USED ON FINANCIAL RESEARCH

The Main technology used currently for the financial research is the Statistical Methods. The Statistical methods used in financial research are:

- Correlation
- Regression
- Forecasting and Time Series Analysis

CORRELATION

Correlation studies the relationship between two variables in which change of the value of one variable causes the change in the other variables. According to Connor," If two or more quantities vary in sympathy so that movement in the one tends to be accompanied by the corresponding movement in the other, then they are said to be correlated." The intensity of relationship between two variables can be ascertained by the quantitative value of coefficient of correlation which can be found out by computation. The degree of correlation on the basis of coefficient of correlation is determined as:

- Perfect Correlation
- Absence of correlation
- Limited degree of correlation

REGRESSION ANALYSIS

Regression is a function that maps data items into real world predictions. It is used to predicting values of continuous valued variable based on the values of other variables. The main goal of regression analysis is to determine the values of parameters for a function that cause the function to best fit a set of data observations. In linear regression, the function is a linear equation. According to Ya Lun Chou," Regression Analysis attempts to establish the nature of the relationship

International Journal of Computers & Technology
Volume 4 No. 2, March-April, 2013, ISSN 2277-3061

between the variables and there by provide mechanism for prediction or forecasting." It provides a measure of coefficient of the determination which tells us the effect of independent variable on the dependent variable. Various types of regression methods are:

- Linear Regression
- Multiple Regression
- Non Linear Regression

FORECASTING AND TIME SERIES ANALYSIS

The main objective of analyzing time series is to understand, interpret and evaluate changes in the phenomena in order to correctly anticipating the course of future events. The Study of time series helps to forecast the magnitude of variable in future so as to arrive at a desired conclusion for one's future course of action. Time series analysis helps to studying the behavior in an orderly manner. A time series is an observation of a random variable. Time series analysis provides a way for selecting a model that helps to estimate limited resources or to describe random processes. Time series models assume that observations vary with probability about function of time. A data set T called a time series is modeled is attempt to discover its main component like

- Long term trends (LT)
- Cyclic Variations (CT)
- Seasonal Variation (ST)
- Irregular movement (IT).

SHORTCOMING IN THE STATISTICAL TECHNIQUES

Statistical methods have a significant role in every fields for decision making. If data is not properly gathered and not interpreted properly then the observations drawn from data also provides the wrong conclusions. The main shortcomings in the statistical methods are as follows:

- Statistical methods do not study individual's values of the data rather it deals with the aggregates of facts.
- Statistical methods deal only with facts and figures. The qualitative aspect of a variable n not be measured using these methods. This limits its scope.
- Statistical methods are mostly used for the analysis purpose. Sometimes the result obtained from the statistical methods might lead to fallacious conclusion. If appropriate Statistical methods is used it can give the beneficial results but if misused these become harmful and lead to wrong direction
- Statistical methods can be used by the expert person properly. One who is not properly aware of the statistical methods can not made best use of these methods.
- For the best results of statistical method, data should be uniformly and homogeneously distributed. In Statistical methods heterogeneous data is not comparable only the homogenous and uniform data is used for the best results.

2) TECHNIQUE LAYER

This layer is mainly concern with the finding knowledgeable patterns from the available data set. The layer mainly involves:

- Preprocessing
- Data Mining.

Preprocessing helps to improve the quality of the data and quality of data. The quality of data directly affects the results of the Mining. After describing the problem we will extract data from the data base related to the problem. The data is collected by different techniques and methods. Usually data used for the KDD process must be cleaned. Data cleaning helps to clean data i.e. removing bad data and finding hidden correlations in the data and also identifying sources of data that are the most accurate, and determining which columns are the most appropriate for use in analysis The Data Mining is one of the main steps of KDD. Data Mining discovers patterns in a data set previously prepared in a specific way. Data Mining involves a collection of tools and techniques for finding useful patterns relating the fields of very large database. Data Mining is extracting of interesting, non trivial, impact, previously unknown and potential useful information and patterns from the data.

PREPROCESSING

The Preprocessing is the one of the essential part performed before the step of Data Mining. Data in the original form cannot be directly used in the Data Mining process as data is gathered from the different format, symbols or in different file formats. The Preprocessing mainly data cleaning that helps to fill the missing values if any also we smooth noisy data and

International Journal of Computers & Technology
Volume 4 No. 2, March-April, 2013, ISSN 2277-3061

also resolve inconsistencies in the data. Preprocessing helps to improve the quality of the data and quality of data directly affects the results of the Mining. Data cleaning is cleans the data i.e. removing bad data and finding hidden correlations in the data also identifying sources of data that are the most accurate, and determining which columns are the most appropriate for use in analysis. After cleaning data, we merge data that was collected from the different forms. That data may have different formats. Data selection analyses the collected data from the different sources and decide from the collected according to some algorithms. In Data Transformation data is transformed in proper format that can be directly used for the mining process. Data transformation involves Smoothing to remove the noise from the data. Aggregation, Generalization and Normalization reduces the data and then transform this in the proper format.

After performing the Preprocessing the data is ready for the step of Data Mining and now we have the Data that do not contain any incomplete patterens, dirts or any incompatibilities.

DATA MINING

Data Mining is the core process that takes input cleaned & transformed data and searches patterns using some algorithms and then results patterns and relationships. There are various types of data Mining algorithms that can be used during the process are:

- Classification
- Regression
- Clustering
- Association Rules

Classification is commonly used data mining technique, which uses a set of predefined groups or classes to develop a model that can classify the data. Classification method uses the mathematical models like decision trees, neural networks etc. Clustering techniques are used for combining group that are similar to each other. It is used to find groups or clusters in the data that are similar in some context. The best known method for Clustering is K-means Methods. It is simple and iterative method works around one artificial point which represent the average location of the cluster is called Centroid. This algorithm takes an input a number of clusters that is the k form. Means is an average location of all the members of a cluster. Association rule searches for the interesting relationship among the data set.

These different methods of Data Mining help us to discover hidden patterns from the available data. This layer helps us to discover patterns from the data set from which noise has been previously eliminated and has been transformed in such a way to enable the pattern discovery process. In common we can say that in the KDD process, Data Mining discovers interesting patterns according to the user's requirement.

3) APPLICATION LAYER

The third and the last step of Conceptual Three phase model of KDD is concluding and reporting. In this step, conclusions are performed on the basis of evidences, Provide reports on the basis of conclusion to determining the truthfulness & fairness of financial statements and communicating the reports to the appropriate authority of the organization. In the three phase model of KDD the last step is Post Processing, the discovered knowledge is visualizes to understand and interpret for humans. That means in Post Processing, the final report of discovered knowledge is presented.

SUMMARY

Conceptual Three Phase KDD model provides us more accurate solution than the traditional model. Traditional Statistical methods have a significant role in every fields for decision making. If data is not properly gathered and not interpreted properly then the observations drawn from data also provides the wrong conclusions. In this paper we discussed three layers of KDD related to financial research field. Conceptual Three phase KDD provides better solution in financial research field.

REFERENCES

Han Jiawei and Kamber, Micheline. "Data Mining: Concepts and Techniques". 2001. Morgon Kaufmann. Sanfransico, CA.

Lee, Sang Jun & Keng, Siau (2001), A Review of Data Mining Techniques, Industrial Management & Data Systems: Volume 101

1. Banerjee A, Merugu S, Dhillon I, Ghosh J (2005) Clustering with Bregman divergences. J Mach Learn
2. Hu T, Sung SY (2006) Finding centroid clusterings with entropy-based criteria. Knowl Inf Syst
3. Elder, John F., IV and Daryl Pregibon, (1996), Advances in Knowledge Discovery & Data Mining, "A Statistical Perspective on KDD.
4. Singh G.N,Bagga Simmi(2011) Three Phase Iterative Model of KDD, International Journal of Information Technology and Knowledge Management, Volume 4, No. 2, pp.695-697.Singh G.N, Bagga Simmi (2011), Clustering Method for categorical and Numeric Data sets, Global Journal in Computer Science.

5 Singh G.N, Bagga Simmi(2012),Applications of Data Mining, International Journal for Science and Emerging Technologies with latest trends.