A Study of Data Mining Techniques And Its Applications

M.Porkizhi
Ph.D., Research Scholar, Rathinam College of Arts & Science, Coimbatore, TamilNadu, India.

Abstract- *Data mining is the computational process of discovering patterns in large data sets. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. The paper discusses few of the data mining techniques, algorithms and some of the organizations which have adapted data mining technology to improve their businesses and found excellent results and focuses on presenting the applications of data mining in the business environment.*

Keywords- Data mining Techniques, Data mining algorithms, Business, Data mining applications.

I. INTRODUCTION

Data mining is the computational process of discovering patterns in large data sets. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. Data mining is most often applied to extraction of useful knowledge from business data however it is also useful in some scientific applications where this more empirical approach complements traditional data analysis. The emergence is due to the growth in data warehouses and the realization that this mass of operational data has the potential to be exploited as an extension of Business Intelligence. In Earlier data Mining used similar manual approaches to review data and provide business projections for many years. Changes in data mining techniques, however, have enabled organizations to collect, analyze, and access data in new ways.

II. OVERVIEW OF DATA MINING

Extraction of information is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Once all these processes are over, we would be able to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration, etc.

The development of Information Technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis. Data mining is a logical process that is used to search through large amount of data in order to find useful data. The unifying goal of the KDD process is to extract knowledge from data in the context of large databases.

The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses.

Three steps involved are

- Exploration
- Pattern identification
- Deployment

Exploration: In the first step of data exploration data is cleaned and transformed into another form, and important variables and then nature of data based on the problem are determined

Pattern Identification: Once data is explored, refined and defined for the specific variables the second step is to form pattern identification. Identify and choose the patterns which make the best prediction

Deployment: Patterns are deployed for desired outcome.

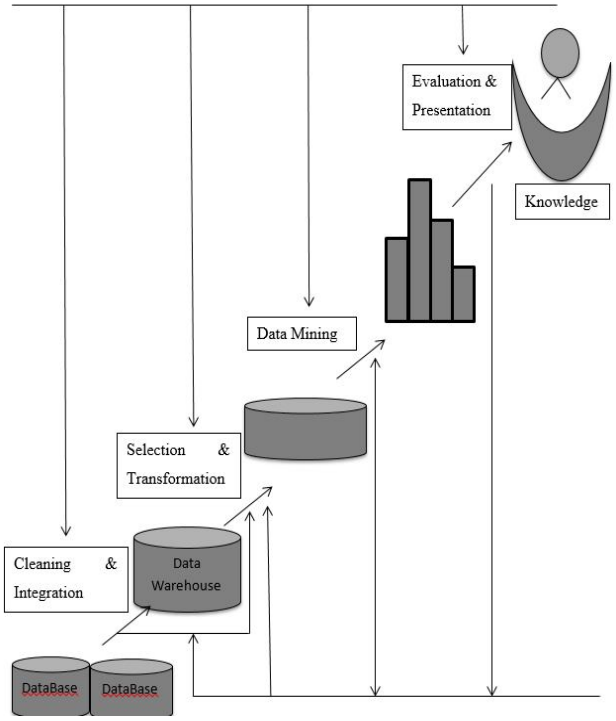

Figure: Knowledge discovery process

II. DATA MINING ALGORITHMS AND TECHNIQUES

A data mining algorithm is a well-defined procedure that takes data as input and produces as output. Output may be the models or patterns. 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 databases.

A. Classification

Classification techniques in data mining are capable of processing a large amount of data. It can be used to predict categorical class labels and classifies data based on training set and class labels and it can be used for classifying newly available data.The term could cover any context in which some decision or forecast is made on the basis of presently available information. Classification procedures recognized method for repeatedly making such decisions in new situations.

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 credit-risk 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
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

B. Clustering

Clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities. Regarding to data mining, this methodology partitions the data implementing a specific join algorithm, most suitable for the desired information analysis. This clustering analysis allows an object not to be part of a cluster, or strictly belong to it, calling this type of grouping hard partitioning. In the other hand, soft partitioning states that every object belongs to a cluster in a determined degree. More specific divisions can be possible to create like objects belonging to multiple clusters, to force an object to participate in only one cluster or even construct hierarchical trees on group relationships.

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 preprocessing 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

- Centroid-based Methods
- Connectivity-based methods
- Density based methods

- Grid-based methods
- Model-based methods

C. Predication

A regression task begins with a data set in which the target values are known. 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
- Nonlinear Regression
- Multivariate Linear Regression
- Multivariate Nonlinear Regression

D. Association rule

Association rule mining is a procedure which is meant to find frequent patterns, correlations, associations, or causal structures from data sets found in various kinds of databases such as relational databases, transactional databases, and other forms of data repositories. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

Types of association rule

- Multilevel association rule
- Multidimensional association rule
- Weighted association rules
- Quantitative association rule
- Multi-relational association rules
- Fuzzy association rules

E. Neural networks

Neural networks have the ability to adapt to changing input so the network produces the best possible result without the need to redesign the output criteria. The concept of neural networks is rapidly increasing in popularity in the area of developing trading systems. Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

Types of neural networks

- Feed Forward Neural Network
- Back Propagation

III. DATA MINING APPLICATIONS

Data mining is highly useful in the following domains: Market Analysis, Management, Corporate Analysis & Risk Management, and Fraud Detection. Apart from these, data mining can also be used in the areas of production control, customer retention, science exploration, sports, astrology, and Internet Web Surf-Aid. 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. Data mining is defined as a business process for exploring large

amounts of data to discover meaningful patterns and rules. Companies can apply data mining in order to improve their business and gain advantages over the competitors. The most important business areas that successfully apply data mining, presented in Figure. below, are:

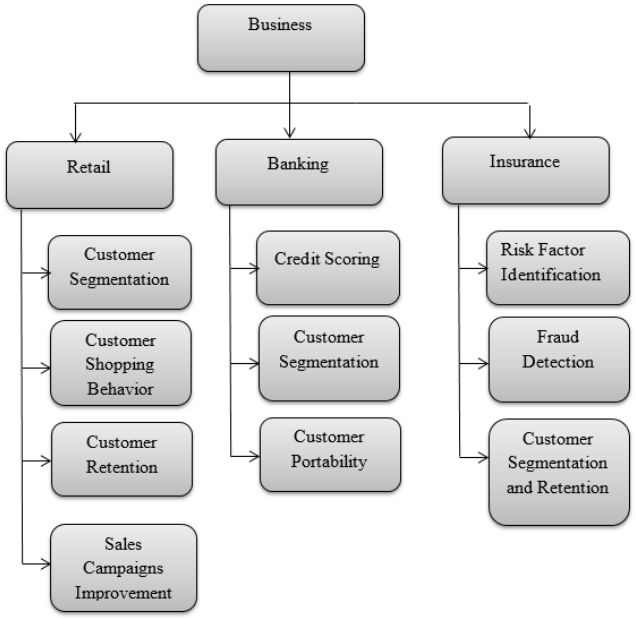

Figure: Business areas that successfully apply data mining

1. Retail

Retail data mining can help identify customer buying behaviors, discover customer shopping patterns and trends, improve the quality of customer service, achieve better customer retention and satisfaction, enhance goods consumption ratios, design more effective goods transportation and distribution policies, and reduce the cost of business.

- Data mining techniques have many applications in the retail industry, including the following:
- Customer segmentation: identify customer groups and associate each customer to the proper group;
- Establish customer shopping behavior: identify customer buying patterns and determine what products the customer is likely to buy next;
- Customer retention: identify customer shopping patterns and adjust the product portfolio, the pricing and the promotions offered;
- Analyze sales campaigns: predict the effectiveness of a sales campaign based on the certain factors, like the discounts offered or the advertisements used.

Retail industry offers a wide area of applications for data mining due to the large amounts of data available for companies.

2. Banking

There are various areas in which data mining can be used in financial sectors like customer segmentation and profitability, credit analysis, predicting payment default, marketing, fraudulenttransactions, ranking investments, optimizing stock portfolios, cash management and forecasting operations, high risk loan applicants, most profitable Credit Card Customers and Cross Selling. The main examples of applications of the data mining techniques in the banking industry are the following:

- Credit scoring: distinguish the factors, like customer payment history, that can have a higher or lower influence over loan payment;
- Customer segmentation: establish customer groups and include each new customer in the right group;
- Customer retention: identify customer shopping patterns and adjust the product portfolio, the pricing and the promotions offered;
- Predict customer profitability: identify patterns based on various factors, like products used by a customer, in order to predict the profitability of the customer.

The information systems for the banking industry contain large amounts of operational and historical data, being a fitted application area for data mining.

3. Insurance.

Data mining can help insurance firms in business practices such as: acquiring new customers, retaining existing customers, performing sophisticated classification or correlation between policy designing and policy selection. In insurance the data mining techniques have the following applications:

- Risk factor identification: analyze the factors, like customer claims history or behavior patterns, that can have a stronger or weaker influence over the insured's level of risk;
- Fraud detection: establish patterns of fraud and analyze the factors that indicate a high probability of fraud for a claim;
- Customer segmentation and retention: establish customer groups and include each new customer to the appropriate group and identify discounts and packages that would increase customer loyalty.

Data mining techniques have many applications in the insurance business and can improve it by analyzing the large amounts of data available for companies.

IV. CONCLUSION

Data mining has importance regarding finding the patterns, forecasting, discovery of knowledge etc., in different business domains. Our current society needs data mining for improving many domains of human life. Business areas like retail, banking and insurance can use data mining methods to improve customer experiences, make optimal decisions, strengthen their market position and achieve competitive advantage. 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. Our current society needs data mining for improving many domains of human life. Business areas like retail, banking and insurance can use data mining methods to improve customer experiences, make optimal decisions, strengthen their market position and achieve competitive advantage.

REFERENCES

[1] Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman, 2nd ed.
[2] Dr. Gary Parker, vol 7, 2004, Data Mining: Modules in emerging fields, CD-ROM.
[3] Crisp-DM 1.0 Step by step Data Mining guide from http://www.crisp-dm.org/CRISPWP-0800.pdf.
[4] Gordon S. Linoff and Michael J. A. Berry, Data Mining Techniques: for Marketing, Sales and Customer Relationship Management. Third Edition, Wiley Publishing , USA, 2011.
[5] Hsu, J. 2002. Data Mining Trends and Developments: The Key Data Mining Technologies and Applications for the 21st Century, The Proceedings of the 19th Annual Conference for Information Systems Educators (ISECON 2002), ISSN: 1542-7382.
[6] Customer Successes in your industry from http://www.spss.com/success/?source=homepage&hpzone=nav_bar.
[7] https://www.allbusiness.com/Technology/computer-software-data-management/633425-1.html last retrieved on 15th Aug 2010.
[8] H. Bhavsar, A. Ganatra, "A Comparative Study of Training Algorithms for Supervised Machine Learning", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231 -2307, Volume-2, Issue-4, September 2012.
[9] http://www.kdnuggets.com/.
[10] Gordon S. Linoff and Michael J. A. Berry, Data Mining Techniques: for Marketing, Sales and Customer Relationship Management. Third Edition, Wiley Publishing, USA, 2011.
[11] Jiawei Han, Micheline Kamber and Jian Pei, Data Mining: Concepts and Techniques. Third Edition, Morgan Kaufmann Publishing, USA, 2011
[12] Vikas Jayasree and Rethnamoney Vijayalakshmi Siva Balan, A Review on Data Mining in Banking Sector, American Journal of Applied Sciences, Vol. 10, Issue 10, 2013, ISSN 1554- 3641, pp. 1160-1165.
[13] A. B. Devale and Dr. R. V. Kulkarni, Applications of data mining techniques in life insurance, International Journal of Data Mining & Knowledge Management Process, Vol.2, Issue 4, July 2012, ISSN 2230-9608, pp. 31-40.