International Journal of Scientific Research in Computer Science and Engineering **Review Paper**
Volume-5, Issue-1, pp.19-23, February (2017) **E-ISSN:** 2320-7639

Data Mining: A Comparative Study of its Various Techniques and its Process

Marie Fernandes

Department of Computer Science, Indore Indira School of Career Studies, Indore, India

Available online at: www.isroset.org

Received 24th Dec 2016, Revised 8th Jan 2017, Accepted 03th Feb 2017, Online 28th Feb 2017

***Abstract* - Data Mining also called as Information Mining or certainty finding is the term which is utilized for removing or finding helpful data from the information that are available in vast databases. It likewise investigates covered up or prescient examples of content that can be said as predictive patterns of text, from databases. This term showed up in 1990's. It is a procedure that examines or analyses information from alternate points of view and compresses it into helpful data. This data can then be utilized for different business purposes by various undertakings. Information mining from that point forward has turned into an essential piece of Knowledge Discovery in Databases (KDD), data Digging, data fishing, and Data Collecting as appropriately termed as Data Dredging, Data Fishing, and Information Harvesting. It turns a large collection of data into knowledge that can fulfill current global challenge because computerization has lead to explosively growing, widely available and gigantic body of data floating through WWW. Data mining methods are expected to change this information into sorted out learning. Keeping in mind the end goal to do as such; capable and flexible tools are required which would reveal important data from the huge measures of information. This need has prompted to numerous strategies, for example, Classical Techniques which incorporates Statistics which provides measurements, Neighborhoods and Clustering which works through grouping and the Cutting edge Procedures incorporates Trees, Networks and Rules. The dominant part of information mining methods manages distinctive information sorts. The scope, purpose and motivation behind this paper is to do a relative investigation of the different procedures accessible in information mining with their preferences, burdens and the field where they can be properly utilized. This paper presents overview of data mining, the different strategies of data or information mining.**

***Keywords* - Data mining, Data Dredging, Statistics, Nearest Neighbor, Decision Trees and Neural Networks.**

I. INTRODUCTION

Data Mining is getting to be a rising exploration point discovering applications in many fields like engineering, Medicine, Business, Education and Science. Data dredging is the use of data mining to uncover patterns in data that can be presented as statistically significant. A lot of information has prompt to expansive databases thus propelled database frameworks, data warehousing and so data mining is additionally progressing. This stockpiling or vault of gigantic volumes of information and has posed like a challenging and testing task in breaking down these stored information. The proficient and viable examinations of information from the enormous volumes of data that have been amassed needs viable data mining strategies. Data mining procedure is concerned with the investigation of data using some product methods or software techniques for finding covered up and unforeseen patterns and connections in sets of information. Data mining concentrates on finding the data that is covered up and is unexpected. It is extraction of new information from expansive databases. Data Mining (DM) is an essential part in the process of Knowledge Discovery in Databases. As there are different information present and in addition many concealed patterns of data are there in the databases thus, it gets to be distinctly important to know the different strategies that can be utilized for data mining.

The rest of the paper is organized as follow: Section II mentions the literature reviews in the form of related work done in data mining. Section III gives details of various data mining techniques. Section IV explains about the classical techniques of data mining with advantages and disadvantages. Section V explains the next generation techniques of data mining with advantages and drawbacks. Section VI deals with the methodology used, Section VII is of Results and Discussions, Conclusion and Future Scope is shown in Section VIII and the references are mentioned in the last section.

II. RELATED WORK

This section gives the summary of the various technical articles and review work carried out in the field of data

mining and its techniques. In [1] Lee, S and Siau, K, has analyzed that some techniques for solving data mining tasks and concluded that the statistical techniques are used to discover patterns and build predictive models, the neural networks are powerful mathematical models suitable for almost all data mining and the Decision trees can naturally handle all types of variables, even with missing values.

In [2] Berson, Alex, evaluated the data sets that are present, different tools that are needed for business data processing and analysis.

In [3] S Mahajan, tries to analyzed the concepts of data mining that could be used in the various fields as part of data collection, data extraction.

In[4] Jain, A.K., Murty, M.N., and Flynn, P.J, analyzed several applications where decision making and exploratory pattern investigation can be performed on expansive informational collections. It concludes that data abstraction that is simple and compact representation of data can be done in decision making rather than using a entire data set.

In [5] P Berkhin, attempts to analyzed that clustering divides the data into groups of similar objects. It disregards details for providing data simplification. It also provides concise summaries of the data.

In [6] Jaskaranjit Kaur and Gurpreet Kaur, described the processes of selected techniques from the data mining point of view. The result of the research is that new research solutions are needed for the problem of categorical data mining techniques for future work.

In [7] J.Sheela Jasmine, attempts to analyzed, neural networks to be a promising data mining tool because they have proven their predictive power through comparison with other statistical techniques using real data sets but due to design problems neural systems need further research before they are widely accepted in industry.

In [8], C Kaur, P Kapoor, M Bala , attempts to analyze the efficiency of neural network algorithms and their effectiveness to produce result as they have self-adjusting nature.

In [9] P Gaur, concludes that neural network is very suitable for solving the problems of data mining because of its characteristics of good robustness, self-organizing adaptive, parallel processing, distributed storage and high degree of fault tolerance.

III. DATA MINING TECHNIQUES

Data Mining is done to prepare the data and distinguish the patterns in the data so that a choice or a judgment can be made. Different data mining methods appeared on the grounds that the span of the data is turning out to be much bigger and this data is more shifted and broad in nature and substance. The business-driven needs additionally have changed basic data recovery mechanisms. As it is impractical for people to prepare huge data to discover significant data opportune, so machine learning tools and advancements are utilized. It being a critical piece of KDD, so knowing the different methods and types of data extraction likewise gets to be distinctly imperative. Knowledge discovery is a procedure that concentrates on certain, possibly helpful or beforehand obscure information from the data. The knowledge discovery process is described in figure -1. The diverse systems utilized for data mining are Classification, Clustering, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method. A large number of modeling techniques are labeled "data mining" techniques [1]. The following section gives a short survey of selected number of these techniques.

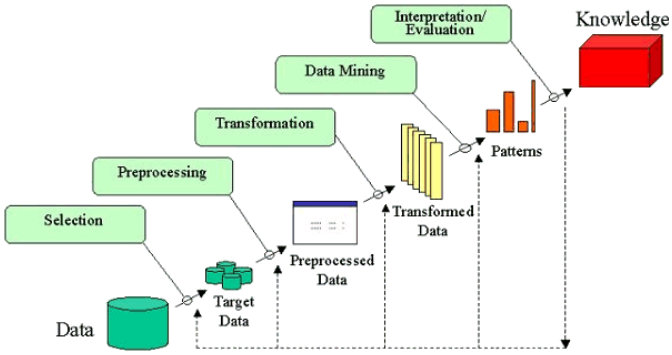

Figure-1 Steps of KDD

IV. CLASSICAL TECHNIQUES

A. Statistical Techniques

Statistics is the traditional field that deals with the collection, quantification, interpretation, analysis and drawing conclusions from data. Data mining is an interdisciplinary field that mines data collectively with the help from computer sciences dealing with data base, machine learning, artificial intelligence, visualization and graphical models, statistics and engineering dealing with pattern recognition, neural networks. Thus, Statistics is a branch of mathematics concerning the collection and the description of data [2].

Presently data mining and statistics has been characterized autonomously however "mining data" for patterns and predictions is precisely what is done through statistics. Some of the procedures that are grouped under data mining, for example, CHAID and CART have been the result of the statistical profession, probability are the foundation on which both data mining and statistics are fabricated. The strategies are utilized in same places for similar sorts of issues

(prediction, classification discovery). The advantages or benefits of Statistical Technique or the Factual Method is that statistics introduces a high level perspective of the database that gives some helpful data to such an extent that it doesn't require each record to be comprehended in detail. For instance, the histogram can rapidly indicate essential data about the database, which is the most incessant or frequent. The disadvantages or we can say the inconveniences of this technique is that for vast piece of data; statistics for the most part is concerned with outlining data and thus numbering issue occurs because of this summarization. Statistical Techniques can't be helpful without specific presumptions about data. The consequences of using the Factual Strategy is that statistics is utilized as a part of the detailing of imperative data from which individuals might have the capacity to settle on helpful choices and to make important decisions. A trivial outcome acquired by a basic strategy is known as a modern method of forecast more appropriately called a sophisticated technique of prediction. It is so called as Naïve Bayes prediction.

B. Nearest Neighbor

Clustering and the Closest Neighbor prediction technique and strategy are a part of the most seasoned strategies utilized as a part of data mining. Many individuals have the reasoning that in clustering records are assembled or grouped together. Nearest neighbor is a prediction technique or a forecast method which is to some degree like clustering yet its significance is that, to foresee or predict estimated value of one record you need to search for records with comparable indicator values in the database(historical) and utilize this prediction value from the record that is "closest" to the unclassified record. A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset. Sometimes its called the k-nearest neighbor technique [3].

The nearest neighbor prediction calculation expresses that "Objects that are "close" to each other will have comparative prediction values". Along these lines, if the estimation of one of the objects is known then you can anticipate it for its closest or nearest neighbors.

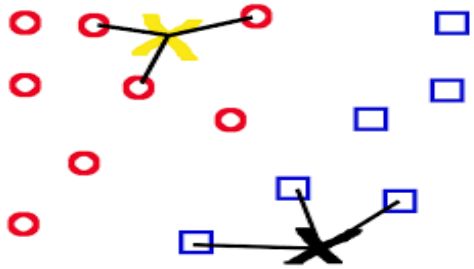

Figure-2 Nearest Neighbor

C. Clustering

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters) [4].Clustering is a technique or is the strategy in which the comparable or like records are assembled together. This is normally done to give a high level perspective of what is happening in the database to the end user or client. Clustering sometimes means segmentation. The nearest neighbor calculation is to some degree refinement of clustering with deference that they both utilize distance in some feature spaces to make or create structure in the information or predictions. The nearest neighbor procedure or calculation is a refinement since some part of the calculation is a method for deciding consequently the weighting of the significance of the predictors and the method for measuring distance inside the feature space. Clustering is one of the uncommon instances of this where the significance of every predictor is thought to be equal or practically comparable. Clustering as applied to data mining applications encounters three additional complications: a). large databases, b).object with many attributes, and c).attributes of different data types [5].

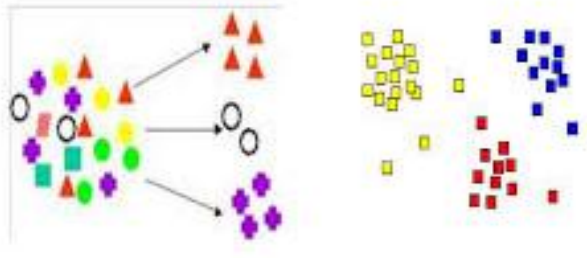

Figure -3 Clustering

V. NEXT GENERATION TECHNIQUES

A. Decision Trees

A decision tree is a predictive model that can be seen as a tree. Every branch of the tree represents a classification question and the leaves of the tree represent partitions of the dataset and their arrangement particularly. Decision trees are utilized for characterization and in addition for estimation tasks. Decision trees can be utilized to assess or discover or to anticipate the result for new sample data. The Decision tree technique can likewise be utilized as a part of investigating the dataset and business issue and has been utilized for preprocessing information for other prediction algorithms.

The Benefits of Decision Trees method is that the Decision trees can normally deal with every type of variables, even which has missing values. The advantageous favorable feature of the Decision tree model is its straightforward nature. The decision tree explicitly specify all possible alternatives and traces each alternative to its conclusion in a single view, allowing for easy comparison among the various alternatives. It uses separate nodes to denote user defined

Figure-3 Clustering

decisions, uncertainties, and end of process which then leads to clarity and transparency to the decision-making process. The Drawbacks of Decision Trees technique is that it doesn't support the extensive number of analytic tests. Decision trees do not specify and impose special restrictions or requirements on the data preparation procedures, that is Decision trees require relatively little effort from users for data preparation. It cannot match the performance of linear regression and consequently, Non-linear connections between parameters don't influence tree execution.

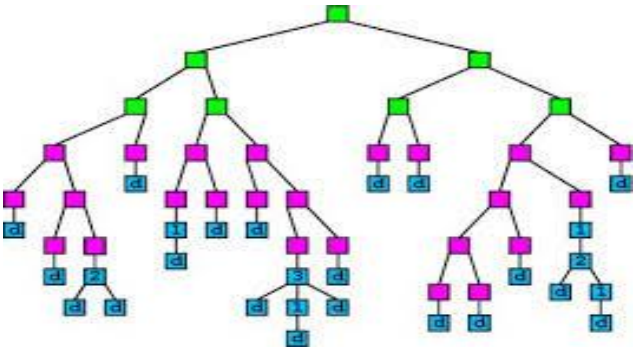

Figure -4 Decision Trees

B. Neural Network Technique

Neural Network is the "Artificial Neural Network". Being artificial in the sense that they are computer programs which implement sophisticated pattern detection and machine learning algorithms on a computer to constuct predictive models from large historical databases. "Artificial neural networks derive their name from their historical development which started off with the previous proposition that machines could be made to think if scientists found ways to mimic the structure and functioning of the human brain on the computer"[6]. "Using neural networks as a tool, data warehousing firms are harvesting information from datasets in the process known as data mining"[7]. There are two main structures of consequence in the neural network: The node loosely corresponds to the neuron in the human brain and the link loosely corresponds to the connections between neurons in the human brain. Along these lines, a Neural Network model is framed as a gathering or collection of interconnected neurons. The course of action of neurons and their interconnections is known as the design of the Network. These interconnections can be a solitary layer or numerous layer and can be unidirectional or bi-directional. A neural network is given a set of inputs and is used to predict one or more outputs. It can be said that Neural network in data mining plays vital role for classification of the complex data[8].Thus, the neural networks are most powerful mathematical models that is suitable for most of the data mining tasks, and the special emphasis lays on classification and estimation problems. Neural networks can be used for outlier analysis, clustering, prediction work and feature extraction. It can even be used in complex classification situations.

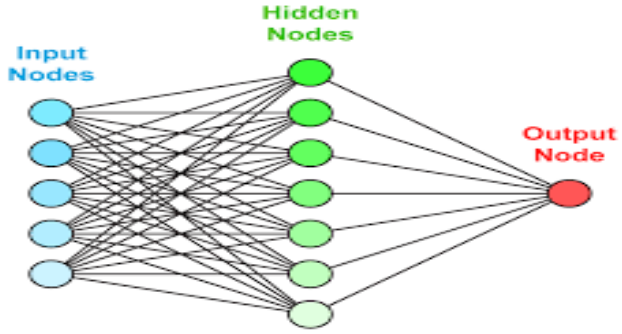

Figure -5 Neural Networks

Neural Networks is capable of producing an randomly complex relationship between inputs and outputs. Neural Networks ought to have the capacity to break down and also arrange information utilizing the inherent elements with no outside support or direction. Neural Networks of various kinds are mostly and can be used for clustering and prototype creation. Neural networks do not work in proper way when there are many hundreds and thousands of input features. Futher more , "neural computing refers to a pattern recognition methodology for machine learning. The resulting model from neural"[9].They do not provide acceptable performance for complex problems. It is difficult to understand the model that neural networks have built upon and how the raw data affects the output predictive result. The Neural Networks can be released on the data straight without having to rearrange or modify the data very much. It is that they are automated to a degree where the user does not need to know that much about predictive modeling or how they work or even the database in order to use them.

VI. METHODOLOGY

Due to shortage of time, this research is based on secondary data sources which comprises of data collected from journals, text books, articles, online and offline mediums. As it becomes necessary to extract hidden information, it is thus, necessary to know the data mining techniques that can be applied on various datasets. Henceforth, the methodology of the study or analyses of data mining is theoretical and has been referred from different literature to reveal the various techniques which can be helpful for extraction of information and hidden patterns.

VII. RESULTS AND DISCUSSION

The consequence or result of the review or the investigation of various data mining techniques and calculations is that there are many apparatuses for dissecting information. Each strategy has a few benefits and negative marks. The decision tree can deal with both persistent and discrete information, it gives great outcomes with the small size tree yet the demerit

is that a little change in information can change the decision tree totally. The nearest neighbor technique of data mining has the benefit of better execution with missing information and it is anything but difficult to actualize and investigate however it requires high count multifaceted nature. The benefit of neural system is that they can group design on which they have not been prepared but rather they have poor interpretation ways.

VIII. CONCLUSION AND FUTURE SCOPE

The general objective of the data mining procedure is to isolate the data from a huge informational collection and change it into a justifiable shape for further utilize. This paper puts a push to portray the procedures of chosen methods from the perspective of data mining and shows the ability of data mining and its distinctive systems. The review presumes that all data mining strategies attempt to fulfill their objectives in immaculate way; however every strategy takes after and has its own attributes, determinations that demonstrate their exactness, inclination and capability. Data mining is consistently substantiating itself as an important device in numerous ranges, however by some parts of the data mining methods are commonly far superior appropriate to some issue zones than to others, subsequently, it is prescribe in many organizations to utilize data mining at any rate to help administrators to settle on right choices as indicated by the data given by data mining. There is not as such any procedure that is and can be totally successful for information mining in considering exactness, constraints, division, outline, forecast, order, application, location and reliance. It is thus, suggested that these methods ought to be utilized as in collaboration with each other.

The current level of the study is empirical research. In terms of future scope, a variety of data mining techniques can be used by researchers to evaluate and extract hidden patterns. In this paper we briefly reviewed the various data mining techniques. This review would be helpful to researchers to focus on the various issues of data mining and the techniques. In future course, we will review the various classification algorithms and tools used in data mining and can put focus on the hot and promising areas of data mining.

REFERENCES

[1] Lee, S and Siau, K. “A review of data mining techniques”, Journal of Industrial Management & Data Systems, Volume-101, Issue-01, pp (41-46), 2001.

[2] Berson, A, Smith, S, and Thearling, K., “Building Data Mining Applications for CRM”, McGraw-Hill Professional, First(1st) edition, 1999.

[3] S Mahajan, "Convergence of IT and Data Mining with other technologies ", International Journal of Scientific Research in Computer Science and Engineering, Volume-01, Issue-04, pp (31-37), Aug 2013

[4] Jain, A.K., Murty, M.N., and Flynn, P.J. “Data Clustering: A Review, Journal ACM Computing Surveys (CSUR)”, Volume-31, Issue-0 3, pp (264-323), 1999.

[5] Jaskaranjit Kaur and Gurpreet Kaur , "Clustering Algorithms in Data Mining: A Comprehensive Study", International Journal of Computer Sciences and Engineering, Volume-03, Issue-07, Page No (57-61), Jul -2015.

[6] B Khalid, N Abdelwahab. “A Comparative Study of Various Data Mining Techniques: Statistics, Decision Trees and Neural Networks”, International Journal of Computer Applications Technology and Research, Volume-5, Issue-03, pp (172 – 175), 2016.

[7] J.Sheela Jasmine, "Application of Fuzzy Logic in Neural Network Using Data Mining Techniques: A Survey", International Journal of Computer Sciences and Engineering, Volume-04, Issue-04, Page No (333-341), Apr -2016.

[8] C Kaur, P Kapoor, M Bala , “Role of Neural network in data mining”, International Journal for Science and Emerging Technologies with Latest Trends, Volume – 02, Issue -01, pp (20-28), 2012

[9] P Gaur, “Neural Networks in Data mining”, International Journal of Electronics and Computer Science Engineering”, Volume -01, Issue -03, pp (1449-1453), 2012

AUTHORS PROFILE

Marie Fernandes has received M.Sc Degree in Electronics and Communication from Devi Ahilya University, Indore (M.P.) in 2005. Presently, she is pursuing MCA Degree from IGNOU University, Delhi. Her area of interest is Operating Systems, Digital Electronics, Operating System, Computer Networking. She has worked as a technical trainer with Jetking Infotrain Pvt. Ltd., till year 2010 in Indore (M.P.) and also worked as Quality Auditor Executive with Jetking Infotrain Pvt. Ltd. for the year 2011 and 2012. She is presently working as Assistant Professor at Indore Indira School of Career Studies,Indore(M.P.)since 2012 till date. Email id-fernandes.marie@gmail.com

