ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

Data Mining: A Study on Various Applications and Techniques

Ashish B. Patel[1]

Research Scholar, Faculty of Computer Science, Pacific Academy of Higher Education and Research University, Udaipur, Rajasthan, India[1]

ABSTRACT: Data Mining is the process of discovering hidden predictive information from large amount of data that can include databases, data warehouses, web, data that are streamed into the system dynamically or other information repositories. Data mining is an essential step in the knowledge discovery in databases process that produces useful patterns or models from data. With increasing growth of data in every application, data mining meets possible future need for effective, scalable, and flexible data analysis in our society. Data mining can be considered as the natural evolution in information technology and joining together several related disciplines and application domains. This paper studies about data mining, knowledge discovery process, different kind of data & patterns can be mined, data mining system classification, applications and techniques of data mining.

KEYWORDS: Data Mining, Knowledge Discovery in Databases (KDD), Information Technology.

I. INTRODUCTION

Data mining is the analysis step of the Knowledge Discovery in Databases process and an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. Apart from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, complexity considerations, post-processing of discovered structures, visualization, and online updating [2] [3] [4]. 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 usage.
Data mining is an iterative and interactive process of discovering novel, valid, useful, comprehensive and understandable patterns and models in massive data sources [5].

The data mining methods can be used in creating new hypotheses to test against the larger data populations. Data Mining is a process of extracting previously unknown, valid, potentially useful and hidden patterns from large data sets [1].

Data Mining is the process of Discovering patterns and knowledge from large amount of data. The data sources can include databases, data warehouses, the web, other information repositories, or data that are streamed into the system dynamically. Data mining is an essential step in the knowledge discovery in databases process that produces useful patterns or models from data. The terms Knowledge Discovery Data (KDD) and data mining are different. The KDD refers to the overall process of discovering useful knowledge from data while data mining refers to discover new patterns from a wealth of data in databases by focusing on the algorithms to extract useful knowledge depicted in the following figure 1 [5].

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

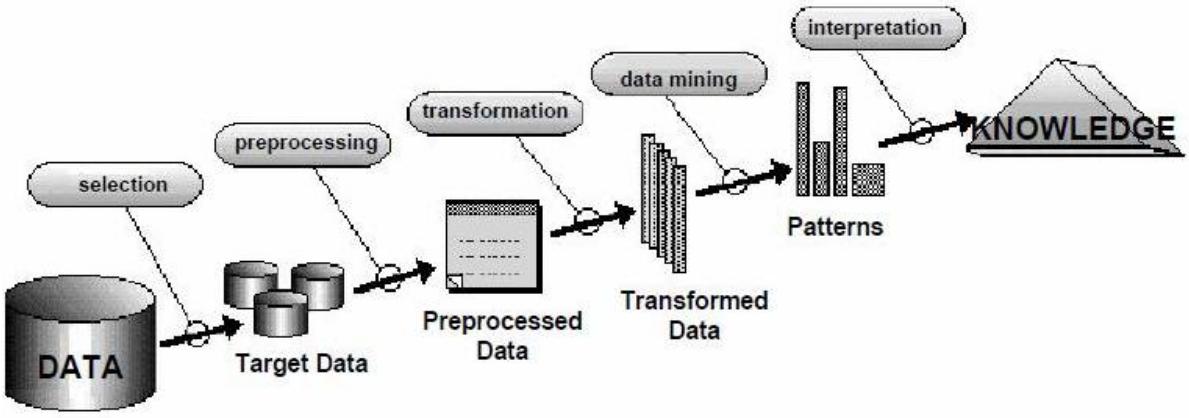

Figure 1: Data Mining and the KDD Process [11].

As seen in figure 1, initially, data are collected from multiple sources and hence requires cleaning and integration of the collected data in order to remove noise and data inconsistency. After cleaning and integration of collected data, it is necessary to select the data relevant to the analysis task from the database. The selected data needs to be transformed into appropriate form so as to perform data mining. It is an essential process where intelligent methods are applied to extract data patterns. In the post processing step of data mining, it is necessary to interpret the patterns into knowledge by removing redundant or irrelevant patterns. The useful patterns shall be translated into terms that humans understand. This requires presentation of knowledge, where visualization and knowledge representation techniques are used to present mined knowledge to the human.

There are different types of data that can be used for mining purpose where data mining can be applied to any kind of data sets as long as the data are meaningful for target application. The most basic forms for data mining are Database data in DBMS and RDBMS, Data Warehouses, Transactional data, Multimedia Data, Spatial databases which includes say maps, Time-Series Database which includes time-series and biological data, historical records, stock exchange data etc., WWW includes information repository by the internet, Data Streams, includes video surveillance and sensor data which are continuously transmitted, and Engineering Design Data includes design of buildings, system components, and integrated circuits [5].

Data Mining can be used widely in the areas like Financial data analysis, Stock market analysis, Biological data analysis, Weblog analysis, DNA mining, Banking, Retail industry, Telecommunication industry, Scientific applications, Intrusion detection, Different applications in Transportation, Medicine, Health Care, Multimedia, Insurance sector and so on.

II. RELATED WORK

CATEGORIES OF DATA MINING

Descriptive: The descriptive model identifies the patterns or relationships in data and explores the properties of the data examined. The few such models or techniques are Clustering, Summarization, Association rule, Sequence discovery and so on. The clustering is similar to classification except that the groups are not predefined, but are defined by the data alone. It is also referred to as unsupervised learning or segmentation. It is the partitioning or segmentation of the data into groups or clusters. The clusters are defined by studying the behaviour of the data by the domain experts. The term segmentation is used in very specific context; it is a process of partitioning of database into disjoint grouping of similar tuples. Summarization is the technique of presenting the summarized information from the data. The

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

association rule finds the association between the different attributes. The association rule mining is a two step process. First is finding all frequent item sets, and second is generating strong association rules from the frequent item sets. The sequence discovery is a process of finding the sequence patterns in data. This sequence can be used to understand the trend [5].

Predictive: The predictive model makes prediction about unknown data values by using the known values. Ex. Classification, Regression, Time series analysis, Prediction etc. Many of the data mining applications are aimed to predict the future state of the data. Prediction is the process of analyzing the current and past states of the attribute and prediction of its future state. Classification is a technique of mapping the target data to the predefined groups or classes, this is a supervise learning because the classes are predefined before the examination of the target data. The regression involves the learning of function that maps data item to real valued prediction variable. In the time series analysis the value of an attribute is examined as it varies over time. In time series analysis the distance measures are used to determine the similarity between different time series, the structure of the line is examined to determine its behaviour and the historical time series plot is used to predict future values of the variable [5].

III. LITERATURE REVIEW

A FEW TECHNIQUES OF DATA MINING

Artificial Neural Networks (ANN): ANN, often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural networks, in other words, it is an emulation of biological neural system. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase [7].

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. A type of neural networks is back propagation [6].

In more practical terms, neural networks are non-linear statistical data modelling tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data. Using neural networks as a tool, data warehousing firms are harvesting information from datasets in the process known as data mining. The difference between these data warehouses and ordinary databases is that there is actual manipulation and cross-fertilization of the data helping users makes more informed decisions. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs [7].

Association Rules: Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses make certain decisions, such as catalogue design, cross marketing and customer shopping behaviour 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 are multilevel association rule, multidimensional association rule, and quantitative association rule [6].

Classification: 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

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

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 are classification by decision tree induction, bayesian Classification, neural networks, support vector machines (SVM), and classification based on associations [6].

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 pre-processing 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 are, partitioning methods, hierarchical agglomerative (divisive) methods, density based methods, grid-based methods, and model-based methods [6].

Decision Trees: A decision tree consists of nodes and branches, starting from a single root node. Each node represents a test or decision. Depending on the outcome of the decision, one chooses a certain branch and when a terminal node (leaf) is reached, a decision on a class assignment is made. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. The topmost decision node in a tree which corresponds to the best predictor is called root node. Decision trees can handle both categorical and numerical data.

Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown on the right. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

Genetic Algorithm: Genetic Algorithms (GAs) are adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetic. The basic concept of GAs is designed to simulate processes in natural system necessary for evolution, specifically those that follow the principles first laid down by Charles Darwin of survival of the fittest. As such, they represent an intelligent exploitation of a random search within a defined search space to solve a problem. GAs are one of the best ways to solve a problem for which little is known. They are a very general algorithm and so will work well in any search space. The Genetic Algorithm was developed by John Holland in 1970. GA is stochastic search algorithm modelled on the process of natural selection, which underlines biological evolution. GA has been successfully applied in many search, optimization, and machine learning problems. GA works in an iterative manner by generating new populations of strings from old ones. Every string is the encoded binary, real etc., version of a candidate solution. An evaluation function associates a fitness measure to every string indicating its fitness for the problem [10].

Standard GA applies genetic operators such selection, crossover and mutation on an initially random population in order to compute a whole generation of new strings. GA runs to generate solutions for successive generations. The probability of an individual reproducing is proportional to the goodness of the solution it represents. Hence the quality of the solutions in successive generations improves. The process is terminated when an acceptable or optimum solution is found. GA is appropriate for problems which require optimization, with respect to some computable criterion. The functions of genetic operators are as follows [10]:

1) Selection: Selection deals with the probabilistic survival of the fittest, in that, more fit chromosomes are chosen to survive. Where fitness is a comparable measure of how well a chromosome solves the problem at hand.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

2) Crossover: This operation is performed by selecting a random gene along the length of the chromosomes and swapping all the genes after that point.
3) Mutation: Alters the new solutions so as to add stochasticity in the search for better solutions. This is the chance that a bit within a chromosome will be flipped (0 becomes 1, 1 becomes 0).

Nearest Neighbor Method: Clustering and the nearest neighbor prediction technique are among the oldest techniques used in data mining. Most people have an intuition that they understand what clustering is namely that like records that are grouped or clustered together. Nearest neighbour is a prediction technique that is quite similar to clustering - its essence is that in order to predict what a prediction value is in one record look for records with similar predictor values in the historical database and use the prediction value from the record that it "nearest" to the unclassified record [10].

The nearest neighbor prediction algorithm works in very much the same way except that "nearness" in a database may consist of a variety of factors not just where the person lives. It may, for instance, be far more important to know which school someone attended and what degree they attained when predicting income. The better definition of "near" might in fact be other people that you graduated from college with rather than the people that you live next to [9].

Nearest neighbor techniques are among the easiest to use and understand because they work in a way similar to the way that people think - by detecting closely matching examples. They also perform quite well in terms of automation, as many of the algorithms are robust with respect to dirty data and missing data [9].

Prediction: The prediction as its name implies is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent 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., decision trees) may be necessary to forecast future values. For instance, prediction analysis technique can be used in sale to predict profit for the future if we consider sale to be an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction [8].

Regression: 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 are linear regression, multivariate linear regression, nonlinear regression, and multivariate nonlinear regression [6].

IV. CONCLUSION

Data mining is the process to extract significant information from large data sets so that various sectors can make better business decisions. Data mining applications can be used widely in almost all Business sectors.

In this paper, I try to understand the data mining process, the process of knowledge Discovery, techniques of data mining. Data mining is useful for both public and private sectors for finding patterns, forecasting and discovering knowledge. We found that Data mining is becoming increasingly common in both the private and public sectors. Data

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 12, December 2018

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] Connolly T., C. Begg and A. Strachan "Database Systems: A Practical Approach to Design, Implementation and Management (3rd Ed.)". Harlow: Addison-Wesley.687
[2] "Data Mining Curriculum". ACM SIGKDD. *2006-04-30*. Retrieved on 23 September 2018.
[3] Clifton, Christopher (2010). "Encyclopædia Britannica: Definition of Data Mining", *Retrieved on 23 September 2018*.
[4] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome *(2009)*. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Retrieved on 23 September 2018.
[5] J. Han and M. Kamber, J Pei "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2016, ISBN 978-0-12-3814479-1.
[6] Bharati M. Ramageri, "Data mining techniques and applictions, Indian Journal of Computer Science and Engineering". ISSN No. 0976-5166 Vol. 1 No. 4 301-305.
[7] Dr. Yashpal singh, alok singh chauhan, "Neural networks in data mining, Journal of Theoretical and Applied Information Technology", volume 5 no 6, ***ISSN no.*** 1992-8645.
[8] Kalyani M Raval, Data Mining Techniques, "International Journal of Advanced Research in Computer Science and Software Engineering", volume 2, Issue 10, October 2012 ISSN: 2277 128X.
[9] Alex Berson, Stephen Smith, and Kurt Thearling. "An Overview of Data Mining Techniques", www.thearling.com/text/dmtechniques/dmtechniques.htm retrieved on 29 November 2018.
[10] Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar, Partha Pratim Sarkar, "Mining Frequent Itemsets Using Genetic Algorithm", International Journal of Artificial Intelligence & Applications, Vol.1, No.4, October 2010.
[11] www.meilingbiereport.wordpress.com retrieved on 29 November 2018.