International Journal of Scientific and Technical Advancements
IJSTA
ISSN: 2454-1532

A Review on Data Mining Techniques

Priyanka Sharma [1], Ajay Abrol[2], Sameru Sharma[3], Manoj Kumar[4]
Department of Electronics and Communication Engineering Government College of Engineering and Technology, Jammu
1. priyankasharma1808@gmail.com 2. ajayabrol17569@rediffmail.com 3. sameru33@gmail.com 4. manoj963@rediffmail.com

Abstract— Data mining is a process which helps in finding useful patterns from large amount of data. Important data can be separated from a large amount of raw data based on certain similar characteristics and is done by recognizing the particular pattern. Pattern recognition leads to formation of clusters of data of similar kind. This paper discusses some of the data mining techniques.

I. Overview of Data Mining

With the development of Information Technology a large amount of databases and huge amount of data in various areas has been generated. Researches made in different databases and information technology has always aroused an approach to save and manipulate this important data for further decision making. Data mining is defined as a process of extracting useful information and patterns from the given large amount of data and is called as knowledge discovery process, knowledge mining from data, knowledge extraction or data analysis or pattern analysis.

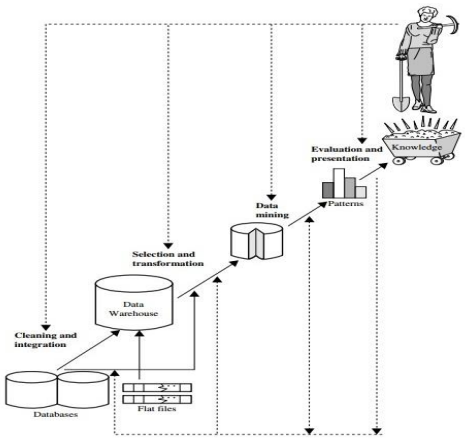

Fig 1: Data mining as a step in the process of knowledge

Data mining is a logical process that finds out useful data from a large amount of raw data. Main goal of this technique is to find previously unknown patterns. Once these patterns are found, they can further be used to make certain decisions for machine learning and predicting analysis.

Data mining involves three steps:
A. Exploration: firstly the data is cleaned and transformed to important variables and then nature of data based on the problem are determined.
B. Pattern Identification: After the exploration, refining and defining of data for the specific variables the second step is to form pattern identification. Identify and choose the patterns which make the best prediction.
C. Deployment: Finally the patterns are put into use for desired outcome [2].

II. Data Mining Algorithms And Techniques

Knowledge is discovered from available databases with the use of different kind of algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbour method etc.

A. Classification

Classification technique for data mining assigns categories to a collection of data in order to attain more accurate predictions and analysis. One of its several methods is decision tree. Its main goal is to set up classification rules that will answer a question, make a decision or predict similar behavior of the data. Hence, to start a set of training data is developed that contains a certain set of similar attributes as well as the likely outcome. Classification algorithm discovers how the set of attributes reaches its desired conclusion. Various types of classification models are classification by decision tree, Neural Networks, Support Vector Machine.

B. Clustering

Clustering is said to be the identification of similar classes of objects. By clustering technique we can further identify dense and sparse regions in object space and discover the overall distribution pattern and correlations among attributes of the given data.

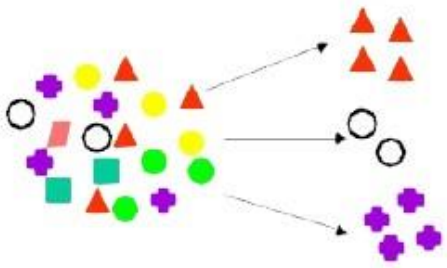

Fig 2: Clustering of data based on similarities

For effective means of distinguishing groups or classes of object clustering approach can also be used. But, it become

Priyanka Sharma, Ajay Abrol, Sameru Sharma, and Manoj Kumar, "A review on data mining techniques," *International Journal of Scientific and Technical Advancements*, Volume 4, Issue 1, pp. 125-126, 2018.

costly, therefore, clustering can be used as pre-processing approach for the selection and classification of attribute subset. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality. Partitioning Methods, Hierarchical Agglomerative (divisive) methods Density based methods, Grid-based methods Model-based methods are the different types of clustering methods [2].

C. Regression

Another technique for data mining is regression technique which can be adapted for predication. Regression analysis can be used to establish the relationship between one or more independent and dependent variables. In data mining, attributes already known are independent variables and what we want to predict are the response variables. Unfortunately, many real-world problems are not simply predicted. 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. Different types of regression methods are Linear Regression, Multivariate Linear Regression, Nonlinear Regression, and Multivariate Nonlinear Regression

D. Association rule

Association and correlation is used to find frequent item set findings among a large set of data. Findings of this type helps 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.

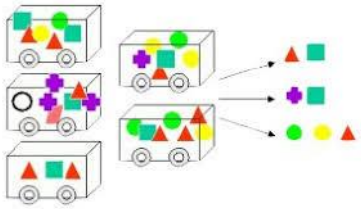

Fig3: Association Rules for a given data set

However, the number of possible Association Rules for a given set of data is generally very large and a high proportion of the rules are usually of little value. Different types of association rule are Multi-level association rule, Multidimensional association rule and Quantitative association rule.

E. Neural networks

A set of connected input/output units is known as a neural network and each connection has a weight present with it. During the phase of learning, network learns by adjusting weights to predict the correct class labels of the input tuples.

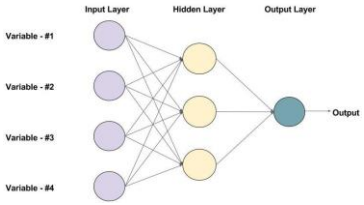

Fig. 4: A feed-forward neural network with one hidden layer(3 neurons)

Neural networks have the noteworthy ability to derive meaning from complicated or vague data and can be used to pull out patterns and detect trends that are complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

III. Conclusion

Data mining is an essential process where intelligent methods are applied to extract data patterns. It has an important significance regarding finding the patterns, forecasting, discovery of complete knowledge etc., in different field of Information Technology. Data mining techniques and algorithms such as classification, clustering etc., helps in finding the patterns in accordance with the certain similar characteristics of the data. Data mining has wide application domain almost in every industry where the data is generated, this is why data mining is considered to be one of the most important frontiers in database and information systems and also the most promising interdisciplinary developments in Information Technology.

IV. References

[1] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, published by Morgan Kauffman, 3rd edition.

[2] Mrs. Bharati M. Ramageri, “Data Min-Ing Techniques And Applications” ,Indian Journal of Computer Science and Engineering Vol. 1 No. 4, ISSN : 0976-5166 pg: 301-305.

[3] Ke Jie, Dong Hongbin, Tan Chengyu and Liang Yiwen, ”PBWA: A Provenance-Based What-If Analysis Approach for Data Mining Processes” Chinese Journal of Electronics Vol.26, No.5, Sept. 2017

[4] LiHua Wang BeiHang Zijun Zhou, “Congestion Prediction for Urban Areas by Spatiotemporal Data Mining”, International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery 978-1-5386-2209-4/17 2017 IEEE

[5] Sagardeep Roy Anchal Garg,” Analyzing Performance of Students by Using Data Mining Techniques A Literature Survey” 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON) GLA University, Mathura, Oct 26-28, 2017, 978-1-5386-3004-4/17

Priyanka Sharma, Ajay Abrol, Sameru Sharma, and Manoj Kumar, “A review on data mining techniques,” *International Journal of Scientific and Technical Advancements*, Volume 4, Issue 1, pp. 125-126, 2018.