I J R B A T, Issue (XI) Vol (I) Jan 2023: 138-142
A Double-Blind Peer Reviewed & Refereed Journal
OPEN ACCESS
e-ISSN 2347 – 517X
Original Article
INTERNATIONAL JOURNAL OF RESEARCHES IN BIOSCIENCES, AGRICULTURE AND TECHNOLOGY
 www.ijrbat.in

ROAD MAP FOR DATA MINING TECHNIQUES

M. N. Quadri

Department of Computer Science, NilkanthraoShinde Science & Arts College, Bhadrawati, Dist. Chandrapur, Maharashtra, India
Corresponding Email: quadrimn@gmail.com

Communicated :10.12.2022 | Revision: 20.01.2023 & 24.01.2023 Accepted: 26.01.2023 | Published: 30.01.2023

ABSTRACT:
Now a day's data is very important part for each sector of the world. Each sector trying to extract the meaningful data form very large amount of database and store for future used. For extracting hidden information from data we have to accept any one technique which takes minimum time to retrieve the meaningful data. In this paper the different techniques are studied and introduced for retrieving the data and try to show which one is the best to retrieve the data. It shows overview detail regarding techniques which perform predictive and descriptive data mining task. I have gone through the existing research work in the area of data mining techniques which will promote us to have a fair analysis in the field of data mining techniques.

Keywords: - Data Mining, Decision Tree, Genetic Algorithm, Clustering, Classification.

INTRODUCTION :

Data mining is used for extract and analyzing the information from the database and provide the valuable knowledge. Applications of data mining by analyzing data is used for decision making, control of business, education systems and market analysis [1].Different methods or techniques are used in data mining for processing and analysis of different data patterns among which well known are decision tree, association, clustering, rules mining, summarization and classification[15]. Today in each sector of the electronic environment, data is collected and store in one center, from this data any firm or organization need the meaningful data anytime and anywhere. So there should be a technique to handle this hidden and meaningful data to the organizer of firm so that they can take quick decision and proceed for their work. Following figure 1 shows the details about data mining process takes place.

In this data mining process the first stage is to collect the different information among different data sources. The data warehouse is created by using data cleaning and integration of the data. The data set is prepared after selection and transformation of data from data warehouse. Extracted patterns are maintained by data mining process. Lastly the end user gets knowledge and information from the data sources by post processing and visualization from extracted patterns. In this paper the data mining technique were study and gives the suggestion for use of better technique to handle the data.

DATA MINING TECHNIQUES :

The descriptive and predictive are the two Data mining techniques used to process the data. The information about input data is provided by the descriptive technique where as in predictive technique hidden or unknown information will be predicted. Data mining used the different techniques such as clustering, classification, association rule etc. [33][7][19].

A. Clustering Technique

In data clustering technique the common or logically similar data collected and processed in

http://doi.org/10.29369/ijrbat.2023.010.1.0023

a group [24]. This technique preforms the common type of data separately. It is one of the effective techniques to find out the patterns in the dataset within the complex data [19][31]. It is one of the popular techniques for data mining. In this technique the similar data forms one group and dissimilar data forms another group [26].

B. Classification Technique

It is observed that in current research, the researcher uses the classification technique because of labeled training data, accuracy of result and advance algorithm which is based on supervised learning[30][6][18][5]. Some constraints areused in classification on which the data is classified in different classes. It has different applications such as marketing, business modeling, credit analysis etc. [2]. Classification is used in bank loan application to predict safe or risky [20]. Most of the researcher uses the classification technique because of easily bifurcation of the data and gives better result as compared to other techniques. The classification techniques are decision tree algorithm, Naïve Bayes algorithm etc. as follows

C. Decision Tree Algorithm

It is one of the classification algorithm used in data mining. It has discrete-valued target function and learned function [17]. Decision tree algorithm shows the different ways of dividing the data set into parts like sections [7].

The class leveled training tuples are learning in the decision tree process. The prediction of any model and important information by large amount of data will be classified. Decision tree algorithm represent as a flowchart which having nodes from top to bottom like a tree. The first node or begging node called as root node [29][4].Each leaf node having class label, each internal node shows a test on an attribute, and each branch denotes results of the test.

The three basic algorithms are broadly utilized that are ID3, CART, and C4.5.

ID3:- ID3 stands for (Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan.It is a classification algorithm. In this all model are summarized to particular group according to unique value of condition attribute sets.

CART:-CART stands for classification and regression tree. In 1984 this algorithm was invented by Breiman. It is based on binary separating of the properties in nature. The splitting criterion is used in CART for data gain. The split attribute is selected on the basis of highest data gain. Data gain is active to form tree from training cases. This tree is implemented as classify test data [13].

C4.5:- In 1993 Roos Quinlan introduced the C4.5 algorithm. This algorithm is modified from ID3 induction. The decision tree created by C4.5 algorithm is used for classification referred to as statistical classifier. It accepts data with numerical or categorical values. It is one of the free data mining tools.

D. K Nearest Neighbor Algorithm:-

K-Nearest-Neighbors is a supervised learning algorithm that generates a model using training set until a query of the data set is performed. It is one of the basic classification algorithms in data mining. It is belongs the lazy learners group[12][23][3].

E. Artificial Neural Network:-

In neural network the biological nervous system like brain is used for processing the information [25]. Identifying the complex data structure and big data, artificial neural network is developed. It having the capability of classification, parameter estimation and revealing pattern apprehension [18].It has three layered architecture, input layer is first layer which carry independent variables. Another layer is a hidden layer which carries activation function for enumerating relationship between input and output layers [11].

F. Association Rule Mining [ARM]

In 1993 Agrawal et al. firstly introduced the concept of association rule mining [10]. It uses

http://doi.org/10.29369/ijrbat.2023.010.1.0023

in different databases such as transactional database, relational database as well as data repositories. It is used with different data sets for finding casual structures, frequent patterns and associations [8][32]. It is basically used for business perspective. It has two phase process in which first used for item set mining and another used for rule generation phase.

G. Outlier Analysis Technique

Now a days fraud detection and international marketing outlier analysis technique is used [17]. More importance is given to outlier analysis so the preference is given to outlier analysis as compare to supplementary data mining technique [16]

H. Genetic Algorithm

Darwin's evolution theory comes out in the form of Genetic algorithm. It shows that the fittest species can survive easily and adopt the changes around environment [32]. The Genetic algorithm completely depends on natural selection process shown by the Darwin [14]. In this algorithm random optimized method is used which search solution by bio-based operators like selection, mutation and crossover etc. The fittest individuals are selected among the population which is one of the natural selection. They bring out Offspring which acquire the eccentricity of parents and it will be added to the coming generation. If parents have desirable fitness, their offspring will be upgraded then procreator, and have a better feasibility of surviving. This process go through unless, a generation with fittest or good enough solitary will be accomplish.

Whenever data mining and Genetic algorithm used together then it gives the efficient and optimized solution. Because of their own characteristics and advantages genetic algorithm plays important role in the field of data mining [9]. Genetic algorithm has importance in various data mining fields like fraud detection, risk analysis, agriculture research, biological research etc. Genetic algorithm and Decision tree solves many difficult task of data mining in collaboration. Genetic algorithm also plays crucial role in different data mining techniques like clustering, classification, association, rule prediction etc.

BETTER DATA MINING TECHNIQUE

As the researcher John Holland developed the genetic algorithm, are illustrious tools for solving the complex problem [27]. Number of studies have clear that genetic algorithm works proficiently in the optimization of complex problem's solution [22]. In Genetic algorithm bio-based approaches is used which makes it more efficient than other algorithms like, FP-growth, ant colony, particle swarm and Apriori algorithms in terms of optimization [28][21].

CONCLUSION :

This paper try to show the overview related to data mining, classification, decision tree, genetic algorithm etc. and also provides the concepts. Data mining having different techniques which are systematized and efficient in their ways but Genetic algorithm which is on biological operators gives the optimized solution to complex problems. At the end of this paper got focus on genetic algorithm with their applications in various sector of data mining. It is one of the time saving and strong algorithm for any database. It gives best learning approach and better understanding related to the data mining for complex dataset.

REFERENCES:

Abdelghani Bellaachia and ErhanGuven "Predicting Breast Cancer Survivability Using Data Mining Techniques", Washington DC 20052, vol. 6, pp. 234-239,2010.

Andrew Secker1, Matthew N. Davies2, Alex A. Freitas1, Jon Timmis3, Miguel Mendao3, Darren R. Flower2 "An Experimental Comparison of Classification Algorithms for the

http://doi.org/10.29369/ijrbat.2023.010.1.0023

Hierarchical Prediction of Protein Function".

Aruma Singh, Smita Patel, Shukla, "Applying Modified K-Nearest Neighborn to Detect Threat in Collaborative Information Information Systems", International Journal of Innovative Research in Science Engineering and Technology., vol. 3, no. 6, pp. 14141-14151, 2014.

B. DeVille, "Decision trees for business intelligence and data mining: Using SAS Enterprise Miner", SAS Institute, Cary, 2006.

B. M. 1. D. N. Jana Jarecki, "The Assumption of Class-Conditional Independence in Category Learning".

Baharudin, B., Lee, L. H., & Khan, K. ,"A review of machine learning algorithms for text-documents classification",Journal ofadvances in information technology, 1(1),4-20,2010.

Bendi Venkata Ramana1, Prof. M.Surendra Prasad Babu2, Prof. N. B. Venkateswarlu "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis" International Journal of Database Management Systems (IJDMS), Vol.3, No.2, May 2011.

BERZAL F,CUBERO J-C,MAR N N. TBAR, "An Efficient Method for Association Rule Mining in Relational Databases"[J]. Data & Knowledge Engineering, pp. 37-47-64, 2001.

D. L. Wang, M. Q. Li. "The application of data mining technology based on genetic algorithm," Journal of Nanchang University, vol. 1, A27, pp. 81-84, 2007.

Dong X., Sun F., Han, X., "Study of Positive and Negative Association Rules Based on Multi-confidence and Chi-Squared Test"[C]. LNAI 4093, Springer-Verlag Berlin Heidelberg,: 100-109, 2006.

F. Bonanno, G. Capizzi, G. Graditi, C. Napoli, G.M. Tina, "A Radial Basis Function Neural Network Based Approach for the Electrical Characteristics Estimation of a Photovoltaic Module", Applied Energy, vol. 97 , pp. 956-961, September 2012.

F. C. a. P. Brazdil,"Comparison ofSVM and some older Classification Algorithms in Text Classification Tasks".

Fong, P.K. and Weber-Jhanke, J.H , "Privacy Preserving Decision Tree Learning using Unrealized Data Sets", IEEE Transactions on knowledge and Data Engineering, Vol.24,No.2, February 2012, pp. 353-364.

G. Y. Yu, Y. Z. Wang, "Applied Research of improved genetic algorithms", Machinery, vol. 5, pp. 58-60, 2007.

Gole, S., &Tidke, B. (2015). A survey of big data in social media using data mining techniques.2015 International Conference on Advanced Computing and Communication Systems. doi:10.1109/icaccs.2015.7324059

Gopalan and B. Sivaselan book on, "Data Mining techniques and trends", published by Asoke K. Ghosh, PHI learning private limited.

M. Khan, S.K. Pradhan, M.A. Khaleel, "Outlier Detection for Business Intelligence using data mining techniques", International journal of Computer Applications, vol. 106, no. 2, pp. 0975-8887, November 2014.

M. R. David D. Lewis, "A Comparison ofTwo Learning Algorithms for Text Categorization", Symposium on Document AnalysisandJR, 1994.

M. S. Packianather, A. Davies, S. Harraden, S. Soman, and J. White, "Data Mining Techniques Applied to a Manufacturing SME", Procedia CIRP, vol. 62, pp. 123–128, 2017.

http://doi.org/10.29369/ijrbat.2023.010.1.0023

Mehmed Kantardzic, "Data Mining: Concepts, Models, Methods, and Algorithms", John Wiley &Sons , ISBN: 0471228524, 2003.

Minaei-Bidgoli, B., R. Barmaki, and M. Nasiri, "Mining numerical association rules via multi-objective genetic algorithms", Information Sciences, 233: p. 15-24, 2013.

Olinsky, Alan D., John T. Quinn, Paul M. Mangiameli, and Shaw K. Chen, "A Genetic Algorithm Approach to Nonlinear Least Squares Estimation.", International Journal of Mathematical Education in Science and Technology 35.2.

R. Agrawal, "K-Nearest Neighborn for Uncertain Data", International Journal of Computer Applications (0975–8887), vol. 105, no. 11, pp. 13-16, 2014.

R. Wang et al., "Review on mining data from multiple data sources", Pattern Recognit. Lett., vol. 0, pp. 1–9, Jan. 2018.

S. Haykin, "Neural networks: a comprehensive foundation", (2nd ed.), Prentice Hall, New Jersey, 1999.

Senthilnath, J., S. N. Omkar, and V. Mani, "Clustering using firefly algorithm: performance study", Swarm and Evolutionary Computation 1, no. 3, 2011.

Sivanandam. S. N.. and S. N. Deepa, "Introduction to Genetic Algorithms", Berlin: Springer. 2007.

Srinivasan, S. and S. Ramakrishnan, "Evolutionary multi objective optimization for rule mining: a review", Artificial Intelligence Review,36(3): pp. 205-248, 2011.

SushmitaMitra, &TinkuAcharya, "Data Mining Multimedia, Soft Computing and Bioinformatics", John Wiley & Sons, Inc, 2003.

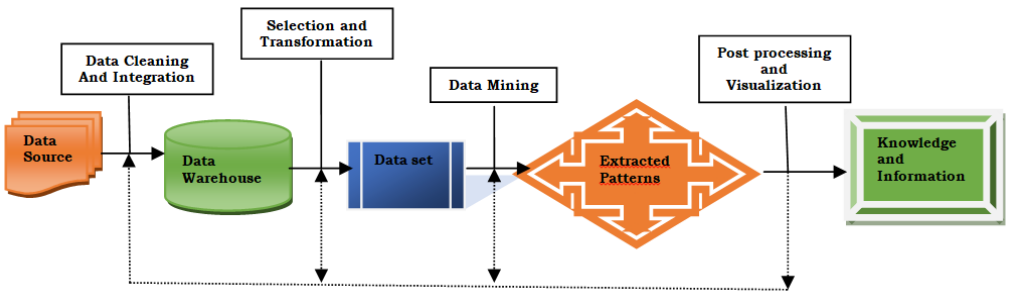
V. C. a. F. Mulier, " Learning From Data", John Wiley & Sons, 1998.

V. Vijay, V. P. Raghunath, A. Singh, and S. N. Omkar, "Variance Based Moving K-Means Algorithm", in 2017 IEEE 7th International Advance Computing Conference (IACC), pp. 841– 847, 2017.

Wu Xindong, Zhang Chengqi, Zhang Shichao,"Mining both Positive and Negative Association Rules", Proceedings of the 19th International Conference on Machine Learning(ICML), San Francisco: Morgan Kaufmann Publishers, 658-665, 2002.

X. Zhong and D. Enke, "A comprehensive cluster and classification mining procedure for daily stock market return forecasting", Neurocomputing, vol. 267, pp. 152–168, Dec. 2017.

Fig.1 Data mining process



http://doi.org/10.29369/ijrbat.2023.010.1.0023