IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

A Survey on Data Mining using Genetic Algorithm

Mariya Khatoon[1*], Abhay Kumar Agarwal[2]

[1,2]Dept. of Computer Science and Engineering K.N.I.T, Sultanpur, U.P, India

Corresponding Author: mariyakhan12301@gmail.com, Tel.: 7607381367

DOI: https://doi.org/10.26438/ijcse/v7i6.888891 | Available online at: www.ijcseonline.org

Accepted: 10/Jun/2019, Published: 30/Jun/2019

Abstract— Growth in the field of data mining rapidly increasing due to its well regulated techniques and efficient algorithms. At present, Genetic algorithm is the bustling research and it allocate the drastic modification in the field of data mining in terms of better optimization of result and the performance of different ventures effectively and efficiently. Because of their accuracy and efficiency, data mining algorithms attract and motivates to researchers to show interest in their technologies and large search space. Genetic algorithm works on bio-responsive operators to evaluate the fittest function in population by the Darwinism. This paper enumerates the enactment of genetic algorithm in frame of reference to data mining algorithms and techniques like decision tree and classification. The main objective of this paper is describes the application and benefits of different data mining techniques related to genetic algorithm.

Keywords— data mining, genetic algorithm, classification, decision tree

I. INTRODUCTION

Currently databases are stiff treasure. Only storage is not the characteristics of database, it also holds the previous knowledge, which is beneficial for current and future reference. It allocates a new market in different discipline and can be a new principle in science and make a million dollars [10]. Non-trivial extrication of implied, previously unknown and prospectively useful information from the data is known as data mining. Data mining is basically a process of inquiring from large repository of data to found new patterns and inclination for interpretation [7]. Data mining uses advanced mathematical algorithms to implode the data and specify the possibility of future occurrences. Knowledge discovery in data is almost identical as data mining.

Among different data mining algorithms, we use genetic algorithm because of its characteristics and specific advantages accompanying to data mining. Searching mechanism is based on enormous composition of parameters in genetic algorithm with the help of bio based terminologies like crossover, mutation, chromosome, fitness function etc. they can search through different combinations and designs to find the perfect combinations of both which could result in a lighter, stronger and overall better final product [25].

There are different techniques in data mining. We use classification technique with decision tree algorithm. Classification technique is used to classify set of data into predefined set of classes or groups. In data analysis task, classification works as a model or classifier to predict the class label attributes. Classification is a data mining function that assigns items in a categories or classes [22]. Decision tree are used to speculate the classes of a categorical based variable from their measurements on one or more predictor or independent variables [17].

II. LITERATURE SURVEY

Shobha et al., proposed that based on the mechanics of natural selection and genetics, genetic algorithm are adaptive heuristic search algorithm. It may be used for various optimization, statistics and complex problems. Global search is the main advantage of genetic algorithm over other knowledge discovery in databases.

Shruti kapil et al., proposed that clustering has been many application areas like software engineering, machine learning, data mining, statistics, image analysis, web cluster engines, and text mining in order to deduce the groups in large volume of data. Clustering defines the objects having same characteristics and attributes belongs to similar group and different which shows different characteristics belongs to different group.

V. Shanmugarajeshwari et al., describes the Feature selection method is used for performance evaluation. For feature selection, there are number of techniques are available in data mining some of them are Chi-Squared Feature selection (CFS), Information gain feature selection (IGFS), Gain Ratio Feature Selection (GRFS) and Correlation based feature selection (CBFS).

Ramandeep Kaur et al., proposed efficient technique of data mining to reduce the multidimensional attacks related network and internet. Support vector machine provides efficient way to detect anomaly and misuse related to internet.

Abhishek Rairikar et al., describes that k nearest neighbor algorithm play a significant role to predict the heart related diseases. K-NN is known as lazy learner and instance based learner approach. It is the simplest algorithms and solves the complex problems.

III. DATA MINING TECHNIQUES

Data mining techniques can be broadly categorized in two different groups those are descriptive and predictive. Descriptive technique provides the information about the input data. Predictive technique basically performs inferences in the input data to generate or predict the information which is previously unknown or hidden. The different techniques are association rule, outlier analysis, classification, clustering, prediction technique etc. [41][8][21].

A. ASSOCIATION RULE MINING

The concept of association rule was firstly introduced by Agrawal et al. in 1993 [12]. Association rule mining allocates the connection between groups of objects in the dataset and it is used for finding frequent patterns, casual structures, associations from various data sets which found in different types of databases such as relational database, transactional database and different data repositories [9][40]. ARM can be said as the two phase process- the first one is the frequent item set mining and the second is rule generation phase. Association rule technique is basically used for business perspective.

B. OUTLIER ANALYSIS TECHNIQUE

Outlier analysis technique is mainly shows the interest in the field of fraud detection and international marketing [19]. In today's world more emphasis is given on outlier analysis as outlier might be indicators of interesting events that have never been known before hence more attention is provided on outlier analysis than the supplementary data mining technique [15].

C. PREDICTION TECHNIQUE

Prediction technique derives the relationship between a thing you know and a thing you need to predict for future reference. Prediction is a tactic used to adumbrate the missing or unavailable data. It is also the two step process equivalent to data classification technique the only difference lies in between them is rather than giving the labels to the data classes whose labels are not known the class labels are anticipated. Abbreviations and Acronyms.

D. CLUSTERING TECHNIQUE

Data clustering is an art of storing the logically similar data composed in a group [30]. Clustering refers to categorizing similar kind of objects. It is a method of exploring the data, a technique of finding out patterns in the dataset and effective technique to manage the complex data [21][39]. It is one of the most vital research fields in the data mining. In clustering we aim at making collections of objects in such a manner that the objects having same attributes belong to same group and objects with different behaviours in dissimilar groups [33].

E. CLASSIFICATION TECHNIQUE

Nowadays classification enhances the accuracy of result due to its advance algorithm and different classifiers and it is based on supervised learning because it uses labelled training data to learn [37][6][20][5]. It is used for classifying data into different classes according to some constrains. Classification has many applications in customer segmentation, business modelling, marketing, credit analysis, and biomedical and drug response modelling [2]. Classification is a form of data analysis that extracts model describing important data classes. Those models are called classifiers; predict categorical class labels. For example, a classification model can be built to categorize bank loan applications as either safe or risky [23]. We use classification technique because it is best suited to our work and it provides the better result as compared to other data mining techniques. Several classification techniques are, k nearest neighbors algorithm, Naïve Bayes algorithm, Support vector machine algorithm, decision tree algorithm, k- means clustering algorithm etc. which are as follows:-

K Nearest Neighbor Algorithm:-

K-Nearest-Neighbors is one of the most basic yet essential classification algorithms in data mining. The k-nearest neighbor belongs to the lazy learners group, is a supervised learning algorithm that means it does not generate a model using the training set until a query of the data set is performed [14][28][3].

Naïve Bayes algorithm:-

The base for the naïve Bayes classifier is Bayes theorem. A hypothesis is generated for the given set of classes. In Naïve-Bayes algorithm independence assumption is made. Based on the target value, the values of the attribute are chosen and it is independent to one another [1]. Let us describe by example why this algorithm is called a naïve. We judge a fruit is an apple when its characteristics are: round 3 inches in diameter and red, even it depends on each other or on other features all of these properties independently contribute to the probability to judge that fruit is apple [32].

Artificial Neural Network:-
Neural network is aroused by biological nervous system like, processing of information through brain [31]. Investigating the complex data structure and big data, artificial neural network is developed. Capability of artificial neural network is parameter estimation, revealing pattern apprehension and classification [18]. Artificial neural network is three layered architecture, first is input layer which carry independent variables. The next one is hidden layer, which carry activation function for enumerating relationship between input and output layers [13].

Support Vector Machine Algorithm:-
Statistical learning theory is the basis of support vector machine, which was firstly developed by Vapnik [38]. Support vector machine supports supervised learning model with associated learning algorithms, which analyze data and recognize patterns. It takes the set of input data and predicts possible classes from the output, for each given inputs, and making it non probabilistic binary linear classifier. Neural network and radial basis functions shows similar kind of functional characteristics from SVM models [8].

Decision Tree Algorithm:-
Decision tree is one of the main Classification Algorithm in data mining. Decision tree a way of approximating discrete-valued target function and in this, learned function is represented by a decision tree [27]. Decision tree algorithm recognizes the different ways of fractionating the data set into branch like sections [4]. Decision tree is a process of learning from class levelled training tuples. It is used for prediction of any model and find out the important information through the large amount of data classification. It is basically like a flowchart having nodes in the form of tree. Topmost node is root node [36]. Each leaf node (terminal node) holds a class label, each internal node (non-leaf node) shows a test on an attribute, and each branch denotes results of the test.

E. GENETIC ALGORITHM
Genetic algorithm is inspired by the "evolution theory of Darwin". That means, the fittest species can survive easily and adopt the changes around environment [25]. It totally based on natural selection process of Darwinism [16]. It is a random optimized method, which finds the solution by bio-based operators like crossover, mutation and selection etc. in Genetic Algorithm the process of natural selection based on fittest individuals from a population. 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.

Application of Genetic Algorithm in Data Mining
When Genetic algorithm and data mining work together it gives the efficient and optimized solution. Due to their own characteristics and advantages genetic algorithm plays vital role in the field of data mining [11]. Genetic algorithm used in different data mining fields like agriculture research, biological research, fraud detection, risk analysis, feature selection air quality etc. hypothesis and cleansing is the main task of genetic algorithm in the contextual relationship of data mining, and it is attained by beginning hypothesis and then permitting all parts of it to alter. Encryption of hypothesis and in the assembling of the evaluation function for fitness is the main facet of genetic algorithm. Decision tree and genetic algorithm solves many difficult task of data mining in collaboration. Genetic algorithm also plays crucial role in different data mining techniques like classification, clustering, rule prediction, association etc.

Why Genetic Algorithm is better than other
John Holland developed the genetic algorithm, are illustrious tools for solving the complex problem [34]. Various studies have manifest that genetic algorithm works proficiently in the optimization of complex problem's solution [26]. There are bio-based approaches in genetic algorithm which makes it more efficient than other meta-heuristic algorithms like, ant colony, particle swarm, FP- growth and Apriori algorithms in terms of optimization [35][24].

Table.1- Application of different meta-heuristic Algorithm.

Sr	Application	Genetic algorithm	Ant colony optimization algorithm	Particle swarm optimization algorithm
1	Natural science	Yes	No	No
2	Mathematics and computer science	Yes	Yes	Yes
3	Biological science	Yes	No	No
4	Bio informatics	Yes	No	No
5	Finance, economic	Yes	No	No
6	Earth science	Yes	No	Yes

IV. CONCLUSION

This paper deliver the overview affiliated to data mining, genetic algorithm, classification, decision tree etc. and also provides the concepts, advantage, limitations, scope, methods, challenges etc. data mining having different algorithms which is systematized and efficient, Genetic algorithm is one of them that optimized the solution with the help of different biological operators. Wind up of this paper focuses on genetic algorithm and their applications in data mining in various ways. It is the most vigorous and time

saving algorithm for any dataset. In fact it gives the understandable and optimal learning approach in context of data mining.

REFERENCES

[1] Aiwen Han, Micheline Kamber, "*Data Mining Concepts and Techniques*".

[2] Andrew Secker1, Matthew N. Davies2, Alex A. Freitas1, Jon Timmis3, Miguel Mendao3, Darren R. Flower2 "*An Experimental Comparison of Classification Algorithms for the Hierarchical Prediction of Protein Function*‖".

[3] 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**.

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

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

[6] 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**.

[7] Bakar, R.S., and Yacef, K, "*The state of educational data mining in 2009: A review and future visions*." JEDM-Journal of Educational Data Mining pp.**1.1 ,3-17, 2009**.

[8] 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**.

[9] 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**.

[10] Chen, M.S., Han, J., and Yu, P.S, " *Data mining: An overview from a database perspective*" IEEE Transactions, Knowledge and Data Engineering, Vol**. 8**, Issue **6**, pp. **866-883, 1996**.

[11] D. L. Wang, M. Q. Li. "*The application of data mining technology based on genetic algorithm*," Journal of Nanchang University, vol. **1**, A**27,** pp. **81-84, 2007**.

[12] 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**.

[13] 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**.

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

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

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

[17] Jiawei Han and Micheline Kamber, "*Data Mining: Concepts and Techniques*",Data Preprocessing, Third Edition, **2011**.

[18] M. Kayri, "*An Intelligent Approach to Educational Data: Performance Comparison of the Multilayer Perceptron and the Radial Basis Function Artificial Neural Networks*", Educational Sciences: Theory & Practice, vol. **15**, no. **5**, pp. **1247-1255**, December **2015**.

[19] 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**.

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

[21] 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**.

[22] Matthew N. Anyanwu, Sajjan G. Shiva, "*Comparative Analysis of Serial Decision Tree Classification Algorithms*", International Journal of Computer Science and Security, (IJCSS) Volume **3** Issue **3** pp.**230** N.P.

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

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

[25] MitsuoGen, Runwei Cheng, "*Genetic Algorithms and Engineering Optimization",* John Wiley and Sons, **2000**.

[26] 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**.

[27] Quinlan, J. R, "*Induction of Decision Trees*", Machine Learning, Vol.**1**, **1986** pp.**81-106. (2004): 207-17**.

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

[29] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "*Analyzing undergraduate students*", performance using educational data mining," Comput. Educ., vol**. 113**, pp**. 177–194**, Oct**. 2017**.

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

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

[32] S. S. Nikam, "*Acomparative Study of Classification Techniques in Data Mining Algorithms*", Orient. J. Comput. Sci. Technol., vol. **8**, no. **1**, pp. **13-19**, **2015**.

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

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

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

[36] Sushmita Mitra, & Tinku Acharya, "*Data Mining Multimedia, Soft Computing, and Bioinformatics*", John Wiley & Sons, Inc, **2003**.

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

[38] V. Vapnik, "*The nature of statistic learning Theory*", Springer, New York, **1995**.

[39] 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**.

[40] 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**.

[41] 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**.

