Novel Approach for Implementing Decision Trees Using Backpropagation for Prediction Analysis

Harpreet kaur1, Shruti Aggarwal2
12Sri Guru Granth Sahib World, University, Fatehgarh Sahib

Abstract- Data mining is an approach which is applied to extract useful information from the raw data. The prediction analysis approach is based on the classification which can predict future possibilities from the current data. This research work is based on the heart disease prediction using back propagation algorithm. The output of the back propagation algorithm is given as input to decision tree classifier. The proposed algorithm is implemented in Anaconda and proposed results show that it performs well in terms of certain parameters.

Keywords- Prediction, Decision tree, Back propagation

I. INTRODUCTION

The process of extraction of interesting knowledge and patterns to analyze data is known as data mining. In data mining there are various data mining tools available which are used to analyze different types of data. Decision making, market basket analysis, production control, customer retention, scientific discoveries and education systems are some of the applications that use data mining in order to analyze the collected information. First step is data cleaning which is used to remove the noise and irrelevant data. Second step is data integration which is used to combine multiple data sources. In third step data are retrieved from the database that comes under the step of data selection. In the fourth step data transformation or consolidation to form appropriate data is done by performing aggregations and summary operations [1]. Data mining is an essential process to extract data patterns by applying different intelligent methods from that knowledge based interesting patterns are identified using pattern evaluation. In the last step the mined knowledge is presented to users by knowledge representations and visualization techniques. It is used to extract large amount of data in order to acquire knowledge which is termed as a misnomer. In today's world large amount of data is collected on daily basis and to analyze such kind of data is very important. The knowledgeable data is extracted from the raw data using the process of data mining. For example, gold is obtained from rocks and sand and is referred as gold mining rather than rock or sand mining [2]. Mining term is characterized as the process of extracting raw material. Many terms are used such as knowledge extraction, data archaeology, knowledge mining from databases and data dredging. This is a world where having a lot of information leads to power and success and this is possible only because of sophisticated technologies such as satellites, computers. The collection of large amount of information has become possible by the advent of means for mass digital storage and computers that helps in storing different types of data. Data mining has great success in many applications and is primarily used today by many companies such as communication, financial, retail and marketing organizations. For the specific customer segments, a retailer can use records of the customer purchase in order to develop product and promote it using data mining. It plays a critical role when it is impossible to enumerate all applications. Image processing, market research, data analysis and pattern recognition, are some of the applications that use cluster analysis [3]. The customer categorized group and purchasing patterns done by clustering can be used by marketer to discover their customer's interest. In biology, mining can be used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in populations. In a city, similar houses and lands area can be identified by employing clustering in geology. To discover new theories, information clustering can be used that classify all documents available on Web [4]. The basic partitioning based method which is used by various clustering tasks that are performed within the low dimensional data sets is known as k-means clustering algorithm. K is utilized as a parameter here and the k clusters are generated by partitioning n objects. Supervised and unsupervised learning are the two methodologies utilized by the data mining. In order to learn the parameters of the model, a training set is utilized in supervised learning while in case of unsupervised learning no training set is utilized, for example k-means clustering.

Classification and prediction are the main objective of the data mining [5]. Disc rate and unordered values or data are classified by the classification models while continuous value is predicted by the prediction models. The examples of classification models are the Decision trees and Neural Networks and example of prediction algorithm is Regression, Association Rules and Clustering. In the classification of data mining, the decision tree approach is considered as the most powerful technique [6]. In this method all the models are build in the form of tree structure. Datasets are broken into small sets and help in the formulation of an associated decision tree. Both the numerical data and categorical data are handled by the decision trees. In the Neural network large numbers of elements are organized in different number of layers that are interconnected to each other. It is the method in which all the multi processing units are combined together using adaptive non-linear data processing algorithms [7]. The simple probabilistic classifier that depends on Bayes' theorem is known as Naive Bayes classifier with strong independent naïve assumption. This algorithm is also known as the independent feature model.

II. LITERATURE SURVEY

Min Chen, et.al (2017) proposed a novel convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm. The data was gathered from a hospital which included within it both structured as well as unstructured types of data. In order to make predictions related to the chronic disease that had been spread within several regions, various machine learning algorithms were streamlined here [8]. It was seen through the various comparisons made amongst existing and the proposed technique that none of the previously existing methods dealt with both types of data that was gathered from medical fields. 94.8% of prediction accuracy was achieved here along with the higher convergence speed in comparison to other similar enhanced algorithms.

Marjia Sultana, et.al (2016) presented that most of the deaths every year are caused due to heart disease, it is the fatal disease [9]. All this process is done on the basis of the data mining techniques. For the investigation of the heart disease various experiments were performed by the author. KStar, J48, SMO, Bayes Net and Multilayer Perceptron were used for this purpose that can be possible through Weka software. Data mining techniques performance is compared with the standard data set in terms of predictive accuracy, ROC curve and AUC value. The SMO and Bayes Net technique shows the optimal performance as compared to the performance of KStar, Multilayer Perceptron and J48 techniques.

M. A. Jabbar, et.al (2016) presented that coronary heart disease is the most fatal heart disease as large amount of deaths occur due to this disease worldwide. Author in this paper discussed the usage of data mining techniques in the medical system [10]. These techniques provide the idea to the doctors whether the patient is suffering from any heart disease or not. The conditional independence assumption of traditional method, in the data mining is relaxed by using this model. For the classification and prediction of heart disease Hidden Naïve Bayes has been utilized in accordance with the proposed model. On the basis of the performed experiments, it is concluded that Hidden Naïve Bayes (HNB) is better than naïve bayes in terms of optimal accuracy.

Theresa Princy, et.al (2016) discussed various data mining techniques have been utilized to detect the rate of the heart disease [11]. Various technologies and different number of attributes has been utilized by many authors for their study. On the basis of number of attributes taken different accuracy was provided by the different technologies. The risk rate of heart disease was detected with the help of KNN and ID3 algorithm and it also provides the accuracy level for different number of attributes. It is concluded from the observation that using new algorithms the numbers of attributes could be reduced that increase the accuracy for the detection of the heart disease.

S.Rajathi, et.al (2016) proposed a technique to enhance the performance of k-Nearest Neighbor (kNN) algorithm that is the integration of Ant Colony Optimization technique. With the help of this method prediction of the heart disease becomes easy [12]. In this technique there are two different phases. kNN algorithm was utilized in the initial phase for the classification of the test data. For the optimized solutions, the ACO technique was utilized as it initializes the population and search to get desired result. In order to present a dataset, Acute Rheumatic Fever (ARF) disease has been utilized that is related to data set. kNNACO algorithm which is an integrated technique is proposed in this paper that is experimented and accuracy is evaluated in terms of accuracy and error rate performance.

Jagdeep Singh, et.al (2016) proposed health care services provide various medical facilities as well as protection against various diseases. Many frameworks have been developed in this paper for the prediction of the heart disease at the early

stage using heart dataset [13]. This Cleveland heart disease is a machine learning repository in the University Of California Irvine (UCI). For the diagnosis of the heart disease there are various parameters such as gender, age, chest pain, blood pressure, blood sugar and many more. As per the performed experiments, it is concluded that a hybrid technique has been utilized for the classification associative rules (CARs) that provides the optimal accuracy.

Ankita Dewan, et.al (2015) presented the neural network technique that is considered as the best among all the classification techniques when comparison done on the basis of the prediction or classification of a non-linear data. The best classifier of Artificial Neural Network is the BP algorithm in which updating technique of weights is used [14]. The errors backward is propagated in this method also. There is limitation in this method that is local minima solution. To solve this issue an efficient optimizing technique was used in this paper that improves the accuracy and used in various application for further prediction.

Tülay Karayilan, et.al (2017) proposed heart disease is the fatal disease from which large number of population is currently suffering as its detection and prevention is major and required to diagnose at the early stage. The process of diagnosis for this diseases is complicated as it requires proper monitoring therefore, early detection of this disease is necessary and accurately. This disease cause maximum numbers of casualities [15]. In the traditional methods there are various limitations as analyzed doing experiments, therefore enhanced methods have been proposed in this paper. On the basis of machine learning medical diagnosis system for the prediction of heart disease has been developed. For the prediction of the heart disease a Back propagation algorithm has been proposed for artificial neural network. Input used has the clinical features in which all the networks were trained using back propagation algorithm. This is done for the neural network in order to determine the condition of the patient whether patient is suffering from heart disease or not.

Ms. Tejaswini U. Mane, et.al (2017) presented the survey performed by the world health organization worldwide for the heart disease in which every year more than 12 million deaths occur due to this fatal disease, therefore maximum casualties are caused due to which detection of this disease is necessary. Heart disease is sometimes referred as the big data approach and for the reduction of such big data, Hadoop Map platform has been utilized. The improvement in the clustering K-means and decision tree algorithm in case of hybrid approach is done by using ID3 for the classification purpose [16]. Heart disease can be diagnosed at the early stage using various parameters such as gender, age, chest pain, blood pressure, blood sugar and so on. As per performed experiments it is concluded that proposed technique provide optimal results for the prediction of the heart disease as compared to other techniques as it improve the treatment process and provide better clinical decision making.

III. PROPOSED METHODOLOGY

This research work is based on the prediction analysis of heart diseases. The prediction analysis is the technique in which future possibilities can be predicted based on the current dataset. In this research work, technique of decision tree is applied previously for the prediction analysis. One of the simplest algorithms amongst all the learning machine algorithms is the decision tree algorithm. Since there are no assumptions made on the underlying data distribution, decision tree is known to be a non-parametric supervised learning algorithm. Here, on the basis of nearest training samples present within the feature space, the samples are classified. The feature vectors are stored along with the labels of training pictures within the training process. Towards the label of its k-nearest neighbors, the unlabelled question point is doled out during the classification process. Through majority share cote, based on the labels of its neighbors, the object is characterized. The object is classified essentially as the class of the object that is nearest to it in the event when k=1. K is known to be an odd integer in case when there are only two classes present. During the performance of multiclass categorization, there can be tie in case when k is an odd whole number. The classification of samples based on the majority class of its nearest neighbor is the major task of decision tree algorithms.

$$Class = arg_v max \sum_{(X_i,y_i)\in D_z} I(v = y_i) \quad \dots (1)$$

Assign all network inputs and output

Initialize all weights with small random numbers, typically between -1 and 1

Repeat

For every pattern in the training set

Present the pattern to the network

for each layer in the network

for every node in the layer

1. Calculate the weight sum of the inputs to the node
2. Add the threshold to the sum

3. Calculate the activation for the node
end
end
for every node in the output layer
calculate the error signal
end
for all hidden layers
for every node in the layer
1. Calculate the node's signal error
2. Update each node's weight in the network
end
end
Calculate the Error Function
End
While ((maximum number of iterations < than specified) AND
(Error Function is > than specified))

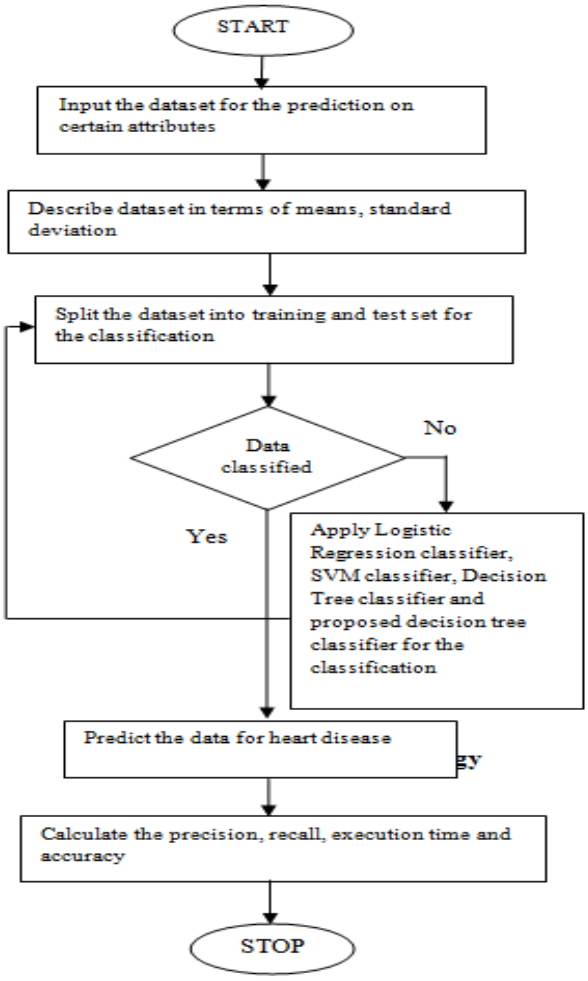

Fig.1: Proposed Methodology

IV. EXPERIMENTAL RESULTS

The proposed approach has been implemented in Python and the results are analyzed in terms of various different parameters.

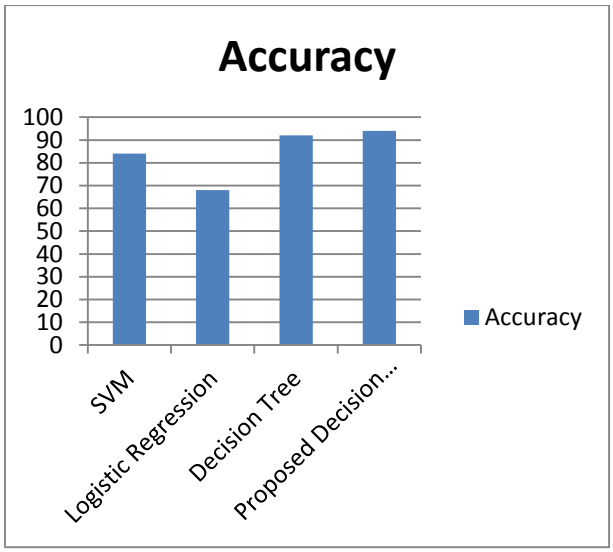

Fig.2: Accuracy Comparison

As shown in Figure 2, the accuracy comparison of existing and proposed algorithm is shown. The accuracy of proposed algorithm is high as compared to existing algorithm.

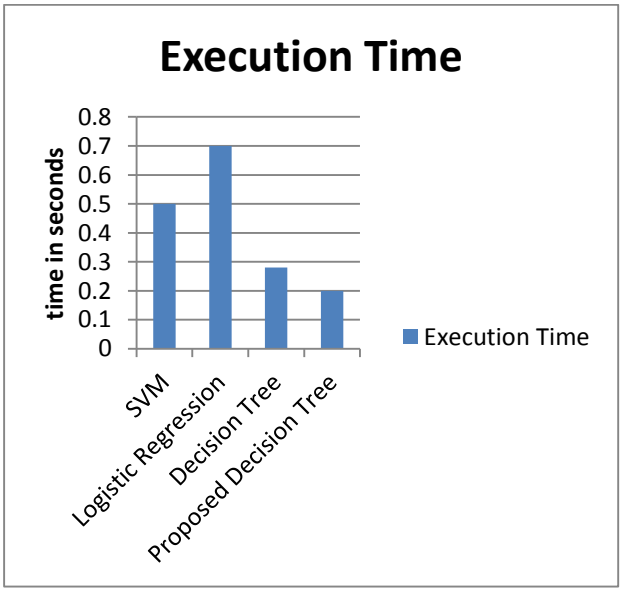

Fig.3: Execution time

As shown in Figure 3, the execution time of proposed and existing algorithm is shown. The execution time of proposed algorithm is less as compared to existing algorithm.

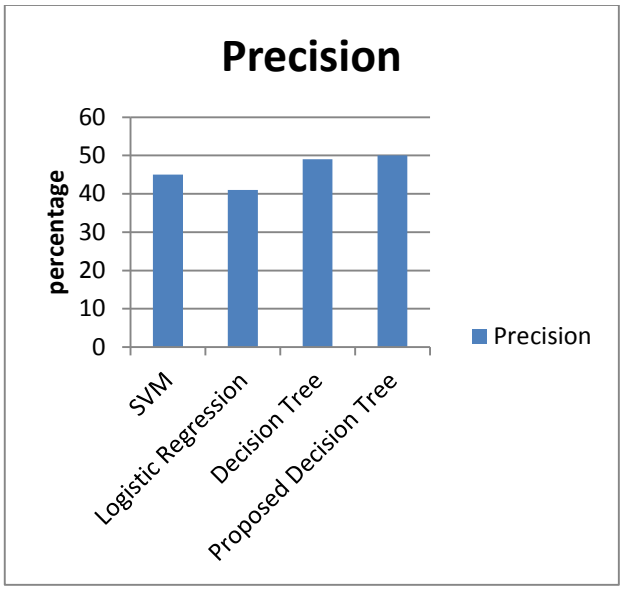

Fig.4: Precision Comparison

As shown in Figure 4, the precision value of the SVM, logistic regression and decision tree is compared. It is analyzed decision tree classifier has maximum precision value as compared to other classifier.

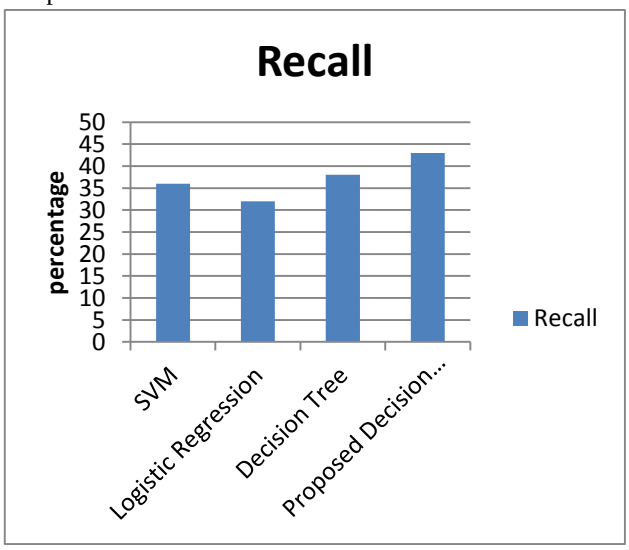

Fig.5: Recall Comparison

As shown in Figure 5, the recall values of SVM, logistic regression and decision tree is compared. It is analyzed that recall value of decision tree classifier is maximum as compared to other classifiers.

V. CONCLUSION

The relevant information is fetched from rough dataset using data mining technique. The similar and dissimilar data is clustered after calculating a similarity between input dataset. The SVM is used to classify both similar and dissimilar data type in which central point is calculated by calculating an arithmetic mean of the dataset. The central point calculated from Euclidean distance is used to calculate a similarity between different data points. In this research work, the technique of back propagation is applied with the decision tree algorithm for the prediction. The back propagation is the technique of neural networks which learn from the previous experience and drive new values. The back propagation is the approach of neural networks which take input attribute number and attribute value as input. The back propagation algorithm drives relationship between the attributes to increase classification of heart diseases. The output of the back propagation algorithm is given as input to decision tree algorithm which increases accuracy of classification. The proposed algorithm is implemented in anaconda and results are analyzed in terms of accuracy, execution time, precision and recall. It is analyzed that proposed algorithm performs well in terms of certain accuracy, execution time, precision and recall.

VI. REFERENCES

[1]. T.John Peter, K. Somasundaram, "AN EMPIRICAL STUDY ON PREDICTION OF HEART DISEASE USING CLASSIFICATION DATA MINING TECHNIQUES", IEEE-International Conference On Advances In Engineering, Science And Management (ICAESM -2012), vol. 4, issue 1, pp. 13-38, 2012

[2]. Theresa Princy. R, J. Thomas, "Human Heart Disease Prediction System using DataMining Techniques", 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT], vol. 4, issue 1, pp. 23-48, 2016.

[3]. Chew Li Sa, Bt Abang Ibrahim, D.H., Dahliana Hossain, E. and bin Hossin, M. (2014), "Student performance analysis system (SPAS)", in Information and Communication Technology for The Muslim World (ICT4M),The 5th International Conference on, vol.15, issue 6, pp.1-6, 2014.

[4]. Qasem A. Al-Radaideh, Adel Abu Assaf and Eman Alnagi (2013), "Predicting Stock Prices Using Data Mining Techniques", The International Arab Conference on Information Technology (ACIT'2013), vol. 23, issue 17, pp. 32-38, 2013.

[5]. Oyelade, O. J, Oladipupo, O. O and Obagbuwa, I. C (2010), "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", International Journal of Computer Science and Information Security, vol. 7, issue 4, pp. 123-128, 2010.

[6]. Tetiana Gladkykh, Taras Hnot and Volodymyr Solskyy, Fuzzy Logic Inference for Unsupervised Anomaly Detection, IEEE

First International Conference on Data Stream Mining & Processing vol. 4, issue 1, pp. 42-47, 2016.

[7]. Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Din, "Data Mining With Big Data",IEEE Transactions on Knowledge and Data Engineering, Vol. 26, issue 1, pp. 23-34, 2014

[8]. Min Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang (2017), "Disease Prediction by Machine Learning over Big Data from Healthcare Communities", 2017, IEEE, vol. 15, issue 4, pp- 215-227, 2017.

[9]. Marjia Sultana, Afrin Haider and Mohammad ShorifUddin, "Analysis of Data Mining Techniques for Heart Disease Prediction", IEEE, vol. 14, issue 1, pp. 123-138, 2016.

[10]. M. A. Jabbar, Shirina samreen, "Heart disease prediction system based on hidden naïve bayes classifier", vol. 4, issue 11, pp. 23-48, 2016.

[11]. Theresa Princy. R, J. Thomas, "Human Heart Disease Prediction System using DataMining Techniques", 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT], vol. 4, issue 1, pp. 23-48, 2016.

[12]. S.Rajathi, Dr.G.Radhamani, "Prediction and Analysis of Rheumatic Heart Disease using kNN Classification with ACO", IEEE, vol. 4, issue 7, pp. 223-248, 2016.

[13]. Jagdeep Singh, Amit Kamra, Harbhag Singh, "Prediction of Heart Diseases Using Associative Classification", IEEE, vol. 7, issue 9, pp. 23-48, 2016.

[14]. Ankita Dewan, Meghna Sharma, "Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification", IEEE, vol. 43, issue 6, pp. 13-24, 2015.

[15]. Tülay Karayilan Tülay Karayilan, "Prediction of Heart Disease Using Neural Network", IEEE, vol. 14, issue 1, pp. 423-468, 2017.

[16]. Ms. Tejaswini U. Mane, "Smart heart disease prediction system using Improved K-Means and ID3 on Big Data", 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI), vol. 8, issue 11, pp. 123-148, 2017.