Propose and Evaluate Classification Technique for Heart Disease Prediction

Rupinder Kaur Gurm, Navjot Kaur

Asst. Professor, M Tech Scholar

Department of Computer Science
RIMT University, Mandi-Gobindgarh.

Abstract: The data mining is the technique to analyze the complex data. The prediction analysis is the technique which is applied to predict the data according to the input dataset. In the recent times, various techniques have been applied for the prediction analysis. In the base paper, SVM technique is applied for the prediction analysis. In the technique whole data is divided into testing and training part. The test data is classified into two classes' means first class is of data which have less chances of heart disease and second which have high chances of heart disease. In this research work, further improved will be proposed in this existing method using decision tree classifier. The proposed improvement increase accuracy of classification and reduce execution time.

Keywords: SVM, KNN, Neural Networks, Back Propagation

1. 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. The multimedia, object relational, relational and data ware houses are some of the databases for which data mining has been studied. First step is data cleaning which is used to remove 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. Data mining is an essential process to extract data patterns by applying different intelligent methods then knowledge based interesting patterns are identified using pattern evaluation. In the last step the mined knowledge are presented to users by knowledge representations and visualization techniques. The group membership for data instances can be predicted with the help of classification technique within the data mining. In order to predict the data for example classification can be utilized by the applications on a specific day to identify the weather which can be either “sunny”, “rainy” or “cloudy”. Two steps are followed within this process. Image processing, market research, data analysis and pattern recognition, are some of the applications that use cluster analysis. The customer categorized group and purchasing patterns done by clustering can be used by marketer to discover their customer’s interest. In biology, it 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. The unsupervised data clustering classification method create clusters, group of objects in such a way that objects in different clusters are distinct and that are in same cluster are very similar to each other. In data mining, cluster analysis is considered as one of the traditional topics that is the first step in discovery of knowledge. The data objects are grouped into a set of disjoint classes which is known as clusters. Now, objects within a class have high resemblance to each other and in the meantime objects in separate classes are more unlike. Unsupervised clustering means that clustering does not rely on predefined classes and trainings. Pattern recognition and unsupervised learning is different from each other in the area of statistics known as decision analysis. The other name is discriminate analysis that classifies objects from given set of object. 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. It ensures that the similar types of objects are grouped within one cluster and dissimilar objects are placed in separate clusters. The cluster centres are identified here with the help of this algorithm. It is made sure that there is a reduction of the sum of the squared distances of each data point to the nearest centre of cluster. Initially, the k objects are chosen on random basis. Here, a cluster mean or centre is represented by each of these randomly chosen objects. Further, the nearest cluster centre gets assigned to it, by each object present within the data set. New mean for each of the clusters is computed further by this algorithm. On the basis of the given output prediction for the certain outcomes can be done by classification process. The outcome is predicted by the processed algorithm as different set of attributes are present in the training set and in the respective outcome which is called as goal or prediction attribute. The relationship between the attributes is discovered by the algorithm that will be helpful in the prediction of the outcome. Genetic programming (GP) has been widely used in the area of research as it solves data mining classification problems. The major fact that genetic programming is so widely used is because it predict the rules effectively that are naturally represented in GP. Results are produced by the GP with global search problems like classification [6]. Several ‘peaks’ are present in the
search space for classification this cause local search algorithms also known as simulated annealing that performs badly. Neural networks are the interconnectivity between the processing elements also called units, nodes, or neurons. These networks are designed after the cognitive processes of the brain. These networks are used to predict new outcomes from the previous observations. In order to produce an output function all the present neurons within the network work together. Ant Colony algorithms is the natural inspired technique and by the behavior of ants as they help in finding the optimal path from the colony to food. They use the good paths within a graph in order to find optimal ways.

II. LITERATURE REVIEW

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 [7]. 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 is caused due to heart disease, it is the fatal disease. It is necessary diagnose this disease at the early stage as maximum number of causalities are occurred from this disease [8]. 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. 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) discussed the use of data mining techniques in the medical system [9]. These techniques provide the idea to the doctors whether the patient is suffering from any heart disease or not. Hidden Naïve Bayes is the extended version of the traditional Naïve Bayes method in the data mining. 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 superior to 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. For the effective and efficient diagnosis of the heart disease various Data mining techniques and classifiers has been utilized so far, discussed in this paper. All the obtained results compared to provide effective technique [10]. 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 in order to enhance the performance of k-Nearest Neighbor (kNN) algorithm is the integration of Ant Colony Optimization technique [11]. 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 techniques is proposed in this paper that is experimented and accuracy is evaluated in terms of accuracy and error rate performance.

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 causalities [12]. 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.

III. RESEARCH METHODOLOGY

The prediction analysis is the technique which can predict the future possibilities from the existing data. The prediction analysis techniques are based on the clustering and classification. In the basepaper, medical data is analyzed to predict the regional diseases. The data is collected from the central china in 2013-2015 to prepare modal for predicted analysis. The base paper modal for the prediction analysis is based on the neural networks. The clustered data is given as input to the classification algorithm which can divide the dataset into two parts testing and training. The SVM classifier is used to classify the data into certain number of classes. In the k-mean clustering algorithm, the centered points are calculated by taking arithmetic mean of the whole dataset which can reduce accuracy of prediction analysis. When the dataset is complex, it is difficult to establish relationship between the attributes of the dataset. In this research work, the decision tree classifier is applied which can classify the wheat production in certain number of classes. The decision tree classifier can be replaced with some other classifiers to improve accuracy of classification.
A. Proposed Pseudo Code

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
  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
    Calculate the Error Function
  end
while ((maximum number of iterations < than specified) AND (Error Function is > than specified))

IV. EXPERIMENTAL RESULTS

The proposed work has been implemented in Python and its performance has been evaluated by making comparisons with existing algorithm in terms of various parameters.

Figure 1: Accuracy Comparison
As shown in figure 1, the accuracy comparison of existing and proposed algorithm is shown. The accuracy of proposed algorithm is high as compared to existing algorithm.

Figure 2: Execution time
As shown in figure 2, the execution time of proposed and existing algorithm is shown. The execution time of proposed algorithm is less as compared to existing algorithm.

Figure 3: Success Ratio
As shown in figure 3, the success ratio of proposed and existing algorithm is compared for performance analysis. The success ratio of proposed algorithm is high as compared to existing algorithm.
As shown in figure 4, the CAP analysis is shown in this figure. On the axis- of this curve the training dataset is given as input and on the y-axis the test data is given as input. The blue line shows that CAP curve which represents accuracy of the classifier.

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 k-mean clustering is used to cluster both similar and dissimilar data type in which central point is calculated by calculating an arithmetic mean of the dataset. The central point calculated Euclidian distance is used to calculate a similarity between different data points. According to the type of input dataset a clustered data is classified using SVM classifier scheme in the last step. The clustering accuracy get reduced when some of the data points remain uncluttered that has been concluded in this work. In this work technique will be proposed which calculated Euclidian distance in the iterative manner and increase clustering accuracy.

REFERENCES