Tuned Random Forest Algorithm for Improved Prediction of Cardiovascular Disease

P.Nancy, B.Swaminathan, K.Navina, B.Nandhine, P.Lokesh

Abstract-Data mining is becoming more and more popular and essential in the field of medicine. The large amounts of data produced everyday by the medical industry are very complex and voluminous to be processed and analyzed by the usual traditional means. In such cases data mining comes into play. Despite the presence of several prediction algorithms, the efficiency is questionable due to the presence high error rate. Therefore it is necessary to choose a prediction algorithm that gives higher accuracy with fewer errors. The aim of this paper is to create a system for efficient and accurate prediction of cardiovascular disease. The datasets for the process is taken from UCI machine learning repository. The datasets are tested for accuracy using ANOVA technique. The algorithms are investigated using the WEKA tool. The best features for prediction are obtained from feature selection algorithms. Various classification algorithms are applied on the datasets to identify the most efficient algorithm. We observe that random forest gives consistently better accuracy than other algorithms. Tuning is done on the random forest algorithm to further improve the accuracy of prediction system.

Keywords-Random forest, data mining ,classification algorithms, WEKA, tuning ,accuracy

I. INTRODUCTION

The World Health Organisation (WHO) states that cardiovascular diseases are one of the most prevalent causes of death globally and a large number of people die due to this disease every year. Cardiovascular diseases include deep vein thrombosis, peripheral arterial disease, coronary heart disease, congenital heart disease, cerebrovascular disease, and rheumatic heart disease. Most of these cardiovascular diseases can be prevented by addressing certain behavioural risk factors like unhealthy diet, usage of tobacco, obesity, physical inactivity and unhealthy use of alcohol .It is the need of the hour to create a system that allows accurate prediction of heart diseases in early stages to take appropriate measures to halt the disease.

Data mining, as the name suggests, is nothing but the process of discovering patterns in large amounts of data. Data mining aims to extract meaningful information. It is the process of analyzing hidden patterns of data for categorization into information which will be useful. It is used to facilitate decision making and efficient analysis. The medical data mining is said to be an important field in terms of research due to its significance in the development of various applications in the healthcare domain. Several researches have been carried out earlier to build classification models by combining or using individually the data mining algorithms and techniques such as Naïve Bayes, neural network, unsupervised algorithms such as support vector machine, etc. Confusion matrices are constructed with respect to the dataset which helps the researchers to estimate the performance of each machine learning algorithm.

II. RELATED WORKS

Numerous researches and works have been done related to heart disease diagnosis **using** different data mining techniques and methodologies.

H.Benjamin's paper [1] is a comparative analysis of algorithms such as random forest, decision trees, clustering and Naive Bayes. The results of this paper present that the random forest algorithm is best suited for prediction of heart disease than the other algorithms considered, namely decision tree and Naïve Bayes classification algorithms. Princy. R and Thomas (2016) focused on the data mining techniques such as Naïve Bayes algorithm, Decision Tree, Neural Network and KNN to predict the risk of heart disease in patients. It is found that the result is highly accurate when large number of attributes are used. M.A.Jabbar's [2] model uses random forest algorithm as classification algorithm. For feature selection, chi square and genetic algorithms are used as measures to predict heart disease. In order to build and train the classifier, 75% of the data set is. The remaining 25% of the data from the data set is used for testing with 10-fold cross validation. Another paper [3] uses the data from Cleveland Heart Disease database consisting of 303 records & Statlog Heart Disease database which consists 270 records for performing the comparative analysis of three algorithms Neural Networks, Decision tree and Naive Bayes. S.Kiruthika Devi, S. Krishnapriya and Dristipona Kalita‘s[7] paper is as follows. This paper is for perfect analysis of heart disease. The output of each algorithm is combined. After combining, the output will be compared. Different algorithms are used such as Decision tree, ANN, Nave Bayes and SVM algorithms are used. Vishal Jadhav, Devendra Ratnaparakhi, Tusdhar Mahajan‘s [8] paper gives us details about Heart Disease Prediction System which is a web application. This web application fetches the data from stored database and compares them with the stored dataset.

Revised Manuscript Received on May 06, 2020.

Dr.P.Nancy, Assistant Professor, Department of Computer Science and Engineering,Rajalakshmi Engineering College, Chennai -602105. nancy.p@rajalakshmi.edu.in

Dr.B.Swaminathan, Professor, Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai -602105. swaminathan.b@rajalakshmi.edu.in

K.Navina, UG Scholars ,Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai -602105. navina.k.2016.cse@rajalakshmi.edu.in

B.Nandhine, UG Scholars ,Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai -602105. , nandhine.b.2016.cse@rajalakshmi.edu.in

P.Lokesh, UG Scholars ,Department of Computer Science and Engineering, Rajalakshmi Engineering College, Chennai -602105. lokesh.p.2016.cse@rajalakshmi.edu.in

III. PROPOSEDCARDIOVASCULARDISEASE PREDICTION FRAMEWORK

The proposed system is to use feature selection and tuning of random forest algorithm to obtain better cardiovascular prediction system.

The datasets for prediction are taken from UCI machine learning repository. There are three datasets namely Cleveland, Switzerland and Hungarian. Cleveland dataset is most popularly used in various research papers. We have analysed the performance of algorithms in all these datasets.

The framework for the proposed system is given in Figure 1. The basic processes involved are

i) Data collection
ii) Pre-processing
iii) Feature selection
iv) Classification and
v) Prediction

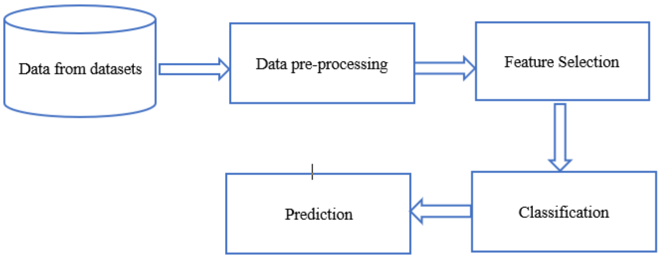


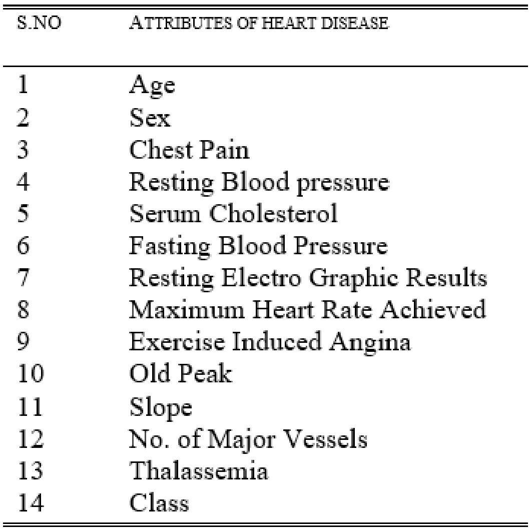
Fig 1: Cardiovascular Disease Prediction framework

A. Data Set Description

We have obtained our data from UCI heart-disease directory which contains 4 databases related to heart disease diagnosis. All attributes in these datasets are numeric-valued. The data was collected from the following locations:

1. Cleveland Clinic Foundation (cleveland.data)
2. Hungarian Institute of Cardiology, Budapest (hungarian.data)
3. University Hospital, Zurich, Switzerland (switzerland.data)

The data in each of these databases have the same instance format. 76 raw attributes are present in these datasets. We have used only 14 among them



S.NO	ATTRIBUTES OF HEART DISEASE
1	Age
2	Sex
3	Chest Pain
4	Resting Blood pressure
5	Serum Cholesterol
6	Fasting Blood Pressure
7	Resting Electro Graphic Results
8	Maximum Heart Rate Achieved
9	Exercise Induced Angina
10	Old Peak
11	Slope
12	No. of Major Vessels
13	Thalassemia
14	Class

Table 1:Attribute Description

a) Cleveland dataset

The Cleveland dataset has 299 instances with 6 missing values. There are 14 attributes in total which describes the risk of cardiovascular disease.

b) Hungarian dataset

The Hungarian dataset contains 294 instances including a total of 491 missing values. Original database has 14 columns and 294 rows each having eight categorical columns such as sex, cp(chest pain), fbs(fasting blood pressure), restecg(resting electro graphic results), exang(exercise-induced angina), slope, thal(thalassemia), num(number of major vessels) and five numeric columns which are age, trestbps(resting blood pressure), chol(serum cholesterol), thalach and oldpeak.[9]

c) Switzerland dataset

The Switzerland dataset has 123 instances with 273 missing values. This dataset also contains 14 attributes

B. Data Preprocessing

Initially, ANOVA test is done on the datasets to validate the accuracy of the data. All records in the dataset with missing values are removed which helps to improve the accuracy of the prediction of various algorithms. Then the dataset is split into two parts.

Training data: The machine is trained with this data. 60% of the data from the dataset is taken as training data. This data is used to train various classification algorithms.

Validation data: The validation data is used for the cross-validation of the data mining results. It is used to find the rate of accuracy of each classification algorithm. The remaining 40% of the data is taken as validation data.

C. Feature Selection

Feature selection, also known as variable selection, attribute selection or variable subset selection, is known as the process of selecting a subset of relevant features for use in data mining model construction.

There are 14 attributes that are helpful in predicting the presence of cardiovascular disease. Feature selection algorithms help to identify the attributes that are more relevant than other attributes. This helps to increase the accuracy.

We used CFS-Subset feature selection algorithm with BestFirst search method and FilteredSubsetEval with GreedyStepwise search method. 9 and 10 attributes have been selected by the algorithms respectively, out of the 14 attributes, which could be used to better describe the model and increase the accuracy of the prediction systems.

a) Correlation based Feature Selection(CFS)

A good feature will always be highly correlated to the class or category and not repetitive to other features which may be relevant. There are two stages in correlation based feature selection. They are

i) selecting relevant features from the class and finding the redundant features
ii) removing such redundant features from the original dataset.[7]

CFS [6] is said to be a fully automatic algorithm. This is because, it is not required by the user to specify the number of features to be selected or specify any thresholds, although they can be incorporated easily if required. CFS operates on the original feature space, which means that any knowledge created by a learning algorithm, using features selected by CFS, can be understood in terms of the original features and not in the terms of a transformed space. Most significantly, CFS is a filter, and does not acquire the high computational cost that is associated with repeatedly applying a learning algorithm.

D. Classification

Classification [4] is the process of finding a group of models that describe and distinguish data classes. This is done to realize the goal of having the ability to use the model to predict the classes whose label is not known. This phase involves the execution of the classification algorithms to spot the best performing algorithm among others. The classification accuracy obtained by percentage split as described in the data pre-processing phase, was calculated and a comparison was done amongst the classifiers. The algorithms that yielded the highest accuracies are described below.

a) Random Forest

In random forest [2] a randomly selected set of attributes is used to split each node.

Every node is split using the best split among a subset of predictors which are purposely chosen at random at the node. This methodology is different from the ones followed in standard trees, in which each node is split using the best split among all attributes which are available in the dataset considered. Further, new values are predicted by combining the predictions of many constructed decision trees.

Random forest represents an ensemble model or an algorithm because it derives its final prediction from numerous individual models. These individual models could be of either similar type or different type. In the case of random forest algorithm, since decision trees are used, the individual models are of the same type.

Algorithm

i) A bootstrap sample is selected from the training set

ii) An un-pruned tree is grown on this bootstrap sample.

iii) A number of nodes are randomly selected at each internal node and best split is determined.

iv) The majority vote from all the trees is taken as the overall prediction.

b) Naïve Bayes

Naïve Bayes (NB)[1] is a statistical classifier which assumes no enslavement between attributes. Bayes rule is the basis for Naive Bayes and it assumes that attributes are independent of each other. The principle of Naïve Bayes classifier is as follows:

• Training Step: By assuming predictors to be conditionally independent given for a class, the method estimates the parameters of a probability distribution which is known as the prior probability.

• Prediction Step: This step finds the posterior probability for unknown test data of the dataset which belongs to each class. This method is then used to classify the test data based upon the largest posterior probability.

c) Multilayer Perceptron

An MLP [10] (or Artificial Neural Network - ANN) with a single hidden layer can be represented graphically as shown in figure 2.

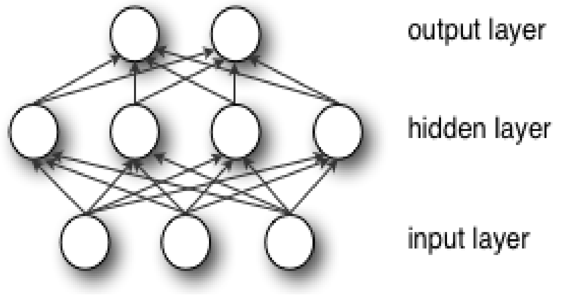


Fig 2: Neural Network

Formally, a one-hidden-layer MLP is a function

$f : R^D \rightarrow R^L,$

where D is the size of input vector x and L is the size of the output vector f(x), such that, in matrix notation:

$$f(x) = G(b^{(2)} + W^{(2)}(s(b^{(1)} + W^{(1)}x))),$$

vith bias vectors b^(1), b^(2); weight matrices N^(1), W^(2) and activation functions G and s.

E. Tuning

Machine learning algorithms can be tuned to give better accuracy than the traditional methods. We can alter and play with the features of algorithm in order to discover the combination of parameters that result in the best performance for our problem. There are a number of parameters that can be changed. The main parameters to adjust when tuning are max_features and n_estimators. The max_features is the size of the random subsets of features to be considered when a node is split. The lower value of max_feature, the greater is the reduction of variance, but this also increases the in bias. The n_estimators is the number of trees in the random forest. A larger number of trees are better, but also it will take too long compute. We will also find that the results will stop getting significantly better beyond a critical number of trees. [11]

In WEKA tool, max_features are given as numFeatures. The number of random features to split the tree on can be decided on the basis of Gini index of the dataset.

IV. RESULT ANALYSIS

The data is loaded is trained in the model using various classification algorithms. A percentage split of 66% is done to separate the training and test data. The accuracy is calculated and obtained.

Accuracy rate = (No. of correctly classified observations / Total No. of observations) x 100

The following tables depict the accuracy of algorithms in various datasets.

SNO	ALGORITHM	WITHOUT FEATURE SELECTION	CfsSubsetEval FEATURE	FilteredSubsetEval FEATURE SELECTION

1	BayesNet	85	83	86
2	Naïve Bayes	83.67	84	86
3	Naïve Bayes Multinomial text	60	60	60
4	Naïve Bayes Updateable	83.67	84	86
5	Multilayer perceptron	80	84	77
6	Simple Logistic	85	82	87
7	SMO	83	83	78
8	Classification Via Regression	84	85	85
9	AdaBoost M1	82	82	85
10	Iterative Classification	84	84	85
11	LogitBoost	85	84	85
12	Decision table	79	83	79
13	Random Forest	83	84	76
14	Random Tree	77	84	80
15	REP Tree	80	81	76

Table 2 a: Result of Hungarian dataset

SNO	ALGORITHM	WITHOUT FEATURE SELECTION	CfsSubsetEval FEATURE SELECTION	Filtered SubsetEval FEATURE SELECTION	ConsistencySubsetEval FEATURE SELECTION
1	BayesNet	60	60	57.84	
2	Naïve Bayes	61.764	62.72	60.784	60.784
3	Naïve Bayes Multinomial text	53.92	53.92	53.92	62.745
4	Naïve Bayes Updateable	61.764	62.74	60.784	53.92
5	Multilayer perceptron	53.92	62.74	61.764	62.745
6	Simple Logistic	60.78	62.74	62.745	57.84
7	SMO	62.74	58.82	59.803	58.83
8	Classification Via Regression	56.86	60.784	59.803	60.78
9	AdaBoost M1	53.92	53.9215	53.92	58.82
10	Iterative Classification	59.8	62.74	64.705	53.19
11	LogitBoost	54.9	62.74	63.725	62.74
12	Decision table	58.82	57.8431	57.843	59.80
13	Random Forest	62.5	60.78	56.862	58.82
14	Random Tree	53.92	56.86	56.862	57.92
15	REP Tree	59.80	59.8039	57.843	55.80

Table 2 b: Result of Cleveland dataset

SNO	ALGORITHM	WITHOUT FEATURE SELECT	CfsSubset Eval FEATURE SELECTION	Filtered Subset Eval FEATURE

		ION		SELEC TION
1	BayesNet	33.33	38.095	33.33
2	Naïve Bayes	26.19	38.095	33.33
3	Naïve Bayes Multinomial text	38	37	38.095
4	Naïve Bayes Updateable	26.19	38.095	33.33
5	Multilayer perceptron	35.78	33.33	28.57
6	Simple Logistic	42.80	38.095	28.571
7	SMO	30.95	38.095	35.714
8	Classification Via Regression	21.42	38.095	30.95
9	AdaBoost M1	38.09	35.714	35.71
10	Iterative Classification	35	35.71	33.33
11	LogitBoost	40.47	35.71	28.51
12	Decision table	35.714	35.714	35.71
13	Random Forest	38.095	38.095	23
14	Random Tree	35.71	38.095	21.428
15	REP Tree	38.09	38.095	28.57

Table 2 c: Result of Switzerland dataset

We find that random forest gives consistently better accuracy and therefore it is chosen as the classifying algorithm for our system.

When the random forest algorithm is tuned with numFeatures=12 , it gives us the best accuracy. The change in accuracy with the numFeatures is given in table 3.

Table 3: Variation of accuracy with numFeatures

numFeatures	ACCURACY
1	56.86
2	59.80
3	60
4	62.7457
5	63.7255
6	64.7059
7	64.7059
8	63.7255
9	62.7457
10	62.7457
11	63.7255
12	**65.68**
13	62.7457
14	62.7457

It can be observed that the tuning the algorithm gives better results that features selection in cases. Hence tuning can be done to an algorithm if feature selection does not yield high accuracy.

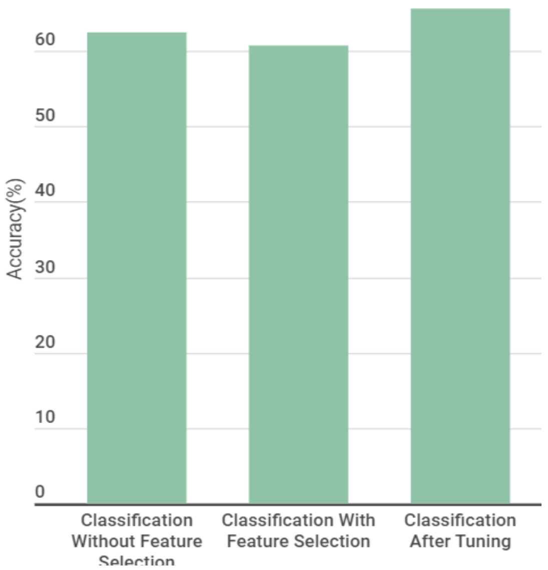


Figure 3: Accuracy Vs Methods of classification

V. CONCLUSION

Datamining in the field of medicine is an emerging trend and a very important one. It is necessary to find better ways to diagnose diseases with accuracy in order to prevent and cure them. In our system, we have proved that feature selection increases accuracy than traditional methods of classification. A combination of feature selection algorithm and classification is needed for efficient prediction of heart diseases. It is also seen that Random forest ,Naïve Bayes, and Neural networks have better accuracy than other algorithms. The classification produces greatest accuracy when tuned. Hence tuning can be done if feature selection does not yield expected high accuracy. Further researches to predict cardiovascular diseases with higher accuracies will benefit the medical community and the whole of mankind.

REFERENCES

1. H. Benjamin Fredrick David and S. Antony Belcy, .“Heart disease prediction using data mining techniques” ICTACT journal ,October 2018.
2. M.A.Jabbar, B.L.Deekshatulu and Priti Chandra, “Intelligent heart disease prediction system using random forest and evolutionary approach”, Journal of Network and Innovative Computing ISSN 2160-2174 Volume 4 (2016)
3. Chaitrali S. Dangare and Sulabha S. Apte , “Improved Study of Heart Disease Prediction System using datamining classification system “ ,IJCA journal, June 2012.
4. Shashaank D.S, Sruthi.V, Vijayalashimi M.L.S and Shomona Garcia Jacob , “Improved turnover prediction of shares using hybrid feature selection“ ,International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.6, November 2015 .
5. Mark A. Hall, “Correlation-based Feature Selection for Machine Learning” ,The University of Waikato ,NewZealand(1999)
6. Kowshalya, A.M., Madhumathi, R. & Gopika, N, “Correlation Based Feature Selection Algorithms for Varying Datasets of Different Dimensionality.”, Wireless Pers Commun 108, 1977–1993 (2019).
7. S. Kiruthika Devi, S. Krishnapriya and Dristipona Kalita, “Prediction of Heart Disease using Data Mining Techniques”, Department of CSE, SRM University.
8. Vishal Jadhav, Devendra Ratnaparakhi, Tusdhar Mahajan, “iDiagnosis -The Intelligent Medical Diagnostic System”, (July 2019)
9. R. Misir, R.K. Samanta , “A Study on performance of UCI Hungarian dataset using missing value management techniques”, International Journal of Computer Sciences and Engineering”, vol. 5, issue 3, pp. 40-44.
10. http://www.deeplearning.net/tutorial/mlp.html
11. https://scikit-learn.org/stable/modules/ensemble.html#parameters