International Journal of Engineering & Technology, 7 (2.33) (2018) 1068-1070

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Prediction of Heart Disease Using Regression Tree

Nagaraj M. Lutimath, Arathi B N,Shona M

*Department of Computer Science and Engineering,Sri Venkateshwara College of Engineering, Bengaluru, India, Department of Computer Science and Engineering, Sri Venkateshwara College of Engineering,Bengaluru India, * Corresponding author E-mail: nagarajlutimath@gmail.com , aradvg@gmail.com,sonasuresh04@gmail.com*

Abstract

Decision trees are an important paradigm in machine learning. They are simple and very effective classification approach. The decision tree identifies the most important features of a given problem. The heart disease has a set of distinct values affecting the heart. It includes blood vessel disorders such as irregular heart beat issues, weak heart muscles, congenital heart defects, cardio vascular disease and coronary artery disease. Coronary heart disorder is a familiar type of heart disease. It reduces the blood flow to the heart leading to a heart attack. In this paper the available data set of the patients suffering from heart disease is analyzed. R language is used to predict the accuracy.

Keywords: *Decision Tree, Machine Learning, Random forest, R Studio*

1. Introduction

Machine Learning is the technology to improve the performance of the machines by developing efficient algorithms so that they learn by experience for a given task. Classification is one such method that makes the machines to learn. The well known procedure for classification is decision trees which has the capability to recognize and split the data into separate classes. Some of the other classification learning techniques are C4.5, ID3, boosted decision trees [1][2].

Every classification procedure or the classifier checks for the best splits using the data attributes in the data sets. The data that is weakly shown by previous classifiers are given higher priority by the succeeding classifiers. Erroneous and missing data in the data set reduction is prime objective by the classifier.

In selecting the data attribute for decision making the data set is recursively segregated into separate classes with similar types Every machine learning procedure executes into two phases the training stage and testing stage. In training phase, the decision tree is constructed based on the current examples in the training dataset. In testing stage examples are identified using the model constructed by the training dataset.

Many related works are done taking the medical data for diagnosing heart attack. Coronary heart disease (CHD) is vital heart disability in grownups that causes to death in most of the developed countries in the world. Prediction model taking C4.5 decision tree was used on the heart data set for classification [3]. A prediction model that uses combination of both pre pruning and post pruning of decision tree learning improved the classification accuracy by reducing the tree size [4]. It improved the classification accuracy in comparison with other classification learning methods such as the CART, ID3 and C4.5 decision trees. Other decision learning approaches such as the neural networks, Bayesian networks, association rule learning have their own importance. These learning techniques were compared using available data set for heart disease [5]. There advantages and limitations were made in the research study.

In this paper we taken a heart disease data set and analysis using regression tree learning. The paper is organized as follows section II discusses classification. Section III describes classification methods. Section IV briefs feature engineering. Section V deals with prediction analysis and section VI summarizes the conclusion.

2. Classification

Classification is an important procedure for machine learning. It has three forms, supervised learning, unsupervised learning and semi-supervised learning. In supervised learning process the procedure works with the group of examples whose labels are known.

The classification learning approach considers categorical values but the regression procedure takes numerical values. In the unsupervised learning method the class labels are not known in advance but are grouped into clusters as per their attribute characteristics. Semi-supervised learning utilizes both labeled and unlabelled class data.

The classification learning is normally a supervised procedure that takes an example in the data set and identifies to it to a class attribute. An example has two parts the predictor attribute values and target attribute values respectively. The predictor attribute values are used to predict the values of target attribute value. It is also used to predict the class of an example.

In the classification learning procedure the dataset is divided into two disjoint sets. The training data set and the test data set. The classification process consists of two stages. They are the training stage and the testing stage. The model is obtained by using training data set at the training stage. The testing stage uses the model on the test data set to predict the target attribute value.

The values of both predictor attributes and the target attributes are used to obtain the classification model. This model gives the relationship between predictor attribute values and classes which predicts the classes of an example. In the testing phase, the

prediction is made by the algorithm to see the actual class of just classified example. When classifying examples in the test set are unseen during training, the classifier maximizes the predictive accuracy. The knowledge learned by a classification procedure can be represented in many different ways such as the association rule learning,, decision tree learning and artificial neural network learning to name a few.

3. Classification Methods

3.1 Decision Tree

Decision tree is utilized for classification in the decision making process. It consists of two distinct nodes, the internal node and the leaf node. An internal node has the root node. The internal nodes are related to attributes, where as the leaf nodes are connected to classes. Every non-leaf node has an outgoing branch .To find the class for the new example in the data set. The search process starts at the root node. The subsequent internal nodes are covered till the leaf node is arrived. To find the right class for a leaf node, testing is done for every internal node from a given root node. The result of the test is to search the branch travelled from the root node to the leaf node visiting each and every internal node between them. The class for the example is the class of the final leaf node.

3.2 Random Forest

Random forests are the important aspect of classification. They are the group of decision trees. Bagging is used as selection parameter in these forests. This approach is used in a cycle with random forests. Random feature selection with replacement is used taking the training data set. Thus the tree grows with the new training data set. The concept of building forests is to construct many decision trees classified by example of the class .Giving a suitable vote to each class of the tree the root of the random forest is selected. Generally the decision tree class that gets the highest vote is selected as the root [5][6].Random forests are normally unpurned. They are used to predict the prediction accuracy.

4. Feature Engineering

For studying classification process data set from UCI machine learning repository for heart disease at Cleveland is taken. The dataset is divided into two sets, the test data set and the training data set. The procedure and the related feature engineering is done on the training data, and model thus obtained is utilized on the test data to predict the results.
The problem statement for the following the problem is stated as follows:"
Problem Statement: "To predict the value for the patients suffering from heart disease"
To group the historical usage patterns with heart disease predictable data in order to analyze the number of patients suffering from heart disease.
Data Set used is the "Heart disease diagnosis from the Cleveland dataset taken from UCI Machine Repository". The variables are defined as data fields as shown below.
Data attributes are,
a_age- attribute age in expressed in years
a_sex- attribute sex is expressed in male with value 1 and female with value 0.
a_cp- attribute for chest pain category is expressed with values 1,2,3 and 4 in terms of typical angina, atypical angina, non-anginal pain, asymptomatic respectively.
a_trestbps- attribute for resting blood pressure (BP) expressed in mm Hg, when the person is admitted to the hospital.
a_chol- attribute for serum cholesterol in mg/dl
a_fbs-attribute for expressing fasting blood sugar > 120 mg/dl with true and false indicated numerically by 1, 0.
a_restecg- attribute for resting electrocardiographic outcome expressed with values 0,1 for normal and S T-T wave abnormality(T wave inversions and/or ST elevation or depression of > 0.05 mV), 2= showing probable or definite left ventricular hypertrophy by Estes' criteria)
a_thalach- attribute for maximum heart rate of the patient.
a_exang-attribute for exercise induced angina indicated numerically by 1 and 0 for yes and no categorical values.
a_oldpeak-attribute for ST depression induced by exercise relative to rest
a_slope-attribute for the slope of the peak exercise ST segment expressed in terms of up sloping, flat and down sloping with values 1,2 and 3 respectively.
a_ca- attribute for count of major vessels with a range from (0-3) with flourosopy coloring.
a_thal- attribute for type of heart defect with expressed 3 for normal, 6 for fixed defect and 7 for reversable defect
a_num- attribute for predicting the patients suffering from heart disease."
The Cleveland data set containing 303 tuples is divided into 212 tuples for training set and remaining 91 tuples into test data. The sample for training is executed in R and is taken using the equation (1),

training <-sample (1:303, 212, replace=FALSE). (1)

The training data set and test data set are then calculated using equation (1).

The formula is then calculated using equation (2) below,

formula<- (a_num~a_age+a_sex+a_cp+ a_trestbps+ a_chol+ a_fbs+a_restecg+a_thalch+a_exang+a_oldpeak+ a_slope+ a_ca+ a_thal) (2)

In the equation (2) a_num is the predictor attribute, a_age, a_sex, a_cp, a_trestbps, a_chol, a_fbs, a_restecg, a_thalch, a_exang, a_oldpeak, a_slope, a_ca and a_thal are the response attributes.

The regression tree is then constructed using the equations (3).

fit<-rpart(formula, data= trainingset, method="anova") (3)

The parameters used in the rpart function of equation (3) are the formula which is utilized from equation (2), data is the trainingset calculated from equation (2) and the method is anova is used for regression analysis.

The regression tree is plotted using equation (4) below.

rpart.plot (fit, type=3, digits=3, fallen. leaves=TRUE) (4)

The regression tree is shown in Fig 1. The random forest is plotted using the equation (5).
xyz.rf=randomForest (formula, data= trainingdata) (5)

The formula and data parameters are described above in the equation (5).

4.1 Performance Measures

We have utilized three performance measures for analysis. They are the Mean Absolute Error (MAE),Sum of Squared Error(SSE) and Mean Squared Error(MSE). MAE is defined as the mean of the absolute difference between the actual and predicted values of the instances in the data set.SSE is defined as the sum of the squares of the actual and predicted values of the instances in the data set. MSE is defined as the mean of the squares of the actual and predicted values in the data set.

5. Prediction Analysis

Before prediction analyses the data is preprocessed and missing data are evaluated using mean of the attribute. The MAE, SSE and MSE are calculated for the overall heart disease data set and are listed in Table I. In the Table I the value of MAE is lower than MSE. Now analyzing Table II, we find the lowest value of MAE and MSE are 0.43 and 0.49 respectively, this occurs when a_sex is female. We also observe that SSE is lower when a_sex is female, which also supports the evidence that the model predicts with higher accuracy when a_sex is female.

Now observing Table III we see that minimum value of MAE and MSE is 0.31. This occurs when a_cp has 2 as its value. Thus the model predicts better in this case. We also observe that the highest value of MAE and MSE are 0.78 and 1.12. Thus the prediction model deviates from the actual values in this case.

Now observing Table IV we see that the lowest value of MAE and MSE are 0.43 and 0.41. This occurs when the a_slope has 3 as its value. Hence the prediction accuracy of the model is better in this case. The highest value of the MAE and MSE in Table IV are 0.72 and 1.01, this happens when the value of a_slope is 2, thus the prediction model deviates from the actual values in this case. The prediction model behaves moderately when the value of a_slope is 1.Using the tables Table II, Table III and Table IV, we find that the minimum MAE considering the attributes a_sex, a_cp and a_slope we get 0.31. The occurs for attribute a_cp and the value of attribute for this minimum MAE is 2.Thus the model predicts better for this value of the attribute a_cp. From tables Table II, Table III and Table IV, the minimum value of SSE is 0.31 for the same value of attribute a_cp. Hence the model predicts better for the attribute a_cp with value 2.

We have taken different cases for the values of some of the attributes of the heart disease data set and have observed the model prediction using regression tree.

We now proceed to random forest analysis. In the Fig 2 we observe that as the number of trees increases the error decreases. At 200 trees the error is 0.8 and at 500 trees, error is 0.7. The value of error decreases from 1.8 to 0.8.Thus the more the number of trees lesser the error in random forest.

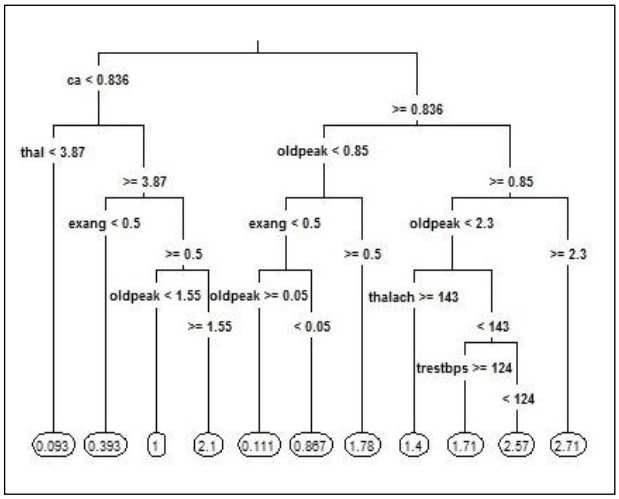

Fig 1:.Regression tree for given heart disease data set

Table I:. Mae, Sse And Mse For Overall Data Set

Error Type	Value
MAE	0.77
SSE	105.24
MSE	1.16

Table Ii: Mae , Sse And Mse For Male And Female For A_Sex

a_sex	MAE	SSE	MSE
male	0.65	173.01	0.84
female	0.43	47.07	0.49

Table Iii.: Mae, Sse And Mse For Male And Female For A_Cp

Type of Error	Value of a_cp=1	Value of a_cp=2	Value of a_cp=3	Value of a_cp=4
MAE	0.57	0.31	0.40	0.78
SSE	13.52	15.42	30.37	160.76
MSE	0.59	0.31	0.35	1.12

Table Iv. . Mae, Sse And Mse For Male And Female For A_Slope

Type of Error	a_slope=1	a_slope=2	a_slope=3
MAE	0.46	0.72	0.43
SSE	70.28	141.16	8.62
MSE	0.49	1.01	0.41

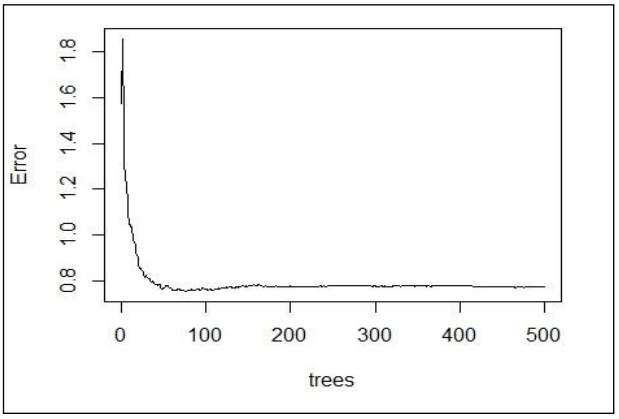

Fig 2:. Error vs. Number of Trees Graph for training data

6. Conclusion

In this paper the regression tree prediction is made for the heart disease taking the UCI machine learning Cleveland data set repository. The MAE, SSE and MSE are calculated. The MAE for the male is more than the female for the given attribute in the data set. From the analyses of regression tree we find the prediction model performs better for a_cp attribute in the data set in terms of MAE and MSE. The random forest for the given data is also analyzed. We find the error decreased as the number of trees increase. In future other techniques of machine learning such as deep learning, association rule learning and neural networks will be studied and prediction accuracy will be improved by constructing effective prediction models.

7. References

[1] Manish Varma Datla. "Bench Marking of Classification Algorithms: Decision Trees and Random Forests using R –A Case Study", International Conference on Trends in Automation, Communications and Computing Technology (I-TACT-15), Bangalore, Dec 21-22, 2015, pp.1-7.

[2] J. R. Quinlan, "Learning decision tree classifiers," ACM Computing Surveys, 28(1),Volume 28, Issue 1, March 1996, NY, USA. pp. 71-72.

[3] Minas A. Karaolis, Member, IEEE, Joseph A. Moutiris, Demetra Hadjipanayi, Constantinos S. Pattichis, "Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees", IEEE transactions on information technology in biomedicine, Vol. 14, No. 3, May 2010, pp.559-566.

[4] Ali Mirza Mahmood1, 2* Mrithyumjaya Rao Kuppa, "Early detection of clinical parameters in heart disease by improved decision tree algorithm", Second Vaagdevi International Conference on Information Technology for Real World Problems, 2010, pp. 24-29.

[5] František Babič, Jaroslav Olejár, Zuzana Vantová, Ján Paralič, "Predictive and Descriptive Analysis for Heart Disease Diagnosis", Proceedings of the Federated Conference on Computer Science and Information Systems, Prague, 2017, ISSN 2300-5963 ACSIS, Vol. 11,, DOI: 10.15439/2017F219, pp. 155–163.