International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Predictive Analysis of Heart Disease using Stochastic Gradient Boosting along with Recursive Feature Elimination

V Kakulapati[1], Ankith Kirti[2], Vaibhav Kulkarni[3], Charan Pandit Raj[4]

Sreenidhi Institute of Science and Technology, Yamnampet, Ghatkesar, Hyderabad -501301

Abstract: ***Coronary heart disease (CHD) is an illness in which plaque constructs up inside the coronary arteries.CHD descriptions for over 15.9% of all deaths creates it the most regular origin of death globally. Health professionals are facing tough to anticipate the heart alignment as it is capable of medical practitioners that require experience and knowledge. Over the past few years machine learning has proved to be a very successful tool in clinical diagnosis which take up the huge amount of data. This contains hidden information that can be used effectively in making informed decisions. The investigation such type of data utilizes maximum time in terms of execution and utilization of resources. Data features do not support for the outcomes. Therefore, it is especially significant to recognize the features that add further in recognizing diseases. The aim of this work is to predicting the heart disease stage level of a patient by employ machine learning algorithms. In this regard we used the stochastic gradient boosting algorithm along with Recursive Feature Elimination (RFE) for selecting the best features in the data.***

Keywords: predict, heart attack, hidden, stochastic, gradient, recursive.

1. Introduction

In the current situation there is various precise innovations give knowledge to specialists in taking clinical decisions, however, they are not yet to be correct. Given the success of machine learning in the medical field [1] in the past few years we felt that if harnessed effectively the power of Machine Learning (ML) can help doctors in making informed decisions. Heart disease prediction system can provide valuable insights to medical professionals in making decisions about the state of heart of patients. Medical professionals may neglect to take exact decision [2] while diagnosing the heart alignment of a patient, in this way a heart alignment prediction method which utilize machine learning algorithms aid such cases to get precise outcomes. There are many devices accessible which utilize prediction algorithms, but they have few deficiencies. Many tools cannot deal with huge information and most are not incorporated, not deployed on the web and subsequently not available on the online. There are numerous medical facilities and medicinal services which gather data of patient details which becomes hard to deal with present existing frameworks [3].

Heart disease prediction framework can help medicinal experts in anticipating heat alignment in view of the clinical information of patients [4]. Subsequently by implementing a heart alignment prediction framework utilizing machine learning algorithms and doing some kind of investigation on different heart related issues, it can have the capacity to predict the more probabilistically that the patients will be diagnosed with heart problem patients. We choose the Stochastic Gradient boosting algorithm. It is one of the most powerful classification algorithms. It constructs a calculation model in the form of an ensemble of weak prediction models, typically decision trees. It constructs the representation in a stage-wise approach similar to further boosting methods [5] do, and it simplifies them by allowing optimization of a subjective differentiable failure task.

2. Related Work

Coronary Heart diseases (CHD), which is also referred to as Coronary artery disease (CAD) [6] has been a predominant and a persistent problem in the field of medical diagnosis. Like every muscles in human body, human heart needs a constant supplying of oxygen and supplements, these are conveyed to it by the blood in the coronary veins. In the event that these coronary conduits end up plainly limited or stopped up because of the development of the plaque and neglect to supply enough blood to the heart, it brings about CHD. If the blood [7] carrying the oxygen and nutrients does not reach the heart, the heart may respond with pain called angina. The soreness is usually felt in the chest or sometimes in the human left side. The supply of blood is completely cut-off outcomes with a heart attack.CHD can be caused due to a wide range of factors and it requires complex medical tests and a medical practitioner with knowledge and experience to diagnose it. Given the success of machine learning in the medical field in the past few years we felt that if harnessed effectively the power of Machine Learning(ML) can help doctors in making informed decisions.ML is able of automating manual procedures conceded out by practitioners, which are generally time-consuming and subjective.

The utilization of machine learning can accumulate time for practitioners and afford unbiased, repeatable outcomes. A lot of researchers have previously worked on this problem. W. Nor Haizan, et. al,[8][9] performed classification on the basis of clustering. They performed a comparative study using the pruning method in decision tree algorithms. The paper delves deep into the REPTree decision tree technique. Chosen a few attributes which have well-built relationship in the medical dataset and recommended a system which facilitates patients avoid auxiliary health check up. Aqueel Ahmed and Shaikh Abdul Hannan performed data mining on heart disease dataset using SVM and rough set techniques based on association rules and clustering [10] .From the

above one can understand that various classifiers are used for medical data analysis in an effective way. Therefore we decided to use Stochastic Gradient Boosting algorithm as a classifier.

3. Dataset

The dataset we used is a part of a large collection of datasets of heart disease from the large repository of UCI. We utilized the Cleveland dataset; the dataset has 14 variables. Those are

- age: age in years
- sex: sex (1 = male; 0 = female)
- cp: chest ache category
  Assessment 1: typical angina
  Assessment 2: atypical angina
  Assessment 3: non-anginal pain
  Assessment 4: asymptomatic
- trestbps: remaining blood pressure (in mm Hg on admit into the hospital)
- chol: serum cholesterol in mg/dl
- fbs: (pasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

- respiteecg: remaining electrocardiographic outcomes
  Assessment 0: common
  Assessment 1: having ST-T signal irregularity (T signal inversions and/or ST altitude or gloominess of > 0.05 mV)
  Assessment 2: proving feasible or specific left ventricular hypertrophy by Estes criteria
- thalach: maximum heart rate achieved
- exang: exercise induced angina (1 = yes; 0 = no)
- oldpeak = ST depression induced by exercise relative to rest
- slope: the slope of the peak exercise ST segment
- ca: number of most important containers (0-3) colored by fluoroscopy
- thal: 3 = normal; 6 = fixed defect; 7 = reversible defect
- num: diagnosis of heart disease (angiographic disease status)
  -- Value 0: < 50% diameter narrowing
  -- Value 1: > 50% diameter narrowing
  (in any most important container: features 59 through 68 are containers)

The "num" meadow pass on to the presence of heart ailment in the patient. It is a factor of classes 'Yes' and 'No'.

4. Our Method

The heart ailment information contains the screening clinical statistics of heart patients. The package we used extensively in this system is the caret [6] package (short for Classification and Regression Training) .This package provides wide range of algorithms for both categorization and regression; it has variety of feature selection techniques. Initially, the data set is pre-processed to construct the learning procedure well-organized. In the early phase of our proposed study, we used preprocessing in order to normalize the data and handle missing values. Then we used a process called Recursive Feature Elimination for selecting the most useful attributes from the information. Finally, data split into training and a test set and then utilized Stochastic Gradient boosting algorithm to prepare a representation.

5. Implementation Result

Understanding the data

To understand the dataset we used various methods in R language, the head () function to which is used to look at the first few rows of data. To review the dimensions of our data we used the dim () function. And we reviewed the distribution of our data with the summary () function. Visualizing the data: We used the barplot () function to create bar plots of all the attributes of our data.

Pre-processing of data

As the analysis revealed there were missing values in our data, so the first step was to address this issue. We did this using the mean () function. The mean values of each column ('thal' &'ca') were used to fill up these missing spots.

To rework the statistics caret bundle affords the pre-process () characteristic that takes a technique argument to suggest the form of pre-processing to carry out. To this function we pass two arguments 'Center' and 'Scale'. The middle rework calculates, implies for an attribute and subtracts it from every fee. The scale change ascertains the standard deviation for a trait and partitions each incentive on that standard deviation.

Once this is done we split the dataset into training and test set. We tried out various split ratios but the ratio that gave us the best results was 92.5 % in the training set and the remaining in the test set.

Feature selection

To select the most useful features we used a process called Recursive Feature Elimination (RFE). It is a mainstream programmed technique for feature determination given by the caret R package is called Recursive Feature Elimination or RFE.A Random Forest calculation is utilized on every cycle to assess the model. The algorithm utilizes 10-fold Cross Validation. In this, the first informational collection is separated into 10 subsets, the model preparing and testing is rehashed 10 times. Each time, one of the 10 subsets is utilized as the test set and rest are assembled to frame a preparation set. When this is done the normal blunder over each of the 10 trials is ascertained. The benefit of this strategy is that it doesn't rely on upon how the information gets isolated. Each subset of the dataset gets the chance to be in a test set precisely once, and gets the chance to be in a preparation set 9 times. The change on the subsequent gauge is lessened as the quantity of folds is expanded. The inconvenience of this strategy is that the preparation calculation must be a rerun starting with no outside help 10 times, which implies it takes 10 fold the amount of calculation to make an assessment. Out of the13 attributes the algorithm reduced the attributes to just 3 attributes of most significant to dependent variable (num).

```
RFEControl (functions=rfFuncs, tech-
        nique="repeatedcv",
          Quantity=10)
```

The outcomes of the RFE algorithm are the variables.

- cp: chest ache category,
- ca: quantity of chief containers (0-3),
- Thal: 3 = regular; 6 = attached deficiency; 7 = reversible deficiency.

The outcomes of the RFE process:

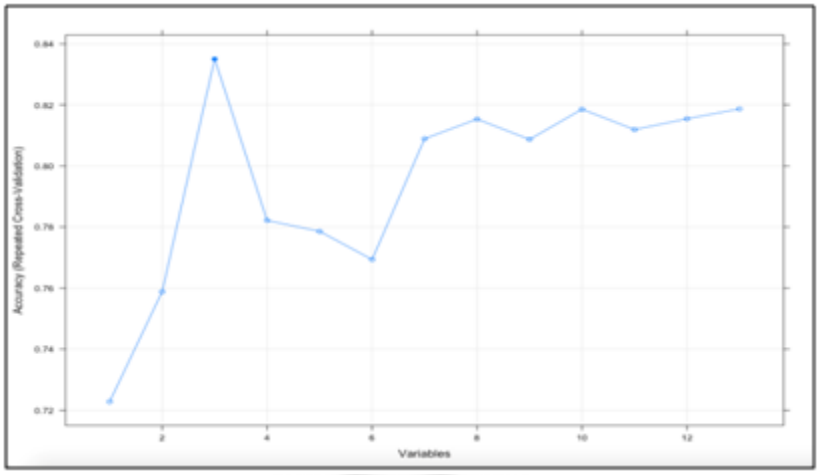

Figure 4.1: Cross validation for variables.

```
       cp              ca              thal          num
 Min.   :1.000   Min.   :0.0000   Min.   :3.000   No :164
 1st Qu.:3.000   1st Qu.:0.0000   1st Qu.:3.000   Yes:139
 Median :3.000   Median :0.0000   Median :3.000
 Mean   :3.158   Mean   :0.6766   Mean   :4.736
 3rd Qu.:4.000   3rd Qu.:1.0000   3rd Qu.:7.000
 Max.   :4.000   Max.   :3.0000   Max.   :7.000
```

Figure 4.2: Mean and median for the reduced dataset

Spot checking

This algorithm is about in receipt of a speedy assessment of a group of different algorithms on the machine learning problem so we can compare the results of our algorithm to others. Spot-checking algorithm is a technique utilized in applied machine learning. We perform test eight different algorithms and the preeminent results were producing by our enhanced algorithm

Train the model

The algorithm which we implementing in this work is Stochastic Gradient boosting algorithm. This is one of the all the more intense techniques for building predictive model. The presentation of Gradient boosting Friedman expected a slight adjustment to the calculation, enlivened by Breiman's stowing strategy. Friedman suggested that at each cycle of calculation, a construct learner ought to be fit in light of a subsample of the preparation set drawn aimlessly by replacement (Cross-validation). A generous change in slope boosting's precision with this adjustment. Subsample size is some steady part f of the span of the preparation set. The calculation turns out to be substantially quicker as relapse trees must be fit to littler datasets at every emphasis. Friedman acquired that prompts great outcomes for little and direct estimated preparing sets. We developed a improved version of the Gradient boosting algorithm by slight modification of the original algorithm.

```
Control<-trainControl                    (me-
thod="repeatedcv", number=10,
Repeats=30)
```

To tune the performance of our algorithm we used a tuneGrid.

```
gbmGrid <- enlarge.grid (interface.intensity
= c(1,3,5),
n.trees = (1:20)*50,
shrinkage = 0.001,
n.minobsinnode = 50)
```

We tested the interaction depth at three levels 1, 3 and 5. The quantity of trees between 50 to 1000, shrinkage at 0.001 and the minobsin node at 50.

This algorithm also uses repeated Cross-validation.

```
Tuning parameter 'shrinkage' was held constant at a value of 0.001
Tuning parameter 'n.minobsinnode' was
 held constant at a value of 50
Accuracy was used to select the optimal model using  the largest value.
The final values used for the model were n.trees = 1000, interaction.depth = 3, shrinkage = 0.001
 and n.minobsinnode = 50.
```

Figure 4.3: Level wise algorithm

Predictable result

We used a variable to store the predicted values by the GBM model and then created a confusion matrix. A confusion matrix is used to explain and analyse the presentation of a categorization representation, or a classifier on a set of investigation information for which the accurate values are acknowledged. The confusion matrix is proportionally simple to understand, but the related vocabulary can be

```
pred = predict(fit.gbm2,newdata=testing[1:3]
ConfusionMatrix (pred, testing$num)
```

```
Confusion Matrix and Statistics

          Reference
Prediction No Yes
       No  11   0
       Yes  1  10

               Accuracy : 0.9545
                 95% CI : (0.7716, 0.9988)
    No Information Rate : 0.5455
    P-Value [Acc > NIR] : 3.126e-05

                  Kappa : 0.9091
 Mcnemar's Test P-Value : 1

            Sensitivity : 0.9167
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9091
             Prevalence : 0.5455
         Detection Rate : 0.5000
   Detection Prevalence : 0.5000
      Balanced Accuracy : 0.9583

       'Positive' Class : No
```

Figure 4.4: Confusion matrix for performance analysis of classification

5. Conclusion

We designed a system which efficiently integrates various aspects of machine learning and produces an accurate result. We achieved classification accuracy 95.45%. The use of the machine learning algorithms gives dependable and precise results in the field of healthcare, stochastic gradient boosting algorithm attained maximum accuracy in terms of performance as a powerful classifier, and the least execution time, and also used for the future prediction of classification of medical field.

6. Future Enhancement

For more precisely predictive analytics on heat disease with comparative analysis of numerous efficient machine learning algorithms. And we will analyse online patient data to give better diagnosis. This system can be integrated with a large volume of data to make it more users' friendly and accurate results.

References

[1] https://medicinex.stanford.edu/program-2015.
[2] Prerana T H M et al ., "Prediction of Heart Disease Using Machine Learning Algorithms- Naïve Bayes, Introduction to PAC Algorithm, Comparison of Algorithms and HDPS", Volume 3, Number 2 – 2015, ©IJSE. ISSN: 2347-2200, PP: 90-99.
[3] A systematic analysis for the Global Burden of Disease Study 2015.". *Lancet (London, England)*. https://www.ncbi.nlm.nih.gov/pubmed/27733281
[4] Aswathy Wilson et al., "Heart Disease Prediction Using the Data mining Techniques" International Journal of Computer Science Trends and Technology (IJCST) – Volume 2 Issue 1, Jan- Feb 2014.
[5] https://en.wikipedia.org/wiki/Gradient_boosting.
[6] https://patient.info/health/the-heart-and-blood-vessels.
[7] https://quizlet.com/762158/health-flash-cards
[8] W Nor Haizan W et al., "A Comparative Study of Reduced Error Pruning Method in Decision Tree Algorithms" 2012 IEEE International Conference on Control System, Computing and Engineering, 23 - 25 Nov. 2012, Penang, Malaysia.
[9] Franck Ohlhorst, January 2013' Big Data Analytics: Turning Big Data into Big Money', ISBN: 978-1-118-14759-7, pp 176.
[10] C. S. Dangare and S. S. Apte, "Improved study of heart disease prediction system using data mining classification techniques," International Journal of Computer Applications, vol. 47, no. 10, pp. 44–48, 2012.
[11] Ashfaq Ahmed K, Sultan Aljahdali and Syed Naimatullah Hussain, (2013) "Comparative Prediction Performance with Support Vector Machine and Random Forest Classification Techniques", International Journal of Computer Applications Vol. 69, No.11, pp 12-16