# Empirical Evaluation of the Performance of Feature Selection Approaches on Random Forest

Smitha S Kumar
School of Mathematical and Computer Science
Heriot Watt University
Email:smitha.kumar@hw.ac.uk

Talal Shaikh
School of Mathematical and Computer Science
Heriot Watt University
Email:t.a.g.shaikh@hw.ac.uk

*Abstract*—**Medical data contain very valuable information which can save many lives if it is analyzed and utilized efficiently. Efficient analysis of this large volume of data demands the right choice of predictors and this in turn can impact the accuracy of the decision support system. Dimensionality reduction and feature subset selection are two techniques to reduce the number of features used in classification. In this paper we perform an empirical evaluation of four feature selection methods when applied in conjunction with Random Forest classifier. The feature selection techniques applied are Relief feature selection algorithm, Random forest selector, Recursive feature elimination and Boruta Feature selection algorithm. Results show that feature selection methods boosts the performance of the classifiers and in this case the features selected by the Boruta feature selection algorithm gives the best results.**

***Keywords—Data mining;Classification;Heart disease;Random Forest;Feature selection***

## I. INTRODUCTION

According to World Health Organization (WHO) cardiovascular disease including heart disease and strokes are the leading cause of death in the world [1]. Clinical database have massive information about patients and their medical conditions. Datamining can uncover new knowledge for decision making from this massive medical data. Identifying high risk patients, Hospital infection control, Ranking Hospitals are some of successful mining applications in the healthcare arena [2]. Several data mining techniques are used in heart disease detection, such as Naïve Bayes, Decision Tree, Neural Networks, Support Vector machines and Random Forest [3].

In processing medical data it is important that we select the best set of attributes which improves the performance of the model, increases the computational efficiency, decrease the storage requirements. Most of the time the data set includes a lot of features with different qualities which can influence the performance of the classifiers. Noisy features can affect the performance of the classifier. The reduction of original feature set to a smaller one preserving the relevant information while discarding the redundant one is referred to as feature selection [FS] [4].

This paper is organized as follows. Section II describes the various feature selection approaches and a brief overview of the related work in the heart disease detection. Section III focuses on the research methodology. The testing platform, implementation details and results are discussed in Section IV followed by Section V which includes conclusion.

## II. LITERATURE REVIEW

Researchers have carried out a lot of experiments in applying various data mining techniques in heart disease prediction. Zriqat, I et al. (2016) conducted a study on the performance of various data mining algorithms including decision tree. Discriminant, Random Forest, and Support Vector Machine. Results show classification algorithms can effectively diagnose cases of heart disease and decision tree outperforms the other classifiers [5]. Zhang et al. (2013) applied decision tree to identify patients with heart failure [6].

The motivation to pursue this research is to investigate the effect of feature selection on ensemble methods, in particular Random forest classifier. Random Forest, an ensemble –based method (supervised classification algorithm) builds multiple decision trees based on random samples and averages the result obtained from those decision trees (majority voting). Thus it tends to be less prone to over fitting. It is one of the most widely used machine learning algorithm for classification. The two important parameters used in the Random Forest classifier are ntree(number of trees) and mtry(the number of features to choose the best subset). The number of attributes randomly chosen is sqrt(M), if M is the total number of attributes. Fig1.shows an example of majority voting in random forest classifier.

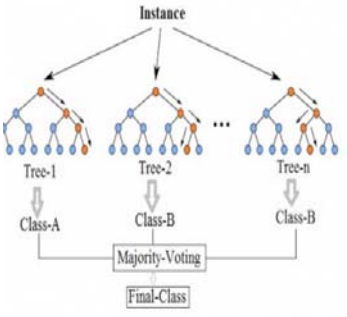


Fig. 1. Random Forest Classifier (from community.tibco.com)

According to Langley [1994][3] , the FS method is based on four main steps.

1. Generation procedure
2. Evaluation function;
3. Stopping criterion
4. Validation procedure.

Generation procedure deals with generating subsets from the given set of attributes. If the set is of size N, there are $2^N$ subsets. So if the N value is large, the task of testing and training the attribute also becomes harder. There are multiple strategies such as forward strategy (starts with an empty set and then add a new feature at each iteration), backward strategy (start with all attributes and remove one at each iteration). Evaluation function measures the quality of the subset. Stopping criteria is used to avoid time consuming search operations done on the attribute space. Validation procedure checks the performance of the DM algorithms using the selected subset of attributes [4].

The feature selection approaches are classified into four categories, filter, wrapper, embedded and hybrid. Filter methods (Independent methods) generate subsets and evaluate the subset variable without model construction. In Wrapper approaches the subset is generated, and is evaluated based on performance of the model. Embedded approach suggests a subset of variables and it is evaluated during the construction of the model. Hybrid methods are a combination of the filter and wrapper methods [4].

Tan et al. (2009) proposed an approach with Genetic Algorithm and Support Vector machine (SVM). Analyses show that this hybrid approach generated good classification accuracy when compared with the existing approaches [7].

Dag, H et al. (2012) analyzed three feature selection algorithms ((Information Gain, Gain Ratio and Correlation Based Feature Selection (CFS)). Results show that the three feature selection algorithms produce almost the same subset of attributes differing in the order of the attributes slightly. The performance of J48 was more or less the same with the three subsets [8].

### *A. Relief Feature Selection Algorithm*

This is a feature selection algorithm inspired by instance base learning. The original RELIEF algorithm works only with a two-class problem. This algorithm is extended to deal with noisy, incomplete and multi-class data set. The pseudocode of the Relief algorithm is given below [9].

Require**:** $\{I_1 \ldots I_n\}$ –dataset of $n$ examples comprised of $a$ features. Each example belong to one of $c$ classes
$m$ – number of examples to be selected by the algorithm
$p_c$ – prior probability distribution of classes in the data set
$k$ – number of neighbors from each class to be selected by the algorithm
$Wj$: =0, $j \in [1 \ldots a]$ – weights of individual attributes
$diff(a, I_i, I_j)$ – differences between values of the attributes $a$ between $I_i$ and $I_j$

```
for l := 1 … m do
  Ir := randomly selected example from {I1…In}
  cir – class of the selected example
  for t :=1 …c do
      Mt= k closest examples from class t
       for h:=1…a do
         for IM ∈Mt do
          if cir = t then Wh = Wh – diff(h, Ir, IM) / m k
          if cir ≠t then Wh =Wh + pt/(1-pt) × diff(h, Ir, IM) /mk
         end for
       end for
  end for
end for
```

### *B. Recursive Feature Elimination*

This algorithm ranks the predictors and the removes the least importance ones sequentially. The goal is to find the best subset which can design an accurate model. It does a backward selection of predictors based on the ranking done. The pseudocode of the Recursive Feature Elimination (RFE) is given below [10].

1. for Each Resambling Iteration do
   Partition data into training and test/hold-back set via resampling
   Tune/train the model on the training set using all predictors
   Predict the held-back samples
   Calculate variable importance or ranking
   for each subset size $S_i, i=1 \ldots S$ do
      Keep the $S_i$ most important variables
      [Optional] Pre-process the data
      Tune/train the model on the training set using Si predictors
      Predict the held-back samples
      [Optional] Recalculate the ranking for each predictor
   end
  end
2. Calculate the performance profile over the Si using the held-back samples
3. Determine the appropriate number of predictors
4. Estimate the final list of predictors for the final model
5. Fit the final model based on the optimal Si using the original training set.

### *C. Boruta Feature Selection Algorithm*

Boruta Algorithm is a feature selection implemented as wrapper algorithm around random forest which works well with classification and regression problems. This method removes attribute iteratively.

The Boruta algorithm consists of following steps [11]:
1. Extend the set(information system) by adding copies of all variables

2. Remove the attribute correlations by shuffling with the response.
3. Run the random forest classifier on the extended information system and calculate the Z scores.
4. Calculate the maximum Z score among shadow attributes (MZSA), and assign a hit to the attribute that scored better than MZSA.
5. For each attribute with undetermined importance perform a two-sided test of equality with the MZSA.
6. The attributes which have importance lesser than MZSA is marked as 'unimportant' and permanently remove from the information system.
7.The attributes which have importance higher than MZSA as 'important'.
8. Remove the shadow attributes.
9. Repeat the procedure until the importance is assigned for all the attributes or the algorithm has reached the previously set limit of the random forest runs.

## III. RESEARCH METHODOLOGY

The dataset used for this experiment is acquired from the UCI machine learning repository [12]. The Cleveland database includes 303 observations and 13 attributes. The dependent variable is the attribute num categorized as diagnosed with heart disease or without heart disease

TABLE I. CLEVELAND DATA SET

| | Attribute | Description |
|---|---|---|
| | Age | age in years |
| | Sex | (1 = male; 0 = female) |
| | Cp | Chest pain type {1,2,3,4} |
| | Trestbps (mmHg) | Resting blood pressure |
| | Chol (mg/dl) | Serum cholesterol |
| | Fbs | Fasting blood sugar {1,0} |
| | Restecg | Resting electrocardiographic results |
| | Thalach | Maximum heart rate achieved |
| | Exang | Exercise induced angina |
| | Oldpeak | ST depression induced by exercise relative to rest |
| | Slope | The slope of the peak exercise ST segment |
| | Ca | Number of major vessels (0-3)colored by flourosopy |
| | Thal | Heart status 3 = normal; 6 = fixed defect; 7 = reversible defect |
| | num | Diagnosis of heart disease {predicted attribute} |

Some of the instances included missing values. These instances are removed from the original data set. The number of observations after data pre-processing is 299. In our experiment, we evaluated the performance of feature selection approaches using the statistical package R. The caret (Classification and Regression Training), random Forest, FSelector, DoParallel, Boruta are the packages used in implementing the experiment [13].

The dataset is partitioned into training and test data (70% and 30% respectively).
To avoid overfitting 10-fold cross-validation is applied on the train data. Train data is used to build the classifier and the test data is used to validate the model.
The proposed approach is described below:

Load the data set
Apply the feature selection technique on the train data set.
Use this subset of features to train the classifier
Test the result on the test data

## IV. RESULTS

The performance of Random Forest classifier is evaluated with the 13 attributes. randomForest() function in randomForest package is used to create the model using the training data set .The default values are used for the number of trees (500) and the number of variable tried at each split (3). Fig 2. shows the feature importance, in relation to MeanDecreaseAccuray. MeanDecreaseAccuracy measures the impact of each feature on the accuracy of the model.

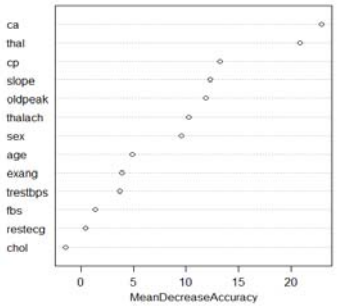


Fig. 2. MeanDecreaseAccuracy

The RandomForest-feature selector approach uses the top features identified by the Random Forest classifier to train the model. The features selected include ca, thal, cp,slope, oldpeak, thalach, sex,age. The model accuracy is 82.02%.

Recursive feature elimination is done using the rfe() function of caret package.

```
control<-
rfeControl(functions=rfFuncs,method="cv",number=10)
featurelist<-rfe(traindata[,1:13],
traindata[,14],sizes=c(1:13),rfeControl = control)
```

The model accuracy based on this approach 84.2%. The features identified by this approach are ca, thal, cp,slope, oldpeak, thalach, sex,age,exang. Fig3. shows the variable importance based on RFE.

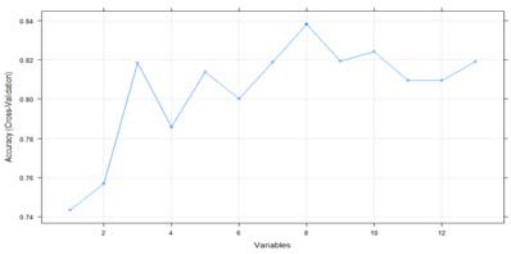

Fig. 3. Variable Importance-RFE

An implementation of Relief () function of FSelector package in R is used .The sample size is 20 and the neighbours.count=5 is applied. The cuttof weight applied is 5.
weights <- relief(num~., traindata, neighbours.count=5, sample.size = 20)
subset <- cutoff.k(weights, 5)
The top 5 features listed by this approach is cp, sex, restecg, ca, thal. The model accuracy is 82%.

In feature selection with the Boruta algorithm, the features identified by this algorithm are age, sex, cp, thalach, exang, oldpeak, slope,ca,thal. The variable importance is shown in the below graph Fig 4.

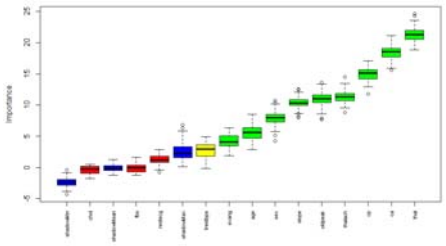

Fig. 4. Variable Importance-Boruta Algorithm

Colour codes for attribute decisions, respectively Confirmed (Green), Tentative(Yellow), Rejected(Red) and shadow. Boruta performs 99 iterations in 8.303525 secs.

In order to evaluate the prediction rate, we use a confusion matrix to obtain performance measures such as accuracy, sensitivity, specificity, Kappa. Accuracy is calculated as number of correct predictions divided by total number of observations. Sensitivity is calculated as number of true(correct) positive predictions divided by total count of positive predictions. Specificity is the number of correct negative predictions divided by the total count of negative predictions. Kappa is a metric which compares the actual accuracy with an expected accuracy.

Confusion matrix provide by the caret package is used to generate the confusion matrix. Table 2 list the metrics related to the performance of various approaches.

TABLE II. MODEL EVALUATION

| | **Sensitivity** | **Specificity** | **Kappa** | **Accuracy** |
|---|---|---|---|---|
| Random Forest-Original data set | 0.8333 | 0.7805 | 0.6149 | 0.809 |
| Random Forest Selector Approach | 0.8542 | 0.7805 | 0.6369 | 0.8202 |
| RF-Recursive Feature Selection | 0.8542 | 0.8293 | 0.6834 | 0.8427 |
| RF-Relief | 0.8958 | 0.7317 | 0.6343 | 0.8202 |
| RF- Boruta- features including Tentative feature | 0.8333 | 0.8293 | 0.6614 | 0.8315 |
| RF- Boruta-confirmed features | 1 | 0.9756 | 0.9773 | 0.9888 |

## V. CONCLUSION

Based on the investigated methods, it is found that all the feature selection approaches are beneficial for boosting the performance of the learning algorithm. Results show that Relief and Random Forest Selector approach based approach when used with Random forest generated the same accuracy (82%). Boruta algorithm with Tentative features included in the feature list achieves an accuracy which is close to the accuracy of Recursive feature elimination approach (84%). The algorithm outperforms recursive feature elimination when the confirmed 9 features are used and produces an accuracy of 98%. These are ca, thal, cp,slope, oldpeak, thalach, sex,age. Results show that the four feature selection approaches produce almost the same subset of attributes even though the orders are slightly different.Future work includes applying the feature selection approaches with other ensemble methods, applying evolutionary computing for feature selection phase and analyse the results.

## REFERENCES


[1] World Health Organization(2016,Sept.22). *Cardiovascular Disease.* [Online] Available: http://www.who.int/cardiovascular_diseases/en/

[2] Obenshain, M. (2004). Application of Data Mining Techniques to Healthcare Data. Infection Control & Hospital Epidemiology, 25(8), 690-695. doi:10.1086/502460

[3] Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg. (2013). Data Mining in Clinical Decision Support Systems for Diagnosis, Prediction and Treatment of Heart Disease. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 2(1), 218-223.

[4] [Langley, 1994] P. Langley. Selection of relevant features in machine learning. In Proceedings of the AAAI Fall Symposium on Relevance. AAAI Press, 1994.

[5] Zriqat, I., Altamimi, A., & Azzeh, M. (2017). A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods.

[6] Zhang, Goode, Rigby, Balk, & Cleland. (2013). Identifying patients at risk of death or hospitalisation due to worsening heart failure using decision tree analysis: Evidence from the Trans-European Network-Home-Care Management System (TEN-HMS) Study. International Journal of Cardiology, 163(2), 149-156.

[7] Tan, Teoh, Yu, & Goh. (2009). A hybrid evolutionary algorithm for attribute selection in data mining. *Expert Systems With Applications, 36*(4), 8616-8630.

[8] Dag, H., Sayin, K., Yenidogan, I., Albayrak, S., &Acar, C. (2012). Comparison of feature selection algorithms for medical data. *Innovations in Intelligent Systems and Applications (INISTA), 2012 International Symposium on,* 1-5.

[9] Čehovin, L, &Bosnić, Z 2010, 'Empirical evaluation of feature selection methods in classification', Intelligent Data Analysis, 14, 3, pp. 265-281, Business Source Premier, EBSCOhost, viewed 23 May 2017

[10] Khun.Max.(2016,Nov.29). *The caret package*, *Recursive feature elimination incorporating resampling*,[Online] Available: https://topepo.github.io/caret/recursive-feature-elimination.html#resampling-and-external-validation

[11] Miron B. Kursa, & Witold R. Rudnicki. (2010). Feature Selection with the Boruta Package. Journal of Statistical Software, 36(11), 1-13.

[12] UCImachine learning Repository. *Heart Disease Data Set,[Online]* Available: http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[13] R Development Team(n.d), *The R manuals* [Online] Available: https://cran.r-project.org/manuals.html