Detecting Heart Abnormality using ECG with CART

Paurakh Rajbhandary
Department of Electrical Engineering
paurakh@stanford.edu

Benjamin Zhou
Department of Mathematics
bzhou2@stanford.edu

Gaspar Garcia Jr
Department of Computer Science
gaspar09@stanford.edu

Abstract

Cardiovascular disease (CVD) is the leading cause of global deaths. Electrocardiogram (ECG or EKG) is the most widely used first line clinical tool for checking electrical activity in the heart. Hence using ECG recordings to automatically identify arrhythmias accurately and efficiently can be an important tool for cardiologists. We use the UC Irvine (UCI) Machine Learning Repository containing an arrhythmia data set to implement a multinomial classification for different types of heart abnormalities. We show that a decision tree learning algorithm (already widely used in many medical diagnostics) is well suited for this application. We achieved 80% classification accuracy with our decision tree, which is relatively high compared to other studies.

I. Introduction

In the US alone, more than half a million people die of heart disease, accounting for 1 in every 4 deaths. About 720,000 Americans have a heart attack, of which 515,000 are a first heart attack, and 205,000 are recurrent. Coronary heart disease alone costs the US $108.9 billion each year. The impact of heart disease on lives and the cost of healthcare is a growing concern.

Electrocardiogram (ECG) is one of the first line tests cardiologists use to check for problems with electrical activity of the cardiac muscles of patients. ECG is also sometimes performed as part of a physical examination, and is portable, making its use for pre-diagnosis of heart abnormalities favorable. Hence, it is of great interest to be able to accurately predict arrhythmia using patient ECG data.

There has been much previous research on arrhythmia classification. Given the large dimensional size of arrhythmia data features, one approach involved transforming the features to a lower dimension using Principle Component Analysis, and then applying Support Vector Machines (1). Another approach involved analyzing the performance of a Naive Bayes learning algorithm given varied learning times (2). The motivation behind previous work has been centered on developing a model that can perform non-invasive risk assessment of arrhythmia in a patient. Using machine learning principles, patient information can be analyzed to determine the features most indicative of arrhythmia.

However, to our knowledge, there has not been any published result in arrhythmia classification using CART (Classification and Regression Tree) analysis. Our goal is to accurately predict different types of arrhythmia in patients. Intuitively speaking, since doctors infer some medical condition (children nodes) based on symptoms (parent nodes), we decided to design a classifier based on a tree structure, which graphically makes biological sense. Thus, we present an arrhythmia classifier based on CART and Decision Tree Analysis.

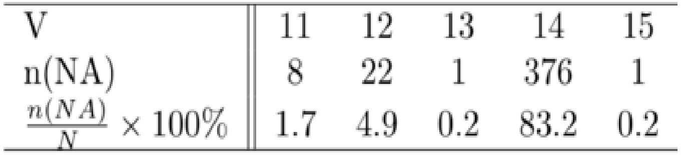
II. Methods

1. *Data Set*

We used the UC Irvine Machine Learning Repository (9) which has a data set containing arrhythmia information for 452 patients (rows). Each of these patients has 279 features (columns), and was classified into one of 16 categories (15 abnormal and 1 normal heart conditions). Our feature space comprised of 279 dimensions, including patient information such as age, sex, height, the PQRST wave signal, and channel information.

2. *Imputation*

In the data set, there were many missing or NA values. When we observe the numbers of missing data for each feature, we find the following.



V	11	12	13	14	15
n(NA)	8	22	1	376	1
$\frac{n(NA)}{N} \times 100\%$	1.7	4.9	0.2	83.2	0.2

Table 1. Frequency of NA values in features 11-15 (2nd row). Percentage of patients with missing values for each feature 11-15 (3rd row)

We saw that 376 patients, 83.2% of all 452 patients, had a missing value for feature 14. Since this is a significant proportion of the data, we chose to omit feature 14, since it was mostly NA values and offered no real training benefit. For the remaining NA values, which were in features 11, 12, 13, and 15, we had to choose to either (a) remove all patients (~30 of them) with these missing values, or (b) impute the missing data. Since ~30 patients (5% of total patients) had missing values, removing these patients could lead to significant information degeneration.

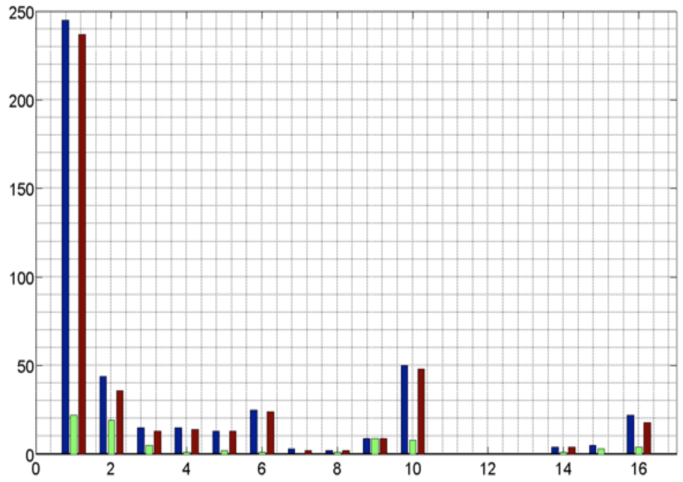


Table 2. Histogram of Classification Frequencies

This histogram compares the frequencies of each classification after removal of patients with NA values (a) and imputation (b). The majority of the patients were classified as 1 or having normal heart conditions. There were no classifications from 11-13. The blue columns represent the frequencies of each classification with imputed data. The red columns represent the frequencies of each classification after removing ~30 patients with NA values. The green columns represent the frequencies of each classification after removing all patients with feature 14 missing. We notice that the frequency for each classification is less for (a) than (b). If we examine more closely, we see that class 15 has positive frequency after imputation, but zero frequency after removal of patients. Therefore, training on data after removal of the ~30 patients would give us less leverage in predicting category 15. Thus, we instead imputed the data for the missing features using the R package rpart (3).

3. *CART Model*

A decision or classification tree represents a multi-stage decision process, where a decision is made at each node. When it is at some node D in the building process, it asks: which feature at D would give an optimal split to the children of D? Mathematically, the CART solves the following optimization problem for each node as follows (11). Given a predetermined value x_j^R, called the scalar splitting value for x_j ,the problem is to find a feature x_j, where x_j exists in the feature space, that optimally maximizes the separation of the data at node D. Here, P_l refers to the fraction of points that will be partitioned to the left child of D, and similarly P_r is the fraction partitioned to the right child. t_p is the parent node, t_l is the left child and t_r is the right child, i denotes the impurity function used to calculate impurity at the given node. We want to choose our features in the tree that minimize the impurity at each node. The objective function below maximizes the decrease in impurity from the parent node and its children.

$$\underset{x_j \leq x_j^R,\ j=1,\ldots,M}{\arg\max} \quad [i(t_p) - P_l i(t_l) - P_r i(t_r)]$$

Figure 1. Optimization Objective Function for CART

3.1 *Building the tree*

When building the classification tree, decisions that led to a compact tree with few nodes were preferred. Using the R package rpart (3), we built the decision tree using an initial complexity parameter of 0.0001. The complexity parameter controls the size of the tree, and is a way for the CART algorithm to know when to stop partitioning into subset trees. This parameter allowed us to prune the tree to prevent overfitting.

4. *Evaluation metric*

To evaluate our tree accuracy, we randomly split our data set into training data and test data in three ways, 50% training data and 50% test data, 70% training data and 30% test data, 100% training data and 100% test data. In each case, we trained our tree on the training data and tested on the test data. We also performed 10 fold cross validation.

5. *Grid of evaluation*

With the imputed data, and using R packages such as ‘e1071’ (5), we implemented Support Vector Machine with different kernels including linear, radial etc., and Naive Bayes to compare and contrast accuracies obtained by those metrics with our decision tree accuracies. We also compared with the AIRS algorithm done by Polat (1). We also implemented feature selection on our arrhythmia data to obtain a subset of important features, and consequently created a decision tree, and ran SVM, Naive Bayes, etc. on those features. Thus, in our grid of evaluation, we compared the accuracies obtained from each metric on imputed data, with the accuracies obtained from each metric from feature selection.

III. RESULTS

1. *CART*

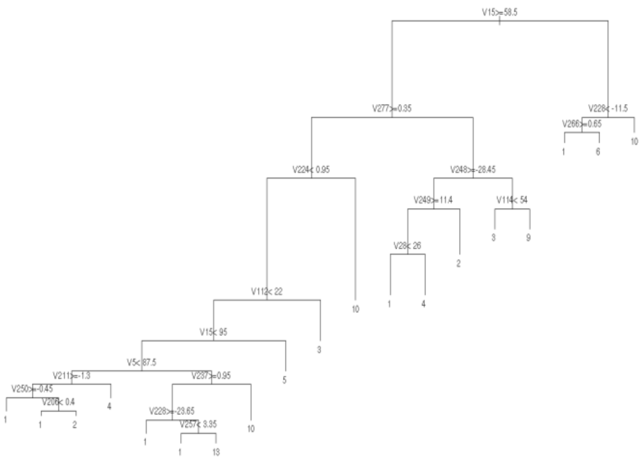
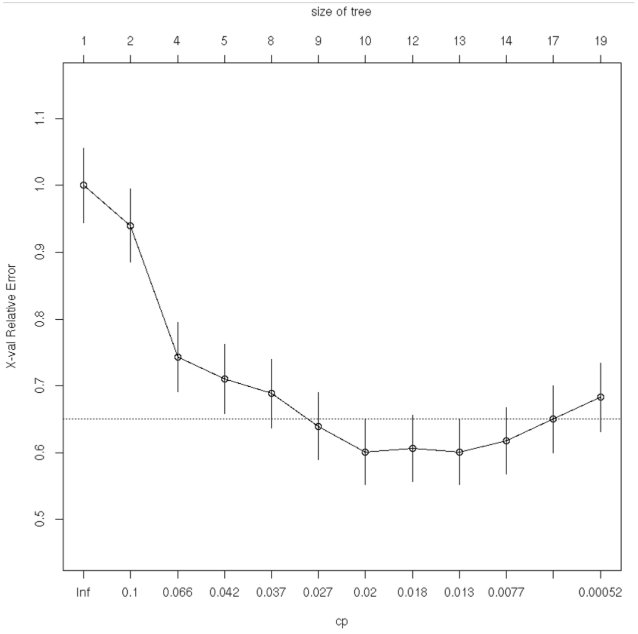


Figure 2. Decision Tree from UCI Arrhythmia data

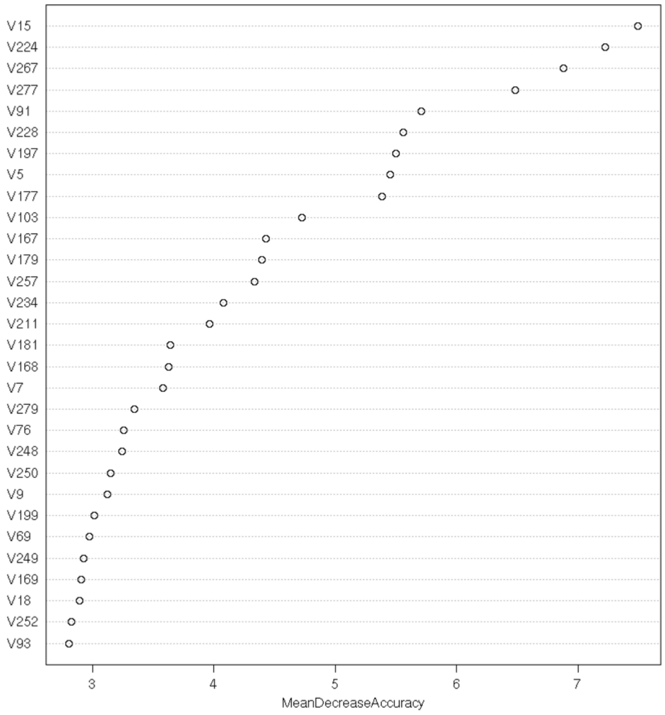
Figure 2 is the decision tree for the arrhythmia data. There are classifications 1-10 represented, and the tree is about ~7 levels deep. The root of the tree was feature 15 or the heart rate in number of heart beats per minute.

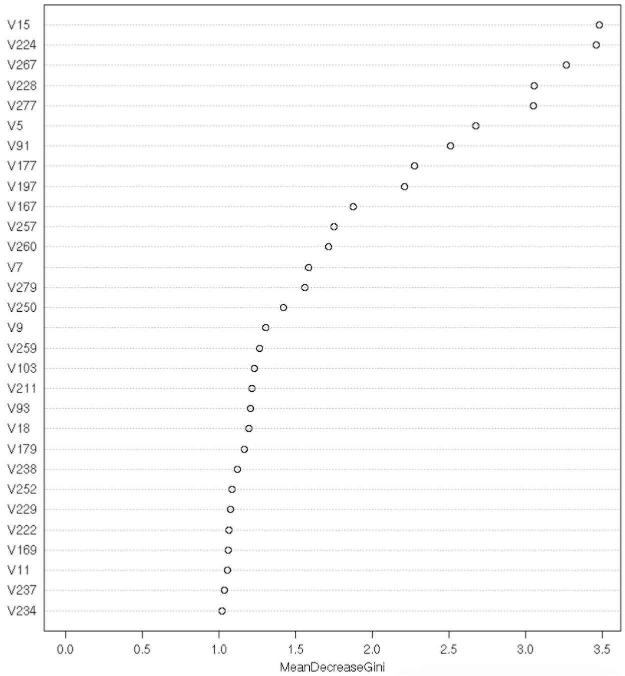


Graph 1. Complexity Parameter vs. Error

The above graph shows the optimal complexity parameter. The complexity parameter that minimized the relative error of the tree was ~0.027.

2. *Accuracy Measure*





Graph 2. Importance Measures of the Arrhythmia Features

The above two graphs give two types of importance measures of the features. The left graph shows how worse the decision tree would perform without each feature, i.e. the measured decrease in overall accuracy of the tree. Hence, a high decrease in accuracy would be expected for very predictive features. The right graph measures the Gini decrease, which is another way of measuring the importance of features. We see that in both graphs, feature 15 has the highest scores. This makes sense, since feature 15 is the root of the decision tree.

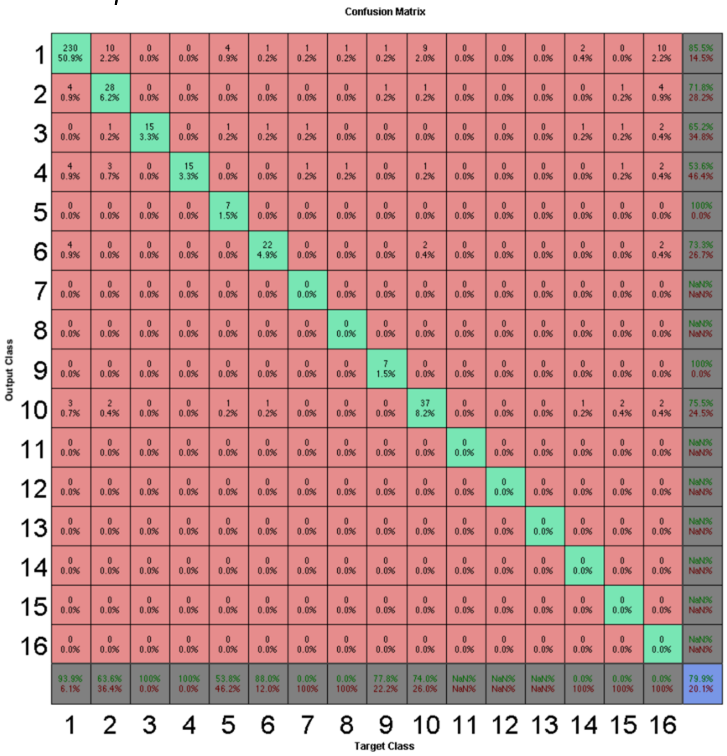
50 \50 %	70 \30 %
100 \100 %	10-fold x-validation

Algorithm	Feature Reduced		$-V_{14}, Imputed$	
Naive Bayesian	54.11	56.12	62.60	62.10
	57.39	56.10	64.23	62.91
ANN	57.12	61.11	-	-
	66.70	62.32	-	-
AIRS-Fuzzy Weighting	-	-	76.34	75.34
	-	-	-	76.34
SVM-Linear	66.67	73.34	61.23	68.34
	81.20	75.28	100.0	68.28
SVM-Radial	63.4	72.87	60.11	62.22
	79.04	73.02	79.12	66.45
SVM-Polynomial	64.33	68.81	62.3	64.23
	81.6	72.90	89.98	67.00
CART-Information Split	68.54	71.27	65.27	66.11
	81.67	79.81	80.10	74.56
CART-Gini Split	68.58	69.3	68.10	68.46
	81.67	78.81	79.86	79.65

Table 3. Overall accuracies from all metrics.

Table 3 compares accuracies of all the different metrics. In each box, we divide the box into a quadrant of values. The top left cell of the quadrant refers to the accuracy of splitting the data 50% training data and 50% test data. The top right cell is 70% training data and 30% test data. The bottom left cell is 100% training data and 100% test data. The bottom right is the result of 10-fold cross validation. We see that the CART algorithm has ~80% accuracy, which is relatively high compared to other algorithms. Comparing the accuracies of CART on imputed data and feature selection, they are approximately the same, while the accuracies for SVM on imputed data and feature selection are significantly different. This suggests that the CART algorithm already does well in finding the most important features in the arrhythmia data, which is represented by the pruning of the tree and the complexity parameter.

3. *Confusion Matrix*


Confusion Matrix

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
1	230 50.9%	10 2.2%	0 0.0%	0 0.0%	4 0.9%	1 0.2%	1 0.2%	1 0.2%	1 0.2%	9 2.0%	0 0.0%	0 0.0%	0 0.0%	2 0.4%	0 0.0%	10 2.2%	85.5% 14.5%
2	4 0.9%	28 6.2%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.2%	1 0.2%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.2%	4 0.9%	71.8% 28.2%
3	0 0.0%	1 0.2%	15 3.3%	0 0.0%	1 0.2%	1 0.2%	1 0.2%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.2%	1 0.2%	2 0.4%	65.2% 34.8%
4	4 0.9%	3 0.7%	0 0.0%	15 3.3%	0 0.0%	0 0.0%	1 0.2%	1 0.2%	0 0.0%	1 0.2%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.2%	2 0.4%	53.6% 46.4%
5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	7 1.5%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
6	4 0.9%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	22 4.9%	0 0.0%	0 0.0%	0 0.0%	2 0.4%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 0.4%	73.3% 26.7%
7	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
8	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
9	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	7 1.5%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
10	3 0.7%	2 0.4%	0 0.0%	0 0.0%	1 0.2%	1 0.2%	0 0.0%	0 0.0%	0 0.0%	37 8.2%	0 0.0%	0 0.0%	0 0.0%	1 0.2%	2 0.4%	2 0.4%	75.5% 24.5%
11	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
12	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
13	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
14	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
15	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
16	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	NaN% NaN%
	93.9% 6.1%	63.6% 36.4%	100% 0.0%	100% 0.0%	53.8% 46.2%	88.0% 12.0%	0.0% 100%	0.0% 100%	77.8% 22.2%	74.0% 26.0%	NaN% NaN%	NaN% NaN%	NaN% NaN%	0.0% 100%	0.0% 100%	0.0% 100%	79.9% 20.1%

Output Class

Target Class

Table 4. Confusion Matrix of Classifications by CART

The additional information provided by the confusion matrix indicates the weaknesses and strengths of our model. For instance, we can infer with relatively high confidence that a prediction indicating class 1, normal heart conditions, is correct since we have 93.9% true positive rate for normal heart conditions. However, we can see that we only had a 53.8% true positive rate for class 5, Sinus Tachycardy. The confusion matrix allows us to understand which specific classification inaccuracies are occurring so that we can fine tune our model against such errors in the future. More specifically we can observe our data with attention to attributes that might sway the model towards the errors made apparent by the confusion matrix, something we could not do with the overall accuracy.

IV. Conclusion

Our CART analysis performed with ~80% accuracy, which did relatively well compared to other classifiers such as SVM or AIRS (1). From our CART graphical model, we found that the root of the tree was feature 15, or the heart rate in number of beats per minute. As the root of the tree is suggested to be a determining feature of the data, this makes sense because we would expect normal or abnormal heart rates to be strongly correlated with arrhythmia. Looking at the top nodes of the tree, we saw that these nodes were related to a multiple of different features, but mainly features in the 200 range i.e. channel values.

V. Discussion and Future Work

Future work could involve further investigating top features suggested by CART using PCA, features selection, etc. We could try fitting more models, and see if those models also suggest that features in the 200 range are important. We could use this new information to dig deeper into the data to understand what relationship these features have with other features. Also, as the data set had a lot of missing values, one could also work on gathering more robust training data, or using another method of imputation such as Fuzzy K-means Clustering, since we can expect training examples with the same classification to have similar characteristics or features (10). Ideally, in the future, we want to validate or decide which features are important for doctors to decide between normal or abnormal heart conditions.

Acknowledgment

We would like to thank Dave Deriso who advised this work. We also want to acknowledge Andrew Ng, without whom this wouldn’t be possible.

References

[1] Polat, Kemal, and Salih Güneş. "Detection of ECG Arrhythmia using a differential expert system approach based on principal component analysis and least square support vector machine." *Applied Mathematics and Computation*186.1 (2007): 898-906.

[2] Soman, Thara, and Patrick O. Bobbie. "Classification of arrhythmia using machine learning techniques." *WSEAS Transactions on computers* 4.6 (2005): 548-552.

[3] Therneau, Terry M., Beth Atkinson, and Brian Ripley. "rpart: Recursive partitioning." *R package version* 3.3.8 (2010)..

[4] "The Caret Package." *The Caret Package*. 15 Aug. 2014. Web. 28 Nov. 2014.

[5] "Package ‘e1071’." *R Software package, avaliable at http://cran.r-project.org/web/packages/e1071/index.html*

[6] Polat, Kemal, Seral Şahan, and Salih Güneş. "A new method to medical diagnosis: Artificial immune recognition system (AIRS) with fuzzy weighted pre-processing and application to ECG arrhythmia." *Expert Systems with Applications* 31.2 (2006): 264-269.

[7] Arrhythmia data set can be downloaded from webpage at *https://archive.ics.uci.edu/ml/datasets/Arrhythmia*

[8] Li, Dan, et al. "Towards missing data imputation: A study of fuzzy k-means clustering method." *Rough Sets and Current Trends in Computing*. Springer Berlin Heidelberg, 2004.

[9] Robertson, B. L., C. J. Price, and M. Reale. "Nonsmooth optimization using classification and regression trees." Proceedings of the 18th IMACS World Congress and MODSIM09 International Congress on Modelling and Simulation, Cairns, Australia. 2009.