Prediction and Classification of Cardiac Arrhythmia

Pammi Sairam, Akhilesh Yadav, Samujjal Das, Rajesh Kumar Jha

Dept. of CSE, Indian Institute of Technology Guwahati

CS 561 Project Report

***Abstract* -- Cardiac Arrhythmia[1] is a cluster of conditions in which the electrical activity of the heart is not regular or is more relaxed or quicker than the usual. Many cardiac arrhythmias are not of any harm: but, if they are particularly abnormal, or result from a weak or unstable heart, they can cause significant and even potentially fatal symptoms[2]. It is the major cause of causality for both men and women around the globe.**

We aim to predict and classify Cardiac Arrhythmia based on a patient's medical record, in this very project. Our primary objective is to classify a patient into one of the Arrhythmia classes like Tachycardia and Bradycardia based various measurements like his ECG measurements and help us in understanding the application of machine learning in medical domain. After appropriately selecting the features we aim to resolve this problem by using Machine Learning based Algorithms namely Support vector machine, Softmax Regression, Naïve Bayes, Decision Tree, Adaptive Boosting (AdaBoost) and Random Decision Forest.

I. Introduction

The aggregate number of causalities caused because of cardiovascular illnesses[3] read 17.3 million a year as per the WHO reasons for death. In this way, how to anticipate heart arrhythmia, in actuality, is of extraordinary centrality.

We have chosen this project since arrhythmia has no symptoms and is a well-known silent killer, so it can be a boon for people who do not have any clue at their disposal regarding their medical condition. In this project, we plan to develop a machine learning system that can classify a patient into different cardiac arrhythmic classes.

The diagnosis and classification of cardiac arrhythmia can be categorised into various classes based on the various parameters like age, Blood Pressure level, Electrocardiogram measurements and other attributes.

The first class will allude to the typical patient with no disease while different classes might speak to various classes of cardiovascular arrhythmia like Tachycardia, Bradycardia and Coronary related artery problems. The problem here is of Supervised learning.

2. Data Set

The dataset[4] for the for the same is being taken from the UC Irvine Machine Learning Repository https://archives.ics.uci.edu/mla/dataset/Arrhythmia. There are a total of (452) rows, each representing medical record of a unique patient. There are 279 columns or attributes like blood pressure, age, weight and patient's ECG related data. Out of 279 attributes, 206 are linear valued or co-related and the rest are trivial. General attributes like age and weight have discrete values which are integers while other ECG features are real valued. The dataset has provided a platform for training and testing different models used and furthermore setting a benchmark for the same.

The plan is to differentiate between the existence and absence of heart arrhythmia and to categorise it among one of the 13 groups. The first class refers to 'normal' ECG classes 02 to 12 refers to different classes of arrhythmia and class 13 refers to the rest which haven't been classified yet. The 13 classes are as following -

Number	Class	Instances
1	Not Arrhythmia	245
2	Ischemic changes (Coronary Artery)	44
3	Old Anterior Myocardial Infarction	15
4	Old Inferior Myocardial Infarction	15
5	Sinus tachycardia	13
6	Sinus bradycardia	25
7	Ventricular Premature Contraction (PVC)	3
8	Supraventricular Premature Contraction	2
9	Left bundle branch block	9
10	Right bundle branch block	50
11	Left Ventricle hypertrophy	4
12	Atrial Fibrillation or Flutter	5
13	Rest	22

Table 1: Different Classes of Arrhythmia.

3. SURVEY

The plan is to differentiate between the presence and absence of heart arrhythmia and to categorise it in one of the 13 classes. For now, there exists a computer program that makes such a categorisation or classification. In many case, there are contrasts between the cardiology's and the projects classification. Taking the cardiology's as a best quality level we intend to limit this distinction by methods for machine learning calculations and instruments.

4. SCOPE

These machine learning procedures can be sent in doctor's facilities where a substantial dataset is accessible and can help the specialists in settling on more exact choices and to chop down the quantity of causalities because of heart infections later on.

Furthermore, the same can be used and employed in wearable devices that can track the patients' conditions and report depending on the complexity. Moreover, on the unavailability of doctors the machine can handle the situation for a given amount of time.

5. METHODOLOGY

A. Feature Selection[5]*:*

At the outset, we eradicated some of the categorical features that were 96% of time referring either all 0's or all 1's. On the off chance that any training data has a missing an incentive for a given column, we set it as the mean of the value give or take the standard deviation for that attribute identified with the class it has a place with. The focal introduce when utilizing a feature selection technique is that the information contains numerous highlights that are either excess or immaterial, and would thus be able to be evacuated without causing much loss of data.

If for a given column if most of values are missing, then we discard that attribute and eradicate it from our training set.

The features can be classified into five blocks –

- features that are related to biographical characteristics, i.e., height, gender, age, weight and heart rate.
- features regarding the average wave durations of each interval (PR and ST interval and QRS complex).
- features that are concerned with vector angles of each of the waves.
- features that are concerned with the widths and amplitude of each wave

B. Principal Component Analysis[6]*:*

The main idea of Principle component analysis is mainly useful to reduce the dimensionality of dataset, which consisting of many attributes correlated with each other. i.e. It is a procedure that transforms a number of correlated attributes (actual features) into a smaller number of uncorrelated attributes (Principle components).

Typically, principle components are more important features and less or equal in number to original attributes. The first principle component which we find has the highest possible variance and next principle component has the next highest possible variance and so on. All principle components are orthogonal to each other.

C. Random Forests and Decision Trees[7]*:*

Decision trees are a sort of model that are utilized for both classification and regression issues. Decision trees offer response to successive inquiries which have a tendency to send us down a specific course of the Decision tree, given the appropriate response.

We have executed a Random Forest classifier. The model works by reliably sampling with substitution a fragment of the training dataset, and fitting a decision tree to it. The quantity of trees alludes to the circumstances the dataset is randomly tested. Besides, in each sampling emphasis, a random arrangement of highlights are chosen.

In decision trees, every hub alludes to one of the input variables, which has edges to children for all possible values that the input can take. Each leaf relates to an estimation of the class label given the estimations of the information factors spoke to by the way from the root node to the leaf node. The number of trees and the number of leaves are learned through cross validation.

Random forests are a solid modelling technique and considerably more vigorous than a solitary decision tree. They accumulate many decision trees to limit overfitting as well as error due to bias and therefore yield useful results with respect to usual decision trees.

5. Models

A. *Naïve – Bayes Classifier*[8]*:*

It is a classification method in light of Bayes Theorem[9] with a presumption of independence among predictors, that is they are absolutely free. Naive Bayes strategies are an arrangement of supervised learning calculations in view of applying Bayes' theorem with the "naive" supposition of independence between each combine of features. Using the formula of conditional probability, we have the probability of whole training set:

$$P(x,y) = \prod_{i=1}^{m}\left(\prod_{j=1}^{n}\phi_{j,x_j^{(i)}|y=y^{(i)}}\right)\phi_{y^{(i)}}$$

We have employed Naive Bayes binomial and multinomial classifiers using Python. And have used Naive Bayesian equations to calculate the corresponding posterior probabilities for each class present out there. The class with the highest posterior probability amongst the rest is the result of the prediction.

Naive Bayes learners and classifiers can be to a great degree quick contrasted with rest of the modern techniques. The decoupling of the class conditional feature distributions allude that every dispersion can be evaluated autonomously as a one-dimensional distribution. This thus annihilates issues originating from the issue of enormous measurements.

B. *SVM*[10] *(Support Vector Machines):*

Support Vector Machines rely upon the possibility of decision planes that describe decision boundaries. A decision plane is one that confines between a course of action of things having unique class memberships.
Here our aim is to minimize the following equation:

$$\left[\frac{1}{n}\sum_{i=1}^{n} max\left(0,1 - y_i\left(\underset{w}{\rightarrow}.\underset{x_i}{\rightarrow} - b\right)\right)\right] + \lambda \left\|\underset{w}{\rightarrow}\right\|^2$$

where the parameter λ regulated the trade-off among increasing the margin-size and reinsuring that the $\underset{x_i}{\rightarrow}$ do not lie on the incorrect side of the margin.

In SVM, a hyperplane is chosen to best separate the focuses in the info variable space by their class, either class 0 or class 1. In two-dimensions you can imagine this as a line. You can make classifications utilizing this line. By putting the information esteems into the line condition, we find whether another point is above or beneath the line. We have attempted both the polynomial and the linear kernels for the SVM and results show that the linear kernel beat the polynomial kernel, and have actualized linear kernel leaving the polynomial one.

C. *Softmax Regression*[11]*:*

Here we use a generalised version of logistic regression that is Softmax regression. This can be used for multi-class classification (Provided classes are mutually exclusive). In logistic regression number of classes was binary but here it can take any number of classes. We just here replace the sigmoid function of logistic regression with so called Softmax function:

$$P\left(k = j|a^{(i)}\right) = \phi_{softmax}\left(a^{(i)}\right) = \frac{e^{a^{(i)}}}{\sum_{i=0}^{m} e_m^{a^{(i)}}}$$

More precisely logistic regression uses sigmoid function as activation function but it uses softmax function as activation function. Softmax function applies at output layer, it calculates the probability of given input x being in class k.
For an example, in our case we have 13 categories to classify arrhythmia in, so here the value that K will take is 13. For different cases the value of K changes. For the purpose of digit recognition, we will have the value of K to be 10 for ten digits to be recognised.
Here we have provided with the training cost of Softmax regression. The training accuracy was same for both the splits that is for 80%-20% and 70%-30%. The learning rate chosen was .00001.

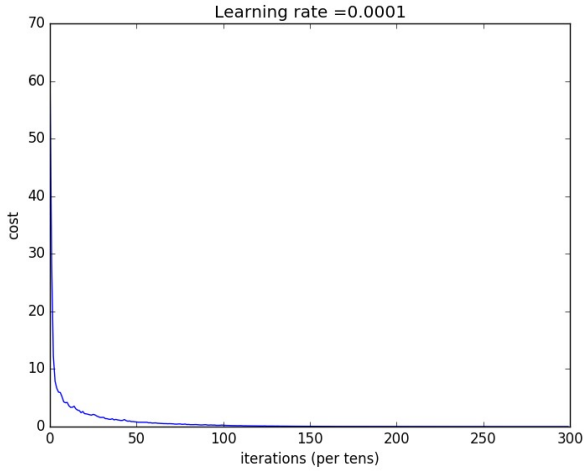

Image 1: Softmax Regression training error rate.

D. *Decision Tree*[12]*:*

We utilize a non-parametric supervised strategy utilized for classification called Decision Trees, it predicts the estimation of an objective variable by learning basic decision rules surmised from the information features.
A decision tree normally begins with a solitary node, which branches into conceivable results. Every one of those results prompts extra nodes, which branch off into different potential outcomes. This gives it a treelike shape.
There are three unique sorts of nodes: chance nodes, decision nodes, and end nodes. A possibility node, being spoken to by a circle, showed the probabilities of specific outcomes. A decision node, being spoken to by a square, demonstrates the decision to be made, and in conclusion an end node demonstrating the ultimate result of a decision way. Decision tree has ended up being an effective strategy and has given us a precision more than that of the normal exactness.

E. *AdaBoost*[13] *(Adaptive Boosting):*

AdaBoost infers to a specific method of training a classifier that is boosted. A boosted classifier can be of the form:

$$F_T(k) = \sum_{t=1}^{T} f_t\,(j)$$

where each f_t is a learner that's weak and that takes an object k as input and returns a value referring the class of the variable.

Adaboost classifier consolidates weak classifier calculation to frame solid classifier. A solitary calculation may classify the items inadequately. In any case, on the off chance that we consolidate numerous classifiers with determination of training set at each cycle and appointing appropriate measure of weight in definite voting, we can have great precision score for general classifier.
Adaboost is an ensemble learning technique to create strong classifier with the help of number of weak classifier like

decision stump (decision tree with maximum depth 1). In this method first each instances of training data set is given same weight:

$$w(i)=1/n$$

where w(i) is the weight of i'th training instance and n is the total number of instances in training dataset. Then a decision stump is trained on weighted sample and prediction error is calculated for each instance. Let the prediction error for i'th instance is terror(i) and this is equal to:

terror = 0 if (y == p),
otherwise 1

The misclassification rate is calculated with the help of the weight of each training instances.

$$error = sum(w(i) * terror(i)) / sum(w)$$

Stage value for current decision tree is also calculated to make the final prediction:

$$stage = ln((1-error) / error)$$

where ln() is natural logarithm. Now weight of each training instance is calculated using:

$$w = w * exp (stage * terror)$$

Another decision tree is trained using the new assigned weights and in this manner, various decision tree are added sequentially using the weight training data. This process is repeated until a pre-set number of decision stump have been created or no further improvement can be made on the training data. The final prediction is made using the weighted average of prediction made from different decision stump and stage values.

F. *Random Forests*[14]*:*

Random forests or random decision forests classifier is additionally an ensemble learning strategy. In this technique number of training set (around couple of hundreds) is made from unique training set by random choosing variable and occurrences. Random forests can be used to rank the criticalness of variables in a regression or classification issue regularly.

This procedure is known as bootstrapping. A decision tree is prepared on each bootstrap. The final prediction can be the normal of all forecast or poll result can be utilized to make final prediction.

The training algorithm for random forests used the general system of bootstrap collecting or bagging to tree learners. As a bit of their change, random forest predictors consistently incite a uniqueness measure between the perceptions. Furnished the training set with reactions bagging over and over chooses a random example with replacement of the training set and fits trees to these examples.

6. RESULTS

The key focus of this project was to develop a system that could generously recognize an arrhythmia. The second focus of this venture was to build up a technique to vigorously classify an Electrocardiogram measurement trace into one of 13 wide arrhythmia classes. We report our execution for every one of the six methods utilizing one methodology.

We show results for each algorithm, as well as vary other parameters for better results. The following are two tables with results. At the outset, the testing-training data was split 20% - 80% and in the latter case, the testing-training data was split 30% - 70%.

Model	Training Accuracy	Test Accuracy
Naïve-Bayes Classifier	55.57%	54.52%
Support Vector Machine	100%	65.83%
Softmax Regression	100%	66.26%
Decision Tree	85.12%	71.32%
Adaptive Boosting	100%	67.65%
Random Forests	100%	70.92%

Table 2: For 80% - 20% split.

Model	Training Accuracy	Test Accuracy
Naïve-Bayes Classifier	54.63%	50.73%
Support Vector Machine	100%	70%
Softmax Regression	100%	60.62%
Decision Tree	87.39%	72.63%
Adaptive Boosting	100%	78.16%
Random Forests	100%	72.52%

Table 3: For 70% - 30% split.

7. ANALYSIS

Being a non-trivial problem, the data set available for the problem is limited and skewed. So only accuracy can't be taken as the measure of performance for the same.

In classification problems precision is the fraction of relevant instances among the retrieved instances, while recall is the fraction of relevant instances that have been recognised over the aggregate sum of relevant instances.

The F1 score can be interpreted as a weighted average of the precision and recall. It is the harmonic average of precision and recall and an F1 score reaches its best value at 1 and worst at 0.

The literal meaning of recall and precision can be expressed as, recall expresses the ability to find all relevant instances in a dataset, precision expresses the proportion of the data points our model says was relevant actually were relevant. Having high precision is when we say someone has arrhythmia, we

are usually right and having high recall is we can identify most people with arrhythmia out there.
We found the precision and recall scores for the models which performed well than the rest.

Model	Precision Score	Recall Score	F_1 Score
Random Forest	0.7314935144	0.7703703704	0.752445448
Decision Tree	0.6357205153	0.6632653061	0.649158888
AdaBoost	0.5359531773	0.6739130435	0.597029277
SVM	0.7305269187	0.6907216495	0.710062726

Table 4: Precision and Recall scores.

We can see our analysis shows Random forest has the highest recall score and with respect to Recall both SVM and Random Forest perform well. And the highest F_1 score is again possessed by Random Forest and is being followed by SVM.

8. CONCLUSION

It is clear from the above data that the SVM and Softmax Regression algorithms are able to automatically detecting and classify arrhythmias with a great accuracy (Training Data = 100 % and Testing Data=65%). Furthermore, Random Forests consistently performs better than PCA in terms of feature selection. The general approach we followed in this project was as per the following. We started with Random Forests and we tried to obtain maximum accuracy but we got around 71% by having of split of the data into 80% training and 20% testing, which clearly indicated lack of training data.
Then we used Softmax Regression which uses the softmax function and we ran it using Adam optimizer. Softmax regression gave comparatively poor results with average accuracy around 66 %. Naïve-Bayes classifier gave poor results due to problem of lack of enough training examples (452) and excessive number of features.
Adaptive boosting Using PCA gave the best results with average accuracy of classification around 100 % for training set and 78 % for testing set.

9. ACKNOWLEDGEMENT

We are profoundly thankful to our professor Dr. Rashmi Dutta Baruah for her continued guidance with direction and support over the span of this course.

10. REFERENCES

[1] Labib Khadra, Amjed S.Al-Fahoum and Saed Binajjaj, "A Quantitative Analysis Approach for Cardiac Arrhythmia Classification Using Higher Order Spectral Techniques", IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 52, NO. 11, NOVEMBER 2005.

[2] Constanze Schmidt, Jana Kissebach, Patrick A Schweizer, Hugo A Katus and Dierk Thomas, "The pathology and treatment of cardiac arrhythmias: focus on atrial fibrillation", 2011 Mar 31.

[3] Cristina Basso, Fiorella Calabrese, Domenico Corrado, Gaetano Thiene, "Postmortem diagnosis in sudden cardiac death victims: macroscopic, microscopic and molecular findings", Cardiovascular Research 50 (2001) - 290-300, accepted 29 January 2001.

[4] Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[5] Chulmin Yun, Donghyuk Shin, Hyunsung Jo, Jihoon Yan and Saejoon Kim, "An Experimental Study on Feature Subset Selection Methods",Seventh International Conference on Computer and Information Technology.

[6] Jonathan Shlens, "A Tutorial on Principal Component Analysis", Dated: December 10, 2005; Version 2.

[7] Simon Bernard, Laurent Heutte and Sebastien Adam, "On the Selection of Decision Trees in Random Forests", Proceedings of International Joint Conference on Neural Networks, Atlanta, Georgia, USA, June 14-19, 2009.

[8] I. Rish, "An Empirical Study of the naive bayes classifier", T.J. Watson Research Centre

[9] Bradley Efron, "Bayes' Theorem in the Twenty-first Century", published in Science 340, June 7, 2013, doi:10.1126/science.1236536.

[10] Durgesh K. Srivastava, Lekha Bhambhu, "Data Classification using Support Vector Machine", Journal of Theoretical and Applied Information Technology.

[11] Chao-Ying Joanne Peng, Kuk Lida Lee, Gary M. Ingersoll, "An Introduction to Logistic Regression Analysis and Reporting".

[12] Himani Sharma, Sunil Kumar, "A Survey on Decision Tree Algorithms of Classification in Data Mining", International Journal of Science and Research (IJSR)

[13] Tae-Ki An and Moon-Hyun Kim, "A New Diverse Adaboost Classifier", 2010 International Conference on Artificial Intelligence and Computational Intelligence

[14] Zachary Jones and Fridolin Linder, "Exporatory Data Analysis using Random Forests"