CARDIAC ARRHYTHMIA DETECTION USING CLASSIFICATION TECHNIQUE

Sritharan Mahendra Babu
Mahendrababu.s@husky.neu.edu

Abstract

Cardiac Arrhythmias are any of a group of conditions in which the electrical activity of the heart is irregular or is faster or slower than normal. Cardiac arrhythmia, also known as cardiac dysrhythmia or irregular heartbeat, is a group of conditions in which the heartbeat is irregular, too fast, or too slow. It is the leading cause of death related to heart for both men and women in the world. Heart diseases kill more than 400,000 people annually. In the United States, someone has a heart attack every 34 seconds. This paper aims to detect and classify arrhythmia into 16 different variants using different types of classifier.

1. Introduction

Irregularity in heart beat may be harmless or life threatening. Cardiac Arrhythmia acts as a precursory diagnosis to the Cardiac/Heart Attack. Cardiac Arrhythmia can be diagnosed by measuring the heart activity using an instrument called ECG or electrocardiograph. Different parameter values can be extracted from the ECG and can be used along with other details about the patient like age, medical history, etc. to detect cardiac arrhythmia. Using machine learning, cardiac arrhythmia diagnosis can be very helpful. The project aims at using different machine learning algorithms like Logistic Regression, SVM for predicting and classifying cardiac arrhythmia into different categories.

1.1. Existing Methods

The most traditional method of Cardiac Arrhythmia detection is done by doctors or physicians to look at these long duration ECG recordings and find minute irregularities in the obtained recordings. There are various other hardware devices to do the analysis of the ECG recordings to detect Cardiac Arrhythmia but it involves human involvement or assistance to understand the ECG recordings to analyze the disease.

1.2. Overall Approach

I am planning on approaching this problem using machine learning algorithms which I learnt during my course work. I am primarily using two major Classifiers to predict or classify the dataset for Cardiac Arrhythmia. The classifiers are namely:

- Logistic Regression

Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function.

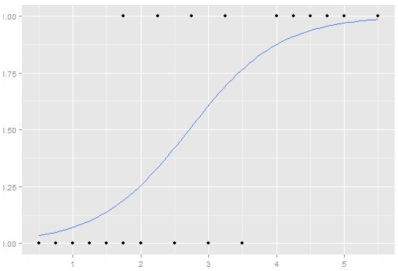

Figure 1. Logistic Regression

- SVM (Support Vector Machines)

Support vector machines (SVMs, also support vector network) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other.

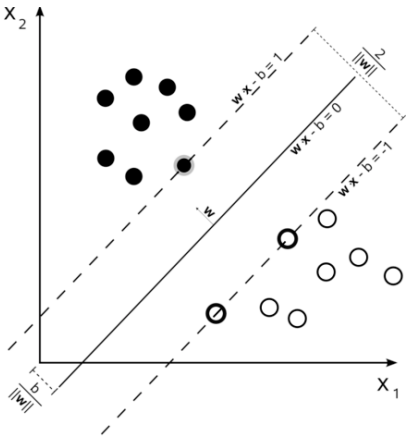

Figure 2. SVM (Support Vector Machine)

Using these two algorithms / classifiers is a very high level approach of this project. The detailed approach involves more data preprocessing and data analysis such feature selection and dimensionality reduction to increase the accuracy of the classifier / estimator.

2. Proposed or Studied Approach

I am using Logistic Regression and SVM as the primary estimator / classifier for performing the prediction. But the data needs to be pre-processed even before it is fed as the train or test data to the estimators.

2.1. Details of the Study

I am using a mid-scale dataset which has been obtained from University of California Irvine. This dataset is recorded on various patients using ECG machine. The data cannot be directly fed into the estimator as it contains unknown variables which is totally not usable by the estimators.

2.2. Data Pre-Processing

- Firstly, I am reading the dataset from a csv using pandas to transform the data into a panda's data frame
- Secondly I had to clean the data as it contains unknown variables and the data type varies between integer and float
- Then I have replaced the unknown variables with the median of the feature vector through a concept calling imputation

2.3. Technical Approach

I am using Scikit Learn machine learning package which gives Logistic Regression and SVM as an out of box functions for performing the inference as well as prediction, numpy for matrix computation and other mathematical computation and matplot library for plotting the accuracy graphs.

- After data cleaning, the data is split into multiple types of train and test vectors.
- The data is first split in a very naïve approach with the top 90% as the train data and the next 10% as the test data
- Second set of data is split with a random seed of 100 with 90%-10% approach to eliminate the bias or skewed nature of the dataset

Basically, two different types of dataset are constructed from the parent file to understand the nature / behavior of various estimator to achieve maximum accuracy.

After two sets of dataset being constructed, both the dataset is tested with Logistic Regression and SVM estimators for comparing the accuracy as well the trueness of the prediction.

I am using a high dimensional dataset with 279 feature vectors thus introducing overhead of understanding those feature vectors that contributes to higher accuracy.

So, to avert this high dimensional feature vector problems, I introduced another layer of data processing called dimensionality reduction. I am using PCA for dimensionality reduction for increasing the accuracy of the prediction.

In the inference phase, after the data pre-processing the training data is applied to both the Logistic Regression and SVM estimators. After inference of different estimators, prediction of the test data is performed to analyze the accuracy of the estimators.

I have performed estimation of the given dataset using 3 estimators (Logistic Regression, LinearSVM, anova SVM).

- Logistic Regression is used as a benchmark algorithm / estimator for getting an insight into prediction and understand the accuracy of the dataset from a regression point of view
- SVM is used as an actual estimator or inference algorithm for classification of high dimensional dataset. I tried using SVM with various kernels and for various PCA component levels.
- Since I am performing PCA dimensionality reduction of the dataset, various iterations of experiments were performed on both Logistic Regression and SVM at different component reduction levels.
- I tried with another data pre-processing method called SelectKBest. This is basically selecting those features that have highest impact on the prediction accuracy by analyzing the consistency of variance for all the feature vectors.

3. Experiments

I have performed various levels of experiments on obtaining the highest accuracy for the given dataset. The high-level experiments are:

- Data pre-processing
- Estimation / Inference Training and Testing

3.1. Details of the Dataset

The data has been taken from a well-maintained ECG (Electrocardiography) database (https://archive.ics.uci.edu/ml/datasets/Arrhythmia). It contains 279 attributes (ECG/ Patient related variables) and 452 instances. The variable 'Class' is our target variable. Class 01 refers to 'normal' ECG, classes 02 to 15 refer to different classes of Arrhythmia and class 16 refers to the rest of unclassified classes. The below figure represents the distribution of the given dataset.

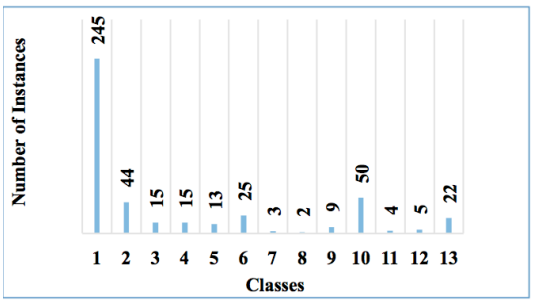

Figure 3. Input / Given Dataset split

3.2. Experiments Details Steps

3.2.1. Data Pre-Processing

- Import of CSV data using Pandas and converting it into a matrix
- Replacing the unknown variables with 'NaN' for all the feature vectors
- Performing Imputation of those feature vectors that contain unknown variable (i.e. replace unknown variables with the mean of the feature column)
    - *(Note: I have enquired with professor that performing imputation for handling unknown variables is an appropriate approach for data pre-processing)*
- Applying PCA to the high dimensional dataset for dimensionality reduction
- Train and Test data split based on randomization
    - The train and test data are split into 90%-10% for inference and test purpose

3.2.2. Logistic Regression

I am using logistic regression for benchmarking and to understand the accuracy of the dataset for various PCA components. I am applying 'sag' solver and 'multinomial' for multi-class analysis.

Experiments:

- Logistic Regression had a good accuracy for the training dataset and but for test dataset the accuracy is around **63%**. This an expected accuracy level as the dataset is high dimensional and there is a tendency of the regression curve to get skewed due to the high dimensionality. Hence I ran the Logistic Regression inference model for various levels of PCA components range. The accuracy plot and graph are shown below. For Logistic regression, the highest accuracy for the given dataset is **63%** which is reached at a maximum PCA components of 80 (i.e. the 279 feature vectors are reduced to 80 components / features)

PCA Component Range	Accuracy of the Test Data
50	63.04347826
60	60.86956522
70	60.86956522
80	63.04347826
90	58.69565217

Table 1. Logistic Regression PCA vs Accuracy

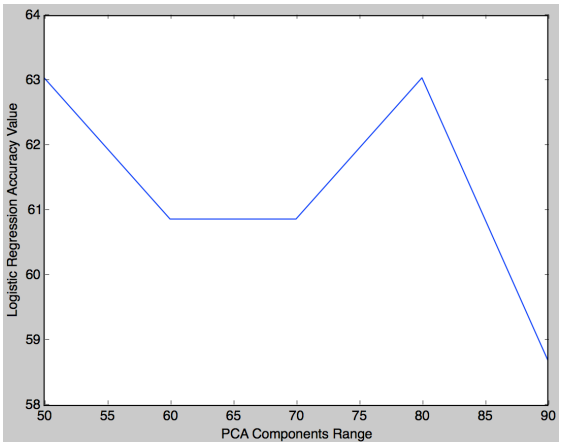

Figure 4. Logistic Regression PCA vs Accuracy graph

3.2.3. SVM

SVM is a very effective estimator / classifier algorithm for performing estimation / prediction of dataset with very high dimensional range. Since the given dataset is with a high dimensional feature vectors of about 279 features, there is a possibility of the estimator to get skewed or to get biased to a column. I performed the SVM inference for various PCA component range to obtain the highest accuracy level of the prediction. I also performed the SVM training for various kernels such 'Linear', 'rbf', 'polynomial' kernels. Various observations are obtained:

- For the present given dataset 'rbf', 'polynomial' kernels didn't perform well for any levels of PCA components and the accuracy didn't increase more than 60%
- Linear kernel SVM was promising with the accuracy improving over different PCA component range. Maximum accuracy was **81%** obtained at a PCA component range of 80. The accuracy score and the PCA component range are shown in table and graph.

PCA Component Range	Accuracy of the Test Data
50	76.08695652
60	78.26086957
70	78.26086957
80	80.43478261
90	78.26086957

Table 2. SVM PCA vs Accuracy

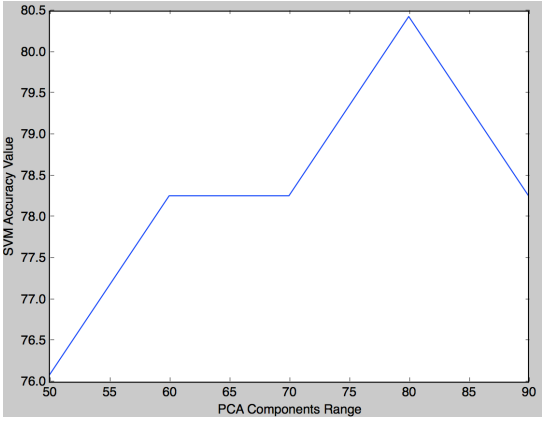

Figure 5. SVM PCA vs Accuracy graph

3.2.4. ANOVA SVM

Anova SVM is basically a feature selection model. The model select's 'k' best features that contributes to the accuracy of the prediction of the estimator. I tried with various range of 'k' values but the accuracy didn't exceed more that 70% maximum. I performed this approach to understand the scope of difference between PCA and Feature selection in the data extraction process. It is seen that for this given dataset the integrity of the dataset and the accuracy of prediction is higher for PCA linear kernel.

4. Conclusion

This paper presents the study and implementation of few techniques of machine learning to perform the prediction of Cardiac Arrhythmia dataset to achieve maximum accuracy. In this paper / project I have implemented Logistic Regression which is an appropriate estimator for benchmarking of the feature understanding and accuracy threshold analysis. Then I implemented SVM which is a very good estimator for dataset of high dimensional space. In this project, I would like conclude that for the given Cardiac Arrhythmia dataset I could achieve a maximum of 81% accuracy at PCA of 80 features with Linear SVM.

5. Future Work

Several other advanced prediction models can be implemented to achieve higher accuracy. From my perspective, more hybrid prediction models such as Random Forest + SVM or ensemble classifiers pipeline can be implemented for better classification accuracy. I would also like to envisage into advanced Neural Network at various levels of hidden perceptron with different activation functions.

References

[1].http://en.wikipedia.org/wiki/Cardiac_dysrhythmia