Meghana Ravishankar
A25248100

Dimensionality Reduction for Arrhythmia

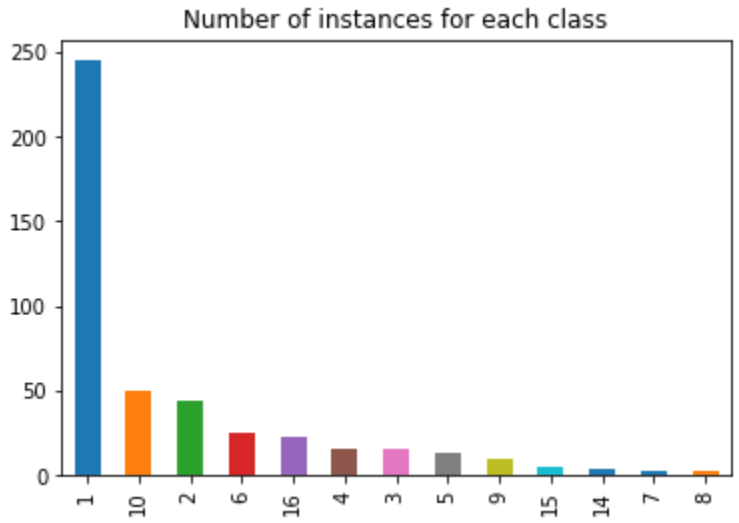
(Project for CS696)

Introduction

Cardiac Arrhythmia is a heart condition in which heartbeat is irregular. About 80% of sudden cardiac death is the result of ventricular arrhythmias. Arrhythmias may occur at any age but are more common among older people. The diagnosis of cardiac arrhythmia type is based on ECG(Electrocardiogram) readings.

The dataset for this project is taken from UCI Machine Learning Repository. It consists of 452 rows and 279 attributes, this is an ideal dataset to understand dimensionality reduction and curse of dimensionality. The dataset consists of 16 classes out of which 3 classes have no recorded instances and thus there are only 13 classes.

The instances for normal condition are more than any other classes as shown in below graph:



The goal of this project is to reduce dimensionality in the most efficient way so as to get better accuracy using scikit learn for python.

Methods

The data was loaded and the columns were re-named by exporting values from an excel file.

For the KNN model, the dimension of the data was not reduced, and the NaN's were replaced with 0. This model is made so that a comparison to old and new models can be made.

For the second KNN model the data was loaded, and the missing data denoted with '?' was changed to NaN values. The columns(features) with NaN values where identified and dropped from the dataset. Doing this reduced 5 features. Other methods for replacing NaN included filling it with zeros or imputing it with feature mean, variance or previous or most repeated value. But

these methods were not required as number of features with NaN were less and imputing or replacing with zero would mean tampering or changing data. This is not advised as the instances are very few. Then features with zero variance were dropped which reduced the features by 17. From the statistical point of view a zero variance means all data points for the feature lie on the same path i.e. a straight line is connecting all points in a plane, this also means that this won't contribute to any learning process. By the end of this a total of 22 features were dropped.

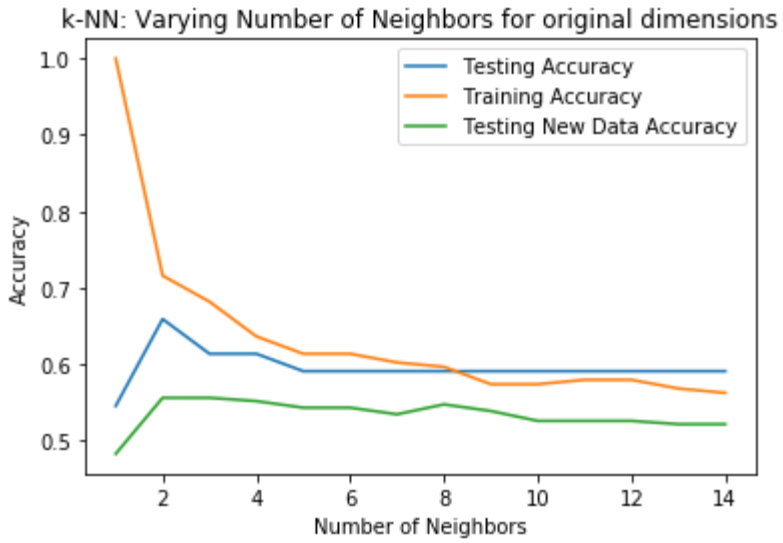
Further feature reduction is done using Correlation filter which dropped 40 features from original number of features. And Random Forest was used to select features.

. The data was then split into data1 and data2 consisting of 220 and 232 instances respectively. Data2 is used as test data, whereas data1 will be again split into train and test data. A supervised clustering algorithm (K-nearest neighbor) was used to test the classification accuracy of the data.

Results:

The maximum efficiency was of approx.55%.
The original output is:

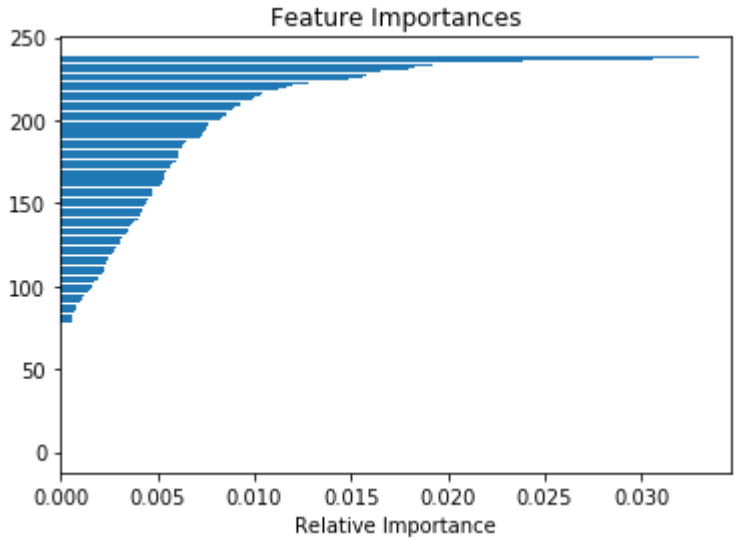
```
Training Accuracy
[ 1.         0.71590909  0.68181818  0.63636364  0.61363636  0.61363636
  0.60227273  0.59659091  0.57386364  0.57386364  0.57954545  0.57954545
  0.56818182  0.5625    ]

 Testing Accuracy
[ 0.54545455  0.65909091  0.61363636  0.61363636  0.59090909  0.59090909
  0.59090909  0.59090909  0.59090909  0.59090909  0.59090909  0.59090909
  0.59090909  0.59090909]

 Testing Accuracy on new data
[ 0.48275862  0.55603448  0.55603448  0.55172414  0.54310345  0.54310345
  0.53448276  0.54741379  0.5387931   0.52586207  0.52586207  0.52586207
  0.52155172  0.52155172]
```



The Random forest selected the following features:

Columns chosen using Random Forest:
[('QRS_duration',), ('Q-T_interval',), (' T_interval',), ("V1_ R'",), ('V1_Number_of_intrinsic_deflections',), ("V2_ R'",), ('D1_a_T_wave',), ('D2_a_T_wave',), ('AVR_a_T_wave',), ('AVR_a_QRSTA',), ("V1_a_R'_wave",), ('V1_a_P_wave',), ('V1_a_QRSA ',), ('V2_a_S_wave',), ("V2_a_R'_wave",), ('V2_a_QRSA ',), ('V3_a_R_wave',), ('V3_a_T_wave',), ('V3_a_QRSTA',), ('V4_a_T_wave',), ('V4_a_QRSTA',), ('V5_a_JJ_wave',), ('V5_a_QRSA ',), ('V6_a_R_wave',), ('V6_a_T_wave',)]

Shape of data after choosing important features: (452, 25)

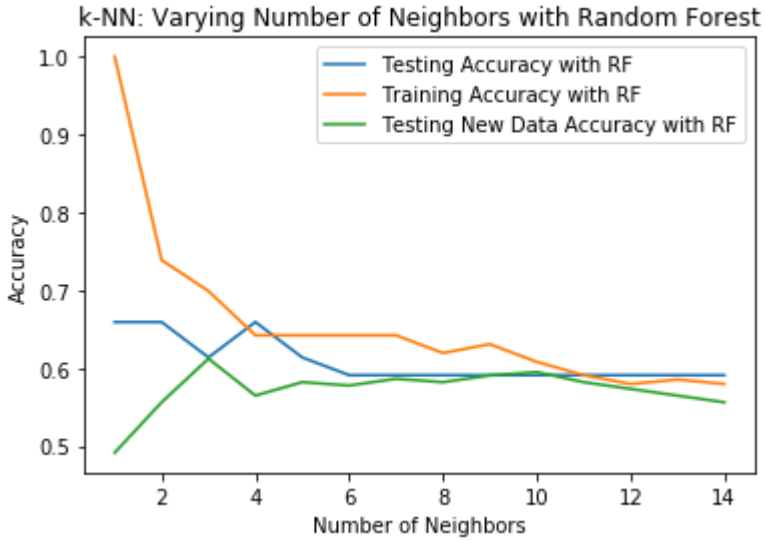


KNN for the reduced dimension: approx.. 61% accuracy for test data.

Training Accuracy
[1. 0.73863636 0.69886364 0.64204545 0.64204545 0.64204545
0.64204545 0.61931818 0.63068182 0.60795455 0.59090909 0.57954545
0.58522727 0.57954545]

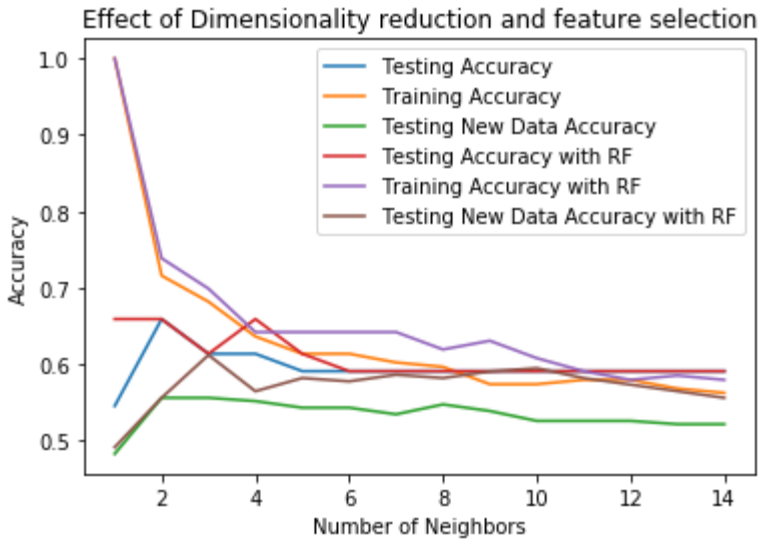
Testing Accuracy
[0.65909091 0.65909091 0.61363636 0.65909091 0.61363636 0.59090909
0.59090909 0.59090909 0.59090909 0.59090909 0.59090909 0.59090909
0.59090909 0.59090909]

Testing Accuracy on new data
[0.49137931 0.55603448 0.61206897 0.56465517 0.58189655 0.57758621
0.5862069 0.58189655 0.59051724 0.59482759 0.58189655 0.57327586
0.56465517 0.55603448]



Conclusion:

There was at least 5% increase in efficiency even when the dimensions were reduced, since the number of features is few, we can also conclude the time taken was less. It is important to reduce dimensions carefully as we do not want to remove any important features. The efficiency can be improved if we have more samples and when other algorithms are tried.



Reference

- https://www.analyticsvidhya.com/blog/2015/07/dimension-reduction-methods/
- http://scikit-learn.org/stable/
- https://dzone.com/articles/what-is-exploratory-data-analysis
- https://www.datacamp.com/community/tutorials/exploratory-data-analysis-python
- https://github.com/starlordvk/Prediction-Of-Cardiac-Arrhythmia/blob/master/CardiacArrhythmiaResults.pdf
- Classification of Cardiac Arrhythmias Patients by Azar Fazel,Fatima Algharbi, Batool Haider
- https://scientificdg.com/index.php/en/2017/01/07/standardization-vs-normalization/