Performance of different classifiers on heart disease detection dataset

Diwakar Sharma
Masters in Software Engineering
Department of Computer Science
Drexel University
ds3222@drexel.edu

Malhar Deshpande
Masters in Software Engineering
Department of Computer Science
Drexel University
md3239@drexel.edu

Abstract

The given problem is prediction of heart disease, to classify users from dataset of UCI data repository into 5 classes. Goal of the experiment is to classify users into classes which is non-zero(1, 2, 3 or 4) for severity of presence and zero for absence of heart disease and also to measure the correctness of our classifier.

Keywords: PCA Dimensionality Reduction, Linear SVM, RBF SVM, Stratified K-Means Algorithm

1 Introduction

There are 4 datasets that are provided by UCI data repository for Machine Learning for Heart Disease diagnosis. The datasets are Cleveland Clinic Foundation (cleveland.data) [4], Hungarian Institute of Cardiology, Budapest (hungarian.data) [1], V.A. Medical Center, Long Beach, CA (long-beach-va.data) [3] and University Hospital, Zurich, Switzerland (switzerland.data) [2]. All these datasets are real time datasets generated from patient records and they have around 76 attributes including a prediction attribute, which suggests regarding the severity of disease on scale of 1 to 4 or no presence at all. For our experiment purposes, we have used Cleveland database [4]. This database contains 76 attributes but for our experiment purposes we are using best 14 of those attributes. All attributes are numeric-valued.

2. Machine Learning Concepts Used

PCA Dimensionality Reduction:

PCA converts the data from high dimensional space to a lower dimensional space. It is a statistical method in which the observed raw data set, which might be correlated in nature, is converted into linearly uncorrelated variables. These uncorrelated variables are principle components. The components are ordered in the decreasing order of their variance in this representation. PCA can be computed by mean centering the data and performing an SVD on it and taking the first few columns (principal components) of the V matrix produced by SVD as the principal components.

$$X=U^VT$$

where, X is the data matrix, is the diagonal eigenvalue matrix and U and V are unitary matrices.

Linear SVM:

Support Vector Machines (SVM) is a supervised machine learning algorithm to solve multiclass classification problem. Given set of training examples, marked as one of the two classes, SVM algorithm builds model that assign each new example of the dataset to one of those classes.
Now, there are 2 kinds of problem. One those are linearly separable and other are non-linearly separable. For linearly separable problems, SVM uses a linear kernel which classifies dataset among different classes using a linear hyperplane.

RBF kernel SVM:

For non-linearly separable problem, SVM uses a RBF kernel which is a non-linear kernel function because no hyperplane is sufficient enough to accurately classify data.

Stratified k-fold Cross Validation:

Stratification is the process of rearranging the data as to ensure each fold is a good representative of the whole. The folds are selected so that the mean response value is approximately equal in all the folds.

3 Our Approach

Our approach was to try different classifier and try and compare which one of those works better for the given dataset of heart disease. The classifiers that we have compared are Linear SVM, Non-linear SVM and Stratified K-Mean on the given vector representation of Cleveland dataset [4]. For our experimental purposes, we have divided our experiment into 2 problems. For both problems, we try to run our classifiers for 60/40 and 80/20 splits where we use 60% and 80% for training our classifiers and 40% and 20% for testing their predictions respectively. Cleveland dataset [4] has 303 instances and 14 attributes. Our first step is to apply dimensionality reduction and for that purpose we have use Principal Component Analysis (PCA) with 5 components. Once the PCA has been applied on the original X value where X is the feature set, our feature set is reduced to X_new which is a vector representation of 303 samples x 5 features. First, we try out 60/40 split of X_new where 60% of X_new is used to train the SVM classifier and 40% of X_new is used to test. Similarly, next we try 80/20 split of X_new, which is problem 1 of our experiment.

Problem 1

We classify data using a Linear SVM and predict likelihood of disease belonging to a particular class of severity ranging from 1 to 4 i.e. least to most severe with value of C=0.001. Here, C is the penalty that the classifier incurs every time there is a misclassification that takes place so job of the classifier is to incur penalty as minimal as possible while classifying data in order to keep cost of classification at minimum. In order to check if other type of classifiers work better for this dataset than Linear SVM, we use a RBF i.e. non-linear kernel for the SVM classifier and classify data keeping value of C same. Similarly as last part of our problem 1, we use Stratified k-fold cross validation with 5 folds with a RBF kernel and keeping value of C same as for above classifiers in our search to find which classifier works better for this dataset. The results for this have been shown below in **Fig 1 below.**

Problem 2

For problem-2 of our experiment, we go a step further by predicting absence (zero) or presence (non-zero) of heart disease. This is possible because we group all severity classes (1 to 4) together which mean that a non-zero would indicate presence of heart disease and a zero would indicate absence of heart disease. Problem-2 of the experiment follows same procedure as that of problem-1. First step is dimensionality reduction for which we use PCA with 5 components that picks best 5 components out of 14 attributes. Now what we get is a vector representation as we obtained in problem-1, which basically implies 303 samples x 5 features. For problem-2, we use an 80/20 split where 80% of data is used to train classifier and 20% is used to test. Now, we follow the same procedure as we did for problem-1 we apply 3 classifiers i.e. Linear SVM, Non-Linear SVM with RBF kernel and Stratified k-means cross validation with 5 folds, all for a value of C=0.001. The results are shown in **FIG 2**

Table and figures below describe classifiers that we used and interprets results that they generated.

3 Result

Table 1: Problem-1 60/40 Data Split

Classifier	**Classification Accuracy (X_new)**	**Classification Accuracy (X_new- Split)**
Linear SVM	80%	55%
Non-Linear SVM (kernel 'RBF')	100%	49%
Stratified k-mean cross validation (kernel 'RBF')	100%	55%

Table 2: Problem-1 80/20 Data Split

Classifier	**Classification Accuracy (X_new)**	**Classification Accuracy (X_new- Split)**
Linear SVM	81%	59%
Non-Linear SVM (kernel 'RBF')	100%	57.37%
Stratified k-mean cross validation (kernel 'RBF')	100%	54.23%

Table 3: Problem-2 80/20 Data Split

Classifier	**Classification Accuracy (X_new)**	**Classification Accuracy (X_new- Split)**

Linear SVM	100%	73.77%
Non-Linear SVM (kernel 'RBF')	100%	57.37%
Stratified k-mean cross validation (kernel 'RBF')	100%	54.23%

Figure 1: Problem-1 PCA Dimensionality Reduction (n_components=5)

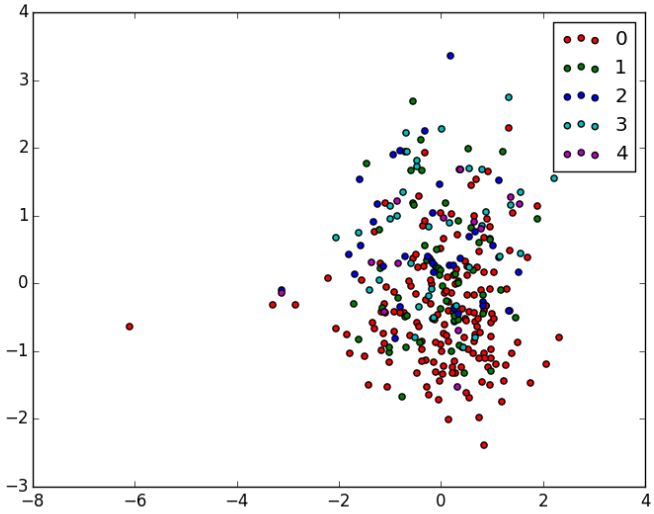
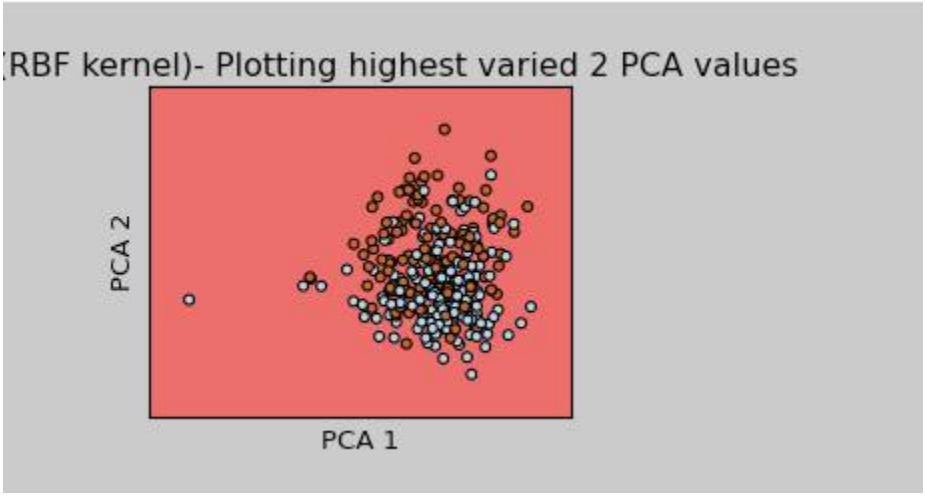


Figure 2: Problem-2 Integrating all severity classes into one i.e. presence (non-zero)



To our observation, with 80% of data for training and 20% data for testing and integrating all severity classes into one that would indicate presence of heart disease improved accuracy of our classifiers as compared to what we observed in problem-1.

Experimental Setup

We did our work entirely using the machine learning repositories provided by Scikit-Learn module, numpy module and matplotlib module of the Python programming language. The

Python version used was 2.7.5. In order to run the programs, just verify if Python Programming language version 2.7.3 and above is installed with the mentioned Machine Learning module dependencies in your system.

5 Conclusion

Based on the results shown above and experiments performed, it is evident that input data plays an important role in prediction along with machine learning techniques. As is seen in the dataset, provided, we have labels from 0 to 4 where the labels of 4 are hardly 13 and when we split the data into train and test, the number become very less which is nothing but noise and can be totally removed from the dataset by using filtering techniques and hence the linear model will be available to predict the outcome much better with absence of noise. Moreover, PCA has again proven that we can get rid of similar feature set and still obtain predictions with great efficiency. Moreover, we have conducted tests using non linear RBF kernel which is a normal first choice and then validating against linear SVC kernel which outperformed RBF in split case. Most importantly, the above experiment not only helped us in predicting the outcome but also gave us valuable insights about the nature of data, which can be used in future to train our classifiers in a much better way.

Acknowledgments

We would like to acknowledge and thank the colleagues at " Hungarian Institute of Cardiology, University Hospital, Zurich, University Hospital, Basel, V.A. Medical Center, Long Beach and Cleveland Clinic Foundation" for providing dataset which laid the foundation of our work. Also, we would like to thank Professor Rachel GreenStadt of Drexel University, who encouraged us to dive deep into the topics of machine learning. We are greatly thankful to the Victor Vapnik for the SVM concepts and to the team maintaining the Sklearn module of Python which helped us in using Machine learning repositories in tandem with Python Language.

References

[1] Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.

[2] University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.

[3] University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

[4] V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:Robert Detrano, M.D., Ph.D.

[5] http://scikit-learn.org/0.13/index.html

[6] http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[7] Jolliffe, I. T. (1986). Principal Component Analysis. Springer-Verlag. p. 487. doi:10.1007/b98835. ISBN 978-0-387-95442-4

[8] Andreas Müller (2012). Kernel Approximations for Efficient SVMs (and other feature extraction methods)

[9] Hsu, Chih-Wei; Chang, Chih-Chung; and Lin, Chih-Jen (2003). A Practical Guide to Support Vector Classification (Technical report). Department of Computer Science and Information Engineering, National Taiwan University.

[10] http://www.cs.cmu.edu/~schneide/tut5/node42.html