Comparison of Heart Disease Classifiers Using Principal Component Analysis

By: Rayhaan Rasheed

BME 6850: Pattern Recognition

Professor: Murray Loew, Ph.D

12/12/2018

Introduction

Heart disease, also known as Coronary Artery Disease (CAD), is the leading cause of death in the entire world. It is a condition in which blood flow is obstructed because the patient's arteries begin to narrow down. This leads to the patient having a myocardial infarction (MI), otherwise known as a heart attack. According to the American Heart Association, heart disease is the root cause for 1 in every 3 deaths in the United States It has claimed more lives than all forms of cancer combined.[1] There have many studies highlighting the various causes of heart disease like sex, age, cholesterol, high blood pressure, smoking, and much more.[2,3] Over the years, physicians have been able to combine knowledge about these causes with other patient-specific metrics to predict whether or not a patient has CAD. If doctors can predict heart disease early on, and automate that process, then there may be a chance to prolong a patient's life and mitigate the negative side effects.

Many techniques have been used to acquire the appropriate information as well accurately diagnose a patient. Currently, the most popular choice is machine learning methods, especially supervised machine learning. Yeh et al. in 2011 compared decision trees, Bayesian classifiers, and backpropagation neural networks to investigate the problem of heart disease diagnosis by data mining.[4] Another study, by Jabber et. al in 2013 used a K-Nearest Neighbor approach to classify heart disease based on similar genetic attributes.[5] This paper will continue the theme of utilizing a machine learning approach for diagnosing heart disease but look at the various levels of heart disease as well. The goal of this paper is to use multiple classification models in a multi-class setting using the original and a dimensionality reduced dataset.

Data

The data used in this project is the Heart Disease Dataset generated by Robert Detrano, M.D., Ph.D. at the V.A. Medical Center, Long Beach and Cleveland Clinic Foundation. The full database was pulled from the Machine Learning Repository created by the University of California Irvine. There are 76 attributes in total, yet only 14 are preferred in most of the literature.[6] These 14 features include 13 attributes and 1 target. The 13 attributes were chosen based on two feature selection methods: Computerized Feature Selection (CFS) and Medical Feature Selection (MFS).[3] The CFS features are derived from an algorithm-based feature selection. MFS is when a physician deems a specific attribute important to determining heart disease.[7] Nahar et al., in 2013 contributed the medically important features and proved that a combination of CFS and MFS features is important for effective classification of CAD.[8] The target feature contains five possible outcomes: 0,1,2,3,4. A patient in class 0 does not show any signs of heart disease; however, patients in class 1 through 4 show various levels of heart disease, with 4 being the most severe case.

Methods

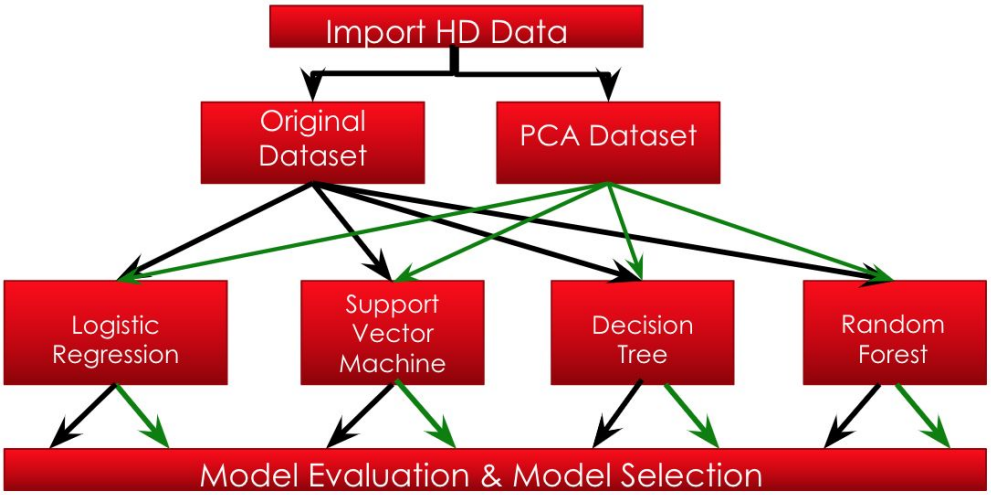

Figure 1: Flowchart of Approach

Classifier Construction

The objective of this project is to answer the question: based on the given dataset, does multiclass classification of heart disease work best on the original dataset or on a dimensionally reduced version. Since we want to compare one class against the other remaining four, a One-vs-All (OvA) approach is taken when constructing the classifier. For linear classifiers, this produces 4 OvA hyperplanes in the feature space. Based on literature and the characteristics of certain classifiers, this paper will use Logistic Regression, Support Vector Machine (SVM), Decision Tree, and Random Forest as the classifiers. As can be seen in Figure 1, eight classifiers are constructed and evaluated: four from the original dataset and the rest from the transformed dataset.

Principal Component Analysis (PCA)

PCA will be used as the dimensionality reduction technique to create the second dataset. This procedure converts a set of variables into a new set of literally, uncorrelated variables called principal components. The first principal component explains the largest possible variance, and the variance decreases for each principal component after. Dey et. al. in 2016 trained heart disease classifiers on various amounts of principal components and showed using six principal components produced the best accuracy.[9]

Training and Hyperparameter Tuning

Using methods such as 10-fold cross-validation, each classifier will be trained on different slices of the dataset to increase prediction power. Each classifier has a set of hyperparameters that can be optimized to ensure the classifiers are properly trained and constructed. These hyperparameters are unique to the classifier type and can vary based on the training data.

Model Evaluation

Each classifier will produce its own Receiver Operating Characteristic (ROC) curve. The curve plots the true positive rate (TPR) against the false positive rate (FPR). Since this is a multiclass setting, a ROC curve will be generated for each class. The goal is to evaluate overall model performance, so a micro-average ROC curve is constructed by aggregating the contributions of all classes to compute the average metric. The micro-average ROC is preferable in a multiclass setting because it mitigates error when there is a class imbalance. The area underneath the ROC curve, called the AUC, is an indicator for model performance. The greater the AUC the better the performance.

Results

Logistic Regression:

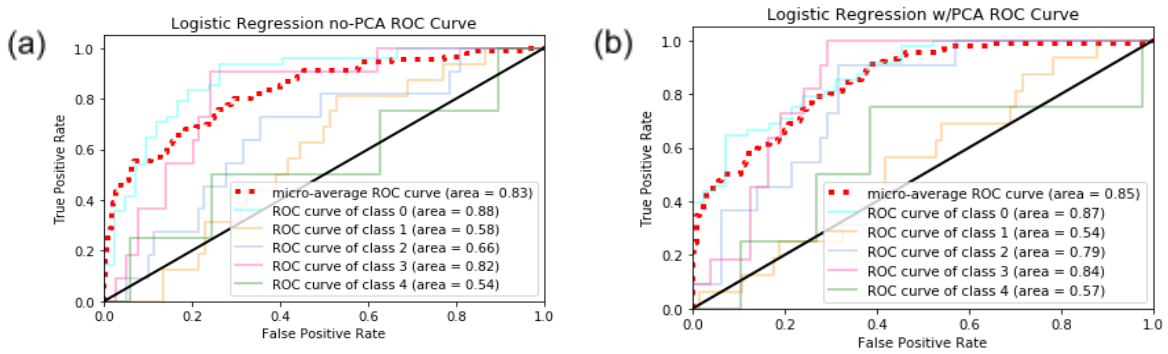

Figure 2: ROC Curve for Logistic Regression (a) original dataset (b) PCA dataset

Logistic Regression is the most widely used classification method for OvA problems. The model connects the two classes using a sigmoid function. A weighted sum is sent into the sigmoid function and then sent to a threshold function. The classifier works fairly well on the original dataset (see Figure 2a) with an average AUC of 0.83. This curve is a great example of how the micro-average ROC curve helps give an average calculation of model performance because the curve for class 0 is good, but it gets worse as go through different classes. Class 4 has the least amount of patients so the model did not have many to train with. Figure 2b has an average AUC of 0.85. This is only slightly better than the logistic regression using the original dataset. When looking at the individual classes, using the original dataset gave a better AUC for classes 0 and 1, but the PCA dataset produced better AUCs for classes 2,3, and 4.

Support Vector Machine:

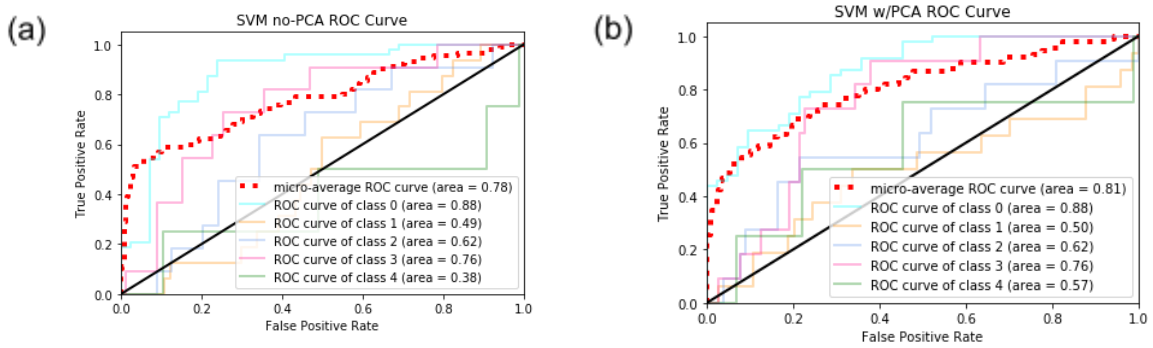

Figure 3: ROC Curve for Support Vector Machine (a) original dataset (b) PCA dataset

A Support Vector Machine utilizes a hyperplane that traverses across all available dimensions. This model is said to be useful in OvA situations especially when using high dimensional data. Figure 3a shows that using a SVM on the original dataset gives an average AUC of 0.78. Similarly, the average AUC for the SVM used on the PCA dataset is 0.81 (see Figure 3b). The SVM used on the PCA dataset not only has the better average AUC, but it also has the best AUCs for the individual classes.

Decision Tree:

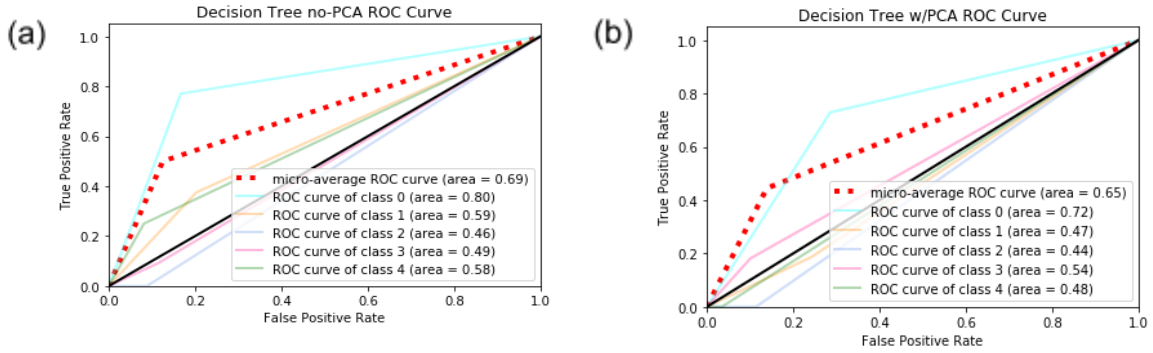

Figure 4: ROC Curve for Decision Tree (a) original dataset (b) PCA dataset

Decision trees use a tree model to come to conclusions based on a set of features. It learns a series of questions to infer the class labels of the samples. There are many types of decision trees, but only a Gini decision tree is used. The Gini tree splits the data to depending on what removes the most impurity. This continues until all that is left are the pure leaf nodes representing the class labels. Figure 4a contains an average AUC of 0.69, and Figure 4b contains an average AUC of 0.65. These values are not that great and do not produce good AUCs for the individual classes.

Random Forest:

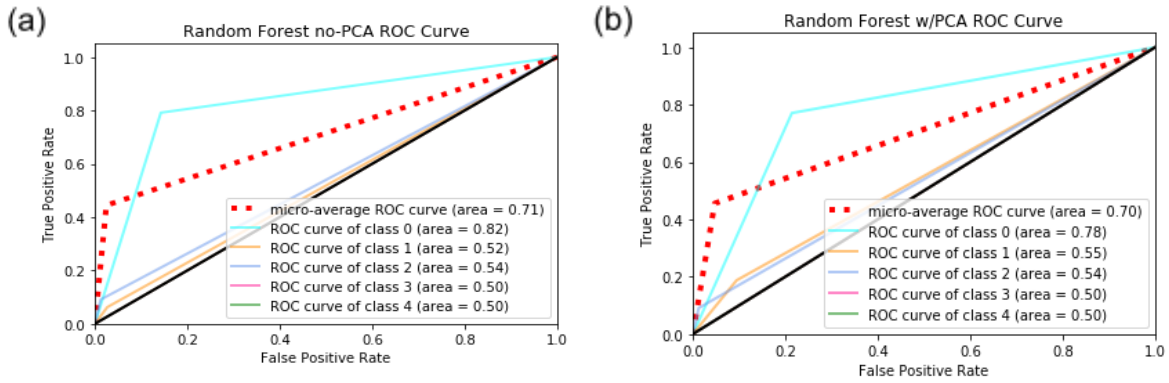

Figure 5: ROC Curve for Random Forest (a) original dataset (b) PCA dataset

Random Forest classifiers are an ensemble of decision trees working together to assign certain class labels to a set of samples. Figure 5a has an average AUC of 0.71, and Figure 5b has an average AUC of 0.70. There is no significant difference between the two values, and there is not much of a difference between the individual class ROC curves as well.

Conclusion

The Logistic Regression classifier outperformed all the other classifiers on the original and transformed dataset. Given more samples to train with and reduce the class imbalance, the classifiers could have better performance and accuracy. Looking at the difference in performance between the original dataset and the PCA dataset, it is pretty apparent that there is no significant difference between classifier performance given these datasets. This is most likely due to the fact that the 14 features have already been pre-selected based on their high impact on classifying a patient's CAD. If the raw 76 attribute dataset was available, then comparing the three datasets would be an interesting project. Many models have been made to tackle the problem of whether a person is positive or negative for heart disease. Taking the problem one step further and determining the severity of a patient's heart disease will eventually lead to better patient care and personalized medical treatment.

References

[1] Emelia, B. Correction to Heart Disease and Stroke Statistics—2018 Update: A Report From the American Heart Association. Circulation 135, (2018).

[2] Heller, R. F., Chinn, S., Pedoe, H. D. & Rose, G. How well can we predict coronary heart disease? Findings in the United Kingdom Heart Disease Prevention Project. Bmj 288, 1409–1411 (1984).

[3] Shahwan-Akl. Cardiovascular Disease Risk Factors among Adult Australian-Lebanese in Melbourne. International Journal of Research in Nursing 1, 1–7 (2010).

[4] Yeh, D.-Y., Cheng, C.-H. & Chen, Y.-W. A predictive model for cerebrovascular disease using data mining. Expert Systems with Applications 38, 8970–8977 (2011).

[5] Jabbar, M. A., Deekshatulu, B. & Chandra, P. Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm. Procedia Technology 10, 85–94 (2013).

[6] Detrano R (1988), Heart Disease Data Set, V.A. Medical Center, Long Beach, and Cleveland Clinic Foundation, Retrieved from https://archive.ics.uci.edu/ml/datasets/Heart+Disease.

[7] N.Bethel, T. Viswanadha, S. A Knowledge driven Approach for Efficient Analysis of Heart Disease Dataset. International Journal of Computer Applications 147, 39–46 (2016)

[8] Nahar, J., Imam, T., Tickle, K. S. & Chen, Y.-P. P. Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. Expert Systems with Applications 40, 96–104 (2013).

[9] Dey, A., Singh, J. & Singh, N. Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of Attributes using Principal Component Analysis. International Journal of Computer Applications140,27–31 (2016).