Topic 8: Ensemble Methods

Jonathan Ibifubara Pollyn

GitHub Repository: https://github.com/JonathanPollyn/Machine-Learning-for-Data-Science

College of Science, Engineering and Technology, Grand Canyon University

DSC-540: Machine Learning for Data Science

Dr. Aiman Darwiche

December 22, 2021

Introduction

This project requires the implementation of an experiment by obtaining the data and developing a classifier (in Python) that implements an ensemble architecture, following the procedures outlined in the article. Fulfilling all the models specified in the report are not required due to how challenging it is. As stated by the authors, the purpose of the paper is to see if using ensemble learning algorithms improves physical activity recognition accuracy compared to using single classifier algorithms. The authors affirmed that three different data sets were employed in the investigation, one of which contained wrist-worn accelerometer data. A four-step classification framework was created for each data set: data preprocessing, feature extraction, normalization, feature selection, and classifier training and testing. For the custom ensemble, three decision fusion approaches were used to combine the results of the separate classifiers: weighted majority vote, nave Bayes combination, and behavior knowledge space combination. Classifiers were cross-validated using leave-one-subject cross-validation, and average F1 scores were used to compare them.

The authors stated that the ensemble learning approaches consistently beat individual classifiers in all three data sets. Random forest models always offered excellent activity recognition among the standard ensemble approaches; however, the custom ensemble model utilizing weighted majority voting had the greatest classification accuracy in two of the three data sets.

Application

In an effort to re-implement this report, the first action was to gain access to the dataset, which was provided as a link to the article. Once the dataset has been loaded, next was to load all nine .dat files into python. After randomly checking some files, all nine files do not have headers, which indicates that there is a lot of work to get the data suitable for building a model.

By using the PDF provided regarding information about the dataset, each data file contains 54 columns per row, and the columns are listed below1: timestamp (s)

2: activityID (see II.2. for the mapping to the activities)
3: heart rate (bpm)
4-20: IMU hand
21-37: IMU chest
38-54: IMU ankle

Focusing on one file (subject101.dat), I used various statistical methods in python to understand the data when it shows that almost 90 percent of the heart_rate information is missing and missing some of the other data. I was forced to drop the heart_rate column from the dataset with this knowledge. Next, null values were checked to understand how many null values are in the data, and it was clear that most of the values in the datasets are null in similar column types. Using these columns will not add any value to the model, so excluding such columns (IMU_hand, IMU_chest, IMU_ankle) from the model will provide accurate reporting.

Finally, after pre-processing the data, the dataset was split into training and test (X and y) which the training contained 70 percent of the data. There is now a target variable and data to predict the target variable. The data is then standardized in other to build the ensemble model. After importing the required modules, like Random Forest Classifier, AdaBoost Classifier, Gradient Boosting Classifier, xgboost, and bagging. By using the accuracy_score from the sklearn matrices, the accuracy of each model was obtained, which shows that the RandomForest Classifier model has an accuracy of 0.9999557228 as the best model when compared to the other model. The obtained accuracy aligns with the author's obtained result, which states that random forest models consistently offered excellent activity recognition among the standard ensemble approaches; however, the custom ensemble model utilizing weighted majority voting had the most classification accuracy in two of the three data sets used. Figure 1 shows the plot of all

created models, and we see that the RandomForestClassifier came out as the best Ensemble model compared to the others created.

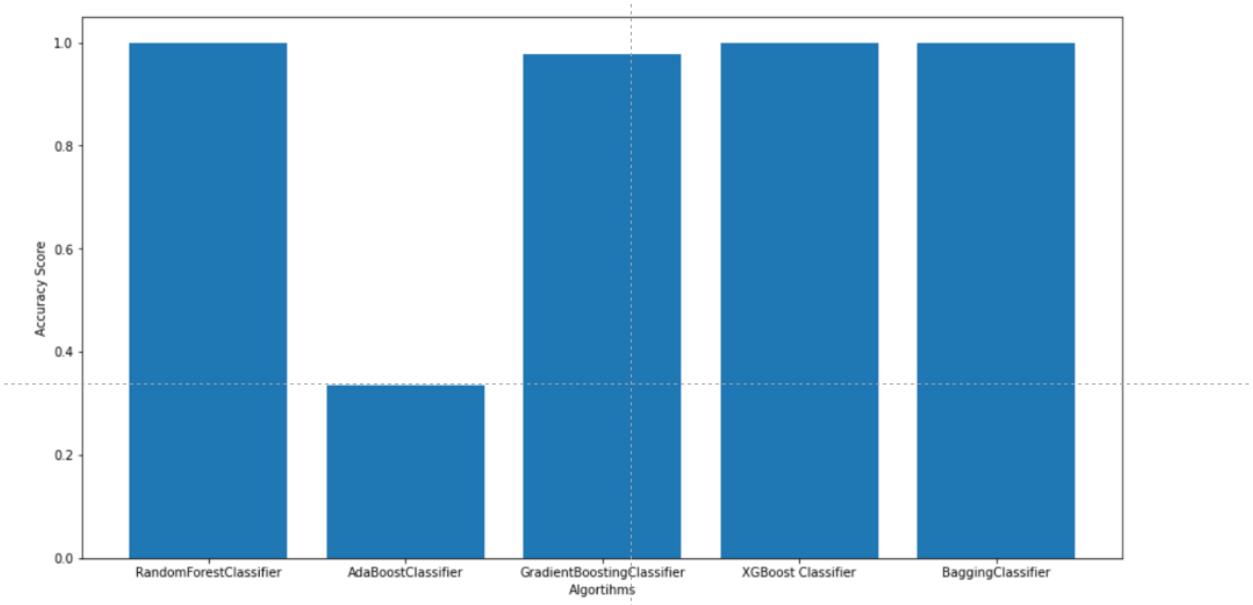

Figure1: Plot of all created models

References List

Chowdhury, A., Tjondronegoro, D., Chandran, V., & Trost, S. G. (2017). Ensemble methods for classification of physical activities from wrist accelerometry. *Medicine & Science in Sports & Exercise*, *49*(9), 1965–1973. https://doi.org/10.1249/mss.0000000000001291