Shengxin Ding
Shijin Kuai
Yaruo Zhang

Project 1 Report

Methodology:

To compare the performance of the classification algorithms, we first clean the data, and the subtract training data and test data from all data set. Then we build model to compare the performance of the seven classification algorithms on the 54 datasets. For each model we use gridSearch to find the best parameter, with cross-validation = 5. To pick the best model, we run each model on the test data and record the test score, which is Accuracy and compare the test score among the classification algorithms. We choose algorithms that have highest accuracy for each data and put all results into a table. Multiple models can all have high accuracy. At last, we find the classification algorithms that appear most in the final results (shown as figure 1-1).

First we take a look of data to see if there is anything to be paid attention to when we clean them, like some data sets begin their content in line 6 so we have to use skiprows function in pandas.read_csv to subtract information in the correct line.

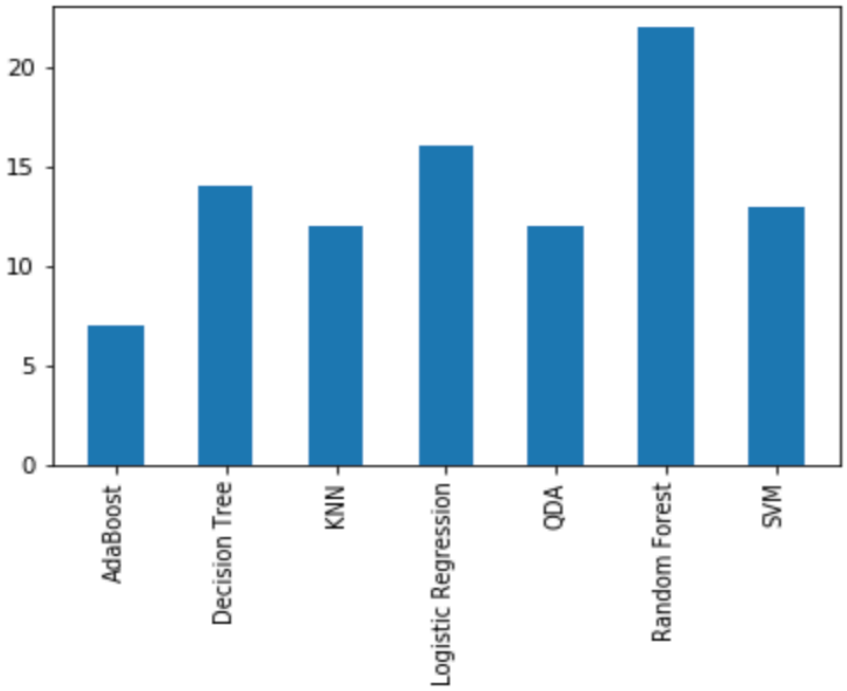

figure 1-1

After considering all situations, we should take care of such as skiprows, NA value and separators, we wrote two functions process 54 datasets to make sure all of them are illustrated as names together with first line of each one. The first function is to extract information we need to show data sets such as header, categoric indices, header and so on, the second one is to use above information within this function to open each data set. After running these two functions, we could obtain information of each data set with a for loop. Among the seven models' performance, we find that the random forest model works the best.

Why Random Forest works so good on most data sets (Key findings):

We think the reason why Random Forest is among the best models is it produces a classifier with high accuracy. The model can handle a large number of input variables and can assess the importance of variables when deciding on the category. In addition, the model contains a good way to estimate the lost data and, if a large portion of the data is lost, it can still maintain accuracy. At last, it provides an experimental method to detect variable interactions. And for the disequilibrium classification data set, it can balance the error.

Figure 2-2 shows the accuracy of how each classification algorithm fit in each dataset. Most data sets achieve best accuracy in random forest model.

Dataset	AdaBoost	Decision Tree	KNN	Logistic Regression	QDA	Random Forest	SVM
acute-inflammations-1	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
acute-inflammations-2	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
balance-scale	0.824000	0.752000	0.872000	0.832000	0.896000	0.824000	0.864000
banknote-authentication	0.974545	0.992727	1.000000	0.992727	0.978182	0.996364	1.000000
blood-transfusion-service-center	0.793333	0.746667	0.740000	0.746667	0.440000	0.740000	0.726667
breast-cancer-wisconsin-diagnostic	0.894737	0.964912	0.929825	0.964912	0.956140	0.956140	0.631579
breast-cancer-wisconsin	0.970803	0.948905	0.985401	0.985401	1.000000	1.000000	0.992701
breast-cancer-wisconsin-prognostic	0.666667	0.564103	0.589744	0.743590	0.717949	0.666667	0.717949
car-evaluation	0.806358	0.927746	0.852601	0.907514	0.170520	0.907514	0.893064
chess-king-rook-vs-king-pawn	0.925000	0.993750	0.967187	0.978125	0.582812	0.993750	0.925000

figure 2-2

Shengxin Ding
Shijin Kuai
Yaruo Zhang

Limitations in our project could be listed as follows:

First of all, how to set parameters that make sense given context each data set in has been a big problem through whole model process. Parameters we chose are based on information we found on http://scikit-learn.org/stable/index.html, those parameters are commonly used and welcomed in real-world application of testing models, but they might not make sense in data set we runs models. Also, the range of parameters in our models might not wide enough to generate radical change in model accuracy. With analysis it is reasonable to see a large increase in accuracy with range of 1 to 100 for a specific parameter, but range is only set between 1 to 10 in our model. We are not sure about every threshold of parameters in our model, so we conclude that such issue is a highly potential limit of our project.

In addition, we do have chances of improving performance of models if we are given more time to study on models and run them several more times. It is possible that we could obtain better scores and results with more time and efforts on that. But with limited time, this is the best model and function, we come up with, to run data sets given.

Compare the use of ML methods in this project against typical ML applications:

Compared to typical ML applications, the ML methods in this project aim to put each individual variable from the population under study into many classes. The goal is to help predict the result. Classification helps analysts to use measurements of an object to identify the category to which that object belongs. Data consists of many examples of objects with their correct classification During the project, we use data to establish an efficient rule. We use functions to grab the best parameters, in other words, independent variables. We use cross validation to check the process. The ML methods may work well in classification and prediction fields.