What *is* the Best Classifier?

By: David Rady

Abstract

This report investigates one main goal. For a given selection of data sets, can we say what is the 'best' classifier in terms of good predictions? How much does the answer depend on the particular selection of data sets? How much does the answer depend on our computational constraints? We investigate these questions using data sets from the UCI repository.

1. Introduction

This report seeks to provide insight into the main goal of the project at hand.

Those who are unfamiliar with the complexity of machine learning might wonder *"What is the best classifier?"*. The goal of this project was to deep dive into how simple this question is to answer. The performance of eight different classifiers were investigated on numerous data sets. These models were fine-tuned to perform as best as possible across all the data sets, while keeping in mind the computational and time constraints.

Consistency in the code was maintained as best as possible, especially between the classifiers. The two pipelines were performed in the same way and many functions were reused or altered while maintaining maximum similarity.

2. Methodology & Experimental Results

The first steps were to download the data sets and load them into our scripts, either using functions or a few lines of code. The data was then preprocessed the same way, using a train-test-split with a 70% training - 30% testing split, with random state of 0 for reproducibility. This was followed by scaling of the data using sklearn's StandardScaler.

Prior to determining the best model of each type, the models first had to be tuned to improve performance. A few parameters were chosen, generally 2 per model, with multiple suitable values for each parameter. Then a grid search with 3-fold cross validation was performed on each type of model, for each data set. Once the grid search was completed, the scores of the best estimator of each grid search were used for analysis.

As for analyzing the classification performance, accuracy was used as the main metric. Although metrics such as 'recall' are much preferred when dealing with medical diagnoses due to the desire for less false negatives, accuracy was used so that the scores could be compared equally across all data sets. To be able to determine the performance gains from parameter tuning, a dummy model with default values was trained for each type of model for comparison.

2.1 Classification

The classification experiments were performed over eight data sets; 'Diabetic Retinopathy', 'Default of Credit Card Clients', 'Breast Cancer Wisconsin', 'Statlog (German Credit Data)', 'Adult', 'Yeast', 'Thoracic Surgery', and 'Seismic Bumps'. An additional data set, 'Zoo Animal Classification', was added as part of the novelty component. The eight different classifier models investigated were Logistic Regression, SVM, Decision Trees, Random Forests, K-nearest neighbors,

AdaBoost, Gaussian Naive Bayes, and Neural Network.

For the observations, it was noticed that the scores for the models varied mostly between the datasets themselves instead of between the models. Figure 1 in section A. of this report shows this variation between data sets very well using a heatmap. This can possibly be attributed to the fact that some of the data sets were unbalanced, having many more instances of a given target class than others.

As for the comparison of the scores between models, Table 1 shows that the Random Forest Classifier has the highest average score over all the data sets (83.894%). However, this score was only a negligible percentage higher than some of the others such as Logistic Regression (83.834%). The models performed fairly consistently compared to each other, with the exception of the Gaussian Naive Bayes Classifier, which generally had the lowest score of all the models.

In terms of the comparisons of the tuned models vs. the dummy models with default values, the tuned models generally performed much better than the dummy models. However, there were some instances such as the Random Forest Classifier where the dummy model outperformed the tuned model. This might be because the default values have no limit on the depth, whereas the tuned model did. This might also be attributed to the cross validation, which withholds some of the training data each time and returns the mean score of this.

In general, the classification experiments led to the conclusions that different models and parameters perform differently on different data sets. Some models such as Support Vector Machines perform better on binary output data such as the Breast Cancer Wisconsin data set, which can be seen in Table 1.

3. Conclusions

In conclusion, it was found that the best classifiers cannot simply be determined. Some of the data sets had very high or low scores for all models. It was clear that some models were better than others depending on the data, for example if the target classes had more than 2 classes or not. The scores of the models were also limited by the computational constraints of not being able to test many more parameters or parameter values in the grid search. Looking back, this could be resolved by taking multiple sequential subsets of the larger datasets and studying them together.

Essentially, the main goal of this project was achieved, since we now know the question *"Which classifier is best?"* does not have a straightforward answer. While the best performing model was Random Forest, certain models perform better on certain datasets, such as the SVM's on binary data. Therefore, the answer would really depend on many aspects, mainly the data being used.

A. Detailed experimental results

	Logistic Regression	SVM	Decision Tree	Random Forest	K-nearest	AdaBoost	Gaussian NB	Neural Network
Diabetic Retinopathy	73.41	73.121	63.584	67.341	66.763	67.341	45.087	73.41
Default of Credit Clients	81.078	82.344	82.211	82.378	81.022	82.478	80.144	82.111
Breast Cancer Wisconsin	96.491	98.246	94.152	97.076	95.322	97.661	91.813	94.737
Statlog (German Credit Data)	77.333	77.667	71.333	75.333	73.0	75.667	75.667	78.333
Adult	81.678	76.377	83.472	83.048	80.499	77.919	80.216	79.614
Yeast	59.417	58.52	59.641	64.574	58.072	43.498	58.296	59.193
Thoracic Surgery	93.617	93.617	93.617	94.326	93.617	93.617	90.071	94.326
Seismic Bumps	94.716	94.716	94.716	94.201	94.459	94.716	88.918	94.459
Zoo Animals	96.774	96.774	93.548	96.774	96.774	93.548	93.548	96.774
Average Score	**83.834**	**83.486**	**81.808**	**83.894**	**82.169**	**80.716**	**78.195**	**83.661**

Table 1: Classifier Test Scores on Different Data Sets (figures shown in %)

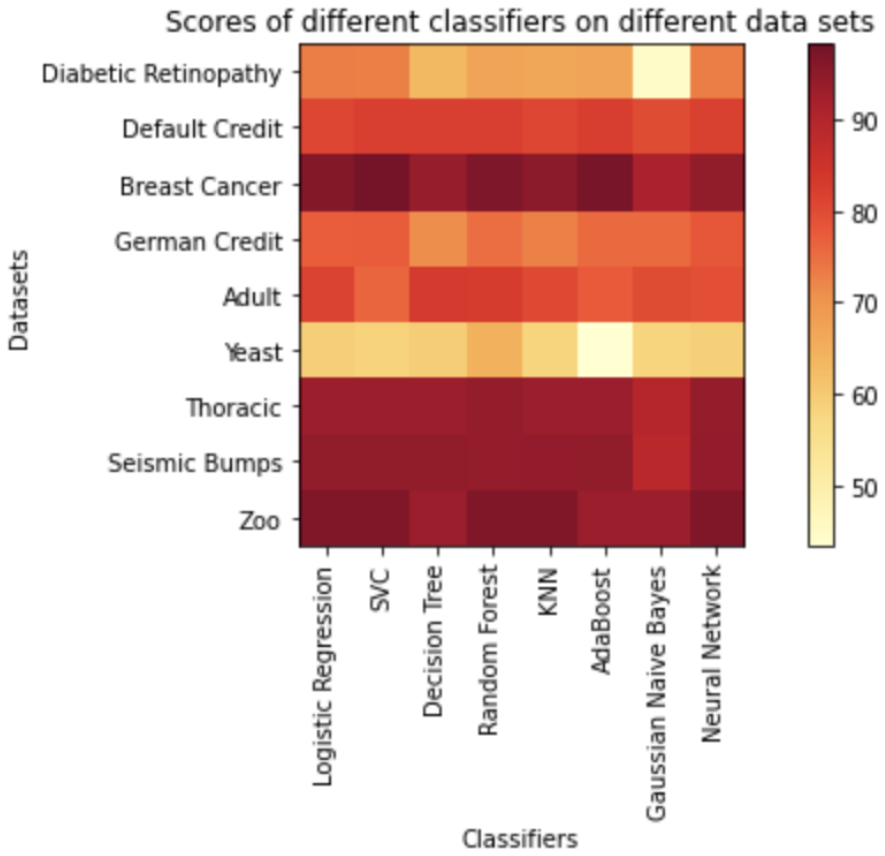

Figure 1: Heatmap of Classifier Test Scores on Different Data Sets