A Replication of "An Empirical Comparison of Supervised Learning Algorithms"

By: Adrianna Hohil
PID: A15139284

Abstract

In recent years of machine learning research, there have been a multitude of new supervised machine learning algorithms introduced to the field. They have revolutionized the ways we explore and model data, make mathematical and real-world predictions, and greatly enhanced our everyday technology. However, little research has been done to evaluate the performance of supervised machine learning algorithms. Therefore, this project's purpose is to assess and compare the performances of supervised machine learning algorithms such as KNN, logistic regression, and random forests across various datasets and analyze performance through accuracy and t-tests.

Introduction

The purpose of this project is to assess the performance of KNN, logistic regression, and random forests by replicating the conditions of the Caruana paper on a much smaller scale. This project uses three of the UCI datasets cited in the paper known as the adult dataset, cover type dataset, and letter recognition dataset. Classifier performance will be assessed based on their ability to classify binary data on the majority of the dataset after two trials of training on 5000 random samples of the data. It is hoped that my project will not only be able to replicate that of the Caruana paper, but will validate their findings and conclusions.

Method

To limit the scale of this project in relation to the Caruana paper, evaluation of performance of only three algorithms was assessed across 3 datasets instead of eight algorithms across 11 datasets. Each algorithm was required to complete a binary classification task for each dataset.

Algorithms

For KNN, 26 values for k were used ranging from k = 1 through k = 500 with a step of every 20. As stated in the paper, "we use KNN with Euclidean distance and Euclidean distance weighted by gain ratio." The string "uniform" represents uniform weights, meaning all points in each neighborhood were weighted equally. For the string "distance," closer neighbors of a query point will weigh points by the inverse of their distance.

For logistic regression, I trained unregularized and regularized models, varying the ridge regularization parameter by factors of 10 from 10^-8 to 10^4.

For random forests, the forests have 1024 trees. The size of the feature set considered at each split is 1,2,4,6,8,12,16 or 20 - except for the letters dataset, which will be explained in the experiment section.

All model means were set to 0 and all standard deviations were set to 1.

Dataset

All datasets were converted to binary for algorithm classification.

For the adult dataset, the "Yearly Outcome" column stating whether a person's salary was above 50,000 dollars or below 50,000 was converted to binary using the label binarizer. This was known as the test set, while all other columns excluding "Yearly Outcome" were the training data.

For the letter recognition dataset, the "Capital Letter" column that identifies which letter there is was remapped using the Letter 2 technique from the paper - the values of letters A through M were set to 0 whereas the values of letters N through Z were set to 1.

For the cover type dataset, the "Cover_Type" column contained 7 different types. In the paper, the type with the highest number of attributes was set to one value while all other types were set to another value. However, the number of columns in the dataset was so large (581,011 columns) that to shrink the dataset I found it better to keep the type with the highest number of attributes (type 2) and set it to 1, but then drop 5 other types and keep 1 (type 1) to set to 0. This change only shrunk my dataset to 363,427 columns.

After all 27 trials were completed, each algorithm's binary classification performance was determined based on the highest mean accuracy across datasets.

Experiment

In this project, the three classifiers chosen were K nearest neighbors (KNN), logistic regression, and random forests which were each trained on a random 5000 samples out of each UCI dataset used in the Caruana paper: the adult dataset, the cover type dataset, and the letter recognition dataset. Next, the optimal hyperparameters were found for each classifier through a 5 fold cross validation. Then, each classifier was retrained on another random 5000 samples before measuring each model's performance on the test set. This series of 3 trials per dataset was repeated for each classifier, collecting a total of 27 trials. However, I was curious to see if the results would vary if an extra trial was added, thus 4 trials x 3 classifiers x 3 datasets = 36 trials. After comparing the accuracies of each classifier, it was found that the random forest classifier performed best on 2 out of the 3 datasets.

NOTE: After setting the hyperparameters for each classifier and attempting to run each one on all three of my datasets, I ran into a methodological problem. I had to decrease my max features of my random forest classifier to 16 for my letter recognition dataset. This dataset only had 16 columns, so a split of 20 was not possible to run. This finding and adjustment is curious, because the paper reports using 20 features for all of the datasets from the UCI repository, yet it is computationally impossible to do so on

the letter recognition dataset. This made it impossible to fully replicate the paper with the same hyperparameters, and made me question what other flaws the paper might have.

Results and discussion

Main Results

Table 1: Mean Test Set Performance for Classifiers across Datasets

	KNN	Random Forest	Logistic Regression
Accuracy	0.843216	0.8634416	0.7764083

Table 2: Mean Test Set Performance across Classifiers per Dataset

		KNN	Random Forest	Logistic Regression
Adult	Accuracy	0.817525	0.8106	0.8224
Cover Type	Accuracy	0.792	0.831975	0.781725
Letters	Accuracy	0.920125	0.94775	0.7251

Secondary Results

Table 3 (Main Matter Table): Mean Optimal Train Set Performance per Trial per Dataset

		KNN	Random Forest	Logistic Regression
Adult	Accuracy	Trial 1: 0.8140 Trial 2: 0.8248 Trial 3: 0.8204 Trial 4: 0.8142	Trial 1: 0.8694 Trial 2: 0.8760 Trial 3: 0.8770 Trial 4: 0.8666	Trial 1: 0.8274 Trial 2: 0.8346 Trial 3: 0.8310 Trial 4: 0.8278
Cover Type	Accuracy	Trial 1: 0.9582 Trial 2: 0.9554 Trial 3: 0.9574 Trial 4: 0.9550	Trial 1: 0.9598 Trial 2: 0.9670 Trial 3: 0.9638 Trial 4: 0.9640	Trial 1: 0.7828 Trial 2: 0.7900 Trial 3: 0.7820 Trial 4: 0.7808
Letters	Accuracy	Trial 1: 0.8140 Trial 2: 0.8248 Trial 3: 0.8204 Trial 4: 0.8142	Trial 1: 0.8694 Trial 2: 0.8760 Trial 3: 0.8770 Trial 4: 0.8278	Trial 1: 0.8274 Trial 2: 0.8346 Trial 3: 0.8310 Trial 4: 0.8666

Appendix 1: Raw Classifier Test Scores per Trial per Dataset

		KNN	Random Forest	Logistic Regression
Adult	Accuracy	Trial 1: 0.8196 Trial 2: 0.8187 Trial 3: 0.8146 Trial 4: 0.8172	Trial 1: 0.8086 Trial 2: 0.8119 Trial 3: 0.8125 Trial 4: 0.8094	Trial 1: 0.8220 Trial 2: 0.8219 Trial 3: 0.8196 Trial 4: 0.8261
Cover Type	Accuracy	Trial 1: 0.7914 Trial 2: 0.7926 Trial 3: 0.7873 Trial 4: 0.7967	Trial 1: 0.8308 Trial 2: 0.8307 Trial 3: 0.8342 Trial 4: 0.8322	Trial 1: 0.7811 Trial 2: 0.7831 Trial 3: 0.7837 Trial 4: 0.7795
Letters	Accuracy	Trial 1: 0.9167 Trial 2: 0.9078 Trial 3: 0.9395 Trial 4: 0.9165	Trial 1: 0.9447 Trial 2: 0.9422 Trial 3: 0.9602 Trial 4: 0.9439	Trial 1: 0.7236 Trial 2: 0.7244 Trial 3: 0.7287 Trial 4: 0.7237

In all cases, each classifier was trained to a high level of accuracy. For example, as seen in Table 1, the mean test set performance across datasets for KNN was 84.32%, for random forest was 86.34%, and for logistic regression was 77.64%. In those cases, the testing data was validated to a very high degree but was limited to the performance of the classifier on the training data set. That is to say, the classifier still performed well, but just not as well as the training dataset (see notebook). This is expected because there is obvious variability in the dataset that was used for validation considering that the classifiers were trained on such a small percentage of the dataset. Also, I can conclude that the algorithms were not overtrained (overfitting did not occur) because results from the validation were extremely high which demonstrates that the classifier generalized extremely well against data that it hadn't been presented with before.

For the adult dataset, the logistic regression classifier outperformed the KNN classifier and random forest classifier. However, for both the letters and cover type datasets, the random forest classifier outperformed the KNN classifier and the logistic regression classifier.

The Random Forest classifier outperformed the KNN classifier and the logistic regression classifier on two out of the three datasets. This was most likely because it performs well on binary data and the input features were binary for the testing data.

Appendix 2.1: P-values for Classifier Comparison across Datasets

	KNN & Random Forest	Random Forest & Logistic Regression	Logistic Regression & KNN
Accuracy	p = 0.744313	p = 0.163730	p = 0.238473

Appendix 2.2: P-values for Classifier Comparison per Dataset

		KNN & Random Forest	Random Forest & Logistic Regression	Logistic Regression & KNN
Adult	Accuracy	p = 0.00303	p = 0.000377	p = 0.031024
Cover Type	Accuracy	p = 1.35787	p = 1.69308	p = 0.0033268
Letters	Accuracy	p = 0.013361	p = 3.75724	p = 1.28807

When comparing the p-values for pairs of classifiers across datasets, there were no values less than 0.05, and therefore no statistically significant findings. Thus we cannot reject the null hypothesis, and all means are the same.

However, there were varied results when comparing classifier p-values per dataset (as seen in Appendix 2.2). P-values for KNN & Random Forest for the cover type dataset, Random Forest & Logistic Regression for the cover type dataset, and Logistic Regression & KNN for the letters dataset were too large, thus not allowing us to reject the null hypothesis. However, p-values for KNN & Random Forest for the letters dataset and Logistic Regression & KNN for the adult dataset appear statistically significant (with scores of 0.031 and 0.013), thus leading us to believe we may be able to reject the null hypothesis. However, p-values for the remaining comparisons of KNN & Random Forest for the adult dataset, Random Forest & Logistic Regression for the adult dataset, and Logistic Regression & KNN for the cover type dataset were all so close to zero that it's unlikely the statistical significance present is due to chance.

Conclusion

Overall, it was challenging yet exciting to replicate the structure and design of the Caruana paper to explore the performance of KNN, logistic regression, and random forests. Given the few statistically significant findings in my results section, the methodological flaw with random forests, and finding that random forests outperformed the other two classifiers, I would be curious to explore further. I'm most curious about the statistical significance of the p-values for each algorithm's performance, as there was a lot of variability in results but promising statistical significance and I want to know why. As for exploring algorithm performance further, I think using more performance metrics to compare classifier performance as well as different types of datasets that were not random, but instead tended to each classifier, would be interesting to study.

References

Caruana, R.. “Multitask Learning.” *Machine Learning* 28 (2004): 41-75.