Classification Comparisons of Supervised Learning Algorithms

Chih Hsien Chen
University of California, San Diego
chc455@ucsd.edu

Abstract

In the following work we revisit a subset of the comparisons made by Caruana and Niculescu-Mizil[2] on a set of five supervised machine learning algorithms to measure varying degrees of performance on three different data sets. Our results remained largely consistent with the earlier proven results with Random Forest performing the highest across the board which indicates reliability in application to different data.

1 Introduction

The comparisons made by Caruana and Niculescu-Mizil were conducted to demonstrate the No Free Lunch Theorem measuring the effectiveness of several vetted machine learning algorithms[4]. Although there were varying datasets where different algorithms shined, the report did provide a look at a few consistently high performing algorithms.

In this report, we compared five of the ten implemented algorithms and controlled for a few specific parameters optimized using grid search and cross validation. Our comparisons wanted to look at the foundational differences between the algorithms so several of the bagged and boosted methods that scored consistently high were omitted. We also avoided neural networks because we predicted that the limited scope of our datasets would not yield appropriately correlated results with the many specific parameters that would need adjusting for the model implementation.

2 Methodology

2.1 Learning Algorithms

We explored the accuracy scores of each algorithm with the following range of parameters. Each model was tuned with grid search and cross validation over three runs selecting the highest occurring parameter. These models were all selected from the highly dependable Scikit-learn library which provided ease of implementation and high levels of customization[1].

SVM: A 'C' parameter of a range from 10^{-3} to 10^{3} was chosen to test several margins in which the best results would arise.

Linear Regression: A 'C' parameter of a range from 10^{-8} to 10^{5} was chosen to test several regularization strengths in which the best results would arise.

Random Forest: We tried all subsets of attribute lengths for each dataset to see the best number of features to split for the best results.

KNNs: A range of 1 to 7 neighbors was selected for our *K* value to select well performing boundaries that would yield the best results.

Decision Trees: We tested several values for our decision trees from a range of 1 to 6 to find the optimal depth that would yield the best results.

2.2 Data Sets

Our chosen datasets from the UCI Machine Learning repository[3] were selected by the following criterion: reliability between various algorithms mean that the labeled data should be used for binary classification purposes or modified to perform classification purposes using a *One vs Rest* setting. The ratio of the dataset itself against the number of attributes had to be in a limited range due to computational constraints. Several larger datasets using over 50,000 data points and over 200 attributes exceeded the capacity of the training machine.

Each dataset was divided into sets of 20/80, 50/50, and 80/20 respectively to training and testing data. The data was shuffled between each split to eliminate any potential confounds between trials. Any categorical labels were converted to numerical values for appropriate calculations.

The HTRU2 dataset describes a sample of 17,898 pulsar candidates with 9 different attributes describing observed shapes to perform a classification analysis [5].

The mushroom dataset contained 8124 instances of different mushroom samples and 22 attributes corresponding to 23 species of gilled mushrooms. The attributes were converted from categorical values to numerical values before use in training.

The wilted dataset contained 4889 instances of aerial images of diseased and healthy tree areas with 6 attributes describing different landscape features. This dataset also contained categorical data converted into appropriate numerical values for training use.

3 Results and Discussion

Table 1.

	20/80	50/50	80/20	Avg.
SVM	.989	.9906	.9903	.990
LR	.9881	.9902	.9894	.9892
KNN	.9869	.9894	.9893	.9888
RF	.9904	.9921	.993	.9918
DT	.9863	.9892	.9906	.9887

In order of highest overall performance, the mean accuracy calculated over the three dataset and trials ranked random forests first followed by SVMs, logistic regression, K nearest neighbors, and decision trees. Our overall results remained rather consistent with the findings by Caruana and Niculescu-Mizil with the exception of logistic regression rounding out the least accurate in the earlier report.

What is perhaps most interesting is how well each model performed on the mushroom data set over each trial where Decision Trees fared the worst but not by much. This particular dataset had mostly categorical features which we one hot encoded expanding our columns to 96 which may explain the high accuracy given the feature to case ratio. To see details of reported accuracies of each dataset over a series of trials, please refer to the tables in the *Extras* section.

4 Conclusion

The comparisons made in this report show relatively strong accuracies across the board on these vetted datasets but an important takeaway is how significant the role of some of the hyperparameters play in achieving higher accuracies. Larger training sets will undoubtedly yield better performance but important care must be taken in consideration of appropriate parameter measures for various ratio sizes of samples and features.

References

[1] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

[2] Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings Of The 23Rd International Conference On Machine Learning - ICML '06*. doi:10.1145/1143844.1143865

[3] Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[4]Caruana, R., Karampatziakis, N., & Yessenalina, A. (2008). An empirical evaluation of supervised learning in high dimensions. *Proceedings Of The 25Th International Conference On Machine Learning - ICML '08*. doi:10.1145/1390156.1390169

[5] R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, J. D. Knowles, Fifty Years of Pulsar Candidate Selection: From simple filters to a new principled real-time classification approach, Monthly Notices of the Royal Astronomical Society 459 (1), 1104-1123, DOI: 10.1093/mnras/stw656

[6] Johnson, B., Tateishi, R., Hoan, N., 2013. A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees. International Journal of Remote Sensing, 34 (20), 6969-6982.

Extras

HRTU2 Accuracy

	20/80	50/50	80/20	Avg.
SVM	.9797	.9800	.9779	.9792
LR	.9795	.9799	.9773	.9789
KNN	.9714	.9727	.9715	.9719
RF	.9784	.9796	.9802	.9794
DT	.9770	.9773	.9754	.9766

Mushroom Accuracy

	20/80	50/50	80/20	Avg.
SVM	1	1	1	1
LR	1	1	1	1
KNN	1	1	1	1
RF	1	1	1	1
DT	.9948	.9978	.9988	.9971

Wilt Accuracy

	20/80	50/50	80/20	Avg.
SVM	.9873	.9917	.9931	.9907
LR	.9847	.9908	.9908	.9889
KNN	.9893	.9954	.9965	.9937
RF	.9928	.9968	.9988	.9961
DT	.9870	.9926	.9977	.9924