An Empirical Comparison of Supervised Learning Algorithms: Revisited

Meyhaa Buvanesh
Department of Cognitive Science
University of California, San Diego

Abstract

For supervised learning classification problems, choosing the optimal classifier is not only a challenge, but also is critical for success. The work by Caruana and Niculescu-Mizil pushed for the systematic evaluation of the performance of classification algorithms. This work revisits the original work by Caruana and Niculescu-Mizil and evaluates the performance of 4 classifiers across 3 datasets from the UCI Machine Learning Repository: Logistic Regression, Support Vector Machine, Random Forest, K-Nearest Neighbor. The datasets explored include classification problems related to wine quality, salaries based on adult census data, and banking marketing.

Introduction

Supervised learning involves input and output variables and using an algorithm to learn the mapping between the input and output variables. Supervised learning problems can be grouped into regression and classification problems. The purpose of both regression and classification is to develop a model that can be used for mapping new examples by predicting and generalizing based on the set of training data [6]. The success of a classification problem depends on the classifier chosen and the corresponding parameters used.

Although there was development of machine learning algorithms since the 1990s, Caruana and Niculescu-Mizil in 2006 conducted a comprehensive and systematic evaluation of 10 classifiers to quantitatively assess and rank the performance of the different methods across 11 datasets under 8 metrics [1]. Although there was significant variability across the performance of the models across the datasets, some models, on average, were generally superior. Calibrated boosted trees and random forests were the best learning algorithms, overall, followed by uncalibrated bagged trees, calibrated Support Vector Machines (SVM), and un-calibrated neural nets. And naive bayes, logistic regression, decision trees, and boosted stumps had a generally lower performance [1].

Methodology

This paper revisits the theme of Caruana Niculescu-Mizil's work by using their methodology to evaluate a set of 3 classifiers across 3 datasets. The 3 classifiers being evaluated include random forests, SVM, and logistic regression. Parameters and ranges were modeled after the work of Caruana and Niculescu-Mizil; however, this was not always feasible due to implementational constraints. Each dataset includes 3 trials. In each of the trials, the data is shuffled and split into training and testing sets with different partitions. The first split saves 20% of the data for testing, the second saves 50%, and the third saves 80%.

Support Vector Machine: Wine Quality dataset implementation includes a grid search exploration.

Random Forest: 10-fold cross validation used.

KNN: Searched through 1-25 iterations to find optimal n.

Datasets

All three datasets were retrieved from the University of California, Irvine Machine Learning Repository [2]-[4]. The datasets included a dataset on wine quality, adult census income, and banking-marketing.

Wine Quality

The purpose of this dataset is to use the 11 numerical physiochemical properties of wine provided to model wine quality. The size of this dataset is 1599 x 12 features [2].

Adult Census Income

The purpose of this data is to classify whether a person has an income greater than $50,000 based on the 5 numerical and 9 categorical demographic and career information features provided. For the categorical variables, the data was transformed using the factorize method form the Pandas library to obtain a numerical representation. The size of this dataset is 31,561 x 12 features [3].

Banking Marketing

The purpose of this data is to classify whether a client will subscribe to a term deposit or not based on the 7 numerical and 9 categorical financial information features provided. For the categorical variables, the data was transformed using the LabelEncoder method from sklearn. And for the actual outcome of whether or not the client subscribed, the yes and no responses were mapped to 1 and 0, respectively. The size of this dataset is 45,211 x 17 features [4].

Results

Table 1 shows the average algorithm performance values for each classifier for each of the 3 dataset classification problems: wine quality, adult salary, and banking marketing.

Table 1: Algorithm Performances [2]-[4]

Classifier	Wine Quality	Adult Salary Census	Banking Marketing	Average
Logistic Regression	0.8735	0.8208	0.8899	0.8614
SVM	0.8688	0.8226	0.8826	0.8580
Random Forest	0.8964	0.8573	0.9032	0.8856
KNN	0.8736	0.8341	0.8876	0.8651

As expected, the random forest algorithm performed the best followed by KNN. And logistic regression, on average, performed better than SVM.

Discussion

This paper aims to revisit a subset of the thorough evaluation conducted by Caruana and Niculescu-Mizil. Compared to the original studies, this report only compares 4 classifiers on 3 datasets with only one main metric of accuracy. Although the general trend aligns with the consensus from previous work [1], some of the values are different amongst the datasets individually. Across all the datasets, the random forest classifier performed the best. For the wine quality dataset, the logistic regression and KNN classifiers only differed in performance by a margin of 0.0001. For the adult salary census dataset, the SVM classifier performed better than logistic regression. And for the banking marketing dataset, the logistic regression classifier performed better than the KNN implementation. This just highlights the theme of the From the No Free Lunch Theorem for search and optimization. There are no short cuts in finding the optimal algorithm because the optimal algorithm may not exist. When averaged over a variety of problems and metrics, each algorithm has its variability which is demonstrated through this data as well [5].

There were computational and time constraints which limited the depth of this report; mainly, artificial neural networks (ANN), boosting and bagging classifiers were left out of this report. Additionally, even amongst the classifiers used in this report, a Bayesian optimization could have been utilized for optimizing the hyper-parameters. The code attached with this report also shows exploration of different means of optimizing the hyper-parameters for the SVM implementation using a grid search.

Conclusion

Overall, the performance of 4 classifiers was evaluated across 3 datasets. Each classifier was tested in 3 independent trials with different data partitions for training and testing. The testing accuracies were then averaged across the classifiers for each partition. Overall, performance improved with the use of the random forest and KNN algorithms, similar to the work of Caruana and Niculescu-Mizil [1]. However, the accuracies of the SVM and logistic regression methods were not far behind. From the No Free Lunch Theorem, it is evident that each classifier is unique in its execution and performance [5]. Extensive future work needs to be conducted in order to better systematically evaluate the performance of these classifiers and understand how to rank such classifiers for general use across datasets.

Bonus

In addition to the 3 required classifiers, this report also evaluated a fourth KNN classifier.

References

[1] Caruna, R., and Niculescu-Mizil, A., 2006. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine Learning*,* 161-168. ACM.

[2] Cortez, P., Cerdeira, A., Matos, T. and Reis, J. (2019). *Wine Quality Data Set*. [online] UCI Machine Learning Repository. Available at: http://archive.ics.uci.edu/ml/datasets/Wine+Quality.

[3] Kohavi, R. and Becker, B. (2019). *Adult Data Set*. [online] UCI Machine Learning Repository. Available at: http://archive.ics.uci.edu/ml/datasets/Adult.

[4] Moro, S., Cortez, P. and Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62, pp.22-31.

[5] Sewell, M. (n.d.). *No Free Lunch for Supervised Machine Learning*. [online] No-free-lunch.org. Available at: http://www.no-free-lunch.org.

[6] Shetty, B. (2018). *Supervised Machine Learning: Classification*. [online] Towards Data Science. Available at: https://towardsdatascience.com/supervised-machine-learning-classification-5e685fe18a6d.