Linear SVM, Decision Tree, and Random Forest On Three Datasets

Justin Nguyen
University of California, San Diego
jhn074@ucsd.edu

Abstract

In accordance to the paper published by Caruana & Niculescu-Mizil on various supervised machine learning methods, it was concluded that some methods result in stronger performance in various accuracy metrics. In this paper, a subset of the methods tested in the Caruana & Niculescu-Mizil study, namely the linear support vector machine, the decision tree, and the random forest methods, are tested and compared in terms of their performance with regard to testing accuracy. Whereas Caruana & Niculescu-Mizil measure performance in more accuracy metrics, this study will be limited to testing accuracy, but also measure the testing accuracy of different training/testing partitions. Experiments on three different datasets seem to confirm the findings of the aforementioned paper, but also reveal the effects of three different training/testing partitions.

1. Introduction

Caruana & Niculescu-Mizil's extensive study of various supervised machine learning algorithms provides an excellent foundation from which to gauge the relative performances. Whereas their study had used a large list of performance metrics to determine, such as testing accuracy, f-score, lift, etc, this experiment seeks to study the relative performances of SVM, Decision Tree, and Random Forest solely on testing accuracy. From Caruana & Niculescu-Mizil's study, it is expected that Random Forest should result in the best accuracy with SVM following, and Decision Tree with the worst performance.

The goal of this study is to replicate the before mentioned ranking using similar classifications parameters as those used to train the SVM, DT, and RF used in the Caruana & Niculescu-Mizil study. However, due to limitations in both computation and time, as this experiment will be carried out on a laptop, some of these parameters may not be identical to those used in the Caruana study due to differences in datasets, normalization and scaling, subsetting, as well as implementation of each machine learning method. The datasets used to obtain the accuracies will also differ from those used in the Caruana & Niculescu-Mizil study. The Adult, Wine Quality, and Abalone datasets have all been obtained from the UCI Machine Learning Repository. Whereas the Adult dataset has been shared between the two studies, this study will use a smaller subset of 2000 instances due to a lack of computational resources.

This study also seeks to explore the correlations between training and testing dataset splits and the resulting testing accuracy. Intuition leads us to believe that having a larger training set will allow the algorithms to create better predictions on the testing set, and the empirical results obtained

from this experiment tend to reveal this trend, with small variances that can be accounted for variances in the properties of the datasets used, as well as the smaller sample sizes used in this experiment.

2. Method

This section elaborates on the methodology used in this experiment, including the machine learning methods used and their parameters, the datasets used and the procedure used to prepare the datasets for a binary classification task, as well as the means to which accuracy scores are obtained.

2.1 Machine Learning Algorithms

SVM: the linear SVM implementation from the sklearn package is used for the purposes of this experiment. The regularization parameter is varied by tens from 10^-3 to 10^7. This is different form the Caruana study due to the rescaling of values with MinMaxScaler, as well as a different subset.

Decision Tree: the decision tree implementation for this experiment is also from the sklearn package. The max depth is varied from 1 to 8 branches and performance is determined using the entropy criterion.

Random Forest: Again, the sklearn package of random forest is used for this experiment. The parameters varied in this experiment are the number of estimators, max depth, max features, and the minimum of samples per split.

2.2 Datasets

Adult: The adult dataset was a dataset with 48842 data points and 14 attributes but has been shortened to a subset of 2000 instances for the sake of a lack of computational resources. This is a dataset with both integers and categorical variables, so the categorical variables have been encoded using One Hot encoding from Python's pandas get_dummies method. The classification task for this dataset is to predict whether an individual makes above 50k a year. This column was initially a categorical variable but was reworked into a binary column with 1 representing a salary above 50k and 0 representing a salary below 50k.

Wine Quality: Not to be confused with the wine dataset, this dataset contains a total of 4898 instances and 12 attributes. However, this dataset is split into two datasets, one containing red wines and the other with white ones. The classification task for this dataset was initially multiclass, ranking wine quality on a scale of 1 to 10. In order to transform this dataset into a binary classification task, the red wine and white wine datasets were concatenated into a single wine quality dataset. The column containing the wine quality rating was split into a binary rating, with one class containing wines rated 6 or above, and the other containing wines rated 5 or below. It should be noted that after this procedure, the classes were extremely unbalanced, with the majority of the wines being in the 6 or above category. In order to save on computational resources, as well as to balance the classes, a subset of the dataset containing only 2000 instances was used instead.

Of the 2000 instances, exactly 1000 are in the 6 or above class, and the other 1000 are in the 5 or below class.

Abalone: The final dataset used contains 4177 instances and 8 attributes. The classification task for this dataset was to predict the number of rings an abalone has. Upon transforming the dataset into a binary classification task, it was found that splitting the abalone by those with 10 or more rings against those with less than 10 rings yielded an extremely balanced dataset. As such, I did not feel the need to use a subset of this dataset, and instead used the dataset in its entirety.

2.3 Hyperparameter Search and Accuracy

In order to find the appropriate hyperparameters to use in order to yield the best testing and validations accuracies, the GridSearchCV package from sklearn was used to obtain the best parameters through cross validation for the linear SVM and the Decision Tree. The grid search was visualized through use of heatmaps. The training, validation, and testing accuracies were then obtained using the best parameters, as well as the scoring function.

For the Random Forest, the RandomizedSearchCV was instead used for its increased efficiency and speed when paired with the random forest algorithm. Otherwise, the procedure is the same as that used in the grid search, but instead of heatmaps being used, the accuracies are reported with Randomized search's best scores function.

3. Experiment

Each dataset is run by each of the three algorithms, through three partitions, with three trials for each partition for a grand total of 243 scores. The resulting training, validation, and testing accuracies are then averaged per trial, consolidating the scores seen to 81 scores. The scores are all obtained using 5-fold cross validations and three trials.

3.1 Performance by Classifier

The mean test score is obtained simply by taking the average of a classifier's testing scores for all three of its partitions. As shown in table 1, each mean test score is denoted by up to three asterisks. Each asterisk represents its ranking relative to the other mean tests scores from the same dataset. For example, one asterisk for the random forests implementation on the adult dataset indicates that the score is rank one relative to the other scores, making it the best classifier for that dataset.

Following this notation, we can see that the mean testing scores obtained for each classifier follow a very clear pattern. The random forest classifier consistently outperforms the other classifiers, no matter which dataset it is performed on. The linear support vector machine follows second, consistently outperforming the decision tree in every dataset. This leaves the decision tree classifier as the consistent worst classifier in terms of testing accuracy out of the three. These results are reassuring, as they are consistent with those found in the Caruana & Niculescu-Mizil study, following the same exact order as their study.

Table 1. Obtained Average Scores Over Three Datasets

Classifier/ Partition	Average Training	Average Validation	Average Testing	Mean Test Score
SVM Adult 80/20	.85166	.827*	.8425*	
SVM Adult 50/50	.85833*	.826	.83299	
SVM Adult 20/80	.85583	.80267	.80313	.82621**
SVM Wine 80/20	.74104	.738	.71167	
SVM Wine 50/50	.732	.72767	.733*	
SVM Wine 20/80	.75583*	.74433*	.71896	.72121**
SVM Abalone 80/20	.79198*	.78933	.78349	
SVM Abalone 50/50	.79191	.78933	.78794*	
SVM Abalone 20/80	.79800	.79333*	.78646	.78596**
DT Adult 80/20	.83000	.817*	.84250*	
DT Adult 50/50	.84267	.814	.81067	
DT Adult 20/80	.85417*	.79433	.79917	.81745***
DT Wine 80/20	.80979	.72767*	.70833*	
DT Wine 50/50	.86267*	.723	.69933	
DT Wine 20/80	.80917	.71267	.69396	.70054***
DT Abalone 80/20	.80335	.78033*	.77951*	
DT Abalone 50/50	.80348	.76833	.77421	
DT Abalone 20/80	.84591*	.773	.75982	.77118***
RF Adult 80/20	.923125	.83729*	.85083*	
RF Adult 50/50	.93733*	.83231	.842	
RF Adult 20/80	.93417	.82427	.82500	.83928*
RF Wine 80/20	.9975	.80643*	.82250*	
RF Wine 50/50	1.0*	.78264	.78366	
RF Wine 20/80	.98833	.76608	.73791	.78136*
RF Abalone 80/20	.93425	.79408	.79306*	
RF Abalone 50/50	.96775*	.79338	.79097	
RF Abalone 20/80	.96766	.80360*	.78775	.79059*

3.2 Performance by Partitions

Within each classifier/dataset pair, there are three scores denoted by an asterisk. These scores represent the best within their column, within the classifier/dataset pair. For example, the score denoted by an asterisk within the SVM Adult blocks under the testing score represents the best testing score within the three partitions in the SVM experiment on the Adult dataset.

The results found between partitions do not lead to as nearly a clear-cut conclusion as the findings from the mean test scores on each classifier. For the decision tree and random forest classifier, the best validation and testing scores are almost always obtained by using a partition of 80% training and 20% testing. However, when examining the linear SVM, it is found that only the Adult dataset yielded the best accuracy under a 80/20% partitions, whereas the Wine Quality and Abalone

dataset actually faired marginally better under the 50/50% partition over the 80/20 and 20/80 partitions.

I speculate that these findings are not due to partitions themselves, but rather the interactions between the linear SVM classifier and the Wine Quality and Abalone datasets. This discrepancy between testing scores and partitions could be explained by the linear SVM's inability to correctly classify instances of the Wine Quality and Abalone datasets, as these datasets may simply have a poor fit for a linear classification. I reason this is the case due to the relationship between training scores and testing scores for these two sets under the linear SVM. Particularly in the Abalone dataset, the values between training and testing scores are very close to each other, and hover around .78-.79 accuracy range. This is also likewise in the Wine Quality set, but it sits around a lower accuracy of .74-.75. This suggests that the linear SVM is just a poor fit for the data in general.

Conclusion

The findings in this study suggest that the rankings obtained in the Caruana & Niculescu-Mizil for different learning classifiers is appropriate for the datasets chosen for this experiment. Of the linear SVM, decision tree, and random forest, it is found that the random forest classifier performs consistently better than the other two, where as the decision tree is a rather poor classifier when uncalibrated, leaving the SVM to be the middle of the two.

With respect to the effect of partitioning on the testing accuracy, it was found that having a larger percentage of the dataset allocated to training the classifiers, i.e. an 80/20% split, generally led to greater test scores for most classifiers. The only classifier that did not consistently generate the result is the linear SVM. However, this could possibly be explained by the low training and validation scores granted by the linear SVM on the Wine and Abalone datasets, suggesting that the lower test scores are a symptom of a poor linear fit on the datasets in general. It can be concluded then that having a higher percentage weighted towards training leads to better testing scores in general, but only if the classifier performed on the data is that of an appropriate fit.

References

Caruana, R., & Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning* (pp. 161-168). ACM.

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Becker, B and Kohavi, R. (1996). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets/Adult]. Silicon Graphics, Data Mining and Visualization

Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn and Wes B Ford (1994) "The Population Biology of Abalone (_Haliotis_ species) in Tasmania. I. Blacklip Abalone (_H. rubra_) from the North Coast and Islands of Bass Strait",
Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288)

See Jupyter Notebook for heatmaps