Comparing Classifiers: Insights on Test Sizes & Accuracy

Keshav Tiwari

Abstract

The rapid increase of Machine Learning applications in recent years has brought with it a growing set of supervised learning tools, algorithms and models aimed towards extracting the most accurate predictions from the given labeled data. However, developers are now faced with the challenge of comparing these models to evaluate the best one. This distracts from the actual goals of the developer and forces them to incur computational, financial, and time costs. Thus, this paper serves as an empirical evaluation of 5 prominent classifiers - Random Forest, SVMs, Logistic Regression, K-Nearest Neighbor, and Decision Trees - on 4 binary classification tasks. Moreover, it discusses the relationship between test set sizes and model accuracy scores (validation, training, and testing). Using Platt Scaling and Standard Scaling along with cross-validating through GridSearchCV, the paper aims to provide a holistic and considered comparison of these models' accuracy scores.

1. Introduction

1.1 Background

A few papers in the past have assessed and compared state-of-the-art algorithms for learning. Two of the most prominent studies include STATLOG (King et al., 1995) and "An Empirical Comparison of Supervised Learning Algorithms" (Caruana & Niculescu-Mizil, 2006). Both studies were comprehensive for their time, in fact, (Caruana & Niculescu-Mizil, 2006) built on top of the empirical evaluation conducted in STATLOG, given the emergence of newer learning algorithms such as SVMs and Random Forests after 1995. This paper aims to recreate the results presented in (Caruana & Niculescu-Mizil, 2006), albeit at a much smaller scale, to validate its findings and contextualize them in 2024.

1.2 Approach

While that paper reviewed 10 classifiers averaged over 8 performance metrics across 11 binary classification tasks, this paper will focus on 5 models that were spread across the performance rankings stated in the paper, thus providing a representative sample of classifiers. Along with the intentional choice of classifiers, this paper further chose datasets (explained further in section 2.1) which, by default or by stable manipulation, posed binary classification problems to further match (Caruana & Niculescu-Mizil, 2006). To those ends, all algorithms mentioned in this paper also underwent Platt's Scaling, so that algorithms that are not designed to predict probabilities can still be fairly compared to the others. Specifically, SVMs, KNN, and Logistic Regression sets also underwent standard scaling, given their sensitivity to feature scaling. Limiting the number of datasets to 4 and performance metrics to just 1, i.e., accuracy, allows us to focus on the goal of merely validating the trends proposed in 2006, minimizing superfluous time and computational costs. In the interest of those computational costs, this paper also seeks to notice the relationship between increasing training and decreasing testing data with a model's accuracy score.

2. Method

2.1 Datasets & Preprocessing

All 4 datasets used to train, and test were obtained from the UC Irvine Machine Learning Repository, specifically used for classification tasks with clear target columns. The selection of these datasets was made primarily motivated by a variety of

domains, features and a computationally feasible number of instances. Two of them were randomly sampled to obtain a representative subset of instances that minimized computational costs while being significant enough to recreate (Caruana & Niculescu-Mizil, 2006)'s results. Note that most target columns for the datasets were originally multi-label and were reasonably mapped into binary labels. The following figure presents key characteristics of the datasets used.

Table 1.0: Description of Problems

Datasets	# Instances	# Features	Feature Types
OBESITY	2,111	16	Integer
STUDENT	4,424	36	Real, Categorical, Integer
CREDIT	30,000*	23	Integer, Real
BANK	45,211*	16	Categorical, Integer

**instances were limited to 4,424 by random sampling due to unfeasible computational and time costs*

Table 1.0 summarizes the key characteristics of the datasets used to train each classifier. OBESITY refers to a dataset that estimated the obesity levels of people in Mexico, Peru and Columbia, based on their eating habits and physical condition. The target variable for this dataset was originally multi-label with 7 different stages of body weight (3 stages of obesity, 2 stages of being overweight, normal and underweight), however, they were binarily mapped to be obese (1) or not obese (0). STUDENT refers to a dataset which predicts student dropout and academic success based on personal information. Like the OBESITY dataset, it had multi-labels (dropout, enrolled, and graduate) which were mapped to be dropout (1) or not dropout (0). CREDIT refers to a dataset that is used to predict the probability of a customer defaulting on their payments next month, based on personal information and payment history. BANK refers to a dataset that is used to predict if customers are going to subscribe to a term deposit as a result of a marketing campaign. Categorical 'yes' or 'no' values were mapped to 1 and 0 values respectively. Furthermore, given the large set of instances in CREDIT and BANK datasets, they were randomly sampled in order to minimize computational and time costs, since they had to be trained over 5 classifiers, 3 partitions and 3 trials, along with 2 other datasets. Lastly, the categorical features within these datasets were transformed into numerical variables using the OneHotEncoder, in order to be trained and tested later.

2.2 Classifiers, Calibration & Cross-Validation

The classifiers used in this paper were chosen to represent the trend found in (Caruana & Niculescu-Mizil, 2006). For each of them, the aim was to explore the range of parameters and standard variations as holistically as computationally possible. All classifiers underwent Platt Scaling, to create an even platform for them to be compared to each other, since algorithms like SVMs are not designed to predict probabilities. Moreover, algorithms that are sensitive to feature scaling underwent Standard Scaling, including Logistic Regression, SVMs, and KNN. The following section details the sets of hyperparameters for each classifier which were cross validated to find the best possible model.

SVMs: a list of regularization parameters was used (C) = [0.1, 1, 10] which varied by factors of ten; linear and the radial kernel basis function kernels were used; underwent Platt Scaling and Standard Scaling.

K-Nearest Neighbors (KNN): different numbers of neighbors were used [3, 5, 7]; underwent Platt Scaling and Standard Scaling.

Random Forest (RF): number of estimators were varied from the list [100, 200, 500]; along with the max depth [None, 10, 20]; underwent Platt Scaling only.

Logistic Regression (LR): regularization parameters were varied by factors of 10 [0.1, 1, 10]; underwent Platt Scaling and Standard Scaling.

Decision Trees (DT): max depth was varied from [None, 5, 10]; and minimum sample splits were varied from [2, 5, 10]; underwent Platt Scaling only.

A parameter grid was initialized with all the variations for each learning algorithm. Then GridSearchCV was used to cross-validate across all hyper parameter combinations. Thus, the optimal hyperparameters to be used, i.e., the best possible classifier model was used, to further enable a valid comparison amongst all learning algorithms.

2.3 Evaluation Strategy

The datasets were all split into training and testing sets in 3 different proportions: 80/20, 50/50, and 20/80. This was necessary to confirm trends regarding test sizes and model accuracy. The simplest and most common performance metric that could have been used was the accuracy score. Thus, for each trial, partition, test size, and classifier, accuracy scores were calculated for both training and testing sets. Moreover, in order to get validation accuracy, the internal validation of hyperparameters was used by calling 'grid.best_score_'. Thus, testing, training and validation accuracies were found for 180 trials in total (4 datasets x 5 classifiers x 3 test sizes x 3 trials). Specifically, the mean testing accuracies were calculated across 3 trials, and then over 4 datasets with 3 test sizes each (12 columns) to rank order the 5 classifiers.

3. Experiment

This section will discuss the results of the experiment conducted and will aim to find key findings to answer the key questions posed at the beginning of this paper regarding relative performances of binary classifiers and the trends of accuracy with increasing training and decreasing testing datasets. Please note that while testing, training and validation accuracies were found for all 180 trials, testing accuracy is the chosen metric of evaluation given its direct relationship to the model's effectiveness as a classifier for unforeseen data.

3.1 Classifier Performance across Datasets

As each classifier was run on each dataset, the results were stored and averaged over all 3 trials per test size split. The table below displays how average testing accuracies varied across classifiers, datasets and test sizes. The final column is the mean testing accuracy across all columns, and is the column used to provide a final rank order to the classifiers.

Table 1.1: The average testing accuracy of each classifier across datasets (classification tasks) and splits (in test size)

Clf	OBESITY			STUDENT			CREDIT			BANK			Mean
	0.2	**0.5**	**0.8**	**0.2**	**0.5**	**0.8**	**0.2**	**0.5**	**0.8**	**0.2**	**0.5**	**0.8**	
RF	.0779	0.787	0.786	0.865	0.867	**0.866**	0.811	**0.813**	**0.810**	0.887	0.884	0.884	**0.837**
SVM	0.779	0.787	**0.789**	0.863	0.863	0.862	0.805	0.805	0.801	**0.890**	**0.887**	**0.885**	0.835
LR	**0.803**	0.788	0.765	**0.868**	**0.869**	**0.866**	0.793	0.796	0.791	0.884	0.882	**0.885**	0.833
DT	0.778	**0.789**	0.780	0.852	0.861	0.845	**0.813**	0.811	0.800	0.885	0.879	0.882	0.831
KNN	0.770	0.781	0.782	0.834	0.832	0.828	0.783	0.792	0.790	0.879	0.878	0.878	0.819

As per Table 1.1, Random Forest proved to be the most effective learning algorithm of across all binary classification tasks. This is in alignment with the findings of (Caruana & Niculescu-Mizil, 2006), wherein it ranked topmost of all the classifiers the two papers have in common. SVMs ranked second, further in alignment with the target paper which ranked it second amongst all common classifiers too. However, the rankings of classifiers cease to be common with (Caruana & Niculescu-Mizil, 2006) beyond RF and SVMs. As per which, LR should be the least effective learning algorithm out of the remaining three. In fact, the order in the paper is KNN > DT > LR, which is the exact opposite of that found in Table 1.1. There can be several reasons for this, considering the variation in dataset sizes, a difference in hyperparameter tuning and arrangements, varying effects of calibration techniques etc. There are extremely fine margins amongst the top 4 models which suggests that there was realistic room for DT to rank above LR in this paper.

Interestingly, LR had the most highest-ranking classification splits, since it ranks highest in 5 of the 12 columns above. However, it performs extremely poorly in the tasks it is not the highest ranking in, causing its fall to third position. This aligns with the paper which suggests that across datasets, there was severe variation in which classifier would be the best performing. It should also be noted that KNN performed considerably worse than the other 5 models, given its gap of 0.11 in mean testing accuracy to DT, and it not ranking as the best model in any of the classification tasks.

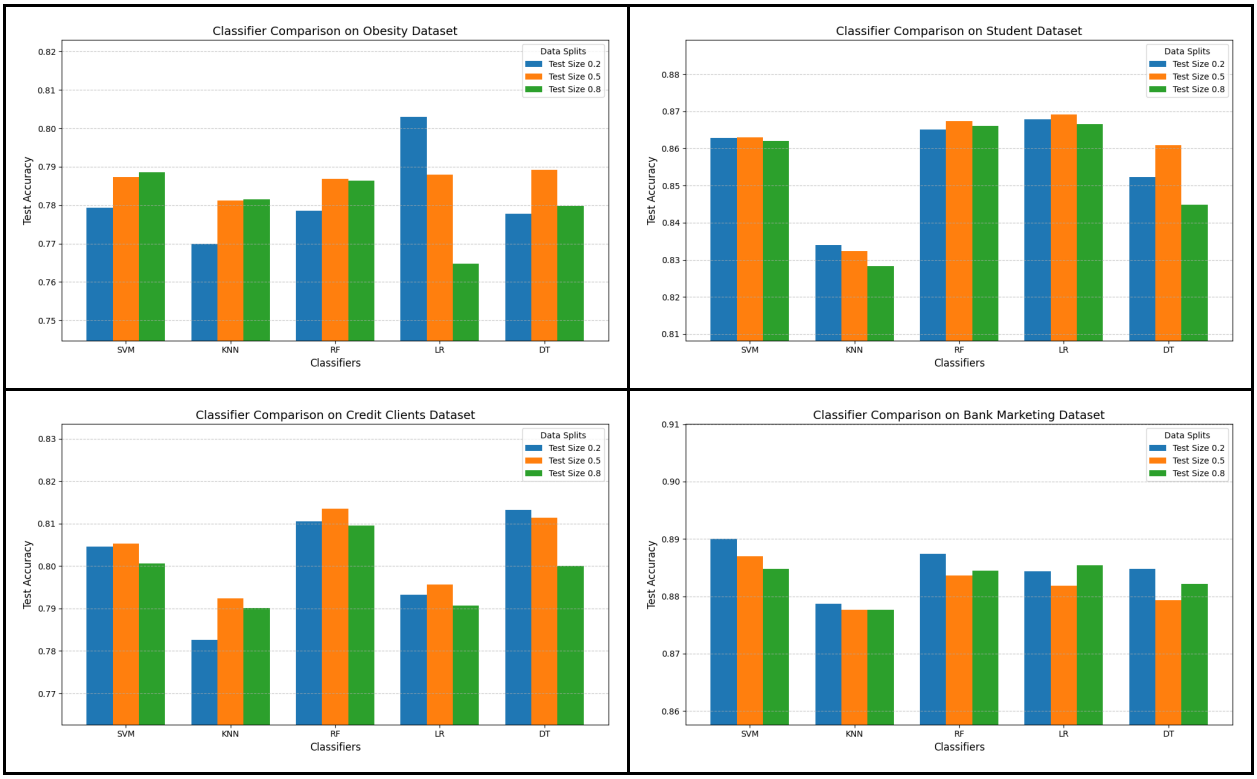
3.2 Classifier Performance across Test Size (Training Set/Testing Set Splits)

The next significant research question for this paper was to explore the relationship between training set sizes and model accuracies. Testing accuracy remains the performance metric for this section of the results as well, as we see how accuracies varied for each of the classifier's datasets across test split sizes.

Figure 1.0 shows the trends obtained when testing accuracy were plotted for each dataset and classifier, and each of their splits. As is evident from the graphs, there was severe variation in which split performed best across the datasets. For most cases, the lower test size (0.2) performed with greater accuracy than the larger test size (0.8), albeit narrowly. This trend is in perfect agreement with the hypothesis that with an increasing training set, i.e., falling test size, the model accuracy will increase. However, in the case of the OBESITY dataset, other than LR, no other classifier seemed to have followed this trend. This was surprising to me, along with the 50/50 split being the best performing split in a significant number of trials.

This can be attributed to a lack of instances, and perhaps could have been resolved with a greater degree of feature selection and scaling. The variations for almost all of these testing accuracies is minimal so making any concrete statements about the trend is not entirely possible. Smaller test sets can lead to high variance within instances and their consequent accuracy measurements, leading to more unreliable results.

Figure 1.0: The trends between testing accuracy and train/test splits across classifiers and datasets



Conclusion

The aim of this paper was to validate the findings of the (Caruana & Niculescu-Mizil, 2006) paper and confirm trends between test sizes and classifier performances. After observing the results of this paper, while both papers do not agree with each other entirely, significant trends and results are confirmed by both. Even though the classifier rankings were not the same, both papers agreed on ranking RF and calibrated SVMs as the two highest ranking classifiers, while also citing the need for calibration, Pratt Scaling and cross validation in both papers to attain the best results. They both also confirmed the varying nature of performance with varying datasets amongst the classifiers.

As far as test split vs testing accuracy goes, the trends were clear in that in most trials higher test set sizes performed worse than lower test set sizes. While most cases proved the paper's initial hypothesis, a significant number of cases had the 50/50 split ranked as the best performing split, which can be attributed to the unreliability of small test sizes, lack of instances, smaller training data, poor feature selection etc. Notably, the difference amongst the testing accuracies was close to minimal and therefore, significant claims regarding trends were not possibly declared.

Some significant traits that limited the scope and the effectiveness of this paper relative to (Caruana & Niculescu-Mizil, 2006) were unfeasible computational costs, the over-reliance on accuracy scores as a performance measure, and the remapping of multi-label target variables to binary targets. Solving these fundamental issues, along with potentially experimenting with different scaling/calibrating techniques or even different learning algorithms may prove to make such a validation study more effective in the future. Crucially, calibration and scaling techniques made a significant contribution to the cross-validation processes and made the models perform a lot better, underlining their importance in supervised learning methods.

References

- Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, 161–168. https://doi.org/10.1145/1143844.1143865
- Estimation of Obesity Levels Based On Eating Habits and Physical Condition [Dataset]. (2019). UCI Machine Learning Repository. https://doi.org/10.24432/C5H31Z.
- Realinho, V., Vieira Martins, M., Machado, J., & Baptista, L. (2021). Predict Students' Dropout and Academic Success [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5MC89.
- Moro, S., Rita, P., & Cortez, P. (2014). Bank Marketing [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5K306.
- Yeh, I. (2009). Default of Credit Card Clients [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C55S3H.