A Comparative Analysis of Supervised Learning Models

Paul J. Losiewicz
University of California, San Diego
Cognitive Science Department
plosiewi@ucsd.edu

Abstract

This report aims to extend directly off of the methodology performed by Caruana and Niculescu-Mizil in 2006 and compare the performance of three popular yet incredibly different supervised machine learning models: the support-vector machine (SVM), random forest classifier, and logistic regressor. This project is exploratory in nature and smaller in scope when compared with the original. The three models were each evaluated across three datasets, each of which with significantly different structural features. The experiment produced similar results, as within datasets, the three models would perform relatively close to one another. The major conclusions made from this experiment revolve around the significant metric differences from dataset to dataset, in which the results can be explained by the drastic difference in feature set sizes.

1 Introduction

As an undergraduate mathematics student and aspiring researcher, it is imperative to be able to assess and determine the appropriate classifier for various contexts and domains. This report is an extension of Caruana and Niculescu-Mizil's comparative paper on supervised learning algorithms and an attempt to reproduce results based on simplified, yet similar, methodology [2].

While accuracy is a popular and intuitive metric choice for evaluating the efficacy of supervised machine learning models, it lacks versatility, especially when employed in imbalanced binary classification problems [7]. Since machine learning has developed rapidly as a field in the past two decades, several metrics have emerged, each of which are tailored to their own specific training data environments. In Caruana and Niculescu-Mizil's original study, eight metrics and eleven datasets were used, and in each pairing, results were seen to fluctuate [2]. It could be deduced that certain metrics and algorithms are tailored for very specific contexts, while other algorithms, like Random Forests, are fairly versatile and robust.

Since this report has a much smaller scope than the original, I selected the F1 Score as the sole metric for this comparison, as it is better equipped for unbalanced contexts by accounting for both accuracy (the percentage of correct predictions) and recall (the percentage of positives correctly identified) [6]. By no means is the F1 Score universally applicable to all binary classification problems; however, in this instance, I feel that it provides the most objective comparison and easily interpretable results between the three classifiers I have studied.

Out of the eight original classifiers tested in the paper that inspired this report, I have chosen the support vector machine (SVM), logistic regressor, and random forest classifier to compare across three different datasets. SVMs are known for their versatility and ability to identify and manage non-linear decision boundaries, making them well-suited for high-dimensional, real-world contexts [5]. Logistic regression is typically viewed as a baseline model in classification tasks that performs efficiently in terms of computing power, but tends to struggle when applied to noisy datasets with

outliers [3]. Drawing from the results of Caruana and Niculescu-Mizil, random forests tend to perform well across both metrics and datasets, but like all classifiers, they can be susceptible to inconsistent results, particularly due to overfitting [2]. Using one metric and three datasets, I aim to reproduce results from the former paper with a similar methodology yet smaller scope, or even perhaps arrive at results that are inconsistent. I want to remark that, regardless of the results I arrive at, this study is simplified and exploratory, and by no means is this an attempt to contradict the conclusions drawn by Caruana, Niculescu-Mizil, or the broader machine learning academic community.

2 Methodology

2.1 Hyperparameters and Algorithms

This section outlines the hyperparameters tuned for each algorithm to achieve optimum results. For each iteration of an algorithm, a GridSearchCV object from the scikit-learn library was employed. As the results show, the respective hyperparameters may vary from trial to trial, even within the same partition, likely due to randomized train-test splits.

2.1.1 SVM

For the SVM, a radial basis function (RBF) kernel was used for each trial. The RBF is known as a universally applicable kernel that is often employed in nonlinear contexts [4]. Degrees 2 and 3 were included in the grid search but were later determined to be arbitrary, as the polynomial kernel was removed from the search during preliminary testing. Gamma values of 0.001, 0.1, 1, and 2 were included in the search. Gamma controls the slope of the kernel function, with lower values creating smoother decision boundaries and higher values resulting in sharper, more localized boundaries [8]. Finally, five values of the regularization parameter C were used, ranging from 0.01 and increasing by factors of 10 up to 100.

2.1.2 Random Forest Classifier

For the Random Forest Classifier, only one hyperparameter was optimized in the grid search. Values of 2, 4, 8, and 16 were tested for the max_features parameter, representing the size of the feature set that is considered when splitting a node in each decision tree within the forest.

2.1.3 Logistic Regression

The only hyperparameter optimized for this classifier was the regularization constant C, which essentially controls how strongly the model fits to the training data. Values of $10^{-6}, 10^{-4}, 10^{-2}, 1,$ and 10 were tested.

2.2 Performance Metrics

As mentioned in the introduction, F-1 score was chosen as the sole performance metric due to its intuitiveness and ability to assess unbalanced datasets, as opposed to accuracy.

2.3 Datasets

Three datasets were used to test the supervised learning algorithms, all of which were sourced from the UC Irvine Machine Learning Repository [1].

2.3.1 Cannabis Use Dataset

The first dataset details drug consumption data and originally came with several potential target variables. Cannabis consumption was selected as the target variable because it could be transformed into a relatively balanced binary classification problem. The data was processed to assign -1 to individuals who had not consumed cannabis in the past year and $+1$ to those who had. This split resulted in 999 values labeled as $+1$ and 886 labeled as -1. The dataset contained 12 continuous features used for training.

2.3.2 Obesity Dataset

The second dataset focused on obesity data. Like the cannabis dataset, it required manipulation to create a binary classification problem. Patients categorized as "obese" to any degree were assigned $+1$, while others were labeled as -1. The dataset included five categorical features, which were one-hot encoded, resulting in a total of 1,283 features. Out of the 2,111 samples, 1,139 were labeled as -1.

2.3.3 Income Dataset

The third dataset concerned income levels, specifically whether individuals earned more or less than $50,000 annually. The original dataset contained over 40,000 samples, but for computational efficiency during grid search, 5,000 samples were randomly selected. After encoding seven categorical features, the final feature matrix had 101 features.

2.4 Partitions and Trials

For each dataset and algorithm pair, three train-test splits were evaluated: 20/80, 50/50, and 80/20 (train%/test%). Three trials were conducted for each split, resulting in a total of 27 F1 scores per model-dataset pairing. To simplify analysis, the average F1 score across trials is reported in the tables. Additionally, hyperparameters selected during grid search were recorded for each trial, resulting in nine sets of hyperparameters per model-dataset pairing.

3 Results by Dataset

3.1 Cannabis Use Dataset

3.1.1 SVM

For the cannabis dataset, the SVM consistently achieved high F1 scores across all partitions and metrics, with all scores exceeding 0.80. The highest testing F1 score of 0.832 was recorded in the 50/50 partition, using hyperparameters $\gamma = 0.1$ and $C = 0.1$. Interestingly, the highest cross-validation score (0.837) and testing score (0.860) were achieved in the 20/80 partition. Despite variations in hyperparameters, the F1 scores remained relatively stable across partitions.

3.1.2 Random Forest

The Random Forest classifier performed similarly to the SVM, with all F1 scores above 0.80. Its best testing F1 score of 0.830 occurred in the 20/80 partition. Metrics across partitions were nearly identical, indicating robust performance. The hyperparameter `max_features` varied significantly across trials and partitions, ranging from 2 to 16, but this did not lead to noticeable inconsistencies in F1 scores.

3.1.3 Logistic Regression

Logistic Regression achieved its highest testing F1 score of 0.826 in the 50/50 partition. However, differences in scores across partitions were minimal. Cross-validation scores consistently hovered around 0.820, with a range smaller than 0.01. The regularization constant C varied across trials, but this variability had little to no impact on performance metrics.

3.2 Obesity Dataset

3.2.1 SVM

For the obesity dataset, significant differences were observed between training and testing F1 scores. Training scores ranged from 0.870 to 0.936, with smaller training sizes yielding higher scores. Testing F1 scores ranged from 0.748 to 0.757, while cross-validation scores ranged from 0.760 to 0.780. The highest testing and cross-validation scores were observed in the 20/80 partition.

3.2.2 Random Forest

Similar trends were observed with the Random Forest classifier. Training F1 scores ranged from 0.927 to 0.942, while testing scores ranged from 0.731 to 0.755. Cross-validation scores ranged from 0.735 to 0.759. Despite these variations, the Random Forest performed consistently across partitions. The `max_features` parameter continued to fluctuate across and within partitions, ranging from 2 to 16.

3.2.3 Logistic Regression

Logistic Regression exhibited a noticeable drop between training and testing F1 scores. Training scores ranged from 0.871 to 0.909, while testing scores ranged from 0.728 to 0.751. The highest testing score was achieved in the 80/20 partition. Cross-validation scores ranged from 0.710 to 0.740, with the highest score in the 80/20 partition.

3.3 Income Dataset

3.3.1 SVM

The SVM performed less effectively on the income dataset compared to the cannabis and obesity datasets. Training F1 scores ranged from 0.634 to 0.746, and testing scores were lower, ranging from 0.550 to 0.573. Cross-validation scores ranged from 0.575 to 0.582, with the best results achieved in the 80/20 partition.

3.3.2 Random Forest

The Random Forest classifier displayed a significant gap between training and testing F1 scores. Training scores ranged from 0.770 to 0.830, while testing scores ranged from 0.578 to 0.629. Cross-validation scores were slightly higher than testing scores, ranging from 0.576 to 0.593. The best performance was observed in the 80/20 partition.

3.3.3 Logistic Regression

Logistic Regression showed consistent performance across training, testing, and cross-validation metrics. Training scores ranged from 0.621 to 0.651, testing scores from 0.595 to 0.611, and cross-validation scores from 0.596 to 0.606. Variability across partitions was minimal.

4 Conclusion

Based on the experimental results of this report, the greatest differences in supervised learning models can be observed from dataset to dataset.

In the cannabis dataset, where the least amount of features were used, all three models performed exceptionally well with testing and cross-validation scores consistently above 0.800 and negligible differences both across partitions and models. It's worth pointing out that the hyperparameters would vary within the same model/partition pairing for all three models, outlining the robustness and versatility of the models of choice. Additionally, the cannabis dataset was the smallest dataset in terms of entries, but it was significantly smaller in terms of features. Some interesting results from this dataset were the incredibly high training scores for the Random Forest Regressor, which did not result in overfitting when looking at the data collected from testing and cross-validation.

Our models started to perform significantly worse in the obesity dataset, where we saw our highest number of features (1283) across the three datasets. Once again, there were fairly negligible differences in test-scoring from model to model, but to point out any difference, the SVM had slightly smaller variability compared to the Random Forest and Logistic Regressor. The same trends can be seen in terms of hyperparameters as they varied greatly from trial to trial. Although there is a slight dropoff in F-scores from our trials in this dataset compared to the last, it's certainly notable that our models performed consistently well across partitions despite the significantly greater magnitude in terms of features set.

Our models performed significantly worse in the last dataset, which featured a large number of features, similarly to the obesity dataset, but there were also more than double the entries. While the SVM seemed to perform more reliably in the cannabis trials, we can see that the Random Forest and Logistic Regressor perform slightly better across the board, with the logistic regression even featuring less variability in terms of testing range across partitions. In addition, we can see a significant difference in the testing and training F-1 scores for the Random Forest here, which is an indicator of overfitting. While we observed this same trend in the cannabis dataset, the random forest still scored exceptionally high marks in terms of testing metrics, so that insight is less significant to point out.

Despite visual differences between the dataset, my main takeaway from this experiment is the negligible differences between supervised training models within datasets and the significant differences across datasets. Consistent with the work of Caruana and Niculescu-Mizil, the random forest regressor is robust but prone to overfitting. The logistic regressor is an incredibly reliable classifier across most contexts, especially when considering its baseline nature and lack of computational demand. All three models displayed versatility as they adjusted their optimal hyperparameters to the context of the trial, as we saw the hyperparameters changed within the same partition and dataset from trial to trial. The similarity in testing results across the three models certainly leads to future questions about the versatility of other models and how they would adjust to the data tested in this report.

References

[1] C. Blake and C. Merz. Uci repository of machine learning databases. *University of California, Irvine*, 1998.

[2] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 161–168. ACM, 2006.

[3] Jiashi Feng, Huan Xu, Shie Mannor, and Shuicheng Yan. Robust logistic regression and classification. In *Advances in Neural Information Processing Systems*, volume 27, pages 253–261, 2014.

[4] S. Han, C. Qubo, and H. Meng. Parameter selection in svm with rbf kernel function. In *World Automation Congress*, pages 1–4, 2012.

[5] Ankan Kar, Nirjhar Nath, Utpalraj Kemprai, and Aman. Performance analysis of support vector machine (svm) on challenging datasets for forest fire detection. Chennai Mathematical Institute, 2020.

[6] Yang Lu, Yiu-ming Cheung, and Yuan Yan Tang. Bayes imbalance impact index: A measure of class imbalanced dataset for classification problem. *arXiv preprint arXiv:1901.10173*, 2019.

[7] A. Menon, H. Narasimhan, S. Agarwal, and S. Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pages 603–611. PMLR, 2013.

[8] Intisar Shaheed, Jwan Alwan, and Dhafar Abd. The effect of gamma value on support vector machine performance with different kernels. *International Journal of Electrical and Computer Engineering (IJECE)*, 10:5497–5506, 2020.