Comparing the Performance of Various Supervised Learning Algorithms

Stephen Li
University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093
sxl002@ucsd.edu

Abstract

In this project, the primary goal was to replicate the results in Rich Caruana's and Alexandru Niculescu-Mizil's paper on the comparison of supervised learning algorithms. For this project, the SVM, KNN, and Random Forests classification were chosen and used to analyze three datasets from the UCI database repository. Further experiments involved varying the size of the dataset analyzed as well as changing the ratio of the training and testing split.

1. Introduction

Years after Rich Caruana's and Alexandru Niculescu-Mizil's original comparisons of various supervised learning algorithms, it becomes a challenging and interesting task to accomplish. The methods used in this project, Support Vector Machines, K-Nearest Neighbors, and Random Forests were chosen because of their ease of use as well as the fact that they provide a well-balanced spread of methods to be tested, each with their own pros and cons. Although the paper lacks many critical details on how to follow the exact procedure, sufficient knowledge can be leveraged to get a good idea on how to conduct the study. In addition, a bit of this project is dedicated to expanding on the goals of the study but in a different direction.

1.1. Data Description

Three different datasets were analyzed, and they are the same datasets used in the reference paper. Adult_Dataset, Cover_Type_Dataset, and Letter_Dataset are all downloaded from the UCI dataset repository.

For the Adult_Dataset, every categorical variable used was converted to dummy variables. The dataset contains 32,561 data points, with 14 variables (6 continuous, 8 categorical) that increase to 107 when using dummy variables.

The Cover_Type_Dataset was converted to a binary classification problem by treating the most commonly occurring cover type as positive and the rest as negative. The data contains 581,012 data points total, but only about 30,000 are used for analysis. There are 54 total attributes used for analysis.

For the Letter_Dataset, analyzing the data was divided into two parts. In Letter_Dataset - p1 the letter 'O' was treated as positive and the rest of the letters were treated as negative. In Letter_Dataset – p2 letters 'A-M' were treated as positive and the rest as negative, for a more balanced analysis. The data contains 20,000 points total, with 16 attributes used for analysis.

1.2. Problem Description

In this experiment, Rich Caruana's and Alexandru Niculescu-Mizil's paper is used as a guideline. Like in that paper, the main goal of this experiment is to compare the effectiveness of various supervised learning algorithms at a smaller scale and with less precision, while also adding some new comparisons.

2. Methods

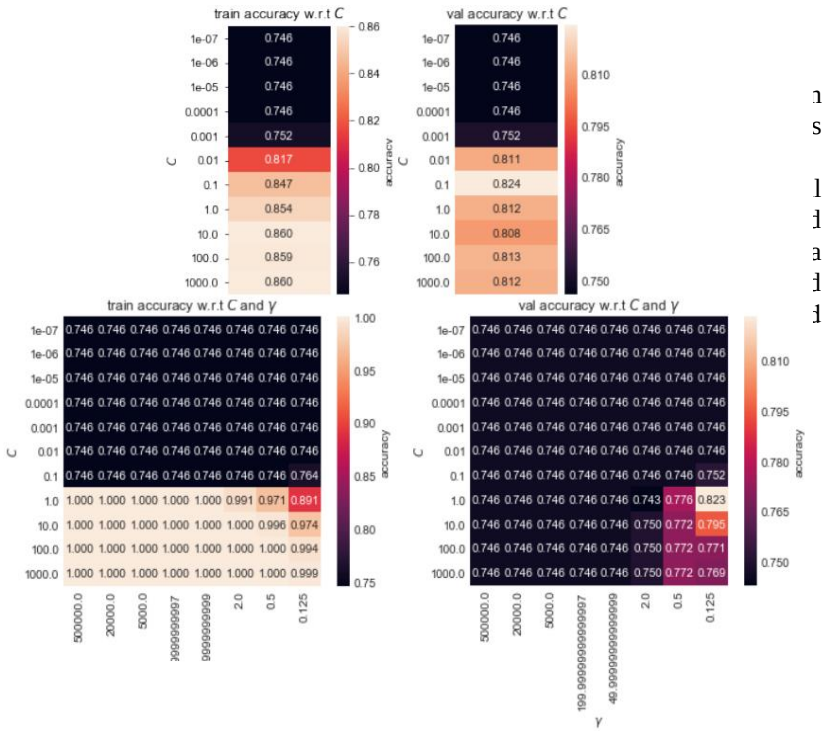
All classification methods were coded in python using Jupyter Notebook.

Both linear and rbf kernels were used for SVM classification. Both kernels used five cross validations and varied the regularization parameter from 10^{-7} to 10^{3}. However, the rbf kernel also used radial widths of length {0.001,0.005,0.01,0.05,0.1,0.5,1,2}, which were converted to gamma values. Because SVM analysis takes an extraordinary long to complete, the sklearn function StandardScaler had to be used to reduce the amount of time to process the data.

KNN used 26 values for K ranging from K = 1 to K = 26.

Random Forests implemented the Breiman-Cutler method, using 1024 trees.

These methods offer a well-balanced distribution of performances. Random Forests is considered a higher performing method, SVM is reasonably effective method, and KNN is a simple to use but the lowest performing method. Additionally, they are simple to understand and explain the average user.



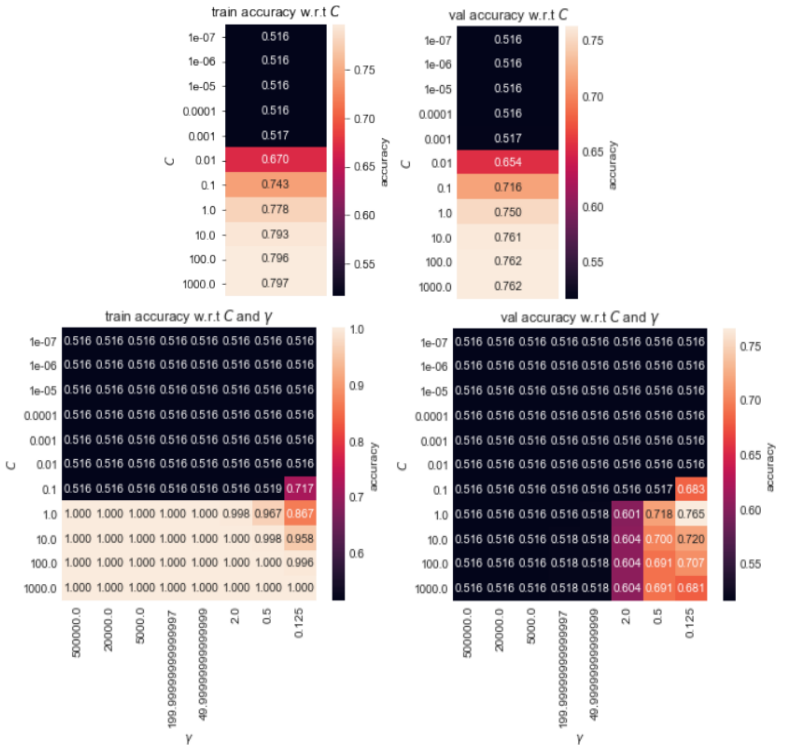
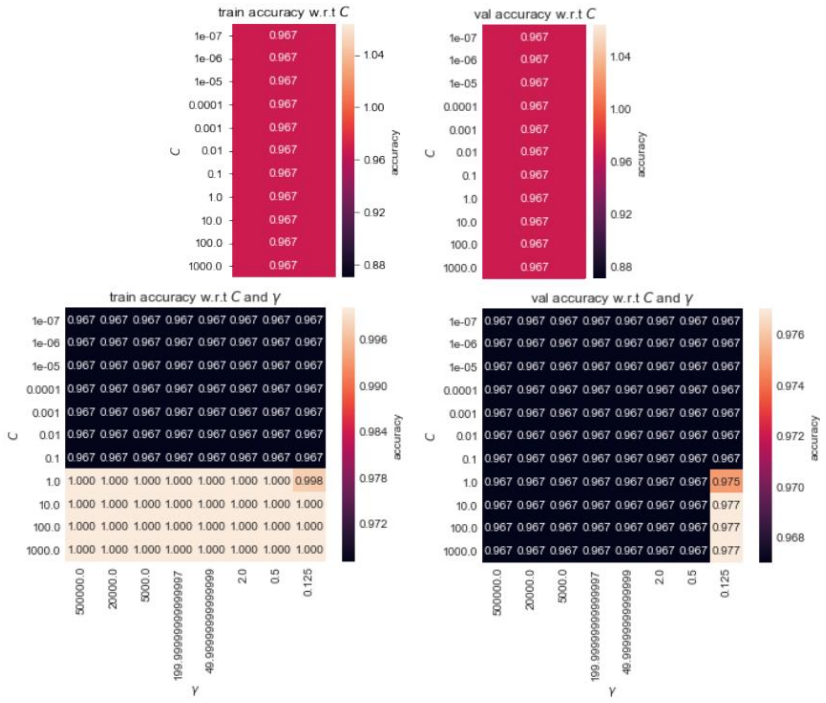


Figure 1: The left image shows training and testing accuracies for SVM algorithms in the Adult dataset, 20:80 split.
Figure 2: The right image shows training and testing accuracies for SVM algorithms in the Cover Type dataset, 20:80 split.



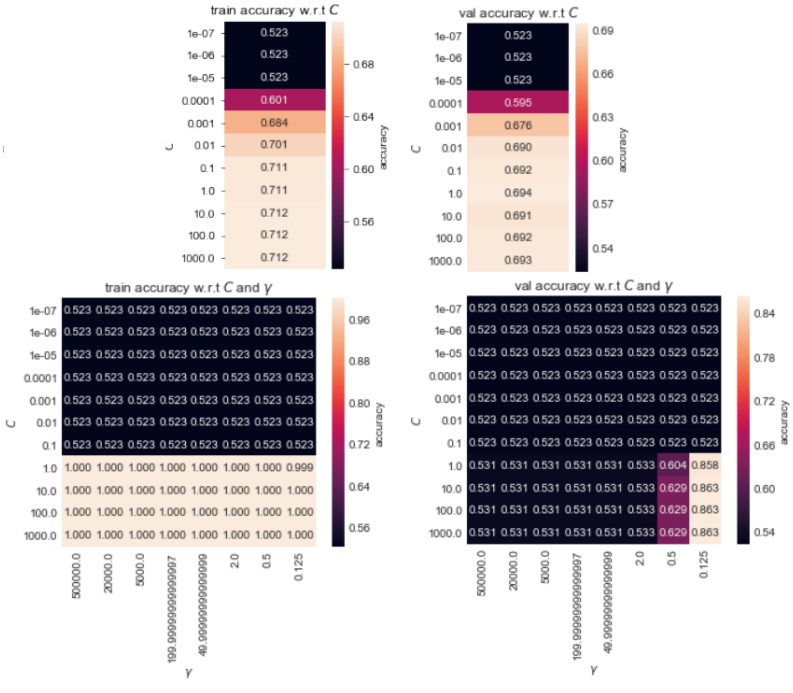


Figure 3: The left image shows training and testing accuracies for SVM algorithms in the Letter - p1 dataset, 20:80 split.
Figure 4: The right image shows training and testing accuracies for SVM algorithms in the Letter - p2 dataset, 20:80 split.

	Adults	Cover Type	Letter-p1	Letter-p2	**Averaged Results**
Linear SVM	0.851	0.766	0.962	0.962	0.885
RBF SVM	0.853	0.807	0.992	0.967	0.905
KNN	0.838	0.757	0.981	0.894	0.867
RF	0.848	0.821	0.986	0.948	0.901
	Adults 20:80	Cover Type 20:80	Letter-p1 20:80	Letter-p2 20:80	**Averaged Results**
Linear SVM	0.831	0.751	0.962	0.717	0.815
RBF SVM	0.834	0.754	0.975	0.884	0.861
KNN	0.821	0.716	0.964	0.772	0.818
RF	0.840	0.765	0.971	0.887	0.866
	Adults 50:50	Cover Type 50:50	Letter-p1 50:50	Letter-p2 50:50	**Averaged Results**
Linear SVM	0.852	0.750	0.965	0.726	0.823
RBF SVM	0.855	0.783	0.986	0.932	0.889
KNN	0.819	0.746	0.980	0.853	0.850
RF	0.838	0.805	0.987	0.930	0.890
	Adults 80:20	Cover Type 80:20	Letter-p1 80:20	Letter-p2 80:20	**Averaged Results**
Linear SVM	0.842	0.756	0.964	0.740	0.826
RBF SVM	0.841	0.802	0.990	0.954	0.897
KNN	0.838	0.775	0.975	0.880	0.867
RF	0.846	0.820	0.983	0.945	0.899
Mean Values					
	Adults	Cover Type	Letter-p1	Letter-p2	**Averaged Results**
Linear SVM	0.844	0.756	0.963	0.786	0.837
RBF SVM	0.846	0.786	0.986	0.934	0.888
KNN	0.829	0.748	0.975	0.850	0.850
RF	0.843	0.803	0.982	0.927	0.889

Figure 5: A table showing the obtained testing accuracies for all four partitions, as well as the averaged results. The bottom table contains the mean values for all the results, and the averaged results from that table gives us the overall best classifier.

The Cover Type dataset loaded the data file and removed any unnecessary commas. Before dividing the data, the Counter function was used to discover what the largest class size was, and then the data was modified so that the largest class would be treated like a ‘1’ and the rest would be treated as ‘0’.

Letter - p1 and Letter - p2 used similar methods performed on the same dataset. Both codes loaded the data and removed the commas from them. However, Letter - p1 has a function that replaces all mentions of ‘O’ with ‘1’ and converts the rest of the letters as ‘0’, while Letter – p2 replaces all occurrences of letters ‘A’ through ‘M’ with ‘1’ and converts the rest of the letters as ‘0’.

For all datasets, the data was shuffled randomly and partitioned the same way as the original paper. Each dataset aimed to predict an outcome in the datasets. For the Adult dataset, the variable to predict was whether income was greater than 50,000 a year. For the Cover Type dataset, the variable to predict was the Forest Cover Type Designation. For the Letter - p1 dataset, the variable to predict was whether the letter was ‘O’ or not. For the Letter - p2 dataset, the variable to predict was whether the letter was within the range of ‘A’ through ‘M’.

Finally, additional processing was done on the data by taking 5,000 points from the entire, randomized dataset and examined using 20:80, 50:50, and 80:20 ratios for the division of training and testing data. While not ideal, this reduced the time to analyze the data by a large amount, and was a necessary sacrifice to get the project done at a reasonable pace. Besides these differences, the procedures performed were the same for the various partitions.

3. Results

A general overview of the results can be seen in Fig. 5. Here it can be seen that RF has the highest testing accuracy by a very small margin, SVM-rbf has the second highest, KNN is third and SVM-linear ends up last. When looking at the mean values individually, SVM-rbf ends up beating

RF by slight margins, but loses in the Cover Type dataset. RF turns out to be the most consistent method, and moments where it loses out in accuracy to SVM-rbf is compensated by its speed.

Following the original paper's split between training and testing data yielded the best results overall. Overall, it seems to be a better idea to train the dataset by increasing the training dataset, rather than increasing the testing dataset. More testing would be required to make a conclusive judgement about this though.

Looking at the data more closely shows some unexpected results. Figures 1, 2, 3, and 4 all show symptoms of very similar accuracies within the results no matter what the regularization parameter was or the radial width. Although the diagrams only show the results from the 20:80 split, similar results can be seen for all the other splits. Another oddity in the results that stood out was the unusually high testing accuracies for Letter - p1, which can be seen in Fig. 5. Results were very consistent across all training and testing splits, and across all classification methods.

3.1. Comparisons to Original Paper

Although the results are different from the original paper, the same conclusions were obtained. In the original paper, RF was the second best, SVM was fourth place, and KNN ended up in sixth place, which was the same order obtained in this project. However, accuracies were higher across the board, except for Random Forests, where similar results were obtained. The biggest difference between the results obtained in the paper and the results obtained in this project was from KNN.

4. Conclusions

Despite the large amount of difficulties in doing this project, when averaged the results obtained are comparable to what is seen in the original project.

Without any additional knowledge or detailed results from the original paper, it is difficult to tell whether the results obtained in Figures 1, 2, 3, and 4 are intentional or not. The exact cause of the results remains unknown, but unless there was a large error with the way the data was preprocessed or with the parameters used, the most likely explanation is that it is just the nature of the data. Binary classification can result in some imbalanced results at times, which can be seen most prominently in the results for Letter - p1 in Fig. 5.

Admittedly, this experiment has many flaws that could have been amended with more time and experience. Because the original paper did not go into extensive detail with their procedures, much of this paper's procedures were carried out with guesswork and existing knowledge. Hardware and time constraints also played a large role in the differences in results. This paper's analysis was done without cloud computing or using a super computer so some sacrifices had to be made by drastically reducing the size of the original datasets when partitioning the data. Additionally, this project did not go into as much specific detail into analyzing the most optimal classification methods. In the future, possible improvements could have been made by further tuning the parameters for the classification methods and preparing more metrics to analyze.

5. Bonus Points

Personally, I believe this project deserves bonus points due to an above average effort being made to analyze the data. I effectively analyzed the data 64 times, because I used four classifiers (SVM-linear, SVM-rbf, KNN, RF), on four datasets (Adults, Cover Type, Letter-p1, Letter-p2), with four different partitions (original values, 20:80, 50:50, and 80:20). The detail and analysis this paper goes through is much higher than expected as well.

References

[1] Rich Caruana and Alexandru Niculescu-Mizil. 2006. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (ICML '06). ACM, New York, NY, USA, 161-168. DOI: https://doi.org/10.1145/1143844.1143865.