Comparing SVM, Logistic Regression, and Random Forest Classifiers

EJ Ozyazgan
A13833934

Abstract

I looked at the performance of three different classifiers: SVM, Logistic Regression, and Random Forests. In order to test their performance, I compared them using three datasets with three partitions each. I tested with a variety of parameters on each classifier to best optimize them. Overall the Random Forests did the best followed by Logistic Regression and SVM.

Introduction

There have been multiple studies outlining the performance of these classifiers. Rich Caruana's, *Empirical Comparison of Supervised Learning*, compared SVM, Logistic Regression, Random Forests, as well as others. However, I only looked at Linear SVM, Logistic Regression, and Random Forests for my study.

Similarly, to Rich Caruana's study I looked at multiple parameters such as the regularization parameter and radial width. More detailed implementation about each classifier implementation bellow in methods section.

These classifiers are used in all sorts of modern systems form self-driving cars to predicting your next purchase and optimizing development systems.

It is vital to understand when these classifiers perform best given the data and results required, in order to have the most accurate systems.

Data and Problem

I used three different datasets from the UCI Machine Learning Repository: Mushrooms, Wine Quality, and Dota2.

The mushroom dataset consists of 23 species of gilled mushrooms each with a multitude of different attributes. There is a total of 22 attributes used to determine if the mushroom is poisonous or not.

Wine Quality had two datasets, one for white and one for red wines, however, I choose to focus on white wines. This dataset used 11 different attributes to determine the quality of the wine from 0 – 10. For the study I simply wanted to predict if the wine had good quality > 5 or poor <= 5.

The last dataset, Dota2, was a compilation of Dota2 game data. Each game had 116 attributes in order to account for the variety of playable champions in the game as well as different game modes. The data was used to predict if a team would win a game given the champions being played and game mode.

Using these three datasets I tested each one against my three classifiers to see how they compared.

Mythology

1. Classifiers

I choose to use the following three classifiers: SVM, Logistic Regression, and Random Forests. Below is the breakdown of how I parameterized each one.

SVM: I used an SVM with a linear and kernel, radial width {0.001, 0.005, 0.01, 0.05}, as well as a variety of regularization parameters ranging from 10^{-3} to 10^{0}.

Logistic Regression: For my logistic regression classifier I varied my regularization parameters from 10^{-3} to 10^{0}.

Random Forests: I used random forest with 100 estimators.

2. Data Set Up

To set up my data, I created three partitions for each data set of 80/20, 50/50, and 20/80. This allowed me to see the effect of more training data vs test accuracy for each classifier. I also created three trials within each partition to be able to getter test the data by having three separately shuffled sets of data for each partition.

3. Performance Metrics

When comparing the different classifiers and datasets, I used their best accuracies from each trial and each partition to calculate the overall best accuracy for each dataset and partition for each classifier. I also used f-score to get a better understanding of each classifiers precision and recall with each dataset.

Experiments

The first classifier I tested was SVM. For the mushroom dataset, this classifier performed with perfect accuracy on all three partitions. Although this may be due to the way this data set was structured. Since it was all categorical data, I had to one hot encode it to better fit the algorithms. This encoding could have caused the data to be to easily predictable.

SVM did not perform as well on the wine data set with an average accuracy of 0.49 for 80/20, 0.49 for 50/50, and 0.48 for 20/80.

For the Dota data SVM performed roughly the same as the wine data with 0.62 average accuracy for 80/20, 0.62 for 50/50, and 0.65 for 20/80.

The linear regression classifier performed similarly as SVM on the mushroom data

On the Wine dataset linear regression yielded an average accuracy of 0.72 for 80/20, 0.73 for 50/50, and 0.71 for 20/80 as well.

For the Dota data linear regression resulted in an average accuracy of 0.61 for 80/20, 0.62 for 50/50, and 0.64 for 20/80

For the random forests classifier, the mushroom data performed like the other two.

Random forests classifier was able to predict an average accuracy of 0.89 on the wine data for 80/20, 0.86 for 50/50, and 0.84 for 20/80.

For the Dota set, random forests predicted average accuracies of 0.85 for 80/20, 0.85 for 50/50, and 0.84 for 20/80.

Conclusion

In conclusion the Random Forests classifier defiantly performed the best across all data sets. Followed by Linear Regression as the second best and SVM as worst.

There was a general trend that with less training data there was a decrease in accuracy, although this change was not very drastic. In some instances, there was an increase in accuracy between the 80/20 and 50/50 partition with the Wine and Dota datasets in linear regression.

SVM performed quite worse than the other classifiers. This could be due to parameters used. I also tried using a rbf kernel with SVM along with a variant of C and gamma values, however there was no change in the accuracy.

Random forest performed better than the other classifiers as expected. This classifier got decent amount of accuracy with roughly 0.85 for both Wine and Dota across all partitions.

The mushroom data consistently got a 1.0 accuracy on all partitions across all classifiers. I tried cleaning the data to better fit binary classification as well as trying to focus on less attributes. I thought maybe the one hot encoding was casing the data to be to easy to predict since all the data got encoded to be 1 or 0. Since all 22 attributes needed to be encoded, I ended up with 96 attributes after encoding.

Although the Dota data had similar 116 attributes which were also mainly 1 and 0, so I do not believe the about and type of attributes are the reasoning for the 1.0 accuracy.

I also tried to focus on only half the attributes, however I still ended up with a nearly perfect 0.99 accuracy for all classifiers and partitions.

References

Caruana, Rich, and Alexandru Niculescu-Mizil. “An Empirical Comparison of Supervised Learning Algorithms.” Proceedings of the 23rd International Conference on Machine Learning - ICML '06, 2006, doi:10.1145/1143844.1143865.

Mushroom records drawn from The Audubon Society Field Guide to North American Mushrooms (1981). G. H. Lincoff (Pres.), New York: Alfred A. Knopf

Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal @2009