Logistic Regression, SVM, and Random Forrest Comparison

Pedro Grimaldo Jr.
University of California, San Diego
Department of Cognitive Science
pgrimaldo@ucsd.edu

Abstract

In this study, we take a look at which machine learning classifier is best out of Logistic Regression, SVM, and Random Forrest. We took datasets from the UCI Machine Learning Repository in order to run these experiments. During our experiments we are testing different train/test splits in order to determine which classifier performs. As we run multiple tests along different trials and splits we are able to see that all classifiers perform really well.

1 Introduction

In machine learning we are learning as we go. We are able to take the tools that have been developed and improve on them. Since data is so diverse and complex, we are able to use different tools in order to help us understand it. When we look at data, there are many ways that we can interpret it and different ways to test things. We wanted to take a look at some datasets and run them through classifiers to see how they would work. The classifiers that we chose are, Logistic Regression, Support Vector Machine (SVM), and Random Forrest.

We believe that these classifiers would be the best in order to test out how well they work with different types of data. We tested each of them by using three different train/test splits (20/80, 50/50, 80/20) to see how the amount of training data would affect the performance. By testing them like this we are able to get the accuracy of each classifier and we are able to determine which one works better for what type of data.

2 Methodology

2.1 Classifiers

We chose three different classifiers that could cover different types of modeling like linear model, kernel-based model, and a tree-based model. This allowed us to test and compare the different types of classifiers.

2.1.1 Logistic Regression

Logistic regression classifier would help us look at data that is more linear and easily separable.

2.1.2 SVM (RBF)

Support vector machine with a radial basis function classifier would help us look at data that can be a bit more complex. It usually performs well when the data is structured.

2.1.3 Random Forrest

Random forrest helps us look at nonlinear patterns and is able to perform really well when there is a lot more data.

2.2 Datasets

We chose three different datasets that had a different amount of entries to test out the classifiers. Each of them would allow us to look at the strengths of each classifier with different data.

2.2.1 Wine Quality

This dataset had a total of 1599 entries for red wine. The way that we took a look at this data is that red wine was given a score from 0-10 and we made it so that if it had a score of 7 or more it would be considered "good" and if it had a score less than 7 then it would be considered "not good". We wanted to make sure that the data was split into two groups for when we tested the data.

2.2.2 Heart Disease

This dataset had a total of 297 entries for Cleveland. This dataset had a lot of instances but we only focused on the presence instance from the Cleveland study. It had a few missing values but we made sure to exclude them before we tested. If the patient had heart disease it was assigned a "1" and if the patient did not it was assigned a "0". This one was our smallest sample size to see how it would affect the classifiers.

2.2.3 Breast Cancer

This dataset had a total of 569 entries. For this dataset we also separated them into two groups being: malignant as a "1" and benign as a "0". This dataset was chosen because it seemed to be the cleanest looking dataset that can help test each classifier.

3 Experimental Results

The following tables are the accuracy of the tests with different train and test splits:

Table 1: 20/80 Split: Test Accuracy Across Datasets

Classifier	**Breast Cancer**	**Heart Disease**	**Wine Quality**
Logistic Regression	0.9671	**0.8221**	0.8706
Random Forest	0.9415	0.8137	**0.8750**
SVM (RBF)	**0.9708**	0.8137	0.8628

Table 2: 50/50 Split: Test Accuracy Across Datasets

Classifier	**Breast Cancer**	**Heart Disease**	**Wine Quality**
Logistic Regression	0.9731	0.8345	0.8775
Random Forest	0.9520	0.8300	**0.8992**
SVM (RBF)	**0.9731**	**0.8345**	0.8867

Table 3: 80/20 Split: Test Accuracy Across Datasets

Classifier	**Breast Cancer**	**Heart Disease**	**Wine Quality**
Logistic Regression	**0.9825**	0.8333	0.8740
Random Forest	0.9561	0.8222	**0.9250**
SVM (RBF)	0.9795	**0.8444**	**0.9250**

4 Discussion and Conclusion

Based on the results, we are able to see that they all performed different for every dataset. Logistic regression performed well across all of the datasets, but performed better on the heart disease dataset. SVM performed the best on the breast cancer dataset since it was very structured. Random forrest performed the best on the wine quality dataset, since it was a bigger dataset it allowed the classifier to form the different trees in order to make up the results. By changing the train and test split we can see the difference and as we grow our training size we are able to see better results.

5 Potential Problems

5.1 Small Sample Size

Having a small sample size is a problem because not enough data is being used to train and test these classifiers. Usually we would want a bigger sample size in order to make sure that over fitting does not happen.

5.2 Overfitting

Overfitting is also a problem that can arise. By using random forrest, it can overfit some sections not making it a valid score because when we train the data there may not be enough data causing it to have a lot of trees with not much data in them.

5.3 Two Classes

A problem that could happen with having a lot of classes like our wine dataset having groups of 11 groups ranging from 0-10 and converting that to 2 groups may cause us to lose potential information. We have it a hard cut at 7 so it could potentially limit the classifier.

6 References

Caruana, R. & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd International Conference on Machine Learning (ICML '06), pp. 161–168.

A Appendix

The following tables are the Train and Test Results from each of the classifiers. They are separated by their respective split.

Table 4: 20/80 Train-Test Split: Average Accuracy Across Datasets

Classifier	**Breast-train**	**Breast-test**	**Heart-train**	**Heart-test**	**Wine-train**	**Wine-test**
Logistic Regression	0.985	0.967	0.881	0.822	0.885	0.870
SVM (RBF)	0.982	0.971	0.938	0.814	0.961	0.863
Random Forest	1.000	0.941	0.994	0.813	0.984	0.875

Table 5: 50/50 Train-Test Split: Average Accuracy Across Datasets

Classifier	**Breast-train**	**Breast-test**	**Heart-train**	**Heart-test**	**Wine-train**	**Wine-test**
Logistic Regression	0.989	0.973	0.851	0.834	0.884	0.877
SVM (RBF)	0.993	0.973	0.864	0.834	0.968	0.887
Random Forest	0.998	0.952	0.982	0.829	0.999	0.899

Table 6: 80/20 Train-Test Split: Average Accuracy Across Datasets

Classifier	**Breast-train**	**Breast-test**	**Heart-train**	**Heart-test**	**Wine-train**	**Wine-test**
Logistic Regression	0.986	0.982	0.850	0.833	0.881	0.873
SVM (RBF)	0.990	0.979	0.855	0.844	0.979	0.925
Random Forest	0.997	0.956	0.974	0.822	0.998	0.925