A Student's Comparison of Three Supervised Learning Algorithms

Kailey Ayala, *Student, Cogs 118a*

***Abstract*— This paper is a comparative analysis of three supervised learning algorithms on three datasets collected from the UCI machine learning repository. It attempts to replicate the study done by Caruana and Niculescu-Mizil (2006). The algorithm considered are SVM (linear and RBF kernels), Decision Tree, and KNN. A comparison of different training and testing ratios as they effect the testing accuracy of classification is included.**

***Index Terms*— SVM (Support Vector Machine), RBF (Radial Basis Function), KNN (K-Nearest Neighbor).**

I. Introduction

THIS is a partial replication of the study done by Caruana and Niculescu-Mizil (2006) where-in they took a variety of algorithms available at that time and compared their performance on ~10 different datasets. In this study three algorithms from that paper were chosen, namely a SVM (Support Vector Machine), Decision Tree, and KNN (K-Nearest Neighbor). These were each run on three datasets from the UC (University of California) Irvine Machine Learning Repository. The datasets chosen were for predicting breast cancer, wine quality, and location of wine production. The use of multiple datasets allows for the generalization of the algorithm's efficacy to a variety of real world problems. The metric used to determine efficacy is testing accuracy.
In addition, three different training and testing set size ratios were used and the average error by algorithm over all datasets used in this study are presented. This is to illustrate the general trend presented by variation in the training and testing ratio.

II. Methods

Three datasets were acquired from the UCI Machine Learning Repository. These were chosen for being datasets suited to supervised learning and, in particular, classification. They were the Wine Recognition Data set (Forina 1998), Wine Quality data set (Cortez, et al. 2009), and the Wisconsin Diagnostic Breast Cancer data set (Lichman 2013). The Wine Recognition Data set has 178 instances with 13 features including the label feature. This data set was designed for the prediction of the location of the production of the wine. The label feature was original three locations for the reason of computational time cost these were grouped into a binary label with location 1 being the positive label and location 2 and 3 being the negative label. The Wine Quality data set has 4898 instances with 12 features including the label feature. The feature label originally was a discrete ranking from 0 to 10 however due to time and computational time cost this was grouped into a binary label where the positive label was all scores above and including 5. The Wisconsin Diagnostic Breast Cancer data set has 699 instances and 10 useable features including the label feature. All data processing was done in Jupyter Notebook using python 2.7. For the SVM and Decision Tree algorithms scikit-learn methods were used. The KNN algorithm was hand coded. The hyper parameters for these algorithms were tuned by being run for a variety of different hyper parameters. Those with the presumed best generalization were chosen to be used. While tuning these parameters fivefold cross-validation was used for all models. For the SVM two different kernels were used: linear and RBF (Radial Basis Function). For the training of parameters C and gamma for the RBF SVM a method similar to grid search was used. The difference was the testing values of C were not uniformly spaced. The output of one of these grid searches is shown in figure 1. The after the parameters were tuned the testing accuracy was found with a variety of different training testing ratios. For the wine location and breast cancer data sets training/testing ratios from 0.1 to 0.9 increasing at an interval of 0.1 were found. This is because they were shorter and fit the study's time requirements. However, for the wine quality dataset the only training/testing ratios are from 0.6 to 0.8 due to the size of the dataset and therefore the computational time cost.

```
[[   0.     0.1    0.2    0.3    0.4    0.5    0.6    0.7  ]
 [   1.     0.041  0.056  0.064  0.074  0.083  0.094  0.104]
 [  10.     0.04   0.052  0.06   0.069  0.079  0.087  0.094]
 [  20.     0.04   0.053  0.059  0.069  0.08   0.086  0.095]
 [  50.     0.039  0.052  0.06   0.07   0.079  0.086  0.095]
 [ 100.     0.041  0.052  0.059  0.069  0.079  0.086  0.094]
 [ 150.     0.04   0.053  0.059  0.07   0.079  0.086  0.095]
 [ 200.     0.041  0.052  0.059  0.069  0.079  0.086  0.095]
 [ 250.     0.041  0.052  0.059  0.069  0.079  0.087  0.095]
 [ 300.     0.04   0.052  0.059  0.07   0.079  0.086  0.094]]
```

Figure 1. An example of the hyper parameter search used to find C and gamma in the RFB SVM. This one was from the breast cancer data. The top row is gamma parameters. The first column is C parameters. The remainder is the error from the SVM for those parameters.

III. Experiment

After each algorithm's hyper-parameter's were tuned to best classify the data using accuracy produced from a cross-validation, each algorithm was then run on the dataset with a random 60/40 training/testing split. The algorithm that performed best on average across the datasets was the linear SVM with an average accuracy of 0.97. The SVM RBF would have taken second had it not been for the Wine Recognition

Submitted 17 June 2017.
Student in the Department of Cognitive Science at UCSD. (e-mail: kaayala@ucsd.edu).

data set. Instead KNN came in second with an average accuracy of 0.946. Decision Trees came in third with an average accuracy of 0.943. In last came the RBF SVM with an accuracy of 0.872. The RBF SVM had the greatest discrepancy from the Caruana and Niculescu-Mizil (2006) paper. They had found that regardless of Kernel SVMs would have at least been ranked above Decision Trees.

For testing/training ratio [60:40]				
Testing Accuracy	KNN	Decision Tree	SVM Linear	SM RBF
Wine Recognition (location)	0.909	0.928	0.98	0.683
Wine Quality (white wine)	0.96	0.962	0.966	0.966
Wisconsin Diagnostic Breast Cancer	0.969	0.941	0.964	0.968
Average	0.946	0.943	0.97	0.872

Figure 2. This is a comparison of the performance of different classifiers on three different datasets. The performance metric is the accuracy of the classifiers on testing data with a 60/40 training/testing data ratio.

A variety of different training/testing ratios were used on each classifier. The purpose of this was to show the general trend of the training/testing split on the accuracy. The breakdown by the individual classifier better illustrates the general trend than the average over all the classifiers. This trend could have been shown better if there had been a normalization step before averaging.

Training/Testing ratio	KNN	Decision tree	SVM linear	SVM rbf	average
[60:40]	0.946	0.943	0.97	0.872	0.933
[70:30]	0.941	0.95	0.968	0.87	0.932
[80:20]	0.944	0.945	0.969	0.883	0.935

Figure 3. This is the breakdown of the training and testing ratio as it effects the accuracy of the classifier on testing data.

IV. Conclusion

From this investigation, it appears linear SVMs have the best accuracy when classifying data in general. However, this may be due to the limited number of algorithms being used for this investigation. The Caruana and Niculescu-Mizil (2006) paper suggests that the best algorithm would be a boosted decision tree. It must be considered however that recently ANN's have made significant advances and it is likely would vastly out perform any algorithm in this paper or the Caruana and Niculescu-Mizil paper.

With respect to the training/testing ratio variation, it should be noted that the while on average the accuracy increases as the ratio increases that this may not generalize to future test data because of the potential for increased variability in real world test data.

References

Caruana, Rich, and Alexandru Niculescu-Mizil. "An Empirical Comparison of Supervised Learning Algorithms." *Proceedings of the 23rd International Conference on Machine Learning* (2006): n. pag. Web. 16 June 2016.

Wine Quality Data Set
P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Wine (location) Data Set
Forina, M. et al, PARVUS -
An Extendible Package for Data Exploration, Classification and Correlation.
Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno,
16147 Genoa, Italy.

Wisconsin Breast Cancer Data Set
Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.