An Empirical Comparison of Supervised Learning Algorithms

Sahib Athwal

sahibathwal@gmail.com

Abstract

A number of supervised learning methods have been introduced to us throughout the course of COGS 118A. The objective of the study is to investigate a few Machine Learning Algorithms and determine if any of them are better than others at binary classification tasks. We present a large-scale empirical comparison between three supervised learning methods: Random Forests, Decision Trees, and KNN. An important aspect of our study is the use of a variety of performance criteria to evaluate the learning methods. This includes the training accuracy and the validation accuracy, which can analyzed visually through heat maps. The analysis will allow us to determine which of the following supervised learning methods are better for our binary classification tasks. The hope is to gain a better insight as to why this will be useful in real world scenarios. The project can definitely be taken further with more datasets and classifiers as this is merely a simple version.

1. Introduction

In the project, there were 5 datasets used to come to some sort of conclusion based on the data. I chose these specific data sets for two reasons because i found they might have some practical use and the other reason was because they related to me in someway making me more inclined to be invested into this project. All of the data sets were used from Kaggle, https://www.kaggle.com/datasets. The 5 datasets being used include the following: Company Bankruptcy Prediction from Taiwan Economic Journal (Liang and Tsai 2021), Students Performances on Exams (Seshapanpu 2018), Angry Birds Review (Cantekin 2020), Parkinson's Disease Speech Features (Sakar 2018), and Spam Email Classification (Naidu 2021). They will be referred to as the following: BANK, STUDENT, AB, PD, SP in that respective order. The supervised machine learning algorithms ran were KNN, Random Forests, and Decision Tree through google collaborate ipython notebook.

2. Methodology

The main idea was to implement each of the supervised machine learning techniques on the datasets in order to test the performance and accuracy of each classifier. The data was manipulated to use binary classification. Any cases of ordinal and nominal data were encoded using one hot encoding library to be able to fit the data for our binary classification. The general procedure was to go ahead and clean each respective dataset, then once ready run each of the techniques mentioned above for various random data values for 5 trials. Then we calculated all the training accuracy and the testing accuracy for each parameter to see which would be optimal in our solution. There were three main splits I used that include: 80:20, 50:50, and 20:80 for the trained set and validation set respectively. In each of my respective classifiers I maintained using the K 5 folds cross validation with variants to the size. I also went on to use the gridsearch when using

my K 5 folds cross validation, so I would be able to use the best possible parameters when running each classifier. For my random forest and decision tree, I found that changing the parameter to run 'Entropy' made the code run faster on google collaborate.

3. Results

Most of the data had relatively high training and testing accuracy spanning from ~0.6 all the way to 1. BANK by far had the highest training and testing accuracy nearing off 0.9-1. STUDENT had training and testing accuracy ranging from 0.77-0.82, which was not favorable in our case as well. PD had training and testing accuracy ranging from 0.7-0.82, which was better than STUDENT, but still not favorable. SP maintained a training and testing accuracy ranging from 0.65-0.9, which was the most accurate out of all of our data sets for all the respective classifiers. AB dataset also matched the range of the SP dataset evenly and through various trials some were higher than others and others were lower respectively. However, the analysis from our P and T tests guided us to disregard the data because comparing the various classifier methods for each dataset showed no statistical significance, which followed the results of our high accuracy. Overall, it seemed like the most efficient algorithm was our Decision Tree classifier (**Table 3.1**), but it was still too high of an accuracy to be considered truly accurate. The T and P Test results are also available in the **Appendix** section, and the data was all saying there is no statistical significance among each of the respective classifiers for each dataset.

Main Results

Table3.1 Mean Performance Accuracy For Each Classifier

	Decision Tree	Random Forest	KNN
Mean	0.803	0.812	0.851

Table3.2 Mean Performance Accuracy For Each Classifier Based on Data

	Decision Tree	Random Forest	KNN
BANK	0.9656	0.9663	0.9642
STUDENT	0.8275	0.7818	0.7709
PD	0.8243	0.7793	0.7668
SP	0.8146	0.7460	0.7593
AB	0.8246	0.7871	0.7557

4. Discussion

The reasoning behind such high training and testing accuracy stemmed from the fact that the data was limited and already relatively accurate, so there was not much to do in terms of the classification. As a result most of our training and testing was on an upper bound, which in a real world scenario would not be representative of how things are. To better further this experiment more comprehensive data that would be needed in order to be truly representative of each of the populations the data modeled. This must be the modern struggle of most machine learning scientists as well as data scientists because they do not always have comprehensive data.

5. Conclusion

This merely served as a warm up to more research to be done to better the classifiers being optimized. The hope is to get more data and have the data be truly classified for each of their respective features in order to decisively come to a conclusion that we can make some sort of statement about our data. I hope to further this research on my own in order to better my knowledge and to be able to use more classifiers for better training and accuracy. That way I can help contribute further to the machine learning community.

6. Extra Credit

I independently tested more data sets than required, and I tried to make the visuals for the experiment as appealing as can be so it would be readable. My hope is that with this and my algorithmic implementation of the classifiers I used, excluding the SVM which I did not have enough time to run for all my data sets but works completely fine, at O(n) runtime would compensate for some form of extra credit on this assignment. See the **Appendix** below for the visual tables and heat maps for every trial and overall the classifiers as well. My last case for extra credit was the fact that I did some research on what makes companies go bankrupt from a reputable source website that gave me insight as to which features I should maintain in order to run my classifier more efficiently.

References

UCSD COGS 118A Lectures, Videos, and Discussions

Caruana, Rich., & Niculescu-Mizil, Alexandru, (2006). *An Empirical Comparison of Supervised Learning Algorithms*. Department of Computer Science, Cornell University, Ithaca. https://www.cs.cornell.edu/~caruana/ctp/ct.papers/caruana.icml06.pdf

Deron Liang and Chih-Fong Tsai. (2021, February). *Company Bankruptcy Prediction*, Version 2. Retrieved March 18, 2021 from https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction

Sakar, C.O., Serbes, G., Gunduz, A., Tunc, H.C., Nizam, H., Sakar, B.E., Tutuncu, M., Aydin, T., Isenkul, M.E. and Apaydin, H., 2018. *A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform*. Applied Soft Computing, https://doi.org/10.1016/j.asoc.2018.10.022

Chandramouli Naidu. (2021, February). Spam *Classification for Basic NLP*, Version 1. Retrieved March 18, 2021 from https://www.kaggle.com/chandramoulinaidu/spam-classification-for-basic-nlp

Ahmettez Cantekin. (2020, July). *beginner_datasets*, Version 1. Retrieved March 18, 2021 from https://www.kaggle.com/ahmettezcantekin/beginner-datasets

Jacki Seshapanpu. (2018, November). *Students Performance in Exams*, Version 2. Retrieved

March 18, 2021 from https://www.kaggle.com/spscientist/students-performance-in-exams

Research Website For Bankruptcy

https://infimoney.com/signs-that-show-a-company-is-declining/

Stack Overflow + Documentation Used for Coding

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html

https://stackoverflow.com/questions/43211239/valueerror-unknown-label-type-continuous

https://stackoverflow.com/questions/41925157/logisticregression-unknown-label-type-continuous-using-sklearn-in-python/41925957

https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

https://stackoverflow.com/questions/37292872/how-can-i-one-hot-encode-in-python

https://stackoverflow.com/questions/60153981/scikit-learn-one-hot-encoding-certain-columns-of-a-pandas-dataframe

https://stackoverflow.com/questions/29576430/shuffle-dataframe-rows

https://scikit-learn.org/stable/modules/svm.html

https://scikit-learn.org/stable/modules/cross_validation.html

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

https://stackoverflow.com/questions/54057011/google-colab-session-timeout

https://www.kite.com/python/answers/how-to-convert-a-pandas-dataframe-column-from-object-to-int-in-python

https://medium.com/analytics-vidhya/using-the-corrected-paired-students-t-test-for-comparing-the-performance-of-machine-learning-dc6529eaa97f

Special Thanks

This is for the Professor mainly because I had a very rough quarter, and I appreciate him working so closely with me despite my conditions. I also want to thank the TAs for helping facilitate many questions during the sections and on the piazza posts because I saved a lot of time from their hard work answering questions I would have not thought of initially.