Final Project: An Empirical Comparison of Supervised Learning Algorithms

Anthony Martinez, atm010@ucsd.edu
COGS 118A, Winter 2020

1. Abstract

There is now a plethora of supervised learning methods. It is important to understand which method works best depending on the circumstances. This paper will try to recreate the results from a previous study conducted by Caruana and Niculescu-Mizel titled "An Empirical Comparison of Supervised Learning Algorithms"[1]. However, this paper will be on a smaller scale. For this paper, I will analyze only three methods: SVM, K-Nearest Neighbors, and Logistic Regression. It is important to try to replicate the results from Caruana and Niculescu-Mizil's paper in order to have a better understanding of supervised learning methods

2. Introduction

Throughout winter quarter in COGS 118A, we were introduced to many different supervised learning methods. We were taught these methods in hopes of having a better understanding of the two-class classification problem. Recently, there has been more interest in conducting a study comparing the effectiveness between the different supervised learning methods. The report done by Caruana and Niculescu-Mizil is probably the most famous now. It is important because it included the analysis of new learning algorithms at the time such as SVM, boasting, and random forest.

Each learning method can be used for a different purpose. For example, SVM is best used for a classification problem and/or regression problems. KNN is best used for the classification problem as well, but prefers smaller datasets since it lowers the runtime. There are tradeoffs between each learning method, so it is important to know what method works for which circumstance.

In Caruana and Niculescu-Mizil's paper, they do a large scale empirical comparison of ten supervised learning algorithms. In this paper, I will do a small scale empirical comparison of three supervised learning algorithms. For each algorithm, I will be exploring multiple parameters such as: training and testing data split, different data sets, and more. The goal will be to see if I can get the same results as Caruana and Niculescu-Mizil's paper. More specifically, I will be comparing the accuracy scores for each learning method from their paper with the accuracy scores I get.

3. Methods

1. Learning Algorithms

Since this is a small scale comparison, I will compare SVM, KNN, and Logistic Regression. I chose these three learning methods because they were my favorite methods I learned about in 118A. For each method, I will shortly explain what it is, why I chose it, and the parameters used for it.

SVM: SVM works by mapping data to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable[2]. I chose this method because it is so powerful and can be used for many circumstances. The kernel I use is linear and the random_state I set to 0.

KNN: KNN works best when the data is easily separable, therefore labeling each point as a certain label to classify the data. I chose this method because I found it very interesting, especially since it's considered a "lazy learning algorithm". Parameters are set at k=5(default), metric='minkoski', and p=2 which basically measures distance as euclidean distance.

Logistic Regression: Logistic Regression is used to estimate the probability of an event occurring or not(classification). I chose this method because it was the first learning method I learned. Parameters are set to default from the sklearn package in python.

2. Data Sets

I compare the algorithms on three different data sets that come from the UCI Machine Learning repository. The data sets I use are IRIS, BREAST_CANCER, and DIGITS[3]. Each data set is a classification problem, meaning that there are only two classes. IRIS contains 150 instances, BREAST_CANCER contains 569 instances, and DIGITS contains 5620 instances.

3. Main Algorithm

So as to achieve the results for this experiment, I used a Jupiter notebook to write my code. I used the programming language python and utilized the sklearn package to access different learning algorithms. In order to to perform a 3 x 3 x 3 x 3 experiment, that is 3 trials X 3 classifiers X 3 datasets X 3 partitions (20/80, 50/50, 80/20), we need to follow the pseudo code that was provided by the professor. The pseudo code is as follows:

For i in three different datasets

For j in three types of different partitions

For t in three different trials/repeats (shuffling or performing random splits for each type j (20/80,50/50,80/20))

For c in three different classifiers

cross validate

find the optimal hyper-parameter

train using the hyper-parameter above

obtain the training and validation accuracy/error

test

obtain the testing accuracy

compute the averaged accuracy (training, validation, and testing) for each classifier c out of three trials/repeats

rank order the classifiers

4. Experiments

For each trial in the experiment, I chose three different partitions of the data sets which are the following: 20/80, 50/50, 80/20. Compared to Caruana and Niculescu-Mizil's 5 trials, I did only 3 trials. However, it is still enough to discriminate the accuracy scores between classifiers. After running my algorithm in python, I get accuracy scores for each trial on each classier on each data set for each partition. So that is a total of 81 unique accuracy scores. When taking the average accuracy score for all uses of the SVM classifier, the accuracy is 95.4%. When taking the average accuracy score for all uses of the KNN classifier, the accuracy is 95.2%. When taking the average accuracy score for all uses of the logistic regression classifier, the accuracy is 94.3%. SVM has the highest accuracy score, however, KNN has almost the same score. Similarly, Logistic Regression is only a percent lower in accuracy than the other two classifiers. These accuracy scores are all really high which means these are strong working supervised learning methods. Compared to Caruana and Niculescu-Mizil's paper, my learning methods achieve higher average accuracies. However, the order of highest to lowest accuracy between methods is the same. Although, the No Free Lunch Theorem suggests, there is no universally best learning algorithm[1]. My results are unique because in each iteration, SVM was the best classifier followed by KNN and logistic regression. My results differ from Caruana and Niculescu-Mizil in that my classifiers are consistently in the same order for accuracy score. This could be caused by my implementation or size of data sets. Ultimately, my experiment does yield similar (although not completely similar) results as Caruana and Niculescu-Mizi.

Note: It was hard for me to make a table containing all of the accuracy scores, so they are included with my code in a Jupiter notebook.

5. Conclusions

Supervised learning methods have come a long way in the last 20 years. In COGS 118A we were able to learn about a lot of these methods. Ultimately, my results showed that SVM had the highest average accuracy score, followed closely by KNN and Logistic Regression. It is important to note that in each trial that SVM had the greatest accuracy followed by KNN and Logistic Regression. This is not consistent with Caruana and Niculescu-Mizil's paper, as sometimes other methods performed better than others. However, in my experiments all of the accuracy scores are above 90%, while in Caruana and Niculescu-Mizil's paper, the accuracy scores for these learning algorithms range from 65-83%. The accuracy scores in my experiment could be much higher in my experiment possibly due to the fact that two of my data sets are relatively small. Similarly, the training data accuracy scores are greater than testing data accuracy scores. It is clear that I received similar results as Caruana and Niculescu-Mizil's paper, but on a much smaller scale. It is also important to note that sometimes the best models can perform poorly and the worst models can perform very good. It mainly depends on the problem that you are working on. All in all, learning these supervised learning algorithms have been very interesting and I really enjoyed my time in COGS 118A.

6. References

[1] Rich Caruana and Alexandru Niculescu-Mizil, "An Empirical Comparison of Supervised Learning Algorithms", ICML 2006.

[2] (n.d.). Retrieved from https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/svm_howwork.htm

[3] (n.d.). Retrieved from https://archive.ics.uci.edu/ml/index.php

[4] A Statistical Learning Approach to Modal Regression

Yunlong Feng, Jun Fan, Johan A.K. Suykens; (2):1–35, 2020.

[5] Consistent Algorithms for Clustering Time Series

Azadeh Khaleghi, Daniil Ryabko, Jérémie Mary, Philippe Preux; (3):1–32, 2016.