Cogs 118A Final Paper

Dongyoung Kim

University of California, San Diego

An evaluation of different supervised learning algorithms

Abstract

Data collection methods have drastically improved over the past few decades, leading to the Artificial Intelligence revolution and the introduction of a number of supervised learning algorithms. Rich Caruna and Alexandru Niculescu-Mizil are notable contributors to new supervised learning algorithms, with their most popular work being that of a paper named “An Empirical Comparison of Supervised Learning Algorithms. In their paper, Rich Caruna and Alexandru Niculescu-Mizil compare the performance of multiple different supervised learning algorithms by testing each of them through a number of different datasets and comparing the prediction accuracies. In this paper, three of their algorithms and testing strategies are implemented and tested on three different datasets each, in order to further verify the comparative performances of these various algorithms and to test their generalizability. While this paper does not replicate their algorithms to the exactly, the method of implementation for these algorithmic learning strategies are largely representative of the same used by Rich Caruna and Alexandru Niculescu-Mizil and allow for comparative analyses thereof.

1 Introduction

Machine Learning researchers have made breakthroughs throughout the 21st century. The first wave of high powered, fine tuned supervised learning algorithms was brought to prominence in the late 1990s and early 2000s. However, Rich Caruna and Alexandru Niculescu-Mizil brought together different supervised learning algorithms run across different datasets and measuring the success under different metrics. Their study in 2010 showed that some supervised learning algorithms performed better than others when compared on aspects such as performance in higher dimensions and scale. They compare 10 supervised learning algorithms on 11 datasets using 8 unique metrics

This report entails a replica of their work, by applying three of their selection of supervised learning algorithms (boosted tree, random forests, and KNN) to three different datasets from the UCI machine learning repository (Adult, Bank, and Tic-Tac-Toe). The reasons for choosing some supervised learning algorithms over others is in the next section. One of the most notable absence is that of Artificial Neural Networks, and the reason for this omission is also stated in the next section. Moreover, this report does not include both model calibration and bootstrap analysis. Three datasets will be used to train three different models, with each model receiving an accuracy score on predicting the testing set by using the learned weights fitted from the training set. The three models each being tested on three datasets each will result in 9 different prediction accuracy scores.

2 Methodology

2.1 Learning Algorithms

This section of the report delves into the various supervised learning algorithms used in the work. These algorithms were chosen on a number of criterion, namely prominence, effectiveness, availability and ease of use. The algorithms utilize libraries such as scikit-learn and pandas, and the hyper-parameters are optimized with 3-fold grid search. The hyper-parameters used in this work is not exactly the same as that used in the original study as constraints in this experiment meant exact replication would not be possible. However, the algorithmic learning strategies of the models are comparably similar, allowing parallels to be drawn from the performance metrics of this experiment to the original study.

K-Nearest Neighbors (KNN): K-nearest neighbors is a non-parametric model used for supervised classification and regression. It takes as input the specified k number of training examples to the current datapoint and outputs a classification based on the majority labels of the k-closest neighbors. We perform this algorithm by utilizing the Euclidean distance and $1/r^2$ weight, where r is the distance from the point of interest to our chosen prototype. We choose the values of k to be of the following list: [1,10,20,30,40,50,100,150]

Random Forests (RF): Random Forests is an ensemble learning, in this case for classification, that operates through constructing multiple decision trees at a time and returning a label depending on the mode of the classes from each of the individual trees. A cross-validation method is used here, where the only hyper-parameter tuned is that of feature size at the time of splitting. The random forests method used here utilizes the Out-Of-Bag (OOB) estimate for cross validation.

AdaBoost: Gradient Boosting/boosted trees is a classification algorithm that procures its prediction model in the form of ensembles of weaker prediction models. It builds a stepwise model and generalizes said model by allowing for the optimization of an arbitrary differential loss function. Here, we use gradient boosting in decision trees as the base learners. There is no condition on the decision tree, but the number of learners and learning rate were tuned via cross validation.

2.2 Datasets

In this paper, we compare the performance of three supervised learning algorithms mentioned in the earlier section on three datasets. All the datasets used in this experiment were taken from the UC Irvine Machine Learning Repository. The datasets I have chosen in this experiment are banking marketing dataset, tic-tac-toe dataset and adult dataset. The banking-marketing dataset has features that are mixed categorical and real-valued, with a binary target of whether a prospective client will subscribe to a term deposit. The tic-tac-toe dataset lists every single possible outcome of a game of tic tac toe, with the board conditions encoded and represented through having each dimension represent a location on the 3x3 board; the label for this dataset is whether the player with the 'x' mark is victorious in that instance of the game. Lastly, the adult dataset contains features that are continuous in nature, with a binary target of whether a person has an annual income of at least $50,000.

In the case of the banking-marketing dataset, 10 base features were categorical in nature, and were hence represented using one-hot encoding technique. The rest of the features were real-valued and were Z-scored.

Table 1: Properties of datasets used

Dataset	**Number of attributes**	**Training set**	**Testing set**
Bank Marketing	17	36168	9042
Tic-Tac-Toe	9	766	191
Adult	14	26049	6512

2.3 Experimental Procedure

The classifiers mentioned in the earlier section were all trained and testing using 80-20 splits of the three datasets described above, with the parameters selected through gridsearch and the weights obtained through training the models on the datasets. The models used on the test set would be evaluated by using the hyper-parameters that yielded the best performance in the cross validation. We simply calculate the raw accuracy for classification to test each classifier. For each classifier, we train and test it on the three different datasets mentioned above. Each time, we perform cross validation to find the relevant hyper-parameters that correspond to the classifier being used. After the model has been fitted, the testing split set is used to obtain an accuracy score of the weights, and the average of these scores are taken to represent the performance of each model on a given dataset.

3 Results

Table 2: Classifier Performance

Classifier	Bank Dataset accuracy	Tic-Tac-Toe	Adult
KNN	**0.9353**	**0.9427**	**0.9401**
AdaBoost	**0.9659**	**0.9727**	**0.9482**
Random Forests	**0.9514**	**0.9666**	**0.9398**

4 Conclusion

All three of the algorithms performed reasonably well on all three of the datasets, achieving close to 95% mean accuracy scores on the testing set each time. As a whole, boosted trees and random forests performed slightly better than the KNN model, with a slight edge going to boosting method. The tic-tac-toe dataset returned the best results, as the target labels are almost directly related to the attributes. It also took the least time to train on, due to the smaller dataset size. These results are all in line with the results shown in Rich Caruna and Alexandru Niculescu-Mizil's paper, verifying their analyses of the performances of these supervised learning algorithms and increasing the generalizability of their findings.

References

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning - ICML 06*. doi:10.1145/1143844.1143865

Mertayak, C. (n.d.). AdaBoost. Retrieved from https://www.mathworks.com/matlabcentral/fileexchange/21317-adaboost

Cao, Y. (n.d.). Efficient K-Nearest Neighbor Search using JIT. Retrieved from https://www.mathworks.com/matlabcentral/fileexchange/19345-efficient-k-nearest-neighbor-search-using-jit