Replication of An Empirical Comparison of Supervised Learning Algorithms

Zhi Jia Teoh
zteoh@ucsd.edu

Abstract

For the past decade, quite a number of supervised learning methods have been introduced. Fortunately, various empirical studies was done comparing these methods. One of them was by Rich Caruana and Alexandru Niculescu-Mizil [1]. In this paper, we will be doing a smaller similar empirical study. Unsurprisingly, the results obtained were similar to the original report.

1. Introduction

During the era of the rise of machine learning, various algorithms have been created. Various scientists and researchers have created many supervised learning algorithms and eventually adopted into many programming languages in the form of external libraries.

However, with so many learning methods, one begs the question of which is the best. This question is exactly why comprehensive empirical studies have been performed. As noted in the abstract above, the study by Caruana and Niculescu-Mizil in 2006 is what we are trying to replicate here, albeit into a much smaller scale. Most notably, their study spanned across ten classifiers using eleven datasets. Furthermore, their study has shown that boosted decision trees and random forests to be of superior classifier. Predictably, classifiers such as support vector machines and logistic regression did not perform as well.

We will be performing a replication of their work using five different classifiers spanning across six datasets. The details of them will be explained in the subsequent sections.

2. Methodology

2.1 Procedure

For each of the classifiers listed below, we would explore the space of parameters as thoroughly as possible to obtain the best model possible. All classifiers except logistic regression will be using sklearn implementation of GridSearchCV with cross-validation size of 3. Additionally, all classifiers will also be using sklearn implementations [2]. For each dataset, we repeat each data splits of [20:80], [50:50] and [80:20] for three times to calculate the average accuracy. Lastly, it should be noted that the data is shuffled each time for better generalization.

2.2 Classifiers

Support Vector Machines: For support vector machines, we will be using poly kernel with the penalty parameter of the error term 'C': {.001,.01,1,10,100}. The degree of the kernel would be defaulted at 3.

Logistic Regression: For the multi-label classification, we will be employing 'newton-cg' method, whereas for binary-label classification, 'liblinear' method will be used. For 'newton-cg' method, L2 penalty term is used and for 'liblinear' method, L1 penalty term is used. The inverse regulation strength for both methods is with 'Cs': {0.01,0.1,1.0,10.0,100.0}.

K-Nearest Neighbor: For this classifier, the choice of K will be spanned across 26 equally spaced intervals, with the maximum being the the size of the training set. The weights will be uniformed therefore all points in neighborhood are weighed equally.

Random Forest: For this classifier, the criterion set is 'gini' and the n_estimators (number of trees) will be {64,128,256,512,1024,2048}. The maximum features considered will be equally spaced until the max features of the dataset.

Gradient Boosting: For this classifier, the loss function to be optimized will be 'defiance' (= logistic regression). The n_estimators (number of boosting stages) will be {64,128,256,512,1024,2048} and the learning rate will be {.001,.01,.1}.

2.3 Datasets

Adult (Census Income) [2]
The original dataset has 48 842 instances and 14 attributes. However, for time complexity sake, we will be instead using only 10 000 instances. The attributes range from age, workclass, race and so on. The predictive label (classification label) will be to predict the income of <50k or => 50k. Therefore, it has binary classification. Some attributes we trimmed down such as age.

Bank Marketing [2]
The original dataset has 45 221 instances and 17 attributes. This dataset contains a lot of missing information valued as "unknown". Therefore, any instances with unknown values were dropped. The modified dataset therefore has 7842 instances which is sufficient. Various attributes were trimmed down such as age, duration, balance and so on. The predictive label is whether the client will subscribe (yes/no) a term deposit. Therefore, it has binary classification.

Connect-4 [2]
This dataset contains 67 557 instances and 42 attributes. It contains all legal 8-ply positions in the game of connect-4 in which neither player has won yet, and that the next move is not forced. The classification label will be the theoretical value for the first player. This dataset is not binary classification as it has three class i.e. win, loss or draw.

Letter Recognition [2]
This dataset contains 20 000 instances with 16 attributes. The objective of this dataset is to identify the capital of the alphabets (A-Z). Therefore it's a multi-class classification problem. It has various features ranging from width, height, positions and so on.

Nursery [2]
This dataset contains 12 960 instances and 8 attributes. The objective is to identify the recommendation status of the nursery school. The features range from parents, children and so on. It is a multi-label classification problem and the target has three classes.

Skin Segmentation [2]
Our final dataset contains 245 057 instances with 4 attributes. For complexity sake, only 10 000 instances will be used. The features are the color code (R,G,B) and the decision class label is binary (1,2)

3. Results

Note that for each values in the table below, the accuracy has been averaged out of the three tries

3.1 Final Results (all splits combined)

	LOGREG	SVM	KNN	RF	GB
Adult	0.757	0.757	0.78765	0.79625	0.82725
Bank	0.81893	0.82333	0.82052	0.78817	0.82084
Connect-4	0.66475	0.66575	0.66788	0.71363	0.74713
Letter	0.61438	0.77583	0.76855	0.82188	0.77861
Nursery	0.75897	0.76060	0.84895	0.96817	0.99874
Skin	0.91875	0.930875	0.997	0.992625	0.99025
Mean accuracy	0.75546	0.78556	0.81509	0.84679	0.86047

3.2 Classifiers Ranking based on mean accuracy

Classifier	Mean Accuracy
Gradient Boosting	0.86047
Random Forest	0.84679
K-Nearest Neighbor	0.81509
Support Vector Machines	0.78556
Logistic Regression	0.75546

3.3 Discussion

From the table above, skin dataset gives very high accuracy for all classifiers. Intuitively this makes sense since the dataset features are based on human skin color and the fact that color spectrum is very distinctive, one could easily predict the outcome. For nursery dataset, we can see that random forests and boosting outperforms other classifier by a large margin particularly boosted trees that yields

nearly perfect accuracy. Next, the letter dataset is interesting as it shows that almost all classifiers beside logistic regression yields somewhat similar accuracy. In particular, most classifiers yield approximately 77% accuracy, but LOGREG classifier yields only 61% accuracy. The connect-4 dataset also yields some interesting result as all three SVM, LOGREG and KNN only has 66% accuracy whereas RF and GB yields around 71% accuracy. The bank dataset has good accuracy for all classifiers, with all performing decently good at 82% accuracy except RF at 78%. Lastly, for the adult dataset, we could see that again that LOGREG and SVM perform much worse (75%) compared to the remaining three classifiers that is nearing 80% accuracy.

4. Conclusion

From the results obtained, we could see that empirically, random forests and boosted trees gives the best accuracy. K-nearest neighbors classifier accuracy lies somewhat in between the rest. As expected, SVM and LOGREG classifiers produces the worse accuracies among them, in particular LOGREG.

We managed to somewhat replicate the results obtained in the original studies, albeit this report is not rigorous enough in the sense that the number of classifiers and datasets used is much lesser than the original.

5. References

[1] Caruana Rich, Niculescu-Mizil Alexandru (2006) - An Empirical Comparison of Supervised Learning Algorithms

[2] Blake, C., & Merz, C. (1998) - UCI repository of machine learning databases.