An Empirical Comparison of Supervised Learning Algorithms

Fan Ding
University of California, San Diego
COGS 118 A
fading@ucsd.edu

Abstract— A number of supervised learning methods have been introduced in the last decades. In this report, we are interested in evaluating the performance of 3 supervised learning algorithms --- Random Forest, SVM, and logistic regression. We test on three datasets. Another important aspect of our study is the use of a variety of performance partition to obtain better results.

I. INTRODUCTION

Machine learning is becoming more and more popular. Much research has been done to evaluate different supervised machine learning algorithms. In this study, we are specifically interested in measuring the performance of 3 supervised learning algorithms, namely random forest, SVM and logistic regression. We used the latest SK-Learn package. Since the performance of an algorithm depends not only on the algorithm itself but also on the dataset, we measured the performance of these three algorithms on three totally different datasets. We presented our results here: The average accuracy for Random Forest classifier for three datasets are around 0.63, 0.84, 0.56. The average accuracy for SVM kernel classifier for three datasets are around 0.72, 0.75, 0.54. The average accuracy for logistic regression classifier for three datasets are around 0.72, 0.79, 0.53. The SVM kernel and logistic regression have more stable performance.

II. METHOD

Datasets

a) Car

This data set contains 1728 instances with 7 attributes, namely overall price, buying price, maintenance price, number of doors, capacity, luggage size and safety rating. The purpose of this dataset is to determine whether a car's safety level based on the given parameters. We digitize 'high" to 1 and "med" or "low" to 0. In addition, we one-hot-code all the nominal columns in the dataset into binary number 0 and 1.

b) Adult:

This data set contains 48842 instances with 14 attributes. namely, age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, and so on. The purpose of this dataset is to determine whether a person can make annual income greater than 50k based on the given parameters. We pick '>50" to 1 and "<= 50" to 0. We also one-hot-code all the nominal columns in the dataset into binary numbers 0 and 1.

c) Forest fires

This dataset contains 518 instances with 13 attributes. The goal of this dataset is to determine that the burn area base on all parameters. We one-hot-code all the rest parameter to 1 and 0.

2. Training Classifiers

a) Random forest

Random forests is a joint learning method for classification, regression, and other tasks that operate by building a multitude of decision trees at the time of training and generating the class that is the mode of classes (classification) or mean prediction (regression) of individual trees. Random decision forests correct decision trees' habit of overfitting their training set. We train our Random forest classifier with max feature{1,2,4,6,12,16,20},number of trees{1024}

b) SVM RBF

Gaussian RBF is another popular Kernel method used in SVM models. The RBF kernel is a function whose value depends on the distance from the origin or from some point. Our SVM classifier takes the kernel “rbf”, C list{0.001,0.01,0.1,1 and 10} and the gamma list {0.001,0.01,0.1,1}

c) Logistic regression

The logistic model is used to model the probability that a certain class or event exists. This can be extended to model various kinds of events, such as determining whether an image contains objects. Each object detected in the image will be assigned a probability between 0 and 1 and the sum by adding one. The logistic regression classifier takes C list {0.001,0.01,0.1,1,10,100,1000}

III EXPERIMENTS

1.Data preprocessing

Right after loading every dataset, we first modify the data size by cutting some parts of it to make sure we can divide it into 3 parts. Then, we do 3 trials. Then each trial, we randomly shuffle the entire dataset. If the dataset contains nominal columns, we apply one-hot-coding method over those columns to make categorical variables into numerical variables.

2.Training set and Testing set

Each data set is evaluated in 3 partition(table 1):

- 20% training set + 80% testing set
- 50% training set + 50% testing set
- 80% training set + 20% training set

3. Measures of classifiers' performance

In our research, we chose to use accuracy determined by percentages to measure the performance and effectiveness of our algorithms. Then we report the testing accuracy controlled by **Cross-Validation** of 5 folds for both the training and testing data set. For each dataset, we run each partition 3 times for all classifiers (Random forest, SVM RBF, Logistic regression). Then compute the average score to avoid potential outlier results. Then, the best hyperparameters for each dataset and classifier are reported.

4.Performance of classifiers

(See table 2 for the average performance for all classifiers over 3 datasets)

As we can see, the Random Forest has the best performance for the car data. However, when Random Forest faces the large data, “adult”, it has the worst performance. The accuracy is around 0.84, which is the best over all three classfiter. On the other hand, SVM has the most stable performance. It has 0.79, 0.72, 0.53 for 2 datasets. Besides that, Logistic regression also gives great performance. It has 0.79, 0.72, 0.53 for three datasets. All in all, when faced with a large dataset, SVM and Logistic regression have better performance.

IV CONCLUSION

In this study, we are specifically interested in measuring the performance of 3 supervised learning algorithms, namely Random Forest, SVM, and Logistic Regression by using the latest SK-Learn package. Since the performance of an algorithm depends not only on the algorithm itself, but also on the data set, we measure the performance of these three algorithms on three completely different data sets. As we can see in the table 3. The SVM kernel and logistic regression have more stable performance. The average accuracy for logistic regression classifier for three datasets are around 0.72, 0.79, 0.53. The average accuracy for SVM kernel classifier for three datasets are around 0.72, 0.75, 0.54. The average accuracy for Random Forest classifier for three datasets are around 0.63, 0.84, 0.56.

Table 1: partition of train and test size

Problem	Train, Test size 1	Train, Test size 2	Train, Test size 3
Car	1360,340	850,850	340,1360
Adult	4000,1000	2500,2500	1000,4000
forestfires	400,100	250,250	100,400

Table 2: average accuracy for each learning algorithms over 3 partitions

Model	Train & Test size	Random forest(train/test)	SVM rbf(train/test)	Logistic regression
Car	1360,340	0.79/0.64	0.78/0.72	0.75/0.74
Car	850,850	0.85/0.62	0.77/0.72	0.76/0.72
Car	340,1360	0.92/0.64	0.78/0.71	0.73/0.71
Adult	4000,1000	1.0/0.83	0.92/0.74	0.80/0.78
Adult	2500,2500	1.0/0.85	0.91/0.76	0.79/0.80
Adult	1000,4000	1.0/0.83	0.91/0.76	0.79/0.79
Forestfires	400,100	0.99/0.58	0.94/0.58	0.70/0.54
Forestfires	250,250	0.99/0.56	0.81/0.54	0.91 /0.54
Forestfires	100,400	0.99/0.53	0.84/0.52	0.81/0.51

Table 3: Rank for classifiers

Classifier	1st	2nd	3rd
Random forest	0.84	0.63	0.56
SVM RBF	0.75	0.72	0.54
Logistic regression	0.79	0.72	0.53