An Empirical Comparison of Supervised Learning Algorithms

Jiaying He, University of California, San Diego

jiayingucsd@gmail.com

Abstract

Supervise learning methods are being rapidly developed and applied in recent decade. Caruana's research on empirical comparison of supervised learning algorithms inspires me to use some of the method that he used in his paper to examine his results and evaluate the learning algorithm. I use k nearest neighbor, decision tree and neural network to predict the dataset ADULT, COV_TYPE and WINE. I use accuracy as criteria to evaluate the learning methods.

1. Introduction

In this paper, I want to reproduce Caruana's research on empirical comparison on supervised learning algorithm in high dimension, and I also want to study the supervised learning algorithm in low dimension. I use ADULT, COVTYPE, and WINE dataset from UCI Machine Learning Dataset to examine Decision tree, K Nearest Neighborhood, and Neural Network by comparing their prediction accuracy. I randomly select 5000 as training data and the rest as testing data for ADULT and COVTYE dataset. Also, I use the same parameters along with Caruana's paper. The result is really similar to Caruana's research. Neural Network has the highest accuracy, followed by Decision tree, then KNN.

2. Methodology

2.1 Data sets

I compare three algorithms on three binary classification problems. ADULT, COV_TYPE, AND LETTER are from the UCI Repository (Blake & Merz, 1998). The strings such as education level and country in ADULT dataset has been converted to number. COV_TYPE has also been converted by treating the largest class as positive and the rest as negative (Caruana). Because the dataset is too huge, I randomly select 5000 as training set and use the rest as testing set. Table one describes the dataset.

DATA SET	#AFTER	TRAIN SIZE	TEST SIZE	%POZ
ADULT	14/105	5000	5000	100%
COV_TYPE	54	5000	576012	8%
WINE	13	53	127	41%

Table one: Description of problem

2.2 Learning Algorithms

The goal of this research is to find the optimal parameters for these three algorithms through a 5-fold cross-validation.

KNN

I use 3 values of K, ranging from 1 to 5 as parameters. I use KNN with Euclidean distance. Table two shows the different k value affect different dataset accuracy.

K value	1	3	5
ADULT	0.76	0.76	0.76
COV_TYPE	0.88	0.86	0.85
WINE	0.96	0.94	0.93

Table two

KNN works best for WINE especially when K = 1, followed by COV_TYPE when k =1, and then ADULT.

Decision Trees (DT):

I use cross validation to select the optimal leaf size (the maximum depth D) to build decision tree.

ANN: Neural networks are very good at pattern recognition problems, such as forest cover type and different categories of wine. So in this experiment, ANN in these two dataset performs better than Adult.

2.3 Performance Metric

I use accuracy as metric to evaluate the performance. The average accuracy of three algorithms in three dataset is compared. The neural network is the best performance method, followed by decision tree and k nearest neighbor.

3 Performances Comparison

Table three shows the normalized score for each algorithm on each of the 3 problems. Each entry is an average over accuracy. KNN with Euclidean distance works fairly well for binary problems, the accuracy of KNN slightly below decision tree. Decision tree works pretty stably among all three problems. Neural Network works the best in each problem set. Table three shows the accuracy of each learning algorithm on each problem set.

MODEL	ADULT	COVTYPE	WINE	MEAN
KNN	0.76	0.74	0.94	0.81
DT	0.83	0.86	0.84	0.84
ANN	0.87	0.86	0.93	0.89

Table three: normalized score for each algorithm

4. Conclusions

Due to the metric, the accuracy, ANN ranks the best model, Decision Tree is the second, and KNN is the third. According to No Free Lunch Theorem, there is no "best" algorithm (Caruana 5). The best models in average perform weakly in some problem, and the models that perform weakly in general perform well on some problem. For instance, KNN works poorly compare to the other two, but it works best in WINE dataset. Neural network performs best in average, but in COVTYPE, decision tree works as well as ANN.

The comparison might not as accuracy and fair as reference paper due to the selection of dataset. Both ADULT and COVTYPE are high dimensional dataset, but WINE is a low dimension dataset. Use another low dimensional dataset to do a 2 to 2 comparison instead of 2 to 1 will be better for this study. And use other metrics such as F-values will be more accurate on comparison. Also, it might be benficial to use the Bootstrap Analysis that Caruana did in the reference paper due to the fact that reduction of specificity of data sets.

5. References

Caruana, R., & Niculescu-Mizil, A. (2006). *An empirical comparison of supervised learning algorithms*. ICML '06, 161–168.