Extremely randomized trees, a Better model in Binary classification

YAN Ning,Student ID:20441822

***Abstract*—This article introduce an application of Extremely randomized trees in binary classification problem and compare the 10-CV mean error rate of a number of classifiers(Logistic Regression, Neroual Network, Adaboost, Decision Tree, Random Forest, Extremely randomized trees) on a binary classification data set.**

***Keywords*—Supervised learning, Extremely Randomized Trees, Binary classification, Cross Validation.**

I. Introduction

Classification method aim to identify the classes of object from its descriptive features. The aim is to design a classifier to do classify more accurate than human or approximate human. Decision tree is a very effective supervised learning model to solve classification problems. In balance dataset, decision tree's accuracy always good. Random Forest, a strong classifier, integrates several sub decision trees and decide the class by bagging.It's a very effective model with low error rate by correct the habit(overfitting the training set) of decision tree.

Extremely randomized trees method is similar to Random Forest but there still two main difference. First, Random forest randomly choose sub-features or sub-samples to get each decision tree, and Extremely randomized trees uses every training sample to get each decision tree, that is, each decision tree applies the same whole training sample. Second, Random forests get the best bifurcation properties in a random subset, while Extremely randomized trees are completely randomized to obtain bifurcation values in order to bifurcate the decision tree.

A. Extremely Randomized Trees Algorithm

The Extremely randomized trees algorithm builds a set of unprocessed decision or regression trees based on the traditional process from top to down.(In this article we only concern decision trees.) Table 1. shows how Extremely Randomized Tree build trees.

B. Cross Validation

K fold Cross Validation is a statistic method randomly split samples into K sets with the same number of samples. Each time choose (K-1) sets as training data and the remained one as test data. Comparing the predict result and its' label, we could compute the error rate. 10 fold cross validation is the most common one.

Split a node(S)
Input: the local learning subset S corresponding to the node we want to split Output: a split $[a < a_c]$ or nothing If Stop split(S) is TRUE then return nothing. Otherwise select K attributes $a_1, ..., a_K$ among all non constant (in S) candidate attributes; $DrawKsplitss_1, ..., s_K$,where $s_i = Pick$ a random $split(S, a_i)$,$i = 1, ..., K$; Return a split s such that $Score(s, S) = max_i = 1, ..., K\ Score(s_i, S)$.
Pick a random $split(S, a)$
Inputs: a subset S and an attribute a Output: a split Let a^S_{max} and a^S_{min} in denote the maximal and minimal value of a in S; Draw a random cut-point ac uniformly in $[a^S_{max}, a^S_{min}]$; Return the split $[a < a_c]$.
Stop split(S)
Input: a subset S
Output: a boolean If $\lvert S \rvert < n_min$, then return TRUE; If all attributes are constant in S, then return TRUE; If the output is constant in S, then return TRUE; Otherwise, return FALSE.

TABLE I. Extra-Trees splitting algorithm.

C. data set Introduction

The dataset concludes 52 features and binary labels $L_i = 0$ or $L_i = 1$. There are 39.53% labels belongs to 1, and 60.57% belongs to 0. It seems balance. So we could choose binary classifiers rather than One class classifiers.

II. Experiments

We classify data set by LogisticRegression, Neroual Network, Adaboost, Decision Tree, RandomForest, Extremely randomized trees, and use 10-CV to measure the accuracy of classifiers.

A. Parameter choices of Extra-Trees

- $n\_estimators = 100$, build 100 weak classifiers as component
- $max\_depth = 52$,
- $random\_state = 0$

III. Conclusion

As the table two shown, Extremely Randomized Trees is the best classifier of this data set.

Classifier	Accuracy	classifier	Accuracy
Decision Tree	90.83%	Random Forest	93.97%
ExtraTrees	95.40%	Neural Network	91.99%
Adaboost	94.16%	Logistic Regression	92.61%

TABLE II. 10-CV Mean Accuracy of Classifiers.

Thanks:Yu Zhang, the teacher of 6000B