Assignment Report

Implementation of 3 supervised classification algorithms

Praneeth Varma
940423-9213
prbh16@student.bth.se
Blekinge institute of technology

I. Introduction

In this report, I try to represent three supervised classification algorithms which I have considered to implement on the spam dataset are Decision tree, Random forest, and logistic regression. I have calculated the training time, accuracy, and F measures of all the algorithms. The accuracy and F measures are being represented in the form of tables. The training time for each algorithm is also given after it is being implemented in R. I also represent the values of Friedman test and nemeyi test.

II. Implementation

Initially I download the dataset from the given link and I set the path for the dataset to be imported to the R environment. Once the dataset is being imported. I try to import random Forest and caret from the library if they are not available in R library I download the necessary packages like install. Packages("caret"), install. Packages("randomForest") for the implementation of the algorithms. Now I write the code so that the algorithms are run on stratified ten-fold cross-validation tests and they are respectively showed in the console window along with the training times. In the code the stratified ten-fold cross validation measurements represent Decision tree algorithm, Random forest and logistic regression respectively. Random forest takes a bit of time so we must wait for the measurements to be printed on the console window. It takes around 10 to 15 seconds for training. I represent the results in a tabular form and accordingly calculate the F measure by using the formula.

F = 2(precision)(recall)/precision+recall
Precision = tp/tp+fp
Recall = tp/tp+fn
Accuracy = tp + tn/tp+tn+fp+fn
In the above formula
tp - true positive (non spam ,nonspam)
tn - true negative (non spam ,spam)
fp – false positive (spam , nonspam)
fn – false negative(spam , spam)

After representing them in the table we calculate the average mean and standard deviation. Then I perform the Friedman test and I calculate the Friedman statistic. In my case the Friedman statistic is 18.2 which is greater than chi-square value of 5.991. I reject the null hypothesis(H0) because my value is greater than chi-square value. As I reject the null hypothesis I calculate the critical difference by using the nemeyi test. The qo value for 3 classifies for alpha=0.05 is 2.343. Now substituting k = 3 in the critical difference formula and there are 3 classifiers we get a critical difference of 1.04782.

III. Results

Fold	Decision tree	Random forest	Logistic regression
1.	0.8673913	0.9521739	0.9370933
2.	0.8627451	0.9457701	0.9237473
3.	0.8391304	0.9543478	0.9434783
4.	0.8500000	0.9521739	0.9370933
5.	0.8785249	0.9521739	0.9152174
6.	0.8586957	0.9608696	0.9282609
7.	0.8652174	0.9500000	0.9065217
8.	0.8627451	0.9152174	0.9391304
9.	0.8416486	0.9544469	0.9000000
10.	0.8741866	0.9586057	0.9413043
Average	0.8600285	0.9495833	0.9271846
Standard Deviation	0.1298	0.01278	0.01537

F Measure			
Fold	**Decision tree**	**Random Forest**	**Logistic regression**
1	0.74173	0.73121	0.74282
2	0.74209	0.74672	0.74354
3	0.73489	0.79612	0.73751
4	0.73833	0.74341	0.74198
5	0.74425	0.74707	0.74209
6	0.75035	0.75414	0.75312
7	0.73266	0.74162	0.74017
8	0.74302	0.75428	0.75190
9	0.74803	0.75810	0.75709
10	0.74474	0.74722	0.74433

References:

[1] P. Flach, Machine learning: The art and science of algorithms that make sense of data. Cambridge: Cambridge university press, 2012.