Assignment 2 Report

Tanveer Feisal Snigdho
MSc in Computer Science
Blekinge Institute of Technology
Karlskrona, Sweden
t.f.snigdho@gmail.com

I. Introduction

This report contains a brief description of assignment 2, the objectives, dataset and the procedures.

II. Objective

The objectives of the assigment are as following:

- To implement three supervised classification learning algorithms. Here, I choose: Support Vector Machine (SVM), K-Nearest Neighbors (KNN) and Decision Tree (DTree)
- Compare their computational performance with training time.
- Compare their predictive performance with accuracy and F-measure.
- To perform Friedman test to check if there is any significant difference. If so, then perform Nemenyi test to check the critical difference amont those algorithms.

III. Dataset

The spambase dataset was used for this assignment. The source of the dataset is available in the assignment instructions manual [1]. According to this dataset, each row contains data of one email. The last column indicates if it is a spam or not. The first 57 columns represents different value of specific properties.

IV. Procedures

First, I loaded the data. I iterated through all the lines and split all the values with a comma. Then shuffled the data as it was necessary for this assignment.

Then I converted all the string values into float. After that, I separated each row of data from labels. The first 57 columns were saved in variable X and the last column which is the label was saved in variable Y.

I wrote separate functions for implementing SVM, KNN and DTree. I used libraries from sklearn to implement those algorithms.

I also calculated how much time were taken for computing everything about each of those three algorithms.

I used my own custom display_scores function to display the values of each algorithm.

After implementing each algorithm, I used cross_val_score method with five metrics e.g., accuracy, recall, precision, f-measure and training time. I used all those metrics to implement 10-Fold cross validation by sklearn library.

Then I showed comparison of 'Accuracy' among all of the algorithms. Right After that, I called the friedman_test function. In this function, I implemented the Friedman test. First I printed the ranked values. Then I used the following formula to calculate the Friedman statistic (Fr) value.

$$Fr = \frac{12}{nk\,(k+1)} (T_1^2 + T_2^2 + T_3^2 + \ldots T_n^2) - 3n\,(k+1)$$

Here, n = number of cases which is 10 in all the cases.

k = number of algorithms which is 3 in all the cases.

T_i is the sum of ranks of a particular algorithm.

After calculating Fr, I compared it with the corresponding critical value. The critical value for k = 3 and n = 10 at the α = 0.05 level is 7.8 (from the Chi-square table). In all the cases in this assignment, Fr values were greater than 7.8. So the null hypothesis was rejected all the times.

To determine the critical differences between the algorithms, I need to determine the critical difference first. The critical difference was found from the following formula:

$$CD = q_a \sqrt{(k\,(k+1) / 6n)}$$

Here, q_a is 2.343 and hence the CD = 1.0478

If the difference between mean of rank of two algorithms is more than CD then the significant different is found here. This is how I calculated the significant differences.

V. References

[1] V. Boeva, "ML-Instructions."