Course DV2542: Assignment-2

Venkata Satya Sai Ajay Daliparthi
Department of Computer Science
BTH
Karlskrnona, Sweden
veda18@studen.bth.se

I. Introduction

Aim of this study is to experimentally compare the computational and predictive performance of three classification algorithms on an email spam detection task.

II. Dataset

Spambase Data Set used for this study is a part of UCI machine learning repository which contains 4601 instances with 57 attributes. There are 2 classes present in target attribute. They are spam(1) and ham(0).[2].

III. Importing Data and libraries

Python programming language is chosen to implement this task. Various python standard libraries like pandas, numpy, time, math, sklearn are imported along with the dataset.

IV. ALGORITHMS

The dataset is loaded and shuffled. Three classification algorithms Logistic Regression, Support Vector Machines and Decision Trees are imported by using using sklearn module.

V. VALIDATION

Stratified ten-fold validation is used to divide dataset into ten equal parts. Each one of the three classification algorithms is applied on each fold. Then the Metrics like accuracy, recall, precision and F-measure are computed for each algorithm. Then largest value in the each K-fold value is assigned Rank based on the higest value(given 1).

VI. ResultS

TABLE I. Accuracy

Fold	Algorithms		
	Logistic Regression	*SVM*	*Decision Tree*
1	0.9371(1)	0.9328(2)	0.9241(3)
2	0.9046(2)	0.9132(1)	0.9196(1)
3	0.9111(3)	0.9154(2)	0.9176(1)
4	0.9109(2)	0.9130(1)	0.8935(3)
5	0.9522(2)	0.9587(1)	0.9174(3)
6	0.9174(3)	0.9217(2)	0.9261(1)
7	0.9435(2)	0.9543(1)	0.9261(3)
8	0.9326(1)	0.9239(2)	0.9000(3)
9	0.9412(1)	0.9390(2)	0.9150(3)
10	0.9303(2)	0.9325(1)	0.9194(3)
Avg rank	1.9	1.5	2.6
avg	0.9281	0.9305	0.9133
stdev	0.0153	0.0155	0.0121

TABLE II. F-Measure

Fold	Algorithms		
	Logistic Regression	*SVM*	*Decision Tree*
1	0.9171(1)	0.9141(2)	0.9072(3)
2	0.8778(2)	0.8889(1)	0.8693(3)
3	0.8852(3)	0.8920 (2)	0.8944(1)
4	0.8832(2)	0.8870 (1)	0.8658(3)
5	0.9382(2)	0.9465 (1)	0.8950(3)
6	0.8902(3)	0.8966 (2)	0.9050(1)
7	0.9282(2)	0.9421 (1)	0.9056(3)
8	0.9132(1)	0.9025 (2)	0.8708(3)
9	0.9244(1)	0.9218 (2)	0.8926(3)
10	0.9111(2)	0.9146(1)	0.8992(3)
Avg rank	1.9	1.5	2.6
avg	0.9071	0.9106	0.8905
stdev	0.0203	0.0202	0.0151

TABLE III. Training Time

Fold	Algorithms		
	Logistic Regression	*SVM*	*Decision Tree*
1	0.1317(2)	151.3594(3)	0.0533(1)
2	0.0491(1)	294.3539(2)	0.0528(3)
3	0.0870(2)	175.3428(3)	0.0477(1)
4	0.0474(1)	303.0067(2)	0.0552(3)
5	0.0573(2)	213.2388(3)	0.0556(1)
6	0.0730(2)	144.3701(3)	0.0524(1)
7	0.0722(2)	604.8594(3)	0.0517(1)
8	0.0781(2)	153.9641(3)	0.0548(1)
9	0.1022(2)	162.3246(3)	0.0505(1)
10	0.0360(1)	221.0076(2)	0.0459(3)
Avg rank	1.7	2.7	1.3
avg	0.0734	242.3827	0.0520
stdev	0.0271	132.5463	0.0030

VII. Friedman test

Null Hypothesis: All the algorithms perform similarly. Friedman test is done. The critical value for $k = 3$ and $n = 10$ at the $\alpha = 0.05$ level is 7.8 .The Friedman Statistic values for accuracy, f-measure (f1 score) and training time are 6.20, 6.20 and 15.18 respectively. Therefore, null hypothesis is rejected for Testing time only.

VIII. Nemenyi test

Nemenyi test is performed using the formula provided in the textbook [1]. Q-alpha value is taken 2.343. The critical difference is 1.047.

Accuracy- ***All algorithms perform similarly***

*F-measure****: All algorithms perform similarly***

Training Time: ***SVM and Decision Tree exceeds critical difference.***

References

[1] P. Flach, *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press, 2012.

[2] D. Dua and C. Graff, *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences, 2017.