Machine learning assignment 2

Harshal Patel hapt18@student.bth.se

Department of computer science

Blekinge institute of technology,

Karlskrona, Sweden

INTRODUCTION

This report aims to present the implementation of three different algorithms, namely, Support vector machines, k-nearest neighbor and random forest. For each algorithm, accuracy, recall, precision, f-measure, average rank and computation time is calculated and presented in a tabular format. Stratified 10-fold cross validation with Friedman test and Nemenyi test is performed.

Mentioned algorithms are traditional algorithms in machine learning and are convenient to use, compare and gain insight about their performances. All three algorithms have their unique idea to find the accuracy of classification. SVM uses support vectors to draw hyperplane which can be used for classification. K-NN use 'k' number of nearest neighbors to classify new data points, whereas, random forest is a collection of different decision trees combined to classify new data. These algorithms possess high difference in implementation, hence makes it interesting to compare their accuracy and f-measure.

DATA

Spambase dataset used in this experiment consists of 4600 Instances and 58 attributes. Features from 0 to 57 contains the information about the mail and 58th feature is the class to which it belongs (0-not spam, 1-spam). Stratified 10-fold cross validation is used to split the data.

METHOD

Firstly, data loaded is converted into data frames using python library. These data frames contain features and class labels separately. Data is then divided into training and testing (stratified 10-fold cross validation). 10% of data is used for testing in each fold. Secondly, this data is used in all the three mentioned algorithms along with calculating training time for each algorithm. Accuracy and f-measure is calculated using confusion matrix. Thirdly, Friedman test and Nemenyi test is conducted for each algorithm using accuracy, f-measure and training time.

Next, average rank for each algorithm is calculated by adding the rank of each fold and dividing by number of folds. Sum of squared differences (SD) is calculated, if SD<7.8, null hypothesis is accepted by, otherwise, we reject it by printing all three algorithms perform differently.

Critical difference (CD) is calculated for Nemenyi to figure out the pair of algorithms performing differently.

$$CD = qalpha * \sqrt[2]{((k * (k + 1))/(6 * n))}$$

Where qalpha = 2.343, k = 3, n = 10.

Mean rank of each algorithm is then compared with CD to find the algorithms performing differently and exceeds the critical difference.

RESULTS

Results obtained, are shown below in a tabular form for SVM, K-NN and random forest. Null hypothesis is rejected as all the three algorithms perform in a different manner.

```
FriedmenTestandNemeyiTest(Accuracy,0)
```

Fold	SupportVectorMachine	KNN	Random Forest
1	0.66 (3)	0.74 (2)	0.95 (1)
2	0.65 (3)	0.77 (2)	0.95 (1)
3	0.73 (3)	0.78 (2)	0.94 (1)
4	0.75 (3)	0.82 (2)	0.95 (1)
5	0.70 (3)	0.81 (2)	0.96 (1)
6	0.68 (3)	0.82 (2)	0.96 (1)
7	0.76 (3)	0.82 (2)	0.97 (1)
8	0.71 (3)	0.83 (2)	0.97 (1)
9	0.75 (2)	0.72 (3)	0.89 (1)
10	0.71 (3)	0.76 (2)	0.86 (1)

```
avearge rank  2.9        2.1         1.0
reject null hypothesis all three algorithms perform differently
Critical difference for Nemenyi test is  1.0478214542564015
SVM and random forest perform algorithms differently exceeds critical difference
Knn and random forest perform algorithms differently exceed critical difference
```

Fig 1. Shows Friedman and Nemenyi test for accuracy of three algorithms, average rank and critical difference.

```
In [42]: FriedmenTestandNemeyiTest(Fmeasure,0)
```

Fold	SupportVectorMachine	KNN	Random Forest
1	0.48 (3)	0.66 (2)	0.94 (1)
2	0.48 (3)	0.71 (2)	0.93 (1)
3	0.56 (3)	0.71 (2)	0.92 (1)
4	0.63 (3)	0.76 (2)	0.94 (1)
5	0.53 (3)	0.76 (2)	0.94 (1)
6	0.52 (3)	0.78 (2)	0.95 (1)
7	0.62 (3)	0.75 (2)	0.96 (1)
8	0.55 (3)	0.78 (2)	0.96 (1)
9	0.59 (3)	0.67 (2)	0.87 (1)
10	0.54 (3)	0.70 (2)	0.82 (1)

```
avearge rank  3.0        2.0         1.0
reject null hypothesis all three algorithms perform differently
Critical difference for Nemenyi test is  1.0478214542564015
SVM and random forest perform algorithms differently exceeds critical difference
```

Fig. 2 Shows Friedman and Nemenyi test for F-measure of three algorithms, average rank and critical difference.

```
In [43]: FriedmenTestandNemeyiTest(Computationaltrainingtime,1)
```

Fold	SupportVectorMachine	KNN	Random Forest
1	2.03 (3)	0.09 (1)	0.90 (2)
2	1.82 (3)	0.08 (1)	0.81 (2)
3	1.75 (3)	0.07 (1)	0.81 (2)
4	1.78 (3)	0.07 (1)	0.81 (2)
5	1.81 (3)	0.07 (1)	0.81 (2)
6	1.83 (3)	0.07 (1)	0.86 (2)
7	1.82 (3)	0.07 (1)	0.85 (2)
8	1.83 (3)	0.08 (1)	0.87 (2)
9	1.80 (3)	0.07 (1)	0.80 (2)
10	1.81 (3)	0.07 (1)	0.81 (2)

```
avearge rank  3.0        1.0         2.0
reject null hypothesis all three algorithms perform differently
Critical difference for Nemenyi test is  1.0478214542564015
Knn and SVM perform algorithms differently exceeds critical difference
```

Fig. 3 Shows Friedman and Nemenyi test for computation time of three algorithms, average rank and critical difference.

CONCLUSION

Algorithms like support vector machines, K-nearest neighbor and random forest are implemented. Also, Friedman test and Nemenyi test are performed and compared based on accuracy, f-measure and training time

of algorithms. Mean rank is calculated for each algorithm. Null hypothesis is rejected since all the algorithms perform differently.

REFERENCE

Flach, P., 2012. Machine learning: the art and science of algorithms that make sense of data. Cambridge University Press.