CIS6930: Introduction to Data Mining

Project 1: Classification

Yash Sinha
UFID: 15618171

1. Dataset preparation

This classification project implements the following four classification algorithm on the Life Expectancy dataset from Wikipedia ("List by the CIA(2016)").

a. K-NN (K-Nearest Neighbour)
b. SVM (Support Vector Machine)
c. C4.5 (Decision Trees)
d. RIPPER (Decision Trees)

Initial dataset preparation included adding the Continent column in the table which includes the continent values for the countries. This acts as the class labels which is used for prediction. I have also removed the 'Country' column because it isn't useful for training the model. I then searched for null values in the dataframe and then removed them. However, there are no null values in our dataset.

I have used different seeds to randomly shuffle the rows and partition it in the raio of 80% and 20% for creating training and testing set respectively. I have used the parameter "repeatedcv" for cross-validation in training of K-NN and SVM. I have also used the parameters "center" and "scale" for preprocess argument in K-NN and SVM for centering and scaling of the data. For K-NN, I have used the parameter 'number' as 10 in K-NN for trControl which tries 10 ks and selects the best k, based on the accuracy obtained. An example of the output is:

```
k   Accuracy   Kappa
 5  0.5042541  0.3852745
 7  0.5182433  0.4010550
 9  0.5333511  0.4160999
11  0.5205406  0.3981553
13  0.5080358  0.3825816
15  0.5284193  0.4042553
17  0.5277566  0.4026683
19  0.5397925  0.4165485
21  0.5471121  0.4251843
23  0.5341262  0.4075009
```

So, in the above example, the best k chosen is 21.

Finally, this step is repeated for 5 times to get 5 different random samples. The accuracies are then appended to a list and finally the average accuracy is calculated for the four models.

2. Description of classification methods

Four different classification methods have been used:

a. K-NN
   K nearest neighbors is a simple classification algorithm that trains the model based on the distance of one data point from the other. One of the popular distance measures is Euclidian distance. A data point is classified by most common amongst its K nearest neighbors measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.

   Caret package in R provides the implementation for K-NN using the method "knn" which also takes other arguments like preprocess and trControl. The "number" argument in trControl specifies the number of Ks to try.

b. SVM
   A Support Vector Machine (SVM) performs classification by finding the hyperplane that maximizes the margin between the two classes. The vectors (cases) that define the hyperplane are the support vectors.

   Caret package in R provides the implementation for SVM using the method "svmLinear" which also takes other arguments like preprocess and trControl.

c. C4.5
   A classification tree is a model that predicts the class label of data items. The tree is built by repeatedly dividing the data into groups based on attribute values. The attribute on which to divide is selected by information gain, a statistical technique for determining which attribute split will most cleanly divide the data.

   Caret package in R provides the implementation for C4.5 using the method "J48".

d. RIPPER
   RIPPER is based in association rules with reduced error pruning (REP), a very common and effective technique found in decision tree algorithms. In REP for rules algorithms, the training data is split into a growing set and a pruning set.

First, an initial rule set is formed that over ts the growing set, using some heuristic method. This overlarge rule set is then repeatedly simplified by applying one of a set of pruning operators typical pruning operators would be to delete any single condition or any single rule. At each stage of simplification, the pruning operator chosen is the one that yields the greatest reduction of error on the pruning set. Simplification ends when applying any pruning operator would increase error on the pruning set.

Caret package in R provides the implementation for RIPPER using the method "JRip".

3. Classification analysis and results

I have created 5 different training and testing sets. The following results correspond to one of the iterations for each of the models. The formulaes are:

Accuracy: (TP + TN)/ (TP + TN + FP + FN)
Precision: (TP / TP + FP)
Recall: (TP / TP + FN)
F Measure: 2 * ((precision * recall) / (precision + recall))

a. K-NN

Accuracies for different K values:

k	Accuracy	Kappa
5	0.5090339	0.3885995
7	0.5010615	0.3773276
9	0.5025893	0.3756336
11	0.5088693	0.3837239
13	0.5047718	0.3797857
15	0.5349064	0.4159755
17	0.5235138	0.4009133
19	0.5278155	0.4024767
21	0.5244042	0.3964720
23	0.5222119	0.3939527

```
[1] "Confusion Matrix"
               Predicted
Actual          Africa Asia Europe North America Oceania South America
  Africa             9    2      0             0       0             0
  Asia               1    5      1             1       0             1
  Europe             0    2      6             2       0             0
  North America      0    2      2             1       0             1
  Oceania            0    2      0             1       1             0
  South America      0    1      1             0       0             0
```

```
               precision    recall        f1
Africa         0.8181818 0.9000000 0.8571429
Asia           0.5555556 0.3571429 0.4347826
Europe         0.6000000 0.6000000 0.6000000
North America  0.1666667 0.2000000 0.1818182
Oceania        0.2500000 1.0000000 0.4000000
South America  0.0000000 0.0000000       NaN
```

"Average accuracy for KNN: 0.538095238095238"

b. SVM

```
[1] "Confusion Matrix"
               Predicted
Actual          Africa Asia Europe North America Oceania South America
  Africa             9    1      1             0       0             0
  Asia               1    6      2             0       0             0
  Europe             0    3      7             0       0             0
  North America      0    2      4             0       0             0
  Oceania            0    4      0             0       0             0
  South America      0    1      1             0       0             0
```

```
               precision    recall        f1
Africa         0.8181818 0.9000000 0.8571429
Asia           0.6666667 0.3529412 0.4615385
Europe         0.7000000 0.4666667 0.5600000
North America  0.0000000       NaN       NaN
Oceania        0.0000000       NaN       NaN
South America  0.0000000       NaN       NaN
```

"Average accuracy for SVM: 0.561904761904762"

c. C4.5

```
[1] "Confusion Matrix"
               Predicted
Actual          Africa Asia Europe North America Oceania South America
  Africa             8    1      2             0       0             0
  Asia               1    5      3             0       0             0
  Europe             0    0     10             0       0             0
  North America      0    2      4             0       0             0
  Oceania            0    2      2             0       0             0
  South America      0    1      1             0       0             0

               precision    recall    f1
 Africa        0.7272727 0.8888889 0.800
 Asia          0.5555556 0.4545455 0.500
 Europe        1.0000000 0.4545455 0.625
 North America 0.0000000       NaN   NaN
 Oceania       0.0000000       NaN   NaN
 South America 0.0000000       NaN   NaN
```

"Average accuracy for C45: 0.59047619047619"

d. RIPPER

```
[1] "Confusion Matrix"
               Predicted
Actual          Africa Asia Europe North America Oceania South Americ
  Africa            10    0      1             0       0
  Asia               4    4      1             0       0
  Europe             2    0      8             0       0
  North America      2    1      3             0       0
  Oceania            2    2      0             0       0
  South America      1    1      0             0       0

              precision    recall        f1
Africa        0.9090909 0.4761905 0.6250000
Asia          0.4444444 0.5000000 0.4705882
Europe        0.8000000 0.6153846 0.6956522
North America 0.0000000       NaN       NaN
Oceania       0.0000000       NaN       NaN
South America 0.0000000       NaN       NaN
```

"Average accuracy for RIPPER: 0.504761904761905"

4. Conclusion

From the accuracy measures for the different classification algorithms, it can be observed that C4.5 gives the best results. SVM comes in a close second. RIPPER decision tree model is the least accurate among the four algorithms. Regarding the performance, I observed that KNN and SVM are pretty fast. The decision trees

algorithms, RIPPER and C4.5 are slow. Also, the classes Oceania and South America were often misclassified. This might be due to class imbalance as not many entries correspond to these classes in the training set.

5. References

- https://en.wikipedia.org/wiki/List_of_countries_by_life_expectancy
- https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Classification/JRip
- https://www.r-bloggers.com/computing-classification-evaluation-metrics-in-r/
- https://rpubs.com/kjmazidi/195428
- https://rpubs.com/njvijay/16444
- http://dataaspirant.com/2017/01/19/support-vector-machine-classifier-implementation-r-caret-package/