CIS6930 Fall 2017: Introduction to Data Mining
Project I: Classification

By Jay Shah (1461-3930)

10/8/2017

1 Description of Dataset Preparation

Life Expectancy Dataset is obtained from Wikipedia which contains five columns. It contains the life expectancy of each country. Now, we are classifying data based on the continents. Continent dataset is obtained from Wikipedia which contains country and its continent.

With the use of both dataset new dataset is created. It has country, life expectancy and the continent in which country belongs to. Below is the sample of combined data.

Rank	Country	Overall Life	Male Life	Female Life	Continent
1	Monaco	89.5	85.6	93.5	Europe
2	Japan	85	81.7	88.5	Asia
3	Singapore	85	82.3	87.8	Asia
4	Macau; China	84.5	81.6	87.6	Asia
5	San Marino	83.3	80.7	86.1	Europe
6	Iceland	83	80.9	85.3	Europe
7	Hong Kong, China	82.9	80.3	85.8	Asia
8	Andorra	82.8	80.6	85.1	Europe
9	Switzerland	82.6	80.3	85	Europe
10	Guernsey	82.5	79.9	85.4	Europe
11	Israel	82.4	80.6	84.4	Asia
12	South Korea	82.4	79.3	85.8	Asia
13	Luxembourg	82.3	79.8	84.9	Europe
14	Australia	82.2	79.8	84.8	Oceania
15	Italy	82.2	79.6	85	Europe
16	Sweden	82.1	80.2	84.1	Europe
17	Canada	81.9	79.2	84.6	North America
18	Jersey	81.9	79.4	84.5	Europe
19	Liechtenstein	81.9	79.7	84.6	Europe
20	France, metropolitan	81.8	78.7	85.1	Europe
21	Norway	81.8	79.8	83.9	Europe
22	Spain	81.7	78.7	84.9	Europe
23	Austria	81.5	78.9	84.3	Europe
24	Anguilla	81.4	78.8	84.1	North America
25	Bermuda	81.3	78.1	84.5	North America
26	Netherlands	81.3	79.2	83.6	Europe
27	Cayman Islands	81.2	78.5	84	North America
28	Isle of Man	81.2	79.5	83	Europe
29	New Zealand	81.2	79.1	83.3	Oceania

Continent is used as a class label. We don't need country. We will consider the numerical features of the Dataset only. Rank is just the index showing highest to lowest life expectancy of each country. That is why rank is not considered as a feature for classification of the Dataset.

Using the sample function in R dataset is divided randomly in two parts. One part is 80% of dataset which is used to build/train model and other 20% data is used to test the model.

2 Description of the Classification Methods.

There are four classification algorithms we will be using in this project. Each algorithm and its parameters are described as follow.

1. **k-Nearest Neighbor (kNN)**
   KNN is lazy learner. It takes 4 parameters in the argument.
   - Training Data: 80% of the data that was divided earlier is used as training data.
   - Test Data: 20% of the data that was remaining is used as test data.
   - Classification Label: Here we use Continent (4^{th} column) as a classification label for KNN.
   - Value of K: Generally, we choose square root of total number of observations to get the efficient result.

2. **C4.5**
   C4.5 is a decision tree based algorithm which takes 3 parameters in the argument.
   - Name of the row which we want to classify (Continents in our case).
   - Control: It takes 3 more parameters in the argument.
     - C Pruning Confidence: It is the upper bound on the error rate at leaf node.
     - R: It stands for the Reduced Error pruning which is FALSE by default.
     - M: It sets the minimum number of instances at a leaf node.

     This are tuning parameters which are used to get the maximum efficiency of an algorithm.
   - Data: It takes training data as an input to build the model.

   After implementing this function, prediction function is used to predict the output of test data.
   Predict function takes C4.5 model and test data as an input to generate the output.

3. **Ripper**
   Ripper is a decision tree based algorithm which takes 3 parameters in the argument.
   - Name of the row which we want to classify (Continents in our case).
   - Control: It takes 3 parameters in the arguments,
     - O: Number of runs require for the optimization. Default value is 2. We set it 0.
     - F: Number of folds where one-fold is used for pruning and rest for growing rules.
     - N: It sets the minimum weight of the instances in a rule.

     This are tuning parameters which are used to get the maximum efficiency of an algorithm.
   - Data: It takes training data as an input to build the model.

   After implementing this function, prediction function is used to predict the output of test data.
   Predict function takes Ripper model and test data as an input to generate the output.

4. **Support Vector Machine (SVM)**
   Support Vector Machine is supervised learning algorithm which takes 6 parameters in the argument.
   - Name of the row which we want to classify (Continents in our case).
   - Data: It takes training data as an input to build the model.
   - Type of Classification we need to do in SVM method.
   - Type of Kernel (Linear, Polynomial etc.) I used linear kernel for our classification algorithm.
   - Cost: It specifies the upper bound of a cost of constraints violation.
   - Scale: Indicates if the variables to be scaled or not.

   We can also specify other parameters like gamma (needed for all kernel except linear), degree (needed for polynomial kernel), etc.
   After implementing this function, prediction function is used to predict the output of test data. Predict function takes SVM model and test data as an input to generate the output.

3 Classification and Result Analysis

Dataset is divided 5 times randomly and then each algorithm is applied on the dataset. Thus, for each algorithm we get 5 different results depending on the partition of the dataset. Below is the result of the dataset for each algorithm. There are four types of classification algorithms used in this project. Each algorithm and its parameters are described as follow.

1. k-Nearest Neighbor (kNN)

	1st Dataset	2nd Dataset	3rd Dataset	4th Dataset	5th Dataset	Average
Accuracy	56.41	61.9	64.86	59.52	62.79	61.096

- Best Accuracy: 64.86%
- Below is the classification table I got for best accuracy of KNN algorithm.

```
> confusionMatrix(Test_Prediction,TestData1[,4])       #Result and analysis of KNN
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa             8    1      0             0       1             0
  Asia               0    6      0             1       1             1
  Europe             1    0      9             2       0             0
  North America      1    0      0             0       1             0
  Oceania            0    1      0             2       0             0
  South America      0    0      0             0       0             1

Overall Statistics

               Accuracy : 0.6486
```

- Upper column shows the name of the continents. Left column of continent is the output classification result. e.g. There are total 10 continents of Africa in Testset out of which 8 are correctly classified and other 2 are incorrectly classified in Europe and North America.

2. C4.5

	1st Dataset	2nd Dataset	3rd Dataset	4th Dataset	5th Dataset	Average
Accuracy	53.85	52.38	59.46	52.38	53.49	54.312

- Best Accuracy: 59.46%
- Below is the classification table I got for best accuracy of C4.5 algorithm.

```
> confusionMatrix(p1,TestData1[,4])                  #Result and analysis of C45
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa             8    0      0             0       0             0
  Asia               1    5      2             1       0             1
  Europe             1    0      5             0       0             0
  North America      0    1      0             1       1             0
  Oceania            0    2      1             3       2             0
  South America      0    0      1             0       0             1

Overall Statistics

               Accuracy : 0.5946
```

- Upper column shows the name of the continents. Left column of continent is the output classification result.

e.g. There are total 10 continents of Africa in Testset out of which 8 are correctly classified and other 2 are incorrectly classified in Asia and Europe.

3. Ripper

Input	1st Dataset	2nd Dataset	3rd Dataset	4th Dataset	5th Dataset	Average
Accuracy	58.97	52.38	62.16	52.38	62.79	57.736

- Best Accuracy: 62.79%
- Below is the classification table I got for best accuracy of Ripper algorithm.

```
> confusionMatrix(p2,TestData1[,4])                #Result and analysis of Ripper
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa            13    0      1             1       0             0
  Asia               1    7      0             1       1             0
  Europe             1    1      6             3       2             2
  North America      0    1      0             1       0             0
  Oceania            0    0      0             1       0             0
  South America      0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.6279
```

- Upper column shows the name of the continents. Left column of continent is the output classification result. e.g. There are total 14 continents of Africa in Testset out of which 13 are correctly classified and other 2 are incorrectly classified in Europe and Asia.

4. Support Vector Machine

	1st Dataset	2nd Dataset	3rd Dataset	4th Dataset	5th Dataset	Average
Accuracy	69.23	64.29	59.46	64.29	55.81	62.616

- Best Accuracy: 69.23%
- Below is the classification table I got for best accuracy of SVM algorithm.

```
> confusionMatrix(p,TestData1[,4])                 #Result and analysis of SVM
Confusion Matrix and Statistics

               Reference
Prediction      Africa Asia Europe North America Oceania South America
  Africa             7    0      0             0       0             0
  Asia               1    6      1             0       0             1
  Europe             0    3     13             3       1             1
  North America      0    1      0             1       0             0
  Oceania            0    0      0             0       0             0
  South America      0    0      0             0       0             0

Overall Statistics

               Accuracy : 0.6923
```

- Upper column shows the name of the continents. Left column of continent is the output classification result. e.g. There are total 8 continents of Africa in Testset out of which 7 are correctly classified and other 1 is incorrectly classified in Asia.

4 Conclusion

1. Result of each algorithm highly depends on how we partition data. Good partition can lead us to efficient result and bad partition can lead us to very low efficiency.

2. Efficiency of Algorithm also depends on the tuning parameters we set. For example, K value in KNN algorithm, Confidence Interval value in C45 algorithm, Cost value in Support Vector Machine Algorithm, etc.

3. Among the classification algorithms implemented, Support Vector Machine yields the best result.

4. Below is the order of algorithm which is more efficient to less efficient as we goes from top to down.

    - Support Vector Machine
    - K Nearest Neighbor
    - Ripper
    - C4.5

5 References

1. https://cran.r-project.org/web/packages/RWeka/RWeka.pdf

2. https://www.analyticsvidhya.com/blog/2015/08/learning-concept-knn-algorithms-programming/

3. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

4. https://www.youtube.com/watch?v=JFJIQ0_2ijg&t=163s