Project 1: Classification

Raheen Mazgaonkar
UFID: 4714-4316
October 10, 2017

Dataset Preparation

The following data preprocessing methods were applied to the data:

1) Dimensionality Reduction
   Data mining algorithms work better on data of lower dimensions. Hence to improve accuracy it is important to ensure that there are no redundant parameters n the dataset. In the given dataset the parameter plays no role in obtaining continent from average age. Hence it removed before classification.

2) Random sampling for creation of training and testing set
   Random samples of training and testing sets were created from the given data in roughly 80:20 ratio. In order to ensure that each time different sets are created different seed values were used.

Several other preprocessing methods are present. However, they weren't applied as they didn't add any more value to the current dataset. For example, normalization wasn't applied as the given data is present in a specific range.

In addition to this, some functions in R require the predictors to be numeric hence we have factorized Continent column of the Dataset.

Classification Methods

A) KNN

- Method used: train
- Package: caret
- Optimization method: Used 10-fold cross validation repeated 5 times using trainControl() to obtain best value for k.

Results:

Seed Values	2016	9850	6230	4150	123	Average
Accuracy	0.5128	0.5581	0.5283	0.5	0.5952	0.5388

For seed: 2016

Class	Africa	Asia	Europe	North America	Oceania	South America
Precision	0.9166	0.5	0.4166	0.25	0	0
Recall	0.8461	0.3333	0.7142	0.25	0	0
F-Measure	0.88	0.4	0.5263	0.25	NaN	NaN

B) RIPPER

- Method used: train
- Package: caret
- Optimization method: Used 10-fold cross validation repeated 5 times using trainControl() to obtain best model. Preprocessing methods "center" and "scale" were also applied

Results:

Seed Values	2016	9850	6230	4150	123	Average
Accuracy	0.5897	0.5581	0.5	0.5789	0.5526	0.5558

For seed: 2016

Class	Africa	Asia	Europe	North America	Oceania	South America
Precision	0.9167	0.6	0.3181	NaN	NaN	NaN
Recall	0.8461	0.3333	1	0	0	0
F-Measure	0.88	0.4285	0.4827	NaN	NaN	NaN

C) C4.5

- Method used: J48
- Package: Rweka
- Optimization method: Set reduced error pruning option to TRUE and used 10-fold cross validation to obtain optimized fit.

Results:

Seed Values	2016	9850	6230	4150	123	Average
Accuracy	0.5641	0.6976	0.5625	0.5526	0.5952	0.5944

For seed: 2016

Class	Africa	Asia	Europe	North America	Oceania	South America
Precision	0.9166	0.625	0.3333	NaN	0	NaN
Recall	0.8461	0.5556	0.8571	0	0	0
F-Measure	0.88	0.588	0.48	NaN	NaN	NaN

D) Support Vector Machine

- Method used: svm()
- Package: e1071
- Optimization method: Used tune.svm to get optimal cost and gamma and recalculated model using this cost and gamma parameter

Results:

Seed Values	2016	9850	6230	4150	123	Average
Accuracy	0.5897	0.7209	0.5833	0.5789	0.5952	0.6136

For seed: 2016

Class	Africa	Asia	Europe	North America	Oceania	South America
Precision	0.9166	0.625	0.5	0.2	NaN	NaN
Recall	0.8461	0.5555	1	0.25	0	0
F-Measure	0.88	0.5882	0.6666	0.2222	NaN	NaN

Conclusion

Several methods for classification in R were reviewed. In to be acquainted with the various classification and optimization methods present in R, different methods were used for different classification techniques.

The average and standard deviation for accuracy were:

Classification Technique	Average Accuracy	Standard Deviation
KNN	0.5833	0.0382
RIPPER	0.5558	0.0346
C4.5	0.5944	0.0598
SVM	0.6136	0.0603

In general, it was observed that the accuracy of the classifier is dependent on how the data is divided into training and testing sets and the parameters use for tuning the classifiers.

References

i. https://en.wikipedia.org/wiki/List_of_countries_by_life_expectancy
ii. https://topepo.github.io/caret/train-models-by-tag.html
iii. http://topepo.github.io/caret/model-training-and-tuning.html
iv. https://topepo.github.io/caret/pre-processing.html
v. https://cran.r-project.org/web/packages/RWeka/RWeka.pdf
vi. https://rdrr.io/cran/RWeka/man/Weka_control.html
vii. https://cran.r-project.org/web/packages/e1071/e1071.pdf
viii. http://www.calculator.net/standard-deviation-calculator.html