Glass Classification

Can you correctly identify glass type?

Amirsh.nll@gmail.com

Student:

Amir Shokri

Student id:

9811920009

Teacher:

Dr.Farzin Yaghmaee , Mr.Alireza Pourreza

Course:

Machine Learning

Abstract

This is a Glass Identification Data Set from UCI. It contains 10 attributes including id. The response is glass type (discrete 7 values)

1. Data

Attribute Information:

Id number RI: refractive index

Na: Sodium (unit measurement: weight percent in corresponding oxide, as are attributes 4-10)

Mg: Magnesium

Al: Aluminum

Si: Silicon

K: Potassium

Ca: Calcium

Ba: Barium

Fe: Iron

Type of glass: (class attribute) -- 1 building_windows_float_processed -- 2 building_windows_non_float_processed -- 3 vehicle_windows_float_processed -- 4 vehicle_windows_non_float_processed (none in this database) -- 5 containers -- 6 tableware -- 7 headlamps

2. Features

Data exploration of this dataset reveals two important characteristics: 1) The variables are highly correlated with each other including the response variables: So which kind of ML algorithm is most suitable for this dataset Random Forest, KNN or other? Also since dataset is too small is, there any chance of applying PCA or it should be completely avoided?

2) Highly Skewed Data: Is scaling sufficient or are there any other techniques which should be applied to normalize data? Like BOX-COX Power transformation?

3. Models

The documentation uses the KNN algorithm, Decision Tree, Naïve Bayes, MLP, and Logistic Regression; details are as follows:

3.1. KNN algorithm

Best test score is 0.6448598130841121 and K = 1

Best train score is 1.0 and K = 1

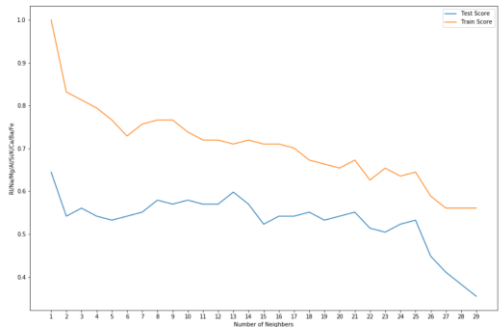

Fig 1: blue (test score), orange (train score)

3.2. Decision Tree

Decision Tree Algorithm test accuracy:

0.6355140186915887

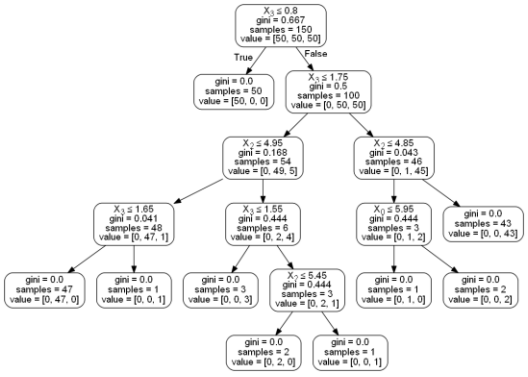

Fig2: decision tree figure

3.3. Naïve Bayes

Naive Bayes test accuracy:

0.34579439252336447

3.4. MLP

hidden_layer_sizes= (7, 2) Accuracy:

0.5700934579439252

hidden_layer_sizes= (5, 2) Accuracy:

0.2616822429906542

hidden_layer_sizes= (9, 8) Accuracy: 0.

0.6355140186915887

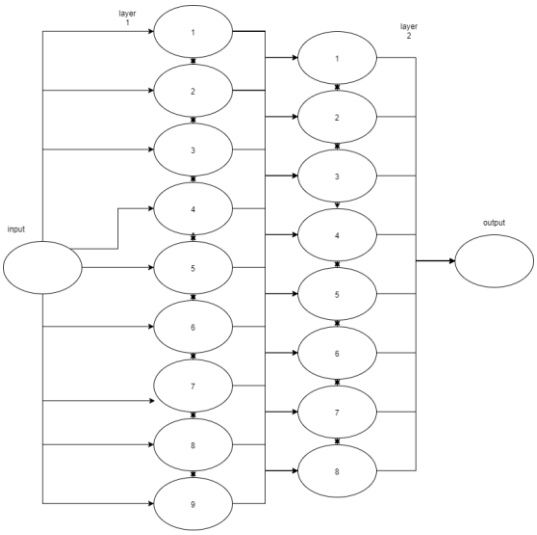

Fig3: mlp schema (1:9:8:1)

3.5. Logistic Regression

Logistic Regression test accuracy (score):

0.42990654205607476

4. Results

According to the results, the highest accuracy is 0.6355140186915887 which is based on the decision tree algorithm and mlp with the number of two hidden layers that are possible in 9 rows and 8 rows, respectively.

Summary: You can see the summary in Table 1 that the Naïve Bayes algorithm is weak accuracy.

KNN algorithm	0.6448598130841121
Decision Tree	0.6355140186915887
Naïve Bayes	0.34579439252336447
MLP	0.6355140186915887
Logistic Regression	0.42990654205607476

Table 1: summary of accuracy

5. Discussion

This exercise was a great help to get acquainted with Classification algorithms. The most interesting part was that by examining each algorithm and changing the inputs and parameters of each algorithm, new results were obtained.

The point here was that the higher the number of data records, the more interesting the results can be, and the better choices and percentages can be made.

6. Acknowledgments

https://archive.ics.uci.edu/ml/datasets/Glass+Identification Source:

Creator: B. German Central Research Establishment Home Office Forensic Science Service Aldermaston, Reading, Berkshire RG7 4PN

Donor: Vina Spiehler, Ph.D., DABFT Diagnostic Products Corporation (213) 776-0180 (ext 3014)

Mr. A

Many thanks to Dr.Farzin Yaghmaee , Mr.Alireza Pourreza for the Machine Learning class.

7. References

https://www.kaggle.com/uciml/glass

https://stackoverflow.com/questions/10059594/a-simple-explanation-of-naive-bayes-classification/20556654

https://github.com/aaqibqadeer/UCI-glass-detection