Machine Learning and Data Mining project: Leaf identification

Sirianne MADON KENGNE

Course of AA 2022-2023 - Data Science and Scientific Computing

1 Problem statement

This work aims to propose a method for identifying leaves based on leaf attributes using a suitable learning method. Some leaves have been collected and organized in group. For each of them, there's a species label that has been assigned. Assuming that the labels have been assigned correctly, it is appropriate to use a supervised learning tool to solve this problem.

Our learning tool will be able to take as input the leaf attributes (Solidity, Aspect ratio, etc.) which are numerical vectors and output the leaf name which is a categorical variable. Hence, we have a multi-classification problem[2].

2 Assessment and performance indexes

Since I have a classification problem, the performance indexes that I used to assess my model are:

- The accuracy;
- The F1 score metric which is good for unbalanced data;
- The Area Under the Curve (AUC): Here, the AUC value is computed for each class. For that, we treat each class as a binary classification problem, with the samples from that class being considered as the positive class and all other samples being considered as the negative class. The final AUC value for the multi-class classification problem is then obtained by averaging the AUC values across all pairs of classes.

3 Proposed solution

Among many learning techniques that we tried for this problem, we choose to present 3 in this report which are:

- Support Vector Machine (SVM)
- Random Forest
- Linear discriminant Analysis(LDA)[1]

They were choose because of their ability to handle with high-dimensional data and unbalanced classes. They are robust to overfitting and easy to use. moreover LDA works by projecting the data onto a lower-dimensional space while maximizing the separation between the different classes.

4 Experimental evaluation

4.1 Data description

The database comprises 40 different plant species corresponding to 30 simple leaves and 10 complex leaves. But only simple leaves have numerical values for their attributes. There are 7 attributes that describe the shape of a leaf, and 7 more that describe what its texture is like.

Initially, we have a table of 340 rows (one for each species) and 16 columns (14 are numerical values for the different attributes and one is the specimen number). The data are unbalanced and have no missing values.

4.2 Procedure

4.2.1 Data preprocessing and partition

I have first removed the useless column that was used just to count the number of leaves for each species. Following this, I changed specimen numbers to the scientific name of the species and converted them to factors. I also assigned each column its own attribute name.

After I preprocessed my data, I divided it into train and test data to be used for all the learning techniques.

4.2.2 Evaluation of the training performance

I conducted 10-folds cross validation (CV) on the training data for each of our learning techniques. I specified a seed to ensure reproducibility of results. The table below shows the accuracy and the kappa of the models in predicting the correct output for a given input based on the training data. As can be seen, LDA has the highest values, followed by random forest and SVM. After performing 10-fold cross-validation to tune the parameters of the random forest model, I determined that using a value of 8 for the number of variables (mtry) resulted in the best prediction performance.

Learning technique	accuracy	Kappa
SVM	0.7342509	0.7228979
Random forest	0.7860774	0.7770040
LDA	0.7949057	0.7862237

Table 1: Models accuracy and Kappa after cross validation.

4.3 Results and discussion

The prediction on the test data using our chosen learning technique gave us the summary result reported in the table2 below.

Learning technique	accuracy	95%CI confidence interval
SVM	0.5714	$(0.4675, 0.671)$
Random forest	0.7245	$(0.625, 0.8099)$
LDA	0.7857	$(0.6913, 0.8622)$

Table 2: Models comparison based on accuracy.

The low accuracy of SVM in this table can indicate that the model is not capturing the underlying patterns in the data. This implies that it is unable to generalize well on the new data. The confidence interval is almost the same for the random forest and the LDA, so I used other performance indexes to choose the best among them.

For the Area Under the curve(AUC), we have $AUC\ RF = 0.8611$ less than $AUC\ LDA = 0.9888$.

The histogram in figure1 show the F1 score analysis of each specie comparing the performance of random forest and LDA. There we can see the information summary in this table3. The accuracy of the LDA is not very different from the

Learning technique	number of species perfectly identified	number of species identified at less than 0.5	number of species unable to identified
Random forest	9	6	0
LDA	12	2	1

Table 3: Information from the F1 score analysis

accuracy of random forest, as it fails to classify the species "Arisarum vulgare" which is perfectly classified by random forest. Other than that, LDA is better than random forest because it is able to identify 12 species and is less effective for just 2 species.

In conclusion, LDA is the learning technique that has the greatest accuracy, the greatest AUC and is a better classifier. So we will propose it as the most appropriate tool for this leaf identification problem.

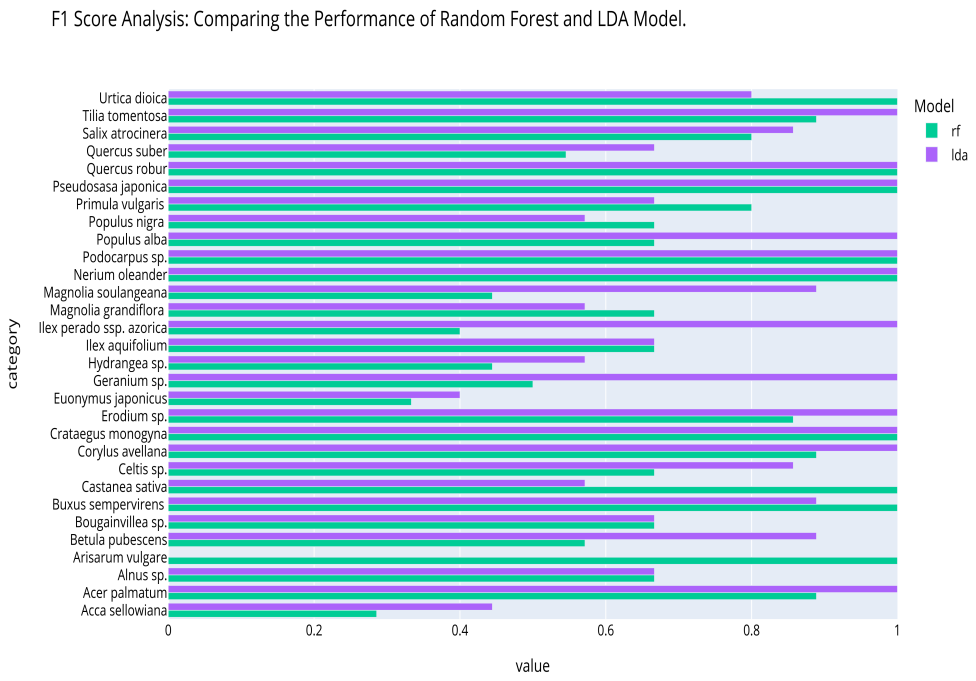

Figure 1: F1 score

References

[1] James Gareth, Witten Daniela, Hastie Trevor, and Tibshirani Robert. *An introduction to statistical learning: with applications in R*. Spinger, 2013.

[2] Hefin Rhys. *Machine Learning with R, the tidyverse, and mlr*. Simon and Schuster, 2020.