Introduction to Machine Learning project: Leaf Identification

Sebastiano Zagatti

Course of AA 2019-2020

1 Problem Statement

[2] Traditionally, the identification of plants has been carried out by specialised technicians called taxonomers, who use several plant attributes in order to distinguish between different species. However, recent years have seen a growing trend in task automation in many fields, plant recognition included.
Currently, there is a shortage on taxonomists and the financial expenditure of this kind of specialized services has been raising. Therefore, the development of an automatic system for plant recognition, using morphometric and photometric informations extracted from digital images of leaves, could provide both specialists and non-specialists with a valuable tool with reduced or no costs.
The goal of this project is to design a method for leaf identification based on a provided set of already classified leaves, each of them characterised by their morphometric and photometric measures. The list of the considered species is reported in Table 1.

Class:	Scientific Name:	Class:	Scientific Name:	Class:	Scientific Name:
1	Quercus Suber	11	Acer Palmatum	21	Ilex Perado Ssp. Azorica
2	Salix Atrocinera	12	Celtis Sp.	22	Magnolia Soulangeana
3	Populus Nigra	13	Corylus Avellana	23	Buxus Sempervirens
4	Alnus Sp.	14	Castanea Sativa	24	Urtica Dioica
5	Quercus Robur	15	Populus Alba	25	Podocarpus Sp.
6	Crataegus Monogyna	16	Primula Vulgaris	26	Acca Sellowiana
7	Ilex Aquifolium	17	Erodium Sp.	27	Hydrangea Sp.
8	Nerium Oleander	18	Bougainvillea Sp.	28	Pseudosasa Japonica
9	Betula Pubescens	19	arisarum Vulgare	29	Magnolia Grandiflora
10	Tilia Tomentosa	20	Euonymus Japonicus	30	Geranium Sp.

Table 1: List of the considered spieces, each with its class reference number.

2 Proposed Solution and Performance Indexes

[1] In order to deal with this multiclass classification problem, a simple base-line needs to be chosen: considering the distribution of the number of observations

for each class, the base-line can be set to be the model whose only response corresponds to the most frequent class.
After that, it's possible to compare three different classification models: Tree, Random Forest and Support Vector Machine, using the accuracy as a performance index. Moreover, further considerations can be made about each single class by taking into account the relative accuracy.

3 Experimental evaluation

3.1 Data

The data used for this project consists of 340 observations of 30 different species of plants; each observation is characterised by 16 different features, including morphometric (attributes 3 to 9) and photometric (attributes 10 to 16) measures:

1. Class (Species)
2. Specimen Number
3. Eccentricity
4. Aspect Ratio
5. Elongation
6. Solidity
7. Stochastic Convexity
8. Isoperimetric Factor
9. Maximal Indentation Depth
10. Lobedness
11. Average Intensity
12. Average Contrast
13. Smoothness
14. Third Moment
15. Uniformity
16. Entropy

Starting from this data, a matrix has been made, where the rows represent each observation and the columns display all the attributes previously listed; it is clear that the second attribute (Specimen Number) is redundant, so it has been removed from the dataset.
In Figure 1 the distribution of the number of observations for each class is displayed. Considering this distribution, there are enough observations for each class in order to apply different classification models, so there is no need to leave some classes out during the process.

3.2 Procedure

After the data management described in the previous section, a trivial computation of the accuracy of the base-line has been carried out, considering that the most frequent class is number 11: *Acer Palmatum.*
Then a 5-fold cross validation has been carried out on three different models: Tree, Random Forest (RF) and a Support Vector Machine (SVM) with radial kernel. Using these techniques, both the accuracy of the model and the accuracy relative to each class have been obtained.
Regarding the cross validation, it is important to point out that the observations are randomly divided into the folds with uniform probability; the consequences

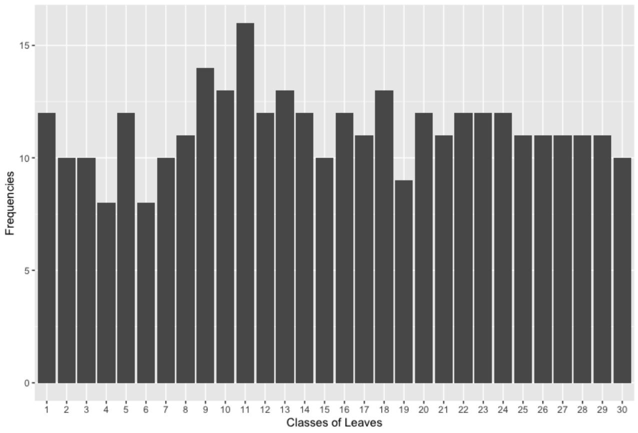

Figure 1: Distribution of the number of observations for each class. Classes are identified using the reference number as displayed in Table 1.

of this choice will be discussed in the next section.
A One-vs-One technique has been chosen for the SVM, due to two different factors:

- The used dataset is not largely populated, so in this case the computational effort implied by this choice is not relevant;
- In general, this technique is less sensitive to the problems related with not uniformly distributed datasets.

3.3 Results and Discussion

The results of the comparison between the base-line and the proposed models are displayed in Table 2.

Model:	Accuracy:
Base-Line	0.05
Tree	0.47
Random Forest	0.76
Support Vector Machine	0.58

Table 2: Models vs Accuracy.

All the three models provide a huge improvement when compared with the considered base-line; it is also clear that Random Forest provides a better accuracy than both Tree and Support Vector Machine.
Besides the global accuracy of the model, it is also important to consider the comparison between the accuracy of the models relative to each class. The results are displayed in Table 3.

	Relative Accuracy:		
Class:	Tree	RF	SVM
1	0.43	0.63	0.28
2	0.10	0.53	0.10
3	0.47	0.73	0.36
4	0.00	0.35	0.00
5	0.67	1.00	0.95
6	0.40	0.80	0.50
7	0.00	0.63	0.30
8	0.96	1.00	1.00
9	0.07	0.63	0.68
10	0.62	0.60	0.45
11	1.00	1.00	1.00
12	0.61	0.46	0.53
13	0.55	0.52	0.34
14	0.26	0.6	0.06
15	0.60	1.00	1.00
16	0.40	0.40	0.70
17	0.40	0.80	0.68
18	0.72	0.78	0.78
19	0.5	0.95	0.75
20	0.37	0.53	0.15
21	0.29	0.63	0.60
22	0.40	0.50	0.30
23	0.52	0.80	0.73
24	0.67	0.93	0.12
25	0.40	0.92	0.68
26	0.60	0.53	0.53
27	0.30	0.63	0.48
28	0.60	0.80	1.00
29	0.43	0.87	0.73
30	0.80	0.73	0.73

Table 3: Class vs Relative Accuracy of each model.

In general RF provides the best accuracy for most of the classes when compared to the other proposed models. For all three models, the accuracy of class number 4 (*Alnus Sp.*) is the worst value. The fact that the value of the relative accuracy of both Tree and SVM is 0.00 could be related either to a problem in the implementation of the models, in particular the cross-validation, or to the fact that that class is inherently difficult to classify. One possible problem in the implementation of cross-validation could be the fact that, as said in the previous section, the observations are chosen with a uniform probability when the folds are built. This problem results in an uncertainty regarding whether the population is correctly represented in the folds or not. This issue has been assessed by considering a different cross-validation procedure, that makes sure that every class is correctly represented in each folder. The new results are displayed in Table 4.

By comparing the values of Table 3 and Table 4, an improvement in the results of the classification of most classes is noticeable when considering the second procedure for cross-validation. Unfortunately, although the problem in the implementation of cross-validation has been solved, it seems that the issues on class number 4 persist. The accuracy values of the three models in this case are presented in Table 5.

The use of a different procedure resulted in a noticeable improvement in both Tree and Support Vector Machine, but Random Forest is still the most accurate model for this classification.

In conclusion, regarding this specific dataset and among the proposed classification techniques, Random Forest provided the best results, both in accuracy and relative accuracy. Regarding the classes, class number 4 turned out to be the most difficult to predict, a possible solution to this issue could be increasing the number of observations of this class.

	Relative Accuracy:		
Class:	Tree	RF	SVM
1	0.07	0.83	0.60
2	0.10	0.60	0.20
3	0.40	0.60	0.60
4	0.00	0.40	0.00
5	0.77	0.90	0.90
6	0.70	1.00	0.70
7	0.20	0.80	0.50
8	0.90	1.00	1.00
9	0.30	0.67	0.60
10	0.63	0.90	0.57
11	1.00	1.00	1.00
12	0.83	0.77	0.77
13	0.47	0.67	0.43
14	0.53	0.70	0.37
15	1.00	1.00	1.00

	Relative Accuracy:		
Class:	Tree	RF	SVM
16	0.57	0.67	0.73
17	0.77	0.93	0.87
18	0.70	0.87	0.60
19	0.10	0.90	0.80
20	0.57	0.60	0.30
21	0.47	0.67	0.73
22	0.40	0.73	0.53
23	0.60	1.00	0.90
24	0.83	0.93	0.33
25	0.80	0.80	0.53
26	0.53	0.43	0.67
27	0.40	0.63	0.57
28	0.70	0.77	1.00
29	0.67	0.77	0.60
30	0.90	0.90	0.90

Table 4: Class vs Relative Accuracy of each model when using a different cross-validation procedure.

Model	Accuracy
Tree	0.57
Random Forest	0.78
Support Vector Machine	0.64

Table 5: Models vs Accuracy using a different cross-validation procedure.

References

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 6. Springer, 2013.

[2] Pedro Filipe Barros Silva. Development of a system for automatic plant species recognition. 2013.