Leaf Identification

Luca Raveri

Introduction to Machine Learning
2019-2020

1 Problem Statement

The goal of the project is to design a classification model that is able to recognize a specific plant species starting from some features that describe the characteristics of leafs in terms of shape and texture. The problem is known in literature as plant identification.
The aim is to compare different possible classification techniques, in the context of supervised learning, in order to provide a solution to the problem.
Several models have been designed and then compared in terms of estimated test error.

2 Proposed Solution

Four classification models are here implemented. First, a Classification Tree has been designed. Then, two models based on the aggregation of trees, with machine learning techniques known as Tree Bagging and Random Forest, were implemented. The last designed model is qualitatively different from the previous ones and is based on the Support Vector Machines technique.

In order to obtain an estimation of the test error of the various models, since test data are not available, the 10-fold cross validation procedure has been adopted. This error estimation has been used as the performance index to rank the effectiveness of the considered models.

3 Experimental Evaluation

3.1 Data Description

As training data we have a dataset made up of 340 samples. This consists of 15 input variables that describe some key physical characteristics of the leaf. The corresponding output variable, which is the species of the plant to which the leaf belongs, is also provided. Each species has been labeled with numbers from 1 to 40.

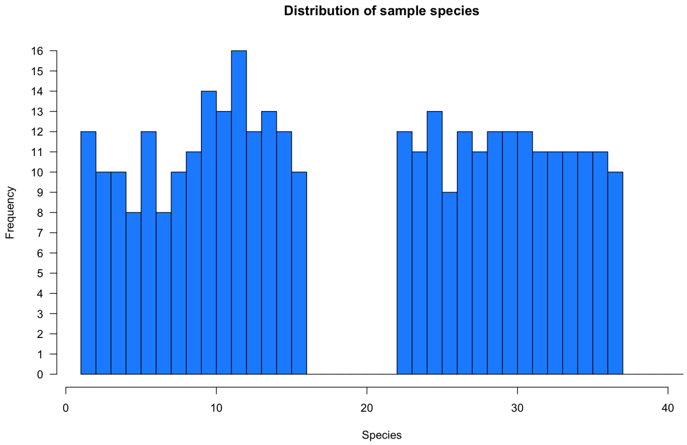

Figure 1: The histogram reporting the distribution of the sample species.

A detailed description of the meaning of each of the 15 predictors and the list of the names of the species is reported in Silva et al. [1].
It has to be noted that samples belonging to species from 16 to 21 and from 37 to 40 are not present in the dataset.

3.2 Experimental Procedure

In this section we discuss some technical details concerning all the adopted design strategies of the implemented Machine Learning techniques, such as parameter tuning.

In order to perform the numerical computation, the programming language R has been used, and the packages ‘rpart’, ‘randomForest’, ‘e1071’ were adopted.

3.2.1 Classification Tree

The first model considered here is a classification tree, that was built by using the recursive binary splitting algorithm. Then, the tree was appropriately pruned.

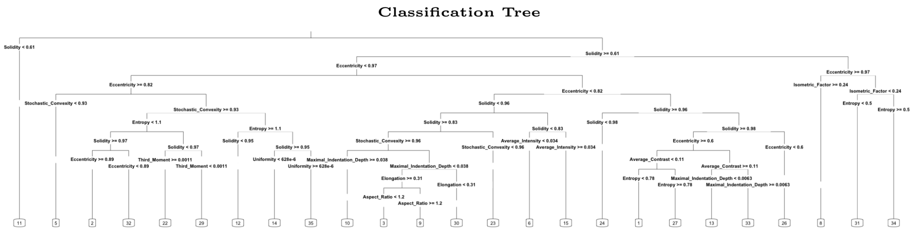

Figure 2: The classification tree plotted using the ‘rpart.plot’ package.

3.2.2 Tree Bagging

The second model, which is based on the aggregation of trees, was built with the Tree Bagging technique.
The procedure is to build B number of classification trees, using B bootstrapped training sets, and then the average of the resulting predictions was performed.
In accordance with James et al., B should be big enough to guarantee a good prediction, and given that overfitting is not occurring, by empirical way, the default value of $B = 500$ has been set [2].

3.2.3 Random Forest

The third method is still based on the aggregation of trees, and was built with the Random Forest technique.
The idea behind the technique is to base each split in the construction of the trees on a sample of m predictors, randomly chosen, instead of basing it on p predictors. This way the built trees will be less correlated.
Regarding the tuning of m, the value has been set in such a way that the OOB error was minimal. The optimal value $m = 3$ was found.

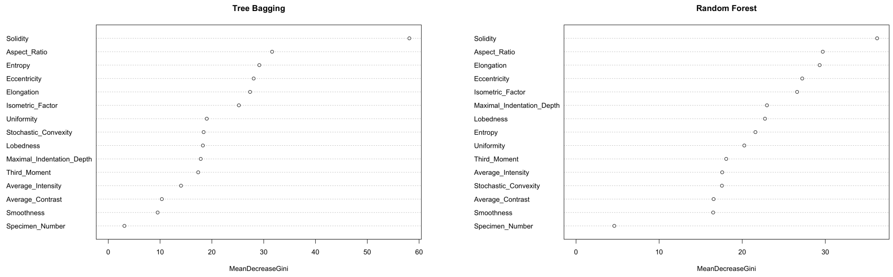

Figure 3: The variable importance, computed using the mean decrease in Gini index.

3.2.4 Support Vector Machines

The fourth model is based on the Support Vector Machines technique.
A radial kernel was selected, and the one versus one approach for the multiclass classification problem was adopted.
As a result of the tuning procedure, the values $\gamma = 0.5$ and $C = 10$ were found.

3.3 Results

The simulation results in terms of test errors, estimated by the 10-fold cross validation procedure, are reported in Table 1.
As we can see, the aggregation of trees has led to a significant increase in performances, from a test error of 0.34 for the Classification Tree based model, to a test error of 0.24/0.23 for the models based on the aggregation of trees.

The Tree Bagging based model achieves a test error of 0.24. A careful choice of the m parameter has brought further improvements, reaching a test error of 0.23 for the Random Forest based model.
Lower performances were achieved by the Support Vector Machines technique, indeed the computed test error is 0.35.
In conclusion, the best classifier over the considered ones is the Random Forest based.

ML technique	**Estimated test error**
Classification Tree	0.3411
Tree Bagging	0.2382
Random Forest	0.2264
Support Vector Machines	0.3470

Table 1: Test error comparison.

4 References

[1] Silva, P.F.B., Marçal, A.R.S., da Silva, R.M.A. (2013). Evaluation of Features for Leaf Discrimination. In: Kamel M., Campilho A. (eds) Image Analysis and Recognition. ICIAR 2013. Lecture Notes in Computer Science, vol 7950. Springer, Berlin, Heidelberg.

[2] James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning with applications in R. Springer, New York.