OTTAWA 2023
64TH WORLD STATISTICS CONGRESS

IPS Abstract

Tree-based statistical learning techniques and explanatory tools

Author: Prof. Rosanna Verde

Submission ID: 1105

Reference Number: 1105

Brief Description

Random Forests [2] represents one of the most popular Machine Learning tools in the field of supervised classification when the number of observations and the number of variables is too large to predict a priori classes.

As known in Statistics and Data Analysis, a good discrimination process requires a selection of predictors in order to avoid problems of the curse of dimensionality.
RF techniques attempt to overcome this problem through a random selection of features and observations.

Thus, they create a large number of decision trees against which to classify elements into a priori classes, according to a 'majority voting' mechanism.

In this way, an assignment rule is established to predict the belonging of an element to a class.

The characteristics that have been selected and that contribute most to the construction of the trees are listed as the most discriminating.

Furthermore, the random choice of predictors solves a redundancy problem in the classification process and provides a very high performance of the Random Forests classifier, measured in terms of accuracy.

In addition, the resampling of a set of observations for the construction of the different trees also involved a process of the robustness of the method.
Boosting methods further improve the performance of tree-based techniques in supervised classification learning by weighing more the best intermediate solutions.
A recent proposal is to apply tree-based techniques to functional data [1] in supervised classification which is a field still little known and underdeveloped.

The proposed contribution focuses on functional classifiers and interpretative tools to improve their performance.

However, as with tree classification methods on classical data, there is a strong automatism in the classification process that represents a challenge to the widely established techniques.

There is no clear description of the decision process, based on the characteristics of the objects to predict the class of belonging.

In particular, the assignment is made with respect to many decision trees and sets of different descriptors that characterize the classes a priori [3, 6].
An interesting contribution could certainly be to provide explanatory and descriptive tools, which, in addition to the accuracy of prediction, allow us to understand the discriminating power of the descriptors selected as the most competitive in constructing the trees.

For this reason, a criterion of recognition of the predictors that most contribute to the separation of the a priori groups should be combined with an embedding procedure that seeks multiple solutions and a final compromise.
The regard of characteristics of the data and, even of aggregated data, such as functional data, is a process of statistical analysis, or learning statistics, which can highlight group structure also in the a priori classes.

The detection of sub-groups in a priori classes can improve interpretation and prediction, as shown in [4].
Furthermore, the description of the separation curves of the classes to be predicted, rather than simple split values, can allow us to interpret the similarity/dissimilarity of the functional characteristics of the curves belonging to the different a priori groups.

In [5] an interpretation of the characteristics of patients with respect to false positives and false negatives was proposed in an application on ECG data.
Aids to the interpretation of the tree-based functional classifiers are still an open frontier.

Some contributions are advanced in the choice of the best transformation of functional data (also by introducing derived functions) to grasp the differences between the classes to be predicted in terms of slope or changing rates.

Applications on real data, in the medical and environmental fields, allow validating of the proposals, related to the interpretative tools in the classification methods based on trees.Main References1.

Ramsay J, Silverman B.

Functional Data Analysis, 2nd edn.

New York: Springer (2005).2.

Breiman L.

Random Forests.

Machine Learning 45(1) 5–32 (2001).3.

Ferraty F, Vieu P.

Curves discrimination: a nonparametric functional approach.

Comput Stat Data Anal.

2003;44(1–2):161-173.4.

Maturo, F., Verde, R.: Pooling random forest and functional data analysis for biomedical signals supervised classification: theory and application to electrocardiogram data.

Statistics in Medicine, 41 (12), 2247–2275 (2022).5.

Maturo, F., Verde, R.: Combining unsupervised and supervised learning techniques for enhancing the performance of functional data classifiers, Computational Statistics (2022) https://doi.org/10.1007/s00180-022-01259-86- Haouij NE, Poggi JM,Ghozi R, Sevestre-Ghalila S, Jaïdane, M.: Random forest-based approach for physiological functional variable selection for driver’s stress level classification.

Statistics Methods and Appl.

28 (1), 157-185 (2018).

Abstract

Random Forests represents one of the most popular Machine Learning tools in the field of supervised classification when the number of observations and the number of variables is too large to predict a priori classes. As known in Statistics and Data Analysis, a good discrimination process requires a selection of predictors to avoid problems of the curse of dimensionality.

RF techniques attempt to overcome this problem through a random selection of features and observations. Thus, they create many decision trees against which to classify elements into a priori classes, according to a 'majority voting' mechanism. In this way, an assignment rule is established to predict the belonging of an element to a class. The characteristics that have been selected and that contribute most to the construction of the trees are listed as the most discriminating. Furthermore, the random choice of predictors solves a redundancy problem in the classification process and provides very high performance of the RF classifier, measured in terms of accuracy. In addition, the resampling of a set of observations for the construction of the different trees also involved a process of the robustness of the method.

Boosting methods further improve the performance of tree-based techniques in supervised classification learning by weighing more the best intermediate solutions.

Tree-based techniques applied to functional data in supervised classification is a field still little known and underdeveloped. The proposed contribution focuses on functional classifiers and explanatory tools to improve their performance. However, the strong automatism in the classification process represents a challenge to the widely established techniques. There is no clear description of the decision process, based on the characteristics of the objects to predict the class of belonging. In particular, the assignment is provided according to a large number of decision trees and different sets of descriptors of the classes a priori. An interesting contribution is to provide new aids which, in addition to the accuracy of prediction, allow to recognize the predictors that most contribute to the separation of the a priori groups. This combines an embedding procedure that seeks multiple solutions and a final compromise.

Moreover, the regard of the characteristics of the functional data can allow the detection of sub-groups in a priori classes can improve interpretation and prediction.

Finally, the description of the separation curves of the classes, rather than simple split values, can allow us to interpret the similarity of the functional characteristics of the curves belonging to the different a priori groups.

Aids to the interpretation of the tree-based functional classifiers is still an open frontier.

Some contributions are advanced in the choice of the best transformation of functional data to grasp the differences between the classes to be predicted in terms of slope or changing rates.

Applications on real data, in medical and environmental fields, have corroborated the proposals.