Supervised classification of curves via a combined use of functional data analysis and random forest

Fabrizio Maturo*, Rosanna Verde**

*University of Campania Luigi Vanvitelli, Caserta, Italy
fabrizio.maturo@unicampania.it
**University of Campania Luigi Vanvitelli, Caserta, Italy
rosanna.verde@unicampania.it

Abstract. Technological advancement led to the development of tools to collect vast amounts of data usually recorded at temporal stamps or arriving over time, like data coming from sensors. Dimensionality reduction and classification techniques have assumed an increasingly important role to deal with the analysis of such data. In recent years, we have witnessed growing research in the fields of functional data analysis on the one hand and machine learning on the other. In this short paper, we propose a new supervised classification strategy that combines statistical FDA techniques and machine learning approaches. Especially, we aim at extended the Random Forest to the case of functional data predictors and a (scalar) response variable. In this context, we focus on the case of a binary dependent variable and thus we concentrate on functional classification trees, particularly on the scalar-on-function classification problem. New interpretative tools are also furnished to support the classification rule. An application on real data of monthly electrical power demand has allowed of exploiting the potential of the high-dimensional data classification strategy when the data are expressed by curves. The ultimate aim of this research is to provide professionals in the energy supply industry with a methodological tool capable of analyzing data from monitoring sensors that produce high-frequency data.

1 Introduction

Today, dimensionality reduction and classification techniques are among the most studied strategies for dealing with the enormous amount of data we deal with every day. The reason is that technological progress led to the evolution of instruments to collect and manage vast amounts of data. Thus, now, we can continuously accumulate data through telephones, calculators, sensors with different purposes, e.g. to control the environment, health, temperature, pressure, and earthquakes. Hence, the hunt for statistical methodologies to investigate this variety of data is crucial.

One of the most recent approaches to dealing with high-dimensional data is functional data analysis (FDA). Particularly, when we deal with time series with a huge number of observations, FDA has recently been proposed in many application contexts. Of course, the FDA can easily be used even when the reference domain is different from time but, in this context, we focus on the case in which the domain is the temporal one.

In summary, the theory and practice of statistical methods in situations where available data are functions (instead of real numbers or vectors) are often referred to as functional data analysis (FDA) (Ferraty and Vieu, 2003; Ramsay and Silverman, 2005; Ferraty, 2011). This topic has become very popular during the last decades and is now a major research field in statistics. Dealing with functional data have a significant impact on statistical thinking and methods, changing how we represent, model, and predict data. The basic idea of FDA is to deal directly with the function generating the data instead of the sequence of observations, and thus to treat observed data functions as single entities.

All the benefits and motivations of using the FDA have been highlighted in the extensive and recent literature (Ferraty and Vieu, 2003; Ramsay and Silverman, 2005; Ferraty, 2011; Febrero-Bande and de la Fuente, 2012; Aguilera and Aguilera-Morillo, 2013; Cuevas, 2014; Maturo et al., 2019), e.g. exploit tools such as derivatives, take advantage of a non-parametric approach without the need for very restrictive assumptions, reduce the dimensionality of the data with the use of the single entity function, etc. For our purposes, we focus on the last aspect, but bearing in mind that our approach can also be extended to derivatives of the original functions.

In recent years, we are witnessing a strong development of the literature on FDA that seeks to reproduce, in a functional key, a large part of classical statistics based on scalar observations. Based on the technique used to represent the functional data, many possible solutions may exist to reduce the dimensionality of the data and deal with them. The functional principal component decomposition is considered by many scholars (Ramsay2005, Ferraty2006, Aguilera2011, Febrero2012). It allows us displaying the functions by a linear combination of a small number of functional principal components (FPC). The functional data can be rewritten as a decomposition in an orthonormal basis by maximizing the variance. Generally, the advantage of this approach is that it finds a lower-dimensional representation preserving the maximum amount of information from the original data. Another standard method for representing the functional data is the basis approximation. Ramsay (2005) suggested that functions can be obtained using a finite representation on a fixed basis.

Much of the traditional statistical techniques have been applied to functional data by exploiting these dimensionality reduction techniques. To date, however, in the field of supervised and semi-supervised classification applied to functions, we can say that research is still in an embryonic phase, especially as regards the combination of FDA and machine learning. In the literature, very recently, some scholars have begun to produce the first studies on the subject.

For this reason, the aim of this paper is to propose a novel approach for supervised classification of high-dimensional data in a temporal domain via the combined use of FDA and machine learning techniques. Specifically, in this study, we focus on FDA and the related dimensional reduction techniques combined with decision trees, bagging, and random forest (RF). The goal is to propose a method to exploit the potential of both approaches to classify high-dimensional data that can be expressed through curves.

2 Material and Methods

The basic idea of FDA is to handle data functions as single objects. Nevertheless, in practical applications, functional data are often observed as series of point data, and thus the function expressed by $z = f(x)$ reduces to record of discrete observations that are denoted by the T pairs $(x_j; z_j)$ where $x \in \Re$ and z_j are the values of the function computed at the points x_j, $j = 1, 2, ..., T$ (Ramsay and Silverman, 2005). Generalizing the reference framework, we consider that a functional variable X is a random variable assuming values in a functional space ξ, and a functional data set is a sample x_1,...,x_N, also denoted $x_1(t)$,...,$x_N(t)$, drawn from a functional variable X (Ferraty and Vieu, 2003). The first step in FDA is to convert the observed values $z_{i1}, z_{i2}, ..., z_{iT}$ for each unit $i = 1, 2, ..., N$ to a functional form. The most common approach to estimate the functional datum is the basis approximation. The basic idea is that functions can be obtained using a finite representation in a fixed basis (Ramsay and Silverman, 2005). Limiting our attention to the $\mathcal{L}_2$ context (see (Ramsay and Silverman, 2005; Aguilera and Aguilera-Morillo, 2013; Febrero-Bande and de la Fuente, 2012) for more details), a function $x(t)$ can be expressed by a linear combination of these basis functions as follows:

$$x(t) = \sum_{j \in \mathbb{N}} c_j \phi_j(t) \approx \sum_{j=1}^{K} c_j \phi_j(t) = \hat{x}(t) \tag{1}$$

where c is the vector of coefficients defining the linear combination and $\phi_j(t)$ is the j-th basis function, from a subset of $K < \infty$ functions that can be used to approximate the full basis expansion.

Another common approach is the Functional Principal Component Analysis (FPCA). The latter leads to a dimensionality reduction whilst preserving the maximum amount of information from the original data (Ramsay and Silverman, 2005; Aguilera and Aguilera-Morillo, 2013; Febrero-Bande and de la Fuente, 2012). In this case, the functional data can be approximated by:

$$\hat{x}_i(t) = \sum_{i=1}^{K} \nu_{ik} \xi_k(t) \tag{2}$$

where ν_{ik} is the score of the generic FPC ξ_k for the generic function x_i $(i = 1, 2, ..., N)$. These approximations are constructed so that the variance of the principle component scores decreases as K increases (with ν_{i1} having maximal variance), implying that $\hat{x}_i(t)$ is the best possible approximation of $x_i(t)$ using K orthogonal basis functions.

Decision Trees (DTs) are a supervised learning technique that predict values of responses by learning decision rules obtained from features. They can be used in both a regression and a classification context. For this reason, they are sometimes also referred to as Classification And Regression Trees (CART). Detailed information on DTs and how to build a pruned tree can be found in many works (Hyafil and Rivest, 1976; Quinlan, 1986; Hastie et al., 2009).

Bagging (Bootstrap aggregating) is an extension of decision trees (Breiman, 1996). With the goal of reducing the variance of a single decision tree, the basic idea is to create many decision trees and build a classifier for each bootstrap replication of the original dataset. To classify new statistical units on the basis of the observed features, all the trees created are used and the class predicted by the majority of trees is assigned (majority vote).

Random Forest (RF) (Ho, 1998) is one of the most efficient machine learning algorithms and is a particular case of bagging for decision trees. It consists of applying bagging to the data and bootstrap sampling to the predictor variables at each split. This implies that at each splitting step of the tree algorithm, a random sample of n predictors is selected as split candidates from the full set of the predictors. This leads to an improvement of the classic bagging because it allows to obtain a classifier that is not strongly influenced by the correlation among trees, which otherwise would all be dominated by the most discriminating variable.

Our starting idea is that DTs, Bagging, and RF can be extended to the FDA framework, both in the case that the functions are obtained by smoothing high frequency data in the time domain and in the event that the functions depend on other specific parameters. In this work, we focus on the former case.

Exploiting the coefficients of a fixed basis system like those in Equation 1, the Decision Tree (DT) and RF approaches can be extended to the case of functional data of the form $\{y_i, x_i(t)\}$, with a predictor curve $x_i(t)$, $t \in J$, and y_i being the (scalar) response value observed at sample $i = 1, ..., n$. The response variable could be either numeric or categorical, leading to regression or classification trees, respectively; however, here we focus on the case of a binary dependent variable and thus we concentrate on functional classification trees, particularly on the scalar-on-function classification problem. Classification trees consist in recursive binary partitions of the feature space into rectangular regions (terminal nodes or leaves). To build the tree, an optimal binary partition is provided at each step of the algorithm, based on the optimization of cost criterion. The algorithm begins with the full data set composed of the coefficients obtained in Equation 1 and continues until the terminal leaves are obtained. Having obtained the best split in one node, the data are partitioned into two nodes; the rule is replicated to performe the most suitable binary separation on all resulting nodes. Typically, a huge tree is produced at the beginning, which is then pruned according to an optimization criterion.

Therefore, the coefficients of the linear combination are used as new features to predict the response. The interpretation is slightly different with respect to the classical DT because the values of the splits of c_j should be interpreted according to the part of the domain that the single b-spline $\phi_j(t)$ mostly represent. Hence, the joint read of the coefficients and of the plot of $\phi_j(t)$ can help interpreting the classification tree. The great problem with a single tree is that its predictive performance is usually not persuasive, and modest changes in the data may lead to very diverse trees. A useful technique to reduce this kind of variance is to create an ensemble of trees using the RF approach (Ho, 1998). The idea of FRF is quite recent. Few papers are available in the literature and the approaches are considerably different. For example, Möller et al. Möller et al. (2016) propose an approach based on the mean of the function within fixed intervals of the domain whereas El Haouij et al. El Haouij et al. (2019) and also Gregorutti et al. Gregorutti et al. (2015) focus on the wavelet basis decomposition. Our approach is quite different and can be based both on the b-spline decomposition and FPCs decomposition.

Assume the FRF consists of H trees τ_h, $h = 1, ..., H$, where H is chosen to be a large number, such as $H = 200$. The *h-th* tree τ_h is grown on a random subset of the training set, obtained from the original data $D = \{(y_i, x_i(t)), i = 1, ..., n\}$ by drawing, with replacement, a bootstrap sample $D^*_h = \{(y^{(h)}_s, x^{(h)}_s(t)), s = 1, ..., n\}$ of the same size n as the original data set. It is straightforward to replace $x^{(h)}_s(t)$ using its expansion in term of b-spline basis as in Equations 1. Thus, the data points $s = 1, ..., n$ present in the *h-th* bootstrap sample D^*_h are called an in-bag sample, on which the *h-th* tree will be grown. Instead, the out-of-bag (OOB) sample is composed of the remaining data points $\{y_i, x_i(t)\}$ that are not present in D_h^*. Thus, we construct H decision trees using H bootstrapped training sets, and we average the resulting predictions. Because each tree is grown deep and is not pruned, each tree has low bias, but high variance. Averaging these H trees diminishes the variance. This is what we can call the phase of "*functional bagging*" (FB) and gives gains in accuracy with respect to a single DT because it combines hundreds or thousands of trees. For a given test observation, we register the class predicted by each of the H trees, and take a "majority vote". Consequently, the overall prediction is the most commonly occurring class among the H forecasts. Expanding the number of trees H will not lead to overfitting. In practice, we want to use a value of H that is large enough for the test error to have settled down. Therefore, in FB, we build a number of decision trees on bootstrapped training samples. Now, suppose that after the expansion computed using Equation 1, we observe that there are some moderately strong predictors but there is one very strong predictor in the training data; in our case, for example, one basis can explain a specific part of the domain and be dominant in resolving the final classification of the original curves. In FB, most or all of the individual trees will use this powerful predictor in the top split. Consequently, all bagged trees will look quite similar to each other, so the predictions from these DTs will be highly correlated. Averaging highly correlated scores leads to a smaller decrease in variance than averaging uncorrelated quantities. Therefore, FB will not lead to a tangible reduction in variance over a single tree.

Functional Random Forest (FRF) gives an improvement over FB because it involves a small tweak that decorrelates the trees. At each split in the tree-building process, we consider a random sample of π predictors, $\pi < K$, as candidates for the split, where K is the total number of b-spline basis (see Equation 1). A new sample of π predictors is taken at each split, for example of size $\pi \approx \sqrt{K}$. Therefore, at each split in the tree, the algorithm is not even allowed to consider a majority of the available b-spline coefficients. Indeed, on average, $\frac{\pi - K}{\pi}$ of the splits will not even contemplate some predictors. In this way, FRF decorrelates the trees, making the average of the trees less variable and hence more reliable. Thus, the difference between FB and FRF depends on the choice of π. When $\pi = K$, FRF is equivalent to FB.

3 Application

In the following, we limit our attention to Equation 1, and thus we consider only one possible way to reconstruct the original curves, i.e. b-splines. The proposed approach is applied to a real dataset derived from twelve monthly electrical power demand time series from Italy and first used in the paper "Intelligent Icons: Integrating Lite-Weight Data Mining and Visualization into GUI Operating System" (Wei et al., 2005). The classification task is to distinguish days from Oct to March (inclusive) from April to September.

Figure 1 illustrates the original signals. The black signals are the days from Oct to March whereas the red curves are those from April to September. The basic idea is therefore to predict, based on the trend of the curves, whether a new curve belongs to the October-March class or the April-September class.

Figure 2 shows the meaning of building many decision trees on the b-spline decomposition using the functional bagging approach.

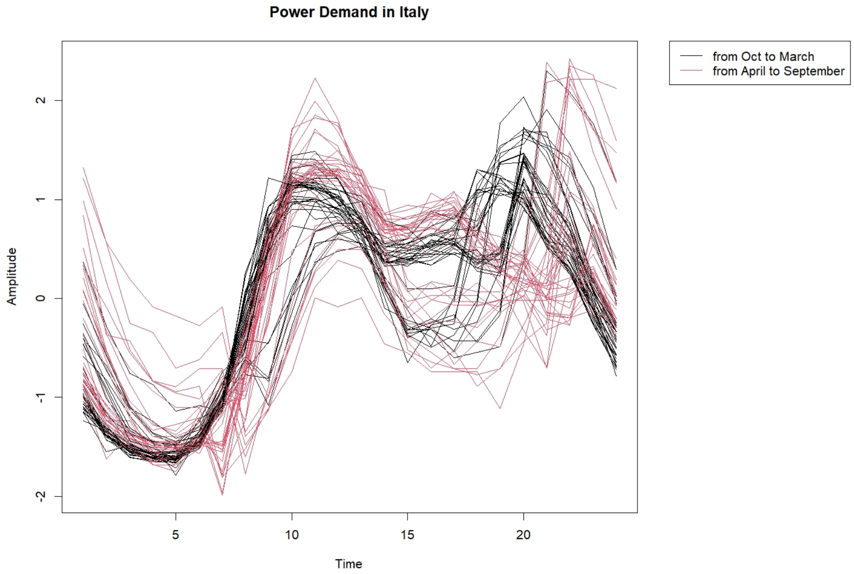

FIG. 1 – *Power Demand in Italy.*

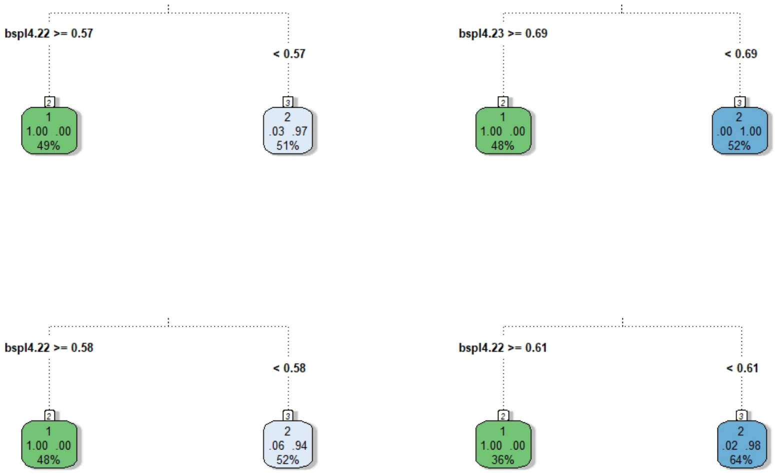

FIG. 2 – *Bagging functional data.*

For the sake of brevity, we limit ourselves to saying that the sample was divided into a training sample and a test sample both of equal size. Using FRF, the percentage of cases correctly classified is 96.60%. This dataset was the subject of a competition to find the best classifier. The best result obtained during the competition is 97.03% accuracy, not very far from the results of our approach, which however has a significant interpretative implication.

Clearly, our approach can be extended to the use of FPCs and also to derivatives as possible new features for building classification tree classifiers with interesting insights.

4 Conclusions

In today's society, thanks to technological advances we are able to collect huge amounts of data. Many of these data, especially in the environmental and industrial fields, come from sensors that produce high frequency observations, for example for monitoring pollution and climate. Consequently, in the last few decades, many methods have developed to deal with this kind of data because traditional statistical methods can fail in many circumstances, for example when we are interested in classification, there is the well-known problem of the curse of dimensionality. This research proposes a new classification approach for high-dimensional data that combines the use of FDA, with the aim of reducing the dimensionality of the data, and some of the most recent tree-based machine learning techniques. In particular, we focused on the application of binary decision trees and random forest using data that are expressed through curves. However, this approach can clearly be extended to bagging and boosting. In this article, we have proposed an application on energy demand in Italy but this methodology can also be successfully extended to data relating to environmental monitoring or other industrial sectors, and more generally, to the management of all those data coming from sensors that produce observations in large quantities, e.g. datastreams. The ultimate goal is to provide practitioners with a useful tool to predict trends and possible perturbations in the environmental and industrial fields.

References

Aguilera, A. and M. Aguilera-Morillo (2013). Penalized PCA approaches for b-spline expansions of smooth functional data. *Applied Mathematics and Computation 219*(14), 7805–7819.

Breiman, L. (1996). Bagging predictors. *Machine Learning*.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*.

El Haouij, N., J. M. Poggi, R. Ghozi, S. Sevestre-Ghalila, and M. Jaïdane (2019). Random forest-based approach for physiological functional variable selection for driver's stress level classification. *Statistical Methods and Applications*.

Febrero-Bande, M. and M. O. de la Fuente (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*.

Ferraty, F. (2011). *Recent Advances in Functional Data Analysis and Related Topics*. Physica-Verlag HD.

Ferraty, F. and P. Vieu (2003). Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis 44*(1-2), 161–173.

Gregorutti, B., B. Michel, and P. Saint-Pierre (2015). Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics and Data Analysis*.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *Springer Series in Statistics The Elements of Statistical Learning - Data Mining, Inference, and Prediction*.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hyafil, L. and R. L. Rivest (1976). Constructing optimal binary decision trees is NP-complete. *Information Processing Letters*.

Maturo, F., A. Balzanella, and T. Di Battista (2019). Building statistical indicators of equitable and sustainable well-being in a functional framework. *Social Indicators Research*.

Möller, A., G. Tutz, and J. Gertheiss (2016). Random forests for functional covariates. *Journal of Chemometrics*.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*.

Ramsay, J. and B. Silverman (2005). *Functional Data Analysis, 2nd edn*. New York: Springer.

Wei, L., E. Keogh, X. Xi, and S. Lonardi (2005). Integrating lite-weight but ubiquitous data mining into GUI operating systems. *Journal of Universal Computer Science*.

Résumé

Les progrès technologiques ont conduit au développement d'outils pour collecter de grandes quantités de données généralement enregistrées sur à des instants temporels ou arrivant en continu, comme des données provenant de capteurs. Les techniques de réduction de dimensionnalité et de classification ont joué un rôle de plus en plus important dans l'analyse de ces données. Ces dernières années, nous avons assisté à une recherche croissante dans les domaines de l'analyse de données fonctionnelles (ADF) d'une part et de l'apprentissage automatique d'autre part. Dans cette article, nous proposons une nouvelle stratégie de classification supervisée qui combine des techniques de données fonctionnelles et les approches d'apprentissage automatique. En particulier, nous visons à étendre la techniques des forêts aléatoires au cas des prédicteurs fonctionnels et nous nous concentrons sur les arbres de classification fonctionnelle lorsque la variables de response est scalaire. De nouveaux outils d'interprétation sont également fournis pour exploiter la règle de classification. Une application sur des données réelles de demande mensuelle de puissance électrique a permis d'exploiter le potentiel de la stratégie de classification des données de grande dimension lorsque les données sont exprimées par des courbes. L'objectif ultime de cette recherche est de fournir aux professionnels de l'industrie de l'approvisionnement énergétique un outil méthodologique capable d'analyser les données des capteurs de surveillance qui produisent des données haute fréquence.