Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Computational Intelligence & Information Management

Interpretable support vector machines for functional data

Belen Martin-Barragan *, Rosa Lillo, Juan Romo

Department of Statistics, Universidad Carlos III de Madrid, Spain

ARTICLE INFO

Article history:
Received 8 November 2011
Accepted 18 August 2012
Available online 12 September 2012

Keywords:
Data mining
Interpretability
Classification
Linear programming
Regularization methods
Functional data analysis

ABSTRACT

Support Vector Machines (SVMs) is known to be a powerful nonparametric classification technique even for high-dimensional data. Although predictive ability is important, obtaining an easy-to-interpret classifier is also crucial in many applications. Linear SVM provides a classifier based on a linear score. In the case of functional data, the coefficient function that defines such linear score usually has many irregular oscillations, making it difficult to interpret.

This paper presents a new method, called *Interpretable Support Vector Machines for Functional Data*, that provides an interpretable classifier with high predictive power. Interpretability might be understood in different ways. The proposed method is flexible enough to cope with different notions of interpretability chosen by the user, thus the obtained coefficient function can be sparse, linear-wise, smooth, etc. The usefulness of the proposed method is shown in real applications getting interpretable classifiers with comparable, sometimes better, predictive ability versus classical SVM.

1. Introduction

The term Functional Data Analysis was already used in [30] two decades ago. Since them, especially in the last decade, it has become a fruitful field in statistic. The range of real world applications where the objects can be thought as functions is as diverse as speech recognition, spectrometry, meteorology or clients segmentation, to cite just a few [19,9,17,20]. The objects of study in Functional Data Analysis (FDA) are functions. A good review of different FDA techniques applied to real world problems can be found in [31]. For a deeper insight into the subject see, e.g., [10,32].

We deal with the problem of classifying functional data. Suppose we observe a binary response Y (the class) to a functional predictor X, where $X \in \mathcal{X}$ is a function defined on the bounded interval $\mathcal{I}$, i.e., $X : \mathcal{I} \mapsto \mathbb{R}$, and $\mathcal{X}$ is a given set of functions. Our aim is to construct a classification rule that predicts Y for a given functional datum X with good prediction ability and some interpretability properties.

The classification rule is based on the sign of the so-called *score function f*. The score function is an operator $f : \mathcal{X} \mapsto \mathbb{R}$ that assigns a real number to a given function X. Since our aim is interpretability, we consider the score function to be a linear operator $T_{\beta,\omega}$ with coefficient function $w \in \mathcal{X}$ and intercept $\beta \in \mathbb{R}$,

$$f(X) = T_{\beta,w}X = \int_{\mathcal{I}} w(t)X(t)dt + \beta = \langle w,X \rangle + \beta, \qquad (1)$$

where $\langle f,g \rangle = \int_{\mathcal{I}} f(t)g(t)dt$. The estimation of the coefficient function w on the whole interval $\mathcal{I}$ is an infinite dimensional problem. This issue is addressed via regularization, which simultaneously allows us to address our other concern: interpretability.

As in standard Support Vector Machines (SVMs), $w(t)$ express the discriminative power of $X(t)$. For example, areas where $w(t)$ is zero or small has none or low discrimination power, whereas for $|w(t)|$ large, one can expect the behavior of $X(t)$ to have influence over the classification. This idea provides a clear interpretation of $w(t)$ at a particular time point t, but getting a general idea about the coefficient function w requires it to be simple: cases where $w(t)$ has unnatural wiggles all along the interval $\mathcal{I}$ are difficult to interpret.

The simplicity of w might be understood in different ways depending on the application. For instance, a coefficient function that is non-zero in just a few points, could detect the few points that are more relevant in classification. This idea has been proposed within a logistic regression model, see [24]. In other situations, one might prefer a coefficient function that is constant over a few subintervals of $\mathcal{I}$ and zero on the rest. A method that detects a few segments with high discriminative power has been proposed in [22] by combining feature selection, classical linear discriminant analysis and SVM. In gene expression analysis, detection of relevant segments are also quite desirable because relevant genes are expected to be located close to each other along the chromosome [33]. All this literature provides different methodologies for different notions of interpretability. Our proposal is to provide a common framework where all this notions can be seen as particular cases.

We use the interpretability notions proposed by [17] for functional linear regression. We consider that a classifier is interpretable

* Corresponding author.
E-mail addresses: belen.martin@uc3m.es (B. Martin-Barragan), lillo@est-econ.uc3m.es (R. Lillo), juan.romo@uc3m.es (J. Romo).

0377-2217/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ejor.2012.08.017

if one or several derivatives of the coefficient function w are sparse, i.e., the derivatives are zero in many points. The choice of the derivatives that are enforced to be sparse depends on the notion of interpretability preferred by the practitioner. In this context, this paper proposes a new method, that we call *Interpretable Support Vector Machines for Functional Data* (ISVMFD), producing SVM-based classifiers for functional data which have high classification accuracy and whose coefficient functions are easy to interpret. The problem is formulated as a linear program, in the framework of L_1-norm SVM.

The seek of interpretability is not new in functional data analysis. A penalized version of the classical Linear Discriminant Analysis (LDA) is proposed in [14] and is denoted as PDA. PDA and ISVMFD share common ideas: regularization and interpretability. However the two methods are different in many aspects. The main difference is the error criteria used: ISVMFD is based in minimizing the hinge loss whereas PDA is based on maximizing the between-class variance relative to the within-class variance. Besides, interpretability in PDA is achieved by using a penalty matrix that imposes a spatial smoothness constraint on the coefficients.

Another approach for finding interpretable classifier is variable clustering techniques, or in a more general framework, variable selection. Methods that use this kind of selection techniques are usually based on a two-phase framework. There is a phase where the variables are clustered or selected, and the classifier is built in a posterior phase. For instance, in [18] a variable clustering phase is embedding into a three-phase classification procedure in order to select ranges in spectra. See, for instance, [12,13] for a review in the wide variety of feature selection methods that can be applied within a two-phase framework. In contrast, when IFSVM is used, the selection phase is done together with the construction of the classifier.

The outline of the paper is as follows: Section 2 reviews classical literature for SVM on multivariate data, its extension to functional data and how interpretability has been addressed for multivariate data. In Section 3 the ISVMFD method is introduced and a proposal to implement it through the use of a basis is provided. Section 4 studies how other methods available in the literature are particular cases of ISVMFD. A wide study with two real-world datasets is presented in Section 5 and finally, in Section 6, several conclusions are driven. An Online Companion Appendix that includes more illustrative examples is provided.

2. Support vector machines

We focus in this paper on the binary supervised classification problem, where two classes $\{-1,1\}$ of curves need to be discriminated. SVM [8,27,38] have become very popular during the last decade. The basic idea behind SVM can be explained geometrically. If the data are living in a p-dimensional space, SVM finds the separating hyperplane with maximal margin, i.e., the one furthest away from the closest objects. This geometrical problem is expressed as a smooth convex problem with linear constraints, solved either in its primal or dual form. Another interpretation can be done in terms of the regularization theory where the hinge loss plus a quadratic regularization penalty is minimized [15,35]. The most popular and powerful versions of SVM embed the original variables into a higher dimensional space [16]. This embedding is usually implicitly specified by the choice of a function called kernel.

Extensions of SVM to functional data have been proposed in [28,34]. In [28], SVM is used to represent the functional data by projecting the original functions onto the eigenfunctions of a Mercer Kernel. Ref. [34] define new classes of kernels that take into account the functional nature of the data. Two types of functional kernels are proposed: projection-based kernels and transformation-based kernels. In projection-based kernels, the idea is to reduce the dimensionality of the input space, i.e., to apply the standard filtering approach of FDA. Transformation-based kernels allow to take into account expert knowledge (such as the fact that the curvatures of a function can be more discriminant than its values).

In the multivariate context, kernels provide an implicit way to get a nonlinear classifier, by projecting the data on the higher dimensional space induced by the kernel. The final classifier is nonlinear in the original space, but linear in the projected space. Functional data are indeed high dimensional and the high dimensionality usually generates problems. Therefore the use of kernels to project data on a higher dimensional space seems to be less crucial. Moreover, the kernel-based classifier would be easy to interpret in the projected space, but not in the original one. We focus on the linear kernel in our method.

The interpretability issue in SVM has already been addressed for multivariate data. The first attempts to make SVM more interpretable make use of a two-step procedure: first, SVM is run, and then a rule, resembling the SVM-classifier but easier to interpret, is built. See, e.g. [1,3,26,25]. One obtains an alternative classifier which hopefully get similar predictions, but is more interpretable. Recently, a two-stage iterated method is proposed for credit decision making [23], which combines feature selection and multi-criteria programming. In [6,7], one-step SVM-based procedures are proposed to get the relevant variables and the relevant interactions between variables. Although one would expect classification rates to deteriorate when looking for interpretable classifiers, the experiments in [6,7] show that their proposals are competitive with SVM. See [2,21,37,39] for other recent references on the topic.

3. Methodology

3.1. Interpretable support vector machines for functional data

Let $\{X_u, Y_u\}_{u=1}^{n}$ be a sample of n functional data $X_u \in \mathcal{X}$ together with its class $Y_u \in \{-1,1\}$. The classical SVM with the linear kernel seeks for the coefficient function w that minimizes

$$min_{w,\beta} \|w\|_q^q + C\sum_{u=1}^{n} h(y_u, \langle w, X_u\rangle + \beta) \tag{2}$$

where $\|\cdot\|_q$ is the q-norm, $h(y,s) = (1-ys)_+$ is the hinge loss and C is a tuning parameter that trades off the regularization term $\|\omega\|_q^q$ and the loss term.

The class is predicted as the sign of the score function given in (1). In case of ties, i.e., $f(X) = 0$, prediction can be randomly assigned or following some predefined order. Throughout this article, following a worst case approach, ties will be considered as misclassifications.

Although the regularization with the Euclidean norm is the most common, other norms have also been applied. For instance, the L_1 norm is known to be good when a sparse coefficient vector is desirable. Ref. [4] demonstrated the usefulness of penalties based on the L_1 norm in classification problems. In regression, LASSO [35] and the Dantzig selector [5] also successfully use the L_1 norm in high-dimensional problems.

In order to get the interpretable classifier, we propose a modified version of SVM that we call Interpretable Support Vector Machines for Functional Data (ISVMFD). Following the concepts of interpretability described in Section 1, we propose to use a different regularization term that depends on the preferences of the user for the interpretability notion. The user must select one or several derivatives to be sparse. For example, if the user is concerned with detecting relevant time points, the zero derivative (the actual w) is selected to be sparse. Sparsity of the first derivative leads to constant-wise w which is useful to identify relevant segments. A user might prefer a coefficient function that is zero over large regions,