First International Workshop on
Functional and Operatorial Statistics.
Toulouse, June 19-21, 2008

Recent advances in the use of SVM for functional data classification

Nathalie Villa*[(1,2)] and Fabrice Rossi[(3)]

Addresses of authors:
[(1)] Université de Toulouse, IMT (Institut de Mathématiques), 118 route de Narbonne, F-31062 Toulouse cedex 9, France
[(2)] IUT de Perpignan, Département STID, Domaine Universitaire d'Auriac, F-11000 Carcassonne, France
[(3)] Projet AxIS, INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, B.P. 105, F-78153 Le Chesnay cedex, France
`nathalie.villa@math.univ-toulouse.fr` and `fabrice.rossi@inria.fr`

Abstract
In the past years, several works were dealing with the use of Support Vector Machine (SVM) for classifying functional data. Here, we propose to give an overview of these works and to introduce a new result based on the use of smoothing conditions on the observed functions. The originality of this approach both lies in the fact that the consistency result allows to work with the derivatives of the function instead of the function itself but also that it is relative to the observed discretization and not to the entire knowledge of the functions.

Introduction

As the number of data coming from continuous recording has increased, the analysis of data taking the form of curves has also developed. After the pioneering work of [6, 3, 14] in the framework of linear models, various statistical methods have been adapted to what is now called *functional data analysis* (FDA): this is the case of nonparametric estimation [8, 7], of neural networks [9, 15] or of k-nearest neighbors [2], to name a few.

SVM were introduced in the past years and they appear to be a competitive tool for solving binary classifications. One of their main interest is that they are less sensitive to

the dimensionality of the predictor than other methods. Then, they are potentially an interesting approach in FDA. In his PhD thesis [11], Lee first uses the SVM for classifying curves: his approach was based on PCA pre-processing and was illustrated by several examples. Unfortunately, no consistency result was given. In [17] and [23], the authors present various ways for dealing with binary classification of curves by the way of SVM: the first article presents a projection approach that is valid for any Hilbert space and the second one uses smoothness constraints by the way of a spline interpolation.

This article intends to summarize the past theoretical results obtained for classification of curves with SVM and to introduce a new consistency result with respect to the discretization of the observations. This approach is original as it allows to work on the derivatives of the observations which can be a relevant task for many kind of problems [7, 16, 5]. In section 1, we recall the SVM algorithm and the existing consistency results in the multi-dimensional context. Then, section 2 presents the adaptation of this algorithm to the FDA context. To that aim, section 2.2 develops a consistency result by a projection method and section 2.3 a consistent method on derivatives which uses smoothing splines approximation of the predictors.

The proof of the results given in this paper as long as several applications on real data sets can be found in [17, 18].

1 SVM classifiers

1.1 Definition

Vapnik [22] introduces a theoretical context to model statistical learning and popularized Support Vector Machines (SVM), particularly in the framework of binary classification. To recall what is the principle of SVM, suppose that a training set of size n, $(z_i, y_i)_i$, of i.i.d. observations is given: (z_i) take their values in a space $\mathcal{X}$ and (y_i) in $\{-1, 1\}$. SVM are classifiers that belong to a family of semi-linear classifiers of the form $\phi_n(z) = \text{Sign}\{\langle w, \psi(z)\rangle_{\mathcal{F}} + b\}$ where $\psi : \mathcal{X} \to \mathcal{F}$ is a given nonlinear function from $\mathcal{X}$ to a Hilbert space $\mathcal{F}$, called *feature space*. Then, w and b are parameters that have to be learnt from the data set: they are chosen by an optimization problem that aims at maximizing the margin between the observations $(\phi(x_i))$ from both classes and the decision frontier. More precisely, they are the solution of:

$$\begin{array}{ll} (P_{C,\mathcal{F}}) & \min_{w,b,\xi} \|w\|^2_{\mathcal{F}} + C\sum_{i=1}^n \xi_i \\ & \text{such that} \quad y_i(\langle w, \psi(z_i)\rangle_{\mathcal{F}} + b) \geq 1 - \xi_i,\ 1 \leq i \leq n, \\ & \qquad\qquad \xi_i \geq 0,\ 1 \leq i \leq n. \end{array}$$

Problem $(P_{C,\mathcal{F}})$ has a dual formulation that doesn't directly use the transformed data $\psi(z_i)$ but the inner product $\langle \psi(z_i), \psi(z_j)\rangle_{\mathcal{F}}$. Thus, the nonlinear transformation ψ and the feature space, $\mathcal{F}$ don't have to be explicitly known: they are implicitly used by defining the scalar product, $\langle \psi(z_i), \psi(z_j)\rangle_{\mathcal{F}}$ by the way of a *kernel trick*. A symmetric and positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is chosen: according to Moore-Aronszajn theorem [1], this ensures that there is a Hilbert space $\mathcal{F}$ and an application $\psi : \mathcal{X} \to \mathcal{F}$ such that $\langle \psi(x_i), \psi(x_j)\rangle_{\mathcal{F}} = K(x_i, x_j)$.

1.2 Universal consistency of SVM

SVM are known to have good generalization properties when $\mathcal{X}$ is a finite dimensional space. More precisely, [19, 20] show that d-dimensional SVM are universally consistent, under some hypothesis i.e., that $\lim_{n\to+\infty} L\phi_n = L^*$ where $L\phi_n$ is the probability of misclassification of the classifier ϕ_n, $L\phi_n = \mathbb{P}(\phi_n(Z) \neq Y)$, and L^* is the *Bayes error*, the optimal misclassification rate for the random pair (Z, Y) having same distribution as (z_i, y_i), $L^* = \inf_{\phi:\mathcal{X}\to\{-1,1\}} \mathbb{P}(\phi(Z) \neq Y)$.

This result is obtained with particular kernels: if $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$, the kernel K used has to be *universal* i.e., the set $\{z \in \mathcal{X} \to \langle w, \psi(z)\rangle_{\mathcal{F}},\ w \in \mathcal{F}\}$ has to be dense in the set of continuous functions on $\mathcal{X}$. Secondly, for $\epsilon > 0$, $\mathcal{N}(\epsilon, K)$ is the *covering number* of the space $\mathcal{F}$ i.e., the minimum number of balls of radius ϵ that are needed to cover $\mathcal{F}$; consistency of SVM also requires that $\mathcal{N}(\epsilon, K) = \mathcal{O}(\epsilon^{-\nu_d})$ for a $\nu_d > 0$. Among others, Gaussian kernels, $K_\gamma^d(u, v) = exp\left(-\gamma\|u - v\|_{\mathcal{X}}^2\right)$, satisfy both assumptions with $\nu_d = 1/d$ (see [20]) but this can't be extended to the case where $\mathcal{X}$ has infinite dimension both because the covering number assumption is not fullfilled for usual kernels (as Gaussian kernel) and because assuming that the variable takes its values in a compact set is too much restrictive in infinite dimensional spaces.

In the following, $K_\gamma^d \in$ **(Acv)** will denote any kernel on $\mathbb{R}^d$ that satisfy these two conditions and that possibly can depend on a parameter γ. Moreover, if the calculation of $K_\gamma^d(u, v)$ is only based on the inner product of u and v in $\mathbb{R}^d$, such a kernel can be generalized into K_γ^∞ which is a kernel on L^2 that has the same form as K_γ^d except that the $\mathbb{R}^d$-inner product is replaced by the L^2-inner product.

2 Using SVM to classify functional data

As was explained above, the consistency result obtained for d-dimensional SVM can't be applied directly to the infinite dimensional case. Moreover, in FDA, the observations are not direct realizations of a random pair having a functional predictor: if (X, Y) is a random couple taking its values in $L^2 \times \{-1, 1\}$, then i.i.d. realizations of (X, Y), (x_i, y_i), are not directly observed as (x_i) are only known through a discretization, $\mathbf{x}_i = (x_i(t))_{t\in\tau}$ where τ is a finite subset of $[0, 1]$.

2.1 Kernels for functional data

To obtain consistency result for functional SVM, a pre-processing is required that takes into account the functional nature of X. Depending on the problem, two kinds of pre-processing are investigated in this paper:

- A *projection approach* (developed in [17]) where the pre-processing step is $\mathcal{P} : x \in \mathcal{H} \to \sum_{j=1}^d \langle x, e_j\rangle_{\mathcal{H}} e_j$ where $(e_j)_{j\geq 1}$ is a Hilbert basis of any Hilbert space, $\mathcal{H}$ which is the space where X is taking its values (e.g., a Fourier basis if $\mathcal{H} = L^2$, as stated above). In this approach, $\mathcal{P}(X)$ is a random variable taking its values in a d-dimensional space; then, as it is usual in FDA, a d-dimensional SVM can be computed on the d coordinates of the projection.

- A *differential approach* where a prior assumption on X is used: X is supposed to be "smooth" and, more formally, it is supposed to belong to the Sobolev space $\mathcal{H}^m = \left\{x \in L^2([0,1]) : \ D^m x \text{ exists (in a weak sense) and } D^m x \in L^2\right\}$. This Sobolev space is a Hilbert space with respect to the inner product $\langle u, v\rangle_{\mathcal{H}^m} = \int_0^1 u^{(m)}(t)v^{(m)}(t)dt + \sum_{j=1}^m B^j u B^j v$ where (B^j) denotes m boundary conditions that defines an infinite dimensional subspace of $\mathcal{H}^m$, $\mathcal{H}_1^m$, such that $\mathcal{H}^m = \mathcal{H}_0^m \oplus \mathcal{H}_1^m$ with $\mathcal{H}_0^m = \text{Ker}D^m$ (see [10]). Thus, in this approach, the pre-processing consists in using the derivatives of the original function: $\mathcal{P}^s(X) = (D^m X, (B^j X)_j)$.

The following sections are dedicated to the presentation of consistency results associated to these two approaches and to the description of their advantages and weaknesses.

2.2 Projection approach

The consistency of the projection approach depends on a validation procedure that aims at choosing optimal parameters of the model. Indeed, three parameters have to be chosen for using the SVM on the pre-processed data $(\mathcal{P}x_i)_i$: the best dimension of projection, d, the best regularization parameter, C, in $(P_{C,\mathcal{F}})$ and the best kernel among a finite set of kernels, $\mathcal{K}_d$. If $\mathcal{A}$ denotes a set of lists of parameters to explore, the choice of the optimal parameters, a^* in $\mathcal{A}$ has to be done by the validation procedure described in Algorithm 1.

Algorithm 1 Functional SVM by projection: a validation approach

1: **for** all $a \equiv d \in \mathbb{N}^*, K_\gamma^d \in \mathcal{K}_d, C \in [0; C_d]$ in $\mathcal{A}$ **do**
2: Split the data set into $\mathcal{B}_1 = (x_i, y_i)_{i=1,\dots,l}$ and $\mathcal{B}_2 = (x_i, y_i)_{i=l\dots,n}$.
3: Solve $(P_{C,\mathcal{F}})$ with $z_i = \mathcal{P}x_i$ for the chosen parameters a; the corresponding classifier will be denoted by ϕ_l^a.
4: **end for**
5: Choose $a^* = \arg\min_{a\in\mathcal{A}} \hat{L}_{n-l}\phi_l^a + \frac{\lambda_d}{\sqrt{n-l}}$ with $L_{n-l} = \frac{1}{n-l}\sum_{i=l+1}^n \mathbb{I}_{\{\phi_l^a(x_i)\neq y_i\}}$ and $\lambda_d \in \mathbb{R}$.
6: Finally, keep the classifier $\phi_n = \phi_l^{a^*}$.

A consistency result can be deduced from this procedure:

Theorem 1. *[17] Suppose that:*

Assumption on X: *X takes its value in a bounded subset of $\mathcal{X}$;*

Assumptions on $\mathcal{A}$: *for all $d \geq 1$, $\mathcal{K}_d$ is a finite set that contains a kernel $K_\gamma^d \in$ **(Acv)** at least, $\mathcal{C}_d > 1$ and $\sum_{d\geq 1} |\mathcal{K}_d| e^{-2\lambda_d^2} < +\infty$;*

Assumptions on the training and the validation sets: $\lim_{n\to+\infty} l = +\infty$, $\lim_{n\to+\infty} n - l = +\infty$ *and* $\lim_{n\to+\infty} \frac{l\log(n-l)}{n-l} = +\infty$.

Then, ϕ_n is universally consistent: $\lim_{n\to+\infty} L\phi_n = L^*$ *where* $L\phi_n = \mathbb{P}(\phi_n(X) \neq Y)$ *and* $L^* = \inf_{\phi:\mathcal{H}\to\{-1,1\}} \mathbb{P}(\phi(X) \neq Y)$.

Two applications of this approach in the context of voice recognition are given in [17]. Moreover, [12] also uses this approach to classify gene expression data into functional groups but with a linear kernel.

2.3 Differentiation approach

The projection pre-processing shows interesting results on real data but is somehow restrictive: the form of the representation of X is constrained by an Hilbert basis and the derivatives of X, that are known to be relevant in some practical applications (such as spectrometric data), don't lead to a consistent result with this approach. Moreover, the problem of using a discretization of the observations isn't addressed.

1 Representing X

In the differential approach, x_i is expressed directly in function of its discretization: that allows to obtain its derivatives directly from $\mathbf{x}_i$. In [23], we investigated a method that is close to this one by relying on interpolating splines. But, as the observations of X can be noisy, smoothing splines can be usefull to provide more relevant representations of x_i.

Suppose that $(\tau_d)_d$ is a series of distinct discretization points such that $\tau_d \subset \tau_{d+1}$, then representing x_i by a smoothing spline, from its discretization $\mathbf{x}_i^d = (x_i(t))_{t\in\tau_d}$, consists in solving the optimization problem $x_i^{\lambda,d} = \arg\min_{h\in\mathcal{H}^m} \frac{1}{d}\sum_{t\in\tau_d}(x_i(t)-h(t)) + \lambda\int_0^1 (h^{(m)}(t))^2 dt$ (see [10, 4, 13, 21] for several consistency results of this approximation to the real x_i). The most interesting point of this approach is that it links the derivatives of the smoothing spline estimate with the discretization of the observation: it exists a matrix $\mathbf{M}_d$, symmetric and positive definite, such that

$$\langle \hat{x}_i^{\lambda,d}, \hat{x}_j^{\lambda,d}\rangle_{\mathcal{H}^m} = \mathbf{x}_i^T \mathbf{M}_d \mathbf{x}_j. \tag{1}$$

2 Differentiation kernel for consistent functional SVM

Therefore, using equation (1), a kernel on the derivatives of (x_i) can be defined that is directly computed from the discretizations $\mathbf{x}_i^d$. The following theorem links SVM computed on the derivatives of (x_i) with a more usual kernel affected by the matrix $\mathbf{M}_d$:

Theorem 2 (Consistency of differentiation SVM)**.** *The SVM classifier on $(z_i)_i = (D^m x_i^{\lambda,d}, (B^j x_i^{\lambda,d})_j)_i$ obtained with kernel $K_\gamma^\infty \otimes K_\gamma^m$ is equivalent to the SVM classifier on $(\mathbf{x}_i)_i$ obtained with kernel $K_\gamma^d \circ \mathbf{M}_d^{-1/2}$.*

If this classifier is denoted by $\phi_{n,d}$, and if

Assumptions on the discretization points: *for all d, $(B^j)_j$ are linearly independant from $\{h \to h(t)\}_{t\in\tau_d}$ and, if F is the limit of $F_d(\zeta) = \frac{1}{|\tau_d|}\sum_{t\in\tau_d}\mathbb{I}_{\{\zeta=t\}}$ for the norm $\|u-v\|_\infty = \sum_{t\in[0,1]}|u(t)-v(t)|$, then F is $\mathcal{C}^\infty$,*

Assumption on X: *$X[0,1]$ is a bounded subset of $\mathbb{R}$,*

Assumptions on the kernel: *$K_\gamma^d \in$* **(Acv)**,

Assumptions on the parameters: *if $S_d = \|F_d - F\|_\infty$ then $\lim_{d\to+\infty}\lambda_d = 0$ and $\lim_{d\to+\infty} S_d\lambda_d^{-5/(4m)} = 0$ and the regularization parameter C of the optimization problem $(P_{C,\mathcal{X}})$ is such that $C_{n,d} = \mathcal{O}\left(n^{1-\beta_d}\right)$ where $0 < \beta_d < \nu_d$,*

then, $\lim_{d\to+\infty}\lim_{n\to+\infty} L\phi_{n,d} = L^*$ for $L\phi_{n,d}$ and L^* defined as in theorem 1.

Remark. Assumptions on (τ_d) are fullfilled by $\tau_d = \left\{\frac{j}{2^d}\right\}_{j=0,\dots,2^d}$, for example (see [13]).

References

[1] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[2] G. Biau, F. Bunea, and M. Wegkamp. Functional classification in Hilbert spaces. *IEEE Transactions on Information Theory*, 51:2163–2172, 2005.

[3] H. Cardot, F. Ferraty, and P Sarda. Functional linear model. *Statistics and Probability Letters*, 45:11–22, 1999.

[4] D.D. Cox. Multivariate smoothing splines functions. *SIAM Journal on Numerical Analysis*, 21:789–813, 1984.

[5] S. Dejean, P.G.P. Martin, A. Baccini, and P. Besse. Clustering time-series gene expression data using smoothing spline derivatives. *EURASIP Journal on Bioinformatics and Systems Biology*, 2007:Article ID70561, 2007.

[6] J.C. Deville. Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'INSEE*, 15(Janvier–Avril):3–97, 1974.

[7] F. Ferraty and P. Vieu. The functional nonparametric model and application to spectrometric data. *Computational Statistics*, 17:515–561, 2002.

[8] F. Ferraty and P. Vieu. *NonParametric Functional Data Analysis*. Springer, 2006.

[9] L. Ferré and N. Villa. Multi-layer perceptron with functional inputs: an inverse regression approach. *Scandinavian Journal of Statistics*, 33(4):807–823, 2006.

[10] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1):82–95, 1971.

[11] H.J. Lee. *Functional data analysis: classification and regression*. PhD thesis, Department of Statistics, Texas, A&M University, 2004.

[12] C. Park, J.Y. Koo, S. Kim, I. Sohn, and J.W. Lee. Classification of gene functions using support vector machine for time-course gene expression data. *Computational Statistics and Data Analysis*, 2007. Article in Press. doi:10.1016/j.csda.2007.09.002.

[13] D.L. Ragozin. Error bounds for derivative estimation based on spline smoothing of exact or noisy data. *Journal of Approximation Theory*, 37:335–355, 1983.

[14] J.O. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer Verlag, New York, 1997.

[15] F. Rossi and B. Conan-Guez. Functional multi-layer perceptron: a nonlinear tool for functional data anlysis. *Neural Networks*, 18(1):45–60, 2005.

[16] F. Rossi and N. Villa. Classification in Hilbert spaces with support vector machines. In *ASMDA 2005 proceedings*, pages 635–642, Brest, France, 2005.

[17] F. Rossi and N. Villa. Support vector machine for functional data classification. *Neurocomputing*, 69(7-9):730–742, 2006.

[18] F. Rossi and N. Villa. Consistency of derivative based functional classifiers on sampled data. 2007. Submitted.

[19] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2001.

[20] I. Steinwart. Support vector machines are universally consistent. *Journal of Complexity*, 18:768–791, 2002.

[21] F.I. Utreras. Boundary effects on convergence rates for tikhonov regularization. *Journal of Approximation Theory*, 54:235–249, 1988.

[22] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[23] N. Villa and F. Rossi. Un résultat de consistance pour des SVM fonctionnels par interpolation spline. *Comptes Rendus Mathématique. Académie des Sciences. Paris*, 343(8):555–560, 2006.