ESANN'2005 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 27-29 April 2005, d-side publi., ISBN 2-930307-05-6.

Support Vector Machine For Functional Data Classification

Nathalie Villa[1] and Fabrice Rossi[2]

1- Université Toulouse Le Mirail - Equipe GRIMM
5 allées A. Machado, 31058 Toulouse cedex 1 - FRANCE

2- Projet AxIS, INRIA, Domaine de Voluceau, Rocquencourt, B.P. 105
78153 Le Chesnay Cedex - FRANCE

Abstract. Functional data analysis is a growing research field and numerous works present a generalization of the classical statistical methods to function classification or regression. In this paper, we focus on the problem of using Support Vector Machines (SVMs) for curve discrimination. We recall that important theoretical results for SVMs apply in functional space and propose simple functional kernels that take advantage of the nature of the data. Those kernels are illustrated on a spectrometric real world benchmark.

1 Introduction

In many real-world applications, data should be considered as discretized functions rather than as standard vectors. In these applications, each observation corresponds to a mapping between some conditions (that might be implicit) and the observed response. A well studied example of those functional data [1] is given by spectrometric data (see section 5): each spectrum is a mapping that associates an response (an absorbance for instance) to a light with a given wavelength. Other natural examples can be found in meteorological problems (for instance geographic mappings between coordinates and local weather conditions) and more generally in multiple time series analysis where each observation is a complete time series.

The goal of Functional Data Analysis (FDA) is to use in data analysis algorithms the fact that the studied data are discretized functions. Many data analysis methods have been adapted to functions (see [1] for a comprehensive review of linear methods).

In the present paper, we adapt Support Vector Machines (SVM, see e.g. [2, 3]) to functional data classification. The paper is organized as follows. Section 2 presents the Functional Data Analysis and explains why it usually leads to particular problems, section 3 presents the theoretical SVM for functional data, section 4 explains what kind of problems involve in practice functional data and proposes solutions to overcome them and finally, section 5 illustrate the proposed approach on a real world benchmark.

2 Functional Data Analysis

2.1 Functional Data

To simplify the presentation, this article focuses on functional data for which each observation is described by one real valued function. Extension to the case of several real valued functions is straightforward. More formally, if μ denotes a finite positive Borel measure on $\mathbb{R}$, an observation is an element of $L^2(\mu)$ the Hilbert space of square integrable real valued functions defined on $\mathbb{R}$. The inner product in $L^2(\mu)$ is denoted $\langle .,. \rangle$ and is given by $\langle f, g \rangle = \int fg \mathrm{d}\mu$.

Our goal is to classify functional data into predefined classes. We assume given a learning set, i.e. N examples $(x_1, y_1), \ldots, (x_N, y_N)$ which are i.i.d. realizations of the random variable pair (X, Y) where X has values in $L^2(\mu)$ and Y in $\{-1, 1\}$, i.e. Y is the class label for X which is the functional data.

2.2 Data analysis methods for functional data

Most of the theoretical and practical difficulties in FDA are linked to the fact that $L^2(\mu)$ is an infinite dimensional vector space. As a consequence, some simple problems in $\mathbb{R}^d$ become ill-posed in $L^2(\mu)$, even on a theoretical point of view.

Let us consider for instance the linear regression model in which a real valued target variable U is modeled by $E(U|X) = H(X)$ where H is a linear continuous operator defined on the input space ($L^2(\mu)$ for FDA). When X has values in $\mathbb{R}^d$, H can be easily estimated thanks to least square methods that lead to the inversion of the covariance matrix of X. In practice, problems might appear when d is not small compared to N the number of available examples and regularization techniques can be used (e.g., ridge regression). When X has values in a functional space, the problem is ill-posed because the covariance operator of X is not one-to-one in general and even in particular cases were it is, it has no continuous inverse (see [4]).

To overcome the infinite dimensional problem, most of FDA methods so far have been constructed thanks to two general principles: either use representation methods that allow to work in finite dimension, or introduce regularization constraints that have comparable dimension reduction effects (see [1]). For example, [4] and [5]) develop functional models for linear regression. In the same way, lot of data analysis algorithms have been successfully adapted to functional data. This is the case, for instance, of most neural network models ([6, 7, 8, 9]).

3 Support Vector Machines for FDA

3.1 Large Margin Linear discrimination

The most basic SVM is an affine discrimination function with maximal margin. When the studied data are linearly separable, the parameters (w, b) of the SVM

are obtained by solving the following quadratic programming problem:

$$(P_0)\ \min_{w,b}\langle w,w\rangle,\ \text{subject to } y_i(\langle w,x_i\rangle+b)\geq 1,\ 1\leq i\leq N.$$

In $L^2(\mu)$, (P_0) has always a solution, as long as input functions are in general position (i.e., span a N dimensional subspace of $L^2(\mu)$). It might seem therefore that neither soft margin SVM, nor non linear kernel are needed for functional data. In practice however, it is well known, see e.g. [10], that the solution of (P_0) in high dimensional spaces is not adequate: some regularization is needed to obtain good generalization. Therefore, (P_0) is replaced by its soft margin version, i.e., by the problem:

$$\begin{aligned}(P_C)\ &\min_{w,b,\xi}\langle w,w\rangle + C\sum_{i=1}^N \xi_i,\\ &\text{subject to }\ y_i(\langle w,x_i\rangle+b)\geq 1-\xi_i,\ 1\leq i\leq N,\\ &\qquad\qquad \xi_i\geq 0,\ 1\leq i\leq N.\end{aligned}$$

3.2 Theoretical properties

The use of an infinite dimensional space in both problems might seem related to what is done in general when original data are not linearly separable and are therefore mapped into a high dimensional feature space by the use of a kernel: for some of them (e.g., the gaussian RBF kernel), this feature space has an infinite dimension. But in this case, it is a Reproducing Kernel Hilbert Space (RKHS) and has therefore some regularity properties.

In the considered setting, the original data space is $L^2(\mu)$ which is not a RKHS. Nevertheless, the most important properties of SVM are still satisfied. First of all, it is possible to replace the optimization problem (P_C) by a dual problem (D_C) in which only dot products in the feature space are needed (see [11]):

$$\begin{aligned}(D_C)\ &\min_{\alpha}\sum_{i=1}^N \alpha_i - \sum_{i=1}^N\sum_{j=1}^N \alpha_i\alpha_j y_i y_j\langle x_i,x_j\rangle,\\ &\text{subject to }\ \sum_{i=1}^N \alpha_i y_i = 0,\\ &\qquad\qquad 0\leq\alpha_i\leq C,\ 1\leq i\leq N.\end{aligned}$$

The advantage of using (D_C) rather than (P_C) is obvious: the former is solved in $\mathbb{R}^N$, whereas the latter is solved in $L^2(\mu)$. It is obviously much simpler to calculate approximate integrals (by quadrature or Monte Carlo methods) than to implement of constrained optimization in $L^2(\mu)$. In (D_C), it also appears clearly that the usual dot product of $L^2(\mu)$ can be replaced by any positive kernel defined on $L^2(\mu)$ (see section 4.2 for functional kernels).

Another important theoretical result for SVM is that looking for a large margin classifier provides good generalization properties for the obtained classifier. More precisely, the generalization performances of a SVM can be bounded by a quantity which is related to the margin, to the size of the training set and to the radius of the smallest ball containing all the training set (see theorem 4.18 in [3]). A very interesting point of this result is that it applies to any inner space product and therefore in particular to $L^2(\mu)$.

4 Functional data in practice

4.1 Observations

In practice, the functions $(x_i)_{1\leq i\leq N}$ are never perfectly known. The best situation is the one in which d discretization points have been chosen in $\mathbb{R}$, $(t_k)_{1\leq k\leq d}$ and each function x_i is described by a vector of $\mathbb{R}^d$, $(x_i(t_1),\ldots,x_i(t_d))$. In this situation, it might be tempting to apply standard data analysis methods on $\mathbb{R}^d$ vectors, but as explained in Section 3.1, this usually leads to bad solutions because d might be bigger than N and the variables are highly correlated. As we already said, the use of regularization and of special kernels, which take advantage of the function structures underline in this $\mathbb{R}^d$ vectors, can prevent this problem.

Furthermore, in some application domains, especially medical ones (e.g., [12]), each function is in general badly sampled: the number and the location of discretization points depend on the function and therefore a simple vector model is not anymore possible. A possible solution consists in constructing a approximation of x_i based on its observation values (thanks to e.g., B-splines) and then to work with the reconstructed functions (see [1, 9] for details).

4.2 Using the functional nature of the data

In the simplest situation (uniform discretization), data might simply be considered as vectors in $\mathbb{R}^d$ and standard SVM processing of those vectors could be used. Even in this situation, it is interesting to design functional kernels that use the functional nature of the data.

As explained in section 3, the linear kernel corresponds to the inner product in $L^2(\mu)$ which can be easily implemented or even approximated by the scalar product in $\mathbb{R}^d$ if the discretization is uniform. The Gaussian kernel is based on the euclidean norm in the data space and therefore also applies to functional data, again thanks to an approximate calculation of distances in $L^2(\mu)$. In fact every kernel that is defined using the Hilbert structure of $\mathbb{R}^d$ can be readily implemented in $L^2(\mu)$, either directly because the discretization is uniform or thanks to function approximation method.

Another way to define functional kernel is to use a functional pre-processing combined with a standard kernel. An interesting possibility is offered by derivation operators if the considered functions are smooth enough (in some cases, the corresponding functional space is a RKHS). Using an adapted function approximation method (such as a B-spline expansion), an estimation of $x^{(q)}$, the q-th derivative of x, can be obtained (even if the discretization is not uniform). Then any kernel can be used on the derivatives. This method allows to focus on some particular aspects of the underlying functions, such as the curvature for the second derivative. It is well known that in some application domain such as spectrometry, such kind of features might be more interesting than the original curves.

5 Application

We study in this section spectrometric data from food industry[1]. Each observation is the near infrared absorbance spectrum of a meat sample (finely chopped), recorded on a Tecator Infratec Food and Feed Analyser (we have 215 spectra). More precisely, an observation consists in a 100 channel spectrum of absorbances in the wavelength range 850–1050 nm (see figure 1). The classification problem consists in separating meat samples with a high fat content (more than 20%) from samples with a low fat content (less than 20%). The data set is split into 120 spectra for learning and 95 spectra for testing. Meta-parameters (C for the soft margin and σ for the Gaussian kernel) of the SVMs have been determined by a 10-fold cross validation procedure.

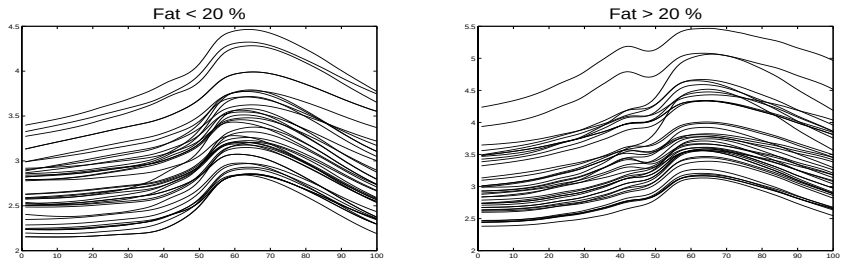

Fig. 1: Spectra for both classes

The problem is used to compare standard kernels (linear and Gaussian kernels) to a derivative based kernel. It appears on figure 1 that high fat content spectra have sometimes two local maximum rather than one: we have therefore decided to focus on the curvature of the spectra, i.e., to use the second derivative.

The following table gives the performances of the considered methods :

Kernel	Learning set error rate	Test set error rate
Linear	0.83%	2.11%
Gaussian	0%	4.21%
Linear on second derivatives	0%	0%
Gaussian on second derivatives	0.83%	1.05%

The results show that the problem is not very difficult as the worst performances (4.21%) corresponds to 4 misclassified spectra among 95. Nevertheless, it also appears that a functional transformation improves the results. The relatively bad performances of the Gaussian kernel on plain data can be explained by the fact that a direct comparison of spectra based on their $L^2(\mu)$ norm is in general dominated by the mean value of those spectra which is not a good feature for classification in spectrometric problems. The linear kernel is less sensitive to this problem. In both cases, the use of a functional kernel introduces expert knowledge (i.e., curvature is a good feature for some spectrometric problems) allows to overcome most of the limitation of the original kernel.

[1]available on statlib: `http://lib.stat.cmu.edu/datasets/tecator`

6 Conclusion

Support Vector Machines use frequently kernels that correspond to mapping original data in an infinite dimensional vector space. While these vector spaces have specific regularity properties, the most important properties of SVM (the dual formulation of the optimization problem and the link between large margin and good generalization performances) are still valid in standard functional spaces. In Functional Data Analysis, observations already live in a functional space and therefore a plain linear kernel (i.e., the inner product of the functional space) is enough in theory to classify functional data if they are in general position. In practice however, the use of adapted kernels is important to obtain good performances. We have shown for instance on real world data that functional transformations such as derivative calculation can improve the quality of classification by allowing the SVM to use more adapted features. The performances obtained are similar to the one reported in [8] and obtained in comparable experimental settings. In [8] classification was made thanks to a multi-layer perceptron that necessitates an order of magnitude more training time than the SVM used in the present paper. SVM appears therefore as a very competitive tool for function classification.

References

[1] Jim Ramsay and Bernard Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer Verlag, June 1997.

[2] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

[3] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2000.

[4] Hervé Cardot, Frédéric Ferraty, and Pascal Sarda. Functional linear model. *Statist. & Prob. Letters*, 45:11–22, 1999.

[5] Hervé Cardot, Frédéric Ferraty, and Pascal Sarda. Spline estimators for the functional linear model. *Statistica Sinica*, 13:571–591, 2003.

[6] Louis Ferré and Nathalie Villa. Multi-layer neural network with functional inputs: an inverse regression approach. *submitted*, 2004.

[7] Fabrice Rossi, Brieuc Conan-Guez, and Aïcha El Golli. Clustering functional data with the som algorithm. In *Proceedings of ESANN 2004*, pages 305–312, Bruges, Belgium, April 2004.

[8] Fabrice Rossi and Brieuc Conan-Guez. Functional multi-layer perceptron: a nonlinear tool for functional data analysis. *Neural Networks*, 18(1):45–60, January 2005.

[9] Fabrice Rossi, Nicolas Delannay, Brieuc Conan-Guez, and Michel Verleysen. Representation of functional data in neural networks. *Neurocomputing*, 64C:183–210, 2005.

[10] Trevor Hastie, Saharon Rosset, Robert Tibshirani, and Ji Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5:1391–1415, October 2004.

[11] Chih-Jen Lin. Formulations of support vector machines: a note from an optimization point of view. *Neural Computation*, 2(13):307–317, 2001.

[12] Gareth M. James and Trevor J. Hastie. Functional linear discriminant analysis of irregularly sampled curves. *Journal of the Royal Statistical Society Series B*, 63:533–550, 2001.