Journal of Multivariate Analysis 146 (2016) 269–281

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

A nonlinear aggregation type classifier

Alejandro Cholaquidis [a], Ricardo Fraiman [a,*], Juan Kalemkerian [a], Pamela Llop [b,c]

[a] Centro de Matemática, Facultad de Ciencias, Universidad de la República, Uruguay
[b] Instituto de Matemática Aplicada del Litoral (UNL - CONICET), Argentina
[c] Departamento de Matemática, Facultad de Ingeniería Química (UNL), Argentina

ARTICLE INFO

Article history:
Received 11 August 2014
Available online 9 October 2015

AMS subject classifications:
62G08
62G20
62M20

Keywords:
Functional data
Supervised classification
Non-linear aggregation

ABSTRACT

We introduce a nonlinear aggregation type classifier for functional data defined on a separable and complete metric space. The new rule is built up from a collection of M arbitrary training classifiers. If the classifiers are consistent, then so is the aggregation rule. Moreover, asymptotically the aggregation rule behaves as well as the best of the M classifiers. The results of a small simulation are reported both, for high dimensional and functional data, and a real data example is analyzed.

1. Introduction

Supervised classification is still one of the hot topics for high dimensional and functional data due to the importance of their applications and the intrinsic difficulty in a general setup. In this context, there is a vast literature on classification methods which include: linear classification, k-nearest neighbors and kernel rules, classification based on partial least squares, reproducing kernels or depth measures. Complete surveys of the literature are the works by Baíllo et al. [1], Cuevas [13] and Delaigle and Hall [14]. In the book *Contributions in infinite-dimensional statistics and related topics* [7], there are also several recent advances in supervised and unsupervised classification. See for instance, Chapters 2, 5, 22 or 48, or directly, Chapter 1 of this issue (Bongiorno et al. [6]). In this context, very recently there has been great interest to develop aggregation methods. In particular, there is a large list of linear aggregation methods like boosting (Breiman [8], Breiman [9]), random forest (Breiman [10], Biau et al. [4], Biau [3]), among others. All these methods exhibit an important improvement when combining a subset of classifiers to produce a new one. Most of the contributions to the aggregation literature have been proposed for nonparametric regression, a problem closely related to classification rules, which can be obtained just by plugging in the estimate of the regression function into the Bayes rule (see for instance, Yang [19] and Bunea et al. [11]). Model selection (select the optimal single model from a list of models), convex aggregation (search for the optimal convex combination of a given set of estimators), and linear aggregation (select the optimal linear combination of estimators) are important contributions among a large list.

* Corresponding author.
E-mail addresses: acholaquidis@cmat.edu.uy (A. Cholaquidis), rfraiman@cmat.edu.uy (R. Fraiman), jkalem@cmat.edu.uy (J. Kalemkerian), pllop@santafe-conicet.gov.ar (P. Llop).

http://dx.doi.org/10.1016/j.jmva.2015.09.022
0047-259X/

In the finite dimensional setup, Mojirsheibani [17,18] introduced a combined classifier showing strong consistency under someway hard to verify assumptions involving the Vapnik Chervonenkis dimension of the random partitions of the set of classifiers, which are non-valid in the functional setup. Very recently Biau et al. [5] introduced a new nonlinear aggregation strategy for the regression problem called COBRA, extending the ideas in Mojirsheibani [17] to the more general setup of nonparametric regression in $\mathbb{R}^d$. In the same direction but for the classification problem in the infinite dimensional setup, we extend the ideas in Mojirsheibani [17] to construct a classification rule which combines, in a nonlinear way, several classifiers to construct an optimal one. We point out that our rule allows to combine methods of very different nature, taking advantage of the abilities of each expert and allowing to adapt the method to different class of datasets. Even though our classifier allows to aggregate experts of the same nature, the possibility of combining classifiers of different characters, improves the use of existing rules as the bagged nearest neighbors classifier (see for instance Hall and Samworth [16]). As in Biau et al. [5], we also introduce a more flexible form of the rule which discards a small percentage α of those preliminary experts that behaves differently from the rest. Under very mild assumptions, we prove consistency, obtain rates of convergence and show some optimality properties of the aggregated rule. To build up this classifier, we use the inverse function (see also Fraiman et al. [15]) of each preliminary expert which makes the proposal particularly well designed for high dimensional data avoiding the curse of dimensionality. It also performs well in functional data settings.

In Section 2 we introduce the new classifier in the general context of a separable and complete metric space which combines, in a nonlinear way, the decision of M experts (classifiers). A more flexible rule is also considered. In Section 3 we state our two main results regarding consistency, rates of convergence and asymptotic optimality of the classifier. Asymptotically, the new rule performs as the best of the M classifiers used to build it up. Section 4 is devoted to show through some simulations the performance of the new classifier in high dimensional and functional data for moderate sample sizes. A real data example is also considered. All proofs are given in the Appendix.

2. The setup

Throughout the manuscript $\mathcal{F}$ will denote a separable and complete metric space, (X, Y) a random pair taking values in $\mathcal{F} \times \{0, 1\}$ and μ the probability measure of X. The elements of the training sample $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, are i.i.d. random elements with the same distribution as the pair (X, Y). The regression function is denoted by $\eta(x) = \mathbb{E}(Y|X = x) = \mathbb{P}(Y = 1|X = x)$, the Bayes rule by $g^*(x) = \mathbb{I}_{\{\eta(x)>1/2\}}$ and the optimal Bayes risk by $L^* = \mathbb{P}\big(g^*(X) \neq Y\big)$.

In order to define our classifier, we split the sample $\mathcal{D}_n$ into two subsamples $\mathcal{D}_k = \big\{(X_1, Y_1), \ldots, (X_k, Y_k)\big\}$ and $\mathcal{E}_l = \big\{(X_{k+1}, Y_{k+1}), \ldots, (X_n, Y_n)\big\}$ with $l = n - k \geq 1$. With $\mathcal{D}_k$ we build up M classifiers $g_{mk} : \mathcal{F} \to \{0, 1\}$, $m = 1, \ldots, M$ which we place in the vector $\mathbf{g_k}(x) \doteq \big(g_{1k}(x), \ldots, g_{Mk}(x)\big)$ and, following some ideas in [17], with $\mathcal{E}_l$ we construct our aggregate classifier as,

$$g_T(x) = \mathbb{I}_{\{T_n(\mathbf{g_k}(x))>1/2\}}, \tag{1}$$

where

$$T_n(\mathbf{g_k}(x)) = \sum_{j=k+1}^{n} W_{n,j}(x) Y_j, \quad x \in \mathcal{F}, \tag{2}$$

with weights $W_{n,j}(x)$ given by

$$W_{n,j}(x) = \frac{\mathbb{I}_{\{\mathbf{g_k}(x)=\mathbf{g_k}(X_j)\}}}{\sum_{i=k+1}^{n} \mathbb{I}_{\{\mathbf{g_k}(x)=\mathbf{g_k}(X_i)\}}}. \tag{3}$$

Here, $0/0$ is assumed to be 0. Like in [5], for $0 \leq \alpha < 1$ a more flexible version of the classifier, called $g_T(x, \alpha)$, can be defined replacing the weights in (3) by

$$W_{n,j}(x) = \frac{\mathbb{I}_{\left\{\frac{1}{M}\sum_{m=1}^{M} \mathbb{I}_{\{g_{mk}(x)=g_{mk}(X_j)\}} \geq 1-\alpha\right\}}}{\sum_{i=k+1}^{n} \mathbb{I}_{\left\{\frac{1}{M}\sum_{m=1}^{M} \mathbb{I}_{\{g_{mk}(x)=g_{mk}(X_i)\}} \geq 1-\alpha\right\}}}. \tag{4}$$

More precisely, the more flexible version of the classifier (1) is given by

$$g_T(x, \alpha) = \mathbb{I}_{\{T_n(\mathbf{g_k}(x),\alpha)>1/2\}}, \tag{5}$$

where $T_n(\mathbf{g_k}(x), \alpha)$ is defined as in (2) but with the weights given by (4). Observe that if we choose $\alpha = 0$ in (4) and (5), we obtain the weights given in (3) and the classifier (1) respectively.