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.

© 2015 Elsevier Inc. All rights reserved.

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/© 2015 Elsevier Inc. All rights reserved.

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.

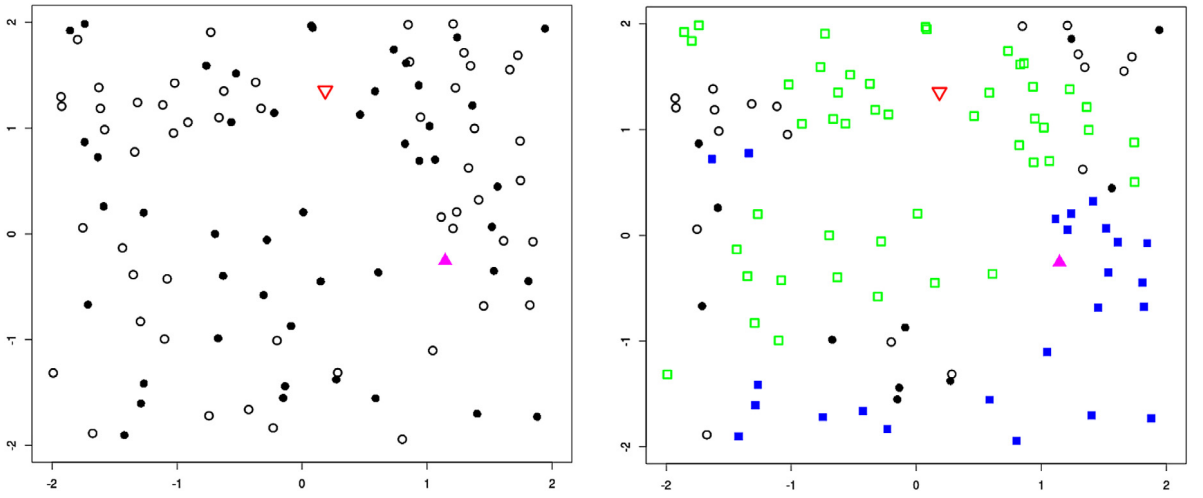


Fig. 1. Left: sample points corresponding to two populations (black filled and empty circles) and two points to classify (red empty and magenta filled triangles). Right: in empty green squares the voters for the red empty triangle and in filled blue squares the voters for the magenta filled triangle. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Remark 1. (a) The type of nonlinear aggregation used to define our classifiers turns out to be quite natural. Indeed, we give a weight different from zero to those X_j which classify x in the same group as the whole set of classifiers $\mathbf{g_k}(X_j)$ (or $100(1-\alpha)\%$ of them).

(b) Since we are using the inverse functions of the classifiers g_{mk}, observations which are far from x for which the condition mentioned in (a) is fulfilled are involved in the definition of the classification rule. This may be very important in the case of high dimensional data to avoid the curse of dimensionality. This is illustrated in Fig. 1, where we show two samples of points: one uniformly distributed in the square $[-2, 2] \times [-2, 2]$ (filled black points) and another uniformly distributed in the L_∞-ring $[-2, 2] \times [-1, 1]$ (empty black points). We also show two points to classify, the empty red and the filled magenta triangles together with their corresponding voters, empty green squares and filled blue squares, respectively. As we can see, observations that are far from the triangles are also involved in the classification.

3. Asymptotic results

In this section we show two asymptotic results for the nonlinear aggregation classifier The first one shows that the classifier $g_T(X, \alpha)$ is consistent if, for $0 \leq \alpha < 0.5$, at least $R \geq (1-\alpha)M$ of them are consistent. Moreover, rates of convergence for $g_T(X, \alpha)$ (and $g_T(X)$) are obtained assuming we know the rates of convergence of the R consistent experts. The second result, shows that $g_T(X)$ behaves asymptotically as the best of the M classifiers used to build it up. Both results are proved under mild conditions. Throughout this section we will use the notation $\mathbb{P}_{\mathcal{D}_k}(\cdot) = \mathbb{P}(\cdot|\mathcal{D}_k)$.

Theorem 1. *Assume that, for every $m = 1, \ldots, R$, the classifier g_{mk} converges in probability to g^* as $k \to \infty$, with $R \geq M(1-\alpha)$ and $\alpha \in [0, 1/2)$. Let us assume that $\mathbb{P}(Y = 1|g^*(X) = 1) > 1/2$ and $\mathbb{P}(Y = 0|g^*(X) = 0) > 1/2$, then*

(a) $\lim_{\min\{k,l\}\to\infty} \mathbb{P}_{\mathcal{D}_k}(g_T(X, \alpha) \neq Y) - L^* = 0$.

(b) *Let $\beta_{mk} \to 0$ as $k \to \infty$, for $m = 1, \ldots, R$ and $\boldsymbol{\beta_{Rk}} = \max_{m=1,\ldots,R} \beta_{mk}$. If $\mathbb{P}_{\mathcal{D}_k}\big(g^*(X) \neq g_{mk}(X)\big) = \mathcal{O}(\beta_{mk})$, then, for k large enough,*

$$\mathbb{P}_{\mathcal{D}_k}(g_T(X, \alpha) \neq Y) - L^* = \mathcal{O}\Big(\max\big\{\exp(-Cl), \boldsymbol{\beta_{Rk}}\big\}\Big), \tag{6}$$

for some constant $C > 0$.

Remark 2. (a) The assumption

$$(1)\ \mathbb{P}(Y = 1|g^*(X) = 1) > 1/2 \qquad (2)\ \mathbb{P}(Y = 0|g^*(X) = 0) > 1/2, \tag{7}$$

is really mild. It just requires that if the Bayes rule $g^*(X)$ takes the value 1 (or 0) the probability that $Y = 1$ is greater than the probability that $Y = 0$ (the probability that $Y = 0$ is greater than the probability that $Y = 1$). Moreover since the Bayes risk $L^* \leq 1/2$ one of the conditions in (7) is always fulfilled.

(b) It is well known that in the finite dimensional case, if the regression function η verifies a Lipschitz condition and X is bounded supported, the accuracy of classical classification rules is $\mathcal{O}(n^{-2/(d+2)})$. Therefore the right hand side of (6) is

$$\mathcal{O}\Big(\max\big\{\exp(-Cl), k^{-2/(d+2)}\big\}\Big),$$

and the optimal rate for $\max\big\{\exp(-Cl), k^{-2/(d+2)}\big\}$ is attained for $l \sim \log(k)$.

(c) The choice of the parameters α and l is an important issue. From a practical point of view, we suggest to perform a cross validation procedure to select the values of the corresponding parameters. See Section 5 for an implementation in a real data example.

In order to state the optimality result we introduce some additional notation. Let $\mathbb{C} \doteq \{0, 1\}^M$ and let us call $\nu \in \mathbb{C}$. Calling $\nu(m)$ the mth entry of the vector ν, we define the following subsets

$$A_\nu^0 \doteq \bigcap_{m=1}^{M} g_{mk}^{-1}(\nu(m)) \times \{0\}, \qquad A_\nu^1 \doteq \bigcap_{m=1}^{M} g_{mk}^{-1}(\nu(m)) \times \{1\},$$

and $\quad A_\nu = A_\nu^0 \cup A_\nu^1$.

For each $\nu \in \mathbb{C}$, we consider the assumption:

$$(\mathcal{H})\ H(\mathcal{D}_k) := \mathbb{P}_{\mathcal{D}_k}\big((X, Y) \in A_\nu^1\big) - \mathbb{P}_{\mathcal{D}_k}\big((X, Y) \in A_\nu^0\big) \neq 0 \quad \text{a.s.}$$

Theorem 2. (1) *For each $m = 1, \ldots, M$,*

$$\mathbb{P}_{\mathcal{D}_k}\big(g_T(X) \neq Y\big) - \mathbb{P}_{\mathcal{D}_k}\big(g_{mk}(X) \neq Y\big) \leq \mathcal{O}_k\big(l^{-1/2}\big),$$

which implies that,

$$\lim_{l\to\infty} \mathbb{P}_{\mathcal{D}_k}\big(g_T(X) \neq Y\big) \leq \min_{1\leq m\leq M} \mathbb{P}_{\mathcal{D}_k}\big(g_{mk}(X) \neq Y\big).$$

(2) *Under assumption $(\mathcal{H})$ we obtain a better approximation rate,*

$$\mathbb{P}_{\mathcal{D}_k}\big(g_T(X) \neq Y\big) - \mathbb{P}_{\mathcal{D}_k}\big(g_{mk}(X) \neq Y\big) \leq \mathcal{O}_k\big(\exp(-K_1 l)\big).$$

4. A small simulation study

In this section we present the performance of the aggregated classifier in two different scenarios. The first one corresponds to high dimensional data while, in the second one, we consider two simulated models for functional data analyzed in Delaigle and Hall [14].

High dimensional setting

In this setting we show the performance of our method by analyzing data generated in $\mathbb{R}^{150}$ in the following way: we generate $n + 200$ i.i.d. uniform random variables in $[0, 1]$, say $Z_1, \ldots, Z_{n+200}$. For each $i = 1, \ldots, n + 200$, if $Z_i > 1/4$, we generate a random variable $X_i \in \mathbb{R}^{150}$ with uniform distribution in $[-2, 2]^{150}$ and set $Y_i = 1$. If $Z_i \leq 1/4$, we generate a random variable $X_i \in \mathbb{R}^{150}$ with uniform distribution in $\tau_v([-2, 2]^{150})$ where τ_v is the translation along the direction $(v, \ldots, v) \in \mathbb{R}^{150}$ for $v = 1/4$ and set $Y_i = 0$. Then we split the sample into two subsamples: with the first n pairs (X_i, Y_i), we build the training sample, with the remaining 200 we build the testing sample. We consider two cases: the homogeneous case, where we aggregate classifiers of the same nature and in the heterogeneous case, where we aggregate experts of different nature.

- Homogeneous case: M k-nearest neighbor classifiers with the number of neighbors taken as follows:
  1. we fix $M = 8$ consecutive odd numbers;
  2. we choose at random $M = 10$ different odd integers between 1 and $\min\{\sum_{i=1}^{k} Y_i, k - \sum_{i=1}^{k} Y_i\}$.

In Table 1, we report the mean and standard deviation (in brackets) of the misclassification error rate for case 1, when compared with the nearest neighbor rules build up with a sample size n taking 5,7,9,11,13,15,17,19 nearest neighbors (these classifiers are denoted by g_{mn} for $m = 1, \ldots, 8$). In Table 2 we report the median and MAD (in brackets) of the misclassification error rate for this case.

In Table 3 we report the mean of the misclassification error rate and standard deviation for case 2, with the original aggregated classifier and the two more flexible versions: $\alpha = 1/8$ and $\alpha = 1/4$. In this table we compare the performance of our rules with the (optimal) cross validated nearest neighbor classifier computed with k and also with n. In Table 4 we report the median and MAD of the misclassification error rate for this case.

- Heterogeneous case: $M = 5$ classifiers: 3 k-nearest neighbor rules with fixed values of k, the Fisher and the random forest classifiers.

Table 1
Mean and standard deviation of the misclassification error rate over 500 replicates for $\mathbb{R}^{150}$ with fixed number of neighbors.

n/k	$g_T(\cdot)$	g_{1n}	g_{2n}	g_{3n}	g_{4n}	g_{5n}	g_{6n}	g_{7n}	g_{8n}
400/300	**0.027** (0.014)	0.045 (0.016)	0.043 (0.017)	0.042 (0.017)	0.042 (0.018)	0.043 (0.018)	0.043 (0.018)	0.043 (0.019)	0.044 (0.019)
600/400	**0.023** (0.012)	0.039 (0.015)	0.036 (0.016)	0.035 (0.015)	0.035 (0.015)	0.035 (0.015)	0.036 (0.015)	0.037 (0.016)	0.037 (0.016)
800/600	**0.020** (0.010)	0.037 (0.014)	0.034 (0.013)	0.033 (0.013)	0.033 (0.013)	0.033 (0.013)	0.033 (0.013)	0.033 (0.013)	0.033 (0.014)

Table 2
Median and MAD of the misclassification error rate over 500 replicates for $\mathbb{R}^{150}$ with fixed number of neighbors.

n/k	$g_T(\cdot)$	g_{1n}	g_{2n}	g_{3n}	g_{4n}	g_{5n}	g_{6n}	g_{7n}	g_{8n}
400/300	**0.025** (0.015)	0.045 (0.015)	0.040 (0.015)	0.040 (0.015)	0.040 (0.015)	0.040 (0.015)	0.040 (0.015)	0.040 (0.015)	0.040 (0.015)
600/400	**0.020** (0.015)	0.035 (0.015)	0.035 (0.015)	0.035 (0.015)	0.035 (0.015)	0.035 (0.015)	0.035 (0.015)	0.035 (0.015)	0.035 (0.015)
800/600	**0.020** (0.007)	0.035 (0.015)	0.035 (0.015)	0.030 (0.015)	0.030 (0.015)	0.030 (0.015)	0.030 (0.015)	0.032 (0.011)	0.030 (0.015)

Table 3
Mean and standard deviation of the misclassification error rate over 500 replicates for $\mathbb{R}^{150}$ with the number of neighbors chosen at random.

n/k	$g_T(\cdot)$	$g_T(\cdot, 1/8)$	$g_T(\cdot, 1/4)$	gcv_n	gcv_k
400/300	**0.029** (0.016)	0.038 (0.019)	0.046 (0.021)	0.040 (0.017)	0.044 (0.018)
600/400	**0.029** (0.016)	0.039 (0.019)	0.047 (0.022)	0.037 (0.016)	0.043 (0.018)
800/600	**0.027** (0.014)	0.036 (0.018)	0.046 (0.020)	0.033 (0.014)	0.036 (0.015)

Table 4
Median and MAD of the misclassification error rate over 500 replicates for $\mathbb{R}^{150}$ with the number of neighbors chosen at random.

n/k	$g_T(\cdot)$	$g_T(\cdot, 1/8)$	$g_T(\cdot, 1/4)$	gcv_n	gcv_k
400/300	**0.025** (0.015)	0.035 (0.015)	0.045 (0.022)	0.040 (0.015)	0.042 (0.019)
600/400	**0.028** (0.019)	0.035 (0.015)	0.045 (0.022)	0.035 (0.015)	0.040 (0.015)
800/600	**0.025** (0.015)	0.035 (0.015)	0.045 (0.022)	0.035 (0.015)	0.035 (0.015)

Here we take 3, 5, 7 nearest neighbors (denoted by g_{mn} for $m = 1, 2, 3$), the Fisher classifier (denoted by g_F) and the random forest classifier (denoted by g_{RF}). In Table 5 we report the averaged misclassification error rates and standard deviation and in Table 6 we report the median and MAD for this case.

Functional data setting

In this setting we show the performance of our method by analyzing the following two models considered in Delaigle and Hall [14]:

- Model I: We generate two samples of size $n/2$ from different populations following the model

$$X_{pi}(t) = \sum_{j=1}^{6} \mu_{p,j}\phi_j(t) + e_{pi}(t), \quad p = 1, 2,\ i = 1, \ldots, n/2,$$

Table 5
Mean and standard deviation of the misclassification error rate over 500 replicates for $\mathbb{R}^{150}$ with fixed number of neighbors, Fisher classifier and random forest.

n/k	$g_T(\cdot)$	g_{1n}	g_{2n}	g_{3n}	g_F	g_{RF}
400/300	0.012 (0.011)	0.049 (0.016)	0.043 (0.017)	0.041 (0.017)	0.020 (0.011)	**0.004** (0.004)
600/400	0.008 (0.007)	0.047 (0.015)	0.040 (0.015)	0.037 (0.015)	0.012 (0.008)	**0.001** (0.002)
800/600	0.007 (0.007)	0.043 (0.015)	0.036 (0.015)	0.034 (0.014)	0.009 (0.007)	**0.000** (0.002)

Table 6
Median and MAD of the misclassification error rate over 500 replicates for $\mathbb{R}^{150}$ with fixed number of neighbors, Fisher classifier and random forest.

n/k	$g_T(\cdot)$	g_{1n}	g_{2n}	g_{3n}	g_F	g_{RF}
400/300	0.010 (0.007)	0.050 (0.015)	0.040 (0.015)	0.040 (0.015)	0.020 (0.015)	**0.000** (0.000)
600/400	0.005 (0.007)	0.045 (0.015)	0.040 (0.015)	0.035 (0.015)	0.010 (0.007)	**0.000** (0.000)
800/600	0.005 (0.007)	0.040 (0.015)	0.035 (0.015)	0.035 (0.015)	0.010 (0.007)	**0.000** (0.000)

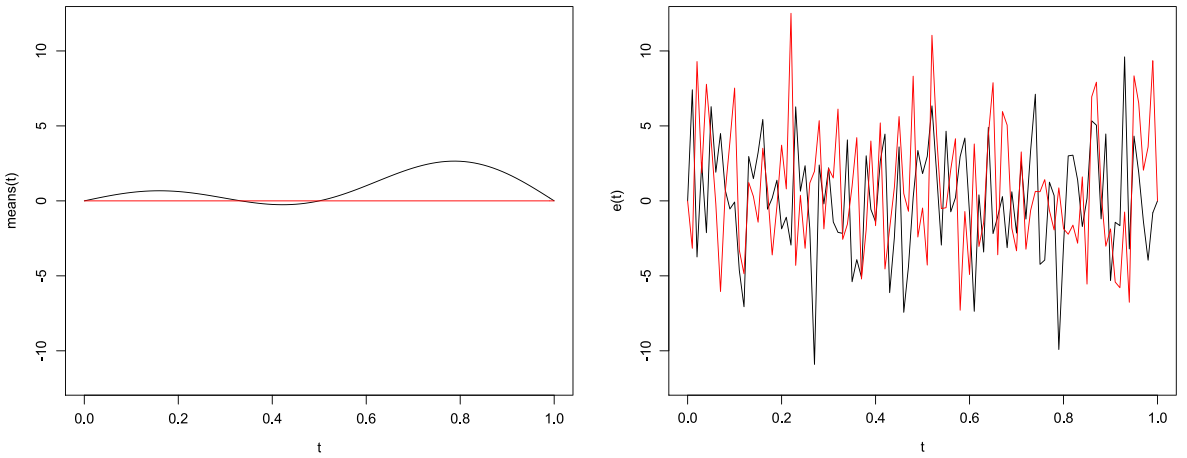


Fig. 2. Mean curve (Left) and Error curve (Right) of the two populations of Model II.

where $\phi_j(t) = \sqrt{2}\sin(\pi jt)$, $\mu_{1,j}$ and $\mu_{2,j}$ are, respectively, the jth coordinate of the mean vectors $\mu_1 = (0, -0.5, 1, -0.5, 1, -0.5)$, and $\mu_2 = (0, -0.75, 0.75, -0.15, 1.4, 0.1)$ while the errors are given by

$$e_{pi}(t) = \sum_{j=1}^{40} \sqrt{\theta_j} Z_{pj} \phi_j(t), \quad p = 1, 2,$$

with $Z_{pj} \sim \mathcal{N}(0, 1)$ and $\theta_j = 1/j^2$.

- Model II: We generate two samples of size $n/2$ from different populations following the model

$$X_{pi}(t) = \sum_{j=1}^{3} \mu_{p,j} \phi_j(t) + e_{pi}(t), \quad p = 1, 2,\ i = 1, \ldots, n/2,$$

where $\mu_1 = 0.75 \cdot (1, -1, 1)$ and $\mu_{2,j}$ the jth coordinate of $\mu_2 \equiv 0$, $\theta_j = 1/j^2$ and the errors are given by

$$e_{pi}(t) = \sum_{j=1}^{40} \sqrt{\theta_j} Z_{pj} \phi_j(t), \quad p = 1, 2,$$

with $Z_{pj} \sim \mathcal{N}(0, 1)$ and $\theta_j = \exp\{-(2.1 - (j-1)/20)^2\}$.

Table 7
Mean and standard deviation of the misclassification error rate over 200 replicates for models I and II.

Model	$g_T(\cdot)$	$g_T(\cdot, 1/5)$	$g_T(\cdot, 2/5)$	$g_T(\cdot, 3/5)$	g_{1n}	g_{2n}	g_{3n}	g_{4n}	g_{5n}
I	0.013 (0.011)	0.005 (0.005)	0.004 (0.005)	0.005 (0.006)	0.017 (0.011)	0.007 (0.007)	0.004 (0.004)	0.003 (0.004)	**0.002** (0.003)
II	0.110 (0.029)	0.074 (0.019)	0.068 (0.018)	0.069 (0.018)	0.124 (0.030)	0.083 (0.020)	0.070 (0.018)	0.066 (0.016)	**0.064** (0.017)

Table 8
Median and MAD of the misclassification error rate over 200 replicates for models I and II.

Model	$g_T(\cdot)$	$g_T(\cdot, 1/5)$	$g_T(\cdot, 2/5)$	$g_T(\cdot, 3/5)$	g_{1n}	g_{2n}	g_{3n}	g_{4n}	g_{5n}
I	0.010 (0.007)	0.005 (0.007)	0.004 (0.007)	0.005 (0.007)	0.015 (0.007)	0.005 (0.007)	**0.000** (0.000)	**0.000** (0.000)	**0.000** (0.000)
II	0.105 (0.030)	0.070 (0.015)	**0.065** (0.015)	0.070 (0.015)	0.120 (0.030)	0.080 (0.022)	0.070 (0.022)	**0.065** (0.015)	**0.065** (0.015)

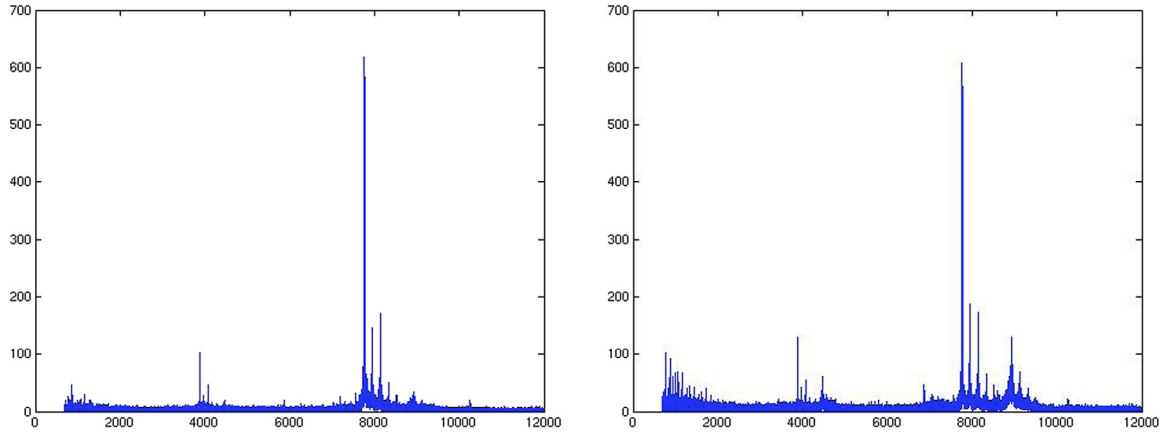


Fig. 3. Spectrogram of a healthy (left panel) and a ovarian cancer suffering woman (right panel).

This second model looks more challenging since although the means of the two populations are quite different, the error process is very wiggly, concentrated in high frequencies (as shown in Fig. 2 left and right panels, respectively). So in this case, in order to apply our classification method, we have first performed the Nadaraya–Watson kernel smoother (taking a normal kernel) to the training sample with different values of the bandwidths for each of the two populations. The values for the bandwidths were chosen via cross-validation with our classifier, varying the bandwidths between 0.1 and 0.7 (in intervals of length 0.05). The optimal values, over 200 replicates, were $h_1 = 0.15$ for the first population (with mean μ_1) and $h_2 = 0.7$ for the second one. Finally, we apply the classification method to the raw (non-smoothed) curves of the testing sample.

In Table 7 we report the averaged misclassification error rate and the standard deviation over 200 replications for models I and II, taking $n = 90$, $k = 60$, $l = 30$, and $\alpha = 0, 1, 2, 3$. In the whole training sample (of n functions) the $n/2$ labels for every population were chosen at random. The test sample consists of 200 data, taking 100 of every population. Here, $g_{mn} = (2m - 1)$-nearest neighbor rule for $m = 1, \ldots, 5$. In Table 8 we report the median of the misclassification error rate and the MAD. For Model I we get a better performance than the PLS-Centroid Classifier proposed by Delaigle and Hall [14]. For model II PLS-Centroid Classifier clearly outperforms our classifier although we get a quite small misclassification error, just using a combination of five nearest neighbor estimates.

5. A real data example: analysis of spectrograms

The data to be analyzed in this section consists in the mass spectra from blood samples of 216 women of which, 121 suffer from an ovarian cancer condition and the remaining 95 are healthy women which were taken as control group. We refer to [2] for a previous analysis of these data with a detailed discussion of their medical aspects, see also [12] for further statistical analysis of these data (see Fig. 3).

A spectrogram is a curve showing the number of molecules (or fragments) found for every mass/charge ratio and, the idea behind spectrograms, is to control the amount of proteins produced in cells since, when cancer starts to grow, its cells

produce a different kind of proteins than those produced by healthy cells. Moreover, the amount of common produced proteins may be different. Proteomics, broadly speaking, consists of a family of procedures allowing researchers to analyze proteins. In particular, here we are interested in some techniques which allow to separate mixtures of complex molecules according to the rate mass/charge (observe that, molecules with the same mass/charge ratio are indistinguishable with a spectrogram).

We have processed the data as follows: we have restricted ourselves to the interval mass charge (horizontal axis) [7000, 9500]. Then, in order to have all the spectra defined in a common equi-spaced grid, we have smoothed them via a Nadaraya–Watson smoother. Finally, every function has been divided by its maximum, in order to have all the values scaled in the common interval [0, 1]. Observe that our interest is to find the location of maxima amount of molecules more than the corresponding heights.

To build the classifier introduced in (5) we have taken 5 nearest neighbor classifiers, with $k = 3, 5, 7, 9$ neighbors. We have implemented the cross validation method in a grid for (α, l), with α taking the values 0, 1/5, 2/5 and l taking 60 values $l = 20, 21, \ldots, 80$. The minimum of the misclassification error was attained for $\alpha = 2$ and $l = 31, \ldots, 36$ in whose case the accuracy obtained was 95%.

6. Concluding remarks

- We introduce a new nonlinear aggregating method for supervised classification in a general setup built up from a family of classifiers $g_{1k}, \ldots, g_{Mk}$. It combines the decision of the M experts according to a "coincidence opinion" with respect to the new data we want to classify.
- The new method, besides being easy to implement, is particularly well designed for high dimensional and functional data. The method is not local, and the use of the inverse functions prevent from the curse of dimensionality that suffers all local methods.
- We obtain consistency and rates of convergence under very mild conditions on a general metric space setup.
- An optimality result is obtained in the sense that the nonlinear aggregation rule behaves asymptotically as well as the best one among the M classifiers (experts) $g_{1k}, \ldots, g_{Mk}$.
- A small simulation study confirms the asymptotic results for moderate sample sizes. In particular it is very well behaved for high-dimensional and functional data.
- In a well known spectrogram curves dataset, we obtain a very good performance, classifying 95%, very close to the best known results for these data.
- Although we have implemented cross validation to choose the parameters (α, l) in Section 5, conditions for the validity of this procedure remain as an open problem.

Acknowledgments

We would like to thank Gerard Biau and James Malley for helpful suggestions. We also thank the referees for their helpful suggestions which improved the presentation of this final version.

Appendix. Proof of results

To prove Theorem 1 we will need the following lemma.

Lemma 3. *Let $f(x)$ be a classifier built up from the training sample $\mathcal{D}_k$ such that $\mathbb{P}_{\mathcal{D}_k}(f(X) \neq g^*(X)) \to 0$ when $k \to \infty$. Then, $\mathbb{P}_{\mathcal{D}_k}(f(X) \neq Y) - L^* \to 0$.*

Proof of Lemma 3. First we write,

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}_k}\big(f(X) \neq Y\big) - L^* &= \mathbb{P}_{\mathcal{D}_k}\big(f(X) \neq Y\big) - P\big(g^*(X) \neq Y\big) \\
&= \mathbb{P}_{\mathcal{D}_k}\big(f(X) \neq Y, Y = g^*(X)\big) + \mathbb{P}_{\mathcal{D}_k}\big(f(X) \neq Y, Y \neq g^*(X)\big) - P\big(g^*(X) \neq Y\big) \\
&= \mathbb{P}_{\mathcal{D}_k}\big(f(X) \neq g^*(X)\big) + \mathbb{P}_{\mathcal{D}_k}\big(f(X) \neq Y, Y \neq g^*(X)\big) - P\big(g^*(X) \neq Y\big) \\
&= \mathbb{P}_{\mathcal{D}_k}\big(f(X) \neq g^*(X)\big) - \mathbb{P}_{\mathcal{D}_k}\big(g^*(X) \neq Y, f(X) = Y\big), \qquad (8)
\end{aligned}$$

where in the last equality we have used that

$$P\big(g^*(X) \neq Y\big) = \mathbb{P}_{\mathcal{D}_k}\big(g^*(X) \neq Y, f(X) \neq Y\big) + \mathbb{P}_{\mathcal{D}_k}\big(g^*(X) \neq Y, f(X) = Y\big),$$

implies

$$\mathbb{P}_{\mathcal{D}_k}\big(g^*(X) \neq Y, f(X) = Y\big) = \mathbb{P}_{\mathcal{D}_k}\big(g^*(X) \neq Y, f(X) \neq Y\big) - P\big(g^*(X) \neq Y\big).$$

Therefore, replacing in (8) we get that

$$\mathbb{P}_{\mathcal{D}_k}\big(f(X) \neq Y\big) - L^* = \mathbb{P}_{\mathcal{D}_k}\big(f(X) \neq g^*(X)\big) - \mathbb{P}_{\mathcal{D}_k}\big(g^*(X) \neq Y, f(X) = Y\big) \leq \mathbb{P}_{\mathcal{D}_k}\big(f(X) \neq g^*(X)\big), \tag{9}$$

which by hypothesis converges to zero as $k \to \infty$ and the lemma is proved. □

Proof of Theorem 1. We will prove part (b) of the Theorem since part (a) is a direct consequence of it. By (9), it suffices to prove that, for k large enough:

$$\mathbb{P}_{\mathcal{D}_k}(g_T(X,\alpha) \neq g^*(X)) = \mathcal{O}\Big(\max\big\{\exp(-C(n-k)), \boldsymbol{\beta_{Rk}}\big\}\Big).$$

We first split $\mathbb{P}_{\mathcal{D}_k}(g_T(X,\alpha) \neq g^*(X))$ into two terms,

$$\mathbb{P}_{\mathcal{D}_k}(g_T(X,\alpha) \neq g^*(X)) = \mathbb{P}_{\mathcal{D}_k}(g_T(X,\alpha) \neq g^*(X), g^*(X) = 1) + \mathbb{P}_{\mathcal{D}_k}(g_T(X,\alpha) \neq g^*(X), g^*(X) = 0) \doteq I + II.$$

Then we will prove that, for k large enough,

$$I = \mathcal{O}\Big(\max\big\{\exp(-C_1(n-k)), \boldsymbol{\beta_{Rk}}\big\}\Big),$$

for some arbitrary constant C_1. The proof that

$$II = \mathcal{O}\Big(\max\big\{\exp(-C_2(n-k)), \boldsymbol{\beta_{Rk}}\big\}\Big),$$

for some arbitrary constant C_2 is completely analogous and we omit it. Finally, taking $C = \min\{C_1, C_2\}$, the proof will be completed. In order to deal with term I, let us define the vectors

$$\mathbf{g_{Rk}}(X) = \big(g_{1k}(X), \ldots, g_{Rk}(X)\big) \in \{0,1\}^R,$$
$$\nu(X) = \Big(1, \ldots, 1, g_{(R+1)k}(X), \ldots, g_{Mk}(X)\Big) \in \{0,1\}^M.$$

Then,

$$\begin{aligned} I &= \mathbb{P}_{\mathcal{D}_k}(g_T(X,\alpha) \neq g^*(X), g^*(X) = 1) \\ &\leq \mathbb{P}_{\mathcal{D}_k}(g_T(X,\alpha) \neq g^*(X), g^*(X) = 1, \mathbf{g_{Rk}}(X) = \mathbf{1}) + \sum_{m=1}^{R} \mathbb{P}_{\mathcal{D}_k}(g_T(X,\alpha) \neq g^*(X), g^*(X) = 1, g_{mk}(X) = 0) \\ &\leq \mathbb{P}_{\mathcal{D}_k}(g_T(X,\alpha) \neq g^*(X), g^*(X) = 1, \mathbf{g_{Rk}}(X) = \mathbf{1}) + \sum_{m=1}^{R} \mathbb{P}_{\mathcal{D}_k}(g^*(X) \neq g_{mk}(X)) \\ &\leq \mathbb{P}_{\mathcal{D}_k}\Big(T_n(\mathbf{g_k}(X),\alpha) \leq 1/2 \Big| g^*(X) = 1, \mathbf{g_{Rk}}(X) = \mathbf{1}\Big) + \sum_{m=1}^{R} \mathbb{P}_{\mathcal{D}_k}(g^*(X) \neq g_{mk}(X)) \\ &\doteq I_A + I_B. \end{aligned}$$

Observe that, conditioning to $\mathbf{g_{Rk}}(X) = \mathbf{1}$ and defining

$$Z_j \doteq \mathbb{I}_{\left\{\frac{1}{M}\sum_{m=1}^{M} \mathbb{I}_{\{g_{mk}(X_j) = \nu(m)\}} \geq 1-\alpha\right\}},$$

we can rewrite $T_n(\mathbf{g_k}(X),\alpha)$ as

$$T_n(\mathbf{g_k}(X),\alpha) = \frac{\sum_{j=k+1}^{n} Z_j Y_j}{\sum_{i=k+1}^{n} Z_i}.$$

Therefore,

$$\begin{aligned} I_A &= \mathbb{P}_{\mathcal{D}_k}\left(\frac{\frac{1}{n-k}\sum_{j=k+1}^{n} Z_j Y_j}{\frac{1}{n-k}\sum_{i=k+1}^{n} Z_i} \leq \frac{1}{2} \Big| g^*(X) = 1, \mathbf{g_{Rk}}(X) = \mathbf{1}\right) \\ &= \mathbb{P}_{\mathcal{D}_k}\left(\frac{1}{n-k}\sum_{j=k+1}^{n} Z_j(Y_j - 1/2) \leq 0 \Big| g^*(X) = 1, \mathbf{g_{Rk}}(X) = \mathbf{1}\right). \end{aligned} \tag{10}$$

In order to use a concentration inequality to bound this probability, we need to compute the expectation of $Z_j(Y_j - 1/2) = Z_jY_j - Z_j/2$. To do this, observe that

$$E(Z_jY_j) = \mathbb{P}_{\mathcal{D}_k}\left(\frac{1}{M}\sum_{m=1}^{M}\mathbb{I}_{\{g_{mk}(X)=\nu(m)\}} \geq 1-\alpha, Y=1\right),$$

and

$$E(Z_j) = \mathbb{P}_{\mathcal{D}_k}\left(\frac{1}{M}\sum_{m=1}^{M}\mathbb{I}_{\{g_{mk}(X)=\nu(m)\}} \geq 1-\alpha\right).$$

Since

$$\{\mathbf{g_{Rk}}(X) = \mathbf{1}\} \subset \left\{\frac{1}{M}\sum_{m=1}^{M}\mathbb{I}_{\{g_{mk}(X)=\nu(m)\}} \geq 1-\alpha\right\} \doteq A_\alpha,$$

we have,

$$\begin{aligned} E(Z_jY_j) - E(Z_j)/2 &= \mathbb{P}_{\mathcal{D}_k}(V=1|A_\alpha)\mathbb{P}_{\mathcal{D}_k}(A_\alpha) - \mathbb{P}_{\mathcal{D}_k}(A_\alpha)/2 \\ &= \mathbb{P}_{\mathcal{D}_k}(A_\alpha)\Big[\mathbb{P}_{\mathcal{D}_k}(Y=1|A_\alpha) - 1/2\Big] \\ &\geq \mathbb{P}_{\mathcal{D}_k}\big(\mathbf{g_{Rk}}(X)=1\big)\Big[\mathbb{P}_{\mathcal{D}_k}(Y=1|A_\alpha) - 1/2\Big]. \end{aligned} \tag{11}$$

Now, since for $m = 1, \ldots, R$, $g_{mk} \to g^*$ in probability as $k \to \infty$,

$$\mathbb{P}_{\mathcal{D}_k}\left(\mathbf{g_{Rk}}(X) = \mathbf{1}\right) \to \mathbb{P}(g^*(X)=1) \doteq p^* > 0. \tag{12}$$

On the other hand, we have that, for k large enough, $\mathbb{P}_{\mathcal{D}_k}(Y=1|A_\alpha) > 1/2$. Indeed, for $m = 1, \ldots, R$, let us consider the events $B_{mk} = \{g_{mk}(X) = g^*(X)\}$ which, by hypothesis, for k large enough verify

$$\mathbb{P}\big(\cap_{m=1}^{R} B_{mk}\big) > 1-\varepsilon,$$

for all $\varepsilon > 0$. In particular, we can take $\varepsilon > 0$ such that $\mathbb{P}(Y=1|g^*(X)=1)(1-\varepsilon) > 1/2$. This implies that

$$\begin{aligned} \mathbb{P}_{\mathcal{D}_k}(Y=1|A_\alpha) &= \frac{\mathbb{P}_{\mathcal{D}_k}(Y=1, A_\alpha, \cap_{m=1}^{R} B_{mk})}{\mathbb{P}_{\mathcal{D}_k}(A_\alpha)} + \frac{\mathbb{P}_{\mathcal{D}_k}(Y=1, A_\alpha, (\cap_{m=1}^{R} B_{mk})^c)}{\mathbb{P}_{\mathcal{D}_k}(A_\alpha)} \\ &\geq \frac{\mathbb{P}_{\mathcal{D}_k}(Y=1, A_\alpha, \cap_{m=1}^{R} B_{mk})}{\mathbb{P}_{\mathcal{D}_k}(A_\alpha)} \\ &> \frac{\mathbb{P}_{\mathcal{D}_k}(Y=1, A_\alpha \big| \cap_{m=1}^{R} B_{mk})}{\mathbb{P}_{\mathcal{D}_k}(A_\alpha)}(1-\varepsilon). \end{aligned} \tag{13}$$

Conditioning to $\cap_{m=1}^{R} B_{mk}$ the event A_α equals C_α given by

$$\left\{R\mathbb{I}_{\{g^*(X)=1\}} + \sum_{m=R+1}^{M}\mathbb{I}_{\{g_{mk}(X)=\nu(m)\}} \geq M(1-\alpha)\right\} \doteq C_\alpha. \tag{14}$$

However, $\alpha < 1/2$ implies that $C_\alpha = \{g^*(X) = 1\}$. Indeed, from the inequality $R \geq M(1-\alpha)$, it is clear that $\{g^*(X)=1\} \subset C_\alpha$. On the other hand, $R \geq M(1-\alpha) > M/2$ and $\alpha < 1/2$ imply that $M - R < M/2 < M(1-\alpha)$, and so the sum in the second term of (14) is at most $M-R$ and consequently, $\{g^*(X)=1\}^c \subset C_\alpha^c$. Then, combining this fact with (13) we have that, for k large enough

$$\begin{aligned} \mathbb{P}_{\mathcal{D}_k}(Y=1|A_\alpha) &\geq \frac{\mathbb{P}_{\mathcal{D}_k}(Y=1, g^*(X)=1 \big| \cap_{m=1}^{R} B_{mk})}{\mathbb{P}_{\mathcal{D}_k}(g^*(X)=1)}(1-\varepsilon) \\ &= \frac{\mathbb{P}_{\mathcal{D}_k}(Y=1, g^*(X)=1)}{\mathbb{P}_{\mathcal{D}_k}(g^*(X)=1)}(1-\varepsilon) \\ &= \mathbb{P}(V=1|g^*(X)=1)(1-\varepsilon) \\ &> 1/2. \end{aligned} \tag{15}$$

Therefore, from (12) and (15) in (11) we get

$$E(Z_jY_j) - E(Z_j)/2 > c > 0.$$

Going back to (10), conditioning to $\nu(X)$ and using the Hoeffding inequality for $|Z_j(Y_j - 1/2)| \leq 1/2$, for k large enough we have

$$I_A = \mathbb{P}_{\mathcal{D}_k}\left(\frac{1}{n-k}\sum_{j=k+1}^{n} -\big(Z_j(Y_j - 1/2) - E(Z_j(Y_j - 1/2))\big) \geq c \,\Big|\, g^*(X) = 1, \mathbf{g_{Rk}}(X) = \mathbf{1}\right)$$
$$\leq \exp\{-C_1(n-k)\},$$

with $C_1 = 2c^2$. On the other hand, by hypothesis we have

$$I_B = \sum_{m=1}^{M} \mathbb{P}_{\mathcal{D}_k}(g^*(X) \neq g_{mk}(X)) = \mathcal{O}(\boldsymbol{\beta_{Rk}}),$$

which concludes the proof. □

Proof of Theorem 2. First we write,

$$\begin{aligned}\mathbb{P}_{\mathcal{D}_k}\big(g_T(X) \neq Y\big) &= \mathbb{P}_{\mathcal{D}_k}\big(T_n(\mathbf{g_k}(X)) > 1/2, Y = 0\big) + \mathbb{P}_{\mathcal{D}_k}\big(T_n(\mathbf{g_k}(X)) \leq 1/2, Y = 1\big)\\ &= \sum_{\nu\in\mathbb{C}} \mathbb{P}_{\mathcal{D}_k}\big(T_n(\mathbf{g_k}(X)) > 1/2, (X,Y) \in A_\nu^0\big) + \sum_{\nu\in\mathbb{C}} \mathbb{P}_{\mathcal{D}_k}\big(T_n(\mathbf{g_k}(X)) \leq 1/2, (X,Y) \in A_\nu^1\big)\\ &\doteq I + II. \end{aligned} \tag{16}$$

Let us take ν fixed. Observe that in this case, $T_n(\mathbf{g_k}(\mathbf{X}))$ depends only on the subsample $\mathcal{E}_l$, therefore the events $(X_j, Y_j) \in A_\nu^i$ and $(X, Y) \in A_\nu^i$ are independent for all $i = 0, 1, j = k+1, \ldots, n$. Then,

$$\begin{aligned} I &= \sum_{\nu\in\mathbb{C}} \mathbb{P}_{\mathcal{D}_k}\big(T_n(\mathbf{g_k}(X), 0) > 1/2, (X,Y) \in A_\nu^0\big)\\ &= \sum_{\nu\in\mathbb{C}} \mathbb{P}_{\mathcal{D}_k}\left(\frac{\sharp\{j : (X_j, Y_j) \in A_\nu^1\}}{l} > \frac{\sharp\{j : (X_j, Y_j) \in A_\nu^0\}}{l}, (X,Y) \in A_\nu^0\right)\\ &= \sum_{\nu\in\mathbb{C}} \mathbb{P}_{\mathcal{D}_k}\left(\frac{\sharp\{j : (X_j, Y_j) \in A_\nu^1\}}{l} > \frac{\sharp\{j : (X_j, Y_j) \in A_\nu^0\}}{l}\right) \mathbb{P}_{\mathcal{D}_k}\left((X,Y) \in A_\nu^0\right)\\ &= \sum_{\nu\in\mathbb{C}} \mathbb{P}_{\mathcal{D}_k}\left(\frac{1}{l}\sum_{j=k+1}^{n} \mathbb{I}_{\{(X_j,Y_j)\in A_\nu^1\}} - \mathbb{I}_{\{(X_j,Y_j)\in A_\nu^0\}} > 0\right) \mathbb{P}_{\mathcal{D}_k}\left((X,Y) \in A_\nu^0\right). \end{aligned}$$

Let us define

$$\begin{aligned} &T_j^\nu \doteq \mathbb{I}_{\{(X_j,Y_j)\in A_\nu^1\}} - \mathbb{I}_{\{(X_j,Y_j)\in A_\nu^0\}},\\ &p_\nu^i \doteq \mathbb{P}_{\mathcal{D}_k}((X,Y) \in A_\nu^i), \quad i = 0, 1,\\ &p_\nu \doteq E(T_j^\nu) = p_\nu^1 - p_\nu^0, K_1 = 2\min_{\{\nu : p_\nu \neq 0\}} p_\nu^2, \quad \text{and} \quad \sigma_\nu^2 \doteq E((T_j^\nu)^2) = p_\nu^1 + p_\nu^0. \end{aligned}$$

To bound term I, we will consider 3 cases, $p_\nu < 0$, $p_\nu > 0$ and $p_\nu = 0$. Let us first assume that $p_\nu < 0$. In this case, using the Hoeffding inequality we have,

$$\mathbb{P}_{\mathcal{D}_k}\left(\frac{1}{l}\sum_{j=k+1}^{n}(T_j^\nu - p_\nu) > -p_\nu\right) = \mathcal{O}_k\big(\exp(-K_1 l)\big). \tag{17}$$

If $p_\nu > 0$, using Hoeffding inequality again we get

$$\begin{aligned}\mathbb{P}_{\mathcal{D}_k}\left(\frac{1}{l}\sum_{j=k+1}^{n} T_j^\nu > 0\right) &= 1 - \mathbb{P}_{\mathcal{D}_k}\left(\frac{1}{l}\sum_{j=k+1}^{n} T_j^\nu \leq 0\right)\\ &= 1 + \mathcal{O}_k\big(\exp(-K_1 l)\big). \end{aligned} \tag{18}$$

If $p_\nu = 0$, since for all ν and j, $E(|T_j^\nu|^3) = 1$, using the Berry–Esseen inequality we get

$$\mathbb{P}_{\mathcal{D}_k}\left(\frac{1}{l}\sum_{j=k+1}^{n} T_j^\nu > 0\right) = \Big[\frac{1}{2} + \mathcal{O}_k\big(l^{-1/2}\big)\Big]\mathbb{I}_{\{\sigma_\nu^2 > 0\}}. \tag{19}$$

Observe that, since $\mathbb{P}(Y=1)=\sum_\nu p_\nu^1$ and $\mathbb{P}(Y=0)=\sum_\nu p_\nu^0$, there exists ν such that $\sigma_\nu^2>0$. Then, from (17)–(19) we get

$$\begin{aligned} I &= \sum_{\nu\in\mathbb{C}} \mathbb{P}_{\mathcal{D}_k}\left(\frac{1}{l}\sum_{j=k+1}^{n} T_j^\nu > 0\right) p_\nu^0 \\ &= \sum_{p_\nu<0} \mathbb{P}_{\mathcal{D}_k}\left(\frac{1}{l}\sum_{j=k+1}^{n} T_j^\nu > 0\right) p_\nu^0 + \sum_{p_\nu>0} \mathbb{P}_{\mathcal{D}_k}\left(\frac{1}{l}\sum_{j=k+1}^{n} T_j^\nu > 0\right) p_\nu^0 + \sum_{p_\nu=0} \mathbb{P}_{\mathcal{D}_k}\left(\frac{1}{l}\sum_{j=k+1}^{n} T_j^\nu > 0\right) p_\nu^0 \\ &= \mathcal{O}_k\big(\exp(-K_1 l)\big)\mathbb{I}_{\{\exists\nu: p_\nu\neq 0\}} + \mathcal{O}_k\big(l^{-1/2}\big)\mathbb{I}_{\{\exists\nu:(p_\nu=0,\sigma_\nu^2>0)\}} + \sum_{p_\nu>0} p_\nu^0 + \frac{1}{2}\sum_{p_\nu=0} p_\nu^0. \end{aligned} \tag{20}$$

Analogously, it is easy to prove that

$$II = \mathcal{O}_k\big(\exp(-K_1 l)\big)\mathbb{I}_{\{\exists\nu: p_\nu\neq 0\}} + \mathcal{O}_k\big(l^{-1/2}\big)\mathbb{I}_{\{\exists\nu:(p_\nu=0,\sigma_\nu^2>0)\}} + \sum_{p_\nu<0} p_\nu^1 + \frac{1}{2}\sum_{p_\nu=0} p_\nu^0, \tag{21}$$

where in the last term we have used that $p_\nu=0$ implies $p_\nu^0=p_\nu^1$. Therefore, with (20) and (21) in (16) we get,

$$\begin{aligned} \mathbb{P}_{\mathcal{D}_k}\big(g_T(X)\neq Y\big) &= \mathcal{O}_k\big(\exp(-K_1 l)\big)\mathbb{I}_{\{\exists\nu: p_\nu\neq 0\}} + \mathcal{O}_k\big(l^{-1/2}\big)\mathbb{I}_{\{\exists\nu:(p_\nu=0,\sigma_\nu^2>0)\}} + \sum_{p_\nu>0} p_\nu^0 + \sum_{p_\nu<0} p_\nu^1 + \sum_{p_\nu=0} p_\nu^0 \\ &= \mathcal{O}_k\big(\exp(-K_1 l)\big)\mathbb{I}_{\{\exists\nu: p_\nu\neq 0\}} + \mathcal{O}_k\big(l^{-1/2}\big)\mathbb{I}_{\{\exists\nu:(p_\nu=0,\sigma_\nu^2>0)\}} \\ &\quad + \sum_{\substack{\nu:\nu(m)=0\\ p_\nu>0}} p_\nu^0 + \sum_{\substack{\nu:\nu(m)=1\\ p_\nu>0}} p_\nu^0 + \sum_{\substack{\nu:\nu(m)=0\\ p_\nu<0}} p_\nu^1 + \sum_{\substack{\nu:\nu(m)=1\\ p_\nu<0}} p_\nu^1 + \sum_{p_\nu=0} p_\nu^0. \end{aligned} \tag{22}$$

On the other hand, for each m we have,

$$\begin{aligned} \mathbb{P}_{\mathcal{D}_k}\big(g_{mk}(X)\neq Y\big) &= \mathbb{P}_{\mathcal{D}_k}\big(g_{mk}(X)=0, Y=1\big) + \mathbb{P}_{\mathcal{D}_k}\big(g_{mk}(X)=1, Y=0\big) \\ &= \mathbb{P}_{\mathcal{D}_k}\left(\bigcup_{\nu:\nu(m)=0}(X,Y)\in A_\nu^1\right) + \mathbb{P}_{\mathcal{D}_k}\left(\bigcup_{\nu:\nu(m)=1}(X,Y)\in A_\nu^0\right) \\ &= \sum_{\nu:\nu(m)=0} p_\nu^1 + \sum_{\nu:\nu(m)=1} p_\nu^0 \\ &= \sum_{\substack{\nu:\nu(m)=0\\ p_\nu<0}} p_\nu^1 + \sum_{\substack{\nu:\nu(m)=0\\ p_\nu>0}} p_\nu^1 + \sum_{\substack{\nu:\nu(m)=1\\ p_\nu<0}} p_\nu^0 + \sum_{\substack{\nu:\nu(m)=1\\ p_\nu>0}} p_\nu^0 + \sum_{p_\nu=0} p_\nu^0, \end{aligned} \tag{23}$$

where in the last equality we used again that $p_\nu=0$ implies $p_\nu^0=p_\nu^1$ to joint

$$\sum_{\substack{\nu:\nu(m)=1\\ p_\nu=0}} p_\nu^0 + \sum_{\substack{\nu:\nu(m)=0\\ p_\nu=0}} p_\nu^1 = \sum_{p_\nu=0} p_\nu^0.$$

Therefore, from (22) and (23) we get

$$\begin{aligned} \mathbb{P}_{\mathcal{D}_k}\big(g_T(X)\neq Y\big) - \mathbb{P}_{\mathcal{D}_k}\big(g_{mk}(X)\neq Y\big) &= \mathcal{O}_k\big(\exp(-K_1 l)\big)\mathbb{I}_{\{\exists\nu: p_\nu\neq 0\}} + \mathcal{O}_k\big(l^{-1/2}\big)\mathbb{I}_{\{\exists\nu:(p_\nu=0,\sigma_\nu^2>0)\}} \\ &\quad + \sum_{\substack{\nu:\nu(m)=0\\ p_\nu>0}} (p_\nu^0 - p_\nu^1) + \sum_{\substack{\nu:\nu(m)=1\\ p_\nu<0}} (p_\nu^1 - p_\nu^0) \\ &\leq \mathcal{O}_k\big(\exp(-K_1 l)\big)\mathbb{I}_{\{\exists\nu: p_\nu\neq 0\}} + \mathcal{O}_k\big(l^{-1/2}\big)\mathbb{I}_{\{\exists\nu:(p_\nu=0,\sigma_\nu^2>0)\}}. \end{aligned}$$

Observe that, if $p_\nu\neq 0$ for all ν, we get $\mathbb{P}_{\mathcal{D}_k}\big(g_T(X)\neq Y\big) - \mathbb{P}_{\mathcal{D}_k}\big(g_{mk}(X)\neq Y\big) \leq \mathcal{O}_k\big(\exp(-lK_1)\big)$. □

References

[1] A. Baíllo, A. Cuevas, R. Fraiman, Classification methods for functional data, in: F. Ferraty, Y. Romain (Eds.), The Oxford Handbook of Functional Data Analysis, Oxford University Press, Oxford, 2011, pp. 259–297.
[2] D. Banks, E. Petricoin, Finding cancer signals in mass spectrometry data, Chance 16 (2003) 8–57.
[3] G. Biau, Analysis of a random forests model, J. Mach. Learn. Res. 13 (2012) 1063–1095.
[4] G. Biau, L. Devroye, G. Lugosi, Consistency of random forests and other averaging classifiers, J. Mach. Learn. Res. 9 (2008) 2015–2033.
[5] G. Biau, A. Fischer, B. Guedj, J. Malley, COBRA: A combined regression strategy, J. Multivariate Anal. (in press). Available online 21 April 2015, http://dx.doi.org/10.1016/j.jmva.2015.04.007.
[6] E. Bongiorno, E. Salinelli, A. Goia, P. Vieu, An overview of IWFOS'2014, in: Enea G. Bongiorno, Ernesto Salinelli, Aldo Goia, Philippe Vieu (Eds.), Contributions in Infinite-Dimensional Statistics and Related Topics, Societá Editrice Esculapio, 2014, pp. 1–6.

[7] E. Bongiorno, E. Salinelli, A. Goia, P. Vieu (Eds.), Contributions in Infinite-Dimensional Statistics and Related Topics, Societá Editrice Esculapio, 2014.
[8] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140.
[9] L. Breiman, Arcing classifiers, Ann. Statist. 24 (1998) 801–849.
[10] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[11] F. Bunea, A.B. Tsybakov, M.H. Wegkamp, Aggregation for Gaussian regression, Ann. Statist. 35 (2007) 1674–1697.
[12] J.A. Cuesta-Albertos, R. Fraiman, T. Ransford, Random projections and goodness-of-fit tests in infinite-dimensional spaces, Bull. Braz. Math. Soc. 37 (2006) 1–25.
[13] A. Cuevas, A partial overview of the theory of statistics with functional data, J. Statist. Plann. Inference 147 (2012) 1–23.
[14] A. Delaigle, P. Hall, Achieving near perfect classification for functional data, J. R. Stat. Soc. Ser. B 74 (2) (2012) 267–286.
[15] R. Fraiman, R. Liu, J. Meloche, Multivariate density estimation by probing depth, in: L_1–Statistical Procedures and Related Topics, in: IMS Lectures Notes—Monograph Series, vol. 31, 1997, pp. 415–430.
[16] P. Hall, R. Samworth, Properties of bagged nearest neighbor classifiers, J. R. Stat. Soc. Ser. B 74 (2) (2005) 267–286.
[17] M. Mojirsheibani, Combining classifiers via discretization, J. Amer. Statist. Assoc. 94 (1999) 600–609.
[18] M. Mojirsheibani, An almost surely optimal combined classification rule, J. Multivariate Anal. 81 (2002) 28–46.
[19] Y. Yang, Aggregating regression procedures to improve performance, Bernoulli 10 (2004) 25–47.