DOI 10.1007/s10559-016-9819-0
Cybernetics and Systems Analysis, Vol. 52, No. 2, March, 2016

AFFINE-INVARIANT CLASSIFIER OF EXTRAPOLATION DEPTH ON THE BASIS OF A MULTILEVEL SMOOTHING STRUCTURE

O. A. Galkin

UDC 519.7

Abstract. *A nonparametric affine-invariant extrapolation depth-based classifier resistant to spikes and extreme values is proposed and investigated. A multilevel smoothing structure is proposed that makes it possible to obtain global properties of density functions and class boundaries under appropriate regularity conditions. The extrapolation depth-based classifier uses kernel density estimates to efficiently classify multidimensional data at different smoothing levels.*

Keywords: *kernel density estimate, smoothing level, depth function.*

INTRODUCTION

The use of classifiers of maximum extrapolation depth allows one to obtain relatively low coefficients of erroneous classification in the case when a priori probabilities of data sets are equal and their distributions differ only in arrangement parameters. However, in practice, distributions of data sets often have different matrices of dispersion and form and also different a priori probabilities. The described features stipulate the topicality of the problem of developing advanced versions of maximum depth classifiers. The existing versions of extrapolation depth-based classifiers allow one to solve applied problems in the case of a monotone relation between depth functions and density functions and also under the condition that data sets have different dispersion matrices [1].

EXTRAPOLATION DEPTH-BASED CLASSIFIER BASED ON ELLIPTIC SYMMETRY OF DISTRIBUTIONS

Let us consider the case when distributions of data sets are elliptic. If $E(z, H_l)$ is the depth of z with respect to H_l, then the Bayesian classifier is specified as follows:

$$\mathfrak{I}_{\mathrm{B}}(z) = \arg \max_{1 \le l \le L} p_l o_l \{E(z, H_l)\},$$

where o_l is a transformation function that monotonically decreases and is the same for all groups of data sets if functions h_l are unimodal and distributions of data sets differ only in arrangement parameters [2]. Moreover, the Bayesian classifier is equivalent to the maximum depth classifier if p_l are equal. However, if at least one of the mentioned conditions is not fulfilled, then a demand arises for the obtainment of information on functional forms o_l.

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine, *galkin.o.a@gmail.com*. Translated from Kibernetika Sistemnyi Analiz, No. 2, March–April, 2016, pp. 64–72. Original article submitted October 6, 2015.

LEMMA 1. If $\xi_l(\cdot)$ is the density function of $F_e(z,H_l)$ and functions $h_1, h_2, \ldots, h_L$ are elliptically symmetric, then the Bayesian classifier is specified as follows:

$$\mathfrak{I}_{\mathrm{B}}(z) = \arg \max_{l \in \{1,\ldots,L\}} \alpha_l \xi_l \{F_e(z,H_l)\}\{F_e(z,H_l)\}^{r-3} / \{1-F_e(z,H_l)\}^{r-1},$$

where α_l is a constant.

Proof. Taking into account that a function h_l is elliptically symmetric, we have

$$h_l(z) = \mathrm{I}(r/2)(2p)^{-r/2} |\Xi_l|^{-1/2} c_l(C(z,H_l))/C(z,H_l)^{r-1},$$

where c_l is the probability density function of $C(z,H_l) = \{(z-\varepsilon_l)' \Xi_l^{-1}(z-\varepsilon_l)\}^{1/2}$ and ε_l and Ξ_l are, respectively, arrangement and scale parameters for h_l.

Hence, it is arguable that

$$\mathfrak{I}_{\mathrm{B}}(z) = \arg \max_{1 \le l \le L} p_l h_l(z) = \arg \max_{1 \le l \le L} \alpha_l \lambda_l \{\mathrm{I}(z,H_l)\} / \{\mathrm{I}(z,H_l)\}^{r-1},$$

where the constant α_l depends on H_l and p_l and λ_l is the density function of $\mathrm{I}(z,H_l)$. Since $F_e(z,H_l)$ $= \{1+\mathrm{I}(z,H_l)\}^{-1}$, the proof follows from the properties of selective distribution.

The lemma is proved.

Note that, until constants α_l change depending on the choice of one-dimensional measures of scale and arrangement, Lemma 1 is true for any definition of extrapolation depth functions.

KERNEL DENSITY ESTIMATES

To construct an advanced version of the extrapolation depth-based classifier, the method of kernel density estimates is used to estimate ξ_l, and also the selective form $F_e(z,H_{lm_l})$ is applied to estimate $F_e(z,H_l)$. In this case, the one-dimensional density is estimated irrespective of the dimension of the space of measurements, which makes it possible to avoid the problem of the so-called "curse of dimensionality" that often takes place when multidimensional nonparametric densities are estimated [3].

Note that the choice of a bandwidth a_l is obligatory to estimate ξ_l, $1 \le l \le L$. This density estimate is specified as follows:

$$\bar{\xi}_{la_l}(\omega) = (m_l a_l)^{-1} \sum_{i=1}^{m_l} \Theta\{a_l^{-1}(\omega - \varpi_{m_l}^{(l)}(z_{li}))\},$$

where Θ is a kernel function and $\varpi_{m_l}^{(l)}(z) = F_e(z,H_{lm_l})$.

LEMMA 2. Let the following assumptions take place:

(a) the function $h_l(z) > 0$ $\forall z \in \mathrm{R}^r$ and $l = 1, 2$;

(b) for $l = 1,2$, the function $H_{\beta,l}(\bar{z}) = P(\beta(Z) \le \bar{z})$ is uniformly continuous at $\bar{z}$, where $\beta(z) = d^{(2)}(z)/d^{(1)}(z)$, $d^{(l)}(z) = \omega_l(\xi^{(l)}(z))(\xi^{(l)}(z))^{r-3} / (1-\xi^{(l)}(z))^{r-1}$, $\xi^{(l)}(z) = F_{np}(z,H_l)$, and Z belongs to the lth class;

(c) for $l = 1,2$, the value of $a_l \to 0$ and value of $m_l a_l^4 \to \infty$ as $m_l \to \infty$.

We also assume that h_1 and h_2 are elliptically symmetric functions. If the sought-for estimate Δ is chosen as a result of minimization of the estimate for the crosscheck of the frequency of errors, then the coefficient of erroneous classification of the extrapolation depth-based classifier $\mathfrak{I}_2(\cdot)$ is converged to Bayes risk as $\min\{m_1, m_2\} \to \infty$.

Proof. It is obvious that

$$|\Psi(r_2) - \Psi_V| \le \sum_{l=1}^{2} \int \left| \prod_{i=1, i \ne l}^{2} \Lambda\left\{ \frac{d_{m_l,a_l}^{(l)}(z)}{d_{m_i,a_i}^{(i)}(z)} \ge v_m \right\} - \prod_{i=1, i \ne l}^{2} \Lambda\left\{ \frac{d^{(l)}(z)}{d^{(i)}(z)} \ge v \right\} \right| h_l(z) dz.$$

It is arguable that $|\Psi(r_2)-\Psi_V|$ converges in probability to zero by the Lebesgue theorem on majorized convergence in which indicators are bounded by corresponding functions. The result can be obtained using selective expectation and repeatedly applying the Lebesgue majorized convergence theorem.

The lemma is proved.

In this investigation, we use the Gaussian kernel assuming that the kernel Θ has a bounded first derivative. In considering a two-class problem in which $d^{(l)}_{m_l,a_l}(z)=\bar{\xi}_{la_l}(\varpi^{(l)}_{m_l}(z))(\varpi^{(l)}_{m_l}(z))^{r-3}/(1-\varpi^{(l)}_{m_l}(z))^{r-1}$ for $l=1,2$ and $\Delta=\log(\alpha_2/\alpha_1)$, it is arguable that the resulting classifier $\mathfrak{F}_2(z)=1$ if $\log[d^{(1)}_{m_1,a_1}(z)]-\log[d^{(2)}_{m_2,a_2}(z)]>\Delta$ and $\mathfrak{F}_2(z)=2$ otherwise. It is obvious that the choice of a_1, a_2, and Δ exerts influence on the performance of the classifier $\mathfrak{F}_2(\cdot)$. Therefore, with increasing the sample size and also according to assumptions (a) and (b), the frequency of errors of the advanced version of the extrapolation depth-based classifier $\mathfrak{F}_2(\cdot)$ converges to Bayes risk under the condition that a_1 and a_2 satisfy assumption (c) and Δ is chosen by minimizing the estimate of the crosscheck of the frequency of errors [4].

THEOREM 1. Let $\bar{\beta}_m(z)=d^{(2)}_{m_2,a_2}(z)/d^{(1)}_{m_1,a_1}(z)$, and let $\beta(z)=d^{(2)}(z)/d^{(1)}(z)$. Then $\exists G_\mu$ such that $P(G_\mu)>1-\mu$ and $\sup_{z\in W_\mu}|\bar{\beta}_m(z)-\beta(z)|\xrightarrow{P}0$ as $\min\{m_1,m_2\}\to\infty$, where $i=1,2$, and Z belongs to the ith class.

Proof. We define $\bar{\xi}_{ia_i}$ and $\bar{\xi}^*_{ia_i}(\omega)=\dfrac{1}{m_i a_i}\sum_{l=1}^{m_i}\Theta\left\{\dfrac{\omega-\omega^{(i)}(z_{il})}{a_i}\right\}$ for $i=1,2$. Note also that

$$\sup_z|\bar{\xi}_{ia_i}(\varpi^{(i)}_{m_i}(z))-\xi_i(\omega^{(i)}(z))|\le\sup_z|\bar{\xi}_{ia_i}(\varpi^{(i)}_{m_i}(z))-\bar{\xi}_{ia_i}(\omega^{(i)}(z))|$$

$$+\sup_z|\bar{\xi}_{ia_i}(\omega^{(i)}(z))-\xi^*_{ia_i}(\omega^{(i)}(z))|+\sup_z|\xi^*_{ia_i}(\omega^{(i)}(z))-\xi_i(\omega^{(i)}(z))|.$$

Using assumption (c) of Lemma 2 and assuming that $N_\Theta=\sup_\kappa|\Theta'(\kappa)|<\infty$, we have

$$\sup_z|\varpi_{ia_i}(\bar{\xi}^{(i)}_{m_i}(z))-\varpi_{ia_i}(\xi^{(i)}(z))|\le N_\Theta\sup_z|\bar{\xi}^{(i)}_{m_i}(z)-\xi^{(i)}(z)|/a_1^2\xrightarrow{P}0 \tag{1}$$

as $m_i\to\infty$. Note that inequality (1) is based on the fact that $\sup_z|\bar{\xi}^{(i)}_{m_i}(z)-\xi^{(i)}(z)|=O_P(m_i^{-1/2})$ and that a function h_i is elliptically symmetric. Thus, from this we have

$$\sup_z|\varpi_{ia_i}(\xi^{(i)}(z))-\omega^*_{ia_i}(\xi^{(i)}(z))|\xrightarrow{P}0 \tag{2}$$

as $m_i\to\infty$.

As a result, using the properties of the uniform continuity of the extrapolation depth function that follow from the elliptical symmetry of h_i, we obtain

$$\sup_z|\omega^*_{ia_i}(\xi^{(i)}(z))-\omega_i(\xi^{(i)}(z))|\xrightarrow{P}0 \tag{3}$$

as $m_i\to\infty$. Note that convergence (3) takes place if assumption (c) of Lemma 2 is fulfilled and the properties of kernel density estimates are used [5].

In fine, uniting convergences (2) and (3), we obtain $\sup_z|\varpi_{ia_i}(\bar{\xi}^{(i)}_{m_i}(z))-\omega_i(\xi^{(i)}(z))|\xrightarrow{P}0$ as $m_i\to\infty$.

For all $\mu>0$, some $\theta=\theta(\mu)>0$ can be found such that the set $G_\mu=\{z:\theta\le\xi^{(1)}(z),\xi^{(2)}(z)\le1-\theta\}$ will have a higher probability than $1-\mu$ with respect to the probability distribution of two classes.

It is obvious that

$$\sup_{z\in G_\mu}\left|\frac{(\bar{\xi}^{(i)}_{m_i}(z))^{r-3}}{(1-\bar{\xi}^{(i)}_{m_i}(z))^{r-1}}-\frac{(\xi^{(i)}(z))^{r-3}}{(1-\xi^{(i)}(z))^{r-1}}\right|\xrightarrow{P}0$$

for $i=1,2$. This implies that

$$\sup_{z\in G_\mu} |d^{(i)}_{m_i,a_i}(z)-d^{(i)}(z)| \xrightarrow{P} 0$$

as $m_i \to \infty$. Thus, since $\inf_{z\in G_\mu} d^{(i)}(z)>0$ for $i=1,2$, we obtain the sought-for result.

The theorem is proved.

THEOREM 2. Let $v_m = \arg\min_\Delta \Psi_m^{VB}(\Delta)$, let $v=\arg\min_\Delta \Psi(\Delta)$, let the assumptions of Lemma 2 be fulfilled, and also let

$$\Psi_m^{VB}(\Delta)=\sum_{i=1,\, l\neq i}^{2} \frac{p_i}{m_i}\sum_{j=1}^{m_i}\Lambda\left\{\frac{d^{(l)}_{m_l,a_l}(z_{ij})}{d^{(i)}_{m_i,a_i}(z_{ij})}\geq \Delta_i\right\},$$

$$\Psi(\Delta)=\sum_{i=1,\, l\neq i}^{2} p_i P\left\{\frac{d^{(l)}(Z)}{d^{(i)}(Z)}\geq \Delta_i\right\},$$

where $m=(m_1,m_2)$, $\Delta_1=1/\Delta$, $\Delta_2=\Delta$, and Z belongs to the ith class. Then $v_m \xrightarrow{P} v$ as $\min(m_1,m_2)\to\infty$ on the condition that v is unique.

Proof. Let us show that $\sup_\Delta|\Psi_m^{VB}(\Delta)-\Psi(\Delta)|\xrightarrow{P}0$ as $\min(m_1,m_2)\to\infty$. Note that the convergence $v_m \xrightarrow{P} v$ takes place that follows from $\sup_\Delta|\Psi_m^{VB}(\Delta)-\Psi(\Delta)|\xrightarrow{P}0$ since $\Psi(\cdot)$ is a unique minimum.

Note that

$$|\Psi_m^{VB}(\Delta)-\Psi(\Delta)|\leq \sum_{i=1,\, l\neq i}^{2}\frac{p_i}{m_i}\sum_{j=1}^{m_i}\left|\Lambda\left\{\frac{d^{(l)}_{m_l,a_l}(z_{ij})}{d^{(i)}_{m_i,a_i}(z_{ij})}\geq\Delta_i\right\}-P\left\{\frac{d^{(l)}(Z)}{d^{(i)}(Z)}\geq\Delta_i\right\}\right|$$

$$\leq \sum_{i=1,\, l\neq i}^{2}\frac{p_i}{m_i}\sum_{j=1}^{m_i}\left|\Lambda\left\{\frac{d^{(l)}_{m_l,a_l}(z_{ij})}{d^{(i)}_{m_i,a_i}(z_{ij})}\geq\Delta_i\right\}-\Lambda\left\{\frac{d^{(l)}(z_{ij})}{d^{(i)}(z_{ij})}\geq\Delta_i\right\}\right|$$

$$+\sum_{i=1,\, j\neq i}^{2}\frac{p_i}{m_i}\sum_{j=1}^{m_i}\left|\Lambda\left\{\frac{d^{(l)}(z_{ij})}{d^{(i)}(z_{ij})}\geq\Delta_i\right\}-P\left\{\frac{d^{(l)}(Z)}{d^{(i)}(Z)}\geq\Delta_i\right\}\right|,$$

where Z belongs to ith class.

Next, we determine the quantities

$$G_m(\Delta_1)=\frac{1}{m_1}\sum_{i=1}^{m_1}|\Lambda\{\beta(z_{1i})\geq\Delta_1\}-P\{\beta(Z)\geq\Delta_1\}|\ ,$$

$$V_m(\Delta_1)=\frac{1}{m_1}\sum_{i=1}^{m_1}|\Lambda\{\bar\beta_m(z_{1i})\geq\Delta_1\}-\Lambda\{\beta(z_{1i})\geq\Delta_1\}|,$$

where Z belongs to the lth class. It can be proved that $\sup_{\Delta_1}|G_m(\Delta_1)|\xrightarrow{ac}0$ by using the Glivenko–Cantelli lemma.

For all $\mu>0$, we obtain some $\xi_\mu>0$ such that

$$\sup_{\Delta_1}|H_{\beta,1}(\Delta_1+\xi_\mu/2)-H_{\beta,1}(\Delta_1-\xi_\mu/2)|<\mu$$

according to assumption (b) of Lemma 2. Moreover, using Theorem 1, we obtain some G_μ such that $P(G_\mu)>1-\mu$ for $l=1,2$, and Z belongs to the lth class.

Then we define the set $W_\mu = \{z : |\beta(z) - \Delta_1| > \xi_\mu / 2\} \cap \{z : z \in G_\mu\}$ using ξ_μ and G_μ.
We have that

$$V_m(\Delta_1) = \frac{1}{m_1} \sum_{\{i:\, z_{1i} \notin W_\mu\}} |\Lambda\{\overline{\beta}_m(z_{1i}) < \Delta_1\} - \Lambda\{\beta(z_{1i}) < \Delta_1\}|$$

$$+ \frac{1}{m_1} \sum_{\{i:\, z_{1i} \in W_\mu\}} |\Lambda\{\overline{\beta}_m(z_{1i}) < \Delta_1\} - \Lambda\{\beta(z_{1i}) < \Delta_1\}|$$

$$\le \frac{1}{m_1} \sum_{i=1}^{m_1} \Lambda\{z_{1i} \notin W_\mu\} + \frac{1}{m_1} \sum_{\{i:\, z_{1i} \in W_\mu\}} |\Lambda\{\overline{\beta}_m(z_{1i}) < \Delta_1\} - \Lambda\{\beta(z_{1i}) < \Delta_1\}|.$$

According to assumption (b) of Lemma 2 and Theorem 1, we have the asymptotic convergence

$$\frac{1}{m_1} \sum_{i=1}^{m_1} \Lambda\{z_{1i} \notin W_\mu\} \to P(Z_1 \notin W_\mu) \le P(|(\beta(Z_1) - \Delta_1| \le \xi_\mu / 2) + P(Z_1 \notin G_\mu) < 2\mu$$

as $\min\{m_1, m_2\} \to \infty$.

It follows from $|\beta(z) - \Delta_1| > \xi_\mu / 2$, $\exists M_0 \ge 1$ that, for all $m = (m_1, m_2)$, where $\min\{m_1, m_2\} \ge M_0$, we have $|\overline{\beta}_m(z) - \Delta_1| > \xi_\mu / 2$.

Thus, we obtain

$$\frac{1}{m_1} \sum_{\{i:\, W(z_{1i})\}} |\Lambda\{\overline{\beta}_m(z_{1i}) < \Delta_1\} - \Lambda\{\beta(z_{1i}) < \Delta_1\}| = 0,$$

which implies that $V_m(\Delta_1) \le 2\mu$.

As a result, this theorem can be proved using a reasoning similar to that used in the case when $i = 2$.

The theorem is proved.

Note that a similar half-space depth-based classifier is complicated and insufficiently efficient in handling zero depths. The experimental modification of the half-space depth function assumes only discrete values, which leads to the loss of information for continuous distributions. As a result, we obtain inexact density estimates with peaks in the neighborhood of these discrete values. Moreover, an essential problem is connected with inequalities at the tail end of the initial estimate of the density $\overline{h}_l$, which is caused by the presence of objects with zero depths. Note that such problems are absent in the case of the experimental modification of the extrapolation depth function, and it is continuous at z. Therefore, the advanced version of the extrapolation depth-based classifier often surpasses the half-space depth-based classifier.

In practice, on a data set, it is necessary to estimate a_l whose optimal asymptotic order is substantiated in Lemma 2 in which the throughput crosschecking method is used for the choice of a_1, a_2, and Δ. To decrease computational expenses, $a_1 = (w_1 / w_2) a_2$ has been chosen since bandwidths must be proportional to dispersions of sets, where w_l $(l = 1, 2)$ is the dispersion measure of evaluation functions of depth $\{\overline{\xi}^{(l)}_{m_l}(z_{l1}), \overline{\xi}^{(l)}_{m_l}(z_{l2}), \ldots, \overline{\xi}^{(l)}_{m_l}(z_{lm_l})\}$. Then we compute

$$d^{(i)}_{m_i, a_i}(z_{lj}) = \overline{\varpi}^{*}_{ia_i}(\overline{\xi}^{(i)}_{m_i}(z_{lj}))(\overline{\xi}^{(i)}_{m_i}(z_{lj}))^{r-3} / (1 - \overline{\xi}^{(i)}_{m_i}(z_{lj}))^{r-1}$$

for a_2, $a_1 = (w_1 / w_2) a_2$, $i, l = 1, 2$, and $j = 1, \ldots, m_l$, where $\overline{\varpi}^{*}$ corresponds to the throughput kernel density estimate for $l = i$ $(l \ne i)$. The constant Δ that depends on a_2 is found based on order statistics $\log[d^{(1)}_{m_1, a_1}(z_{lj})] - \log[d^{(2)}_{m_2, a_2}(z_{lj})]$, $l = 1, 2$, $j = 1, 2, \ldots, m_l$, for the minimization of the frequency of crosscheck errors. Note that the choice of a_2 in the range of values is conditioned by the obtainment of a low coefficient of crosscheck errors. Moreover, the maximum optimizer is chosen from the set of minimizers obtained as a result of the stepped nature of the frequency of crosscheck errors [6].

These results can also be obtained for deep classification using the Mahalanobis depth. Since v_H is a constant that depends on the initial distribution of H, the estimate for the minimum covariance determinant of the dispersion matrix $\Xi_H \to v_H \Xi_H$. However, irrespective of the value of v_H, the form of the Bayesian classifier is similar to that of Lemma 1. This classification method can be adapted to the development of the advanced classifier version based on the Mahalanobis depth, and its asymptotic optimality can be proved based on Lemma 2.

In the case of multiclass classification, $a_1, a_2, \ldots, a_L$ and $\alpha_1, \alpha_2, \ldots, \alpha_L$ are similarly chosen, but, in practice, the minimization of the frequency of crosscheck errors with respect to several parameters is a computationally complicated task. Thus, we perform $\binom{L}{2}$ binary classifications considering a pair of classes where the results of all pairwise classifications are united with the help of the majority voting method. Note that, under corresponding regularity conditions, the consistency of the Bayes risk of the advanced version of the extrapolation depth-based classifier can be proved for multiclass problems on the basis of Lemma 2.

MULTILEVEL SMOOTHING FOR THE CLASSIFICATION OF MULTIDIMENSIONAL DATA

The estimation of the smoothing parameter in kernel density estimates was carried out with the help of the crosscheck method for the advanced version of the depth-based classifier. However, in solving practical classification problems, the model is quite often uncertain in using one pair of bandwidths (a_1, a_2). Along with the problem of selective dependence, the choice of a smoothing parameter that depends on the characteristic object of classification is essential. In this case, a definite smoothing level can determine different behaviors in different regions of the measurement space. Therefore, the problem of investigation of the results of classification for different smoothing scales instead of the use of a fixed pair (a_1, a_2) in certain range is topical. Data indexed according to bandwidths can be pooled by taking the weighed mean value of estimated posterior probabilities [7].

Note that $e^{\rho_{m,a_1,a_2}(z)}$ estimates $p_1 h_1(z) / p_2 h_2(z)$ since the element z belongs to the first class if $\rho_{m,a_1,a_2}(z) = \log[d^{(1)}_{m_1,a_1}(z)] - \log[d^{(2)}_{m_2,a_2}(z)] - \Delta > 0$, where Δ is chosen by minimizing the crosscheck error for fixed (a_1, a_2). Thus, we have $\overline{\pi}_{m,a_1,a_2}(1 \mid z) = e^{\rho_{m,a_1,a_2}(z)} / (1 + e^{\rho_{m,a_1,a_2}(z)})$, which is the estimated posterior probability of a class.

Since $\pi^*_m(l \mid z) = \sum_{a_1,a_2 \in A} q_{a_1,a_2} \overline{\pi}_{m,a_1,a_2}(l \mid z)$, the resulting classifier is formed by the union of the posterior estimates obtained for different values of (a_1, a_2), $\mathfrak{F}_3(z) = \arg\max_{l=1,2} \pi^*_m(l \mid z)$. Note that q_{a_1,a_2} is a weight assigned to the classifier for which a_1 and a_2 are bandwidths of two classes [8].

The union of posterior estimates depends on the weight function q and bandwidth range $A = [a_1^j, a_1^g] \times [a_2^j, a_2^g]$. However, irrespective of the choice of the weight function, the frequency of errors of $\mathfrak{F}_3(\cdot)$ asymptotically converges to the Bayes risk if the upper and lower boundaries a_l^g and a_l^j of a_l satisfy assumption (c) of Lemma 2 for $l = 1, 2$.

THEOREM 3. Assume that, for $l = 1, 2$, the quantities h_1 and h_2 are elliptically symmetric, where $h_l(z) > 0$ $\forall z \in \mathrm{R}^r$, and that $H_{\beta,l}(\bar{z}) = P(\beta(Z) \le \bar{z})$ is a uniformly continuous function at $\bar{z}$, where $\beta(z) = d^{(2)}(z) / d^{(1)}(z)$, $d^{(l)}(z) = \omega_l(\xi^{(l)}(z))(\xi^{(l)}(z))^{r-3} / (1 - \xi^{(l)}(z))^{r-1}$, and $\xi^{(l)}(z) = F_{np}(z, H_l)$ and Z belongs to the lth class. We also assume that, for a_l^g and a_l^j, the following convergences take place: $a_l \to 0$ and $m_l a_l^4 \to \infty$ as $m_l \to \infty$. Then the coefficient of erroneous classification for the multilevel classifier of extrapolation depth $\mathfrak{F}_3(\cdot)$ converges to the Bayes risk as $\min\{m_1, m_2\} \to \infty$.

Proof. The result follows from the Lebesgue majorized convergence theorem under the condition that, for a fixed z, the convergence $\pi^*_m(1\,|\,z) \to \pi(1\,|\,z)$ takes place as $\min\{m_1, m_2\} \to \infty$.

Assume that the convergence $\pi^*_m(1\,|\,z) \xrightarrow{P} \pi(1\,|\,z)$ is not takes place. Thus, $\exists\,\{m_\Delta = (m_{1\Delta}, m_{2\Delta}) : \Delta \ge 1\}$ and $\mu_0 > 0$ such that, $\forall\,\Delta \ge 1$, $|\pi^*_{m_\Delta}(1\,|\,z) - \pi(1\,|\,z)| > \mu_0$. Let $\{A_{m_\Delta}\}$, $\Delta \ge 1$, be the corresponding sequence in the bandwidth range. Taking into account the fact that $\pi^*_{m_\Delta}(1\,|\,z)$ is the weighted mean value of $\overline{\pi}_{m_\Delta, a_1, a_2}(1\,|\,z)$, a subsequence $\{(a_1^{m_\Delta}, a_2^{m_\Delta}) \in A_{m_\Delta}, \Delta \ge 1\}$ can be obtained such that $|\overline{\pi}_{m_\Delta, a_1^{m_\Delta}, a_2^{m_\Delta}}(1\,|\,z) - \pi(1\,|\,z)| > \mu_0$ $\forall\,\Delta \ge 1$. This implies that the convergence $\overline{\pi}_{m_\Delta, a_1^{m_\Delta}, a_2^{m_\Delta}}(1\,|\,z) \xrightarrow{P} \pi(1\,|\,z)$ does not take place. We obtain a contradiction since the sequence of bandwidths satisfies the regularity condition under which, for $l = 1, 2$, the convergences $a_l \to 0$ and $m_l a_l^4 \to \infty$ take place as $m_l \to \infty$.

The theorem is proved.

Based on the proof of Theorem 3, it is arguable that the choice of the weight function q does not exert any considerable influence on the selective performance of the classifier $\mathfrak{I}_3(\cdot)$. However, the choice of A and q is necessary in using an infinite sample. Note that weight must gradually decrease with increasing the frequency of errors with the use of larger scales for classifiers that have smaller frequencies of errors [9].

The frequency of errors Ψ_{a_1,a_2} is estimated by the throughput cross-checking method with the help of the weight function

$$q_{a_1,a_2} = \exp\left[-\frac{1}{2}\frac{(\overline{\Psi}_{a_1,a_2} - \overline{\Psi}_0)^2}{\overline{\Psi}_0(1-\overline{\Psi}_0)/(m_1+m_2)}\right]\Lambda[\overline{\Psi}_{a_1,a_2} \le \min\{p_1, p_2\}],$$

where $\overline{\Psi}_0 = \min\limits_{a_1,a_2} \overline{\Psi}_{a_1,a_2}$.

In the case when the advanced version of the one-level extrapolation depth-based classifier is used for the classification of (m_1+m_2) objects, $\overline{\Psi}_0$ and $\overline{\Psi}_0(1-\overline{\Psi}_0)/(m_1+m_2)$ can be considered as estimates for the mean value and dispersion of the experimental frequency of errors. Moreover, the frequency of errors of the classifier that assigns all objects to the class with the greatest a priori probability is $\min\{p_1, p_2\}$. Note that the weighting scheme of the classifier being investigated demonstrates the zero weight if the classifier with a pair of bandwidths (a_1, a_2) is less efficient than a trivial classifier.

The method based on quantiles of paired distances was used for the choice of A, and also 500 equidistant values of (a_1, a_2) were determined in this interval that satisfies the condition $a_1 = (w_1/w_2)a_2$, where w_1 and w_2 are identical. Note that, as a result of the performed experiments, good results are obtained owing to an appropriate choice of the bandwidth range and weight function.

CONCLUSIONS

In this work, a nonparametric affine-invariant extrapolation depth-based classifier resistant to spikes and extreme values is proposed and investigated. Owing to the connection of the proposed classifier with the Mahalanobis distance and also to the continuity of its experimental form, the extrapolation depth-based classifier exceeds half-space and ordinal depth classifiers in solving a broad spectrum of practical classification problems. Since the extrapolation depth-based classifier is easily modified, it can be applied to the global class of parametric models, whereas linear and quadratic methods of statistics and machine learning are efficiently executed only under the condition of distribution normality. Moreover, the proposed classifier allows one to get rid of the "curse of dimensionality" as to the exponential growth of necessary experimental data depending on space dimension in solving problems of probabilistic-statistical pattern recognition and classification. Hence, in processing small samples in a high-dimensional space, the extrapolation depth-based classifier exceeds usual nonparametric methods when data sets are almost elliptical. Note that a multilevel smoothing structure allows one to investigate global properties of density functions and class boundaries. As a result, in practice, the proposed multilevel method is rather flexible owing to the aggregation of results for different smoothing scales.

REFERENCES

1. J. I. Chacon, T. Duong, and M. P. Wand, "Asymptotics for general multivariate kernel density derivative estimators," Statist., **21**, 810–837 (2011).
2. P. J. Rousseeuw and A. Struyf, "Characterizing angular symmetry and regression symmetry," in: Statist. Plann. Inference, **122**, Issues 1–2, 163–173 (2004).
3. C. C. Holmes and N. M. Adams, "A probabilistic nearest neighbor method for statistical pattern recognition," Journal of the Royal Statistical Society, **64**, 295–306 (2002).
4. H. Oja and D. Paindaveine, "Optimal signed-rank tests based on hyperplanes," J. Statist. Plann. Inference, **135**, 307–321 (2005).
5. T. Lange, K. Mosler, and P. Mozharovskyi, "Fast nonparametric classification based on data depth," Statist. Papers, **55**, 53–67 (2014).
6. F. Godtliebsen, J. S. Marron, and P. Chaudhuri, "Significance in scale space for bivariate density estimation," Journal of Computational and Graphical Statistics, **11**, 1–22 (2002).
7. D. Pollard, Convergence of Stochastic Processes, Springer, New York (1984).
8. Y. Zuo and R. Serfling, "Structural properties and convergence results for contours of sample statistical depth functions," The Annals of Statistics, **28**, 484–497 (2000).
9. Y. Vardi and C. H. Zhang "The multivariate on L_1-median and associated data depth," in: Proceedings of the National Academy of Sciences of the United States of America, **97** (2000), pp. 1423–1426.