EMPIRICAL BAYESIAN CLASSIFICATION
FOR OBSERVATIONS WITH ADMIXTURE

UDC 519.21

O. SUGAKOVA

ABSTRACT. We consider the problem of classification of an object by using observations after a numerical characteristic under the assumption that each object belongs to one of the two given classes. The distribution of the characteristic is unknown for objects of the first class but is assumed to be symmetric. The distribution for the second class is known. We construct an empirical Bayesian classifier and prove a result concerning the asymptotic behavior of the error probability.

1. Setting of the problem

Consider the problem of the classification of an object \( O \) by using observations after its numerical characteristic \( \xi = \xi(O) \in \mathbb{R} \). The object may belong to one of the two populations, either to \( \mathbb{R}_1 \), the population of the main component, or to \( \mathbb{R}_0 \), the population of an admixture. The number of the population which an object \( O \) belongs to is denoted by \( \text{ind}(O) \).

The learning sample \( \xi_1, \ldots, \xi_n \) used to construct a classifier is a sample from a mixture of two components. The members of the sample are independent identically distributed random variables with the probability density

\[
\psi(x) = pf(x - a) + (1 - p)f_0(x),
\]

where \( p \in (0, 1) \) is the concentration of the main component, \( f_0(x) \) the known probability density of the admixture, and \( a \) an unknown median of the distribution of the main component whose density \( f(x - a) \) is unknown but known to be symmetric, that is, \( f(-x) = f(x) \). The parameters \( a \) and \( p \) are also unknown. The consideration of such a model is motivated by the problem of detection of differentially expressed genes under two or more conditions in microarray data (see, for example, [1, 2]).

Our aim is to construct the optimal rule

\[
g: X \to \{0, 1\}
\]

for the classification of objects, where \( X \) is the set of all observed characteristics and where 0 and 1 are the numbers of the classes. The quality of a classifier is determined by the error probability

\[
L(g) = P\{g(\xi(O)) \neq \text{ind}(O)\} = 1 - P\{g(\xi(O)) = \text{ind}(O)\} = 1 - pP\{\xi(O) \in \mathbb{R}_1/\text{ind}(O) = 1\} - (1 - p)P\{\xi(O) \in \mathbb{R}_0/\text{ind}(O) = 0\}.
\]

Fix a certain set \( G \) of admissible classifiers. A classifier \( g_0 \) is said to be Bayesian in the class \( G \) if \( g_0 \in G \) and \( L(g_0) \leq L(g) \) for all \( g \in G \). A Bayesian classifier is denoted by \( g^B \).

2010 Mathematics Subject Classification. Primary 62G05; Secondary 62G20.
It is known that a Bayesian classifier is such that
\[
g^B(x) = I\{pf(x - a) > (1 - p)f_0(x)\},
\]
where \(I\{A\}\) denotes the indicator of the set \(A\) (see [2, p. 149]).

The probability of error for this classifier is equal to
\[
L\left(g^B\right) = 1 - \int_{-\infty}^{\infty} \max(pf(x - a), (1 - p)f_0(x)) \, dx.
\]

Unfortunately one cannot use equalities (1) and (2) in practice to construct a classifier and to estimate its quality, since these equalities contain unknown parameters \(a, p, \) and \(f(x).\)

Thus we construct the so-called empirical Bayesian classifier by substituting the consistent estimators instead of unknown parameters in equalities (1) and (2). Below we study such consistent estimators.

The so-called GEE estimator \(\hat{a}_n\) of the parameter \(a\) is proposed in the paper [3]. The estimator \(\hat{a}_n\) is constructed by the method of generalized estimating equations, namely \(\hat{a}_n\) is a root of the equation
\[
\hat{h}(\hat{a}_n) = 0,
\]
where
\[
\hat{h}(\alpha) = \frac{1}{n} \sum_{j=1}^{n} \left( g_1(x - \alpha)G_2(\alpha) - g_2(x - \alpha)G_1(\alpha) \right).
\]

Here \(g_1(x)\) and \(g_2(x)\) are some given odd functions, and
\[
G_i(\alpha) = \int_{-\infty}^{\infty} g_i(x - \alpha)f_0(x) \, dx, \quad i = 1, 2.
\]

**Lemma 1.1** ([3]). Let \(\hat{a}_n\) be a GEE estimator defined as a root of equation (3). Assume further that

(i) \(\hat{a}_n\) is a consistent estimator;
(ii) the functions \(g_i'(x)\) and \(G_i'(\alpha)\), \(i = 1, 2\), are continuous in \(R\) and such that
\[
E\left((g_i(\xi_1 - a))^2\right) < \infty, \quad i = 1, 2; \quad E\left|\frac{\partial}{\partial a} \hat{h}(\xi_1, \alpha)\right|_{\alpha = a} \neq 0,
\]
and
\[
E \sup_{\alpha: |a - \alpha| < \varepsilon} (g_i'(\xi_1 - \alpha))^{1+\delta} < \infty, \quad i = 1, 2,
\]
for some \(\varepsilon > 0\) and \(\delta > 0\).

Then
\[
\hat{a}_n - a = O_p\left(\frac{1}{\sqrt{n}}\right), \quad n \to \infty.
\]

An estimator \(\hat{p}_n\) of the parameter \(p\) is constructed in the paper [4]:
\[
\hat{p}_n = \frac{1}{\hat{a}_n - m_0}\left(\frac{1}{n} \sum_{j=1}^{n} \xi_j - m_0\right),
\]
where \(m_0 = \int_{-\infty}^{\infty} xf_0(x) \, dx\) is the mean value of the admixture.

**Lemma 1.2** ([4]). Let conditions (i) and (ii) of Lemma 1.1 hold. Then
\[
\hat{p}_n - p = O_p\left(\frac{1}{\sqrt{n}}\right), \quad n \to \infty,
\]
if \(m_0 \neq a\).
The kernel estimators for the density of the main component are constructed in the paper [4]:

\[ \hat{f}_n(x) = \frac{1}{\hat{p}_n} \left( \frac{1}{nh} \sum_{j=1}^{n} K \left( \frac{x + \hat{a}_n - \xi_j}{h_n} \right) - (1 - \hat{p}_n) f_0(x + \hat{a}_n) \right), \]

where \( K(x) \) is a kernel (in other words, \( K(x) \) is a probability density in \( \mathbb{R} \)), and \( h_n, n \geq 1, \) is a smoothing parameter such that \( h_n \to 0 \) and \( nh_n \to \infty \) as \( n \to \infty \).

The paper [1] is the first to consider estimator (7) for the density. Note however that the estimators for \( a \) and \( p \) studied in [1] differ from those used in the current paper.

The following symmetrized estimator for the density \( f(x) \),

\[ \tilde{\hat{f}}_n(x) = \frac{\hat{f}_n(x) + \hat{f}_n(-x)}{2}, \]

is also considered in the paper [4].

2. Construction of the classifier

Now we are ready to construct the empirical Bayesian classifier.

Assume that the set of solutions of the equation \( pf(x - a) = (1 - p) f_0(x) \) is finite. Denote the cardinality of this set by \( m \) and let \( t_1, \ldots, t_m \) be its elements. We assume that all of them belong to the interval \([-R, R]\), where \( R \) is a known number. Let

\[ S(x) = \begin{cases} x & \text{if } x \in [-R, R]; \\ R & \text{if } x > R; \\ -R & \text{if } x < -R. \end{cases} \]

Substituting estimators (3), (5), and (7) in equality (1) instead of the Euclidean parameters \( a, p, \) and the density of the main component \( f(x - a) \), respectively, and using the function \( S(x) \) to control the behavior of the density outside the interval \([-R, R]\) we obtain the truncated empirical Bayesian classifier

\[ \hat{g}(x) = I \left\{ \hat{p}_n \hat{f}_n(S(x) - \hat{a}_n) > (1 - \hat{p}_n) f_0(S(x)) \right\}. \]

The probability of the erroneous classification for the truncated empirical Bayesian classifier is given by

\[ L(\hat{g}) = (1 - p) \int_{-\infty}^{\infty} \hat{g}(x) f_0(x) \, dx + p \int_{-\infty}^{\infty} (1 - \hat{g}(x)) f(x - a) \, dx \]

if a learning sample \( \xi_1, \ldots, \xi_n \) is fixed, while the probability of error for the Bayesian classifier (2) can be written as

\[ L(g) = (1 - p) \int_{-\infty}^{\infty} g(x) f_0(x) \, dx + p \int_{-\infty}^{\infty} (1 - g(x)) f(x - a) \, dx. \]

3. Main results

Our aim is to study the asymptotic behavior of the difference \( L(\hat{g}) - L(g) \) as \( n \to \infty \).

By \( c, c_1, c_2, \ldots, c_k, \) we denote some real constants.

In what follows we use the following assumptions.

(iii) Assumption imposed on the kernel: \( K(x) \) is a function with finite support,

\[ \text{Var}_R K'(x) < c_1, \quad \int_{-\infty}^{\infty} z K(z) \, dz = 0. \]
Let conditions (14) \( \tilde{g}(x) \) and (15) \( L \)
The probability of error for this classifier is equal to

Theorem 3.1. Let conditions (i)–(viii) hold. Then

\[ n^{4/5} (L(\tilde{g}) - L(g)) \Rightarrow \sum_{i=1}^{m} \eta_i^2, \]

where the \( \eta_i \) are independent Gaussian random variables with parameters

\[ N \left( \frac{D^2}{2} \sqrt{\beta_i} \gamma_i, \frac{d^2 \psi(t_i)}{\beta_i^3} \right). \]

Here

\[ \gamma_i = \frac{\psi''(t_i)}{(1-p) f_0'(t_i) - pf'(t_i - a)}, \quad \beta_i = \left| (1-p) f_0'(t_i) - pf'(t_i - a) \right|, \]

\[ D = \int_{-\infty}^{\infty} z^2 K(z) dz, \quad d^2 = \int_{-\infty}^{\infty} K^2(z) dz. \]

Now we construct another truncated empirical Bayesian classifier by substituting the symmetrized estimator \( \hat{g} \) instead of \( \hat{f}(x) \) in (9). Thus the estimator is given by

\[ \tilde{g}(x) = I \left\{ \hat{p}_n \hat{f}_n(S(x) - \hat{a}_n) > (1 - \hat{p}_n) f_0(S(x)) \right\}. \]

The probability of error for this classifier is equal to

\[ L(\tilde{g}) = (1-p) \int_{-\infty}^{\infty} \tilde{g}(x) f_0(x) dx + p \int_{-\infty}^{\infty} (1-\tilde{g}(x)) f(x-a) dx. \]

Theorem 3.2. Let conditions (i)–(viii) hold. Then

\[ n^{4/5} (L(\tilde{g}) - L(g)) \Rightarrow \sum_{i=1}^{m} \zeta_i^2, \]

where the \( \zeta_i \) are independent Gaussian random variables with the parameters

\[ N \left( \frac{D^2 C^2}{4} \sqrt{\beta_i} \alpha_i, \frac{d^2 (\psi(t_i) + \psi(-t_i))}{4C^3 \beta_i} \right). \]

Here

\[ \alpha_i = \frac{\psi''(t_i) + \psi''(-t_i)}{(1-p) f_0'(t_i) - pf'(t_i - a)}. \]
4. Proofs of the results

First we prove some auxiliary results. Denote by \( \hat{m}_n \) the number of roots of the equation

\[
\hat{p}_n \hat{f}_n (t - \hat{a}_n) = (1 - \hat{p}_n) f_0(t)
\]

in the interval \([-R, R]\). Let \( \hat{t}_i, i = 1, \ldots, m \), be the roots of equation (16).

**Lemma 4.1.** Assume that conditions (iv)–(vi) hold. Then

\[
P\{ \hat{m}_n = m \} = 1, \quad n \to \infty.
\]

In addition,

\[
\hat{t}_i - t_i \xrightarrow{P} 0, \quad n \to \infty,
\]

for \( i = 1, \ldots, m \).

**Proof of Lemma 4.1.** Put

\[
h(t) = p f(t - a) - (1 - p) f_0(t), \quad \hat{h}_n(t) = \hat{p}_n \hat{f}_n (t - \hat{a}_n) - (1 - \hat{p}_n) f_0(t).
\]

Using equality (7) in the latter relation we get

\[
\hat{h}_n(t) = \hat{\psi}_n(t) - \psi(t) + 2(\hat{p}_n - p) f_0(t) + h(t) = 0,
\]

where

\[
\hat{\psi}_n(t) = \frac{1}{n \hat{h}_n} \sum_{j=1}^{n} K \left( \frac{t - \xi_j}{\hat{h}_n} \right)
\]

is the kernel estimator for the density of the mixture.

It is proved in [2, p. 172] that

\[
\sup_t \left| \hat{\psi}_n(t) - \psi(t) \right| \xrightarrow{P} 0, \quad n \to \infty,
\]

if conditions (iii), (v), and (vii) hold. Since \( \hat{p}_n - p \xrightarrow{P} 0 \) and \( f_0(t) \) is a bounded function, the second term in (17) also approaches 0 uniformly with respect to \( t \). Thus

\[
\sup_{t \in [-R,R]} \left| \hat{h}_n(t) - h(t) \right| \xrightarrow{P} 0, \quad n \to \infty.
\]

Fix \( \varepsilon > 0 \) and let \( A = \bigcap_{i=1}^{m} \{ t: |t - t_i| \geq \varepsilon \} \). Then there exists \( \delta > 0 \) such that \( \inf_{t \in A} |h(t)| > \delta > 0 \).

\[
P\left\{ \exists n_0: |\hat{h}_n(t) - h(t)| < \frac{\delta}{2} \text{ for all } n > n_0 \text{ and } t \right\} \to 1, \quad n \to \infty,
\]

we have

\[
P\left\{ \inf_{t \in A} |\hat{h}_n(t)| > \frac{\delta}{2} \right\} \to 1, \quad n \to \infty.
\]

Therefore, for all \( \varepsilon > 0 \),

\[
P\left\{ \exists n_0: \forall n > n_0, \forall \hat{t}: |\hat{h}_n(\hat{t})| = 0 \text{ there is } t_i: h(t_i) = 0, |\hat{t} - t_i| < \varepsilon \right\} \to 1 \quad \text{as } n \to \infty.
\]

This means that if \( n \) is sufficiently large, then with probability close to 1, all the roots of equation (16) are as close as one wishes to the roots of the equation \( h(t) = 0 \).

Now we prove that the numbers of the roots of these two equations are the same. The following result is proved in [4].
Lemma 4.2. Assume that conditions (i)–(v) hold. Then
\[
\sup_t |\hat{\psi}_n'(t) - \psi'(t)| = O_p\left(\frac{\ln n}{n^{1/10}}\right), \quad n \to \infty.
\]

For definiteness, assume that \(pf'(t_i - a) - (1 - p)f'_0(t_i) > b > 0\). Then there exists a neighborhood \(\Theta_i\) of the point \(t_i\) such that \(pf'(t_i - a) - (1 - p)f'_0(t_i) > b > 0\) for an arbitrary \(t \in \Theta_i\). It is also obvious that
\[
\hat{p}_n(f_n'(t - \hat{a}_n) - (1 - \hat{p}_n)f'_0(t)) = \hat{\psi}_n(t) - \psi'(t) + 2(\hat{p}_n - p)f'_0(t) + pf'(t - a) - (1 - p)f'_0(t)
\]
whence
\[
> b + O_p\left(\frac{\ln n}{n^{1/10}}\right) + O_p\left(\frac{1}{\sqrt{n}}\right).
\]

This implies that
\[
P\left\{\hat{p}_n(f_n'(t - \hat{a}_n) - (1 - \hat{p}_n)f'_0(t) > b\right\} \to 1, \quad n \to \infty,
\]
for \(t \in \Theta_i\). Hence
\[
P\{\text{equation (16) has a unique root in } \Theta_i\} \to 1, \quad n \to \infty.
\]

Lemma 4.1 is proved. \(\square\)

Lemma 4.3. Let conditions (iv)–(vi) and (viii) hold. Then
\[
n^{1/2}\begin{pmatrix} \hat{t}_1 - t_1 \\ \vdots \\ \hat{t}_m - t_m \end{pmatrix} \Rightarrow \begin{pmatrix} \frac{D^2}{2} \gamma_1 + \frac{\zeta_1}{\beta_1} \\ \vdots \\ \frac{D^2}{2} \gamma_m + \frac{\zeta_m}{\beta_m} \end{pmatrix}, \quad n \to \infty.
\]

Moreover, the random vector \((\zeta_1, \ldots, \zeta_m)^T\) has the Gaussian \(N(0, B)\) distribution, where
\[
B = d^2 \text{diag}(\psi(t_1), \ldots, \psi(t_m)).
\]

Proof of Lemma 4.3. We expand the left hand side of equality (16) with the help of the Taylor formula at the point \(t_i\), and take only the first term of the expansion into account. Then we get
\[
\hat{p}_n(f_n'(t_i - \hat{a}_n) - (1 - \hat{p}_n)f_0(t_i) + (\hat{p}_n(f_n'(t - \hat{a}_n) - (1 - \hat{p}_n)f'_0(t)))(\hat{t}_i - t_i) \approx 0,
\]
whence
\[
\hat{t}_i - t_i \approx -\frac{\hat{p}_n(f_n'(t_i - \hat{a}_n) - (1 - \hat{p}_n)f_0(t_i))}{\hat{p}_n(f_n'(t - \hat{a}_n) - (1 - \hat{p}_n)f'_0(t))}.
\]

Using the equality \(pf(t_i - a) - (1 - p)f_0(t_i) = 0\), we rewrite the latter relation as
\[
(18) \quad \hat{t}_i - t_i \approx \frac{A_{ni} + B_{ni}}{C_{ni}},
\]
where
\[
A_{ni} = \hat{p}_n(f_n'(t_i - \hat{a}_n) - pf(t_i - a)), \quad B_{ni} = (\hat{p}_n - p)f_0(t_i), \quad C_{ni} = \hat{p}_n(f_n'(t - \hat{a}_n) - (1 - \hat{p}_n)f'_0(t)).
\]

As proved above,
\[
(19) \quad B_{ni} = O_p\left(\frac{1}{\sqrt{n}}\right), \quad n \to \infty.
\]
Consider the terms $C_{ni}$. Lemma 4.2 and relation (19) imply that
\begin{equation}
C_{ni} = \hat{\psi}_n'(t_i) - \psi'(t_i) + pf'(t_i) - (1 - p)f'_0(t_i) + (\hat{p}_n - p)f'_0(t_i)
\end{equation}
\begin{equation}
= pf'(t_i - a) - (1 - p)f'_0(t_i) + O_p \left( \frac{1}{\sqrt{n}} \right) + O_p \left( \frac{\ln n}{n^{1/10}} \right)
\end{equation}
\begin{equation}
\rightarrow pf'(t_i - a) - (1 - p)f'_0(t_i).
\end{equation}

Next we consider
\begin{equation}
A_{ni} = \frac{1}{n\epsilon_n} \sum_{j=1}^{n} K \left( \frac{t_i - \xi_j}{\epsilon_n} \right) - pf(t_i - a) - (1 - p)f_0(t_i).
\end{equation}

It is known that the set of conditions (iii), (iv), and (v) imply
\begin{equation}
A_{ni} = n^{-2/5} \left( \frac{D^2}{2} \psi''(t_i) + \zeta_n(t_i) \right) + o(1),
\end{equation}
\begin{equation}
\zeta_n(t_i) = \frac{1}{\sqrt{\epsilon_n n}} \sum_{j=1}^{n} \left[ K \left( \frac{t_i - \xi_j}{\epsilon_n} \right) - KE \left( \frac{t_i - \xi_j}{\epsilon_n} \right) \right]
\end{equation}
(see [5 Chapter 1, Section 10]).

Combining (18)–(21), we see that
\begin{equation}
n^{\frac{1}{2}} \begin{pmatrix}
\hat{t}_1 - t_1 \\
\vdots \\
\hat{t}_m - t_m \\
\end{pmatrix}
\approx \begin{pmatrix}
\frac{D^2}{2} \gamma_1 + O_p \left( n^{-1/10} \right) \\
\vdots \\
\frac{D^2}{2} \gamma_m + O_p \left( n^{-1/10} \right) \\
\end{pmatrix} + \begin{pmatrix}
\frac{\zeta_1}{\beta_1 + O_p \left( \frac{\ln n}{n^{1/10}} \right)} \\
\vdots \\
\frac{\zeta_m}{\beta_m + O_p \left( \frac{\ln n}{n^{1/10}} \right)} \\
\end{pmatrix}.
\end{equation}

The proof of Lemma 4.3 is completed in the same way as in [4] by using the central limit theorem for the scheme of series $(\zeta_n(t_i))_{i=1}^{m}$ (similarly to the proof in [5] Chapter 1, Section 10).

Proof of Theorem 3.1 For definiteness, assume that $pf(x - a) > (1 - p)f_0(x)$ in $(-\infty, t_1)$. By Lemma 4.1 $m_n = m$ with probability close to 1 if $n$ is sufficiently large.

Since
\begin{equation}
I \{ \hat{p}_n f_n(S(x) - \hat{a}_n) > (1 - \hat{p}_n) f_0(S(x)) \} = I \{ \hat{p}_n f_n(R - \hat{a}_n) > (1 - \hat{p}_n) f_0(R) \}
\end{equation}
\begin{equation}
= I \{ pf(R - a) > (1 - p) f_0(R) \}
\end{equation}
for $x > R$ and sufficiently large $n$ and since an analogous equality holds for $x < -R$, as well, the integrals (10) and (11) coincide in the region $|x| > R$.

Condition (vi) implies that
\begin{equation}
L(\hat{g}) - L(g) = \sum_{i=1}^{m} (-1)^{i+1} \int_{t_i}^{t_i} (pf(x - a) - (1 - p)f_0(x)) \, dx.
\end{equation}

We use the Taylor expansion at the point $t_i$ for the above integral. Then
\begin{equation}
\int_{t_i}^{t_i} (pf(x - a) - (1 - p)f_0(x)) \, dx = (pf(t_i - a) - (1 - p)f'_0(t_i)) \left( \hat{t}_i - t_i \right)
\end{equation}
\begin{equation}
+ \frac{(\hat{t}_i - t_i)^2}{2} (pf'(t_i - a) - (1 - p)f''_0(t_i)) + o \left( (\hat{t}_i - t_i)^2 \right).
\end{equation}
The first term on the right hand side of (22) equals 0. Further, the random variables $n^{2/5}(\hat{t}_i - t_i)$ are uncorrelated and asymptotically normal with parameters

$$N \left( \frac{D^2}{2} \gamma_i, \frac{d^2 \psi(t_i)}{\beta_i^2} \right)$$

by Lemma 4.3. The expression

$$(-1)^i + 1 \left( pf'(t_i - a) - (1 - p)f'_0(t_i) \right)$$

is nonnegative and thus it can be replaced by

$$|pf'(t_i - a) - (1 - p)f'_0(t_i)|.$$ 

This completes the proof of Theorem 3.1.

The proof of Theorem 3.2 is analogous.

**BIBLIOGRAPHY**


2. R. E. Maı˘boroda and O. V. Sugakova, *Estimation and classification after observations in a mixture*, Kyiv University, Kyiv, 2008. (Ukrainian)


DEPARTMENT OF MATHEMATICS AND THEORETICAL RADIO PHYSICS, FACULTY OF RADIO PHYSICS, NATIONAL TARAS SHEVCHENKO UNIVERSITY, ACADEMICIAN GLUSHKOV AVENUE 4E, KYIV 03127, UKRAINE

E-mail address: sugak@univ.kiev.ua

Received 21/JUN/2010

Translated by O. I. KLESOV