*LMD konf. darbai,* 187–192
1997, Vilnius

# An asymptotic expansion of the expected regret risk of classification under double inverse sampling scheme

**K. Dučinskas (KU)**

Suppose that individuals comes from one of two distinct populations $\Omega_1, \Omega_2$, with positive prior probabilities $\pi_1, \pi_2$, respectively, where $\sum_{i=1}^{2} \pi_i = 1$. The population of the individual is given by the random 2-dimensional vector $Y = (Y^1, Y^2)'$ of zero-one variables, similar as in remarkable monograph of G. J. McLachlan (see [1]) devoted to the discriminant analysis and statistical pattern recognition. The $i$-th component of $Y$ is defined to be one or zero according as an individual belongs or does not belong to the population $\Omega_i$ $(i = 1, 2)$. Then $Y$ is distributed according to a multinomial distribution consisting of one draw on 2 populations with probabilities $\pi_1, \pi_2$, respectively; that is

$$P(Y = y) = \pi_1^{y^1} \pi_2^{y^2}, \quad \text{or} \quad Y \sim \mathrm{Mult}_2(1, \pi)$$

with $\pi = (\pi_1, \pi_2)'$; $y = (y^1, y^2)'$, where the prime denotes the vector transpose.

Let $X \in \chi \subset R^P$ a $p$-dimensional random feature vector which is measured on each individual. Assume that the distribution of $X$ for the individual from $\Omega_i$ has probability densities $p_i(x)$ which belongs to the parametric family of regular densities $F_i = \{p_i(x; \theta_i),\ \theta_i \in \Theta \subset R^m\}$. In this framework, the topic of classification is concerned with the relationship between the population–membership label $Y$, and the feature vector $X$. A decision is to be made as to which population an individual randomly chosen from $\Omega = \bigcup_{i=1}^{2} \Omega_i$, belongs on the bases of an observed value of $X$. Let $d(\cdot)$ denotes a classification rule formed for this purpose, where $d(x) = i$ implies that an individual with feature vector $X = x$ is to assigned to the population $\Omega_i$ $(i = 1, 2)$. Let $C(i, j)$ denote the cost of allocation when an individual from $\Omega_i$ is allocated to $\Omega_j$ and let $C(i, j)$ always be finite, i.e. $\max_{i,j=1,\dots,L} C(i, j) = C^* < \infty$.

When prior probabilities $\{\pi_i\}$ and densities $\{p_i(x)\}$ are known, the risk $R(d(\cdot))$ associated with rule $d(\cdot)$ can be expressed as

$$R(d(\cdot)) = \sum_{i=1}^{2} \pi_i \int_{\chi} C\big(i, d(x)\big) p_i(x)\, dx.$$

Then Bayes classification rule (BCR) $d_B(x)$ minimising the risk $R(d(\cdot))$ is defined as

$$d_B(x) = \arg\max_{j=1,2} \sum_{i=1}^{2} l_i\, p_i(x),$$

where $l_i = \pi_i(C(i, 3-i) - C(i, j))$, $(i = 1, 2)$.

In practical applications, the density functions $\{p_i(x)\}$ are seldom completely known. Often they are only known up to the parameters $\{\theta_i\}$, i.e., we may only assert that $p_i(x)$ is one of the parametric family of density functions $F_i$.

Suppose that in order to estimate unknown parameters $\theta_1, \theta_2$, there are $M$ individuals of known origin on which feature vector $X$ has been recorded.

That data defined by $T = \{(X_1', Y_1')', \dots, (X_M', Y_M')'\}$ is referred to in literature as training sample, where prime denotes vector transpose. The only case of independent $(X_i', Y_i')'$ $(i = 1, \dots, M)$ will be considered in this paper.

There are two major sampling designs under which the training data $T$ may be realised, separate sampling (SS) and mixture sampling (MS). They correspond, respectively, to sampling from the distributions of $X$ conditionally on $Y = y$ and to sampling from joint distribution of $(X', Y')'$. With separate sampling in practice, the feature vectors are observed for a sample of $M_i$ individuals taken separately from each population $\Omega_i$ $(i = 1, 2)$. The MS design applies to the situation where the feature vector and population of origin are recorded on each of $M$ individuals drawn from mixture of the possible populations.

Sometimes under mixture sampling design we need to continue sampling until both $M_i \geqslant k_i$ with $0 < k_i < +\infty$ $(i = 1, 2)$. This sampling design is called double inverse sampling (DIS) (see [2]). Under DIS total training sample size $M$ and one of $M_i = \sum_{j=1}^{M} Y_j^i$ are random $(i = 1, 2)$.

The distributions of $M_i$ are

$$P(M_1 = m_1) = \begin{cases} \sum_{j=0}^{k_1} C^j_{k_2+j-1}\pi_1^j\pi_2^{k_2}, & \text{if } m_1 = k_1, \\ C^{m_1}_{k_2+m_1-1}\pi_1^{m_1}\pi_2^{k_2}, & \text{if } m_1 > k_1, \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

$$P(M_2 = m_2) = \begin{cases} \sum_{j=0}^{k_2} C^j_{k_1+j-1}\pi_1^{k_1}\pi_2^{j}, & \text{if } m_2 = k_2, \\ C^{m_2}_{k_2+m_2-1}\pi_1^{k_1}\pi_2^{m_2}, & \text{if } m_2 > k_2, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

$$P(M_1 = m_1,\ M_2 = m_2) = \begin{cases} C^{k_1}_{k_1+k_2}\pi_1^{k_1}\pi_2^{k_2}, & \text{if } m_i = k_i,\ i = 1, 2, \\ C^{m_1}_{k_2+m_1-1}\pi_1^{m_1}\pi_2^{k_2}, & \text{if } m_2 = k_2,\ m_1 > k_1, \\ C^{m_2}_{k_1+m_2-1}\pi_1^{k_1}\pi_2^{m_2}, & \text{if } m_2 > k_2,\ m_1 = k_1, \\ 0, & \text{otherwise}. \end{cases} \tag{3}$$

Only the DIS scheme for training is considered in this paper.

Suppose that densities $p_i(x)$ are taken to belong to the same parametric family, i.e., $F_i = F = (p(x; \theta_i),\ \theta_i \in \Theta \subset R^m)$ $(i = 1, 2)$ and assume that prior probabilities $\pi_1, \pi_2$ are known.

Let $\alpha = (\theta_1', \theta_2')$ and $(\theta_1^j \neq \theta_2^j)$, $j = 1, \dots, m$.

The so-called estimative approach to the choice of sample-based classification rule is used. The unknown parameters $\theta_1, \theta_2$ in the BCR are replaced by appropriate estimates $\hat{\theta}_1, \hat{\theta}_2$ obtained from the training data $T$. Hence this sample-based classification rule $d_s(x, \hat{\alpha})$ is defined by

$$d_s(x, \hat{\alpha}) = \arg\max_{j=1,2} \sum_{i=1}^{2} l_i p_i(x, \hat{\theta}_i).$$

The actual risk for the rule $d_s(x, \hat{\alpha})$ is the risk of classifying a randomly selected individual with feature $X$ and is designated by

$$R_A(\hat{\alpha}) = \sum_{i=1}^{2} \pi_i \int_X \big(C(i, d_S(x, \hat{\alpha}))\big) p_i(x)\, dx, \text{ where } \hat{\alpha} = (\hat{\theta}_1', \hat{\theta}_2')'.$$

For notational convenience, $R_A(\hat{\alpha})$ sometimes will be abbreviated to $R_A$.

The expected risk regret (ERR) is defined as the expectation of the difference between actual risk $R_A$ and the Bayes risk $R_B$ that would be obtained if all parameters were known, i.e., $R_B = R_A(\alpha)$. Thus

$$ERR = E_T\big\{R_A(\hat{\alpha})\big\} - R_B,$$

where $E_T\{R_A\}$ denotes the expectation with respect to distribution of $T$.

The purpose of this paper is to find an asymptotic expansions of ERR when maximum likelihood estimates (MLE) of unknown parameters obtained from $T$ under DIS scheme for $T$ are used.

This is an extension of the results of Dučinskas (see [3–4]). In [3] the asymptotic expansion of expected error regret in the situation when only SS scheme was used for training was given, in [4] the asymptotic expansion of expected error regret in classification of mixed random vectors under DIS scheme for training was presented.

The following notation will be used in this paper.

Let $\nabla_\alpha$ be the vector partial differential operator given by

$$\nabla_\alpha^T = \left(\frac{\partial}{\partial \alpha^1}, \ldots, \frac{\partial}{\partial \alpha^m}\right)$$

and

$$|\nabla_\alpha|^2 = \sum_{i=1}^{m} \left(\frac{\partial}{\partial \alpha^i}\right)^2$$

for any $\alpha = (\alpha^1, \ldots, \alpha^m)' \in R^m$.

Similarly, $\nabla_\alpha^2$ denotes the matrix second order differential operator

$$\nabla_\alpha^2 = \left\| \frac{\partial^2}{\partial \alpha_i \partial \alpha_j} \right\|_{i,j=1,\ldots,m}$$

$$G_0(x) = l_1 p_1(x) - l_2 p_2(x),$$
$$\Gamma = \left\{x \in R^p\colon\ G_0(x) = 0\right\},$$

and $\gamma$ is the Lebesque measure on $\Gamma$.

Assume that $I_i$ denotes the $m \times m$ Fisher information matrix based on $\theta_i$, i.e.

$$I_i = E_i\left\{\nabla_{\theta_i} \ln p_i(x) \nabla'_{\theta_i} \ln p_i(x)\right\},$$

where $E_i\{\cdot\}$ represents the expectation based on distribution with density function $p_i(x)$ $(i = 1, 2)$.

Let $M_i$ denote the number of observations from $\Omega_i$ $(i = 1, 2)$. Define the following regularity conditions R:

a) for fixed $M_i$ MLE $\hat{\theta}_i$ admits the following stochastic expansion in some neighbourhood of $\theta_i$

$$\hat{\theta}_i = \theta_i + h_1^i M_i^{-1/2} + o_{P_i}(M_i^{-1/2}),$$

where $o_{P_i}(M_i^{-1/2})/M_i^{-1/2} \xrightarrow{P_i} 0$, $h_1^i$ depends on the distribution $P_i$ with the probability density $p_i(x)$, but does not depend on $M_i$.

b) second-order derivatives of $E_T(R_A)$ with respect to $\alpha$ exist and are continuous in some neighbourhood of $\alpha$.

THEOREM 1([3]). *If regularity conditions R for $\{p_i(x)\}_{i=1,2}$ hold and SS scheme for $T$ is used, then*

$$EER = \frac{1}{2}\sum_{i=1}^{2} \alpha_i/M_i + o(M_0^{-1}), \tag{4}$$

*where*

$$\alpha_i = l_i^2 \int_\Gamma \nabla_{\theta_i}^1 p_i(x) I_i^{-1} \nabla_{\theta_i} p_i(x) |\nabla_x G_0(x)|^{-1}\, d\gamma, \tag{5}$$

$M_0 = \min(M_1, M_2)$.

THEOREM 2. *If regularity conditions R for $\{p_i(x)\}_{i=1,2}$ hold and DIS scheme for $T$ is used,* $\min\{k_i\}_{i=1,2} = k_0 \to \infty$ *and* $\lim k_1/k_2 = \lambda > 0$ *then*

$$EER = \frac{1}{2}\sum_{i=1}^{2} \alpha_i/d_i + o(k_0^{-1}),$$

*where $\alpha_i$ is defined in* (5) *and* $d_i = (1 \vee (\lambda_0/\lambda)^{3-2i} k_i$ $(i = 1, 2)$, $\lambda_0 = \pi_1/\pi_2$.

*Proof*. The proof of the theorem is based on the applications of the rezults of Theorem 1 for the conditional expectation of $R_A$ under the condition $\{M_i = m_i,\ i = 1, 2\}$.

Then

$$E_T\{R_A\} = \sum_{m_1=k_1}^{\infty} \sum_{m_2=k_2}^{\infty} E_T\{R_A \mid \{M_i = m_i\}_{i=1,2}\} \cdot P(M_1 = m_1, M_2 = m_2)$$
$$= R_0 + \frac{1}{2}\sum_{i=1}^{2} \alpha_i E(1/M_i) + o(k_0^{-1}). \tag{7}$$

Using (1), (2) we obtain

$$E(1/M_1) = I_{\pi_2}(k_2, k_1 + 1)/k_1 + \frac{1}{\lambda_0(k_2 - 1)} I_{\pi_1}(k_1 + 2; k_2 - 2) + o\left(\frac{1}{k_0^2}\right), \tag{8}$$

$$E(1/M_2) = I_{\pi_1}(k_1, k_2 + 1)/k_2 + \frac{\lambda_0}{(k_1 - 1)} I_{\pi_2}(k_2 + 2; k_1 - 2) + o\left(\frac{1}{k_0^2}\right), \tag{9}$$

where

$$I_x(p; q) = \frac{1}{B(p; q)} \int_0^x t^{p-1}(1 - t)^{q-1}\, dt$$

is incomplete beta-function.

When $p, q \to \infty$ we can use Wishart approximation

$$I_x(p; q) = \Phi(u) + o(1),$$

where

$$u = \sqrt{\frac{pq}{p+q}} \ln \frac{qx}{p(1-x)}.$$

Then as $p, q \to \infty$ and $\lim q/p = \lambda$,

$$\lim I_x(p; q) = \begin{cases} 0, & \text{if } \lambda < \frac{1-x}{x}, \\ 1/2, & \text{if } \lambda = \frac{1-x}{x}, \\ 1, & \text{if } \lambda > \frac{1-x}{x}. \end{cases} \tag{10}$$

Applying (10) to (8) and (9) we complete the proof of the theorem.

The results of the paper can be use to find optimal training sample scheme (i.e. optimal values of $k_1, k_2$) and optimal parametric structure for the classification of wide range of distributions belonging to the exponential family.

## REFERENCES

[1] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley & Sons, New York, 1992.

[2] C. T. Tu, C. P. Han, Discriminant analysis based on binary and continuous variables, *Journal of American Statistical Association*, **77** (1982), 447–454.

[3] K. Dučinskas, Optimal training sample allocation and asymtotic expansions for error rates in discriminant analysis, *Acta Applicandae Mathematicae*, **38** (1995), 3–11.

[4] K. Dučinskas, Expansions of expected error regret in discriminant analysis of mixed variables, *Klaipėdos Universiteto mokslo darbai*, **2** (1995), 4–10.

### Laukiamo rizikos prieaugio asimptotinis skleidinys, naudojant inversinę mokymo imtį

*K. Dučinskas*

Straipsnyje nagrinėjamas atsitiktinių vektorių klasifikavimo uždavinys, naudojant mokymo imtis, gautas pagal „inversinę“ schemą. Pateikti laukiamo rizikos prieaugio (expected risk regret) asimptotiniai skleidiniai reguliariems pasiskirstymams.