LMD mokslo darbai, 349–354
1998, Vilnius

The risk of classification in discriminant analysis of mixed variables

K. Dučinskas (KU)

1. Introduction

Suppose that individuals come from one of two mutually exclusive and exhaustive populations Ω_1, Ω_2 with positive prior probabilities π_1, π_2, respectively, where $\sum_{i=1}^{2} \pi_i = 1$. The assignment of individual to one of two populations is based on observation of mixed feature variable $U' = (X', Y')$. The prime denotes vector transpose. Here $X \in \mathbf{X}$ is a p-dimensional vector of continuous variables and $Y' = (Y_1, \ldots, Y_r)$ is an incidence vector corresponding to the vector of b discrete variables. assume without loss of generality that b discrete variates are all binary, each taking on zero or one values. Then $r = 2^b$. The observed values of Y serve to break down the classification problem between Ω_1 and Ω_2 into r locations (cells) wherein the classification is based solely on the continuous observations is performed (Krzanowski, 1975). Then

$$P(Y = y/\Omega_i) = \prod_{c=1}^{r} q_{ic}^{y_c}$$

or $Y \sim \text{Mult}_c(1, q_i)$, where $q_i' = (q_{i1}, \ldots, q_{ic})$,

$$\sum_{c=1}^{r} q_{ic} = 1, \quad 0 < q_{ic}, \ i = 1, 2, c = 1, \ldots, r.$$

Let the pdf $p_{ic}(x)$ of X in the cell c of Ω_i belong to the parametric family $F_{ic} = \{p_{ic}(x; \Theta_{ic}), \Theta_{ic} \in R^m\}$, $i = 1, 2, c = 1, \ldots, r$.

Further, the dependence of any functions on any distribution parameters will be suppressed in the cases when functions are evaluated at the true values of these parameters denoted by an asterisk *, e.g. $p_{ic}(x; \Theta_{ic}^*) = p_{ic}(x)$. Let $d_c(.)$ denote a classification rule (CR) for given $Y_c = 1$. Then $d_c(x) = i$ implies that an individual with continuous feature vector $X = x$ and $Y_c = 1$ is to be assigned to the population Ω_i $(i = 1, 2, c = 1, \ldots, r)$. Let $C(i, j)$ denotes the cost of allocation when an individual from Ω_i is allocated to Ω_j and let $C(i, j)$ always be finite, i.e. $\max_{i,j=1,2} C(i, j) = C_0 < \infty$.

When prior probabilities $\{\pi_i\}$, densities $\{p_{ic}(x)\}$ and $\{q_{ic}\}$ are known, the risk $R(\{d_c(\cdot)\})$ associated with rules $\{d_c(\cdot)\}$ can be expressed as

$$R(\{d_c(\cdot)\}) = \sum_{i=1}^{2} \pi_i \sum_{c=1}^{r} q_{ic} \int_{\mathbf{X}} C(i, d_c(x)) \, p_{ic}(x) dx. \tag{1}$$

Then Bayes classification rule (BCR) $d_{B_C}(x)$ minimizing the risk $R(\{d_c(\cdot)\})$ is defined as

$$d_{B_C}(x) = \arg\max_{i=1,2} l_{ic} p_{ic}(x), \tag{2}$$

where

$$l_{ic} = \pi_i q_{ic} (C(i, 3-i) - C(i,i)) \quad (i = 1, 2, c = 1, \ldots, r). \tag{3}$$

Therefore, Bayes risk R_B is

$$R_B = \sum_{i=1}^{2} \pi_i \sum_{c=1}^{r} q_{ic} \int_{\mathbf{X}} C\left(i, d_{B_C}(x)\right) p_{ic}(x) dx = \inf_{\{d_C(.) \in D\}} R(\{d_c(.)\}), \tag{4}$$

where D is the set of all CR $\{d_c(.)\}$ defined before.

In practical applications, the parameters $\{\Theta_{ic}\}$ and $\{q_{ic}\}$ usually are unknown.

Suppose that in order to estimate unknown parameters there are M individuals of known origin on which feature vector U has been recorded. That data is referred to in pattern recognition literature as training sample (TS). The only case of independent observations in TS will be considered in this paper. Suppose that TS realized under sampling design, which consists of two consequent stages. The first stage usually called separate sampling. This sample often is called stratified sample. Then M_i individuals are randomly taken from each population Ω_i $(i = 1, 2)$. The number of individuals M_{ic} taken from cell c in Ω_i are random and $\sum_{c=1}^{r} M_{ic} = M_i$, $i = 1, 2$.

Suppose that there are m_1 elements of all $\{\Theta_{ic}\}$ known a priori to be distinct and let $0c$ be the vector of $m0$ elements known a priori to be equal i.e., $\Theta_{ic} = \left(\theta'_{0c}, \theta'_{ic}\right)' = \left(\theta^1_{0c}, \ldots, \theta^{m_0}_{0c}, \theta^1_{ic}, \ldots, \theta^{m_1}_{ic}\right)'$, where $\theta^{\,k}_{ic} \neq \theta^{\,k}_{jc}$ for i $\neq j$, $(i, j = 1, 2, k = 1, \ldots, m_1$, $c = 1, \ldots, r)$, and $m_0 + m_1 = m$.

Let $q_i^0 = \left(q_{i1}, \ldots, q_{1,r-1}\right)$ $(i = 1, 2)$. Denote α as $n = 2r + m_0 + 2m_1 - 2$-dimensional vector of parameters, i.e.

$$\alpha = \left(\left(q_1^0\right)', \left(q_2^0\right)', \theta'_{01}, \theta'_{11}, \theta'_{21}, \ldots, \theta'_{0r}, \theta'_{1r}, \theta'_{2r}\right)' = \left(\alpha^1, \ldots, \alpha^n\right). \tag{5}$$

Let $P \subset R^n$ be the set of all possible α, such that $\Theta_i \in K$ $(i = 1, 2)$. Then suppose that

$$d_c(x, \alpha) = \arg\max_{i=1,2} (l_{ic} p_{ic}(x, \Theta_{ic}), \tag{6}$$

and

$$R_A(\alpha) = \sum_{i=1}^{2} \pi_i \sum_{c=1}^{r} q_{ic} \int_{\mathbf{X}} C(i, d_c(x, \alpha)) p_{ic}(x) dx. \tag{7}$$

The so-called estimative approach to the choice of sample-based classification rule $d_s(x)$ is used. The unknown parameters q_1^0, q_2^0, $\{\theta_{0c}, \theta_{1c}, \theta_{2c}\}$ are replaced by appropriate estimates from the TS in the BCR, i.e. $d_{sc}(x) = d_c(x, \hat{\alpha})$, where

$$\hat{\alpha} = \left(\hat{q}_1^0, \hat{q}_2^0, \hat{\theta}'_{01}, \hat{\theta}'_{11}, \hat{\theta}'_{21}, \ldots, \hat{\theta}'_{0r}, \hat{\theta}'_{1r}, \hat{\theta}'_{2r}\right)'.$$

The actual risk for the rules $\{d_{cs}(x, \hat{\alpha})\}$ is the risk of classifying a randomly selected individual with feature U and is designated by

$$R_A(\widehat{\alpha}) = \sum_{i=1}^{2} \pi_i \sum_{c=1}^{r} q_{ic} \int_{\mathbf{X}} C\left(i, d_c\left(x, \hat{\alpha}\right)\right) p_{ic}(x) dx. \tag{8}$$

Definition 1. Risk regret (RR) for $\{d_c(x, \hat{\alpha})\}$ is the difference between actual risk $R_A(\widehat{\alpha})$ and Bayes risk R_B, and expected regret risk (ERR) is the expectation of RR, i. e.

$$\mathrm{ERR} = E_T\{R_A(\widehat{\alpha})\} - R_B, \tag{9}$$

where $E_T\{R_A(\widehat{\alpha})\}$ denotes the expectation with respect to TS distribution.

Unfortunately the exact distributions of RR usually are hard to obtain. In those cases, large sample approximations to and asymptotic expansions for the distributions and expectations of RR are required.

The purpose of this paper is to find general expansions of ERR when maximum likelihood estimates (MLE) of unknown parameters of the distributions of mixed feature variables are used. These are used to evaluate the performance of sample-based CR and to find the optimal training sample allocation.

This is an extension of the result of Dučinskas (1995), who presented the asymptotic expansion of expected error regret in the situation when parameter vectors of classified distributions a priori have all components different. T. C. Kao et al. (1991) had also presented the asymptotic distribution of AER and asymptotic expansion for the expectation of AER. However, only in the case of two normal populations with different means and common covariance was considered.

The general asymptotic distribution of RR and asymptotic expansion of ERR in case of several populations and continuous feature variables are derived in paper of Dučinskas (1997). The asymptotic expansion of EER for mixed features with normaly distributed continuous components was derived by Leung (1996).

2. Notations and the main result

Let ∇_a be the vector partial differential operator given by

$$\nabla_\alpha^T = \left(\frac{\partial}{\partial \alpha^1}, \ldots, \frac{\partial}{\partial \alpha^n}\right) \quad \text{and} \quad |\nabla_a|^2 = \sum_{i=1}^{n} \left(\frac{\partial}{\partial \alpha^i}\right)^2$$

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

Similarly, ∇_α^2 denote the matrix second order differential operator

$$\nabla_\alpha^2 = \left\| \frac{\partial^2}{\partial \alpha^i \partial \alpha^j} \right\|_{i,j=1,2}.$$

Let $G_c(x) = l_{1c}p_{1c}(x) - l_{2c}p_{2c}(x)$ and $\Gamma_c = \{x \in R^p : G_c(x) = 0\}$, and γ_c be the Lebesque measure on Γ_c. Assume that I_c^i denotes the $m \times m$ Fisher information matrix for Θ_{ic}, i.e.

$$I_c^i = E_{ic}\left\{\nabla_{\Theta_{ic}} \ln p_{ic}(x)\, \nabla'_{\Theta_{ic}} \ln p_{ic}(x)\right\}, \tag{10}$$

where $E_{ic}\{.\}$ represents the expectation based on distribution with density function $p_{ic}(x)$ $(i = 1, 2, c = 1, \ldots, r)$. It is obvious that matrix I_c^i can be expressed as a block matrix

$$I_c^i = \begin{pmatrix} I_{0c}^i & I_{i0c} \\ I_{0ic} & I_{ic} \end{pmatrix}, \tag{11}$$

where

$$I_{ic} = E_{ic}\left\{\nabla_{\theta_{ic}} \ln p_{ic}(x)\, \nabla'_{\theta_{ic}} \ln p_{ic}(x)\right\},$$

$$I_{0c}^i = E_{ic}\left\{\nabla_{\theta_{0c}} \ln p_{ic}(x)\, \nabla'_{\theta_{0c}} \ln p_{ic}(x)\right\}, \tag{12}$$

$$I_{i0c} = I'_{0ic} = E_{ic}\left\{\nabla_{\theta_{ic}} \ln p_{ic}(x)\, \nabla'_{\theta_{0c}} \ln p_{ic}(x)\right\} \quad (i = 1, 2, c = 1, \ldots, r). \tag{13}$$

Suppose that S is the regularity assumptions for $\{p_{ic}(x)\}$ and $\{q_{ic}\}$ ensuring the following properties of the MLE: MLE $\hat{\alpha}$ from T is consistent estimate and as $M_i \to \infty$, $M_i/M \to r_i > 0$ $(i = 1, 2)$ satisfies

$$\sqrt{M}\,(\hat{\alpha} - \alpha^*) \xrightarrow{L} N_n \quad (0, J_0^{-1}), \tag{14}$$

where

$$J_0 = \text{block diag}(r_1 I(q_1),\ r_2 I(q_2),\ J_1, \ldots, J_r),$$

with

$$I(q_i) = \begin{pmatrix} \frac{1}{q_{i1}} + \frac{1}{q_{ir}} & \frac{1}{q_{ir}} & \cdots & \frac{1}{q_{ir}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{1}{q_{ir}} & \frac{1}{q_{ir}} & \cdots & \frac{1}{q_{i,r-1}} + \frac{1}{q_{ir}} \end{pmatrix}, \quad (i = 1, 2)$$

$$J_c = \begin{pmatrix} \sum_{i=1}^{2} r_i q_{ic} I_{0c}^i & r_1 q_{1c} I_{01c} & r_2 q_{2c} I_{02c} \\ & r_1 q_{1c} I_{1c} & 0 \\ & & r_2 q_{2c} I_{2c} \end{pmatrix}, \quad (c = 1, \ldots, r).$$

Let the random variable $U_i = \left(X_i^c, Y_i\right)$ has the probability mass function $\prod_{c=1}^{r}(p_{ic}\left(x^i\right)q_{ic})^{Y_c^i}$ $(i = 1, 2)$ and

$$V_c = \sum_{i=1}^{2} l_{ic}(-1)^i \left(\nabla_{\theta_{0c}} p_{ic}(x) - I_{0i} I_i^{-1} \nabla_{\theta_{ic}} p_{ic}(x)\right) \tag{15}$$

THEOREM 1. *Let the regularity assumption S hold and let $R_A(\alpha)$ be twice continuously differentiable as function of α in some neighborhood U_{α^*} and let $F(U_1, U_2)$ be real valued function defined on $R^{2(p+r)}$ that satisfies*

$$M(R_A - R_B) < F(U_1, U_2), \tag{16}$$

where $E\{F(U_1, U_2)\} < H$, $0 < H < \infty$.

Then the first order asymptotic expansion of the EER is

$$\text{EER} = \beta/2M + \sum_{i=1}^{2} \rho_i/2M_i + \sum_{i=1}^{2} \eta_i/2M_i + o\left(M^{-1}\right) \tag{17}$$

where

$$\beta = \sum_{c=1}^{r} \int_{\Gamma_c} V_c' \Lambda V_c \, |\nabla_x G_c(x)|^{-1} \, d\gamma_c, \tag{18}$$

$$\rho_{ic} = \sum_{c=1}^{r} \int_{\Gamma_c} l_{ic}^2 \nabla_{\theta_{ic}}' p_{ic}(x) \, I_{ic}^{-1} \nabla_{\theta_{ic}} p_{ic}(x) \, |\nabla_X G_c(x)|^{-1} \, d\gamma_c, \tag{19}$$

$$\Lambda_c = \left(\sum_{i=1}^{2} r_i q_{ic} \left(I_{0c}^i - I_{0ic} I_{ic}^{-1} I_{i0c} \right) \right)^{-1}, \tag{20}$$

$$\eta_i = \sum_{c=1}^{r} l_{ic}^2 \int_{\Gamma_c} p_{ic}^2(x) \, (1 - q_{ic}) \, q_{ic} \, |\nabla_x G_c(x_0)|^{-1} \, d\gamma_c. \tag{21}$$

Proof. The assertion of the stated theorem directly follows from one of Theorem 5 of Dučinskas (1997), after inverting the J_0 and collection the terms at M^{-1} and $\left\{M_i^{-1}\right\}$ $(i = 1, 2)$.

Remark 1. Let $m_0 = 0$, i.e. all true values of components of unknown distribution parameters are distinct. Then $\beta = 0$ in the first order asymptotic expansion of ERR defined in (17).

Remark 2. If $m_0 \neq 0$, but θ_{0c}^* and $\{q_{ic}\}$ are known, then $\beta = \eta = 0$ in (17).

Then training sample allocation problem is viewed as follows. For a fixed value of M, let $W_i = M_i/M$ denote the proportions of observations taken from Ω_i. The design problem is to chose a value W_i^* for W_i that minimizes the AEER $= \sum_{i=1}^{2} \rho_i/2M_i$.

The W_1^* could be expressed explicitly (see Theorem 3 in Dučinskas (1995))

$$W_1^* = 1/\left(1 + \sqrt{\rho_2/\rho_1}\right). \tag{22}$$

REFERENCES

[1] Dučinskas K., Optimal training sample allocation and asymptotic expansions for error rates in discriminant analysis, *Acta Appl. Math.*, **38** (1995), 3–11.

[2] Dučinskas K., An asymptotic analysis of the regret risk in discriminant analysis under various training schemes, *Liet. Matem. Rink.*, **37**(4) (1997), 448–466.

[3] Kao T. C. and McCabe G. P., Optimal allocation for normal discrimination and logistic regression under stratified sampling, *J. Amer. Statist. Assoc.*, **88** (1991), 432–436.

[4] Krzanowski W., Discrimination and classification using both binary and continuous variables, *J. Amer. Statist. Assoc.*, **70** (1975), 782–790.

[5] Leung C. Y., The location linear discriminant for classifying observations with unequal variances, *Statistics and Probability Letters*, **31** (1996), 23–29.

Klasifikavimo rizika diskriminantinėje mišrių kintamųjų analizėje

K. Dučinskas

Nagrinėjamas atsitiktinių vektorių su tolydžiomis ir diskrečiomis komponetėmis diskriminavimo uždavinys. Pateikti laukiamos rizikos pirmos eilės asimptotiniai skleidiniai klasifikavimo taisyklei, naudojančiai parametrų maksimalaus tikėtinumo įverčius pagal stratifikuotas mokymo imtis.