LMD mokslo darbai, III tomas, 399–405

Classification of points in 2-dimensional space based on realizations of Gaussian random fields

Jūratė ŠALTYTĖ (KU), Kęstutis DUČINSKAS (KU)
e-mail: jsaltyte@gmf.ku.lt, duce@gmf.ku.lt

1. Introduction

Suppose Ω_1, Ω_2 are two mutually exclusive and exhaustive classes of objects. Let X be a p-dimensional feature vector, which is measured on each object. For objects randomly chosen from Ω_l, X follows the multivariate distribution with density function $p_l(x;\theta_l) = p_l(x)$, which belongs to the parametric family of regular densities $F_l = \{p_l(x;\theta_l), \theta_l \in \Theta_l \subset R^m\}$, $l = 1, 2$. Discriminant analysis deals with the problem of identifying the class of object for which X is measured. For a zero-one loss function, the Bayes classification rule (BCR) $d_B(x)$, minimizing the probability of misclassification, is equivalent to assigning $X = x$ to Ω_l if

$$\pi_l p_l(x) = \max_{k=1,2} \pi_k p_k(x),$$

where π_l is a prior probability of Ω_l.

Then, BCR $d_B(x)$ is

$$d_B(x) = \arg\max_{k=1,2} \pi_k p_k(x).$$

Let P_B denote the probability of misclassification for BCR $d_B(x)$ or the Bayes error rate (see, e.g., Hand (1997)).

In practical applications, the density functions $\{p_l(x)\}$ are seldom completely known. Often they are known only up to the parameters $\{\theta_l\}$, i.e. we may only assert that $p_l(x)$ is an element of a parametric family of density functions F_l. Under these conditions, it is customary to estimate θ_l from a training sample $T_l = \{X_{l1}, ..., X_{lN_l}\}$ drawn from Ω_l, for $l = 1, 2$. Put $T = T_1 \cup T_2$, $N = N_1 + N_2$.

Let $\widehat{\theta_l}$ be the maximum likelihood estimator (MLE) of θ_l based on T_l $(l = 1, 2)$.

An estimator of the rule $d_B(x)$ is called a *plug-in rule* $d_B\left(x, \widehat{\theta}_1, \widehat{\theta}_2\right)$ and is defined by

$$d_B\left(x, \widehat{\theta}_1, \widehat{\theta}_2\right) = \arg\max_{k=1,2} \pi_k p_k\left(x, \widehat{\theta}_k\right)$$

The actual error rate (P_A) of $d_B\left(x, \widehat{\theta}_1, \widehat{\theta}_2\right)$ is the probability of misclassifying a randomly and independently of T selected object with feature X and is designated by

$$P_A = \sum_{l=1}^{2} \pi_l \int \left(1 - \delta\left(l, d_B\left(x, \widehat{\theta}_1, \widehat{\theta}_2\right)\right)\right) p_l(x)\, dx,$$

where $\delta(\cdot, \cdot)$ is the Kronecker's delta.

DEFINITION. Expected error regret (EER) for $d_B\left(\cdot, \widehat{\theta}_1, \widehat{\theta}_2\right)$ is the expectation of the difference between P_A and P_B with respect to the distribution of $\widehat{\theta}_1$, $\widehat{\theta}_2$, i.e.,

$$EER = E(P_A) - P_B.$$

The purpose of this article is to find the asymptotic expansion for EER. The case of training sample of independent normaly distributed observations from one of two classes with $\Sigma_l = \Sigma$, $l = 1, 2$, was considered in (Okamoto (1963)). Dučinskas (1997) has been made the generalization for the case of arbitrary number of classes ($l \geqslant 2$) and regular class-conditional densities.

2. The main result

Suppose that any point $r = (r_1, r_2) \in D \subset R^2$ can be assigned to one of two classes Ω_1, Ω_2 prescribed above, with positive prior probabilities π_1, π_2, respectively. Here we identify the objects with points in D. The class of the point r is given by the random 2-dimensional vector $Y_r^T = (Y_{1r}, Y_{2r})$ of zero-one variables. The l*th* component of Y is defined to be one or zero according as an class of point r is or not Ω_l $(l = 1, 2)$. Then $Y_r \sim Mult_2(1; (\pi_1, \pi_2))$.

Suppose that X_r means the observation of X at point $r \in D$. A decision is to be made as to which class the randomly chosen point $r \in D$ is assigned on the basis of observed value of X_r.

Let

$$X_r = \sum_{l=1}^{2} Y_{lr}\mu_l + \epsilon_r, \tag{1}$$

where $\mu_1, \mu_2 \in R^p$, $\mu_1 \neq \mu_2$ and the noise $\epsilon_r = \left(\epsilon_r^1, \ldots, \epsilon_r^p\right)$ are the observation of the zero-mean second-order stationary correlated random field at location $r \in D$.

The essential assumption is that $\{\epsilon_r\}$ is Gaussian field with spatially factorized covariance. Hence, the common *class-conditional covariance* between any two observations X_r and X_s at points $r, s \in D$ belonging to Ω_l can be factorized as $\operatorname{cov}(X_r, X_s/r, s \in \Omega_l) = \rho^l(h)\Sigma$, $(r \neq s)$, where $\rho^l(\cdot)$ is the spatial correlation function $(l = 1, 2)$, and $h = r - s$, $\Sigma = \operatorname{cov}(\epsilon_r, \epsilon_r)$.

Also here we assume that the effect of *cross-correlation* between samples from different classes is *negligible*. In this paper we suppose, that it is equal to zero, i. e., $\operatorname{cov}(X_r, X_s/r \in \Omega_1, s \in \Omega_2) = 0$.

Let $D_l = \{s_1^l, ..., s_{N_l}^l\} \subset D$ be the set of points belonging to class Ω_l, $l = 1, 2$. Then X_{lj} means the observation of X at the point s_j^l, i.e. $X_{lj} = X(s_j^l)$, $j = 1, ..., N_l$, $l = 1, 2$.

Then, the expectation for $N_l p \times 1$ stacked vector $T_l^V = \left(X_{l1}', ..., X_{lN_l}'\right)'$ is

$$\mu_l^+ = \mathbf{1}_{N_l} \otimes \mu_l \quad (l = 1, 2), \tag{2}$$

where $\mathbf{1}_{N_l}$ is the N_l-dimensional vector of ones, and $\otimes$ is the Kronecker product. The covariance matrix of T_l^V is

$$\Sigma_l^+ = C_l \otimes \Sigma, \tag{3}$$

where C_l is the spatial correlation matrix of order $N_l \times N_l$, whose (i,j)*th* element is $\rho(s_i^l - s_j^l)$ $(i, j = 1, ..., N_l)$.

Suppose that Σ and C_l are known and μ_l are unknown $(l = 1, 2)$. In this paper maximum likelihood estimators (MLE) $\widehat{\mu}_l$ of μ_l based on T_l are used. Let $C_l^{-1} = \left(c_l^{ij}\right)$.

Lemma. *MLE of* $\{\mu_l\}$ $(l = 1, 2)$ *is*

$$\widehat{\mu}_l = \frac{1}{c_l^{\cdot\cdot}} \sum_{j=1}^{N_l} c_l^{\cdot j} x_{lj},$$

where $c_l^{\cdot j} = \sum_{i=1}^{N_l} c_l^{ij}$ *and* $c_l^{\cdot\cdot} = \sum_{i,j=1}^{N_l} c_l^{ij}$.

Proof. The log-likelihood of T_l is

$$\begin{aligned} \ln L_l = & -\text{const} - \frac{1}{2}\left(N_l \ln|\Sigma| + p \ln|C|\right) \\ & -\frac{1}{2}\left(c^{\cdot\cdot} tr\left(\Sigma^{-1} S_l\right) + c^{\cdot\cdot} tr\left(\Sigma^{-1}\left(\mu_l - \overline{x}_l\right)\left(\mu_l - \overline{x}_l\right)'\right)\right), \end{aligned}$$

where $\overline{x}_l = \frac{1}{c_l^{\cdot\cdot}} \sum_{j=1}^{N_l} c_l^{\cdot j} x_{lj}$ and $S_l = \frac{1}{c_l^{\cdot\cdot}} \sum_{i,j=1}^{N_l} c^{ij}\left(x_{lj} - \overline{x}_l\right)\left(x_{li} - \overline{x}_l\right)'$.

By solving equation $\frac{\partial \ln L_l}{\partial \mu_l} = 0$, we complete the proof of Lemma.

MLE under spatial sampling of Gaussian random fields was studied by Mardia and Marshall (1984). They gave the regularity conditions which ensure consistency and asymptotic normality of parameter estimators. We assume that these conditions hold.

Set $\gamma = \ln \frac{\pi_1}{\pi_2}$, $\Delta\widehat{\mu}_l = \widehat{\mu}_l - \mu_l$ $(l = 1, 2)$ and let $\Delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2)$ be the Mahalanobis distance. Let $\Phi(\cdot)$ and $\varphi(\cdot)$ denote standard normal distribution and density functions, respectively.

The plug-in discriminant function can be written in the form

$$d_B(x; \widehat{\mu}_1, \widehat{\mu}_2) = \left(x - \frac{1}{2}(\widehat{\mu}_1 + \widehat{\mu}_2)\right)' (\widehat{\mu}_1 - \widehat{\mu}_2). \tag{4}$$

Note that (see McLaclan (1974))

$$P_A = \pi_1 \Phi \left(- \frac{\left(\mu_1 - \frac{1}{2}\left(\widehat{\mu}_1 + \widehat{\mu}_2\right)\right)' \left(\widehat{\mu}_1 - \widehat{\mu}_2\right) + \gamma}{\sqrt{\left(\widehat{\mu}_1 - \widehat{\mu}_2\right)' \Sigma \left(\widehat{\mu}_1 - \widehat{\mu}_2\right)}} \right)$$

$$+ \pi_2 \Phi \left(\frac{\left(\mu_2 - \frac{1}{2}\left(\widehat{\mu}_1 + \widehat{\mu}_2\right)\right)' \left(\widehat{\mu}_1 - \widehat{\mu}_2\right) + \gamma}{\sqrt{\left(\widehat{\mu}_1 - \widehat{\mu}_2\right)' \Sigma \left(\widehat{\mu}_1 - \widehat{\mu}_2\right)}} \right).$$

For simplicity, assume $p = 1$.

Approximation of $E\left(P_A\right)$ proposed by Malinovskyi (1979) is

$$E_M\left(P_A\right) \approx \pi_1 \Phi \left(\frac{-\frac{1}{2}\Delta}{\sqrt{1 + \frac{1}{4}\sum_{l=1}^{2} D_l' C_l D_l - D_1' C_{01}}} \right)$$

$$+ \pi_2 \Phi \left(\frac{-\frac{1}{2}\Delta}{\sqrt{1 + \frac{1}{4}\sum_{l=1}^{2} D_l' C_l D_l - D_2' C_{02}}} \right),$$

where $D_l' = (d_{l1}, ..., d_{lN_l})$ with $d_{lj} = \frac{c_l^{jj}}{c_l^{jj}}$, $j = 1, ..., N_l$, and C_{0l} is $N_l \times 1$-dimensional vector of correlations $\rho\left(s_0^l - s_j^l\right)$; here s_0^l denotes the location of point to be classified.

Theorem. *If* $c_l^{jj} \to 0$ $(l = 1, 2)$, *then the first-order asymptotic expansion of EER for* $d_B\left(x, \widehat{\mu}_1, \widehat{\mu}_2\right)$, *using MLE* $\widehat{\mu}_1$, $\widehat{\mu}_2$, *is*

$$EER = \sum_{l=1}^{2} \frac{1}{4c_l^{jj}} \pi_l \varphi \left(-\frac{\gamma}{\Delta} + (-1)^l \frac{\Delta}{2} \right) \left(-\frac{\gamma}{\Delta} + (-1)^l \frac{\Delta}{2} \right)^2 / \Delta$$

$$+ o\left(\frac{1}{\min\left(c_1^{jj}, c_2^{jj}\right)} \right) \tag{5}$$

Proof. Since P_A is invariant under linear transformations of data we use the convenient canonical form of $\sigma = 1$ and $\mu_1 = -\mu_2 = \frac{1}{2}\Delta$ (see Dunn (1971)). Expand P_A in Taylor series about the point $\widehat{\mu}_l = \mu_l$ and then average with respect to the distribution of $\widehat{\mu}_l$ $(l = 1, 2)$. Expansion for $E\left(P_A\right)$ dropping the third order terms is as follows

$$E\left(P_A\right) \cong P_B + \sum_{l=1}^{2} P_l^{(1)} E\left(\Delta\widehat{\mu}_l\right) + \frac{1}{2} \sum_{l,k=1}^{2} tr\left(P_{l,k}^{(2)} E\left(\Delta\widehat{\mu}_l \Delta\widehat{\mu}_k\right) \right), \tag{6}$$

where $P_l^{(1)}$ is the vector of the first-order derivatives of P_A with respect to $\widehat{\mu}_l$ evaluated at μ_l $(l = 1, 2)$. Similarly, $P_{l,k}^{(2)}$ denotes the matrix of the second-order derivatives of P_A with respect

to $\widehat{\mu}_l$ and $\widehat{\mu}_k$ evaluated at μ_l and μ_k, respectively $(l, k = 1, 2)$. In considered situation there was obtained (see Ganesalingam S. and McLaclan G.J. (1978)) that

$$P_B = \pi_1 \Phi\left(-\frac{\Delta}{2} - \frac{\gamma}{\Delta}\right) + \pi_2 \Phi\left(-\frac{\Delta}{2} + \frac{\gamma}{\Delta}\right).$$

From Lemma and the assumptions stated above we have

$$E(\Delta\widehat{\mu}_l) = E(\Delta\widehat{\mu}_l \Delta\widehat{\mu}_k) = 0, \tag{7}$$

$$E\left((\Delta\widehat{\mu}_l)^2\right) = \frac{1}{c_l^{\cdot\cdot}}. \tag{8}$$

Then, using (7), (8) in (6) we complete the proof of the theorem.

COROLLARY. Whether T_l consists of statistically independent X_{lj}, $j = 1, ..., N_l$, then $c_l^{\cdot\cdot} = N_l$ in formula (5).

The corollary holds, since $C_l^{-1} = I$ for statistically independent X_{lj}, $j = 1, ..., N_l$.

The result of the proved theorem can be used in obtaining the optimal sampling design that ensures the minimum of asymptotic EER for fixed training sample size N.

3. Example

We consider an integer regular lattice and use the second-order neighborhood scheme for training sample. Suppose that there are 4 spatially symmetric observations in training sample for each class.

Two spatial correlation functions are considered:

1. $\rho_1^l = \exp\left(-\alpha\sqrt{t^2 h_1^2 + h_2^2}\right)$ – Exponential correlation function $(l = 1, 2)$;

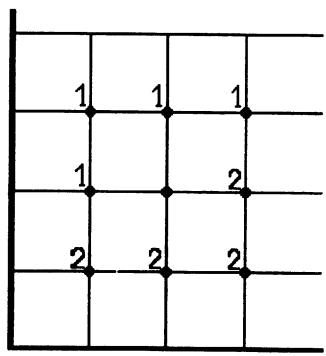

Fig. 1. Second-order neighborhood scheme with symmetric training samples.

Table 1

Exponential correlation function with $\alpha = 0.1$ and $t^2 = 0.5$

Δ	P_B	$AEER$	$AEER_M$	$\frac{AEER}{AEER_M}$	$\frac{INDEP}{AEER_M}$
0.25	0.4503	0.0058	0.0453	0.1291	0.0625
0.75	0.3538	0.0165	0.1329	0.1241	0.0603
1.25	0.2659	0.0242	0.2119	0.1145	0.0559
1.75	0.1908	0.0281	0.2782	0.1012	0.0499
2.25	0.1303	0.0282	0.3299	0.,0854	0.0428
2.75	0.0846	0.0252	0.3668	0.0687	0.0351

Table 2

Ornstein–Uhlenbeck correlation function with $\alpha = \beta = 0.1$

Δ	P_B	$AEER$	$AEER_M$	$\frac{AEER}{AEER_M}$	$\frac{INDEP}{AEER_M}$
0.25	0.4503	0.0056	0.0447	0.1261	0.0634
0.75	0.3538	0.0159	0.1309	0.1212	0.0611
1.25	0.2659	0.0233	0.2086	0.1118	0.0568
1.75	0.1908	0.0271	0.2737	0.0989	0.0508
2.25	0.1303	0.0271	0.3241	0.0837	0.0436
2.75	0.0846	0.0242	0.3598	0.0674	00357

2. $\rho_2^l = \exp\left(-\alpha h_1^2 - \beta h_2^2\right)$ – Ornstein-Uhlenbeck correlation function (see, e.g., Ying Z. (1993)) $(l = 1, 2)$.

Note that ρ_1^l and ρ_2^l are anisotropic correlation functions, when $t^2 \neq 1$ and $\alpha \neq \beta$, respectively. Let

$$AEER \triangleq \sum_{l=1}^{2} \frac{1}{4c_l^{\cdot\cdot}} \pi_l \varphi\left(-\frac{\gamma}{\Delta} + (-1)^l \frac{\Delta}{2}\right)\left(-\frac{\gamma}{\Delta} + (-1)^l \frac{\Delta}{2}\right)^2 / \Delta$$

and

$$AEER_M \triangleq E_M(P_A) - P_B.$$

In Tables 1 and 2 values of $AEER$ and $AEER_M$ with $\pi_1 = \pi_2 = 0.5$ for considered two correlation functions are presented. Also ratios $\frac{AEER_M}{AEER}$ and $\frac{INDEP}{AEER_M}$ are calculated (the ratio $\frac{INDEP}{AEER} = 0.2643$ (for ρ_1^l) and $\frac{INDEP}{AEER} = 0.2747$ (for ρ_2^l) for all Δ), here $INDEP$ means an approximation of the EER in the case of independent observations in neighbouring locations.

Figures in Tables 1 and 2 allow us to conclude that approximation of EER based on asymptotic expansion presented in this paper has smaller values than the approximation proposed by Malinovskyi for all considered cases.

References

[1] D.J. Hand, *Construction and Assessment of Classification Rules*, John Wiley & Sons, New York (1997).
[2] M. Okamoto, An asymptotic expansion for the distribution of the linear discriminant function, *Ann. Math. Statist.*, **34**, 1286–1301 (1963).
[3] K. Dučinskas, An asymptotic analysis of the regret risk in discriminant analysis under various training schemes, *Lith. Math. J.*, **37**(4), 337–351 (1997).
[4] K.V. Mardia, and R.J. Marshall, Maximum likelihood estimation of models for residual covariance and spatial regression, *Biometrika*, **71**, 135–146 (1984).
[5] G.J. McLaclan, The asymptotic distributions of the conditional error rate and risk in discriminant analysis, *Biometrika*, **61**(1), 131–135 (1974).
[6] L.G. Malinovskyi, *Classification of Objects by Methods of Discriminant Analysis*, 162–163 (1979).
[7] O.J. Dunn, Some expected values for probabilities of correct classification in discriminant analysis, *Technometrics*, **13**, 345–353 (1971).
[8] S. Ganesalingam, and G.J. McLaclan, The efficiency of a linear discriminant function based on unclassified initial samples, *Biometrika*, **65**(3), 658–662 (1978).
[9] Z. Ying, Maximum likelihood estimation of parameters under a spatial sampling scheme, *The Annals of Statistics*, Vol. 21, No. 3, 1567–1590 (1993).

Plokštumos taškų klasifikavimas pagal Gauso laukų realizacijas

J. Šaltytė, K. Dučinskas

Straipsnyje nagrinėjamas plokštumos taškų klasifikavimo uždavinys pagal Gauso laukų realizacijas. Gautas pirmos eilės asimptotinis klasifikavimo rizikos padidėjimo skleidinys atvejui, kai į Bajeso klasifikavimo taisyklę įstatome maksimalaus tikėtinumo įverčius. Atliktas skaitinis gauto skleidinio palyginimas su Malinovskio aproksimacija situacijai, kai stebime ir klasifikuojame tik taškus, esančius stačiakampėje gardelėje.