STATISTICAL CLASSIFICATION UNDER DIRECT CHOICE OF INFORMATIVE FEATURES AND ITS RISK

Serikova E.V.

Department of Mathematical Modeling and Data Analysis, Belarusian State University, 4 Fr.Skariny av., 220050 Minsk, Belarus

Abstract. The problem of statistical classification of multivariate normal (Gaussian) observations in the subspace of informative features is studied. The iterative stepwise method for non-informative feature rejection is proposed and efficiency of transition to the selected features is analytically investigated

Mathematical model

Let random observation x (N-vector) from $L \geq 2$ classes $\{\Omega_1,...,\Omega_L\}$ be registered in the feature space R^N. Introduce the notation: $d^0 \in S = \{1,...,L\}$ is an unknown random index of the class, to which the observation x belongs:

$$P\{d^0 = i\} = \pi_i > 0,\ i \in S \qquad (\pi_1 + ... + \pi_L = 1), \tag{1}$$

where $\{\pi_i\}_{i \in S}$ are prior class probabilities [1-5]. Under fixed $d^0 = i$ ($i \in S$) the observation $x \in R^N$ is described by the conditional probability density function:

$$p_i(x) \geq 0,\ x \in R^N: \quad \int_{R^N} p_i(x)dx = 1,\ i \in S.$$

Classes $\{\Omega_i\}_{i \in S}$ are completely determined by the introduced characteristics $\{\pi_i, p_i(\cdot)\}_{i \in S}$. To classify a random observation $x \in R^N$ the well-known Bayesian decision rule (BDR) [1,3], which minimizes the risk (the classification error probability), is used.

However, often in practice the initial feature space is redundant. It means that its dimension N is too large [1-3], and the noninformative feature subset must be rejected from the initial space of N features [2].

In this paper the well-known Fisher model [1, 3] of multivariate normal (Gaussian) distribution mixture is investigated:

$$p_i(x) = n_N(x \mid \mu_i, \Sigma), x \in R^N, \quad i \in S, \tag{2}$$

where

$$n_N(x \mid \mu_i, \Sigma) = (2\pi)^{-N/2}(\det(\Sigma))^{-1/2} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma^{-1}(x-\mu_i)\right)$$

is N-variate Gaussian probability density function with mathematical mean vector $\mu_i = E\{x \mid d^0 = i\} \in R^N$ (so called "centre" [1,4] of the class Ω_i) and non-singular covariance ($N \times N$)-matrix $\Sigma = E\{(x-\mu_i)(x-\mu_i)^T \mid d^0 = i\}$ ($\det(\Sigma) \neq 0$), common for all classes (here "T" is the transposition symbol).

Stepwise method for feature rejection

Now we consider the stepwise transition from R^N to the subspace of informative features. On each step one noninformative feature is rejected from the initial space. With-

out loss of generality we assume that rejected feature is the last one and the observation $x \in R^N$ takes the form:

$$x = (x_1, x_2, ..., x_{N-1}, x_N)^T = ((x^{N-1})^T | x_N)^T \in R^N,$$

and

$$\mu_i = (\mu_{i,1}, \mu_{i,2}, ..., \mu_{i,N-1}, \mu_{i,N})^T = ((\mu_i^{N-1})^T | \mu_{i,N})^T \in R^N,$$

$$\Sigma = \begin{pmatrix} \Sigma_{N-1,N-1} & \sigma_{N-1} \\ \sigma_{N-1}^T & \sigma_{N,N} \end{pmatrix} \begin{matrix} N-1 \\ 1 \end{matrix}$$

are the following expansions for the mathematical mean vector $\mu_i \in R^N$ and the covariance matrix Σ.

Let $\rho_N = \sqrt{\sigma_{N-1}^T \Sigma_{N-1,N-1}^{-1} \sigma_{N-1} / \sigma_{N,N}}$ denote the multiple correlation coefficient [2]. We say that feature is noninformative if ρ_N is close to unity. To estimate the efficiency of transition from the initial feature space let the risk value [1, 3, 4] be used.

Recall [1,3] that under the conditions of Fisher model (1), (2) the BDR in the initial feature space is represented in the form ($x \in R^N$):

$$d_0^N(x) = \arg\max_{i \in S} \{2 \ln \pi_i - (x - \mu_i)^T \Sigma^{-1} (x - \mu_i)\},$$

and the following risk value is determined by the expression:

$$r_0^N = P\{d_0^N(x) \neq d^0\} = \sum_{i \in S} \pi_i P\{d_0^N(x) \neq i | d^0 = i\} = \tag{3}$$

$$= \sum_{i \in S} \pi_i \sum_{\substack{j \in S \\ j \neq i}} \int_{R^N} \prod_{\substack{k \in S \\ k \neq j}} I\left((x - \mu_j)^T \Sigma^{-1} (\mu_j - \mu_k) + \frac{{}_N\Delta_{jk}^2}{2} - \ln \frac{\pi_k}{\pi_j} \right) n_N(x | \mu_i, \Sigma) dx,$$

where $I(z) = \{1, \text{если } z \geq 0; 0, \text{если } z < 0\}$ is the unit function;

$${}_N\Delta_{jk} = \sqrt{(\mu_j - \mu_k)^T \Sigma^{-1} (\mu_j - \mu_k)} \tag{4}$$

is the Mahalanobis distance [1,3] between classes Ω_j, Ω_k in the initial space ($k \neq j \in S$).

It is clear that selected features ($x^{N-1} \in R^{N-1}$) are also determined by the Fisher model, but with parameters $\{\mu_i^{N-1}\}_{i \in S}$ and $\Sigma_{N-1,N-1}$. Hence in the selected feature space the BDR has the form:

$$d_0^{N-1}(x^{N-1}) = \arg\max_{i \in S} \{2 \ln \pi_i - (x^{N-1} - \mu_i^{N-1})^T \Sigma_{N-1,N-1}^{-1} (x^{N-1} - \mu_i^{N-1})\},$$

and the risk $r_0^{N-1} = P\{d_0^{N-1}(x^{N-1}) \neq d^0\}$:

$$r_0^{N-1} = \sum_{i \in S} \pi_i \sum_{\substack{j \in S \\ j \neq i}} \int_{R^{N-1}} \prod_{\substack{k \in S \\ k \neq j}} I\left((x^{N-1} - \mu_j^{N-1})^T \Sigma_{N-1,N-1}^{-1} (\mu_j^{N-1} - \mu_k^{N-1}) + \right. \tag{5}$$

$$\left. + \frac{{}_{N-1}\Delta_{jk}^2}{2} - \ln \frac{\pi_k}{\pi_j} \right) n_{N-1}(x^{N-1} | \mu_i^{N-1}, \Sigma_{N-1,N-1}) dx^{N-1},$$

where

$$ {}_{N-1}\Delta_{jk} = \sqrt{(\mu_j^{N-1} - \mu_k^{N-1})^T \Sigma_{N-1,N-1}^{-1} (\mu_j^{N-1} - \mu_k^{N-1})} \qquad (6) $$

is the Mahalanobis interclass distance in the selected feature space ($k \neq j \in S$).

In the case of two classes ($L = 2$) expressions (3), (5) are simplified [1, 3]:

$$ r_0^N = \pi_1 \Phi\left(-\frac{\Delta_N}{2} - \frac{\ln h}{\Delta_N} \right) + \pi_2 \Phi\left(-\frac{\Delta_N}{2} + \frac{\ln h}{\Delta_N} \right), \quad \Delta_N = {}_N\Delta_{12}, \; h = \frac{\pi_1}{\pi_2}; \qquad (7) $$

$$ r_0^{N-1} = \pi_1 \Phi\left(-\frac{\Delta_{N-1}}{2} - \frac{\ln h}{\Delta_{N-1}} \right) + \pi_2 \Phi\left(-\frac{\Delta_{N-1}}{2} + \frac{\ln h}{\Delta_{N-1}} \right), \; \Delta_{N-1} = {}_{N-1}\Delta_{12}, \qquad (8) $$

where $\Phi(z) = \int_{-\infty}^{z} \varphi(\omega) d\omega$ is the standard Gaussian distribution function with probability density function $\varphi(\omega) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\omega^2}{2} \right)$.

Efficiency of transition to the informative features

Let the relations for the risk (3), (5) (in the case of two classes (7), (8)) be used to investigate the efficiency of transition from initial space R^N to the selected feature space R^{N-1}.

First, let us formulate the following helpful lemmas.

Lemma 1. Under the conditions of Fisher model the following inequalities for the risk (3), (5) are true:

$$ r_0^N \leq \sum_{i \in S} \pi_i \sum_{\substack{j \in S \\ j \neq i}} \Phi\left(-\frac{{}_N\Delta_{ij}}{2} - \frac{\ln \pi_i / \pi_j}{{}_N\Delta_{ij}} \right), \quad r_0^{N-1} \leq \sum_{i \in S} \pi_i \sum_{\substack{j \in S \\ j \neq i}} \Phi\left(-\frac{{}_{N-1}\Delta_{ij}}{2} - \frac{\ln \pi_i / \pi_j}{{}_{N-1}\Delta_{ij}} \right). \qquad (9) $$

Corollary. In the case of two classes ($L = 2$) the exact equalities take place and expressions (9) coincide with (7), (8).

Lemma 2. If ${}_{N-1}\Delta_{ij}$ is the Mahalanobis interclass distance in the selected feature space R^{N-1} from (6) and ${}_N\Delta_{ij}$ is the Mahalanobis interclass distance in the initial feature space R^N from (4), then the following representation takes place ($i \neq j \in S$):

$$ {}_{N-1}\Delta_{ij}^2 = {}_N\Delta_{ij}^2 - (1 - \rho_N^2)^{-1} \sigma_{N,N}^{-1} \left((\mu_{i,N} - \mu_{j,N}) - \sigma_{N-1}^T \Sigma_{N-1,N-1}^{-1} (\mu_i^{N-1} - \mu_j^{N-1}) \right)^2. \qquad (10) $$

Theorem 1. Let Δr_0 be a risk bias: $\Delta r_0 = r_0^{N-1} - r_0^N \geq 0$, where r_0^N, r_0^{N-1} are the risk values from (3), (5). If under the conditions of Fisher model: $\delta_{ij} = \left| (\mu_{i,N} - \mu_{j,N}) - \sigma_{N-1}^T \Sigma_{N-1,N-1}^{-1} (\mu_i^{N-1} - \mu_j^{N-1}) \right| \to 0$, $i \neq j \in S$, then for $L \geq 2$ classes the following inequality is true:

$$ \Delta r_0 \leq \frac{1}{2} (1 - \rho_N^2)^{-1} \sigma_{N,N}^{-1} \sum_{i \in S} \pi_i \sum_{\substack{j \in S \\ j > i}} \varphi\left(-\frac{{}_N\Delta_{ij}}{2} - \frac{\ln \pi_i / \pi_j}{{}_N\Delta_{ij}} \right) \frac{1}{{}_N\Delta_{ij}} \delta_{ij}^2 + O\left((\max_{i,j \in S} \delta_{ij})^4 \right), \qquad (11) $$

and for the case of two classes ($L = 2$):

$$ r_0^{N-1} = r_0^N + \frac{1}{2} (1 - \rho_N^2)^{-1} \sigma_{N,N}^{-1} \frac{\pi_1}{\Delta_N} \varphi\left(-\frac{\Delta_N}{2} - \frac{\ln h}{\Delta_N} \right) \delta_{12}^2 + O(\delta_{12}^4). \qquad (12) $$

We suppose that the rejected feature has a large correlation with other features. This means that the condition of Theorem 1 is well-defined.

Note that the influence of the features rejected on previous steps is not taken into account.

Let R^m be an informative feature subspace obtained on the $(N-m)$-th step. Without loss of generality we assume that the rejected $N-m$ features are the last ones and the observation $x \in R^N$ takes the form:

$$x = ((x^m)^T \,|\, (x^{N-m})^T)^T \in R^N, \quad x^m \in R^m, \quad x^{N-m} \in R^{N-m}.$$

Therefore

$$\mu_i = \left((\mu_i^m)^T \,|\, (\mu_i^{N-m})^T\right)^T \in R^N, \ \mu_i^m \in R^m, \ \mu_i^{N-m} \in R^{N-m}, \ i \in S,$$

$$\Sigma = \left(\begin{array}{c|c} \overset{m}{\Sigma_{m,m}} & \overset{N-m}{\Sigma_{m,N-m}} \\ \hline \Sigma_{m,N-m}^T & \Sigma_{N-m,N-m} \end{array}\right)\begin{array}{l} m \\ N-m \end{array}$$

are the following representations for the mathematical mean vector $\mu_i \in R^N$ and the covariance matrix Σ.

It is easily shown that in the selected feature space the BDR has the form ($x^m \in R^m$):

$$d_0^m(x^m) = \arg\max_{i \in S}\{2\ln\pi_i - (x^m - \mu_i^m)^T \Sigma_{m,m}^{-1}(x^m - \mu_i^m)\},$$

and the risk $r_0^m = P\{d_0^m(x^m) \neq d^0\}$:

$$r_0^m = \sum_{i \in S}\pi_i \sum_{\substack{j \in S \\ j \neq i}} \int_{R^m} \prod_{\substack{k \in S \\ k \neq j}} I\left((x^m - \mu_j^m)^T \Sigma_{m,m}^{-1}(\mu_j^m - \mu_k^m) + \frac{{}_m\Delta_{jk}^2}{2} - \ln\frac{\pi_k}{\pi_j}\right) n_m(x^m \,|\, \mu_i^m, \Sigma_{m,m})dx^m, \quad (13)$$

where

$${}_m\Delta_{jk} = \sqrt{(\mu_j^m - \mu_k^m)^T \Sigma_{m,m}^{-1}(\mu_j^m - \mu_k^m)} \quad (14)$$

is the Mahalanobis interclass distance in the space of m selected features ($k \neq j \in S$).

In the case of two classes ($L = 2$) expression (13) takes the form:

$$r_0^m = \pi_1 \Phi\left(-\frac{\Delta_m}{2} - \frac{\ln h}{\Delta_m}\right) + \pi_2 \Phi\left(-\frac{\Delta_m}{2} + \frac{\ln h}{\Delta_m}\right), \quad \Delta_m = {}_m\Delta_{12}, \ h = \frac{\pi_1}{\pi_2}. \quad (15)$$

Lemma 3. If ${}_m\Delta_{ij}$ is the Mahalanobis interclass distance in the selected feature space R^m from (14) and ${}_N\Delta_{ij}$ is the Mahalanobis interclass distance in the initial feature space R^N from (4), then the following representation takes place ($i \neq j \in S$):

$${}_m\Delta_{ij}^2 = {}_N\Delta_{ij}^2 - \left((\mu_i^{N-m} - \mu_j^{N-m}) - B(\mu_i^m - \mu_j^m)\right)^T \Sigma_{N-m,N-m|m}^{-1}\left((\mu_i^{N-m} - \mu_j^{N-m}) - B(\mu_i^m - \mu_j^m)\right), \quad (16)$$

where $\Sigma_{N-m,N-m|m} = \Sigma_{N-m,N-m} - \Sigma_{m,N-m}^T \Sigma_{m,m}^{-1} \Sigma_{m,N-m}$ is the conditional covariance matrix for random $x^{N-m} \in R^{N-m}$ under the fixed $x^m \in R^m$, and $B = \Sigma_{m,N-m}^T \Sigma_{m,m}^{-1}$ is the matrix of regression coefficients [2].

The relations for the risk (3), (13) (in the case of two classes (7), (15)) be used to investigate efficiency of transition from initial space R^N to the informative feature subspace R^m.

Theorem 2. Let $\overline{\Delta}r_0$ be a risk bias: $\overline{\Delta}r_0 = r_0^m - r_0^N \geq 0$, where r_0^N, r_0^m are the risk values from (3), (13). If under the conditions of Fisher model: $\overline{\delta}_{ij} = \left|\delta_{ij}^*\right| \to 0$, $i \neq j \in S$, where $\delta_{ij}^* = \left(\mu_i^{N-m} - \mu_j^{N-m}\right) - B\left(\mu_i^m - \mu_j^m\right) \in R^{N-m}$, $\left|\delta_{ij}^*\right| = \sqrt{\left(\delta_{ij}^*\right)^T \delta_{ij}^*}$ is the Euclidean norm, then for $L \geq 2$ classes the following inequality is true:

$$\overline{\Delta}r_0 \leq \frac{1}{2}\sum_{i\in S}\pi_i \sum_{\substack{j\in S \\ j>i}} \frac{1}{{}_N\Delta_{ij}}\varphi\left(-\frac{{}_N\Delta_{ij}}{2} - \frac{\ln \pi_i/\pi_j}{{}_N\Delta_{ij}}\right)\times \tag{17}$$

$$\times\left(\delta_{ij}^*\right)^T\left(\Sigma_{N-m,N-m}^{-1/2}\right)^T\left(I_{N-m} - \overline{\rho}_{N-m}^2\right)^{-1}\Sigma_{N-m,N-m}^{-1/2}\delta_{ij}^* + O\left((\max_{i,j\in S}\overline{\delta}_{ij})^4\right),$$

and for the case of two classes ($L = 2$, $\Delta_N = {}_N\Delta_{12}$, $h = \frac{\pi_1}{\pi_2}$):

$$\overline{\Delta}r_0 = \frac{1}{2}\frac{\pi_1}{\Delta_N}\varphi\left(-\frac{\Delta_N}{2} - \frac{\ln h}{\Delta_N}\right)\left(\delta_{12}^*\right)^T\left(\Sigma_{N-m,N-m}^{-1/2}\right)^T\left(I_{N-m} - \overline{\rho}_{N-m}^2\right)^{-1}\Sigma_{N-m,N-m}^{-1/2}\delta_{12}^* + O\left(\overline{\delta}_{12}^4\right), \tag{18}$$

where the $(N-m)\times(N-m)$-matrix $\overline{\rho}_{N-m}^2 = \left(\Sigma_{N-m,N-m}^{-1/2}\right)^T\Sigma_{m,N-m}^T\Sigma_{m,m}^{-1}\Sigma_{m,N-m}\Sigma_{N-m,N-m}^{-1/2}$ is the analog of the multiple correlation coefficient $\rho_N \in R^1$.

Note, in practice results (17), (18) allow to reject features such that the general increment of the risk is less then any predetermined value. On the other hand, the local increment of the risk value on the fixed step for one feature is controlled by expressions (11), (12).

References

[1] S.A. Aivazyan, V.M. Buchstaber, I.S. Yenyukov, L.D. Meshalkin, "Applied statistics: Classification and Dimensionality Reduction", *Finansy i Statistika, Moskow*, (1989).

[2] Y.W. Anderson, "An Introduction to Multivariate Statistical Analysis", *Viley, New York*, (1963).

[3] K. Fukunaga, "Introduction to statistical pattern recognition. second edition", *Academic Press, New York*, (1990).

[4] Yu.S. Kharin, E.E Zhuk, "Stability in Claster Analysis of Multivariate observations", *BSU, Minsk*, (1998).