CLASSIFICATION BASED ON DISTANCE IN MULTIVARIATE GAUSSIAN CASES

KAMEO MATUSITA
THE INSTITUTE OF STATISTICAL MATHEMATICS, TOKYO

1. Introduction

The author previously treated the problem of classification in discrete cases, employing the notion of distance [1]. The purpose of this paper is to treat that problem for multivariate Gaussian cases from the same point of view.

Now, the classification problem is formulated as follows. Let $\{\omega_\nu\}$ be a class of sets of distributions, and let X be a random variable under consideration. Then the problem is to decide which ω_ν is considered to contain the distribution of X. We, of course, assume here that ω_ν and ω_μ have no common distributions when $\nu \neq \mu$. Further, for efficient decision making we assume that for a suitable distance $d(\cdot, \cdot)$ in the space of distributions concerned, we have $d(\omega_\nu, \omega_\mu) > \alpha$ (> 0), $(\nu \neq \mu)$. In some cases, when $d(\omega_\nu, \omega_\mu) = 0$, we can represent each of those ω_ν by a single distribution F_ν so that $d(F_\nu, F_\mu) > 0$. For such F_ν, we can consider the averaged distribution of ω_ν by an adequate distribution over ω_ν.

When the distributions concerned are all known, the decision rule for the above problem runs as follows. Let S_n be an 'empirical' distribution based on n observations on X. We compare the magnitudes of $d(S_n, \omega_\nu)$, and take the set which minimizes $d(S_n, \omega_\nu)$ as the set which contains the distribution. Then the problem is to evaluate the success rate or error rate of this procedure. In this paper, however, we shall treat the case where the distributions concerned are unknown. When the distributions concerned are unknown, we have to estimate them from observations. For that, the number of distributions concerned is required to be finite. Therefore, we assume that each ω_ν consists of a single distribution F_ν and the number of F_ν is finite.

In the present paper, we do not explicitly take into account a priori probabilities and costs of misclassification. However, our procedure will also apply with a slight modification to the case where they need to be considered.

2. Decision rule based on distance

Let X be the random variable under consideration, and S_n an 'empirical' distribution based on n observations on X. Suppose that X has one of $F_1, \cdots, F_l$ as its distribution. Let S_{ν,n_ν} denote the 'empirical' distribution based on a sample of n_ν from F_ν which has the same form as S_n. Then we consider $d(S_n, S_{\nu,n_\nu})$ and take F_{ν_0} when $S_{\nu_0,n_{\nu_0}}$ minimizes $d(S_n, S_{\nu,n_\nu})$.

Since the case of a finite number of distributions can reduce to the case where the number of distributions concerned is 2, we shall confine our consideration to this case.

Let F_1, F_2 be the distributions concerned, and let S'_r, S''_s be the 'empirical' distributions determined by observations on F_1 and F_2, respectively. Then the decision rule for this case is the following:

(i) when $d(S_n, S'_r) < d(S_n, S''_s)$, we decide on F_1;

and

(ii) when $d(S_n, S'_r) > d(S_n, S''_s)$, we decide on F_2.

(iii) For the case $d(S_n, S'_r) = d(S_n, S''_s)$, we determine in advance to take either of F_1, F_2, say F_1.

The success rate is given by

$$P(d(S_n, S'_r) \leq d(S_n, S''_s)|F_1), \tag{1}$$

where "$|F$" in the parentheses means "under the condition that X has F," and

$$P(d(S_n, S'_r) > d(S_n, S''_s)|F_2). \tag{2}$$

Now, when $d(\cdot, \cdot)$ satisfies the triangle axiom, we obtain

$$d(S_n, S'_r) \leq d(S_n, F_1) + d(S'_r, F_1), \tag{3}$$

$$d(S_n, S''_s) \geq d(F_1, F_2) - d(F_1, S_n) - d(F_2, S''_s), \tag{4}$$

and

$$d(S_n, S''_s) - d(S_n, S'_r) \geq d(F_1, F_2) - 2d(S_n, F_1) - d(F_1, S'_r) - d(F_2, S''_s). \tag{5}$$

Therefore, when $d(F_1, F_2) \geq \delta$ (> 0), we have

$$d(S_n, S''_s) - d(S_n, S'_r) \geq \delta - 2d(S_n, F_1) - d(F_1, S'_r) - d(F_2, S''_s), \tag{6}$$

and further, when $2d(S_n, F_1) + d(F_1, S'_r) + d(F_2, S''_s) \leq \delta$, we have $d(S_n, S'_r) \leq d(S_n, S''_s)$. As a result we obtain

$$\begin{aligned} &P(d(S_n, S'_r) \leq d(S_n, S''_s)|F_1) \\ &\geq P(2d(S_n, F_1) + d(F_1, S'_r) + d(F_2, S''_s) < \delta|F_1) \\ &\geq P\left(d(S_n, F_1) < \frac{\delta}{4}, d(F_1, S'_r) < \frac{\delta}{4}, d(F_2, S''_s) < \frac{\delta}{4} \,\middle|\, F_1\right) \\ &= P\left(d(S_n, F_1) < \frac{\delta}{4} \,\middle|\, F_1\right) \cdot P\left(d(F_1, S'_r) < \frac{\delta}{4}\right) \cdot P\left(d(F_2, S''_s) < \frac{\delta}{4}\right). \end{aligned} \tag{7}$$

Thus, for evaluation of the success rate it is sufficient to know about

$$P\left(d(S_n, F) < \frac{\delta}{4} \,\middle|\, F\right). \tag{8}$$

3. Distance and 'test' statistics

Let F_1, F_2 be distributions defined in space R, and let $p_1(x)$, $p_2(x)$ be their density functions with respect to a measure m in R. Then the distance between distributions which we employ here is

$$d(F_1, F_2) = \left[\int_R (\sqrt{p_1(x)} - \sqrt{p_2(x)})^2 \, dm\right]^{1/2}. \tag{9}$$

This distance satisfies the metric space axioms. When we define

$$\rho(F_1, F_2) = \int_R \sqrt{p_1(x)} \sqrt{p_2(x)} \, dm, \tag{10}$$

we have

$$d^2(F_1, F_2) = 2(1 - \rho(F_1, F_2)). \tag{11}$$

The quantity $\rho(\cdot, \cdot)$ expresses the closeness between distributions, and we can use $\rho(\cdot, \cdot)$ in place of $d(\cdot, \cdot)$.

Now, let us turn to the multivariate Gaussian case.

Let R be a k-dimensional space, and let F_1, F_2 be k-dimensional Gaussian distributions with density functions

$$p_1(x) = \frac{|A|^{1/2}}{(2\pi)^{k/2}} \exp\left[-\tfrac{1}{2}(A(x - a), (x - a))\right], \tag{12}$$

$$p_2(x) = \frac{|B|^{1/2}}{(2\pi)^{k/2}} \exp\left[-\tfrac{1}{2}(B(x - b), (x - b))\right], \tag{13}$$

where A, B are positive-definite matrices of degree k, and x, a, b are k-dimensional (column) vectors. Then we obtain

(14)

$$\rho(F_1, F_2) = \frac{|AB|^{1/4}}{|\frac{1}{2}(A + B)|^{1/2}} \exp\left[-\tfrac{1}{4}\{-((A + B)^{-1}(Aa + Bb), (Aa + Bb)) + (Aa, a) + (Bb, b)\}\right]$$

(see [2]). When $A = B$,

$$\rho(F_1, F_2) = \exp\left[-\tfrac{1}{8}(A(a - b), (a - b))\right]. \tag{15}$$

When $a = b$,

$$\rho(F_1, F_2) = \frac{|AB|^{1/4}}{|\frac{1}{2}(A + B)|^{1/2}}. \tag{16}$$

Let $X_1, X_2, \cdots, X_n$ be n (≥ 2) observations on a random variable X with a k-dimensional Gaussian distribution. Define

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i, \tag{17}$$

$$V = \frac{1}{n - 1} \sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})', \tag{18}$$

and let S_n be the k-dimensional Gaussian distribution with mean $\overline{X}$ and covariance matrix V, that is, $S_n = N(\overline{X}, V)$. Set $U = V^{-1}$. Similarly, concerning F_1 and F_2, let

$$S_r' = N(\overline{X}_{(1)}, V_{(1)}), \qquad U_{(1)} = V_{(1)}^{-1}, \tag{19}$$

$$S_s'' = N(\overline{X}_{(2)}, V_{(2)}), \qquad U_{(2)} = V_{(2)}^{-1}. \tag{20}$$

Then we have

$$\rho(S_n, S_r') = \frac{|UU_{(1)}|^{1/4}}{|\frac{1}{2}(U + U_{(1)})|^{1/2}} \exp\left[-\tfrac{1}{4}\{-((U + U_{(1)})^{-1}(U\overline{X} + U_{(1)}\overline{X}_{(1)}), (U\overline{X} + U_{(1)}\overline{X}_{(1)})) + (U\overline{X}, \overline{X}) + (U_{(1)}\overline{X}_{(1)}, \overline{X}_{(1)})\}\right], \tag{21}$$

$$\rho(S_n, S_s'') = \frac{|UU_{(1)}|^{1/4}}{|\frac{1}{2}(U + U_{(2)})|^{1/2}} \exp\left[-\tfrac{1}{4}\{-((U + U_{(2)})^{-1}(U\overline{X} + U_{(2)}\overline{X}_{(2)}), (U\overline{X} + U_{(2)}\overline{X}_{(2)})) + (U\overline{X}, \overline{X}) + (U_{(2)}X_{(2)}, \overline{X}_{(2)})\}\right]. \tag{22}$$

Using these statistics we can make a decision; that is, when $\rho(S_n, S_r') \geq \rho(S_n, S_s'')$, we decide that X has F_1, and when $\rho(S_n, S_r') < \rho(S_n, S_s'')$, we decide that X has F_2. When it is known in advance that $A = B$, we consider

$$\rho_1(S_n, S_r') = \exp\left[-\tfrac{1}{4}(U(\overline{X} - \overline{X}_{(1)}), (\overline{X} - \overline{X}_{(1)}))\right], \tag{23}$$

$$\rho_1(S_n, S_s'') = \exp\left[-\tfrac{1}{4}(U(\overline{X} - \overline{X}_{(2)}), (\overline{X} - \overline{X}_{(2)}))\right] \tag{24}$$

for the case where A $(= B)$ is unknown, and

$$\rho_2(S_n, S_r') = \exp\left[-\tfrac{1}{4}(A(\overline{X} - \overline{X}_{(1)}), (\overline{X} - \overline{X}_{(1)}))\right], \tag{25}$$

$$\rho_2(S_n, S_s'') = \exp\left[-\tfrac{1}{4}(A(\overline{X} - \overline{X}_{(2)}), (\overline{X} - \overline{X}_{(2)}))\right] \tag{26}$$

for the case where A $(= B)$ is known.

When the problem is concerned only with the covariance matrix, we consider

$$\rho_3(S_n, S_r') = \frac{|UU_{(1)}|^{1/4}}{|\frac{1}{2}(U + U_{(1)})|^{1/2}}, \tag{27}$$

$$\rho_3(S_n, S_s'') = \frac{|UU_{(2)}|^{1/4}}{|\frac{1}{2}(U + U_{(2)})|^{1/2}}. \tag{28}$$

For instance, when

$$\frac{|UU_{(1)}|^{1/4}}{|\frac{1}{2}(U + U_{(1)})|^{1/2}} \geq \frac{|UU_{(2)}|^{1/4}}{|\frac{1}{2}(U + U_{(2)})|^{1/2}}, \tag{29}$$

we decide that X has F_1, and when

$$\frac{|UU_{(1)}|^{1/4}}{|\frac{1}{2}(U + U_{(1)})|^{1/2}} < \frac{|UU_{(2)}|^{1/4}}{|\frac{1}{2}(U + U_{(2)})|^{1/2}}, \tag{30}$$

we decide that X has F_2. (For the case where these two statistics are equal, we can, of course, determine in advance to take F_2.)

As to the success rate, we obtain

$$P(\rho(S_n, S_r') > \rho(S_n, S_s'')|F_1) \geq P\left(\rho(S_n, F_1) > \frac{1 - \delta}{16}\middle|F_1\right) \times P\left(\rho(S_r', F_1) > \frac{1 - \delta}{16}\right) \cdot P\left(\rho(S_s'', F_2) > \frac{1 - \delta}{16}\right), \tag{31}$$

and from this relation we can get an evaluation of the success rate, when we have the value of $P(\rho(F, S_n) > \delta | F)$. Thus the next problem is to evaluate $P(\rho(F, S_n) > \delta | F)$ $(\delta < 1)$.

Assume that X is distributed according to $N(a, \Sigma)$. First, concerning $\rho_1(F, S_n)$, $\rho_2(F, S_n)$, we have

$$-8n \log \rho_1(F, S_n) = n(V^{-1}(\overline{X} - a), (\overline{X} - a)), \tag{32}$$

$$-8n \log \rho_2(F, S_n) = n(\Sigma^{-1}(\overline{X} - a), (\overline{X} - a)), \tag{33}$$

and, as is well known, the right-hand sides have a noncentral F and a chi-square distribution, and we have no problem here.

Concerning $\rho_3(F, S_n)$, we have

$$\rho_3(F, S_n) = \frac{|\Sigma^{-1}U|^{1/4}}{|\frac{1}{2}(\Sigma^{-1} + U)|^{1/2}}, \tag{34}$$

and

$$P(\rho_3(F, S_n) > \delta) \geq \left[P\left(\frac{4Z}{(1 + Z)^2} > \delta\right)\right]^k, \tag{35}$$

where Z is a random variable such that nZ has the chi-square distribution with n degrees of freedom (see [2]). Therefore, for given positive δ and ϵ (<1), there exists an integer n_0 such that $P(\rho_3(F, S_n) > \delta) \geq 1 - \epsilon$ uniformly in F for $n \geq n_0$.

Now we will present the general case. Let δ_1, δ_2 be positive numbers such that $\delta = \delta_1 \exp[-(1/4)\delta_2]$, $\delta_1 < 1$. Then we get

$$\begin{aligned} &P(\rho(F, S_n) > \delta) \\ &\quad \geq P\left(\frac{|\Sigma^{-1}W^{-1}|^{1/4}}{|\frac{1}{2}(\Sigma^{-1} + W^{-1})|^{1/2}} > \delta_1\right) P(\beta(\Sigma^{-1}(X - a), (X - a)) < 2\delta_2) \end{aligned} \tag{36}$$

where

$$W = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} X_{i1}^2 & & 0 \\ & \ddots & \\ 0 & & \frac{1}{n}\sum_{i=1}^{n} X_{ik}^2 \end{pmatrix}, \qquad X_i = \begin{pmatrix} X_{i1} \\ \cdot \\ \cdot \\ \cdot \\ X_{ik} \end{pmatrix}, \tag{37}$$

$$\beta = \frac{2}{\delta_1^4} - 1 + \frac{2}{\delta_1^4}\sqrt{1 - \delta_1^4} \tag{38}$$

(see [2]). By taking δ_1 (accordingly δ_2) so that the right-hand side becomes maximum, we can get an evaluation (from below) of $P(\rho(F, S_n) > \delta | F)$. When we want to have $P(\rho(F, S_n) > \delta) > 1 - \epsilon$, let $1 - \epsilon = \alpha_1\alpha_2$, $\alpha_1, \alpha_2 > 0$ and take n large so that

$$P\left(\frac{|\Sigma^{-1}W^{-1}|^{1/4}}{|\frac{1}{2}(\Sigma^{-1} + W^{-1})|^{1/2}} > \delta_1\right) \geq \alpha_1, \tag{39}$$

$$P((\Sigma^{-1}(\overline{X} - a), (\overline{X} - a)) < 2\delta_2) \geq \alpha_2. \tag{40}$$

4. Classification by a linear function of vector components

In this section we consider the classification problem by a linear function of components of a random vector.

Let $N(a^{(1)}, \Sigma_1)$, $N(a^{(2)}, \Sigma_2)$ be k-dimensional Gaussian distributions, and let $X = (X_1, \cdots, X_k)$ be a k-dimensional random vector. The problem is to decide which one of $N(a^{(1)}, \Sigma_1)$, $N(a^{(2)}, \Sigma_2)$ is the distribution of X. For this problem we consider a linear function of the components of X of the form $(c, X) = c_1X_1 + \cdots + c_kX_k$, where c is a constant vector ($\neq 0$). The decision procedure is as follows. Let $X^{(1)}$, $X^{(2)}$ be samples from $N(a^{(1)}, \Sigma_1)$, $N(a^{(2)}, \Sigma_2)$, and let F_{1c}, F_{2c} be the distributions of $(c, X^{(1)})$ and $(c, X^{(2)})$. Further, let E_1, E_2 be optimal regions (on the real line) for classifying an observation from F_{1c}, or F_{2c}. (For instance, E_1, E_2 can be defined by the probability ratio rule.) Then, when (c, X) lies in E_1, we decide that X has $N(a^{(1)}, \Sigma_1)$, and when (c, X) lies in E_2, we decide that X has $N(a^{(2)}, \Sigma_2)$. Therefore, for reducing the probability of misclassification, it is necessary to find an adequate c.

Now, we have

$$
\begin{aligned}
E(c, X^{(1)}) &= (c, a^{(1)}),\\
V(c, X^{(1)}) &= (c, \Sigma_1 c),\\
E(c, X^{(2)}) &= (c, a^{(2)}),\\
V(c, X^{(2)}) &= (c, \Sigma_2 c),
\end{aligned}
\tag{41}
$$

and

$$
\rho(F_{1c}, F_{2c}) = \left[\frac{2(c, \Sigma_1 c)^{1/2}(c, \Sigma_2 c)^{1/2}}{(c, \Sigma_1 c) + (c, \Sigma_2 c)}\right]^{1/2} \cdot \exp\left[-\frac{1}{4}\frac{(c, a^{(1)} - a^{(2)})^2}{(c, (\Sigma_1 + \Sigma_2)c)}\right]. \tag{42}
$$

Therefore, from our standpoint, we should choose a c that minimizes $\rho(F_{1c}, F_{2c})$ when $a^{(1)}$, $a^{(2)}$, Σ_1, Σ_2 are known. When $a^{(1)}$, $a^{(2)}$, Σ_1, Σ_2 are unknown, we use in place of them their estimates obtained from samples.

For example, when it is known beforehand that $\Sigma_1 = \Sigma_2$, we consider

$$
\frac{(c, a^{(1)} - a^{(2)})^2}{(c, \Sigma_1 c)} \tag{43}
$$

and determine c so as to maximize this value. (This is a familiar procedure in multivariate analysis.)

REFERENCES

[1] K. MATUSITA, "Decision rule, based on the distance, for the classification problem," *Ann. Inst. Statist. Math.*, Vol. 8 (1956), pp. 67–77.

[2] ———, "A distance and related statistics in multivariate analysis," *Proceedings of the International Symposium on Multivariate Analysis*, New York, Academic Press, 1966.