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 $d(S_n, S_{\nu,n_\nu})$.