ZASTOSOWANIA MATEMATYKI
APPLICATIONES MATHEMATICAE
XVIII, 3 (1984), p. 403-411

M. KRZYŚKO (Poznań)

SEQUENTIAL CLASSIFICATION IN THE CASE OF MANY POPULATIONS

1. Introduction. Suppose that we consider a certain object belonging to one of m general populations $\pi_1, \ldots, \pi_m$ but we do not know to which of them. Our task is to classify the object into the proper population on the basis of the values of measurements of p characteristics of the object. This is a classification problem where all the characteristics of the object may be observed simultaneously, and it is known as a decision procedure with a fixed number of characteristics (cf. [1] and [4]). In this method the cost of measuring the characteristics is not considered. It is obvious that an insufficiently large number of measured characteristics does not allow us to obtain satisfactory results of the classification procedure. On the other hand, the measurement of an excessive number of characteristics is undesirable from a practical viewpoint. A rational interrelation between the number of misclassifications and the number of observed characteristics may be obtained by the sequential observation of characteristics, where the sequential process is terminated when a satisfactory or required level of accuracy is achieved in the classification.

The problem of sequential classification in the case of an infinite number of observed characteristics was first considered by Reed [5]. He introduced the generalized probability ratio and on this basis he proposed the sequential procedure of elimination of populations. In the case of two populations this procedure reduces to Wald's classical sequential probability ratio test. The two modifications of Reed's procedure in the case of a finite number of observed characteristics were given by Fu [2] and Krzyśko [3].

In this paper another form of the sequential classification method for the case of many populations and a finite number of observed characteristics is presented. This method is a modification of the non-sequential Bayesian classification rule.

2. Method. Assume that $f_i(x_1, \ldots, x_k) = f_{ik}(\boldsymbol{x})$ is a known density function of a k-variate random variable $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ik})'$ whose values can be observed in the objects belonging to a population π_i for $i \in M = \{1, \ldots, m\}$ and $k \in P = \{1, \ldots, p\}$. The known a priori probability of the event of classifying the object into population π_i is denoted by q_i ($q_i > 0$, $q_1 + \ldots + q_m = 1$), and the loss caused by a misclassification of an object into population π_j whilst it belongs really to population π_i is denoted by $L(j|i)$ for $i, j \in M$. If $L(j|i)$ is the so-called simple loss function of the form

$$L(j|i) = \begin{cases} 0 & \text{if } j = i, \\ 1 & \text{if } j \neq i, \end{cases}$$

then the Bayes risk r is expressed as

$$r = 1 - \sum_{i=1}^{m} q_i \int\limits_{W_i^{(p)}} f_{ip}(\boldsymbol{x})\, d\boldsymbol{x}$$

and the optimal (in the sense of minimizing the value of r) non-sequential classification region $W_i^{(p)}$ takes the form

$$W_i^{(p)} = \Big\{\boldsymbol{x}: \bigwedge_{\substack{j=1 \\ j \neq i}}^{m} \big(q_i f_{ip}(\boldsymbol{x}) \geqslant q_j f_{jp}(\boldsymbol{x})\big)\Big\}, \qquad i \in M. \tag{1}$$

The non-sequential classification procedure is described as follows. The object with the observed vector $\boldsymbol{x}_0$ of p characteristic values belongs to population π_i if $\boldsymbol{x}_0 \in W_i^{(p)}$, $i \in M$. Every object can be classified by this method but the probabilities of misclassification may be greater than the admissible ones.

We now describe a sequential classification method related to the Bayesian classification method. At each step of this method we can verify the probabilities of misclassification.

The process of sequential classification takes the following form. At the first step we observe only the value x_{01} of the first characteristic of the given object, and we wish to classify it as belonging to one of m populations $\pi_1, \ldots, \pi_m$. Let $\Pr^{(1)}(\pi_i|\pi_j)$ be the probability of misclassification of the object to population π_i on the basis of only one characteristic when in fact it belongs to population π_j, $i, j \in M$, $j \neq i$. We would like to undertake the classification in such a way that the inequalities $\Pr^{(1)}(\pi_i|\pi_j) \leqslant a_{ij}(1)$ are met for values of $a_{ij}(1)$ selected by us in advance, $i, j \in M$, $j \neq i$. For this purpose we define m non-intersecting classification regions of the form $R_i^{(1)} = W_i^{(1)}$, $i \in M$, where $W_i^{(1)}$ is given by (1). We have

$$\Pr^{(1)}(\pi_i|\pi_j) = \Pr(x_1 \in R_i^{(1)}|\pi_j) = \int\limits_{R_i^{(1)}} f_{j1}(x_1)\, dx_1, \qquad i, j \in M,\ j \neq i.$$

If the inequalities

$$\int_{R_i^{(1)}} f_{j1}(x_1)\,dx_1 \leqslant a_{ij}(1), \qquad i, j \in M,\ j \neq i, \tag{2}$$

are satisfied for all $i, j \in M, j \neq i$, then the observation x_{01} is assigned to population π_i when $x_{01} \in R_i^{(1)}$, and the classification process is terminated.

If for each fixed i there exists at least one $j \in M$, $j \neq i$, for which inequalities (2) are not met, then we define the classification regions

$$T_i^{(1)} = \Big\{x_1 : \bigwedge_{\substack{j=1 \\ j \neq i}}^{m} \big(q_i f_{i1}(x_1) \geqslant A_{ij}(1)\, q_j f_{j1}(x_1)\big)\Big\}, \qquad i \in M_1,$$

$$T_0^{(1)} = \Big\{x_1 : \big(x_1 \notin \bigcup_{i \in M \setminus M_1} R_i^{(1)}\big) \wedge \big(x_1 \notin \bigcup_{i \in M_1} T_i^{(1)}\big)\Big\},$$

where M_1 is the set of those indices of populations for which inequalities (2) are not satisfied, $A_{ij}(1)$ are constants fulfilling the inequality $A_{ij}(1) \geqslant 1$ whose values are related to the probabilities of misclassification. The boundaries $A_{ij}(1)$ are chosen so as to ensure that the inequality

$$\Pr^{(1)}(\pi_i \mid \pi_j) = \Pr(x_1 \in T_i^{(1)} \mid \pi_j) \leqslant a_{ij}(1)$$

is met for the given values of $a_{ij}(1)$, i.e. they satisfy

$$\int_{T_i^{(1)}} f_{j1}(x_1)\,dx_1 \leqslant a_{ij}(1), \qquad i \in M_1,\ j \in M,\ j \neq i.$$

$T_0^{(1)}$ is the region in which no classification can be made on the basis of only one characteristic.

We now verify whether there is such an $i_0 \in M \setminus M_1$ for which $x_{01} \in R_{i_0}^{(1)}$ or such an $i_0 \in M_1$ for which $x_{01} \in T_{i_0}^{(1)}$. If it does exist, then we decide that the object is a member of population π_{i_0}. If however such an i_0 does not exist, i.e. if $x_{01} \in T_0^{(1)}$, we observe x_{02}, the value of the second characteristic of the object being classified.

In the case of two variables we use truncated distributions considered over the region

$$S^{(2)} = T_0^{(1)} \times \{x_2 : -\infty < x_2 < \infty\}.$$

Let us put

$$f_{i2}^{u}(x_1, x_2) = c_{i2}^{-1} f_{i2}(x_1, x_2),$$

where

$$c_{i2} = \int_{S^{(2)}} f_{i2}(x_1, x_2)\,dx_1\,dx_2, \qquad i \in M.$$

We define m non-intersecting regions of the form

$$R_i^{(2)} = \{(x_1, x_2): \bigwedge_{\substack{j=1 \\ j \neq i}}^{m} (f_{i2}^u(x_1, x_2) \geqslant f_{j2}^u(x_1, x_2))\} \times T_0^{(1)}, \quad i \in M.$$

We have

$$\Pr^{(2)}(\pi_i | \pi_j) = \Pr((x_1, x_2) \in R_i^{(2)} | \pi_j) = \int_{R_i^{(2)}} f_{j2}^u(x_1, x_2) dx_1 dx_2, \quad i, j \in M, j \neq i.$$

If

$$\int_{R_i^{(2)}} f_{j2}^u(x_1, x_2) dx_1 dx_2 \leqslant a_{ij}(2), \quad i, j \in M, j \neq i, \tag{3}$$

then the observation (x_{01}, x_{02}) may be classified as belonging to population π_i when $(x_{01}, x_{02}) \in R_i^{(2)}$, and the classification process is terminated.

If for each fixed i there exists at least one $j \in M$, $j \neq i$, for which inequalities (3) are not met, then we define the following two-dimensional classification regions:

$$T_i^{(2)} = \{(x_1, x_2): \bigwedge_{\substack{j=1 \\ j \neq i}}^{m} (f_{i2}^u(x_1, x_2) \geqslant A_{ij}(2) f_{j2}^u(x_1, x_2))\} \times T_0^{(1)}, \quad i \in M_2,$$

$$T_0^{(2)} = \{(x_1, x_2): ((x_1, x_2) \notin \bigcup_{i \in M \setminus M_2} R_i^{(2)}) \wedge ((x_1, x_2) \notin \bigcup_{i \in M_2} T_i^{(2)})\} \times T_0^{(1)},$$

where M_2 is the set of those indices of populations for which inequalities (3) are not met and $A_{ij}(2)$ are constants fulfilling the inequality $A_{ij}(2) \geqslant 1$ whose values are related to the probabilities of misclassification. The boundaries $A_{ij}(2)$ are chosen so as to ensure that the inequality

$$\Pr^{(2)}(\pi_i | \pi_j) = \Pr((x_1, x_2) \in T_i^{(2)} | \pi_j) \leqslant a_{ij}(2)$$

is satisfied for given values of $a_{ij}(2)$, i.e.

$$\int_{T_i^{(2)}} f_{j2}^u(x_1, x_2) dx_1 dx_2 \leqslant a_{ij}(2), \quad i \in M_2, \ j \in M, \ j \neq i.$$

The region $T_0^{(2)}$ is that one in which no classification can be made on the basis of the values of the first two characteristics.

We next verify whether there is an $i_0 \in M \setminus M_2$ for which $(x_{01}, x_{02}) \in R_i^{(2)}$ or an $i_0 \in M_2$ for which $(x_{01}, x_{02}) \in T_{i_0}^{(2)}$. If it exists, then we decide that the given object is a member of population π_{i_0}. If no such i_0 exists, we observe the third feature of the object x_{03}.

The classification process is continued until we have decided that the object belongs to one of the populations $\pi_1, \ldots, \pi_m$ or to the exhaustion of the predetermined number p of characteristics. In general, after observing the value of x_{0k}, the k-th random variable, we use truncated distributions

$$f_{ik}^u(x_1, \ldots, x_k) = c_{ik}^{-1} f_{ik}(x_1, \ldots, x_k)$$

considered over the region

$$S^{(k)} = T_0^{(k-1)} \times \{x_k : -\infty < x_k < \infty\},$$

where

$$c_{ik} = \int_{S^{(k)}} f_{ik}(x_1, \ldots, x_k)\, dx_1 \ldots dx_k, \qquad i \in M,\ k \in P,$$

and

$$T_0^{(0)} \equiv \{x_1 : -\infty < x_1 < \infty\}.$$

Let us define m non-intersecting regions

$$R_i^{(k)} = \Big\{(x_1, \ldots, x_k) : \bigwedge_{\substack{j=1 \\ j \neq i}}^{m} \big(f_{ik}^u(x_1, \ldots, x_k) \geqslant f_{jk}^u(x_1, \ldots, x_k)\big)\Big\} \times T_0^{(k-1)},$$

$$i \in M,\ k \in P.$$

We have

$$\Pr^{(k)}(\pi_i | \pi_j) = \Pr\big((x_1, \ldots, x_k) \in R_i^{(k)} | \pi_j\big) = \int_{R_i^{(k)}} f_{jk}^u(x_1, \ldots, x_k)\, dx_1 \ldots dx_k,$$

$$i, j \in M,\ j \neq i,\ k \in P.$$

If

$$\int_{R_i^{(k)}} f_{jk}^u(x_1, \ldots, x_k)\, dx_1 \ldots dx_k \leqslant \alpha_{ij}(k), \qquad i, j \in M,\ j \neq i,\ k \in P, \tag{4}$$

then the observation $(x_{01}, x_{02}, \ldots, x_{0k})$ may be classified as belonging to population π_i when $(x_{01}, x_{02}, \ldots, x_{0k}) \in R_i^{(k)}$, and the classification process is terminated.

If for each fixed i there exists at least one $j \in M$, $j \neq i$, for which inequalities (4) are not met, then we define the classification regions

$$T_i^{(k)} = \Big\{(x_1, \ldots, x_k) : \bigwedge_{\substack{j=1 \\ j \neq i}}^{m} \big(f_{ik}^u(x_1, \ldots, x_k) \geqslant A_{ij}(k) f_{jk}^u(x_1, \ldots, x_k)\big)\Big\} \times T_0^{(k-1)}, \qquad i \in M_k, \tag{5}$$

$$T_0^{(k)} = \Big\{(x_1, \ldots, x_k) : \big((x_1, \ldots, x_k) \notin \bigcup_{i \in M \setminus M_k} R_i^{(k)}\big) \wedge \big((x_1, \ldots, x_k) \notin \bigcup_{i \in M_k} T_i^{(k)}\big)\Big\} \times T_0^{(k-1)},$$

where M_k is the set of those indices of populations for which inequalities (4) are not met, $A_{ij}(k)$ are constants fulfilling the inequality $A_{ij}(k) \geqslant 1$ whose values are related to the probabilities of misclassification. The constants $A_{ij}(k)$ are chosen so as to ensure that the inequality

$$\Pr^{(k)}(\pi_i|\pi_j) = \Pr\left((x_1, \ldots, x_k) \in T_i^{(k)} \mid \pi_j\right) \leqslant \alpha_{ij}(k)$$

is satisfied for given values of $\alpha_{ij}(k)$, i.e.

$$\int_{T_i^{(k)}} f_{jk}^u(x_1, \ldots, x_k)\, dx_1 \ldots dx_k \leqslant \alpha_{ij}(k), \tag{6}$$

$$i \in M_k,\ j \in M,\ j \neq i,\ k \in P.$$

$T_0^{(k)}$ is the region in which no decision about the classification of the object can be made on the basis of the values of the k characteristics. We now verify whether there exists such an $i_0 \in M \setminus M_k$ for which $(x_{01}, \ldots, x_{0k}) \in R_{i_0}^{(k)}$ or an $i_0 \in M_k$ for which $(x_{01}, \ldots, x_{0k}) \in T_{i_0}^{(p)}$. If it exists, then we decide that the object belongs to population π_{i_0}. If not, i.e. if $(x_{01}, \ldots, x_{0k}) \in T_0^{(k)}$, then we observe the value of the next characteristic of the given object provided $k < p$. If $k = p$ and $(x_{01}, \ldots, x_{0p}) \in T_0^{(p)}$, then we cannot classify the given object within the probabilities chosen in advance. In this exceptional situation, the observation $(x_{01}, \ldots, x_{0p})$ should be classified as belonging to population π_i if $(x_{01}, \ldots, x_{0p}) \in W_i^{(p)}$, $i \in M$, where $W_i^{(p)}$ is given by (1), but in this case the probabilities of misclassification may be greater than those selected in advance.

In order to find the values of $A_{ij}(k)$ for which inequality (6) is satisfied, it is useful to know the intervals within which the $A_{ij}(k)$ vary. We have already known that the $A_{ij}(k)$ must satisfy the inequality $A_{ij}(k) \geqslant 1$. Now we turn to the upper bounds for given values of $\alpha_{ij}(k)$, $i, j \in M$, $k \in P$.

The following relation holds:

$$\sum_{i=0}^{m} \int_{T_i^{(k)}} f_{jk}^u(x)\, dx = 1, \quad j \in M.$$

Hence

$$\int_{T_i^{(k)}} f_{ik}^u(x)\, dx \leqslant 1 - \sum_{\substack{j=1 \\ j \neq i}}^{m} \int_{T_j^{(k)}} f_{ik}^u(x)\, dx, \quad i \in M. \tag{7}$$

Let us integrate the function $f_{ik}^u(x)$ over the region $T_i^{(k)}$. Using (5) we obtain

$$\int_{T_i^{(k)}} f_{ik}^u(x)\, dx \geqslant A_{ij}(k) \int_{T_i^{(k)}} f_{jk}^u(x)\, dx, \quad i, j \in M,\ j \neq i. \tag{8}$$

From (7) and (8) we have

$$1-\sum_{\substack{j=1\\ j\neq i}}^{m}\alpha_{ji}(k) = 1-\sum_{\substack{j=1\\ j\neq i}}^{m}\int_{T_j^{(k)}} f_{ik}^{u}(x)\,dx$$

$$\geqslant \int_{T_i^{(k)}} f_{ik}^{u}(x)\,dx \geqslant A_{ij}(k)\int_{T_i^{(k)}} f_{jk}^{u}(x)\,dx = A_{ij}(k)\,\alpha_{ij}(k).$$

Thus the boundaries $A_{ij}(k)$ fulfill the inequality

$$1 \leqslant A_{ij}(k) \leqslant \Big(1-\sum_{\substack{j=1\\ j\neq i}}^{m}\alpha_{ji}(k)\Big)/\alpha_{ij}(k), \qquad i,j\in M,\ j\neq i.$$

3. Ordering of the characteristics. The order in which the characteristics of the object to be classified are observed is essential. To ensure high effectiveness of the sequential classification it is necessary to choose, at each step, the most strongly discriminant characteristic of the populations $\pi_1,\ldots,\pi_m$. Such a choice guarantees the maximum reduction of the probability of misclassification together with a fast termination of the classification procedure.

Denote by $\Pr^{(k)}(\pi_j^c\,|\,\pi_j)$ the probability of misclassifying an object belonging to population π_j on the basis of k characteristics observed sequentially, where π_j^c stands for all the populations with the exception of π_j, $j\in M$, $k\in P$.

The combination of k characteristics discriminates the populations $\pi_1,\ldots,\pi_m$ more strongly in proportion to the reduction in size of

$$\sum_{j=1}^{m}\Pr^{(k)}(\pi_j^c\,|\,\pi_j).$$

We have

$$\Pr^{(k)}(\pi_j^c\,|\,\pi_j) = \sum_{\substack{i=1\\ j\neq i}}^{m}\Pr^{(k)}(\pi_i\,|\,\pi_j) = \sum_{\substack{i=1\\ i\neq j}}^{m}\int_{T_i^{(k)}} f_{jk}^{u}(x)\,dx \leqslant 1-\int_{T_j^{(k)}} f_{jk}^{u}(x)\,dx,$$

whence

$$\sum_{j=1}^{m}\Pr^{(k)}(\pi_j^c\,|\,\pi_j) \leqslant m-\sum_{j=1}^{m}\int_{T_j^{(k)}} f_{jk}^{u}(x)\,dx, \qquad k\in P.$$

From this inequality we infer that the combination of k characteristics discriminates the populations $\pi_1, \ldots, \pi_m$ more strongly in proportion to the increase in size of the expression

$$C(k) = \sum_{j=1}^{m} \int_{T_j^{(k)}} f_{jk}^{u}(\boldsymbol{x})\, d\boldsymbol{x}.$$

To determine the optimal order of the observed characteristics, we proceed as follows:

The value of the expression $C(1)$ is calculated for each characteristic separately. The characteristic which ensures the maximum value of $C(1)$ is chosen. Successively, all the remaining characteristics are added to the one chosen initially, forming $p-1$ pairs of characteristics. Next, the value of $C(2)$ is calculated for each formed pair. That pair is selected which leads to a maximum value of the expression $C(2)$. This operation is repeated until the set of $p-1$ characteristics, out of the p characteristics, is found which gives the maximum value of the expression $C(p-1)$.

4. Conclusion. If the cost of measuring the characteristics is considered or if the characteristics of the given object appear sequentially, a sequential method of classification should be used. Such problems can arise, e.g., when the given characteristics are to be measured during a production process, where the measurement calls for the interruption of the process, or when the measurement is time -consuming, requires using the complicated measuring equipment, or is associated with complex operations involving risk (as in biomedical applications).

The sequential method of classification described in this paper requires the knowledge of density functions. In practice we often assume that the random vector $\boldsymbol{X}_i$, $i \in M$, has a multivariate normal distribution. However, a situation may occur in which the form of the density function is known but where the parameters are unknown. In this case, the parameters must be estimated from samples. In addition, a situation may occur in which neither the form of the density function nor its parameters are known. Then the density function must be estimated from a sample, without any assumptions made in advance about the function. Many methods for the non-parametric estimation of density functions are known, e.g., the kernel method, the orthonormal series method, or the nearest neighbour method. In the sequential method of classification described in this paper we must choose the admissible values of the probabilities of misclassification. We can use the following procedure. We determine the admissible value of $\Pr^{(k)}(\pi_j^c|\pi_j)$ of the probability of misclassification of an object belonging to population π_j at the level $\alpha_j(k)$, $j \in M$,

$k \in P$. If we have no reason for choosing any particular values of $a_{ij}(k)$, then we may put $a_{ij}(k) = (m-1)^{-1}$ for all $i, j \in M$, $k \in P$.

Acknowledgement. The autor would like to thank Professor T. Caliński for very helpful suggestions.

References

[1] T. W. Anderson, *An introduction to multivariate statistical analysis*, New York 1958.

[2] K. S. Fu, *Sequential methods in pattern recognition and machine learning*, New York 1968.

[3] M. Krzyśko, *A sequential multidecision procedure*, Biom. J. 23 (1981), p. 159-165.

[4] C. R. Rao, *Linear statistical inference and its applications*, New York 1965.

[5] F. C. Reed, *A sequential multidecision procedure*, p. 42-69 in: Proc. Symp. on Decision Theory and Appl. Electron. Equipment Develop., USAF Develop. Center, Rome, New York 1960.

INSTITUTE OF MATHEMATICS
ADAM MICKIEWICZ UNIVERSITY
60-769 POZNAŃ

Received on 20. 3. 1980;
revised version on 2. 8. 1982