APPLIED STOCHASTIC MODELS AND DATA ANALYSIS
Appl. Stochastic Models Data Anal. **15**, 65–75 (1999)

ASYMPTOTICAL OPTIMALITY IN CLUSTER ANALYSIS

YURIJ S. KHARIN* AND EUGENE E. ZHUK

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

SUMMARY

The problem of optimality and performance evaluation for cluster analysis procedures is investigated. For the situations where the classes are described by known or unknown prior probabilities and regular probability density functions with unknown parameters the asymptotic expansions of classification error probability are constructed. The results are illustrated for the case of well-known Fisher classification model.

KEY WORDS: cluster analysis; asymptotical optimality; 'plug-in' decision rule; classification error probability; asymptotic expansions

1. INTRODUCTION: MATHEMATICAL MODEL

The problem of performance evaluation and optimality analysis for statistical classification decision rules remains one of the most important problems in statistical cluster analysis theory and its applications. This fact is caused not only by actual demands from practice but also by a special dominant role of this problem detected by Glick:[1] 'the task of estimating probabilities of correct classification confronts the statistician simultaneously with difficult distribution theory, questions interwining sample size and dimension, problems of bias, variance, robustness and computation cost. But coping with such conflicting concerns enhances understanding of many aspects of statistical classification'.

Majority of results of performance evaluation and optimality analysis for statistical decision rules are concerned with the situation, where the training sample used for decision rule construction is classified and decision rules are being constructed by discriminant analysis methods. Review of these results can be found in the monographs of Lachenbruch,[2] Raudys[3] and McLachlan.[4]

In applied classification problems with poor statistical data the training samples are often unclassified, and methods of cluster analysis (self-training classification) are used for classification of training samples and also for classification of new-registered observations.

The main method of optimality analysis used up to now for these situations is Monte-Carlo simulation. Some analytical results are discussed in monographs of Bock[5] and Kharin.[6] This paper is devoted to optimality analysis for self-training statistical decision rules by asymptotic expansion method.

* Correspondence to Y. S. Kharin, Department of Mathematical Modeling and Data Analysis, Belarusian State University, 4 Fr.Skariny av., Minsk 220050, Belarus.

CCC 8755–0024/99/010065–11$17.50

Received December 1996
Revised October 1998

Define now the mathematical model. Let

$$Q=\{q(x;\theta_*),x\in R^N:\theta_*\in\Theta\subseteq R^m\} \tag{1}$$

be a parametric family of probability density functions in R^N and $L\geqslant 2$ different parameter values $\{\theta_1^o,\ldots,\theta_L^o\}$ be fixed. The set $\{\theta_1^o,\ldots,\theta_L^o\}$ determines L classes $\{\Omega_1,\ldots,\Omega_L\}$. The class Ω_i is described by the probability density function (p.d.f.) $q(\cdot;\theta_i^o)$, $i\in S=\{1,\ldots,L\}$.

Let a sample of n jointly independent random observations $x_1,\ldots,x_n$ from the classes $\{\Omega_i\}_{i\in S}$ be registered in R^N. Introduce the notations: $d_t^o\in S$ is an unknown random index of the class to which the observation x_t belongs:

$$P\{d_t^o=i\}=\pi_i^o,\quad i\in S \tag{2}$$

where $\{\pi_i^o\}_{i\in S}$ ($\sum_{i\in S}\pi_i^o=1$) are prior class probabilities; $D^o=(d_1^o,\ldots,d_n^o)^T$ is the vector of true classification indices for the sample $X=(x_1^T\vdots\ldots\vdots x_n^T)^T$, where T is the transposition symbol. The parameters $\{\theta_i^o\}_{i\in S}$ and the prior probabilities $\{\pi_i^o\}_{i\in S}$ are usually unknown. The cluster analysis problem consists in construction of a decision rule (DR)

$$d=d(x,X):R^N\times R^{nN}\to S \tag{3}$$

for classifying a random observation $x\in R^N$. The DR (3) allows to calculate the estimate $\hat{D}=(\hat{d}_1,\ldots,\hat{d}_n)^T$ for the true classification vector of the sample X: $\hat{d}_t=d(x_t,X)$, $t=1,\ldots,n$, and also to classify the new-registered observations $x_{n+1},x_{n+2},\ldots\in R^N$.

2. 'PLUG-IN' DECISION RULES AND OPTIMALITY MEASURES

Note that if all class characteristics $\{\pi_i^o,\theta_i^o\}_{i\in S}$ are *a priori* known, then the Bayesian DR (BDR)

$$d=d(x;\pi^o,\theta^o)=\arg\max_{i\in S}\{\pi_i^o q(x;\theta_i^o)\} \tag{4}$$

where $\pi^o=(\pi_1^o,\ldots,\pi_L^o)^T$, $\theta^o=(\theta_1^{oT}\vdots\ldots\vdots\theta_L^{oT})^T$, classifies a random observation x from the class Ω_{d^o} with the minimal value of classification error probability:

$$r_o=P\{d\neq d^o\}=1-\int_{R^N}\max_{i\in S}\{\pi_i^o q(x;\theta_i^o)\}\,dx \tag{5}$$

Now consider the situation, where $\{\theta_i^o\}_{i\in S}$ (case A) or $\{\pi_i^o,\theta_i^o\}_{i\in S}$ (case B) are unknown. Denote by $\tilde{\theta}$ and $\{\hat{\pi},\hat{\theta}\}$ any statistical estimators for π^o,θ^o obtained by the sample X for cases A and B, respectively. Instead of the BDR (4) let us use the 'plug-in' DR (PDR):

$$d_A(x,X)=d(x;\pi^o,\tilde{\theta}),\qquad d_B(x,X)=d(x;\hat{\pi},\hat{\theta}) \tag{6}$$

which are derived from the BDR $d(x;\pi^o,\theta^o)$ by substitution of statistical estimators for unknown values of class characteristics.

As the optimality measures of PDR $d_j(\cdot)$, $j\in\{A,B\}$, as in References 6 and 7, let us define:

(a) the classification error probability (CEP):

$$r_j(n)=P\{d_j(x,X)\neq d^o\},\quad j\in\{A,B\}, \tag{7}$$

which means error probability at classifying a new-registered random observation x (x is independent on the random sample X);

(b) the relative bias of the CEP:

$$\kappa_j(n) = \frac{r_j(n) - r_\infty}{r_\infty} > 0, \quad j \in \{A, B\} \tag{8}$$

where $r_\infty > 0$ is defined in (5); the smaller $\kappa_j(n)$ is, the more effective PDR $d_j(\cdot)$ is ($j \in \{A, B\}$). If for increasing sample size n

$$\lim_{n \to \infty} \kappa_j(n) = 0 \tag{9}$$

then the PDR $d_j(\cdot, X)$ is asymptotically optimal;

(c) the δ-admissible sample size:

$$n_j^* = n_j^*(\delta) = \min\{n : \kappa_j(n) \leqslant \delta\}, \quad j \in \{A, B\} \tag{10}$$

where $\delta > 0$ is any predetermined value; if $n \geqslant n_j^*(\delta)$, then the inequality $\kappa_j(n) \leqslant \delta$ holds ($j \in \{A, B\}$).

Note that the optimality measures (7)–(10) make possible to investigate symptotical optimality of PDR (6) with respect to the sample size n by means of risk asymptotic expansion method, which was proposed by Kharin.[6]

3. OPTIMALITY OF 'PLUG-IN' DECISION RULES BASED ON *ML*-ESTIMATES

Let in R^N with prior probabilities π_1^o, $\pi_2^o = 1 - \pi_1^o$ random observations from $L = 2$ classes Ω_1, Ω_2 are being registered (note that the results of this section also can be generalized for $L > 2$ classes).

The training sample $X = (x_1^T : \ldots : x_n^T)^T$ is a random sample of size n from the mixture of two distributions:

$$q_{\pi^o}(x; \theta^o) = \pi_1^o q(x; \theta_1^o) + (1 - \pi_1^o) q(x; \theta_2^o) \tag{11}$$

Because X is a unclassified sample and we want to avoid non-uniqueness, derived by denotations of classes, we assume, that $\theta_2^o \succ \theta_1^o$ (here $\succ$ is the symbol of lexicographic comparison). In this model $q_{\pi^o}(\cdot; \theta^o)$ is an element of the family of mixtures:

$$Q_{\pi^o} = \{q_{\pi^o}(x; \theta), x \in R^N : \theta = (\theta_1^T : \theta_2^T)^T \in \Theta^2 \subseteq R^{2m}, \theta_2 \succ \theta_1\} \tag{12}$$

At first, let us investigate the case A (the prior probability π_1^o of the class Ω_1 is assumed to be known). As an estimator $\tilde{\theta}$ we use the maximum likelihood estimator (MLE):

$$\tilde{\theta} = \arg\max_{\theta} \sum_{t=1}^{n} \ln q_{\pi^o}(x_t; \theta) \tag{13}$$

Note that to solve the multiextremum problem (13) the different numerical methods may be used (see, for example, References 8 and 9).

Assume that the family of p.d.f.s (1) satisfies the following regularity conditions:

(C_1) θ_* is an identifiable parameter of p.d.f. $q(\cdot; \theta_*)$:

$$E_{\theta_*}\{\ln q(x; \theta_*)\} > E_{\theta_*}\{\ln q(x; \theta_{**})\}, \quad \theta_* \neq \theta_{**}$$

where for any function $f(x;\theta_{\bullet\bullet})$

$$E_{\theta_\bullet}\{f(x;\theta_{\bullet\bullet})\}=\int_{R^N} f(x;\theta_{\bullet\bullet})q(x;\theta_\bullet)\,\mathrm{d}x$$

(C_2) For any compact $K\subset\Theta$ and any points $\theta_1^o,\theta_2^o\in K$ some neighbourhoods $U_{\theta_1^o},U_{\theta_2^o}\subset K$ exist, such that for some $a,c>1$, $b>2$, any neighbourhood $U\subset U_{\theta_1^o}$ and any $\theta_1\in U_{\theta_1^o}$, $\theta_2\in U_{\theta_2^o}$ the functions

$$|\ln q(x;\theta_k)|^a,\qquad \left(\sup_{\theta'\in U}|\ln q(x;\theta')|\right)^a$$

$$\left|\frac{\partial^2\ln q(x;\theta_k)}{\partial\theta_{ki}\partial\theta_{kj}}\right|^b,\qquad \left|\frac{\partial\ln q(x;\theta_k)}{\partial\theta_{ki}}\cdot\frac{\partial\ln q(x;\theta_\bullet)}{\partial\theta_{\bullet j}}\right|^b$$

$$\left|\frac{\partial^3\ln q(x;\theta_k)}{\partial\theta_{ki}\partial\theta_{kj}\partial\theta_{kt}}\right|^c,\qquad \left|\frac{\partial\ln q(x;\theta_s)}{\partial\theta_{st}}\cdot\frac{\partial^2\ln q(x;\theta_k)}{\partial\theta_{ki}\partial\theta_{kj}}\right|^c$$

are uniformly integrable with respect to p.d.f. $q(x;\theta_\bullet)$, $x\in R^N$ ($\theta_\bullet\in K$; $k,s\in S=\{1,2\}$; $i,j,t=\overline{1,m}$).

(C_3) $E_{\theta_k^o}\{\nabla_{\theta_k^o}\ln q(x;\theta_k^o)\}=0_m$, $\theta_k^o\in\Theta$,

where 0_m is m-vector, all elements of which are equal to 0;

(C_4) Fisher information matrices

$$H_k=E_{\theta_k^o}\{-\nabla^2_{\theta_k^o}\ln q(x;\theta_k^o)\},\quad k\in S$$

$$J=J(\theta^o)=E_{\theta^o}\{-\nabla^2_{\theta^o}\ln q_{\pi^o}(x;\theta^o)\}$$

where

$$E_{\theta^o}\{g(x)\}=\int_{R^N} g(x)q_{\pi^o}(x;\theta^o)\,\mathrm{d}x$$

are positively defined, so that the minimal eigenvalues of these matrices are separated from zero.

Let us construct asymptotic expansion of the CEP $r_A(n)$ at $n\to+\infty$. Denote by

$$\Gamma=\{x: G(x;\theta^o)=0\}\subset R^N \tag{14}$$

the Bayesian discriminant hypersurface, where

$$G(x;\theta^o)=(1-\pi_1^o)q(x;\theta_2^o)-\pi_1^o q(x;\theta_1^o) \tag{15}$$

$$\alpha=\frac{1}{2}\int_\Gamma(\nabla_{\theta^o}G(x;\theta^o))^{\mathrm{T}}J^{-1}\nabla_{\theta^o}G(x;\theta^o)|\nabla_x G(x;\theta^o)|^{-1}\,\mathrm{d}\mathscr{S}_{N-1}\geqslant 0 \tag{16}$$

$\mathbf{1}(z)=\{1 \text{ if } z\geqslant 0; 0, \text{ if } z<0\}$ is the unit function.

Theorem 1

If the conditions (C_1)–(C_4) are satisfied, p.d.f.s $\{q(x;\theta_k^o)\}_{k\in S}$ have derivatives w.r.t. x ($x\in R^N$, $k\in S$), and surface integral (16) is finite: $\alpha<+\infty$, then the CEP of PDR

$$d_A(x,X)=\mathbf{1}(G(x;\tilde\theta))+1$$

allows the asymptotic expansion:

$$r_A(n) = r_0 + \alpha n^{-1} + O(n^{-3/2}) \tag{17}$$

where

$$r_0 = 1 - \pi_1^0 - \int_{R^1} \mathbf{1}(G(x;\theta^0))G(x;\theta^0)\,dx \tag{18}$$

Proof. Under the regularity conditions (C_1) (C_4) using Chibisov[10] stochastic expansion for the random deviation $\Delta\theta = \tilde{\theta} - \theta^0$ of MLE (13) we find, that $\Delta\theta$ has the third-order moments, and the following asymptotic expansions are true at $\tau = 1/\sqrt{n} \to 0$:

for the bias:

$$E\{\Delta\theta\} = 1_{2m} O(\tau^3) \tag{19}$$

for the covariance matrix of $\tilde{\theta}$:

$$E\{\Delta\theta(\Delta\theta)^{\mathrm{T}}\} = J^{-1}\tau^2 + 1_{2m\times 2m} O(\tau^3) \tag{20}$$

for third-order moments $(k,l,s \in S;\ i,j,t = \overline{1,m})$:

$$E\{|(\tilde{\theta}_{ki} - \theta_{ki}^0)(\tilde{\theta}_{lj} - \theta_{lj}^0)(\tilde{\theta}_{st} - \theta_{st}^0)|\} = O(\tau^3) \tag{21}$$

where 1_{2m}, $1_{2m\times 2m}$ are $(2m)$-vector and $(2m \times 2m)$-matrix, all elements of which are equal to 1.

Now expression (17) is obtained by applying of Taylor formula to the CEP (7):

$$r_A(n) = 1 - \pi_1^0 - E\left\{\int_{R^1} \mathbf{1}(G(x;\tilde{\theta}))G(x;\theta^0)\,dx\right\}$$

in the neighbourhood of θ^0 with respect to $\Delta\theta$, and of the relations (19)–(21). □

Corollary

Under the conditions of Theorem 1 the decision rule $d_A(x, X)$ is asymptotically optimal: $\kappa_A(n) \to 0$, if $n \to +\infty$.

Now let us consider the situation, where both the parameter values θ_1^0, θ_2^0 and the prior probability π_1^0 are unknown. In this case the composed vector $\eta^0 = (\theta_1^{0\mathrm{T}} : \theta_2^{0\mathrm{T}} : \pi_1^0)^{\mathrm{T}}$ of unknown parameters is $(2m+1)$-dimensional vector and its MLE $\hat{\eta} = (\hat{\theta}_1^{\mathrm{T}} : \hat{\theta}_2^{\mathrm{T}} : \hat{\pi}_1)^{\mathrm{T}}$ has the form

$$\hat{\eta} = \arg \max_{\substack{\theta, \\ 0<\pi_1<1}} \sum_{t=1}^{n} \ln q_\pi(x_t;\theta) \tag{22}$$

Theorem 2

If the conditions of Theorem 1 hold, then

$$r_B(n) = r_0 + (\alpha + \beta + \gamma)n^{-1} + O(n^{-3/2}) \tag{23}$$

where

$$\beta = \frac{\pi_1^o}{2(1-\pi_1^o)} \int_{\Gamma} (q(x;\theta_1^o))^2 |\nabla_x G(x;\theta^o)|^{-1} \, d\mathcal{S}_{N-1} > 0$$

and $\gamma = O(r_0) > 0$.

Proof. Is conducted analogously to the proof of Theorem 1. □

It is seen from Theorems 1 and 2 that

$$r_B(n) = r_A(n) + (\beta + \gamma)n^{-1} + O(n^{-3/2})$$

where the term $(\beta + \gamma)n^{-1}$ is derived by random errors of estimation of the prior probability π_1^o in the case B.

Note that the DR $d_B(x, X)$ also is the asymptotically optimal decision rule.

4. OPTIMALITY OF DECISION RULES BASED ON LIKELIHOOD FUNCTION

Let us investigate the case of $L \geqslant 2$ classes and define composite probability distribution of the random sample $X = (x_1^{\mathrm{T}} \vdots \ldots \vdots x_n^{\mathrm{T}})^{\mathrm{T}} \in R^{nN}$ and the random true classification vector $D^o = (d_1^o, \ldots, d_n^o)^{\mathrm{T}} \in S^n$:

$$p(X, D^o | \pi^o, \theta^o) = \prod_{t=1}^{n} \pi_{d_t^o}^o q(x_t; \theta_{d_t^o}^o) \tag{24}$$

The logarithmic likelihood function, which corresponds to (24), has the form

$$l(\pi, \theta, D) = n^{-1} \ln p(X, D | \pi, \theta) = n^{-1} \sum_{t=1}^{n} \ln (\pi_{d_t} q(x_t; \theta_{d_t})) \tag{25}$$

As in Section 3, let us consider the following two cases.

(A) Prior probability vector $\pi^o = (\pi_1^o, \ldots, \pi_L^o)^{\mathrm{T}}$ is known. In this case we need to estimate the vectors D^o and $\theta^o = (\theta_1^{o\mathrm{T}} \vdots \ldots \vdots \theta_L^{o\mathrm{T}})^{\mathrm{T}}$:

$$l(\pi^o, \theta, D) \to \max_{\theta \in \Theta^L, D \in S^n} . \tag{26}$$

(B) All class characteristics π^o and θ^o are unknown, then we need to solve the following problem:

$$l(\pi, \theta, D) \to \max_{\substack{\theta \in \Theta^L, D \in S^n, \\ \pi : 0 < \pi_i < 1, \pi_1 + \cdots + \pi_L = 1}} . \tag{27}$$

Denote by

$$W_n(\pi, \theta) = \max_{D \in S^n} l(\pi, \theta, D) = n^{-1} \sum_{t=1}^{n} f(x_t; \pi, \theta) \tag{28}$$

the statistical estimator for the functional

$$W(\pi,\theta)=\int_{R^N} f(x;\pi,\theta)q_{\pi^0}(x;\theta^0)\,dx \tag{29}$$

where

$$f(x;\pi,\theta)=\max_{i\in S}\ln(\pi_i q(x;\theta_i))$$

$$q_{\pi^0}(x;\theta^0)=\sum_{i\in S}\pi_i^0 q(x;\theta_i^0)$$

The optimization problems (26) and (27) can be rewritten in the form of PDR:

$$d_A(x,X)=\arg\max_{i\in S}\{\pi_i^0 q(x;\tilde\theta_i)\},\qquad \tilde\theta=\arg\max_{\theta} W_n(\pi^0,\theta) \tag{30}$$

$$d_B(x,X)=\arg\max_{i\in S}\{\hat\pi_i q(x;\hat\theta_i)\},\qquad \{\hat\pi,\hat\theta\}=\arg\max_{\pi,\theta} W_n(\pi,\theta) \tag{31}$$

Note, that the PDR (30) and (31) are formal notations of the problems (26) and (27). Unlike Section 3 the processes of unknown class characteristic estimation and classification of the sample X take place simultaneously. But expressions (30) and (31) are more convenient for investigation than expressions (26) and (27).

As in Section 3, to avoid non-uniqueness assume that the components of the estimates $\hat\pi,\hat\theta$ and $\tilde\theta$ are lexicographically ordered and correspond to the true values π^0 and θ^0.

To investigate optimality of the procedures (26) and (27), let us construct the asymptotic expansions of the CEP for the PDR (30) and (31). Introduce the following notations:

$$\rho_+=\max_{\substack{i\neq j\\ i,j\in S}}\int_{R^N}\sqrt{q(x;\theta_i^0)q(x;\theta_j^0)}\,dx \tag{32}$$

$$F_{ij}(x)=\pi_i^0 q(x;\theta_i^0)-\pi_j^0 q(x;\theta_j^0),\quad \Gamma_{ij}=\{x:\ F_{ij}(x)=0\}\subset R^N(i\neq j\in S)$$

$$B_{klij}=\pi_k^0\pi_l^0\int_{\Gamma_{ij}}(\nabla_{\theta_k'}q(x;\theta_k^0))^{\mathrm T}I_{kl}(\theta^0)\nabla_{\theta_l'}q(x;\theta_l^0)|\nabla_x F_{ij}(x)|^{-1}\,d\mathscr{S}_{N-1}$$

where $I_{kl}(\theta^0)$- is the (k,l)th $(m\times m)$-block of the matrix $I(\theta^0)=(I_{kl}(\theta^0))_{k,l\in S}$:

$$I(\theta)=V^{-1}(\theta)\int_{R^N}\nabla_\theta f(x;\pi^0,\theta)(\nabla_\theta f(x;\pi^0,\theta))^{\mathrm T}q_{\pi^0}(x;\theta^0)\,dx V^{-1}(\theta) \tag{33}$$

$$V(\theta)=\nabla_\theta^2 W(\pi^0,\theta)$$

For the case of two classes ($L=2$) and unknown π_1^0 let us introduce

$$\tilde\varkappa=\frac{1}{2}\int_{\Gamma_{12}}(\nabla_{\eta'}F(x;\eta^0))^{\mathrm T}\tilde I(\eta^0)\nabla_{\eta'}F(x;\eta^0)|\nabla_x F(x;\eta^0)|^{-1}\,d\mathscr{S}_{N-1} \tag{34}$$

where

$$F(x;\eta^{o}) = (1-\pi_1^{o})q(x;\theta_2^{o}) - \pi_1^{o}q(x;\theta_1^{o})$$

$$\tilde{I}(\eta^{o}) = \tilde{V}^{-1}(\eta^{o})\int_{R^N} \nabla_{\eta^{o}} f(x;\pi^{o},\theta^{o})(\nabla_{\eta^{o}} f(x;\pi^{o},\theta^{o}))^{T} q_{\pi^{o}}(x;\theta^{o})\,\mathrm{d}x\,\tilde{V}^{-1}(\eta^{o})$$

$$\tilde{V}(\eta^{o}) = \nabla_{\eta}^{2} W(\pi,\theta)|_{\eta=\eta^{o}}, \quad \eta^{o} = (\theta_1^{o\,T}:\theta_2^{o\,T}:\pi_1^{o})^{T}$$

Theorem 3

If under the regularity conditions (C_1)–(C_4) the following asymptotics takes place:

$$\rho_+ \to 0, \quad n \to +\infty, \tag{35}$$

then the estimators $\tilde{\theta}, \tilde{\pi}, \hat{\theta}$ from (30) and (31) are strongly consistent:

$$\tilde{\theta} \xrightarrow{P-1} \theta^{o}, \quad \tilde{\pi} \xrightarrow{P-1} \pi^{o}, \quad \hat{\theta} \xrightarrow{P-1} \theta^{o} \tag{36}$$

If in addition p.d.f. $q(x;\theta_*)$, $\theta_* \in \Theta$, is differentiable w.r.t. $x \in R^N$, the matrices $V(\theta^{o})$, $\tilde{V}(\eta^{o})$ are non-singular and surface integrals in (32) and (34) are finite, then the CEP functionals $r_A(n), r_B(n)$ of the PDR (30) and (31) have the form

$$r_A(n) = r_0 + \frac{1}{2}\sum_{j=2}^{L}\sum_{l=1}^{j-1}(B_{jjlj} + B_{llij} - 2B_{ljlj})n^{-1} + o(n^{-1}) \tag{37}$$

$$r_B(n) = r_0 + \tilde{\varkappa}n^{-1} + o(n^{-1}) \quad (L=2) \tag{38}$$

where the CEP r_0 of the BDR (4) is determined in (5).

Proof. The strong consistency of $\tilde{\theta}, \tilde{\pi}$ and $\hat{\theta}$ (expressions (36)) is proved by applying of the strong law of large numbers to (28):

$$W_n(\pi,\theta) \xrightarrow{P-1} W(\pi,\theta)$$

and the continuity theorems from Reference[11] under asymptotics (35). The asymptotic expansions (37) and (38) are constructed as in the proof of Theorem 1. □

Corollary

The 'plug-in' decision rules (30) and (31) are asymptotically optimal. If according to the conditions of Theorem 3,

$$b_{lj} = \inf_{x\in R^N}\left|\nabla_x \ln\frac{q(x;\theta_l^{o})}{q(x;\theta_j^{o})}\right| \geqslant r, \quad r = O(1), \ l \neq j \in S \tag{39}$$

then

$$r_k(n) = r_0 + o(n^{-1}), \quad k \in \{A,B\} \tag{40}$$

Proof. The asymptotic optimality of the PDR (30) and (31) follows from (37), (38) and (9). Under the conditions (C_1)–C_4 and (39) by means of Cauchy–Schwarz inequality we obtain

$$\max\{|B_{ijlj}|, B_{llij}, B_{jlli}\} < p_1 \rho_+^{1/2}, \quad l \neq j \in S$$

$$\hat{x} < p_2 \rho_-^{1/2}$$

where $p_1, p_2 < +\infty$ are some positive constants. From last inequalities and from the expressions (35), (36) and (38) we have (40). □

Note (see Reference 6) that the asymptotics (35) is of practical value, because under 'large overlapping of classes' (when (35) is disturbed) the value r_0 is large and all DRs are not advisable for applications.

5. FISHER MODEL AND ILLUSTRATION OF OBTAINED RESULTS

Let us illustrate the obtained results for the case with two equiprobable classes ($L = 2$): $\pi_1^0 = \pi_2^0 = 0{\cdot}5$, when the family (1) is Gaussian:

$$q(x; \theta_i^0) = n_N(x \mid \theta_i^0, \Sigma), \quad i \in S = \{1, 2\} \tag{41}$$

where

$$n_N(x, \theta_*, \Sigma) = (2\pi)^{-N/2} (\det(\Sigma))^{-0{\cdot}5} \exp(-0{\cdot}5(x - \theta_*)^{\mathrm{T}} \Sigma^{-1} (x - \theta_*))$$

is N-variate Gaussian p.d.f. with mathematical mean vector θ_* and non-singular covariance $(N \times N)$-matrix Σ ($\det(\Sigma) > 0$). Relation (41) defines the so-called Fisher[12] model.

Denote by

$$\Delta = \sqrt{(\theta_1^0 - \theta_2^0)^{\mathrm{T}} \Sigma^{-1} (\theta_1^0 - \theta_2^0)}$$

the Mahalanobis interclass distance. The CEP r_0 of the BDR (4) has the form

$$r_0 = \Phi\left(-\frac{\Delta}{2}\right)$$

where $\Phi(\cdot)$ is the standard Gaussian distribution function with the p.d.f.:

$$\varphi(z) = \Phi'(z) = (2\pi)^{-1/2} \exp(-z^2/2), \quad z \in R$$

For the PDR based on the MLE (13) and (22) by the results of Section 3 we evaluated main terms in the CEP asymptotic expansions (17) and (23) and determined the δ-admissible sample size (10):

$$n_A^*(\delta) = \left[\frac{1}{2\delta}\left(N + \frac{\Delta^2}{4} + \frac{\Delta^3}{8}\sqrt{\frac{\pi}{2}}\, e^{-\Delta^2/8}\right)\right] + 1 \tag{42}$$

$$n_B^*(\delta) = n_A^*(\delta) + \left[\frac{1}{\delta}\left(\frac{1}{2} + \frac{2}{\Delta^2} + \sqrt{\frac{\pi}{2}}\frac{1}{\Delta}\, e^{-\Delta^2/8}\right)\right] + 1$$

Table 1. Cluster analysis of Iris data

	Case A			Case B		
Sample size, n	Evaluated Δ	$\hat{r}_A$	γ_n	Evaluated Δ	$\hat{r}_B$	γ_n
10	1·713	0·248	0·200	1·692	0·285	0·300
20	1·913	0·193	0·150	1·889	0·209	0·200
30	1·909	0·186	0·167	1·872	0·199	0·200
40	1·882	0·185	0·175	1·882	0·192	0·175

where $[z]$ means the entire part of the value z. If $n = n_j^*(\delta)$, then for the CEP $r_j(n)$ the following approximation is true:

$$r_j(n) \approx \hat{r}_j, \quad \hat{r}_j = \hat{r}_j(n) = (1 + \delta)r_0, \quad j \in \{A, B\}. \tag{43}$$

In practice, this fact allows to estimate the risk (7) when the classification process has finished and when the Mahalanobis distance Δ has been evaluated. As a numerical example let us use the well known Fisher[12] Iris data. Four samples ($n = 10; 20; 30; 40; N = 4$) from two ($L = 2$) equiprobable classes (Iris versicolor, Iris virginica) were used. The experimental results for the case A (prior class probabilities are assumed to be known: $\pi_1^0 = \pi_2^0 = 0.5$) and for the case B ($\{\pi_i^0\}_{i \in S}$ are assumed to be unknown) are presented in Table 1.

Here γ_n is the experimental frequency of error decisions:

$$\gamma_n = \frac{1}{n}\sum_{t=1}^{n}(1 - \delta_{\hat{d}_t, d_t^0}),$$

Note, in real situation, when the true classification vector $D^0 = (d_1^0, \ldots, d_n^0)^T$ is unknown, as the CEP estimates we recommend to use $\hat{r}_j, j \in \{A, B\}$.

Now let us illustrate the results obtained in Section 4 for the PDR based on the likelihood function. Suppose that the Mahalanobis distance Δ is increasing: $\Delta \to +\infty$, $n \to +\infty$, and the asymptotics (35) takes place (because of $\rho_- = \exp(-\Delta^2/8)$ under Fisher model (41)). All conditions of Corollary of Theorem 3 are satisfied:

$$b_{12} = \sqrt{(\theta_2^0 - \theta_1^0)^T \Sigma^{-1} \Sigma^{-1} (\theta_2^0 - \theta_1^0)} \geqslant v, \quad v = O(1)$$

and according to (40) the CEP $r_A(n)$ and $r_B(n)$ of the PDR (30) and (31) have no terms of order $O(n^{-1})$ in their asymptotic expansions:

$$r_j(n) = r_0 + o(n^{-1}), \quad r_0 = \Phi\left(-\frac{\Delta}{2}\right), \quad j \in \{A, B\}$$

ACKNOWLEDGEMENTS

These investigations were partially supported by Belarussian National Grants F40-263, MP94-03.

REFERENCES

1. N. Glick, 'Additive estimators for probabilities of correct classification', *Pattern Recognition*, **1**, 211–222 (1978).
2. P. Lachenbruch, *Discriminant Analysis*, Hafner Press, New York, 1975.

3. Sh. Raudys, 'Small sample effects in classification problems', in *Statistical Problems of Control*, Vol. 18, IMC, Vilnius, 1976 (in Russian).
4. G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York, 1992.
5. H. H. Bock, 'Probabilistic aspects in cluster analysis', *Proc. 13th Conf. of the Gesellschaft für Klassifikation*, Springer, Berlin, 1989.
6. Yu. S. Kharin, *Robustness in Statistical Pattern Recognition*, Kluwer Academic Publishers, Dordrecht, 1996.
7. Yu. S. Kharin and E. E. Zhuk, 'Robustness in statistical pattern recognition under "contaminations" of training samples', in *Proc. 12th IAPR Int. Conf. on Pattern Recognition*, Vol. 2, IEEE Comp. Press, Washington, DC, 1994.
8. J. E. Dennis, Jr., and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood cliffs, NJ, 1983.
9. L. Redner and J. Walker, 'Mixture densities, maximum likelihood and the EM algorithm', *SIAM Rev.* **26** (2), (1984).
10. D. M. Chibisov, 'Asymptotic expansion for some estimate family including *ML*-estimates', *Teor. Ver. Prim.*, **18** (4), 689–701 (1973).
11. A. A. Borovkov, *Mathematical Statistics: Parameters Estimation and Hypotheses Testing*, Nauka, Moscow, 1984 (in Russian).
12. R. A. Fisher, 'The use of multiple measurements in taxonomic problems', *Ann. Eugen.*, **7**, 179–188 (1936).