GMDH-Based Criterion for Optimal Set Features Determination in Nonlinear Discriminant Analysis

Alexander Sarychev [1], **Lyudmyla Sarycheva** [2]

[1] *Institute of Technical Mechanics of the National Academy of Sciences of Ukraine, 15 Leshko-Popel St., Dnipropetrovsk, 49600, Ukraine*
[2] *National Mining University of Ukraine, 19 K. Marks Ave., Dnipropetrovsk, 49027, Ukraine*

Sarychev@prognoz.dp.ua, Sarycheval@nmu.org.ua

Abstract. *The problem of construction and analytical investigation of criterion of nonlinear discriminant functions quality under conditions of unequal covariance matrixes is formulated. The existence of optimum set of features is proved. Conditions of reduction (simplification) of discriminant function which is optimal by structure of features are revealed. It is obtained that this conditions depend on volumes samples and parameters of general sets, i.e. on mathematical expectations and covariance matrices of features for two classes.*

Keywords

Group Method of Data Handling, criterion of nonlinear discriminant functions quality, optimal set of features.

1 Introduction

The problem of construction of a discriminant function under conditions of structural uncertainty in terms of quantity and structure of features is one of objects of researches in a Group Method of Data Handling (GMDH) [1, 2]. As it is known, in GMDH the comparison of quality of models is based on partition of the initial data sample on learning and checking parts: on the learning sample the coefficients of model are estimated, and on checking the quality of model is estimated. This way is popular in applied tasks of the discriminant analysis. Existence of optimum set of features in this way is repeatedly confirmed with a method of statistical tests, but its analytical substantiation is received only in GMDH for a class of linear discriminant functions [1, 2].

However, in tasks of statistical classification the assumption of the linear discriminant analysis of equality covariance matrixes of two classes of observations frequently is not carried out. For example, in technical diagnostics usually consider two classes of states which essentially differ from each other. This distinction is shown not only in an inequality of population means of features, but also in an inequality covariance matrixes of classes, i.e. in distinction of dispersions of features and statistical connections between features in classes.

Because construction and analytical investigation of criterion of nonlinear discriminant functions quality under conditions of unequal covariance matrixes is actual task.

2 Statement of the Problem

Suppose that

$$\mathbf{X}_k = [\mathbf{x}_{k1}, \mathbf{x}_{k2}, \dots, \mathbf{x}_{kn_k}] \quad (1)$$

is a sample of n_k independent observations of m-dimensional random vector $\boldsymbol{\eta}_k$ with a m- dimensional normal distribution with unknown mathematical expectation $\boldsymbol{\chi}_k$ and unknown non-singular symmetric covariance matrix $\boldsymbol{\Sigma}_X$:

$$\boldsymbol{\eta}_k \sim N_m(\boldsymbol{\chi}_k, \boldsymbol{\Sigma}_X), \quad (2)$$

where k is number I or II of general set P_I, P_{II}, and $\boldsymbol{\chi}_I \neq \boldsymbol{\chi}_{II}$.

By the assumption (2) for the observations (1) it is carried out

$$\mathbf{X}_{ki} = \boldsymbol{\chi}_k + \boldsymbol{\xi}_{ki}, \quad k = I, II, \quad i = 1, 2, ..., n_k, \tag{3}$$

where $\boldsymbol{\xi}_{ki} \sim N_m(\mathbf{0}_m, \boldsymbol{\Sigma}_X)$ are independent random vectors distributed according to m-dimensional normal law with zero mathematical expectation and unknown covariance matrix $\boldsymbol{\Sigma}_X$:

$$E\{\boldsymbol{\xi}_{ki}\} = \mathbf{0}_m; \quad E\{\boldsymbol{\xi}_{ki}\boldsymbol{\xi}_{ki}^{\mathrm{T}}\} = \boldsymbol{\Sigma}_X; \quad i = 1, 2, ..., n_k; \quad k = I, II; \tag{4}$$

$$E\{\boldsymbol{\xi}_{ki_1}\boldsymbol{\xi}_{ki_2}^{\mathrm{T}}\} = \mathbf{O}_{(m\times m)}; \quad i_1, i_2 = 1, 2, ..., n_k; \quad i_1 \neq i_2; \tag{5}$$

$$E\{\boldsymbol{\xi}_{Ii_1}\boldsymbol{\xi}_{IIi_2}^{\mathrm{T}}\} = \mathbf{O}_{(m\times m)}; \quad i_1, i_2 = 1, 2, ..., n_k, \tag{6}$$

where $\mathbf{0}_m$ is an m-dimensional zero column vector and $\mathbf{O}_{(m\times m)}$ is an $m\times m$ zero matrix, $E\{\cdot\}$ is a symbol of mathematical expectation.

Suppose that the *a priori* probabilities of appearance of a new observations in the general set P_I and P_{II} are known and equal respectively to π_I and π_{II}, with $\pi_I + \pi_{II} = 1$. Suppose that costs of erroneous classification has been introduced: $c(I/II)$ denotes the cost of erroneous classification of an observation P_{II} as an observation in P_I, while $c(II/I)$ denotes the cost of erroneous classification of an observation in P_I as an observation in P_{II}. A correct classification is not estimated.

The linear discriminant function minimizing, under specified assumptions, the expected expenditures due to erroneous classification has the form

$$R(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\boldsymbol{\delta} - 0{,}5(\boldsymbol{\chi}_I + \boldsymbol{\chi}_{II})^{\mathrm{T}}\boldsymbol{\delta} - \ln c_0, \tag{7}$$

where parameter c_0 is determined by costs of erroneous classifications and a prior probabilities of appear observations

$$c_0 = c(I/II)\pi_{II}(c(II/I)\pi_I)^{-1}, \tag{8}$$

and coefficients of discriminant function $\boldsymbol{\delta}$ is determined by parameters of general sets

$$\boldsymbol{\delta} = \boldsymbol{\Sigma}_X^{-1}(\boldsymbol{\chi}_{\mathrm{I}} - \boldsymbol{\chi}_{\mathrm{II}}). \tag{9}$$

Decision rule has form: if $R(\mathbf{x}^*) \geq 0$ is satisfied, then $\mathbf{x}^* \in P_I$ is valid, and if $R(\mathbf{x}^*) < 0$, then $\mathbf{x}^* \in P_{II}$.

The obtaining of estimates from sample observations (1)–(3) of the coefficient $\boldsymbol{\delta}$ in the discriminant function (7) with statistical analysis is a problem of discriminant analysis *in the narrow sense*.

Fisher's estimates of $\boldsymbol{\delta}$ is

$$\hat{\mathbf{d}} = \mathbf{S}^{-1}(\tilde{\mathbf{x}}_{\mathrm{I}} - \tilde{\mathbf{x}}_{\mathrm{II}}), \tag{10}$$

where s-dimensional column vectors $\tilde{\mathbf{x}}_{\mathrm{I}}$ and $\tilde{\mathbf{x}}_{\mathrm{II}}$ are estimates of the mathematical expectation $\boldsymbol{\chi}_{\mathrm{I}}$ и $\boldsymbol{\chi}_{\mathrm{II}}$:

$$\tilde{\mathbf{x}}_{\mathrm{I}} = (n_{\mathrm{I}})^{-1}\sum_{i=1}^{n_{\mathrm{I}}} \mathbf{X}_{\mathrm{I}i}, \quad \tilde{\mathbf{x}}_{\mathrm{II}} = (n_{\mathrm{II}})^{-1}\sum_{i=1}^{n_{\mathrm{II}}} \mathbf{X}_{\mathrm{II}i}; \tag{11}$$

the $m\times m$ matrix $\mathbf{S}$ is an unbiased estimate of the covariance matrix $\boldsymbol{\Sigma}$:

$$\mathbf{S} = (n_{\mathrm{I}} + n_{\mathrm{II}} - 2)^{-1}[\mathbf{x}_{\mathrm{I}}\mathbf{x}_{\mathrm{I}}^{\mathrm{T}} + \mathbf{x}_{\mathrm{II}}\mathbf{x}_{\mathrm{II}}^{\mathrm{T}}], \tag{12}$$

where $m\times n_{\mathrm{I}}$ and $m\times n_{\mathrm{II}}$ matrix $\mathbf{x}_{\mathrm{I}}$ and $\mathbf{x}_{\mathrm{II}}$ consist of variances observations (1) from estimates $\tilde{\mathbf{x}}_{\mathrm{I}}$ and $\tilde{\mathbf{x}}_{\mathrm{II}}$:

$$\mathbf{x}_{\mathrm{I}} = [\mathbf{X}_{\mathrm{I},1} - \tilde{\mathbf{x}}_{\mathrm{I}}, \mathbf{X}_{\mathrm{I},2} - \tilde{\mathbf{x}}_{\mathrm{I}}, ..., \mathbf{X}_{\mathrm{I},n_{\mathrm{I}}} - \tilde{\mathbf{x}}_{\mathrm{I}}], \tag{13}$$

$$\mathbf{x}_{\mathrm{II}} = [\mathbf{X}_{\mathrm{II},1} - \tilde{\mathbf{x}}_{\mathrm{II}}, \mathbf{X}_{\mathrm{II},2} - \tilde{\mathbf{x}}_{\mathrm{II}}, ..., \mathbf{X}_{\mathrm{II},n_{\mathrm{II}}} - \tilde{\mathbf{x}}_{\mathrm{II}}]. \tag{14}$$

Fisher's estimate (10) is a solution of the maximum determination problem of functional $\Phi(\mathbf{d})$:

$$\Phi(\mathbf{d}) = \frac{\mathbf{d}^{\mathrm{T}}(\tilde{\mathbf{x}}_{\mathrm{I}} - \tilde{\mathbf{x}}_{\mathrm{II}})(\tilde{\mathbf{x}}_{\mathrm{I}} - \tilde{\mathbf{x}}_{\mathrm{II}})^{\mathrm{T}}\mathbf{d}}{\mathbf{d}^{\mathrm{T}}\mathbf{S}\mathbf{d}} \to \max_{\mathbf{d}}, \tag{15}$$

with a restriction

$$\frac{(\tilde{\mathbf{x}}_{\mathrm{I}} - \tilde{\mathbf{x}}_{\mathrm{II}})^{\mathrm{T}}\mathbf{d}}{\mathbf{d}^{\mathrm{T}}\mathbf{S}\mathbf{d}} = 1. \tag{16}$$

The value of the functional (15) under optimal value $\hat{\mathbf{d}} = \mathbf{S}^{-1}(\tilde{\mathbf{x}}_{\mathrm{I}} - \tilde{\mathbf{x}}_{\mathrm{II}})$

$$D^2 = (\tilde{\mathbf{x}}_{\mathrm{I}} - \tilde{\mathbf{x}}_{\mathrm{II}})^{\mathrm{T}} \mathbf{S}^{-1} (\tilde{\mathbf{x}}_{\mathrm{I}} - \tilde{\mathbf{x}}_{\mathrm{II}}) \tag{17}$$

is sample estimate of the Mahalanobis distance

$$\tau_X^2 = (\boldsymbol{\chi}_{\mathrm{I}} - \boldsymbol{\chi}_{\mathrm{II}})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\chi}_{\mathrm{I}} - \boldsymbol{\chi}_{\mathrm{II}}). \tag{18}$$

For mathematical expectation D^2 it is carried out

$$E\{D^2\} = \frac{r}{r-m-1}\tau_X^2 + mc^{-1}, \tag{19}$$

where $r = n_{\mathrm{I}} + n_{\mathrm{II}} - 2$; $c^{-1} = (n_{\mathrm{I}}^{-1} + n_{\mathrm{II}}^{-1})$.

If we don't know in advance just which components in the set X should be included in the discriminant function, we speak of the problem of discriminant analysis *in the wide sense*. It is known that estimate for Mahalanobis distance (18) cannot be used for solving the problem of discriminant analysis, stated in a wide sense, since this estimate increases on addition of new components to discriminant function (into decision rule). In this case setting of "threshold" for the value of increment of estimate of the Mahalanobis distance (18) on addition of a new component into the discriminant function is prevailing method. However such setting inserts subjectivity into the obtained solution.

3 Criterion of Discrimination Function Quality under Conditions Unequal Covariance Matrix with Dividing Observations in to Learning and Checking Samples

Let's the task of discriminant analysis with unequal covariance matrix is stated in the broad sense. The statement this task coincide with statement task of the linear discriminant analysis (1)–(19) under conditions of structural uncertainty in terms of quantity and structure of features but instead common covariance matrix $\boldsymbol{\Sigma}_X$ we have two different matrices $\boldsymbol{\Sigma}_X(I)$ and $\boldsymbol{\Sigma}_X(II)$. Decision rule in this task has form [3]:

$$h(\mathbf{x}^*) = \frac{1}{2}(\mathbf{x}^* - \tilde{\mathbf{x}}(I))^{\mathrm{T}} \mathbf{S}^{-1}(I)(\mathbf{x}^* - \tilde{\mathbf{x}}(I)) - \frac{1}{2}(\mathbf{x}^* - \tilde{\mathbf{x}}(II))^{\mathrm{T}} \mathbf{S}^{-1}(II)(\mathbf{x}^* - \tilde{\mathbf{x}}(II)) +$$
$$+ \frac{1}{2}\ln\left(\frac{\det(\mathbf{S}(I))}{\det(\mathbf{S}(II))}\right) - \ln\frac{c(II/I)\pi(I)}{c(I/II)\pi(II)}, \tag{20}$$
$$\text{if } h(\mathbf{x}^*) < 0 \text{ is satisfied, then } \mathbf{x}^* \in P_I \text{ is valid,}$$
$$\text{if } R(\mathbf{x}^*) > 0 \text{ is satisfied, then } \mathbf{x}^* \in P_{II} \text{ is valid.}$$

In (20) $\mathbf{x}^*$ is analyzed observations; vectors $\tilde{\mathbf{x}}(I)$ and $\tilde{\mathbf{x}}(II)$ are estimations (11) of mathematical expectation $\boldsymbol{\chi}(I)$ and $\boldsymbol{\chi}(II)$; matrices $\mathbf{S}(I)$ and $\mathbf{S}(II)$ are covariance matrices $\boldsymbol{\Sigma}_X(I)$ and $\boldsymbol{\Sigma}_X(II)$:

$$\mathbf{S}(I) = (n(I)-1)^{-1}[\mathbf{x}(I)\mathbf{x}^{\mathrm{T}}(I)], \quad \mathbf{S}(II) = (n(II)-1)^{-1}[\mathbf{x}(II)\mathbf{x}^{\mathrm{T}}(II)], \tag{21}$$

where $\mathbf{x}(I)$ and $\mathbf{x}(II)$ are $m \times n(I)$- and $m \times n(II)$-matrices calculated by (13)–(14).

Besides comparison scheme the solution of this task requires to choose the set features generation method. Suppose that as such method is chosen the complete sorting-out of all possible combinations of features. Suppose that at the s-th step s ($s = 1, 2, \ldots, m$) of complete sorting-out of all possible sets of features only s components from the set X (which constitute the current set V being analyzed) can be included in the discriminant function. In the following are supposed $\mathbf{V}_{\mathrm{I}}$ a $s \times n(I)$ matrix of observations in set $P(I)$, $\mathbf{V}_{\mathrm{II}}$ a $s \times n(II)$ matrix of observations in set $P(II)$, $\mathbf{v}(I)$ a s-dimensional column vector of the mathematical expectations of the observations in $P(I)$, $\mathbf{v}(II)$ a s-dimensional column vector of the mathematical expectations of the observations in $P(II)$, $\mathbf{S}(I,V)$ a covariance $s \times s$ matrix of the observations in $P(I)$ and $\mathbf{S}(II,V)$ a covariance $s \times s$ matrix of the observations in $P(II)$.

Let realize scheme of calculation of classification numbers with partition of the initial data sample on learning and checking subsamples. Let use the estimates of mathematical expectation and covariance matrices which are obtain on learning subsamples A for calculation of classification numbers in space of features V for observations from checking subsamples B of first and second classes by (20):

$$h_{iB}(k,V) = h(\mathbf{v}_{iB}(k)) = \frac{1}{2}(\mathbf{v}_{iB}(k) - \tilde{\mathbf{v}}_A(I))^{\mathrm{T}} \mathbf{S}_A^{-1}(I,V)(\mathbf{v}_{iB}(k) - \tilde{\mathbf{v}}_A(I)) -$$

$$-\frac{1}{2}(\mathbf{v}_{iB}(k)-\tilde{\mathbf{v}}_A(II))^{\mathrm{T}}\mathbf{S}_A^{-1}(II,V)(\mathbf{v}_{iB}(k)-\tilde{\mathbf{v}}_A(II))+ \tag{22}$$

$$+\frac{1}{2}\ln\left(\frac{\det(\mathbf{S}_A(I,V))}{\det(\mathbf{S}_A(II,V))}\right)-\ln\frac{c(II/I)\pi(I)}{c(I/II)\pi(II)},\ i=1,2,...,n_B(k),\quad (k=I,\ II).$$

Let calculate average of this classification numbers for first and second classes and investigated a difference

$$D_{BA}(V)=\bar{h}_B(II,V)-\bar{h}_B(I,V) \tag{23}$$

as criterion of discriminant function quality.

In accordance with principles of modeling MGDH, to prove the adequacy of a criterion it is necessary: 1) to calculate mathematical expectation of researched criterion for given model structure; 2) to research behavior of the mathematical expectation of this criterion depending on structure of models; 3) to prove existence of model of optimum complexity; 4) to obtain of the reduction (simplification) condition for the optimum complexity model.

For random variable $D_{BA}(V)$ it is carry out

$$E\{D_{BA}(V)\}=E\{\bar{h}_B(II,V)-\bar{h}_B(I,V)\}=$$

$$=l_A(I,s)\ \tau_V^2(I)+l_A(II,s)\ \tau_V^2(II)+$$

$$+l_A(I,s)\ \mathrm{tr}[\mathbf{\Sigma}^{-1}(I)\mathbf{\Sigma}(II)]+l_A(II,s)\ \mathrm{tr}[\mathbf{\Sigma}^{-1}(II)\mathbf{\Sigma}(I)]-l_A(I,s)\ s-l_A(II,s)\ s, \tag{24}$$

where

$$\tau_V^2(I)=(\mathbf{v}(I)-\mathbf{v}(II))^{\mathrm{T}}\mathbf{\Sigma}_V^{-1}(I)(\mathbf{v}(I)-\mathbf{v}(II)) \tag{25}$$

is the Mahalanobis distance which must be between general sets $P(I)$ and $P(II)$ in space of features V if $\mathbf{\Sigma}_V(II)=\mathbf{\Sigma}_V(I)$;

$$\tau_V^2(II)=(\mathbf{v}(I)-\mathbf{v}(II))^{\mathrm{T}}\mathbf{\Sigma}_V^{-1}(II)(\mathbf{v}(I)-\mathbf{v}(II)) \tag{26}$$

is the Mahalanobis distance which must be between general sets $P(I)$ and $P(II)$ in space of features V if $\mathbf{\Sigma}_V(I)=\mathbf{\Sigma}_V(II)$;

$$l_A(I,s)=\frac{n_A(I)-1}{n_A(I)-s-2},\quad l_A(II,s)=\frac{n_A(II)-1}{n_A(II)-s-2}. \tag{27}$$

In paper dependence of a mathematical expectation (24) from structure of set of features V is investigated. It is shown, that as against a linear case in the given nonlinear case features at which a dispersion strongly enough differ in the first and second classes, should be included in discriminant function even if they do not differ on a mathematical expectation and are statistically independent.

4 Conclusion

The problem of construction and analytical investigation of criterion of nonlinear discriminant functions quality under conditions of unequal covariance matrixes is formulated. Such criterion is constructed. The existence of optimum set of features is proved. Conditions of reduction (simplification) of discriminant function which is optimal by structure of features are revealed. It is obtained that this conditions depend on volumes samples and parameters of general sets, i.e. on mathematical expectations and covariance matrices of features for two classes.

References

[1] Sarychev A. P.: The Solution of the Discriminant Analysis Task in Conditions of Structural Uncertainty on Basis of the Group Method of Data Handling. Problemy Upravlenia i Informatiki (Ukraine), 3: 100–112, 2008 (Journal of Automation and Information Sciences, Vol. 40, 6: 27–40, 2008).

[2] Sarychev A. P.: Identification of Structural-Uncertain Systems States (Ukraine). Dnipropetrovs'k, NAS Ukraine and NSA Ukraine, Institute of Technical Mechanics, 268, 2008.

[3] Fukunaga K.: Introduction to statistical theory of pattern recognition [Russian translation]. Nauka, Moscow, 368, 1979.