DOI 10.34185/1562-9945-6-125-2019-03
UDC 519.25

A.P. Sarychev, L.V. Sarycheva

GMDH-BASED OPTIMAL SET FEATURES DETERMINATION IN DISCRIMINANT ANALYSIS

Anatation. The task of searching optimum on complexity discriminant function is considered. Criteria of quality of the discriminant functions developed in the Group Method of Data Handling are described: the criterion based on a partition of observations on training and testing samples, and criterion of sliding examination. The tasks of this class belong to pattern recognition problems under the condition of structural uncertainty, which were considered by academician A.G. Ivakhnenko as long ago as 60–70-th of the last century as actual problems of an engineering cybernetics.

Introduction. The decision of task of the discriminant analysis in conditions of structural uncertainty on structure of features assumes acceptance of any way of comparison of discriminant functions, which are constructed on various sets of features. Two ways of comparison are popular in practice. The first way is based on dividing of observations on training and testing subsamples. In this way training subsamples are used for estimation coefficients of discriminant functions, and testing subsamples are used for estimation its qualities of classification. The second way is sliding examination. In this way, observations, which are serially excluded from training subsamples, are used as testing observations. In the literature, these ways are traditionally treated as heuristic methods though the fact of existence in them of optimum set of features repeatedly proved by a method of statistical tests. In the Group Method of Data Handling (GMDH), analytical research of these two ways is carried out [1-6]. For the decision of a task of the discriminant analysis in conditions of structural uncertainty except for a way of comparison discriminant functions it is required to specify algorithm of generation of various combinations of the features included in discriminant functions. Algorithms, which are based on principles GMDH, are developed [7-8]. It is supposed, that as such method is chosen the complete sorting-out of all possible combinations of features.

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

The method of comparison of discriminant functions based on training and testing sample. Suppose that at the step with number s $(s = 1, 2, \ldots, m)$ of algorithm complete sorting-out of all possible sets of features only s components from the set X can be included in the discriminant function and these features form the current set V. In the following we suppose that $\mathbf{V}_{\mathrm{I}}$ and $\mathbf{V}_{\mathrm{II}}$ are $(s \times n_{\mathrm{I}})$- and $(s \times n_{\mathrm{II}})$-matrices of observations from general sets P_{I} and P_{II}, $\mathbf{v}_{\mathrm{I}}$ and $\mathbf{v}_{\mathrm{II}}$ are s-dimensional column vectors of the mathematical expectations in the sets P_{I} and P_{II}, $\mathbf{\Sigma}_V$ is covariance $(s \times s)$ matrix of the sets P_{I} and P_{II}.

Let's consider the estimation of Mahalanobis distance that is constructed with account of dividing of observations on training and testing subsamples. We shall calculate estimations of coefficients discriminant function for set component V on training subsample A and it is used them for estimation Mahalanobis distances as the relation of an intergroup variation to an intragroup variation on testing subsample B

$$D_{AB}^2(V) = \frac{\hat{\mathbf{d}}_A^{\mathrm{T}} (\tilde{\mathbf{v}}_{\mathrm{I}B} - \tilde{\mathbf{v}}_{\mathrm{II}B}) (\tilde{\mathbf{v}}_{\mathrm{I}B} - \tilde{\mathbf{v}}_{\mathrm{II}B})^{\mathrm{T}} \hat{\mathbf{d}}_A}{\hat{\mathbf{d}}_A^{\mathrm{T}} \mathbf{S}_B \hat{\mathbf{d}}_A}. \qquad (1)$$

In formula (1), vector $\hat{\mathbf{d}}_A$ is an estimate of the coefficients of the Fisher function that is calculate on training subsamples A

$$\hat{\mathbf{d}}_A = \mathbf{S}_A^{-1} (\tilde{\mathbf{v}}_{\mathrm{I}A} - \tilde{\mathbf{v}}_{\mathrm{II}A}), \qquad (2)$$

where vectors $\tilde{\mathbf{v}}_{\mathrm{I}A}$ and $\tilde{\mathbf{v}}_{\mathrm{II}A}$ are estimate of the mathematical expectation $\mathbf{v}_{\mathrm{I}}$ and $\mathbf{v}_{\mathrm{II}}$

$$\tilde{\mathbf{v}}_{kA} = (n_{kA})^{-1} \sum_{i=1}^{n_{kA}} \mathbf{V}_{kiA}, \quad k = \mathrm{I}, \mathrm{II}\,; \tag{3}$$

the matrix $\mathbf{S}_A$ is an unbiased estimate of covariance matrix $\mathbf{\Sigma}_V$

$$\mathbf{S}_A = (n_{\mathrm{I}A} - n_{\mathrm{II}A} - 2)^{-1}[\mathbf{v}_{\mathrm{I}A}\mathbf{v}_{\mathrm{I}A}^{\mathrm{T}} + \mathbf{v}_{\mathrm{II}A}\mathbf{v}_{\mathrm{II}A}^{\mathrm{T}}], \tag{4}$$

where $\mathbf{v}_{kA}$ are matrices of deviations of observations $\mathbf{V}_{kA}$ from estimates $\tilde{\mathbf{v}}_{kA}$

$$\mathbf{v}_{kA} = [\mathbf{V}_{k1A} - \tilde{\mathbf{v}}_{kA}, \mathbf{V}_{k2A} - \tilde{\mathbf{v}}_{kA}, \ldots, \mathbf{V}_{kn_kA} - \tilde{\mathbf{v}}_{kA}]. \tag{5}$$

In formula (5) vectors $\tilde{\mathbf{v}}_{\mathrm{I}B}$ and $\tilde{\mathbf{v}}_{\mathrm{II}B}$ calculated analogues (3), and matrix $\mathbf{S}_B$ calculated analogues (4)–(5); $n_{\mathrm{I}A}$ and $n_{\mathrm{II}A}$, $n_{\mathrm{I}B}$ and $n_{\mathrm{II}B}$ are volume of training and testing subsamples respectively, and it is true $n_{\mathrm{I}A} + n_{\mathrm{I}B} = n_{\mathrm{I}}$ and $n_{\mathrm{II}A} + n_{\mathrm{II}B} = n_{\mathrm{II}}$. Using (2), we obtain for $D_{AB}^2(V)$

$$D_{AB}^2(V) == \frac{\left((\tilde{\mathbf{v}}_{\mathrm{I}A} - \tilde{\mathbf{v}}_{\mathrm{II}A})^{\mathrm{T}} \mathbf{S}_A^{-1} (\tilde{\mathbf{v}}_{\mathrm{I}B} - \tilde{\mathbf{v}}_{\mathrm{II}B})\right)^2}{(\tilde{\mathbf{v}}_{\mathrm{I}A} - \tilde{\mathbf{v}}_{\mathrm{II}A})^{\mathrm{T}} \mathbf{S}_A^{-1} \mathbf{S}_B \mathbf{S}_A^{-1} (\tilde{\mathbf{v}}_{\mathrm{I}A} - \tilde{\mathbf{v}}_{\mathrm{II}A})}. \tag{6}$$

Let $\tau_V^2 = (\mathbf{v}_{\mathrm{I}} - \mathbf{v}_{\mathrm{II}})^{\mathrm{T}} \mathbf{\Sigma}_V^{-1} (\mathbf{v}_{\mathrm{I}} - \mathbf{v}_{\mathrm{II}})$ be the Mahalanobis distance for the set V, $r = n_{\mathrm{I}A} + n_{\mathrm{II}A} - 2$, $c_A^{-1} = (n_{\mathrm{I}A}^{-1} + n_{\mathrm{II}A}^{-1})$, $c_B^{-1} = (n_{\mathrm{I}B}^{-1} + n_{\mathrm{II}B}^{-1})$.

Theorem 1. For mathematical expectation of random variable $D_{AB}^2(V)$, we have

$$E\{D_{AB}^2(V)\} = \left(\tau_V^2 - \frac{\tau_V^2\,[s - (r-1)/(r-s)]\, c_A^{-1}}{\tau_V^2 + s c_A^{-1}} + c_B^{-1}\,\frac{r-1}{r-s}\right) \cdot \frac{r-s}{r-1}. \tag{7}$$

The validity of theorem follows from the validity of the following: 1) the estimates obtained on subsamples A and B are independent; 2) the estimate (3) and estimate (4) are independent; 3) matrix $\mathbf{S}_A$ is random $(s \times s)$-matrix which has the Wishart distribution with r degrees of freedom.

Definition 1. The optimal set components (set features) is defined as the set V_{OPT} for which

$$V_{OPT} = \arg\max_{V \subseteq X} E\{D_{AB}^2(V)\}. \tag{8}$$

Definition 2. Optimal discriminant function with respect to the number and composition of the components is defined as the Fisher discriminant function constructed on the set of components V_{OPT}.

We proved that optimal set of components exist and formulated the conditions under which the optimal discriminant function is simplified in number of the features included in it. For this purpose, it was investigated $E\{D^2_{AB}(V)\}$ depending on composition of set V.

It is possible to divide set of components X into the following nonintersecting subsets $X = \overset{o}{X} \cup \overset{o}{R} \cup \widetilde{R} = \overset{o}{V} \cup \widetilde{R}$: so that 1) $\overset{o}{X} \neq \varnothing$ (where $\varnothing$ is the empty set) is the set of components whose mathematical expectation satisfy $\overset{o}{\chi}_{\mathrm{I}h} \neq \overset{o}{\chi}_{\mathrm{II}h}$, $h = 1,2,..., \overset{o}{m}$, where $\overset{o}{m}$ is their number; 2) $\overset{o}{R}$ is the set of components whose mathematical expectation satisfy $\overset{o}{\rho}_{\mathrm{I}h} = \overset{o}{\rho}_{\mathrm{II}h}, h = 1,2,..., \overset{o}{l}$, where $\overset{o}{l}$ is their number and each component in $\overset{o}{R}$ depends statistically on the least one components in the set $\overset{o}{X}$ (the set $\overset{o}{R}$ may be empty); 3) $\widetilde{R}$ is the set of components whose mathematical expectation satisfy $\widetilde{\rho}_{Ih} = \widetilde{\rho}_{IIh}$, $h = 1, 2,..., \widetilde{l}$, where $\widetilde{l}$ is number and each component each component in $\widetilde{R}$ is statistically independent from each set $\overset{o}{X}$ (the set $\widetilde{R}$ may be empty). Relationship between the Mahalanobis distance for the set components $\overset{o}{V} = \overset{o}{X} \cup \overset{o}{R}$ and the Mahalanobis distance for a current analyzed set of components $V \subseteq X$ is formulated in the form of lemmas [1-4].

In case of known parameters of general sets P_{I} and P_{II} it follows from the stated lemmas that: 1) every component from set $\overset{o}{X}$ is necessary in the sense that its inclusion into the current set of components V increase the Mahalanobis distance τ^2_V; 2) every component from the set $\overset{o}{R}$ is necessary in the sense that its inclusion into the current set of components V increase the Mahalanobis distance τ^2_V; 3) every components from the set $\widetilde{R}$ is redundant in the sense, that its inclusion into the current set V does not increase the Mahalanobis distance τ^2_V.

Reduction (simplification) condition of optimal discriminant function. As a rule, in practical applications, parameters of general populations are unknown. However, they can be estimated as statistical estimates on training samples of observations of limited volume. It is known, that if we use constructed rule of classification to the training sample, then estimate of recognition quality will be overstated by mathematical expectation in comparison with the same evaluation of quality on data, independent of training data.

The way for comparison of the discriminant functions based on dividing of the initial data sample on training and testing subsamples give not overstated estimates of recognition quality. Experience of practical applications and test investigations of this way on basis of method of statistical test show that in this way: 1) on increase of size of observations samples increases the number of components in the set, on which the best quality of recognition is attained, and on decrease of size of observations samples the number of components in such set decreases; 2) on increase of the Mahalanobis distance τ_X^2 between general populations (from which observation samples were obtain) the number of components increases in the set, on which the best quality of recognitions is attained, and on decrease of this distance the number of components in such set decreases.

Our analytical investigations confirm these empirically determined regularities about the existence of the discriminant function optimal by the number and composition of components. Let's formulate the conditions of reduction (simplification) optimal discriminant function for a special case of an independent feature. Let the set of V is those, that is carried out $\overset{o}{X} = V \cup \overset{o}{x}$, where $\overset{o}{x} \in \overset{o}{X}$ (one feature is missed). Taking into account (7), we receive

$$\Delta(V) = E\{D_{AB}^2(\overset{o}{X})\} - E\{D_{AB}^2(V)\} =$$

$$= \left(\tau_{\overset{o}{X}}^2 - \frac{\tau_{\overset{o}{X}}^2 \cdot [\overset{o}{m} - (r-1)/(r-\overset{o}{m})] \cdot c_A^{-1}}{\tau_{\overset{o}{X}}^2 + \overset{o}{m} \cdot c_A^{-1}} + c_B^{-1} \cdot \frac{r-1}{r-\overset{o}{m}} \right) \cdot \frac{r-\overset{o}{m}}{r-1} -$$

$$\left(\tau_V^2 - \frac{\tau_V^2\cdot[(\overset{\circ}{m}-1)-(r-1)/(r-\overset{\circ}{m}+1)]\cdot c_A^{-1}}{\tau_V^2+(\overset{\circ}{m}-1)\cdot c_A^{-1}} + c_B^{-1}\cdot\frac{r-1}{r-\overset{\circ}{m}+1}\right)\cdot\frac{r-\overset{\circ}{m}+1}{r-1}. \qquad (9)$$

According to the above-mentioned lemmas for Mahalanobis distances of sets V and $\overset{\circ}{X}$ the ratio $\tau_V^2 = \tau_{\overset{\circ}{X}}^2 - \gamma^2$ is carried out, where $\gamma^2 = \sigma_{\overset{\circ}{x}}^{-2}(\overset{\circ}{\chi}_{\mathrm{I}} - \overset{\circ}{\chi}_{\mathrm{II}})^2$ is the component of Mahalanobis distance, wich caused by the missed independent feature $\overset{\circ}{x} \in \overset{\circ}{X}$. In view of it, having limited to accuracy $(1/n)$, neglecting members of the order $(1/n^2)$, we receive

$$\Delta(V) = \frac{1}{\left(\tau_{\overset{\circ}{X}}^2 + \overset{\circ}{m}\cdot c_A^{-1}\right)\cdot\left[\left(\tau_{\overset{\circ}{X}}^2 - \gamma^2\right) + (\overset{\circ}{m}-1)\cdot c_A^{-1}\right]} \times$$

$$\times\left\{-\left(\tau_{\overset{\circ}{X}}^2\cdot\frac{r-\overset{\circ}{m}+1}{r-1} + \frac{r-\overset{\circ}{m}}{r-1}\cdot\overset{\circ}{m}\cdot c_A^{-1}\right)\cdot\left(\gamma^2\right)^2 + \right. \qquad (10)$$

$$+\tau_{\overset{\circ}{X}}^2\cdot\left(\tau_{\overset{\circ}{X}}^2\cdot\frac{r-\overset{\circ}{m}+2}{r-1} + 2\cdot\frac{r-\overset{\circ}{m}}{r-1}\cdot\overset{\circ}{m}\cdot c_A^{-1}\right)\cdot\gamma^2 -$$

$$\left.\left(\tau_{\overset{\circ}{X}}^2\right)^2\cdot\left(\tau_{\overset{\circ}{X}}^2\cdot\frac{1}{r-1} + \frac{r-\overset{\circ}{m}}{r-1}\cdot c_A^{-1}\right)\right\}.$$

The value $\Delta(V)$ can be both positive, and negative. If $\Delta(V) > 0$, the feature $\overset{\circ}{x}$ is necessary for including in discriminant function. If the $\Delta(V) < 0$, the $\overset{\circ}{x}$ should not be included in discriminant function as it will lead to decreasing of value D_{AB}^2, i.e. addition of an feature $\overset{\circ}{x}$ does not improve quality discriminant function by considered criterion. The condition $\Delta(V) < 0$ is a condition of a reduction (simplification) of discriminant function that is optimal by quantity and structure of features. This condition represents a condition of negative definiteness of a quadratic trinomial relatively γ^2 in braces (10). Reduction of discriminant function is possible when value γ^2 below then threshold value

$$(\gamma^2)_{por} = \tau^2_{\overset{o}{X}} \cdot \frac{\left(\frac{\tau^2_{\overset{o}{X}}}{r-1}\right) + c_A^{-1}}{\tau^2_{\overset{o}{X}}\left(\frac{r-\overset{o}{m}+1}{r-1}\right) + \overset{o}{m} \cdot c_A^{-1}}. \tag{11}$$

In figure 1 dependences of threshold value (11) from volume samples n for a set of Mahalanobis distance $\tau^2_{\overset{o}{X}}$ $(\tau^2_{\overset{o}{X}} = 6, 8, \dots, 18)$ are submitted at fixed $\overset{o}{m} = 6$.

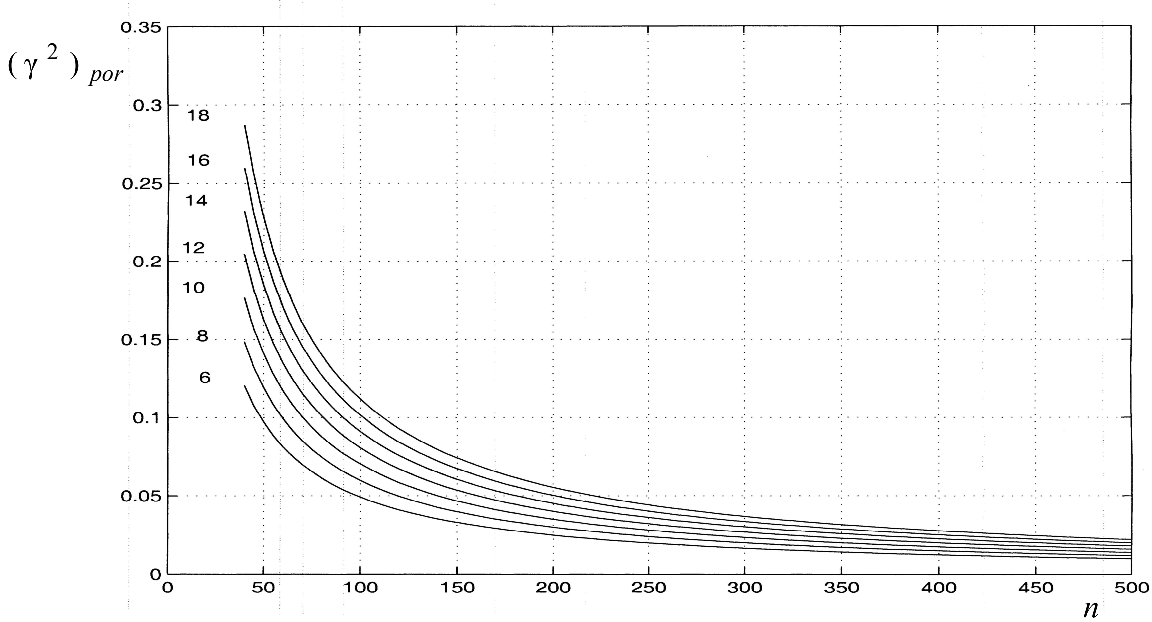

Figure 1 - Dependences of threshold value $(\gamma^2)_{por}$ on volume of sabsamples n for AB-method

Let's note, that in asymptotic at $n \to \infty$ $(r \to \infty,\ c_A^{-1} \to 0)$ the condition of the reduction is not carried out, i.e. $V_{OPT} = \overset{o}{X}$.

The method of comparison of discriminant functions based on sliding examination. Suppose that at the step with number s $(s = 1, 2, \dots, m)$ of algorithm complete sorting-out of all possible sets of features only s components from the set X (which constitute the current set V)

can be included in the discriminant function. In the following are supposed $\mathbf{V}_{\mathrm{I}}$ and $\mathbf{V}_{\mathrm{II}}$ are $(s \times n_{\mathrm{I}})$- and $(s \times n_{\mathrm{II}})$-matrices of observations in general sets P_{I} and P_{II}, $\mathbf{v}_{\mathrm{I}}$ and $\mathbf{v}_{\mathrm{II}}$ are s-dimensional column vectors of the mathematical expectations of the observations in P_{I} and P_{II}, $\mathbf{\Sigma}_V$ is covariance $(s \times s)$- matrix of the observations in P_{I} and P_{II}.

The traditional way of sliding examination is the following: a) one of the observations is eliminated from training sample; b) this observations is classified on basis of discriminant function which was constructed without it; c) observations returns into sample; d) this procedure repeat for second observation, third and so on, until all observations are classified in this way. Usually, a probability of erroneous classification is estimated in applications, that is number of erroneous classified observations is calculated. As opposed to traditional ways of sliding examination in way what we have proposed, a distance is calculated

$$D_S^2(V) = \frac{1}{2}(D_{S\mathrm{I}}^2(V) + D_{S\mathrm{II}}^2(V)), \tag{12}$$

$$D_{S\mathrm{I}}^2(V) = (n_{\mathrm{I}})^{-1} \sum_{i=1}^{n_{\mathrm{I}}} \frac{\mathbf{d}_{(\mathrm{I},i)}^{\mathrm{T}} (\mathbf{v}_{\mathrm{I}i} - \tilde{\mathbf{v}}_{\mathrm{II}})(\mathbf{v}_{\mathrm{I}i} - \tilde{\mathbf{v}}_{\mathrm{II}})^{\mathrm{T}} \mathbf{d}_{(\mathrm{I},i)}}{\mathbf{d}_{(\mathrm{I},i)}^{\mathrm{T}} \mathbf{S}_{(\mathrm{I},i)} \mathbf{d}_{(\mathrm{I},i)}} =$$

$$= (n_{\mathrm{I}})^{-1} \sum_{i=1}^{n_{\mathrm{I}}} (\mathbf{v}_{\mathrm{I}i} - \tilde{\mathbf{v}}_{\mathrm{II}})^{\mathrm{T}} \mathbf{W}_{(\mathrm{I},i)} (\mathbf{v}_{\mathrm{I}i} - \tilde{\mathbf{v}}_{\mathrm{II}}), \tag{13}$$

$$D_{S\mathrm{II}}^2(V) = (n_{\mathrm{II}})^{-1} \sum_{j=1}^{n_{\mathrm{II}}} \frac{\mathbf{d}_{(\mathrm{II},j)}^{\mathrm{T}} (\tilde{\mathbf{v}}_{\mathrm{I}} - \mathbf{v}_{\mathrm{II}j})(\tilde{\mathbf{v}}_{\mathrm{I}} - \mathbf{v}_{\mathrm{II}j})^{\mathrm{T}} \mathbf{d}_{(\mathrm{II},j)}}{\mathbf{d}_{(\mathrm{II},j)}^{\mathrm{T}} \mathbf{S}_{(\mathrm{II},j)} \mathbf{d}_{(\mathrm{II},j)}} =$$

$$= (n_{\mathrm{II}})^{-1} \sum_{j=1}^{n_{\mathrm{II}}} (\tilde{\mathbf{v}}_{\mathrm{I}} - \mathbf{v}_{\mathrm{II}j})^{\mathrm{T}} \mathbf{W}_{(\mathrm{II},j)} (\tilde{\mathbf{v}}_{\mathrm{I}} - \mathbf{v}_{\mathrm{II}j}), \tag{14}$$

$$\mathbf{W}_{(\mathrm{I},i)} = \frac{\mathbf{d}_{(\mathrm{I},i)} \mathbf{d}_{(\mathrm{I},i)}^{\mathrm{T}}}{\mathbf{d}_{(\mathrm{I},i)}^{\mathrm{T}} \mathbf{S}_{(\mathrm{I},i)} \mathbf{d}_{(\mathrm{I},i)}}; \quad \mathbf{W}_{(\mathrm{II},j)} = \frac{\mathbf{d}_{(\mathrm{II},j)} \mathbf{d}_{(\mathrm{II},j)}^{\mathrm{T}}}{\mathbf{d}_{(\mathrm{II},j)}^{\mathrm{T}} \mathbf{S}_{(\mathrm{II},j)} \mathbf{d}_{(\mathrm{II},j)}}. \tag{15}$$

In formula (13), vector $\mathbf{d}_{(\mathrm{I},i)}$ is an estimate of the coefficients of the Fisher discriminant function. Specifically, it is the estimate calculated without the observation number i in the first group:

$$\mathbf{d}_{(\mathrm{I},i)} = \mathbf{S}_{(\mathrm{I},i)}^{-1} (\tilde{\mathbf{v}}_{\mathrm{I}(i)} - \tilde{\mathbf{v}}_{\mathrm{II}}), \tag{16}$$

where the vector $\tilde{\mathbf{v}}_{\mathrm{I}(i)}$ is estimate of the mathematical expectation $\mathbf{v}_{\mathrm{I}}$

$$\tilde{\mathbf{v}}_{\mathrm{I}(i)} = (n_{\mathrm{I}} - 1)^{-1} \left(\sum_{h=1}^{n_{\mathrm{I}}} \mathbf{v}_{\mathrm{I}h} - \mathbf{v}_{\mathrm{I}i} \right); \tag{17}$$

the $\tilde{\mathbf{v}}_{\mathrm{II}}$ is estimate of the mathematical expectation $\mathbf{v}_{\mathrm{II}}$

$$\tilde{\mathbf{v}}_{\mathrm{II}} = (n_{\mathrm{II}})^{-1} \sum_{h=1}^{n_{\mathrm{II}}} \mathbf{v}_{\mathrm{II}h}; \tag{18}$$

the matrix $\mathbf{S}_{(\mathrm{I},i)}$ is an unbiased estimate of covariance matrix $\mathbf{\Sigma}_V$

$$\mathbf{S}_{(\mathrm{I},i)} = (n_{\mathrm{I}} + n_{\mathrm{II}} - 3)^{-1} \left[\sum_{\substack{h=1 \\ (h \neq i)}}^{n_{\mathrm{I}}} \tilde{\tilde{\mathbf{v}}}_{\mathrm{I}h(i)} \tilde{\tilde{\mathbf{v}}}_{\mathrm{I}h(i)}^{\mathrm{T}} + \sum_{q=1}^{n_{\mathrm{II}}} \tilde{\tilde{\mathbf{v}}}_{\mathrm{II}q} \tilde{\tilde{\mathbf{v}}}_{\mathrm{II}q}^{\mathrm{T}} \right], \tag{19}$$

where $\tilde{\tilde{\mathbf{v}}}_{\mathrm{I}h(i)}$ is the observation numbered h in the first group, centered about the estimate $\tilde{\mathbf{v}}_{\mathrm{I}(i)}$

$$\tilde{\tilde{\mathbf{v}}}_{\mathrm{I}h(i)} = \mathbf{v}_{\mathrm{I}h} - \tilde{\mathbf{v}}_{\mathrm{I}(i)}, \quad h = 1, 2, \ldots, n_{\mathrm{I}} \ (h \neq i); \tag{20}$$

and $\tilde{\tilde{\mathbf{v}}}_{\mathrm{II}q}$ is the observations numbered q in the second group, centered about the estimate $\tilde{\mathbf{v}}_{II}$

$$\tilde{\tilde{\mathbf{v}}}_{\mathrm{II}q} = \mathbf{v}_{\mathrm{I}q} - \tilde{\mathbf{v}}_{\mathrm{II}}, \quad q = 1, 2, \ldots, n_{\mathrm{II}}. \tag{21}$$

In formula (14), vector $\mathbf{d}_{(\mathrm{I},j)}$ is an estimate of the coefficients of the Fisher discriminant function. Specifically, it is the estimate calculated without the observation number j in the second group:

$$\mathbf{d}_{(\mathrm{II},j)} = \mathbf{S}_{(\mathrm{II},j)}^{-1} (\tilde{\mathbf{v}}_{\mathrm{I}} - \tilde{\mathbf{v}}_{\mathrm{II}(j)}), \tag{22}$$

where vector $\tilde{\mathbf{v}}_{\mathrm{I}}$ is estimate of the mathematical expectation $\mathbf{v}_{\mathrm{I}}$ and calculated analogous to (18); the vector $\tilde{\mathbf{v}}_{\mathrm{II}(j)}$ is estimate of mathematical expectation $\mathbf{v}_{\mathrm{II}}$ and calculated analogues to (17); the matrix $\mathbf{S}_{(\mathrm{II},j)}$ is an unbiased estimate of the covariance matrix $\mathbf{\Sigma}_V$ and calculated analogous to (19).

From formulas (12)–(22), it is obvious that the statistics $D_{\mathrm{SI}}^2(V)$ is simply the weighed sum of the paired distances between the observations of the first group and estimate of the mathematical expectation $\mathbf{v}_{\mathrm{II}}$ second group, and that statistics $D_{\mathrm{SII}}^2(V)$ – is simply the weighed sum of the paired distances

 ISSN 1562-9945

between the observations of the second group and estimate of the mathematical expectation $\mathbf{v}_{\mathrm{I}}$ first group.

Using (5) and (11), we obtain for $D^2_{S\mathrm{I}}(V)$ and $D^2_{S\mathrm{II}}(V)$

$$D^2_{S\mathrm{I}}(V) = (n_{\mathrm{I}})^{-1} \sum_{i=1}^{n_{\mathrm{I}}} \frac{[(\mathbf{v}_{\mathrm{I}i} - \tilde{\mathbf{v}}_{\mathrm{II}})^{\mathrm{T}} \mathbf{S}^{-1}_{(\mathrm{I},i)} (\tilde{\mathbf{v}}_{\mathrm{I}(i)} - \tilde{\mathbf{v}}_{\mathrm{II}})]^2}{(\tilde{\mathbf{v}}_{\mathrm{I}(i)} - \tilde{\mathbf{v}}_{\mathrm{II}})^{\mathrm{T}} \mathbf{S}^{-1}_{(\mathrm{I},i)} (\tilde{\mathbf{v}}_{\mathrm{I}(i)} - \tilde{\mathbf{v}}_{\mathrm{II}})}, \tag{23}$$

$$D^2_{S\mathrm{II}}(V) = (n_{\mathrm{II}})^{-1} \sum_{j=1}^{n_{\mathrm{II}}} \frac{[(\tilde{\mathbf{v}}_{\mathrm{I}} - \mathbf{v}_{\mathrm{II}j})^{\mathrm{T}} \mathbf{S}^{-1}_{(\mathrm{II},j)} (\tilde{\mathbf{v}}_{\mathrm{I}} - \tilde{\mathbf{v}}_{\mathrm{II}(j)})]^2}{(\tilde{\mathbf{v}}_{\mathrm{I}} - \tilde{\mathbf{v}}_{\mathrm{II}(j)})^{\mathrm{T}} \mathbf{S}^{-1}_{(\mathrm{II},j)} (\tilde{\mathbf{v}}_{\mathrm{I}} - \tilde{\mathbf{v}}_{\mathrm{II}(j)})}. \tag{24}$$

Let $\tau_V^2 = (\mathbf{v}_{\mathrm{I}} - \mathbf{v}_{\mathrm{II}})^{\mathrm{T}} \mathbf{\Sigma}_V^{-1} (\mathbf{v}_{\mathrm{I}} - \mathbf{v}_{\mathrm{II}})$ be the Mahalanobis distance for the set of components V, $n_{\mathrm{I}} = n_{\mathrm{II}} = n$, $r = n_{\mathrm{I}} + n_{\mathrm{II}} - 3 = 2n - 3$, $c^{-1} = n^{-1} + (n-1)^{-1}$.

Theorem 2. For the random variable $D_S^2(V)$, we have

$$E\{D_S^2(V)\} = \left(\tau_V^2 - \frac{\tau_V^2\,[s-(r-1)/(r-s)]\cdot c^{-1}}{(\tau_V^2 + s\cdot c^{-1})} - \frac{(n+1)\,(r-1)}{n(r-s)} \right) \cdot \frac{r}{r-s-3}. \tag{25}$$

The validity of theorem follows from the validity of the following: 1) the observation $\mathbf{v}_{\mathrm{I}i}$, the estimate $\tilde{\mathbf{v}}_{\mathrm{I}}$ and the estimate $\mathbf{S}_{(\mathrm{I},i)}$ are independent; 2) the observation $\mathbf{v}_{\mathrm{II}j}$, the estimate $\tilde{\mathbf{v}}_{\mathrm{II}}$ and the estimate $\mathbf{S}_{(\mathrm{II},j)}$ are independent; 3) matrices $\mathbf{S}_{(\mathrm{I},i)}$ and $\mathbf{S}_{(\mathrm{II},j)}$ are random $(s \times s)$-matrices, which have a Wishart distribution with r degrees of freedom.

Definition 3. The optimal set components (set features) is defined as the set V_{OPT} for which

$$V_{OPT} = \arg\max_{V \subseteq X} E\{D_S^2(V)\}. \tag{26}$$

Definition 4. Optimal discriminant function with respect to the number and composition of the components is defined as the Fisher discriminant function constructed on the set of components V_{OPT}.

We proved that optimal set of components exist in the way that considered and formulated the conditions under which the optimal discriminant function is simplified in number of the features included in it. For this purpose, it was investigated $E\{D_S^2(V)\}$ depending on composition of set V.

Reduction (simplification) condition of optimal discriminant function. As a rule, in practical applications, parameters of general populations are unknown; however they can be estimated as statistical estimates by training samples of observations of limited volume. It is known, that if we use constructed rule of classification to the training sample, then estimate of recognition quality will be overstated by mathematical expectation in comparison with the same evaluation of quality on data, independent of training data.

Way of sliding examination give not overstated estimates of recognition quality. Experience of practical applications and test investigations of this way on basis of method of statistical test show that in this way: 1) on increase of size of observations samples increases the number of components in the set, on which the best quality of recognition is attained, and on decrease of size of observations samples the number of components in such set decreases; 2) on increase of the Mahalanobis distance τ_X^2 between general populations (from which observation samples were obtain) the number of components increases in the set, on which the best quality of recognitions is attained, and on decrease of this distance the number of components in such set decreases.

Our analytical investigations confirm these empirically determined regularities about the existence of the discriminant function optimal by the number and composition of components. Let's formulate the conditions of reduction (simplification) optimal discriminant function for a special case of an independent feature. Let the set of V is those, that is carried out $\overset{\circ}{X} = V \cup \overset{\circ}{x}$, where $\overset{\circ}{x} \in \overset{\circ}{X}$ (one feature is missed). Taking into account (25), we receive

$$\Delta(V) = E\{D_S^2(\overset{\circ}{X})\} - E\{D_S^2(V)\} =$$

$$= \left(\tau_{\overset{\circ}{X}}^2 - \frac{\tau_{\overset{\circ}{X}}^2 [\overset{\circ}{m} - (r-1)/(r-\overset{\circ}{m})] \cdot c^{-1}}{\tau_{\overset{\circ}{X}}^2 + \overset{\circ}{m} \cdot c^{-1}} - \frac{(n+1)(r-1)}{n(r-\overset{\circ}{m})} \right) \cdot \frac{r}{r-\overset{\circ}{m}-3} -$$

$$- \left(\tau_V^2 - \frac{\tau_V^2 [(\overset{\circ}{m}-1) - (r-1)/(r-\overset{\circ}{m}+1)] \cdot c^{-1}}{\tau_V^2 + (\overset{\circ}{m}-1) \cdot c^{-1}} - \frac{(n+1)(r-1)}{n(r-\overset{\circ}{m}+1)} \right) \cdot \frac{r}{r-\overset{\circ}{m}-2}. \quad (27)$$

According to the above mentioned lemmas for Mahalanobis distances of sets V and $\overset{o}{X}$ the ratio $\tau_V^2 = \tau_{\overset{o}{X}}^2 - \gamma^2$ is carried out, where $\gamma^2 = \sigma_{\overset{o}{x}}^{-2}(\overset{o}{\chi}_{\mathrm{I}} - \overset{o}{\chi}_{\mathrm{II}})^2$ is the component of Mahalanobis distance, wich caused by the missed independent feature $\overset{o}{x} \in \overset{o}{X}$. In view of it, having limited to accuracy $(1/n)$, neglecting members of the order $(1/n^2)$, we receive

$$\Delta(V) = \frac{r}{r - \overset{o}{m} - 2} \cdot \frac{1}{\left(\tau_{\overset{o}{X}}^2 + \overset{o}{m} \cdot c^{-1}\right) \cdot \left[\left(\tau_{\overset{o}{X}}^2 - \gamma^2\right) + (\overset{o}{m} - 1) \cdot c^{-1}\right]} \times$$

$$\times \left\{ -\left(\tau_{\overset{o}{X}}^2 + \overset{o}{m} \cdot c^{-1}\right) \cdot \left(\gamma^2\right)^2 + \right. \tag{28}$$

$$+ \tau_{\overset{o}{X}}^2 \cdot \left(\tau_{\overset{o}{X}}^2 \cdot \frac{r - \overset{o}{m} - 4}{r - \overset{o}{m} - 3} + 2\overset{o}{m} \cdot c^{-1} + \frac{2}{r - \overset{o}{m} - 3}\right) \cdot \gamma^2 -$$

$$\left. \left(\tau_{\overset{o}{X}}^2\right)^2 \cdot \left(\tau_{\overset{o}{X}}^2 \cdot \frac{1}{r - \overset{o}{m} - 3} + c^{-1}\right) \right\}.$$

The value $\Delta(V)$ can be both positive, and negative. If $\Delta(V) > 0$, the feature $\overset{o}{x}$ is necessary for including in discriminant function. If the $\Delta(V) < 0$ the $\overset{o}{x}$ should not be included in discriminant function as it will lead to decreasing of value D_S^2, i.e. addition of an feature $\overset{o}{x}$ does not improve quality discriminant function by considered criterion. The condition $\Delta(V) < 0$ is a condition of a reduction (simplification) of discriminant function that is optimal by quantity and structure of features. This condition represents a condition of negative definiteness of a quadratic trinomial relatively γ^2 in braces (28). Reduction of discriminant function is possible when value γ^2 below then threshold value

$$(\gamma^2)_{por} = \tau^2_{\overset{o}{X}} \cdot \frac{\left(\dfrac{\tau^2_{\overset{o}{X}}}{r - \overset{o}{m} - 3} \right) + c^{-1}}{\tau^2_{\overset{o}{X}} + \overset{o}{m} \cdot c^{-1}} . \tag{29}$$

In figure 2 dependences of threshold value (29) from volume samples n for a set of Mahalanobis distance $\tau^2_{\overset{o}{X}}$ $(\tau^2_{\overset{o}{X}} = 6, 8, \dots, 18)$ are submitted at fixed $\overset{o}{m} = 6$.

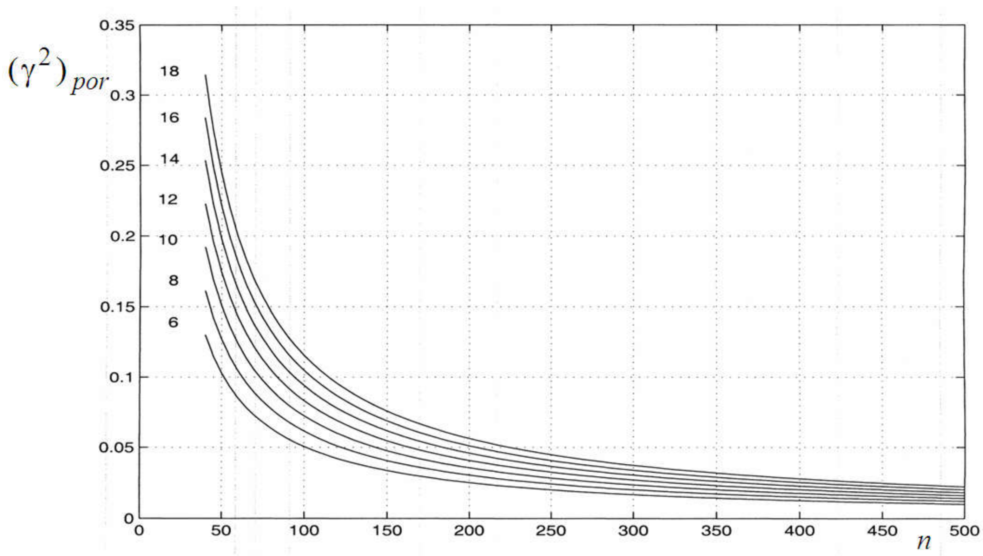

Figure 2 - Dependences of threshold value $(\gamma^2)_{por}$ on volume of subsamples n for S-method

Let's note, that in asymptotic at $n \to \infty$ $(r \to \infty, c^{-1} \to 0)$ the condition of the reduction is not carried out, i.e. $V_{OPT} = \overset{o}{X}$.

Conclusion. The two methods for comparison of the discriminant functions are proved. The first method based on dividing of the initial data sample on training and testing subsamples and second method based on sliding examination. In spite of successful use of these ways in practice and repeated confirmation of its efficiency by the method of statistical test, it was considered traditionally as heuristic method.

It is shown that under condition of structural uncertainty and the absence of a priori estimates of parameters of general sets these methods make it possible to solve the problem of search of the discriminant function of optimal complexity. Conditions of reduction (simplification) of discriminant function, which is optimal by structure of features, are revealed. It is shown, as these conditions depend on volumes samples and parameters of general sets, i.e. on mathematical expectations and covariance matrices of features.

REFERENCES

1. Sarychev A. P. Circuit of Discriminant Analysis with Learning and Checking Subsamples of Observations. Automatica (Ukraine), 1: 32–41, 1990 (Journal of Automation and Information Sciences. – Scripta Technika Inc. – Vol. 23, 1: 29–39, 1990).

2. Miroshnichenko L. V., Sarychev A. P. Scheme of the Sliding Exam for Search of the Optimal Set of Characters in the Problem of Discriminant Analysis. Automatica (Ukraine), 1: 35–44, 1992 (Journal of Automation and Information Sciences. – Scripta Technika Inc., Vol. 25, 1: 33–42, 1992).

3. Sarychev A. P. *S*-Scheme of Sliding Examination for Optimal Set Features Determination in Discriminant Analysis by the Group Method of Data Handling // System Analysis and Modelling Simulation (SAMS), Taylor & Francis. – Vol. 43. – 10: 1351–1362, 2003.

4. 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. – Begell House Inc., Vol. 40, 6: 27–40, 2008).

5. Sarychev A. P. Identification of Structural-Uncertain Systems States. Dnipropetrovs'k, Ukraine, NAS Ukraine and NSA Ukraine, Institute of Technical Mechanics, 268, 2008.

6. Sarychev A. P., Sarycheva L. V. The Solution of the Discriminant Analysis Task on Basis of the Group Method of Data Handling (Ukraine). Upravlyayuschie sistemy i mashiny. 2 (244): 18–27, 2013.

7. Sarychev A. P. An Iterative Algorithm GMDH for Synthesizing the Separating Function in a Linear Discriminant Analysis Problem (Ukraine) // Automatika, 2: 20–24, 1988 (Soviet Journal of Automation and Information Sciences. – Scripta Technika Inc. Volume 21, Number 2, 17–22, 1988).

8. Sarychev A. P., Sarycheva L. V. An Iterative Algorithm GMDH for Synthesizing the Separating Function in a Nonlinear Discriminant Analysis (Ukraine) // Inductive Modeling Complex Systems. – Kyiv : IRTC ITS. 2: 212–225, 2010.

Received 02.12.2019.
Accepted 04.12.2019.

Визначення оптимальної множини ознак у дискримінантному аналізі на основі МГУА

Розглянуто задачу пошуку дискримінантної функції оптимальної складності. Описано два критерії якості дискримінантних функцій, яких розроблено в методі групового урахування аргументів (МГУА): критерій, що заснований на розбивці спостережень на навчальні й перевірочні вибірки, та критерій ковзного іспиту.

GMDH-Based Optimal Set Features Determination in Discriminant Analysis

The task of searching optimum on complexity discriminant function is considered. Criteria of quality of the discriminant functions developed in the Group Method of Data Handling are described: the criterion based on a partition of observations on training and testing samples, and criterion of sliding examination.

Сарычев Александр Павлович – доктор технических наук, ведущий научный сотрудник, Институт технической механики Национальной академии наук Украины и Государственного космического агентства Украины.
Сарычева Людмила Виссарионовна – кандидат физико-математических наук, профессор кафедры геоинформационных систем Национального горного университета.

Саричев Олександр Павлович - доктор технічних наук, провідний науковий співробітник, Інститут технічної механіки Національної академії наук України та Державного космічного агентства України.
Саричева Людмила Віссаріонівна - кандидат фізико-математичних наук, професор кафедри геоінформаційних систем Національного гірничого університету.

Sarychev Alexander Pavlovich – doctor of technical sciences, leading scientific researcher, Institute of Technical Mechanics of the National Academy of Sciences of Ukraine and State Space Agency of Ukraine.
Sarycheva Lyudmila Vissarionovna – professor of Geoinformatics Department of National Mining Universityof Ukraine.