УДК 519.25

DISCRIMINANT FUNCTIONS QUALITY ESTIMATION ON THE BASIS OF TRAINING AND TESTING SAMPLES

Alexander Sarychev [1] and Lyudmyla Sarycheva [2]

[1] *Institute of Technical Mechanics of the National Academy of Sciences of Ukraine, 15 Leshko-Popel St., Dnipropetrovs'k, 49005, Ukraine*
[2] *National Mining University of Ukraine, 19 K. Marks Ave., Dnipropetrovs'k, 49027, Ukraine*
Sarychev@prognoz.dp.ua, Sarycheval@nmu.org.ua

Обгрунтовано спосіб порівняння дискримінантних функцій з розбиттям вибірок спостережень на навчальні й перевірні підвибірки. Отримано умови існування оптимальної множини ознак, які залежать від параметрів генеральних сукупностей і обсягів вибірок. Виявлено закономірності спрощення оптимальної дискримінантної функції при зменшенні обсягів вибірок і при збільшенні дисперсій ознак.
Ключові слова: метод групового урахування аргументів, невизначеність за складом ознак, критерій якості лінійної дискримінантної функції.

The way of comparison of discriminant functions with dividing samples observations on training and testing subsamples is proved. Conditions of existence of optimum set of features which depend on parameters of general sets and volumes samples are received. Laws of simplification of optimum discriminant function at decrease of volumes samples and at increase of dispersions of features are revealed.
Keywords: Group Method of Data Handling, uncertainty on structure of features, criterion of linear discriminant function quality.

Обоснован способ сравнения дискриминантных функций с разбиением выборок наблюдений на обучающие и проверочные подвыборки. Получены условия существования оптимального множества признаков, которые зависят от параметров генеральных совокупностей и объемов выборок. Выявлены закономерности упрощения оптимальной дискриминантной функции при уменьшении объемов выборок и при увеличении дисперсий признаков.
Ключевые слова: метод группового учета аргументов, неопределенность по составу признаков, критерий качества линейной дискриминантной функции.

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 [1-4]. 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. It is supposed, that as such method is chosen the complete sorting-out of all possible combinations of features.

1. Way of comparison of discriminant functions on the basis of training and testing subsamples

Suppose that at the step with number s $(s = 1,2,..., m)$ 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 the component V on the 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:

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

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

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

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

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

In formula (4) $\mathbf{v}_{kA}$ $(k = \mathrm{I}, \mathrm{II})$ are matrices of deviations of observations $\mathbf{V}_{kA}$ from estimate $\tilde{\mathbf{v}}_{kA}$

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

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 $D^2_{AB}(V)$:

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

Let $\tau^2_V = (\mathbf{v}_{\mathrm{I}} - \mathbf{v}_{\mathrm{II}})^{\mathrm{T}} \boldsymbol{\Sigma}_V^{-1} (\mathbf{v}_{\mathrm{I}} - \mathbf{v}_{\mathrm{II}})$ 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. For mathematical expectation of random variable $D^2_{AB}(V)$, we have

$$E\{D^2_{AB}(V)\} = \left(\tau^2_V - \frac{\tau^2_V [s - (r-1)/(r-s)]\, c_A^{-1}}{\tau^2_V + s\, c_A^{-1}} + c_B^{-1} \frac{r-1}{r-s} \right) \frac{r-s}{r-1}. \qquad (7)$$

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

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

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 $X = \overset{\circ}{X} \cup \overset{\circ}{R} \cup \tilde{R} = \overset{\circ}{V} \cup \tilde{R}$: so that 1) $\overset{\circ}{X} \neq \varnothing$ (where $\varnothing$ is the empty set) is the set of components whose mathematical expectation satisfy $\overset{\circ}{\chi}_{\mathrm{I}h} \neq \overset{\circ}{\chi}_{\mathrm{II}h}$, $h = 1, 2, ..., \overset{\circ}{m}$, where $\overset{\circ}{m}$ is their number; 2) $\overset{\circ}{R}$ is the set of components whose mathematical

expectation satisfy $\overset{o}{\rho}_{Ih} = \overset{o}{\rho}_{IIh}, h = 1, 2, ..., \overset{o}{l}$, 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 $\overset{o}{V} = \overset{o}{X} \bigcup \overset{o}{R}$ 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_V^2; 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_V^2; 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_V^2.

2. Conditions of reduction (simplification) of optimum 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 $\overset{o}{X} = V \bigcup \overset{o}{x}$, where $\overset{o}{x} \in \overset{o}{X}$ (one feature is missed). Taking into account (7), we receive

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

$$\left(\tau^2_{\underset{X}{o}} - \frac{\tau^2_{\underset{X}{o}} [\overset{o}{m} - (r-1)/(r-\overset{o}{m})]\, c_A^{-1}}{\tau^2_{\underset{X}{o}} + \overset{o}{m} c_A^{-1}} + c_B^{-1} \frac{r-1}{r-\overset{o}{m}} \right) \frac{r-\overset{o}{m}}{r-1} -$$

$$- \left(\tau^2_V - \frac{\tau^2_V [(\overset{o}{m}-1) - (r-1)/(r-\overset{o}{m}+1)]\, c_A^{-1}}{\tau^2_V + (\overset{o}{m}-1) c_A^{-1}} + c_B^{-1} \frac{r-1}{r-\overset{o}{m}+1} \right) \frac{r-\overset{o}{m}+1}{r-1}. \qquad (9)$$

According to the above mentioned lemmas for Mahalanobis distances of sets V and $\overset{o}{X}$ the ratio $\tau^2_V = \tau^2_{\underset{X}{o}} - \gamma^2$ is carried out, where $\gamma^2 = \sigma^{-2}_{\underset{x}{o}} (\chi_{\mathrm{I}} - \chi_{\mathrm{II}})^2$ is the component of Mahalanobis distance, that 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

$$\Delta(V) = \frac{1}{\left(\tau^2_{\underset{X}{o}} + \overset{o}{m} c_A^{-1}\right) \cdot \left[\left(\tau^2_{\underset{X}{o}} - \gamma^2\right) + (\overset{o}{m}-1) c_A^{-1}\right]} \cdot \left\{ - \left(\tau^2_{\underset{X}{o}} \cdot \frac{r-\overset{o}{m}+1}{r-1} + \frac{r-\overset{o}{m}}{r-1} \cdot \overset{o}{m} \cdot c_A^{-1} \right) \cdot \left(\gamma^2\right)^2 + \right.$$

$$\left. + \tau^2_{\underset{X}{o}} \cdot \left(\tau^2_{\underset{X}{o}} \cdot \frac{r-\overset{o}{m}+2}{r-1} + 2 \cdot \frac{r-\overset{o}{m}}{r-1} \cdot \overset{o}{m} \cdot c_A^{-1} \right) \cdot \gamma^2 - \left(\tau^2_{\underset{X}{o}} \right)^2 \cdot \left(\tau^2_{\underset{X}{o}} \cdot \frac{1}{r-1} + \frac{r-\overset{o}{m}}{r-1} \cdot c_A^{-1} \right) \right\}. \qquad (10)$$

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^2_{AB}, 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 (10). Reduction of discriminant function is possible when value γ^2 below then threshold value

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

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

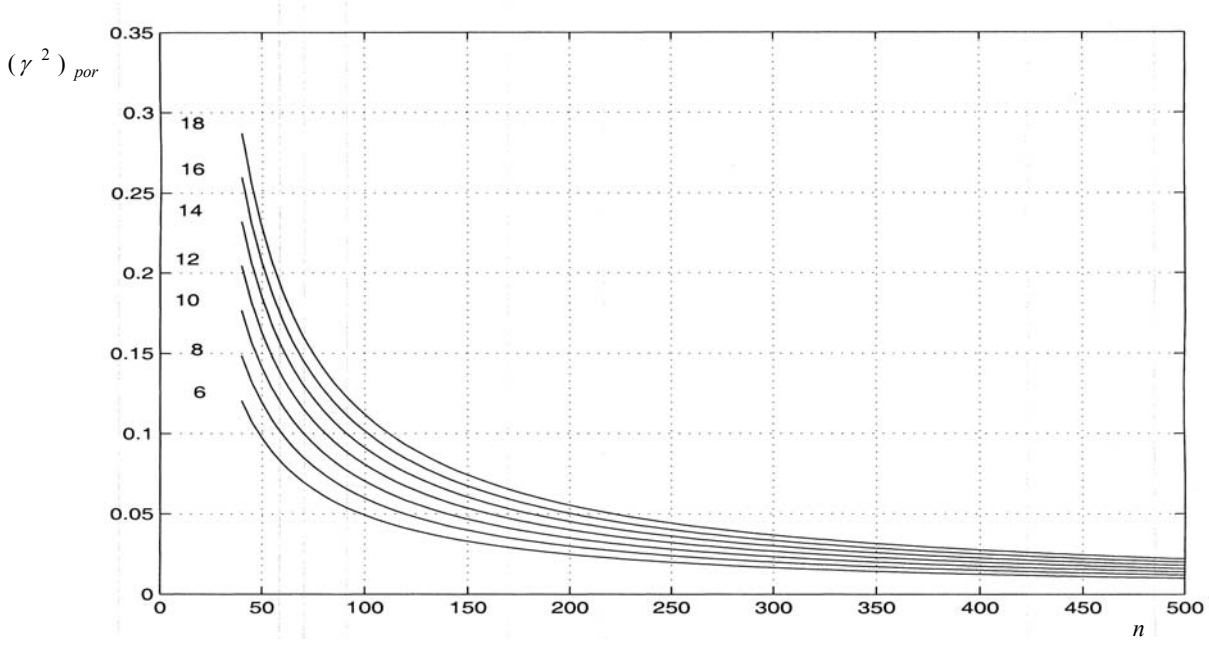

Fig. – Dependences of threshold value $(\gamma^2)_{por}$ on volume of sabsamples n

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. $V_{OPT} = \overset{o}{X}$.

3. Conclusions

The method for comparison of the discriminant functions based on dividing of the initial data sample on training and testing subsamples is proved. In spite of successful use of this way in practice and repeated confirmation of its efficiency by the method of statistical test, it was supposed traditionally as heuristic method. 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. It was shown that under condition of structural uncertainty and the absence of a priori estimates of parameters of general sets this method make it possible to solve the problem of search of the discriminant function of optimal complexity.

References

1. Sarychev A. P. Circuit of Discriminant Analysis with Learning and Checking Subsamplings 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. 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).

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

5. Sarychev A. P., L. V. Sarycheva. The Optimal Set Features Determination in Discriminant Analysis by the Group Method of Data Handling // Systems Analysis and Modeling Simulation (SAMS), Overseas Publishers Association. – 1998. – Vol. 31. – P. 153–167.

6. Sarychev A. P. S-Scheme of Sliding Examination for Optimal Set Features Determination in Discriminant Analysis by the Group Method of Data Handling / A. P. Sarychev // System Analysis and Modelling Simulation (SAMS), Taylor & Francis. – 2003. – Vol. 43. – № 10. – P. 1351–1362.

7. Sarychev A. P., Sarycheva L. V. Quality Estimation of Discriminant Functions by Sliding Examination // The Forth Workshop on Inductive Modelling (IWIM–2011) : July 4–8 2011, Kyiv, Ukraine : Proc. – Kyiv : IRTC ITS, 2011. – ISBN 978-966-02-6078-8. – P. 104–108.