УДК 519.25

# ОЦЕНИВАНИЕ КАЧЕСТВА ДИСКРИМИНАНТНЫХ ФУНКЦИЙ С РАЗБИЕНИЕМ НАБЛЮДЕНИЙ НА ОБУЧАЮЩИЕ И ПРОВЕРОЧНЫЕ ПОДВЫБОРКИ

А.П. Сарычев[1], Л.В. Сарычева[2]

[1] *Институт технической механики НАНУ и ГКАУ, г. Днепропетровск*
[2] *Национальный горный университет МОН Украины, г. Днепропетровск*

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

Обґрунтовано 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.*

## Введение

Решение задачи дискриминантного анализа в условиях структурной неопределенности по составу признаков предполагает принятие какого-либо способа сравнения дискриминантных функций (ДФ), построенных на различных множествах признаков. Два способа сравнения популярны в приложениях. Первый основан на разбиении наблюдений на обучающие и проверочные подвыборки. В этом способе обучающие подвыборки используются для оценивания коэффициентов ДФ, а проверочные – для оценивания ее качества классификации. Второй – способ скользящего экзамена, в котором в качестве проверочных выступают наблюдения, поочередно исключаемые из обучающей выборки. В литературе эти способы традиционно трактуются как эвристические приемы, хотя факт существования в них

оптимального множества признаков неоднократно подтверждался методом статистических испытаний. В рамках метода группового учета аргументов (МГУА) проведено аналитическое исследование этих способов [1–2].

### 1. Способ сравнения дискриминантных функций с разбиением наблюдений на обучающие и проверочные подвыборки

Пусть на этапе с номером $s$ $(s=1,2,...,m)$ алгоритма полного перебора сочетаний признаков в ДФ может быть включено только $s$ компонент из множества $X$, составляющих текущее анализируемое множество $V$. Пусть $V$ соответствуют: 1) $\mathbf{V}_{\mathrm{I}}$ и $\mathbf{V}_{\mathrm{II}}$ – $(s\times n_{\mathrm{I}})$- и $(s\times n_{\mathrm{II}})$-матрицы наблюдений из генеральных совокупностей $P_{\mathrm{I}}$ и $P_{\mathrm{II}}$; 2) $\mathbf{v}_{\mathrm{I}}$ и $\mathbf{v}_{\mathrm{II}}$ – $(s\times 1)$-векторы математических ожиданий для $P_{\mathrm{I}}$ и $P_{\mathrm{II}}$; 3) $\mathbf{\Sigma}_V$ – ковариационная $(s\times s)$-матрица. Рассмотрим оценку расстояния Махаланобиса, рассчитываемую с учетом разбиения наблюдений на обучающие и проверочные подвыборки. Вычислим оценки коэффициентов ДФ для множества компонент $V$ на обучающей подвыборке $A$ и используем их для оценивания расстояния Махаланобиса как отношение межгрупповой вариации к внутригрупповой вариации на проверочной подвыборке $B$:

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

В (1) $(s\times 1)$-вектор $\hat{\mathbf{d}}_A$ представляет собой рассчитанную на подвыборке $A$ фишеровскую оценку коэффициентов ДФ, которая построена в пространстве компонент множества $V$

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

где $(s\times 1)$-векторы $\tilde{\mathbf{v}}_{\mathrm{I}A}$ и $\tilde{\mathbf{v}}_{\mathrm{II}A}$ – оценки математических ожиданий $\mathbf{v}_{\mathrm{I}}$ и $\mathbf{v}_{\mathrm{II}}$:

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

$(s\times s)$-матрица $\mathbf{S}_A$ – несмещенная оценка ковариационной матрицы $\mathbf{\Sigma}_V$

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

В (4) $\mathbf{v}_{kA}$ $(k=\mathrm{I},\mathrm{II})$ – $(s\times n_k)$-матрицы, составленные из отклонений наблюдений $\mathbf{V}_{kA}$ компонент множества $V$ от оценок $\tilde{\mathbf{v}}_{kA}$

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

В (1) $(s\times 1)$-векторы $\tilde{\mathbf{v}}_{\mathrm{I}B}$ и $\tilde{\mathbf{v}}_{\mathrm{II}B}$ вычисляются аналогично (3), а $(s\times s)$-матрица $\mathbf{S}_B$ – аналогично (4)–(5); $n_{\mathrm{I}A}$ и $n_{\mathrm{II}A}$, $n_{\mathrm{I}B}$ и $n_{\mathrm{II}B}$ – объемы обучающих и проверочных подвыборок соответственно такие, что выполняется $n_{\mathrm{I}A} + n_{\mathrm{I}B} = n_{\mathrm{I}}$ и $n_{\mathrm{II}A} + n_{\mathrm{II}B} = n_{\mathrm{II}}$. Используя (2), для (1) получаем

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

**Теорема.** Для математического ожидания случайной величины $D^2_{AB}(V)$ выполняется

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

где $\tau^2_V = (\mathbf{v}_{\mathrm{I}} - \mathbf{v}_{\mathrm{II}})^{\mathrm{T}} \mathbf{\Sigma}_V^{-1} (\mathbf{v}_{\mathrm{I}} - \mathbf{v}_{\mathrm{II}})$ – расстояние Махаланобиса для множества $V$; $r = n_{\mathrm{I}A} + n_{\mathrm{II}A} - 2$, $c_A^{-1} = (n_{\mathrm{I}A}^{-1} + n_{\mathrm{II}A}^{-1})$, $c_B^{-1} = (n_{\mathrm{I}B}^{-1} + n_{\mathrm{II}B}^{-1})$.

Справедливость теоремы следует из следующих утверждений: 1) оценки, полученные на подвыборках $A$ и $B$, независимы; 2) оценки математических ожиданий (3) и ковариационной матрицы (4) независимы; 3) $\mathbf{S}_A$ – случайная $(s\times s)$-матрица, имеет распределение Уишарта с $r$ степенями свободы.

**Определение 1.** Оптимальным множеством компонент (признаков) называется множество $V_{OPT}$:

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

**Определение 2.** Оптимальной по количеству и составу компонент называется фишеровская дискриминантная функция, построенная на множестве $V_{OPT}$.

Доказано существование оптимального множества признаков в способе с разбиением наблюдений на обучающую и проверочную подвыборки, и сформулированы условия, при выполнении которых оптимальная ДФ упрощается по числу входящих в нее компонент.

С этой целью исследована зависимость $E\{D^2_{AB}(V)\}$ от состава множества $V$. Множество компонент $X$ может быть разбито на непересекающиеся подмножества $X = \overset{\mathrm{o}}{X} \cup \overset{\mathrm{o}}{R} \cup \tilde{R} = \overset{\mathrm{o}}{V} \cup \tilde{R}$: 1) $\overset{\mathrm{o}}{X} \neq \varnothing$ ($\varnothing$ – пустое множество) – множество компонент ($\overset{\mathrm{o}}{m}$ – их число), для математических ожиданий которых

выполнено $\overset{o}{\chi}_{Ih} \neq \overset{o}{\chi}_{IIh}, h = 1,2,..., \overset{o}{m}$; 2) $\overset{o}{R}$ – множество компонент, для математических ожиданий которых выполнено $\overset{o}{\rho}_{Ih} = \overset{o}{\rho}_{IIh}, h = 1,2,..., \overset{o}{l}$, где $\overset{o}{l}$ – их число, и каждая компонента из множества $\overset{o}{R}$ статистически зависит хотя бы от одной компоненты из множества $\overset{o}{X}$ (множество $\overset{o}{R}$ может быть пустым); 3) $\widetilde{R}$ – множество компонент, для математических ожиданий которых выполнено $\widetilde{\rho}_{Ih} = \widetilde{\rho}_{IIh}, h = 1,2,..., \widetilde{l}$, где $\widetilde{l}$ – их число, и каждая компонента из множества $\widetilde{R}$ статистически не зависит от любой из компонент множества $\overset{o}{X}$ (множество $\widetilde{R}$ может быть пустым).

Сформулированы в виде лемм соотношения между расстоянием Махаланобиса для множества компонент $\overset{o}{V} = \overset{o}{X} \bigcup \overset{o}{R}$ и расстоянием Махаланобиса для произвольного текущего анализируемого множества компонент $V \subseteq X$ [1–4]. Для случая известных параметров генеральных совокупностей из сформулированных лемм следует: 1) любая компонента из множества $\overset{o}{X}$ необходима в том смысле, что ее включение в текущее множество компонент $V$ увеличивает расстояние Махаланобиса $\tau_V^2$; 2) любая компонента из множества $\overset{o}{R}$ необходима в том смысле, что ее включение в $V$ увеличивает расстояние Махаланобиса $\tau_V^2$; 3) любая компонента из множества $\widetilde{R}$ избыточна в том смысле, что ее включение в множество $V$ не увеличивает расстояния Махаланобиса $\tau_V^2$.

## 2. Условие редукции (упрощения) оптимальной дискриминантной функции

В практических приложениях параметры генеральных совокупностей, как правило, неизвестны, но могут быть получены как статистические оценки по обучающим выборкам наблюдений ограниченного объема. Известно, что если применить построенное правило классификации к обучающей выборке, то оценка качества распознавания будет завышена по математическому ожиданию по сравнению с той же оценкой качества на независимых от обучения данных. Способ с разбиением наблюдений на обучающие и проверочные подвыборки дает незавышенные оценки качества распознавания. Опыт практических применений и тестовые исследования на основе метода статистических испытаний показывают, что в этой схеме: 1) с увеличением объема выборок наблюдений увеличивается количество компонент во множестве $V_0$, на котором достигается наилучшее качество распознавания, а с уменьшением объема

выборок наблюдений количество компонент в $V_0$ уменьшается; 2) с увеличением расстояния Махаланобиса $\tau_X^2$ между генеральными совокупностями (из которых получены выборки наблюдений) увеличивается количество компонент во множестве $V_0$, а с уменьшением этого расстояния количество компонент в таком множестве уменьшается. Проведенные аналитические исследования объясняют эти эмпирически установленные закономерности.

Сформулируем условие редукции (упрощения) оптимальной ДФ для частного случая независимого признака. Пусть множество $V$ таково, что выполняется $\overset{o}{X} = V \bigcup \overset{o}{x}$, где $\overset{o}{x} \in \overset{o}{X}$ (в ДФ пропущен один признак). Учитывая (7), получаем

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

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

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

В соответствии с вышеупомянутыми леммами для расстояний Махаланобиса множеств $V$ и $\overset{o}{X}$ выполняется соотношение: $\tau_V^2 = \tau^2_{\underset{X}{o}} - \gamma^2$, где $\gamma^2 = \sigma^{-2}_{\underset{x}{o}} (\overset{o}{\chi}_{\mathrm{I}} - \overset{o}{\chi}_{\mathrm{II}})^2$ – составляющая расстояния Махаланобиса, обусловленная пропущенным признаком $\overset{o}{x} \in \overset{o}{X}$ (при условии, что $\overset{o}{x}$ статистически не зависим с другими компонентами из $\overset{o}{X}$). С учетом этого, ограничившись точностью $(1/n_k)$ и пренебрегая членами порядка $(1/n_k^2)$, получаем

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

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

Величина $\Delta(V)$ может быть как положительной, так и отрицательной. Если величина $\Delta(V) > 0$, то признак $\overset{o}{x}$ необходимо включать в ДФ. Если величина $\Delta(V) < 0$, то признак $\overset{o}{x}$ не следует включать в ДФ, поскольку это приведет к уменьшению величины $D^2_{AB}$, т.е. добавление признака $\overset{o}{x} \in \overset{o}{X}$ не улучшает качество ДФ по рассматриваемому критерию.

Условие $\Delta(V) < 0$ является условием редукции (упрощения) ДФ, оптимальной по количеству и составу признаков. Оно представляет собой условие отрицательной определенности квадратичного трехчлена относительно $\gamma^2$ в фигурных скобках (10). Пороговым значением для $\gamma^2$, ниже которого возможна редукция ДФ, является значение:

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

На рисунке представлены зависимости порогового значения (11) от объема выборок $n$ для набора расстояний Махаланобиса $\tau^2_{\overset{o}{X}}$ $(\tau^2_{\overset{o}{X}} = 6,8,...,18)$ при $\overset{o}{m} = 6$.

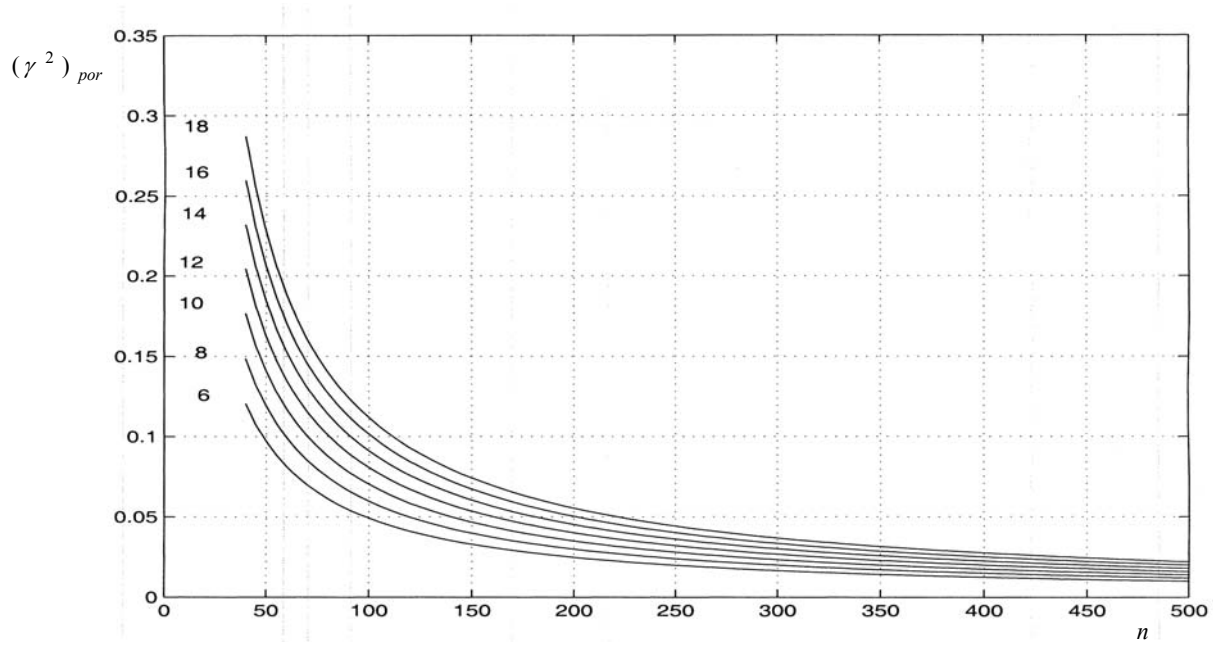


Рисунок 1 – Зависимости порогового значения $(\gamma^2)_{пор}$ от объема выборок $n$

Отметим, что в асимптотике при $n_A \to \infty$ $(r \to \infty)$ условие редукции не выполняется, поскольку $(\gamma^2)_{por} = 0$, а $\gamma^2 > 0$, т.е. $V_{OPT} = \overset{o}{X}$.

**3. Выводы**

Обоснован способ сравнения дискриминантных функций с разбиением выборок наблюдений на обучающие и проверочные подвыборки. Несмотря на успешное применение этого способа на практике и неоднократное подтверждение его работоспособности методом статистических испытаний, он традиционно считался эвристическим приём.

Получены условия существования оптимального множества признаков, зависящие от параметров генеральных совокупностей и объемов выборок. Выявлены закономерности упрощения оптимальной дискриминантной функции при уменьшении объемов выборок и при увеличении дисперсий признаков. Показано, что в условиях структурной неопределенности и отсутствия априорных оценок ковариационной матрицы признаков применение этого способа позволяет решать задачу поиска дискриминантной функции оптимальной сложности.

**Литература**

1. Сарычев А. П. Схема дискриминантного анализа с обучающими и проверочными подвыборками наблюдений / А. П. Сарычев // Автоматика. – 1990. – № 1. – С. 32–41.

2. Мирошниченко Л. В. Схема скользящего экзамена для поиска оптимального множества признаков в задаче дискриминантного анализа / Л. В. Мирошниченко, А. П. Сарычев // Автоматика. – 1992. – № 1. – С. 35–44.

3. Сарычев А. П. Решение задачи дискриминантного анализа в условиях структурной неопределенности на основе метода группового учета аргументов / А. П. Сарычев // Проблемы управления и информатики. – 2008. – № 3. – С. 100–112.

4. Сарычев А. П. Идентификация состояний структурно-неопределенных систем / А. П. Сарычев – Днепропетровск : Институт технической механики НАН Украины и НКА Украины, 2008. – 268 с.