УДК 519.25

А.П. Сарычев, Л.В. Сарычева

Решение задачи дискриминантного анализа на основе метода группового учета аргументов

Рассмотрена задача поиска оптимальной по сложности дискриминантной функции. Описаны критерии качества дискриминантных функций, разработанные в рамках метода группового учета аргументов: критерий, основанный на разбиении наблюдений на обучающую и проверочную выборки, и критерий скользящего экзамена.

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

Розглянуто задачу пошуку оптимальної за складністю дискримінантної функції. Описано критерії якості дискримінантних функцій, розроблених у межах методу групового урахування аргументів: критерій, заснований на розбивці спостережень на навчальну й перевірочну вибірки, і критерій ковзного іспиту.

Введение. Для решения задачи дискриминантного анализа, поставленной в условиях структурной неопределенности по составу признаков, необходимо принять способ сравнения дискриминантных функций, построенных на различных множествах признаков. Два способа сравнения популярны в приложениях. *Первый* основан на разбиении наблюдений на обучающие и проверочные подвыборки, когда обучающие подвыборки используются для оценивания коэффициентов дискриминантной функции, а проверочные – для оценивания ее качества классификации. *Второй* – известный 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 [1–3]. В рамках метода группового учета аргументов (МГУА) разработаны два показателя качества дискриминантных функций в задаче статистической классификации на основе дискриминантного анализа: критерий с разбиением наблюдений на обучающие и проверочные выборки и критерий скользящего контроля [4–6, 15–17].

Постановка задачи

Предположим, что

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

– выборка n_k независимых наблюдений m-мерного случайного вектора $\boldsymbol{\eta}_k$ из генеральной совокупности P_k, имеющего m-мерное нормальное распределение с неизвестным математическим ожиданием $\boldsymbol{\chi}_k$ и неизвестной невырожденной ковариационной матрицей $\sum_X$

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

где k – номер I или II генеральных совокупностей P_I и P_{II}, причем $\boldsymbol{\chi}_{\mathrm{I}} \neq \boldsymbol{\chi}_{\mathrm{II}}$.

Согласно предположению (2) для наблюдений (1) выполняется

$$\mathbf{X}_{ki} = \boldsymbol{\chi}_k + \boldsymbol{\xi}_{ki}, \; k = \mathrm{I}, \mathrm{II}, \; i = 1, 2, \ldots, n_k, \qquad (3)$$

где $\xi_{ki} \sim N_m(\mathbf{0}_m, \sum_X)$ – независимые случайные векторы, распределенные по m-мерному нормальному закону с нулевым математическим ожиданием и ковариационной матрицей $\sum_X$:

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

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

$$E\{\boldsymbol{\xi}_{\mathrm{I}i_1}\boldsymbol{\xi}_{\mathrm{II}i_2}^{\mathrm{T}}\} = \mathbf{O}_{(m\times m)}; \quad i_1, i_2 = 1, 2, \ldots, n_k, \qquad (6)$$

где $\mathbf{0}_m$ – нулевой m-мерный вектор–столбец; $\mathbf{O}_{(m\times m)}$ – нулевая $(m\times m)$-матрица; $\mathrm{E}\{\cdot\}$ – знак математического ожидания.

Будем считать, что априорные вероятности появления наблюдений из генеральных совокупностей P_I и P_{II} известны и соответственно равны π_I и π_{II}, причем $\pi_I + \pi_{II} = 1$. Предположим, что введены цены ошибочных классификаций: $c(I/II)$ – цена ошибочной классификации наблюдения из совокупности P_{II} в качестве наблюдения из P_{I}, а $c(\mathrm{II}/\mathrm{I})$ – цена ошибочной классификации наблюдения из совокупности P_{I} в качестве наблюдения из P_{II}. Правильная классификация не оценивается. Линейная дискриминантная функция (ДФ), для которой ожидаемая ошибка классификации минимальна, имеет вид

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

где параметр c_0 определяется ценами ошибочных классификаций и априорными вероятностями появления наблюдений

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

а коэффициенты ДФ δ определяются параметрами генеральных совокупностей

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

Решающее правило таково: если для наблюдения $\mathbf{x}^*$ выполняется $R(\mathbf{x}^*) \geq 0$, то $\mathbf{x}^* \in P_{\mathrm{I}}$ и относится к первой группе; если $R(\mathbf{x}^*) < 0$, то $\mathbf{x}^* \in P_{\mathrm{II}}$ и относится ко второй группе.

Получение по наблюдениям (1)–(3) оценок коэффициентов δ в дискриминантной функции (7) с последующим статистическим анализом есть задача дискриминантного анализа, поставленная в узком смысле. Фишеровской оценкой δ будет оценка

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

где $\tilde{\mathbf{x}}_{\mathrm{I}}$ и $\tilde{\mathbf{x}}_{\mathrm{II}}$ – оценки математических ожиданий $\boldsymbol{\chi}_{\mathrm{I}}$ и $\boldsymbol{\chi}_{\mathrm{II}}$

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

матрица $\mathbf{S}$ – оценка ковариационной матрицы Σ

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

а $\mathbf{x}_{\mathrm{I}}$ и $\mathbf{x}_{\mathrm{II}}$ – $(m \times n_{\mathrm{I}})$- и $(m \times n_{\mathrm{II}})$-матрицы отклонений наблюдений (1) и (3) от оценок $\tilde{\mathbf{x}}_{\mathrm{I}}$ и $\tilde{\mathbf{x}}_{\mathrm{II}}$:

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

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

Оценка (10) – решение задачи максимизации функционала

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

при ограничении

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

Значение функционала (15) при оптимальном значении $\hat{\mathbf{d}} = \mathbf{S}^{-1}(\tilde{\mathbf{x}}_{\mathrm{I}} - \tilde{\mathbf{x}}_{\mathrm{II}})$ –

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

есть выборочная оценка расстояния Махаланобиса [7, 8]

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

Для математического ожидания D^2 выполняется

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

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

Пусть X – множество m компонент векторов $\boldsymbol{\eta}_{\mathrm{I}}$ и $\boldsymbol{\eta}_{\mathrm{II}}$, над которыми проведены наблюдения; $\overset{\circ}{X}$ – множество $\overset{\circ}{m}$ ($1 \leq \overset{\circ}{m} < m$) компонент векторов $\boldsymbol{\eta}_{\mathrm{I}}$ и $\boldsymbol{\eta}_{\mathrm{II}}$, для математических ожиданий которых выполнено причем множество $\overset{\circ}{X} \subseteq X$ не $\overset{\circ}{\chi}_{\mathrm{I}j} \neq \overset{\circ}{\chi}_{\mathrm{II}j}$, $j = 1, 2, ..., \overset{\circ}{m}$, известно.

Если априори неизвестно, какие именно компоненты из X следует включать в ДФ, то говорят о задаче дискриминантного анализа, поставленной в широком смысле. В этом случае по наблюдениям $\mathbf{X}_{\mathrm{I}}$ и $\mathbf{X}_{\mathrm{II}}$ требуется определить множество компонент, которые необходимо включать в ДФ, и оценить коэффициенты линейной ДФ в пространстве этих компонент. Известно, что оценка расстояния Махаланобиса (17) не может быть использована для решения задачи дискриминантного анализа, поставленной в широком смысле, поскольку эта оценка увеличивается при добавлении новых компонент в ДФ (в решающее правило). Распространенный прием в этом случае – назначение «порога» для величины приращения оценки расстояния Махаланобиса (17) при добавлении новой компоненты. Но такое назначение привносит субъективность в получаемое решение.

Для решения задачи дискриминантного анализа в широком смысле кроме способа сравнения ДФ требуется указать алгоритм генерации различных сочетаний признаков, включаемых в ДФ. В рамках МГУА разработан итерационный алгоритм поиска [4]. Далее предположим, что в качестве такового принят полный перебор всех возможных сочетаний признаков.

Критерий для поиска оптимального множества признаков с разбиением наблюдений на обучающие и проверочные подвыборки

Пусть на этапе с номером s $(s = 1,2,..., m)$ алгоритма полного перебора сочетаний признаков в дискриминантную функцию может быть включено только s компонент из множества X, составляющих текущее анализируемое множество V. Пусть множеству компонент 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_{I} и P_{II};

2) $\boldsymbol{\nu}_{\mathrm{I}}$ и $\boldsymbol{\nu}_{\mathrm{II}}$ – $(s \times 1)$-векторы математических ожиданий для наблюдений из P_{I} и P_{II};

3) $\boldsymbol{\Sigma}_V$ – ковариационная $(s \times s)$-матрица.

Рассмотрим оценку расстояния Махаланобиса, рассчитываемую с учетом разбиения наблюдений на обучающие и проверочные подвыборки

$$D_{AB}^2(V) = \frac{\hat{\mathbf{d}}_A^{\mathrm{T}} (\tilde{\mathbf{v}}_{\mathrm{I}B} - \tilde{\mathbf{v}}_{\mathrm{II}B})(\tilde{\mathbf{v}}_{\mathrm{I}B} - \tilde{\mathbf{v}}_{\mathrm{II}B})^{\mathrm{T}} \hat{\mathbf{d}}_A}{\hat{\mathbf{d}}_A^{\mathrm{T}} \mathbf{S}_B \hat{\mathbf{d}}_A}. \quad (20)$$

В (20) $(s \times 1)$-вектор $\hat{\mathbf{d}}_A$ представляет собой рассчитанную на подвыборке A фишеровскую оценку коэффициентов ДФ, построенную в пространстве компонент множества V

$$\hat{\mathbf{d}}_A = \mathbf{S}_A^{-1}(\tilde{\mathbf{v}}_{\mathrm{I}A} - \tilde{\mathbf{v}}_{\mathrm{II}A}), \quad (21)$$

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

$$\tilde{\mathbf{v}}_{kA} = (n_{kA})^{-1} \sum_{i=1}^{n_{kA}} \mathbf{V}_{kiA}, \quad k = \mathrm{I}, \mathrm{II}; \quad (22)$$

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

$$\mathbf{S}_A = (n_{\mathrm{I}A} - n_{\mathrm{II}A} - 2)^{-1} [\mathbf{v}_{\mathrm{I}A} \mathbf{v}_{\mathrm{I}A}^{\mathrm{T}} + \mathbf{v}_{\mathrm{II}A} \mathbf{v}_{\mathrm{II}A}^{\mathrm{T}}]. \quad (23)$$

В (23) $\mathbf{v}_{kA}$ $(k = \mathrm{I}, \mathrm{II})$ – $(s \times n_k)$-матрицы, составленные из отклонений наблюдений $\mathbf{V}_{kA}$ компонент множества V от оценок $\tilde{\mathbf{v}}_{kA}$

$$\mathbf{v}_{kA} = [\mathbf{V}_{k1A} - \tilde{\mathbf{v}}_{kA}, \mathbf{V}_{k2A} - \tilde{\mathbf{v}}_{kA}, ..., \mathbf{V}_{kn_kA} - \tilde{\mathbf{v}}_{kA}]. \quad (24)$$

В (20) $(s \times 1)$-векторы $\tilde{\mathbf{v}}_{\mathrm{I}B}$ и $\tilde{\mathbf{v}}_{\mathrm{II}B}$ вычисляются аналогично (22), а $(s \times s)$-матрица $\mathbf{S}_B$ – аналогично (23), (24); $n_{\mathrm{I}A}$ 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 $n_{\mathrm{II}A} + n_{\mathrm{II}B} = n_{\mathrm{II}}$. Логика построения статистики $D_{AB}^2(V)$ полностью соответствует принципам МГУА при моделировании в классе регрессионных моделей: коэффициенты ДФ в пространстве компонент V оцениваются на обучающей выборке A, а ее качество – на проверочной выборке B.

Используя (21), для (20) получаем

$$D_{AB}^2(V) = \frac{(\tilde{\mathbf{v}}_{\mathrm{I}A} - \tilde{\mathbf{v}}_{\mathrm{II}A})^{\mathrm{T}} \mathbf{S}_A^{-1} (\tilde{\mathbf{v}}_{\mathrm{I}B} - \tilde{\mathbf{v}}_{\mathrm{II}B})(\tilde{\mathbf{v}}_{\mathrm{I}B} - \tilde{\mathbf{v}}_{\mathrm{II}B})^{\mathrm{T}} \mathbf{S}_A^{-1} (\tilde{\mathbf{v}}_{\mathrm{I}A} - \tilde{\mathbf{v}}_{\mathrm{II}A})}{(\tilde{\mathbf{v}}_{\mathrm{I}A} - \tilde{\mathbf{v}}_{\mathrm{II}A})^{\mathrm{T}} \mathbf{S}_A^{-1} \mathbf{S}_B \mathbf{S}_A^{-1} (\tilde{\mathbf{v}}_{\mathrm{I}A} - \tilde{\mathbf{v}}_{\mathrm{II}A})}. \quad (25)$$

Теорема 1. Для случайной величины $D_{AB}^2(V)$ выполняется

$$E\{D_{AB}^2(V)\} = \left(\tau_V^2 - \frac{\tau_V^2 [s - (r_1 - 1)/(r_1 - s)] c_A^{-1}}{\tau_V^2 + s c_A^{-1}} + c_B^{-1} \frac{r_1 - 1}{r_1 - s} \right) \frac{r_1 - s}{r_1 - 1}, \quad (26)$$

где

$$r_1 = n_{\mathrm{I}A} + n_{\mathrm{II}A} - 2, \quad c_A^{-1} = (n_{\mathrm{I}A}^{-1} + n_{\mathrm{II}A}^{-1}), \quad c_B^{-1} = (n_{\mathrm{I}B}^{-1} + n_{\mathrm{II}B}^{-1}), \quad (27)$$

а $\tau_V^2 = (\boldsymbol{\nu}_{\mathrm{I}} - \boldsymbol{\nu}_{\mathrm{II}})^{\mathrm{T}} \boldsymbol{\Sigma}_V^{-1} (\boldsymbol{\nu}_{\mathrm{I}} - \boldsymbol{\nu}_{\mathrm{II}})$ – расстояние Махаланобиса для множества компонент V.

Справедливость теоремы 1 следует из утверждений:

1. Оценки, полученные на подвыборке A, и оценки, полученные на подвыборке B, независимы.

2. Оценки математических ожиданий (22) и оценка ковариационной матрицы (23) независимы.

3. Операция вычисления математического ожидания и операция вычисления следа случайной матрицы $\mathbf{S}$ коммутируют, т.е. $E\{\mathrm{tr}[\mathbf{S}]\} = \mathrm{tr}[E\{\mathbf{S}\}]$.

4 Если $\mathbf{S}$ – случайная $(s \times s)$-матрица, имеющая распределение Уишарта с r степенями свободы – $\mathbf{S} \sim W_s(r, \mathbf{\Sigma})$ [7, 8], а $\mathbf{b}$ – независимый от $\mathbf{S}$ $(s \times 1)$-вектор, то для распределений квадратичных форм выполняется

$$\frac{\mathbf{b}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{b}}{(\mathbf{b}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{b})\cdot r} \sim (\chi^2(r-s+1))^{-1}, \qquad (28)$$

$$\frac{\mathbf{b}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{\Sigma}\,\mathbf{S}^{-1}\mathbf{b}}{(\mathbf{b}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{b})\cdot r^2} \sim \frac{1}{(\chi^2(r-s+1))^2} \times \times [1 + \frac{\chi^2(s-1)}{\chi^2(r-s+2)}], \qquad (29)$$

где $\chi^2(p)$ – случайная величина, имеющая χ^2-распределение с числом степеней свободы, равным p, а случайные величины $\chi^2(\cdot)$ взаимно независимы.

5. Если случайная величина $\chi^2(p)$ имеет χ^2-распределение с p степенями свободы, то

$$E\{(\chi^2(p))^h\} = p\cdot(p+2)\cdot\ldots \ldots\cdot[(p+2(h-1)], \quad h = 1, 2, \ldots; \qquad (30)$$

$$E\{(\chi^2(p))^{-h}\} = [(p-2)\cdot(p-4)\cdot\ldots \ldots\cdot(p-2h)]^{-1}, \quad h = 1, 2, \ldots (h < p/2). \qquad (31)$$

Утверждения 1 и 2 непосредственно следуют из предположений (4)–(6); утверждение 3 доказывается прямыми вычислениями; утверждение 4, формула (28) – известное свойство рассматриваемых квадратичных форм (например, [7, с. 148; 8, с. 482]); утверждение 4, формула (29) получены А.М. Шурыгиным [9] на основании теоремы А.Д. Деева [10]; утверждения 5, формулы (30) и (31) – известные свойства χ^2-распределения [9]. Из утверждений 1 и 2 следует, что числитель и знаменатель (25) статистически независимы. Для них, используя утверждения 1–5, получаем

$$E\Big\{(\tilde{\mathbf{v}}_{\mathrm{I}A} - \tilde{\mathbf{v}}_{\mathrm{II}A})^{\mathrm{T}}\mathbf{S}_A^{-1}(\tilde{\mathbf{v}}_{\mathrm{I}B} - \tilde{\mathbf{v}}_{\mathrm{II}B}) \times \times (\tilde{\mathbf{v}}_{\mathrm{I}B} - \tilde{\mathbf{v}}_{\mathrm{II}B})^{\mathrm{T}}\mathbf{S}_A^{-1}(\tilde{\mathbf{v}}_{\mathrm{I}A} - \tilde{\mathbf{v}}_{\mathrm{II}A})\Big\} =$$

$$= \left(\tau_V^2\right)^2 \frac{r^2}{(r-s-1)(r-s-3)} + + \left(c_A^{-1} + c_B^{-1}\right)\tau_V^2 \frac{r^2}{(r-s-1)(r-s-3)}\left(\frac{r-1}{r-s}\right) + + c_A^{-1}\, c_B^{-1} \frac{r^2\, s}{(r-s-1)(r-s-3)}\left(\frac{r-1}{r-s}\right), \qquad (32)$$

$$E\Big\{(\tilde{\mathbf{v}}_{\mathrm{I}A} - \tilde{\mathbf{v}}_{\mathrm{II}A})^{\mathrm{T}}\mathbf{S}_A^{-1}\,\mathbf{S}_B\,\mathbf{S}_A^{-1}(\tilde{\mathbf{v}}_{\mathrm{I}A} - \tilde{\mathbf{v}}_{\mathrm{II}A})\Big\} =$$

$$= \tau_V^2 \frac{r^2}{(r-s-1)(r-s-3)}\left(\frac{r-1}{r-s}\right) + + \frac{c_A^{-1} r^2\, s}{(r-s-1)(r-s-3)}\left(\frac{r-1}{r-s}\right). \qquad (33)$$

Учитывая полученные математические ожидания (32) и (33), окончательно получаем (36).

Критерий скользящего экзамена для поиска оптимального множества признаков

Традиционный способ скользящего экзамена состоит в следующем:

- одно из наблюдений исключается из обучающей выборки;
- это наблюдение классифицируется на основе ДФ, построенной на выборке обучения без учета исключенного наблюдения;
- наблюдение возвращается в выборку. Процедура с исключением повторяется для второго наблюдения, третьего и, до тех пор, пока все наблюдения будут классифицированы таким способом. Обычно в приложениях оценивается вероятность ошибочной классификации, т.е. подсчитывается число ошибочно классифицированных наблюдений. В предлагаемом способе вычисляется расстояние

$$D_S^2(V) = \frac{1}{2}(D_{S\mathrm{I}}^2(V) + D_{S\mathrm{II}}^2(V)), \qquad (34)$$

$$D_{S\mathrm{I}}^2(V) = (n_{\mathrm{I}})^{-1}\sum_{i=1}^{n_{\mathrm{I}}} \frac{\mathbf{d}_{(\mathrm{I},i)}^{\mathrm{T}}(\mathbf{v}_{\mathrm{I}i} - \tilde{\mathbf{v}}_{\mathrm{II}})(\mathbf{v}_{\mathrm{I}i} - \tilde{\mathbf{v}}_{\mathrm{II}})^{\mathrm{T}}\mathbf{d}_{(\mathrm{I},i)}}{\mathbf{d}_{(\mathrm{I},i)}^{\mathrm{T}}\mathbf{S}_{(\mathrm{I},i)}\mathbf{d}_{(\mathrm{I},i)}} = = (n_{\mathrm{I}})^{-1}\sum_{i=1}^{n_{\mathrm{I}}} (\mathbf{v}_{\mathrm{I}i} - \tilde{\mathbf{v}}_{\mathrm{II}})^{\mathrm{T}}\mathbf{W}_{(\mathrm{I},i)}(\mathbf{v}_{\mathrm{I}i} - \tilde{\mathbf{v}}_{\mathrm{II}}), \qquad (35)$$

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

$$=(n_{\mathrm{II}})^{-1}\sum_{j=1}^{n_{\mathrm{II}}}(\tilde{\mathbf{v}}_{\mathrm{I}}-\mathbf{v}_{\mathrm{II}j})^{\mathrm{T}}\mathbf{W}_{(\mathrm{II},j)}(\tilde{\mathbf{v}}_{\mathrm{I}}-\mathbf{v}_{\mathrm{II}j}), \quad (36)$$

$$\mathbf{W}_{(\mathrm{I},i)}=\frac{\mathbf{d}_{(\mathrm{I},i)}\mathbf{d}_{(\mathrm{I},i)}^{\mathrm{T}}}{\mathbf{d}_{(\mathrm{I},i)}^{\mathrm{T}}\mathbf{S}_{(\mathrm{I},i)}\mathbf{d}_{(\mathrm{I},i)}};$$

$$\mathbf{W}_{(\mathrm{II},j)}=\frac{\mathbf{d}_{(\mathrm{II},j)}\mathbf{d}_{(\mathrm{II},j)}^{\mathrm{T}}}{\mathbf{d}_{(\mathrm{II},j)}^{\mathrm{T}}\mathbf{S}_{(\mathrm{II},j)}\mathbf{d}_{(\mathrm{II},j)}}. \quad (37)$$

В формуле (35) вектор $\mathbf{d}_{(\mathrm{I},i)}$ – оценка коэффициентов фишеровской ДФ, рассчитанная без наблюдения с номером i из первой группы наблюдений

$$\mathbf{d}_{(\mathrm{I},i)}=\mathbf{S}_{(\mathrm{I},i)}^{-1}(\tilde{\mathbf{v}}_{\mathrm{I}(i)}-\tilde{\mathbf{v}}_{\mathrm{II}}), \quad (38)$$

где вектор $\tilde{\mathbf{v}}_{\mathrm{I}(i)}$ – оценка математического ожидания $\boldsymbol{\nu}_{\mathrm{I}}$

$$\tilde{\mathbf{v}}_{\mathrm{I}(i)}=(n_{\mathrm{I}}-1)^{-1}\Big(\sum_{h=1}^{n_{\mathrm{I}}}\mathbf{v}_{\mathrm{I}h}-\mathbf{v}_{\mathrm{I}i}\Big); \quad (39)$$

вектор $\tilde{\mathbf{v}}_{\mathrm{II}}$ – оценка математического ожидания $\boldsymbol{\nu}_{\mathrm{II}}$

$$\tilde{\mathbf{v}}_{\mathrm{II}}=(n_{\mathrm{II}})^{-1}\sum_{h=1}^{n_{\mathrm{II}}}\mathbf{v}_{\mathrm{II}h}; \quad (40)$$

матрица $\mathbf{S}_{(\mathrm{I},i)}$ – несмещенная оценка ковариационной матрицы $\boldsymbol{\Sigma}_V$

$$\mathbf{S}_{(\mathrm{I},i)}=(n_{\mathrm{I}}+n_{\mathrm{II}}-3)^{-1}\left[\sum_{\substack{h=1\\(h\neq i)}}^{n_{\mathrm{I}}}\tilde{\tilde{\mathbf{v}}}_{\mathrm{I}h(i)}\tilde{\tilde{\mathbf{v}}}_{\mathrm{I}h(i)}^{\mathrm{T}}+\sum_{q=1}^{n_{\mathrm{II}}}\tilde{\tilde{\mathbf{v}}}_{\mathrm{II}q}\tilde{\tilde{\mathbf{v}}}_{\mathrm{II}q}^{\mathrm{T}}\right], \quad (41)$$

где $\tilde{\tilde{\mathbf{v}}}_{\mathrm{I}h(i)}$ – наблюдение с номером h, центрированное относительно оценки $\tilde{\mathbf{v}}_{\mathrm{I}(i)}$

$$\tilde{\tilde{\mathbf{v}}}_{\mathrm{I}h(i)}=\mathbf{v}_{\mathrm{I}h}-\tilde{\mathbf{v}}_{\mathrm{I}(i)},\ h=1,2,...,\ n_{\mathrm{I}}\ (h\neq i), \quad (42)$$

а $\tilde{\tilde{\mathbf{v}}}_{\mathrm{II}q(j)}$ – наблюдение с номером q, центрированное относительно оценки $\tilde{\mathbf{v}}_{\mathrm{II}}$

$$\tilde{\tilde{\mathbf{v}}}_{\mathrm{II}q}=\mathbf{v}_{\mathrm{I}q}-\tilde{\mathbf{v}}_{\mathrm{II}},\ h=1,2,...,n_{\mathrm{II}}. \quad (43)$$

В формуле (36) вектор $\mathbf{d}_{(\mathrm{I},j)}$ – оценка коэффициентов фишеровской ДФ, рассчитанная без наблюдения с номером j из второй группы наблюдений

$$\mathbf{d}_{(\mathrm{II},j)}=\mathbf{S}_{(\mathrm{II},j)}^{-1}(\tilde{\mathbf{v}}_{\mathrm{I}}-\tilde{\mathbf{v}}_{\mathrm{II}(j)}), \quad (44)$$

где вектор $\tilde{\mathbf{v}}_{\mathrm{I}}$ – оценка математического ожидания $\boldsymbol{\nu}_{\mathrm{I}}$, вычисляемая аналогично (40); вектор $\tilde{\mathbf{v}}_{\mathrm{II}(j)}$ – оценка математического ожидания $\boldsymbol{\nu}_{\mathrm{II}}$, вычисляемая аналогично (39); матрица $\mathbf{S}_{(\mathrm{II},j)}$ – несмещенная оценка ковариационной матрицы Σ_V, вычисляемая аналогично (41).

Из выражений (34)–(44) следует, что статистика $D_{S\mathrm{I}}^2(V)$ есть не что иное, как взвешенная сумма парных расстояний между наблюдениями первой группы и оценкой математического ожидания $\boldsymbol{\nu}_{\mathrm{II}}$ второй группы, а статистика $D_{S\mathrm{II}}^2(V)$ – взвешенная сумма парных расстояний между наблюдениями второй группы и оценкой математического ожидания $\mathbf{v}_{\mathrm{I}}$ первой группы.

Используя (38) и (44), получаем

$$D_{\mathrm{SI}}^2(V)=(n_{\mathrm{I}})^{-1}\sum_{i=1}^{n_{\mathrm{I}}}\frac{[(\mathbf{v}_{\mathrm{I}i}-\tilde{\mathbf{v}}_{\mathrm{II}})^{\mathrm{T}}\mathbf{S}_{(\mathrm{I},i)}^{-1}(\tilde{\mathbf{v}}_{\mathrm{I}(i)}-\tilde{\mathbf{v}}_{\mathrm{II}})]^2}{(\tilde{\mathbf{v}}_{\mathrm{I}(i)}-\tilde{\mathbf{v}}_{\mathrm{II}})^{\mathrm{T}}\mathbf{S}_{(\mathrm{I},i)}^{-1}(\tilde{\mathbf{v}}_{\mathrm{I}(i)}-\tilde{\mathbf{v}}_{\mathrm{II}})}, \quad (45)$$

$$D_{\mathrm{SII}}^2(V)=(n_{\mathrm{II}})^{-1}\sum_{j=1}^{n_{\mathrm{II}}}\frac{[(\tilde{\mathbf{v}}_{\mathrm{I}}-\mathbf{v}_{\mathrm{II}j})^{\mathrm{T}}\mathbf{S}_{(\mathrm{II},j)}^{-1}(\tilde{\mathbf{v}}_{\mathrm{I}}-\tilde{\mathbf{v}}_{\mathrm{II}(j)})]^2}{(\tilde{\mathbf{v}}_{\mathrm{I}}-\tilde{\mathbf{v}}_{\mathrm{II}(j)})^{\mathrm{T}}\mathbf{S}_{(\mathrm{II},j)}^{-1}(\tilde{\mathbf{v}}_{\mathrm{I}}-\tilde{\mathbf{v}}_{\mathrm{II}(j)})}. \quad (46)$$

Для упрощения дальнейшего анализа будем полагать: $n_{\mathrm{I}}=n_{\mathrm{II}}=n$. С учетом этого введем обозначения: $r_3=n_{\mathrm{I}}+n_{\mathrm{II}}-3=2n-3$ и $c_3^{-1}=n^{-1}+$ $+(n-1)^{-1}$.

Теорема 2. Для случайной величины $D_S^2(V)$ выполняется

$$E\{D_S^2(V)\}=\left(\tau_V^2-\frac{\tau_V^2\,[s-(r_3-1)/(r_3-s)]\,c_3^{-1}}{(\tau_V^2+s\,c_3^{-1})}-\right.$$
$$\left.-\frac{(n+1)\,(r_3-1)}{n(r_3-s)}\right)\frac{r_3}{r_3-s-3}, \quad (47)$$

где

$$r_3=n_{\mathrm{I}}+n_{\mathrm{II}}-3=2n-3,\quad c_3^{-1}=n^{-1}+(n-1)^{-1}, \quad (48)$$

а $\tau_V^2=(\mathbf{v}_{\mathrm{I}}-\mathbf{v}_{\mathrm{II}})^{\mathrm{T}}\boldsymbol{\Sigma}_V^{-1}(\mathbf{v}_{\mathrm{I}}-\mathbf{v}_{\mathrm{II}})$ – расстояние Махаланобиса для множества компонент V.

Справедливость теоремы 2 следует из того, что:

- наблюдение $\mathbf{v}_{\mathrm{I}i}$, $\tilde{\mathbf{v}}_{\mathrm{I}}$ – оценка математического ожидания (40) и $\mathbf{S}_{(\mathrm{I},i)}$ – оценка ковариационной матрицы (41) независимы;

• наблюдение $\mathbf{v}_{\mathrm{II}j}$, $\tilde{\mathbf{v}}_{\mathrm{II}}$ – оценка математического ожидания и $\mathbf{S}_{(\mathrm{II},j)}$ независимы (эти утверждения следуют из предположений (4)–(6));

• справедливы утверждения 3–5.

Дискриминантная функция, оптимальная по составу признаков

В соответствии с принципами МГУА для обоснования критерия качества моделей необходимо:

• вычислить математическое ожидание критерия для заданной структуры модели;

• исследовать поведение математического ожидания критерия в зависимости от перебираемых структур моделей;

• показать существование модели оптимальной сложности;

• получить условие редукции (упрощения) модели оптимальной сложности.

Отметим, что в рассматриваемом классе линейных ДФ структура модели (дискриминантной функции) однозначно определяется числом и составом признаков, включенных в дискриминантную функцию.

Пусть $D^2(V)$ – один из критериев МГУА, разработанных для решения задачи дискриминантного анализа, поставленной в широком смысле.

Определение 1. Оптимальным множеством компонент (признаков) называется множество

$$V_{\mathrm{opt}} = \arg\max_{V \subseteq X} E\{D^2(V)\}. \qquad (49)$$

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

В работах [4–6, 15–17] проведено исследование критериев, описанных в статье ранее, и показано, что они позволяют решать задачу поиска оптимального множества признаков в задаче дискриминантного анализа, поставленной в широком смысле.

Множество компонент X может быть разбито на непересекающиеся подмножества $X = \overset{o}{X} \bigcup \overset{o}{R} \bigcup \tilde{R} = \overset{o}{V} \bigcup \tilde{R}$:

1) $\overset{o}{X} \neq \varnothing$ ($\varnothing$ – пустое множество) – множество компонент ($\overset{o}{m}$ – их число), для математических ожиданий которых выполнено $\overset{o}{\chi}_{\mathrm{I}h} \neq \overset{o}{\chi}_{\mathrm{II}h}$, $h = 1,2,...,\overset{o}{m}$;

2) $\overset{o}{R}$ – множество компонент, для математических ожиданий которых выполнено $\overset{o}{\rho}_{\mathrm{I}h} = \overset{o}{\rho}_{\mathrm{II}h}$, $h = 1,2,...,\overset{o}{l}$, 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) $\tilde{R}$ – множество компонент, для математических ожиданий которых выполнено $\tilde{\rho}_{\mathrm{I}h} = \tilde{\rho}_{\mathrm{II}h}$, $h = 1, 2,...,\tilde{l}$, где $\tilde{l}$ – их число, и каждая компонента из множества $\tilde{R}$ статистически не зависит от любой из компонент множества $\overset{o}{X}$ (множество $\tilde{R}$ может быть пустым).

Для исследования разработанных критериев в работах [6, 15] 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 произвольных анализируемых множеств компонент $V \subseteq X$. Для случая известных параметров генеральных совокупностей P_{I} и P_{II} из сформулированных лемм следует:

1) любая компонента из $\overset{o}{X}$ необходима в том смысле, что ее включение в текущее множество компонент V увеличивает расстояние Махаланобиса τ_V^2;

2) любая компонента из $\tilde{R}$ избыточна в том смысле, что ее включение в текущее множество V не увеличивает расстояния Махаланобиса τ_V^2;

3) любая компонента из $\overset{o}{R}$ необходима, поскольку ее включение в V увеличивает расстояние Махаланобиса τ_V^2.

В практических приложениях параметры генеральных совокупностей, как правило, неизвестны, но могут быть получены как статистические оценки по обучающим выборкам наблюдений конечного объема. Известно [2, 3], что если применить построенное правило классификации к обучающей выборке, то оценка качества распознавания будет завышена по математическому ожиданию в сравнении с той же оценкой качества на независимых от обучения данных. Способы с разбиением наблюдений на обучающие и проверочные подвыборки и скользящего экзамена не завышают оценки качества распознавания. Опыт практических применений и тестовые исследования этих способов методом статистических испытаний [1–3] показывают, что в них:

1) с увеличением объема выборок наблюдений увеличивается количество компонент во множестве, на котором достигается наилучшее качество распознавания, а с уменьшением объема выборок наблюдений количество компонент в таком множестве уменьшается;

2) с увеличением расстояния Махаланобиса τ_X^2 между генеральными совокупностями (из которых получены выборки наблюдений) увеличивается количество компонент во множестве, на котором достигается наилучшее качество распознавания, а с уменьшением этого расстояния количество компонент в таком множестве уменьшается. Аналитические исследования разработанных критериев МГУА, проведенные в работах [4–6, 15–17], подтверждают эти эмпирически установленные закономерности о существовании ДФ, оптимальной по количеству и составу компонент, и дают условия редукции (упрощения) оптимальной дискриминантной функции при уменьшении объемов выборок и при изменении параметров генеральных совокупностей.

Условие редукции оптимальной дискриминантной функции в способе с разбиением на обучающие и проверочные выборки

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). С учетом (26) получаем

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

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

$$\left. + c_B^{-1}\, \frac{r-1}{r-\overset{o}{m}} \right) \frac{r-\overset{o}{m}}{r-1} -$$

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

$$\left. + c_B^{-1}\, \frac{r-1}{r-\overset{o}{m}+1} \right) \frac{r-\overset{o}{m}+1}{r-1}. \qquad (50)$$

В соответствии с упомянутыми леммами для расстояний Махаланобиса множеств V и $\overset{o}{X}$ выполняется соотношение $\tau_V^2 = \tau_{\overset{o}{X}}^2 - \gamma^2$, где $\gamma^2 = \sigma_{\overset{o}{x}}^{-2} (\overset{o}{\chi}_{\mathrm{I}} - \overset{o}{\chi}_{\mathrm{II}})^2$ – составляющая расстояния Махаланобиса, обусловленная пропущенным признаком $\overset{o}{x} \in \overset{o}{X}$ (при условии, что $\overset{o}{x}$ статистически независим от других компонентов из $\overset{o}{X}$). С учетом этого, ограничившись точностью $(1/n_k)$ и пренебрегая членами порядка $(1/n_k^2)$, получаем

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

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

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

$$-\left(\tau^2_{\overset{\circ}{X}}\right)^2\cdot\left(\tau^2_{\overset{\circ}{X}}\cdot\frac{1}{r-1}+\frac{r-\overset{\circ}{m}}{r-1}\cdot c_A^{-1}\right)\Bigg\}. \qquad (51)$$

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

Условие $\Delta(V)<0$ есть условием редукции (упрощения) ДФ, оптимальной по количеству и составу признаков. Оно представляет собой условие отрицательной определенности квадратичного трехчлена относительно γ^2 в фигурных скобках (51). Пороговым значением для γ^2, ниже которого возможна редукция ДФ, будет значение

$$(\gamma^2)_{\text{пор}}=\tau^2_{\overset{\circ}{X}}\cdot\frac{\left(\dfrac{\tau^2_{\overset{\circ}{X}}}{r-1}\right)+c_A^{-1}}{\tau^2_{\overset{\circ}{X}}\left(\dfrac{r-\overset{\circ}{m}+1}{r-1}\right)+\overset{\circ}{m}\,c_A^{-1}}. \qquad (52)$$

На рис. 1 представлены зависимости порогового значения (52) от объема выборок n при фиксированном значении $\overset{\circ}{m}=6$ для набора расстояний Махаланобиса $\tau^2_{\overset{\circ}{X}}$ ($\tau^2_{\overset{\circ}{X}}=6,8,\ldots,18$).

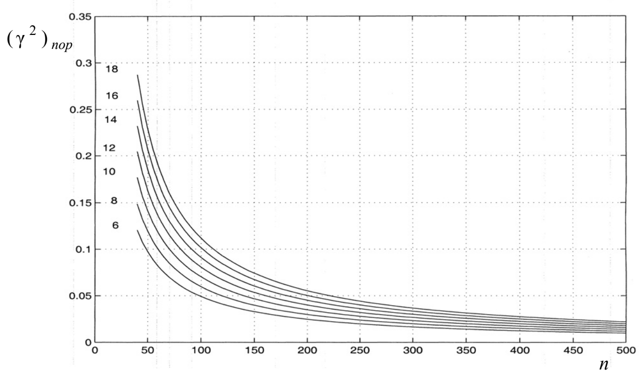

Рис. 1. Зависимости порогового значения $(\gamma^2)_{\text{пор}}$ от объема выборок n

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

Условие редукции оптимальной дискриминантной функции в способе скользящего экзамена

Сформулируем условие редукции (упрощения) оптимальной ДФ для частного случая независимого признака. Пусть множество V таково, что выполняется $\overset{\circ}{X}=V\bigcup\overset{\circ}{x}$, где $\overset{\circ}{x}\in\overset{\circ}{X}$ (в ДФ пропущен один признак). С учетом (47) получаем

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

$$=\left(\tau^2_{\overset{\circ}{X}}-\frac{\tau^2_{\overset{\circ}{X}}\,[\overset{\circ}{m}-(r-1)/(r-\overset{\circ}{m})]\,c^{-1}}{\tau^2_{\overset{\circ}{X}}+\overset{\circ}{m}\,c^{-1}}-\frac{(n+1)(r-1)}{n(r-\overset{\circ}{m})}\right)\times$$

$$\times\frac{r}{r-\overset{\circ}{m}-3}-\left(\tau_V^2-\frac{\tau_V^2\,[(\overset{\circ}{m}-1)-(r-1)/(r-\overset{\circ}{m}+1)]\,c^{-1}}{\tau_V^2+(\overset{\circ}{m}-1)\,c^{-1}}-\right.$$

$$\left.-\frac{(n+1)(r-1)}{n(r-\overset{\circ}{m}+1)}\right)\frac{r}{r-\overset{\circ}{m}-2}. \qquad (53)$$

В соответствии с упомянутыми леммами для расстояний Махаланобиса множеств V и $\overset{\circ}{X}$ выполняется соотношение $\tau_V^2=\tau^2_{\overset{\circ}{X}}-\gamma^2$, где $\gamma^2=\sigma^{-2}_{\overset{\circ}{x}}(\overset{\circ}{\chi}_{\text{I}}-\overset{\circ}{\chi}_{\text{II}})^2$ – составляющая расстояния Махаланобиса, обусловленная пропущенным независимым признаком $\overset{\circ}{x}\in\overset{\circ}{X}$. С учетом этого, ограничившись точностью $(1/n)$, пренебрегая членами порядка $(1/n^2)$, получаем

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

$$\times\left\{-\left(\tau^2_{\overset{\circ}{X}}+\overset{\circ}{m}\,c^{-1}\right)\cdot\left(\gamma^2\right)^2+\right.$$

$$+\tau^2_{\overset{\circ}{X}}\cdot\left(\tau^2_{\overset{\circ}{X}}\cdot\frac{r-\overset{\circ}{m}-4}{r-\overset{\circ}{m}-3}+2\overset{\circ}{m}\,c^{-1}+\frac{2}{r-\overset{\circ}{m}-3}\right)\cdot\gamma^2-$$

$$-\left(\tau^2_{\overset{\circ}{X}}\right)^2\cdot\left(\tau^2_{\overset{\circ}{X}}\cdot\frac{1}{r-\overset{\circ}{m}-3}+c^{-1}\right)\Bigg\}. \qquad (54)$$

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

Условие $\Delta(V)<0$ есть условием редукции (упрощения) ДФ, оптимальной по количеству и составу признаков. Это условие представляет собой условие отрицательной определенности квадратичного трехчлена относительно γ^2 в фигурных скобках (54). Пороговым значением для γ^2, ниже которого возможна редукция ДФ, будет значение

$$(\gamma^2)_{\text{пор}}=\tau^2_{\overset{\circ}{X}}\cdot\frac{\left(\dfrac{\tau^2_{\overset{\circ}{X}}}{r-\overset{\circ}{m}-3}\right)+c^{-1}}{\tau^2_{\overset{\circ}{X}}+\overset{\circ}{m}\,c^{-1}}. \qquad (55)$$

На рис. 2 представлены зависимости порогового значения (55) от объема выборок n при

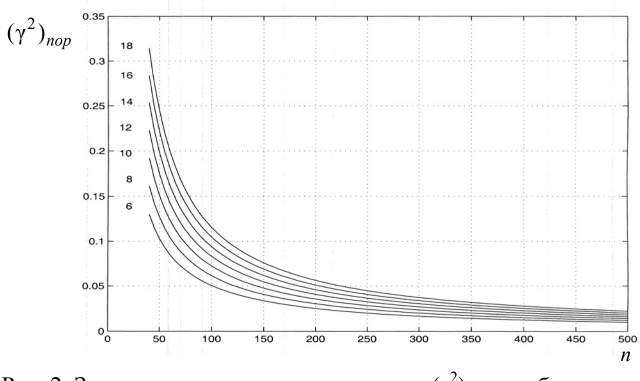

Рис. 2. Зависимости порогового значения $(\gamma^2)_{\text{пор}}$ от объема выборок n

фиксированном значении $\overset{\circ}{m}=6$ для набора расстояний Махаланобиса $\tau^2_{\overset{\circ}{X}}$ $(\tau^2_{\overset{\circ}{X}}=6,8,...,18)$. Отметим, что в асимптотике при $n\to\infty$ $(r\to\infty,\ c^{-1}\to 0)$ условие редукции не выполняется, т.е. $V_{OPT}=\overset{\circ}{X}$.

Заключение. Приведены результаты исследований двух способов сравнения дискриминантных функций в рамках МГУА: способ с разбиением выборок наблюдений на обучающие и проверочные подвыборки и способ скользящего экзамена. Несмотря на успешное применение этих способов на практике и неоднократное подтверждение их эффективности методом статистических испытаний, они традиционно считались эвристическими приемами. Получены условия существования оптимального множества признаков, зависящие от параметров генеральных совокупностей и объемов выборок, и выявлены закономерности упрощения оптимальной дискриминантной функции при уменьшении объемов выборок и при изменении параметров генеральных совокупностей. В условиях структурной неопределенности и отсутствия априорных оценок дисперсий и ковариаций признаков исследованные способы позволяют решать задачу поиска дискриминантной функции оптимальной сложности. Разработанные в рамках МГУА критерии и алгоритмы поиска оптимального множества признаков в задаче дискриминантного анализа в условиях структурной неопределенности применены для решения ряда практических задач. Задачи распознавания образов в условиях структурной неопределенности рассматривались А.Г. Ивахненко еще в 60–70-х годах прошлого века как актуальные задачи технической кибернетики.

1. *Лбов Г.С.* Выбор эффективной системы зависимых признаков // Вычислительные системы. – 1965. – **19**. – С. 21–34.
2. *Вапник В.Н.*, *Червоненкис А.Я.* Теория распознавания образов. – М.: Наука, 1974. – 416 с.
3. *Раудис Ш.Ю.* Ограниченность выборки в задачах классификации // Статистические проблемы управления. – 1976. – **18**. – С. 6–183.
4. *Сарычев А.П.* Итерационный алгоритм МГУА для синтеза разделяющей функции в задаче дискриминантного анализа // Автоматика. – 1988. – № 2. – С. 20–24.
5. *Сарычев А.П.* Схема дискриминантного анализа с обучающими и проверочными подвыборками наблюдений. – Там же. – 1990. – № 1. – С. 32–41.

6. *Мирошниченко Л.В.*, *Сарычев А.П.* Схема скользящего экзамена для поиска оптимального множества признаков в задаче дискриминантного анализа // Там же. – 1992. – № 1. – С. 35–44.
7. *Андерсон Т.* Введение в многомерный статистический анализ. – М.: Физматгиз, 1963. – 500 с.
8. *Рао С.Р.* Линейные статистические методы и их применения. – М.: Наука, 1968. – 548 с.
9. *Шурыгин А.М.* Линейная комбинация простейшего и фишеровского дискриминантов // Ученые записки по статистике. Т. 45. – М.: Наука, 1983. – С. 144–159.
10. *Деев А.Д.* Асимптотические разложения распределений статистик дискриминантного анализа W, M, W^* // Статистические методы классификации. – М.: Изд-во МГУ, 1972. – **31**. – С. 6–51.
11. *Ивахненко А.Г.*, *Сарычев А.П.*, *Аралбаева Г.Г.* Прогнозирование технологических нарушений в производстве алюминия на основе МГУА // Комплексная автоматизация промышленности. Ч. 1.: Третья Польско-советская науч.-техн. конф., 11–14 окт. 1988 г., Вроцлав, Польша: доклады конф. – Вроцлав: Изд-во Вроцлавского политех. ин-та, 1988. – С. 37–47.
12. *Опыт* решения задачи прогноза солнечной активности при точностном и робастном подходах / А.Г. Ивахненко, А.П. Сарычев, П.И. Залевский и др. // Автоматика. – 1988. – № 3. – С. 31–42.
13. *Непараметрические* прогнозирующие модели МГУА. Ч. 3. Модели на языке паттерн- и кластер-анализа для прогнозирования процессов в экономических макросистемах / А.Г. Ивахненко, Н. А. Ивахненко, Ю.В. Костенко и др. // Там же. – 1989. – № 3. – С. 3–17.
14. *Система* многоальтернативного распознавания типов взаимодействия нейронов / А.П. Сарычев, А.Г. Ивахненко, Г.А. Ивахненко и др. // Штучний інтелект. – 2001. – № 2. – С. 66–73.
15. *Сарычев А.П.* Идентификация состояний структурно-неопределенных систем. – Днепропетровск: Ин-т техн. механики НАН Украины и НКА Украины, 2008. – 268 с.
16. *Сарычев А.П.*, *Сарычева Л.В.* Оценивание качества дискриминантных функций на основе скользящего экзамена // Искусственный интеллект. – 2011. – № 3. – С. 271–277.
17. *Сарычев А.П.*, *Сарычева Л.В.* Оценивание качества дискриминантных функций с разбиением наблюдений на обучающие и проверочные подвыборки // Індуктивне моделювання складних систем. – 2011. – **3**. – С. 209–215.
18. *Ивахненко А.Г.* Самообучающиеся системы распознавания и автоматического управления. – Киев: Техніка, 1969. – 392 с.
19. *Ивахненко А.Г.* Системы эвристической самоорганизации в технической кибернетике. – Там же, 1971. – 364 с.
20. *Перцептрон* – система распознавания образов / Под ред. А.Г. Ивахненко. – Киев: Наук. думка, 1975. – 430 с.

Тел. для справок: +38 0562 46-5149, 46-9919,
+38 063 737-0061 (Днепропетровск)
E-mail: *Sarychev@prognoz.dp.ua*, *Sarycheval@nmu.org.ua*

●