УДК 519.25

ИТЕРАЦИОННЫЙ АЛГОРИТМ МГУА ДЛЯ РЕШЕНИЯ ЗАДАЧИ НЕЛИНЕЙНОГО ДИСКРИМИНАНТНОГО АНАЛИЗА

А.П. Сарычев[1], Л.В. Сарычева[2]

[1] *Институт технической механики НАНУ и НКАУ, г. Днепропетровск*
[2] *Национальный горный университет МОН Украины, г. Днепропетровск*
Sarychev@prognoz.dp.ua, Sarycheval@nmu.dp.ua

Розглянуто задачу статистичної класифікації станів складних систем на основі нелінійного дискримінантного аналізу в умовах невизначеності за складом ознак. Побудовано новий ітераційний алгоритм МГУА для розв'язання задачі нелінійного дискримінантного аналізу.
Ключові слова: метод групового урахування аргументів, критерій якості нелінійної дискримінантної функції, ітераційний алгоритм МГУА.

The task of statistical classification of complex systems states is considered on the basis of the nonlinear discriminant analysis in conditions of uncertainty on features structure. New iterative algorithm GMDH for the decision of a task of the nonlinear discriminant analysis is constructed.
Keywords: Group Method of Data Handling, criterion of nonlinear discriminant functions quality, iteration algorithm of GMDH.

Рассмотрена задача статистической классификации состояний сложных систем на основе нелинейного дискриминантного анализа в условиях неопределенности по составу признаков. Построен новый итерационный алгоритм МГУА для решения задачи нелинейного дискриминантного анализа.
Ключевые слова: метод группового учета аргументов, критерий качества нелинейной дискриминантной функции, итерационный алгоритм МГУА.

Введение

Статистические методы математического моделирования сложных систем, для которых отсутствуют точные априорные гипотезы, являются наиболее распространенным современным инструментом научных и инженерных исследований с целью описания и прогнозирования состояния таких систем.

Возможным подходом к описанию состояния исследуемого объекта является отнесение анализируемого состояния к известному классу состояний. Этот подход характерен, как правило, для задач диагностики. Распространенным классом моделей статистической классификации на основе дискриминантного анализа является класс дискриминантных функций (ДФ), линейных по параметрам. Однако, в задачах статистической классификации, которые возникают в диагностике, предположение линейного дискриминантного анализа о равенстве ковариационных матриц двух классов состояний, как правило, не выполняется. Действительно, в диагностике обычно речь идет о классах состояний, которые принципиально отличаются друг от друга, и это отличие проявляется не только в неравенстве математических ожиданий признаков, но и в неравенстве ковариационных матриц классов.

Задача описания и прогнозирования состояний объекта диагностики как задача статистической классификации может характеризоваться структурной неопределенностью по количеству и составу признаков, которые необходимо включать в классификационное правило. Для решения задачи нелинейного дискриминантного анализа в условиях этой неопределенности требуется указать способ сравнения ДФ и алгоритм генерации различных сочетаний признаков, включаемых в ДФ. В статье описан алгоритм решения задачи нелинейного дискриминантного анализа, который построен на основе метода группового учета аргументов [1–2] и относится к классу итерационных алгоритмов МГУА [3–5].

Способ сравнения ДФ, основанный на разбиении наблюдений на обучающие и проверочные подвыборки, является популярным в приложениях. Обучающие подвыборки используются для оценивания коэффициентов ДФ, а проверочные – для оценивания ее качества классификации. В литературе этот способ традиционно трактуется как эвристический прием, хотя факт существования оптимального множества признаков неоднократно подтвержден методом статистических испытаний [6–11] и обоснован аналитически [12–17].

1. Постановка задачи статистической классификации на основе линейного дискриминантного анализа в условиях неопределенности по составу признаков

Предположим, что

$$\mathbf{X}(k) = [\,\mathbf{x}_1(k)\;\; \mathbf{x}_2(k)\;\; \ldots\;\; \mathbf{x}_i(k)\;\; \ldots\;\; \mathbf{x}_{n(k)}(k)\,] \tag{1}$$

– выборка $n(k)$ независимых наблюдений m-мерного случайного вектора $\mathbf{з}(k)$ из генеральной совокупности $P(k)$, имеющего m-мерное нормальное распределение с неизвестным математическим ожиданием $\mathbf{ч}(k)$ и неизвестной невырожденной ковариационной матрицей $\mathbf{У}_X$:

$$\mathbf{з}(k) \sim N_m(\mathbf{ч}(k), \mathbf{У}_X), \tag{2}$$

где k – номер генеральных совокупностей $P(I)$ и $P(II)$, причем $\mathbf{ч}(I) \neq \mathbf{ч}(II)$.

Согласно предположению (2) для наблюдений (1) выполняется

$$\mathbf{x}_i(k) = \mathbf{ч}(k) + \mathbf{o}_i(k),\;\; k = I, II,\;\; i = 1,\, 2, \ldots, n(k), \tag{3}$$

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

$$E\{\mathbf{o}_i(k)\} = \mathbf{0}_m;\;\; E\{\mathbf{o}_i(k)\mathbf{o}_i^{\mathrm{T}}(k)\} = \mathbf{У}_X;\;\; i = 1,\, 2, \ldots, n(k);\;\; k = I, II; \tag{4}$$

$$E\{\mathbf{o}_{i_1}(k)\mathbf{o}_{i_2}^{\mathrm{T}}(k)\} = \mathbf{O}_{(m\times m)};\;\; i_1, i_2 = 1,\, 2, \ldots, n(k);\;\; i_1 \neq i_2; \tag{5}$$

$$E\{\mathbf{o}_{i_1}(I)\mathbf{o}_{i_2}^{\mathrm{T}}(II)\} = \mathbf{O}_{(m\times m)};\ \ i_1 = 1,\ 2,..., n(I);\ \ i_2 = 1,\ 2,..., n(II), \tag{6}$$

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

Будем считать, что априорные вероятности появления наблюдений $p(I)$ и $p(II)$ из совокупностей $P(I)$ и $P(II)$ известны, причем $p(I)+p(II)=1$.

Предположим, что введены цены ошибочных классификаций: $c(I/II)$ означает цену ошибочной классификации наблюдения из совокупности $P(II)$ в качестве наблюдения из $P(I)$, а $c(II/I)$ – цену ошибочной классификации наблюдения из совокупности $P(I)$ в качестве наблюдения из $P(II)$. Правильная классификация не оценивается.

Линейная ДФ, при которой в указанных предположениях ожидаемая ошибка классификации минимальна, имеет вид

$$R(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{д} - 0{,}5(\mathbf{ч}(I) + \mathbf{ч}(II))^{\mathrm{T}}\mathbf{д} - \ln c_0, \tag{7}$$

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

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

а коэффициенты ДФ $\mathbf{д}$ определяются параметрами генеральных совокупностей

$$\mathbf{д} = \mathbf{У}_X^{-1}(\mathbf{ч}_{\mathrm{I}} - \mathbf{ч}_{\mathrm{II}}). \tag{9}$$

Решающее правило для наблюдения $\mathbf{x}^*$ имеет вид:

$$\text{если } R(\mathbf{x}^*) \ge 0, \text{ то } \mathbf{x}^* \in P(I); \tag{10}$$

$$\text{если } R(\mathbf{x}^*) < 0, \text{ то } \mathbf{x}^* \in P(II). \tag{11}$$

Получение по выборочным наблюдениям (1)–(3) оценок коэффициентов $\mathbf{д}$ в дискриминантной функции (7) с последующим статистическим анализом является задачей дискриминантного анализа, поставленной в узком смысле.

Фишеровской оценкой $\mathbf{д}$ является оценка

$$\hat{\mathbf{d}} = \mathbf{S}^{-1}(\bar{\mathbf{x}}(I) - \bar{\mathbf{x}}(II)), \tag{12}$$

где $\bar{\mathbf{x}}(I)$ и $\bar{\mathbf{x}}(II)$ – оценки математических ожиданий $\mathbf{ч}(I)$ и $\mathbf{ч}(II)$:

$$\bar{\mathbf{x}}(k) = n^{-1}(k)\sum_{i=1}^{n(k)}\mathbf{x}_i(k),\ \ k = I, II; \tag{13}$$

матрица $\mathbf{S}$ – оценка ковариационной матрицы $\mathbf{У}_X$

$$\mathbf{S} = (n(I) + n(II) - 2)^{-1}[\tilde{\mathbf{X}}(I)\tilde{\mathbf{X}}^{\mathrm{T}}(I) + \tilde{\mathbf{X}}(II)\tilde{\mathbf{X}}^{\mathrm{T}}(II)], \tag{14}$$

а $\tilde{\mathbf{X}}(I)$ и $\tilde{\mathbf{X}}(II)$ – $(m \times n(I))$- и $(m \times n(II))$-матрицы, составленные из отклонений наблюдений (1) и (3) от оценок $\bar{\mathbf{x}}(I)$ и $\bar{\mathbf{x}}(II)$ соответственно

$$\tilde{\mathbf{X}}(I) = [\mathbf{x}_1(I) - \bar{\mathbf{x}}(I), \mathbf{x}_2(I) - \bar{\mathbf{x}}(I), ..., \mathbf{x}_{n(I)}(I) - \bar{\mathbf{x}}(I)], \tag{15}$$

$$\tilde{\mathbf{X}}(II) = [\mathbf{x}_1(II) - \bar{\mathbf{x}}(II), \mathbf{x}_2(II) - \bar{\mathbf{x}}(II), ..., \mathbf{x}_{n(II)}(II) - \bar{\mathbf{x}}(II)]. \tag{16}$$

Оценка (12) является решением задачи максимизации функционала

$$\Phi(\mathbf{d}) = \frac{\mathbf{d}^{\mathrm{T}}(\bar{\mathbf{x}}(I) - \bar{\mathbf{x}}(II))(\bar{\mathbf{x}}(I) - \bar{\mathbf{x}}(II))^{\mathrm{T}}\mathbf{d}}{\mathbf{d}^{\mathrm{T}}\mathbf{S}\mathbf{d}} \to \max, \tag{17}$$

при ограничении

$$\frac{(\bar{\mathbf{x}}(I) - \bar{\mathbf{x}}(II))^{\mathrm{T}}\mathbf{d}}{\mathbf{d}^{\mathrm{T}}\mathbf{S}\mathbf{d}} = 1. \tag{18}$$

Значение функционала (17) при оптимальных оценках (12)

$$D^2 = (\bar{\mathbf{x}}(I) - \bar{\mathbf{x}}(II))^{\mathrm{T}}\mathbf{S}^{-1}(\bar{\mathbf{x}}(I) - \bar{\mathbf{x}}(II)) \tag{19}$$

является выборочной оценкой расстояния Махаланобиса

$$\phi_X^2 = (\mathbf{ч}_{\mathrm{I}} - \mathbf{ч}_{\mathrm{II}})^{\mathrm{T}}\mathbf{У}^{-1}(\mathbf{ч}_{\mathrm{I}} - \mathbf{ч}_{\mathrm{II}}). \tag{20}$$

Для математического ожидания величины D^2 выполняется:

$$E\{D^2\} = \frac{r}{r - m - 1}\tau_X^2 + mc^{-1}, \tag{21}$$

где $r = n(I) + n(II) - 2$; $c^{-1} = n^{-1}(I) + n^{-1}(II)$.

Пусть X – множество m компонент векторов $\mathbf{з}(I)$ и $\mathbf{з}(II)$, над которыми проведены наблюдения. Если априорно неизвестно, какие именно компоненты из множества *X* следует включать в ДФ, то говорят о задаче дискриминантного анализа, поставленной в широком смысле. В этом случае по наблюдениям $\mathbf{X}(I)$ и $\mathbf{X}(II)$ требуется определить множество компонент, которые необходимо включать в линейную ДФ, и оценить ее коэффициенты.

2. Классификационное правило и критерий качества классификации в нелинейном дискриминантном анализе

Постановка задачи дискриминантного анализа в широком смысле для двух классов с неравными ковариационными матрицами совпадает с постановкой задачи линейного дискриминантного анализа (1)–(21), но вместо

общей ковариационной матрицы $\mathbf{У}_X$ имеются две разные ковариационные матрицы $\mathbf{У}_X(I)$ и $\mathbf{У}_X(II)$, где I и II – номера классов.

В МГУА поиск ДФ оптимальной сложности предполагает: а) указание способа оценивания коэффициентов ДФ; б) указание способа генерирования структур ДФ; в) указание внешнего критерия для оценки качества перебираемых структур; г) исследование поведения математического ожидания внешнего критерия в зависимости от состава признаков в ДФ; д) доказательство существования ДФ оптимальной сложности. Пункты *в, г* и *д* реализованы в [16, 17], где разработан и исследован критерий качества ДФ, основанный на разбиении наблюдений на обучающие и проверочные подвыборки в соответствии с принципами МГУА. Классификационное правило, введенное в [16, 17] для решения задачи дискриминантного анализа для двух классов с неравными ковариационными матрицами, имеет следующий вид:

$$\text{если } h(\mathbf{x}^*) \le 0, \text{ то } \mathbf{x}^* \in P(I); \tag{22}$$

$$\text{если } h(\mathbf{x}^*) > 0, \text{ то } \mathbf{x}^* \in P(II), \tag{23}$$

где

$$h(\mathbf{x}^*) = \frac{1}{2}(\mathbf{x}^* - \bar{\mathbf{x}}(I))^{\mathrm{T}}\mathbf{S}^{-1}(I)(\mathbf{x}^* - \bar{\mathbf{x}}(I)) -$$

$$-\frac{1}{2}(\mathbf{x}^* - \bar{\mathbf{x}}(II))^{\mathrm{T}}\mathbf{S}^{-1}(II)(\mathbf{x}^* - \bar{\mathbf{x}}(II)) +$$

$$+\frac{1}{2}\ln\left(\frac{\det(\mathbf{S}(I))}{\det(\mathbf{S}(II))}\right) - \ln\frac{c(II/I)\pi(I)}{c(I/II)\pi(II)}; \tag{24}$$

$\mathbf{x}^*$ – анализируемое наблюдение; векторы $\bar{\mathbf{x}}(I)$ и $\bar{\mathbf{x}}(II)$ – оценки математических ожиданий $\mathbf{ч}(I)$ и $\mathbf{ч}(II)$

$$\bar{\mathbf{x}}(k) = (n(k))^{-1}\sum_{i=1}^{n(k)}\mathbf{x}_i(k), \quad k = I, II; \tag{25}$$

матрицы $\mathbf{S}(I)$ и $\mathbf{S}(II)$ – оценки ковариационных матриц $\mathbf{У}_X(I)$ и $\mathbf{У}_X(II)$

$$\mathbf{S}(k) = (n(k)-1)^{-1}[\tilde{\mathbf{X}}(k)\tilde{\mathbf{X}}^{\mathrm{T}}(k)], \quad k = I, II, \tag{26}$$

а $\tilde{\mathbf{X}}(I)$ и $\tilde{\mathbf{X}}(II)$ – $(m \times n(I))$- и $(m \times n(II))$-матрицы, составленные из отклонений наблюдений (1) и (3) от оценок $\bar{\mathbf{x}}(I)$ и $\bar{\mathbf{x}}(II)$ соответственно:

$$\tilde{\mathbf{X}}(k) = [\mathbf{x}_1(k) - \bar{\mathbf{x}}(k), \mathbf{x}_2(k) - \bar{\mathbf{x}}(k), \ldots, \mathbf{x}_{n(k)}(k) - \bar{\mathbf{x}}(k)], \quad k = I, II. \tag{27}$$

Реализуем схему расчета классификационных чисел с разбиением данных наблюдений на обучающие и проверочные подвыборки, а именно: классификационные числа для наблюдений проверочных подвыборок B будем рассчитывать по правилу (22)–(24), где используем оценки математических ожиданий и ковариационных матриц, рассчитанные по формулам (25)–(27) по наблюдениям обучающих подвыборок A.

Разность средних значений классификационных чисел (24) для наблюдений проверочных подвыборок B первого и второго класса

$$D_{BA}(X) = \bar{h}_B(II, X) - \bar{h}_B(I, X), \tag{28}$$

$$\bar{h}_B(k, X) = \frac{1}{n_B(k)} \sum_{i=1}^{n_B(k)} h_{iB}(k, X), \quad k = I, II, \tag{29}$$

и является критерием качества нелинейных дискриминантных функций [16, 17].

3. Итерационный алгоритм МГУА нелинейного дискриминантного анализа

В практических приложениях при генерации наборов признаков применяется перебор всех возможных их сочетаний. Как правило, реализации такого способа полного перебора представляют собой многоэтапные процедуры, в которых на каждом этапе с номером s в ДФ допускается не более s признаков. Генерация наборов признаков начинается с отдельно взятых признаков на этапе с номером $s = 1$ и заканчивается на этапе s_{opt}, после которого попытка усложнить ДФ по числу включенных в нее признаков не приводит к улучшению качества решения. Для многих практических приложений полный перебор на современных ЭВМ является реализуемым: общее число вариантов равно $\sum_{s=1}^{s_{opt}} C_m^s$, где $C_m^s = \frac{m!\, s!}{(m-s)!}$ – число сочетаний из m элементов по s элементов. Если же число признаков в исходном множестве достаточно велико, то алгоритм полного перебора заменяют эвристическими процедурами, которые реализуют так называемые "целенаправленные" способы перебора признаков [11]. Применению эвристических процедур предшествует их исследование методом статистических испытаний. Представителем такого класса эвристических процедур является итерационный алгоритм МГУА, описание которого проводится в данной работе.

Запишем классификационное правило (22)–(24) в виде, удобном для реализации разработанной итерационной процедуры. Введем

$$\hat{\mathbf{d}}(I) = \mathbf{S}^{-1}(I)(\bar{\mathbf{x}}(I) - \bar{\mathbf{x}}(II)), \tag{30}$$

$$\hat{\mathbf{d}}(II) = \mathbf{S}^{-1}(II)(\bar{\mathbf{x}}(II) - \bar{\mathbf{x}}(I)) \tag{31}$$

– коэффициенты двух линейных ДФ, каждую из которых можно было бы использовать для решения задачи классификации при условии, что ковариационные матрицы первого и второго класса одинаковы и соответственно равны $\mathbf{У}_X(I)$ в первом случае (30) и $\mathbf{У}_X(II)$ во втором случае (31).

С учетом (30)–(31) в пространстве признаков, образующих множество X, правило классификации (24) для наблюдения $\mathbf{x}^* = (x_1^*, x_2^*, ..., x_m^*)^{\mathrm{T}}$ имеет вид:

$$\text{если } h(\mathbf{x}^*) \le 0, \text{ то } \mathbf{x}^* \in P(I); \tag{32}$$

$$\text{если } h(\mathbf{x}^*) > 0, \text{ то } \mathbf{x}^* \in P(II), \tag{33}$$

где

$$h(\mathbf{x}^*) = -\frac{1}{2}(\mathbf{x}^*)^{\mathrm{T}}\hat{\mathbf{d}}(I) + \frac{1}{2}(\mathbf{x}^*)^{\mathrm{T}}\hat{\mathbf{d}}(II) + d_0(\mathbf{x}^*); \tag{34}$$

$$d_0(\mathbf{x}^*) = \frac{1}{2}(\bar{\mathbf{x}}(I))^{\mathrm{T}}\hat{\mathbf{d}}(I) - \frac{1}{2}(\bar{\mathbf{x}}(II))^{\mathrm{T}}\hat{\mathbf{d}}(II) +$$

$$+\frac{1}{2}(\mathbf{x}^* - \bar{\mathbf{x}}(I))^{\mathrm{T}}\left[\mathbf{S}^{-1}(I) - \mathbf{S}^{-1}(II)\right](\mathbf{x}^* - \bar{\mathbf{x}}(II)) +$$

$$+\frac{1}{2}\ln\left(\frac{\det[\mathbf{S}(I)]}{\det[\mathbf{S}(II)]}\right) - \ln\frac{c(II/I)\pi(I)}{c(I/II)\pi(II)}. \tag{35}$$

Алгоритм позволяет строить первую $\mathbf{x}^{\mathrm{T}}\mathbf{d}(I)$ и вторую ДФ $\mathbf{x}^{\mathrm{T}}\mathbf{d}(II)$ в классе моделей

$$\sum_{q=1}^{p} d_q \prod_{j=1}^{m} x_j^{\alpha(q,j)} + d_0(\mathbf{x}) = 0, \tag{36}$$

где $\mathbf{x} = (x_1, x_2, ..., x_m)^{\mathrm{T}} \in X$ – анализируемый $(m \times 1)$-вектор значений признаков; $X = \{x_1, x_2, ...x_j, ..., x_m\}$ – исходное множество m признаков; j – номер признака; $q = 1,2,..., p$ – номер одночлена в модели; p – число членов в модели; d_q – коэффициенты при одночленах; $\alpha(q, j)$ – целые неотрицательные числа, определяющие структуру одночлена с номером q: число $\alpha(q, j)$ означает степень, в которой признак x_j входит в одночлен с номером q; коэффициент $d_0(\mathbf{x})$ рассчитывается аналогично коэффициенту $d_0(\mathbf{x}^*)$ в (35).

Для упрощения изложения примем, что признаки из множества X могут участвовать в (36) под знаком суммы только линейно, и поэтому для каждого j

сумма $\alpha(q, j)$ по всем q может принимать только два значения: 0 или 1.

Дадим необходимые определения.

Классификационным числом наблюдения $\mathbf{x} = (x_1, x_2, ..., x_m)^{\mathrm{T}}$ для дискриминантных функций $\mathbf{x}^{\mathrm{T}}\hat{\mathbf{d}}(I)$ и $\mathbf{x}^{\mathrm{T}}\hat{\mathbf{d}}(II)$, построенных в пространстве множества признаков X, называется скаляр

$$z(\mathbf{x}) = -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\hat{\mathbf{d}}(I) + \frac{1}{2}\mathbf{x}^{\mathrm{T}}\hat{\mathbf{d}}(II), \qquad (37)$$

где $\hat{\mathbf{d}}(I)$ и $\hat{\mathbf{d}}(II)$ – оценки коэффициентов ДФ (30)–(31).

Введем обозначения $\bar{z}(I)$ и $\bar{z}(II)$ для средних значений классификационных чисел наблюдений, принадлежащих $P(I)$ и $P(II)$ соответственно:

$$\bar{z}(k) = (n(k))^{-1}\sum_{i=1}^{n(k)} z_i(k), \quad k = I, II. \qquad (38)$$

Аналогично классификационные числа наблюдений определяются и для ДФ, построенных в пространстве любого множества признаков $V \subset X$.

Классификационное число z_i принадлежит к k-й $(k = I, II)$ совокупности, если наблюдение i принадлежит к этой совокупности.

Вектором классификационных чисел назовем n-мерный вектор $\mathbf{z} = (z_1, z_2, ..., z_n)^{\mathrm{T}}$, образованный классификационными числами всех n наблюдений.

Структурой вектора классификационных чисел $\mathbf{z}$ назовем набор параметров p и $\alpha(q, j)$, определяющий $\mathbf{z}$ в представлении (36).

Разобьем исходные выборки наблюдений из первой (I) и второй (II) совокупностей на обучающие (A) и проверочные (B) подвыборки. Сгруппируем их в $(m \times n)$-матрице всех наблюдений $\mathbf{X}$ следующим образом:

$$\mathbf{X}^{\mathrm{T}} = \begin{bmatrix} \mathbf{X}_A^{\mathrm{T}} \\ \mathbf{X}_B^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_A^{\mathrm{T}}(I) \\ \mathbf{X}_A^{\mathrm{T}}(II) \\ \mathbf{X}_B^{\mathrm{T}}(I) \\ \mathbf{X}_B^{\mathrm{T}}(II) \end{bmatrix}, \qquad (39)$$

где $\mathbf{X}_A^{\mathrm{T}}$ – $(n_A \times m)$-матрица наблюдений обучающей подвыборки; $\mathbf{X}_B^{\mathrm{T}}$ – $(n_B \times m)$-матрица наблюдений проверочной подвыборки; n_A и n_B – объемы обучающей и проверочной подвыборок, $n_A + n_B = n$; $\mathbf{X}_A^{\mathrm{T}}(k)$ – $(n_A(k) \times m)$-матрица наблюдений обучающей подвыборки из совокупности $P(k)$,

$n_A(I)+n_A(II)=n_A$; $\mathbf{X}_B^{\mathrm{T}}$ – $(n_B(k)\times m)$-матрица наблюдений проверочной подвыборки из совокупности $P(k)$, $n_B(I)+n_B(II)=n_B$, $n_A+n_B=n$.

Построение итерационного алгоритма решения задачи нелинейного дискриминантного анализа, поставленной в широком смысле, выполняется в соответствии с принципами, разработанными в рамках МГУА [1–5]. В нашем случае при построении итерационного алгоритма МГУА для синтеза двух ДФ оптимальной сложности необходимо: 1) указать начальную матрицу классификационных чисел $\mathbf{Z}_0$; 2) определить оператор $\Re$, осуществляющий отображение $\mathbf{Z}_{r-1}\to\mathbf{Z}_r$, r – номер итерации; 3) указать правило остановки.

1. Прежде чем указать начальные матрицы, определим общий вид матриц классификационных чисел на итерации r:

$$\mathbf{Z}_r=[\mathbf{z}_1^r\ \mathbf{z}_2^r\dots\ \mathbf{z}_{F+2+m+s}^r]=\left[\begin{array}{c|c|c|c}\mathbf{z}_{1IA}^r & \mathbf{z}_{2IA}^r & \cdots & \mathbf{z}_{F+2+m+s,IA}^r\\ \hline \mathbf{z}_{1IIA}^r & \mathbf{z}_{2IIA}^r & \cdots & \mathbf{z}_{F+2+m+s,IIA}^r\\ \hline \mathbf{z}_{1IB}^r & \mathbf{z}_{2IB}^r & \cdots & \mathbf{z}_{F+2+m+s,IB}^r\\ \hline \mathbf{z}_{1IIB}^r & \mathbf{z}_{2IIB}^r & \cdots & \mathbf{z}_{F+2+m+s,IIB}^r\end{array}\right], \tag{40}$$

где $\mathbf{z}_v^r$ $(v=1,2,\dots,F+2+m+s)$ – n-мерные векторы; n – общее число наблюдений; F – число лучших по критерию отбора ДФ (см. ниже), передаваемых от итерации к итерации; m – число признаков во множестве X; s – число членов в структуре наилучших (по критерию отбора) ДФ, полученных на итерации $r-1$.

Для удобства дальнейшего изложения разобьем матрицу $\mathbf{Z}_r$ на три части:

$$\mathbf{Z}_r=[\mathbf{G}_r\ \vdots\ \mathbf{C}_r\ \vdots\ \mathbf{D}_r], \tag{41}$$

$$\mathbf{G}_r=[\mathbf{z}_1^r\ \mathbf{z}_2^r\dots\mathbf{z}_F^r], \tag{42}$$

$$\mathbf{C}_r=[\mathbf{z}_{F+1}^r\ \mathbf{z}_{F+2}^r\dots\mathbf{z}_{F+2+m}^r], \tag{43}$$

$$\mathbf{D}_r=[\mathbf{z}_{F+2+m+1}^r\ \mathbf{z}_{F+2+m+2}^r\dots\mathbf{z}_{F+2+m+s}^r]. \tag{44}$$

Укажем теперь начальные матрицы классификационных чисел (для $r=0$ определим $s=0$):

$$\mathbf{Z}_0=[\mathbf{G}_0\ \vdots\ \mathbf{C}_0],\ \mathbf{G}_0=\mathbf{O}_{(n\times F)},\ \mathbf{C}_0=\left[\mathbf{0}_n\ \vdots\ \mathbf{1}_n\ \vdots\ \mathbf{X}^{\mathrm{T}}\right], \tag{45}$$

где $\mathbf{O}_{(n\times F)}$ – нулевая $(n\times F)$-матрица; $\mathbf{0}_n$ – нулевой $(n\times 1)$-вектор; $\mathbf{1}_n$ – единичный $(n\times 1)$-вектор; $\mathbf{X}=[\mathbf{x}_1\ \mathbf{x}_2\dots\mathbf{x}_n]$ – $(m\times n)$-матрица наблюдений исходного множества признаков X, введенная в (39).

2. Определим оператор $\Re$. Пусть $(n\times 1)$-вектор классификационных чисел $\mathbf{z}^r$ определяется по правилу

$$z^r(i) = -\frac{1}{2}\left(\hat{a}(I)\, z_v^{r-1}(i) + \hat{b}(I)\, z_u^{r-1}(i)\right) + \frac{1}{2}\left(\hat{a}(II)\, z_v^{r-1}(i) + \hat{b}(II)\, z_u^{r-1}(i)\right), \qquad (46)$$

где $v, u = 1,2,...,F+2+m+2s$ $(v<u)$; $i = 1,2,\ldots,n$ – номер наблюдения.

Коэффициенты $\hat{a}(I)$, $\hat{b}(I)$ и $\hat{a}(II)$, $\hat{b}(II)$ определим как нормированные оценки

$$\hat{a}(k) = \tilde{a}(k)/(\tilde{a}^{\,2}(k) + \tilde{b}^{\,2}(k))^{1/2}, \quad \hat{b}(k) = \tilde{b}(k)/(\tilde{a}^{\,2}(k) + \tilde{b}^{\,2}(k))^{1/2}, \qquad (47)$$

$$\begin{pmatrix} \tilde{a}(k) \\ \tilde{b}(k) \end{pmatrix} = \mathbf{S}_A^{-1}(k) \begin{pmatrix} \bar{z}_{vIA}^{\,r-1} - \bar{z}_{vIIA}^{\,r-1} \\ \bar{z}_{uIA}^{\,r-1} - \bar{z}_{uIIA}^{\,r-1} \end{pmatrix}, \quad k = I, II. \qquad (48)$$

В (48) оценки коэффициентов ДФ $\tilde{a}(k)$ и $\tilde{b}(k)$ $(k = I, II)$ получены в результате максимизации функционала вида (17) при ограничении вида (18) на обучающей подвыборке A классификационных чисел $z_{vIA}^{r-1}(i)$, $z_{uIA}^{r-1}(i)$, $i = 1,2,...,n_A(I)$ и $z_{vIIA}^{r-1}(i)$, $z_{uIIA}^{r-1}(i)$, $i = 1,2,...,n_A(II)$:

$$\tilde{a}(k), \tilde{b}(k) = \arg\max_{a,b} \Phi_A(k), \qquad (49)$$

$$\Phi_A(k) = \frac{\begin{pmatrix} a \\ b \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \bar{z}_{vIA}^{\,r-1} - \bar{z}_{vIIA}^{\,r-1} \\ \bar{z}_{uIA}^{\,r-1} - \bar{z}_{uIIA}^{\,r-1} \end{pmatrix} \begin{pmatrix} \bar{z}_{vIA}^{\,r-1} - \bar{z}_{vIIA}^{\,r-1} \\ \bar{z}_{uIA}^{\,r-1} - \bar{z}_{uIIA}^{\,r-1} \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} a \\ b \end{pmatrix}}{\begin{pmatrix} a \\ b \end{pmatrix}^{\mathrm{T}} \mathbf{S}_A(k) \begin{pmatrix} a \\ b \end{pmatrix}}, \qquad (50)$$

где в записи скаляра $\bar{z}_{vIA}^{\,r-1}$ индексы внизу указывают: v – номер вектора классификационных чисел в матрице $\mathbf{Z}_{r-1}$, I – принадлежность к первой совокупности наблюдений, A – принадлежность к подвыборке A; черта сверху означает усреднение, например, $\bar{z}_{uIIA}^{\,r-1}$ – среднее значение классификационных чисел вектора $\mathbf{z}_{uIIA}^{r-1}$, рассчитанное по числам из второй (II) совокупности подвыборки A

$$\bar{z}_{uIIA}^{\,r-1} = \frac{1}{n_A(II)} \sum_{i=1}^{n_A(II)} z_{uIIA}^{\,r-1}(i); \qquad (51)$$

$\mathbf{S}_A(k)$ – оценка ковариационной матрицы, рассчитанная по классификационным числам подвыборки A:

$$\mathbf{S}_A(k)=\frac{1}{n_A(k)-1}\left[\begin{array}{c|c}(\tilde{\mathbf{z}}^{\,r-1}_{vkA})^{\mathrm{T}}\tilde{\mathbf{z}}^{\,r-1}_{vkA} & (\tilde{\mathbf{z}}^{\,r-1}_{ukA})^{\mathrm{T}}\tilde{\mathbf{z}}^{\,r-1}_{vkA}\\ \hline (\tilde{\mathbf{z}}^{\,r-1}_{vkA})^{\mathrm{T}}\tilde{\mathbf{z}}^{\,r-1}_{ukA} & (\tilde{\mathbf{z}}^{\,r-1}_{ukA})^{\mathrm{T}}\tilde{\mathbf{z}}^{\,r-1}_{ukA}\end{array}\right]; \tag{52}$$

волной в (52) отмечено центрирование относительно среднего значения, например: $\tilde{z}^{\,r-1}_{ukA}(i)=z^{\,r-1}_{ukA}(i)-\bar{z}^{\,r-1}_{ukA}$, $i=1,2,...,n_A(k)$, $k=I, II$.

Пусть $V=\{v_1,v_2,...,v_s\}$ – множество признаков (s – их число), вошедших в структуру классификационных векторов $\mathbf{z}^r$ в представлении (36)–(37), а $\mathbf{V}$ – $(s\times n)$-матрица наблюдений этих признаков.

Классификационное правило для наблюдений $\mathbf{v}_i$, $i=1,2,...,n$ на основе пары ДФ формулируется аналогично (32)–(35):

$$\text{если } h(\mathbf{v}_i)\le 0, \text{ то } \mathbf{v}_i\in P(I); \tag{53}$$

$$\text{если } h(\mathbf{v}_i)>0, \text{ то } \mathbf{v}_i\in P(II); \tag{54}$$

$$h(\mathbf{v}_i)=z^r(i)+d_0(\mathbf{v}_i),\quad i=1,2,...,n, \tag{55}$$

где $z^r(i)$ – классификационное число, определенное в (46);

$$d_0(\mathbf{v}_i)=\frac{1}{2}\left(\bar{z}^{\,r}_{vIA}\ \bar{z}^{\,r}_{uIA}\right)\left(\begin{array}{c}\hat{a}(I)\\ \hline \hat{b}(I)\end{array}\right)-\frac{1}{2}\left(\bar{z}^{\,r}_{vIIA}\ \bar{z}^{\,r}_{uIIA}\right)\left(\begin{array}{c}\hat{a}(II)\\ \hline \hat{b}(II)\end{array}\right)+$$

$$+\frac{1}{2}\left(z^{\,r-1}_{v}(i)-\bar{z}^{\,r-1}_{vIA},\ z^{\,r-1}_{u}(i)-\bar{z}^{\,r-1}_{uIA}\right)\left[\mathbf{S}_A^{-1}(I)-\mathbf{S}_A^{-1}(II)\right]\left(z^{\,r-1}_{v}(i)-\bar{z}^{\,r-1}_{vIIA},\ z^{\,r-1}_{u}(i)-\bar{z}^{\,r-1}_{uIIA}\right)+$$

$$+\frac{1}{2}\ln\left(\frac{\det[\mathbf{S}_A(I)]}{\det[\mathbf{S}_A(II)]}\right)-\ln\frac{c(II/I)\pi(I)}{c(I/II)\pi(II)}. \tag{56}$$

В соответствии с (39) для $\mathbf{h}=(h_1,h_2,...,h_n)^{\mathrm{T}}$, где $h_i=h(\mathbf{v}_i)$ в (55), справедливо разбиение

$$\mathbf{h}=\left(\begin{array}{c}\mathbf{h}_A\\ \hline \mathbf{h}_B\end{array}\right)=\left(\begin{array}{c}\mathbf{h}_A(I)\\ \hline \mathbf{h}_A(II)\\ \hline \mathbf{h}_B(I)\\ \hline \mathbf{h}_B(II)\end{array}\right). \tag{57}$$

Из всех векторов классификационных чисел $z^r(i)$, $i=1,2,...,n$, генерируемых по правилу (46)–(52), отберем F лучших в смысле качества классификации на подвыборке B, которое оценим разностью

$$CR = \bar{h}_B(II) - \bar{h}_B(I), \tag{58}$$

$$\bar{h}_B(k) = (n_B(k))^{-1} \sum_{i=1}^{n_B(k)} h_{iB}(k), \quad k = I, II. \tag{59}$$

Этот набор F лучших пар векторов классификационных чисел, упорядоченных по возрастанию критерия (58) так, что выполняется

$$CR(\mathbf{z}_1^r) \le CR(\mathbf{z}_2^r) \le \ldots \le CR(\mathbf{z}_F^r), \tag{60}$$

образует матрицу $\mathbf{G}_r$, а матрица $\mathbf{C}_r$ остается без изменений:

$$\mathbf{G}_r = [\mathbf{z}_1^r \ \mathbf{z}_2^r \ \ldots \mathbf{z}_F^r], \tag{61}$$

$$\mathbf{C}_r = \mathbf{C}_0 = \left[\mathbf{0}_n \,\vdots\, \mathbf{1}_n \,\vdots\, \mathbf{X}^{\mathrm{T}}\right], \tag{62}$$

Столбцы матрицы $\mathbf{D}_r$ формируются исключением одночленов из структуры лучшего вектора классификационных чисел по правилу

$$\mathbf{D}_r = [\mathbf{z}_{F+2+m+1}^r \ \mathbf{z}_{F+2+m+2}^r \ \ldots \mathbf{z}_{F+2+m+s}^r], \tag{63}$$

$$z_{i,F+2+m+h}^r = -\frac{1}{2} \sum_{\substack{q=1 \\ (q \ne h)}}^{s} d_q(I) \prod_{j=1}^{m} x_{ij}^{\alpha(q,j)} + \frac{1}{2} \sum_{\substack{q=1 \\ (q \ne h)}}^{s} d_q(II) \prod_{j=1}^{m} x_{ij}^{\alpha(q,j)}, \tag{64}$$

где $i = 1,2,\ldots,n, \ \ h = 1,2,\ldots,s$.

Отображение $\mathbf{Z}_{r-1} \to \mathbf{Z}_r$ осуществлено, оператор $\Re$ определен.

3. Укажем правило остановки. Вычисления заканчиваются на итерации r^*, если выполнено условие

$$CR(\mathbf{z}_F^{r^*+1}) - CR(\mathbf{z}_F^{r^*}) \le \delta_1, \tag{65}$$

где $\delta_1 \ge 0$ – заданная величина.

Вычислительную процедуру (40)–(65), проведенную для фиксированного p (допустимое число членов в модели), назовем этапом алгоритма.

Если априорно число членов в ДФ неизвестно, то поиск начинается с этапа с номером $p_0 = 1$ или с любого заданного p_0. На этапе алгоритма с номером p модели, у которых число членов s больше p, не рассматриваются. Начальная матрица классификационных чисел $\mathbf{Z}_0^{p_0}$ для этапа p_0 совпадает с (45), а для этапа p задается по правилу

$$\mathbf{Z}_0^p = \mathbf{Z}_{r^*}^{p-1}, \tag{66}$$

где $\mathbf{Z}_{r_*}^{p-1}$ – матрица классификационных чисел предыдущего этапа.

Вычисления заканчиваются на этапе с номером p^*, если выполнено условие, аналогичное (65):

$$CR(\mathbf{z}_{F,p^*+1}) - CR(\mathbf{z}_{F,p^*}) \le \delta_2, \qquad (67)$$

где $\mathbf{z}_{F,p}$ – лучший вектор классификационных чисел этапа p (лучший вектор классификационных чисел итерации r^* этого этапа); $\delta_2 \ge 0$ – заданная величина.

Построение алгоритма завершено.

Отметим принципы моделирования метода группового учета аргументов, которые применены при построении алгоритма: а) итерационное оценивание параметров в (40)–(67); б) внешний критерий (58)–(59) для сравнения и отбора моделей; в) неединственность моделей (60), передаваемых от итерации к итерации и от этапа к этапу; г) генерация векторов классификационных чисел по правилу (63)–(64).

4. Выводы

Задача статистической классификации состояний сложных систем рассмотрена как задача нелинейного дискриминантного анализа в условиях неопределенности по составу признаков. Описан критерий качества дискриминантных функций в нелинейном дискриминантном анализе, который основан на разбиении наблюдений на обучающие и проверочные выборки. Построен новый итерационный алгоритм МГУА для решения задачи нелинейного дискриминантного анализа.

Литература

1. Ивахненко А. Г. Помехоустойчивость моделирования / А. Г. Ивахненко, В. С. Степашко. – Киев : Наукова думка, 1985. – 216 с.

2. Ивахненко А. Г. Моделирование сложных систем по экспериментальным данным / А. Г. Ивахненко, Ю. П. Юрачковский. – М. : Радио и связь, 1987. – 120 с.

3. Юрачковский Ю. П. Сходимость многорядных алгоритмов МГУА / Ю. П. Юрачковский // Автоматика, 1981. – № 3. – С. 32–36.

4. Белозерский Е. А. Об одном подходе к построению многорядных алгоритмов МГУА с линейными частными описаниями / Е. А. Белозерский, Н. А. Ивахненко, Ю. П. Юрачковский // Автоматика, 1981. – № 5. – С. 3–7.

5. Сарычев А. П. Итерационный алгоритм МГУА для синтеза разделяющей функции в задаче дискриминантного анализа / А. П. Сарычев // Автоматика. – 1988. – № 2. – С. 20–24.

6. Фукунага К. Введение в статистическую теорию распознавания образов / К. Фукунага; пер. с англ. – М. : Наука, 1979. – 368 с.

7. Раудис Ш. Влияние объема выборки на качество классификации (обзор) / Ш. Раудис // Статистические проблемы управления. – Вильнюс : Ин-т мат. и киб. АН Литвы, 1984. – Вып. 66. – С. 9–42.

8. Распознавание образов: Состояние и перспективы / К. Верхаген, Р. Дейн, Ф. Грун и др.; пер. с англ. – М. : Радио и связь, 1985. – 104 с.

9. Фомин Я. А. Статистическая теория распознавания образов / Я. А. Фомин, Г. Р. Тарловский. – М. : Радио и связь, 1986. – 264 с.

10. Прикладная статистика: классификация и снижение размерности / С. А. Айвазян, В. М. Бухштабер, И. С. Енюков и др. – М. : Финансы и статистика, 1989. – 608 с.

11. Мирошниченко Л. В. Сравнение алгоритмов выбора наилучшего подмножества признаков в распознавании образов / Л. В. Мирошниченко // Статистические проблемы управления. – Вильнюс : ИМК АН Литвы, 1990. – Вып. 93. – С. 78–91.

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

13. 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.

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

15. 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.

16. Sarychev A. P. GMDH-Based Criterion for Optimal Set Features Determination in Nonlinear Discriminant Analysis / A. P. Sarychev, L. V. Sarycheva // III International Conference on Inductive Modelling : ICIM–2010, May 16–22 2010, Yevpatoria : Proc. – Yevpatoria : Ukraine, 2010. – P. 40–43.

17. Сарычев А. П. Определение оптимального множества признаков в задаче нелинейного дискриминантного анализа на основе метода группового учета аргументов / А. П. Сарычев, Л. В. Сарычева // Искусственный интеллект. Интеллектуальные системы. ИИ-2010: материалы Международной научно-технической конференций : 20–24 сентября 2010 г., Кацивели, Украина. – Том 1. –Донецк : ИПИИ «Наука і освіта», 2010. – С. 345–351.