УДК 519:681:51

РАСКИН Л.Г., д.т.н., профессор (НТУ «ХПИ»),
КАТКОВА Т.И., к.пед.н., доцент (Бердянский университет менеджмента и бизнеса)
ГОЛОВКО В.А., ведущий экономист (Укрэксимбанк)

Нечеткий многомерный дискриминантный анализ в задаче диагностики состояния

Рассмотрена задача классификации на множестве нечетко заданных переменных. Решение основано на традиционной технологии многомерного дискриминантного анализа, которая адаптирована к случаю, когда исходные данные – нечеткие величины с известными функциями принадлежности.
Ключевые слова: *диагностика состояния, многомерный дискриминантный анализ, нечеткие исходные данные*

Введение

Задача диагностики состояния объекта принадлежит к обширному классу задач классификации и состоит в отнесении конкретного объекта w к одной из двух совокупностей W_1 или W_2 на основе наблюдения p контролируемых параметров $x_1, x_2, \ldots, x_p$. Стандартная процедура классификации состоит в следующем [1]. Предполагается, что вектор наблюдений X имеет нормальное распределение с параметрами (μ_1, Σ_1), если он принадлежит совокупности W_1, и имеет нормальное распределение с параметрами (μ_2, Σ_2), если этот вектор принадлежит совокупности W_2. Здесь векторы μ_1 и μ_2 задают математические ожидания компонентов X при заданных предположениях, а матрицы Σ_1 и Σ_2 – дисперсионные, причем $\Sigma_1 = \Sigma_2$.

Далее вводится так называемая дискриминантная функция

$$z = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_p x_p . \qquad (1)$$

Теперь объект w относят к W_1, если

$$z > C , \qquad (2)$$

и относят к W_2 - в противном случае, (C – некоторая постоянная).

Необходимые для решения задачи классификации параметры α_j, $j = 1,2,\ldots,p$, и C отыскиваются из следующих соображений.

Если наблюдение $X \in W_1$, то z принадлежит нормальному распределению $N\left(\zeta_1, \sigma_z^2\right)$, если же $X \in W_2$, то z принадлежит распределению $N\left(\zeta_2, \sigma_z^2\right)$, где

$$\zeta_1 = \sum_{j=1}^{p} \alpha_j \mu_{1j} , \; \zeta_2 = \sum_{j=1}^{p} \alpha_j \mu_{2j} ,$$

$$\sigma_z^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} \alpha_i \alpha_j \sigma_{ij}^2 .$$

Набор α_j, $j = 1,2,\ldots,p$, выбирается таким образом, чтобы расстояние Махаланобиса [2] между совокупностями W_1 и W_2, вычисляемое по формуле

$$\Delta^2 = \frac{(\zeta_1 - \zeta_2)^2}{\sigma_z^2} ,$$

было максимальным. Как показано в [3], искомый набор определяется решением системы линейных уравнений

$$\begin{aligned} &\alpha_1\sigma_{11} + \alpha_2\sigma_{12} + \ldots + \alpha_p\sigma_{1p} = \mu_{11} - \mu_{21}, \\ &\alpha_1\sigma_{21} + \alpha_2\sigma_{22} + \ldots + \alpha_p\sigma_{2p} = \mu_{12} - \mu_{22}, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &\alpha_1\sigma_{p1} + \alpha_2\sigma_{p2} + \ldots + \alpha_p\sigma_{pp} = \mu_{1p} - \mu_{2p}. \end{aligned} \qquad (3)$$

Теперь константа C вычисляется по формуле:

$$C = \frac{1}{2}\left(\zeta_1 + \zeta_2\right). \qquad (4)$$

Введем

$P\left(H_1 / X \in W_2\right)$ - вероятность того, что по результатам вычислений по формулам (1), (2) принята гипотеза H_1 о принадлежности X совокупности W_1, когда на самом деле объект принадлежит W_2,

$P\left(H_2 / X \in W_1\right)$ - вероятность того, что по результатам вычислений по формулам (1), (2) принята гипотеза H_2 о принадлежности X совокупности W_2, когда на самом деле объект принадлежит W_1.

Выбор C в соответствии с (4) обеспечивает минимальное значение суммарной вероятности перепутывания.

Практическая реализация описанной стандартной технологии затруднена в связи с следующими обстоятельствами. Во–первых, в реальной ситуации недостаточного числа наблюдений за объектами из W_1 и W_2 получаемые выборочные оценки средних значений и дисперсий контролируемых параметров могут иметь непрогнозируемо большие ошибки. Во-вторых, даже сама гипотеза о нормальности случайных значений этих параметров не может быть обоснованно принята или отвергнута. В этой ситуации естественный путь состоит в использовании описания реальных исходных данных в терминах нечеткой математики [4].

Постановка задачи

Пусть для описания нечетких значений контролируемых параметров для объектов, принадлежащих W_1, используется набор функций принадлежности $\left(\mu_{11}(x_1), \mu_{12}(x_2), \dots, \mu_{1p}(x_p)\right)$, и, соответственно, для описания значений параметров для объектов из W_2 – набор $\left(\mu_{21}(x_1), \mu_{22}(x_2), \dots, \mu_{2p}(x_p)\right)$. Поставим задачу построения процедуры классификации в этих условиях.

Основные результаты

Для каждой из введенных функций принадлежности нечетких величин определим функцию

$$\varphi_{kj}(x_j) = \frac{\mu_{kj}(x_j)}{\int\limits_{-\infty}^{\infty} \mu_{kj}(x_j)\,dx_j}, \qquad (5)$$

$k = 1,2$, $j = 1,2,\dots,p$.

Функции $\varphi_{kj}(x_j)$ неотрицательны, удовлетворяют условию нормировки и, следовательно, являются аналогами плотностей распределения случайных величин. Тогда, в соответствии с концепцией, принятой в [5], могут быть рассчитаны ожидаемые значения соответствующих нечетких величин

$$E_k\left[x_j\right] = \frac{\int\limits_{-\infty}^{\infty} x_j \mu_{kj}(x_j)\,dx_j}{\int\limits_{-\infty}^{\infty} \mu_{kj}(x_j)\,dx_j}, \qquad (6)$$

$k = 1,2$, $j = 1,2,\dots,p$,

а также аналоги дисперсий этих нечетких величин

$$D_k\left[x_j\right] = \frac{\int\limits_{-\infty}^{\infty} x_j^2 \mu_{kj}(x_j)\,dx_j}{\int\limits_{-\infty}^{\infty} \mu_{kj}(x_j)\,dx_j} - \left[E_k(x_j)\right]^2, \qquad (7)$$

$k = 1,2$, $j = 1,2,\dots,p$.

Приведенные соотношения (6) и (7) обеспечивают расчет величин, являющихся естественными аналогами математических ожиданий и дисперсий случайных величин. Пусть, например, функция принадлежности какого-то конкретного контролируемого параметра x (индексы k и j здесь опустим) имеет гауссов вид

$$\mu(\mathrm{x}) = exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}.$$

Тогда:

$$E[x] = \frac{\int\limits_{-\infty}^{\infty} x\, exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}dx}{\int\limits_{-\infty}^{\infty} exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}dx} =$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\int\limits_{-\infty}^{\infty} x\, exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}dx = m;$$

$$E\left[x^2\right]=\frac{\int\limits_{-\infty}^{\infty} x^2\, exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}dx}{\int\limits_{-\infty}^{\infty} exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}dx}=$$

$$=\frac{1}{\sqrt{2\pi}\sigma}\int\limits_{-\infty}^{\infty} x^2\, exp\left\{-\frac{(x-m)^2}{2\sigma^2}\right\}dx=\sigma^2+m^2;$$

$$D[x]=E\left[x^2\right]-(E[x])^2=\sigma^2 .$$

Таким образом, в случае, когда контролируемые параметры объекта есть нечеткие величины с известными функциями принадлежности, задача классификации сводится к расчету аналогов математических ожиданий и дисперсий этих параметров, после чего может быть использована стандартная технология (1) – (4).

Ситуация усложняется, если вследствие, например, малости выборки обоснованное построение функций принадлежности неосуществимо и всей исходной информации достаточно только для оценки функций принадлежности ожидаемых значений и дисперсий контролируемых параметров. Таким образом, будем считать известными наборы

$$\mu_k(m_{kj}),\quad \mu_k(\sigma_{ij}),\ k=1,2,\ i=1,2,\ldots,p,\ j=1,2,\ldots,p .$$

Решение задачи содержит несколько шагов.

На первом шаге решается система линейных алгебраических уравнений (3) с нечетко заданными параметрами. При этом решим вначале порождаемую (3) четкую систему линейных алгебраических уравнений, используя модальные значения $m_{kj}^{(0)},\sigma_{ij}^{(0)}$ нечетких величин $\mathrm{m_{kj}},\sigma_{ij}$, $k=1,2$, $i=1,2,\ldots,p$, $j=1,2,\ldots,p$. Эта система имеет вид:

$$\sum_{j=1}^{p}\alpha_j\sigma_{ij}^{(0)}=m_{1i}^{(0)}-m_{2i}^{(0)}\ ,\ i=1,2,\ldots,p . \tag{8}$$

Пусть $\left\{\alpha_j^{(0)}\right\}$ – решение системы (8).

Введем нечеткие величины U_i

$$U_i=\alpha_1\sigma_{i1}+\alpha_2\sigma_{i2}+\ldots+\alpha_p\sigma_{ip}-m_{1i}+m_{2i}\ ,$$
$$i=1,2,\ldots,p .$$

Понятно, что любому набору $\{\alpha_j\}$ соответствует набор нечетких чисел U_i, $i=1,2,\ldots,m$. Используя правила выполнения операций для нечетких величин [6], найдем функции принадлежности этих нечетких чисел U_i .

$$\mu(U_i)=\sum_{j=1}^{p}\alpha_j\mu(\sigma_{ij})-[\mu_{1i}(m_{1i})-\mu_{2i}(m_{2i})], \tag{9}$$
$$i=1,2,\ldots,p .$$

Сформулируем теперь требования к решению нечеткой системы линейных алгебраических уравнений (3). Во-первых, это решение должно обеспечивать минимальную неопределенность в отношении нечетких значений невязок U_i, то есть максимальную компактность функций принадлежности $\mu(U_i)$, $i=1,2,\ldots,p$. Во-вторых, искомый набор $\{\alpha_j\}$ должен минимально отклоняться от набора $\left\{\alpha_j^{(0)}\right\}$, соответствующего модальному решению системы (8).

Уровень компактности функции принадлежности $\mu(x)$ нечеткого числа x можно оценивать, например, нормированным квадратом площади под кривой $\mu(x)$. С учетом этого оптимизируемый функционал будет иметь вид:

$$I(\{A\})=\sum_{i=1}^{p}\frac{\left[\int\limits_{-\infty}^{\infty}\mu(U_i)dU_i\right]^2}{\left[\int\limits_{-\infty}^{\infty}\mu\left(U_i^{(0)}\right)dU_i\right]^2}+\frac{\left(A-A^{(0)}\right)^T\left(A-A^{(0)}\right)}{\left(A^{(0)}\right)^T A^{(0)}}. \tag{10}$$

Здесь

$$\mu\left(U_i^{(0)}\right)=\sum_{j=1}^{p}\alpha_j^{(0)}\sigma_{ij}^{(0)}-\left(m_{1i}^{(0)}-m_{2i}^{(0)}\right),\ i=1,2,\ldots,p ,$$

$$A^T=(\alpha_1\ \alpha_2\ldots\alpha_p)a ,$$

$$\left(A^{(0)}\right)^T=\left(\alpha_1^{(0)}\ \alpha_2^{(0)}\ldots\alpha_p^{(0)}\right).$$

Если принять, что функции принадлежности нечетких параметров гауссовы, то

$$\mu(U_i)=exp\left\{-\frac{\left(U_i-\overline{U_i}\right)^2}{2\sigma^2}\right\},$$

$$\overline{U_i} = \sum_{j=1}^{p} \alpha_j \sigma_{ij}^{(0)} - \left(m_{1i} - m_{2i}\right),$$

$$\sigma_i^2 = \sum_{j=1}^{p} \alpha_j^2 D_{ij} + \left(D_{1i} + D_{2i}\right),$$

$i = 1,2,\ldots,p$, $j = 1,2,\ldots,p$.

Тогда

$$\int_{-\infty}^{\infty} \mu(U_i) dU_i =$$

$$= \sqrt{2\pi}\sigma_i \frac{1}{\sqrt{2\pi}\sigma_i} \int_{-\infty}^{\infty} exp\left\{-\frac{\left(U_i - \overline{U_i}\right)^2}{2\sigma^2}\right\} dU_i =$$

$$= \sqrt{2\pi}\left(\sum_{j=1}^{p} \alpha_j D_{ij} + \left(D_{1i} + D_{2i}\right)\right)^{0.5},$$

и целевой функционал (10) приобретает вид:

$$I(\{A\}) = \frac{\left(\sum_{j=1}^{p} \alpha_j D_{ij} + D_{1i} + D_{2i}\right)^2}{\left(\sum_{j=1}^{p} \alpha_j^{(0)} D_{ij}^{(0)} + D_{1i}^{(0)} + D_{2i}^{(0)}\right)^2} + \tag{11}$$

$$+ \frac{\sum_{j=1}^{p} \left(\alpha_j - \alpha_j^{(0)}\right)^2}{\sum_{j=1}^{p} \left(\alpha_j^{(0)}\right)^2}.$$

Таким образом, получена обычная задача квадратического программирования минимизации (11) при удовлетворении системе ограничений (3).

Пусть $A^* = \left(\alpha_1^* \;\; \alpha_2^* \ldots \alpha_p^*\right)$ – решение этой задачи.

Далее

$$\zeta_k^* = \sum_{j=1}^{p} \alpha_l^* m_{kj}, \qquad k = 1,2,$$

и

$$C = \frac{1}{2}\left(\zeta_1^* + \zeta_2^*\right) = \frac{1}{2}\sum_{j=1}^{p} \alpha_l^* \left(m_{1j} + m_{2j}\right).$$

Функция принадлежности нечеткого числа C формируется с использованием правил выполнения операций над нечеткими числами [6].

При этом

$$\mu(C) = exp\left\{-\frac{\left(C - \overline{C}\right)^2}{2\sigma_C^2}\right\}, \tag{12}$$

$$\overline{C} = \frac{1}{2}\sum_{j=1}^{p} \alpha_l^* \left(m_{1j}^{(0)} + m_{2j}^{(0)}\right),$$

$$\sigma_C^2 = \frac{1}{4}\sum_{j=1}^{p} \left(\alpha_j^*\right)^2 \left(\sigma_j^{(0)}\right)^2.$$

Теперь для заданного набора $X = \left(x_1\, x_2 \ldots x_p\right)$ вычисляется значение дискриминантной функции

$$z = \sum_{j=1}^{p} \alpha_j^* x_j,$$

которое сравнивается с C. Процедура сравнения нечеткого числа C и четкого значения дискриминантной функции z реализуется следующим образом. Получим функцию принадлежности нечеткой разности $U = z - C$. C учетом (12) имеем:

$$\mu(U) = \mu(z - C) = exp\left\{-\frac{\left[U - \left(z - \overline{C}\right)\right]^2}{2\sigma_C^2}\right\}.$$

Нечеткое число $U = z - C$ естественно считать положительным, если выполняется неравенство $z - \overline{C} > 0$.

В этом случае наблюдаемый объект следует отнести к подмножеству W_1. В противном случае это решение отклоняется.

Выводы

Таким образом, предложен метод решения задачи классификации на основе нечетких исходных данных применительно, в частности, к проблеме диагностики состояния. Процедура классификации реализует традиционные идеи многомерного дискриминантного анализа с учетом неопределенности исходных данных. Стержневым элементом процедуры является методика решения нечеткой системы линейных алгебраических уравнений. Направление дальнейших исследований – учет различий в опасности разных ошибок классификации.

Литература

1. Афифи А. Статистический анализ. Подход с использованием ЭВМ. Пер. с англ. Под ред. Башарина Г.П./А.Афифи, С.Эйзен. – М.: Мир, 1982.-488с.
2. Mahalanobis P.C. On the Generalized Distance in Statistics. Proc. of the National Institute of Sciences of India. 1936, 12, 49-55.
3. Fisher R.A. The Use of Multiple Measurements in Taxonomic problem. Annals of Eugenics, 1936.7,179-188.
4. Zadeh L., Fuzzy Sets/L. Zadeh//Information and Control.-1965.-vol.8(3).-P.338-353
5. Серая О.В. Многомерные модели логистики в условиях неопределенности./О.В. Серая. – Х.: ФОП Стеценко И.И.,2010.-512с
6. Раскин Л.Г. Нечеткая математика/Л.Г. Раскин, О.В. Серая. –Х.: Парус, 2008. - 352с.

Резюме

Розглянуто задачу класифікації на безлічі нечітко заданих змінних. Рішення базується на традиційній технології багатовимірного дискримінантного аналізу, яка адаптована до випадку, коли вихідні дані - нечіткі величини з відомими функціями належності.

The task of classification on a set of fuzzy fixed variables has been considered. The solution is based on a traditional technology of multivariate discriminante analysis which is adapted to the case when the initial data are fuzzy values with known membership functions.

Key words: diagnostics of the condition, multivariate discriminante analysis, fuzzy initial data.

Рецензент д.т.н., профессор Листровой С.В. (УкрГАЖТ)

Поступила 7.03.2013г.