УДК 519.281:621.391

Р.А.Абусев
(Пермский университет)

КЛАССИФИКАЦИЯ МНОГОМЕРНЫХ НОРМАЛЬНЫХ СОВОКУПНОСТЕЙ С ПОМОЩЬЮ ПОТЕНЦИАЛЬНЫХ ФУНКЦИЙ

В случае многомерного нормального распределения строятся классификационные тесты, основанные на потенциальных функциях, и показывается их асимптотическая эквивалентность оптимальному байесовскому правилу. Получена более точная оценка верхней границы вероятности ошибочной классификации.

В работах [1-5], посвященных применению потенциальных функций к решению задач распознавания образов, не предполагается знание законов распределений обучающих выборок. Вместе с тем [6-10] в случае многомерных нормальных совокупностей строятся классификационные тесты, асимптотически эквивалентные оптимальному байесовскому классификационному правилу.

Естественно возникает вопрос: как определить потенциальную функцию, чтобы основанное на ней классификационное правило было также асимптотически эквивалентно байесовскому правилу.

В данной работе потенциальная функция определяется как функция некоторого расстояния в пространстве достаточных статистик. Далее преследуются две цели:

1) построение классификационных правил, основанных на этих потенциальных функциях, асимптотически эквивалентных оптимальному байесовскому правилу;

2) получение оценки верхней границы вероятности ошибочной классификации.

1. Имеются две выборки

$$\tilde{\mathfrak{X}}_i^{(0)} = \{x_{i1}, x_{i2}, \ldots, x_{in_i}\}, \quad i = 1,2 \qquad (1.1)$$

из k -мерных совокупностей $\mathfrak{X}_1$ и $\mathfrak{X}_2$ с неизвестными распределениями.

Вычислим выборочные средние и матрицы ковариаций:

$$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \,, \qquad (I.2)$$

$$S_i = \frac{1}{n_i}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)' \,, \quad i = 1, 2 \,.$$

Для произвольного наблюдения $x_0 \in \mathfrak{X}_1 \cup \mathfrak{X}_2$ и наблюдения $x_{ij} \in \mathfrak{X}_i^{(0)}$ определим расстояние следующим образом:

$$R(x_0, x_{ij}) = (x_0 - x_{ij})' S_i^{-1} (x_0 - x_{ij}) \,, \qquad (I.3)$$
$$(i = 1, 2 \,;\; j = 1, 2, \ldots, n_i) \,.$$

Тогда средним расстоянием между наблюдением x_0 и выборкой $\mathfrak{X}_i^{(0)}$ будет

$$R(x_0, \mathfrak{X}_i^{(0)}) = \frac{1}{n_i}\sum_{j=1}^{n_i}(x_0 - x_{ij})' S_i^{-1}(x_0 - x_{ij}), \; i = 1, 2 \,. \qquad (I.4)$$

На основе (I.4) построим классификационное правило:

$$x_0 \in \mathfrak{X}_1 \quad \text{, если} \quad R(x_0, \mathfrak{X}_1^{(0)} > R(x_0, \mathfrak{X}_2^{(0)}) \,, \qquad (I.5)$$

$$x_0 \in \mathfrak{X}_2 \quad \text{, в остальных случаях.}$$

Теорема I. Пусть совокупности $\mathfrak{X}_i$ имеют k-мерное нормальное распределение с плотностями

$$p_{\mu_i, \Sigma}(x) = \frac{1}{(2\pi)^{\frac{k}{2}} |\Sigma|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i)} \,, \; i = 1, 2 \,, \qquad (I.6)$$

с неизвестными векторами средних μ_1, μ_2 и неизвестной общей матрицей ковариаций Σ. Тогда классификационное правило (I.5) асимптотически эквивалентно оптимальному байесовскому классификационному правилу с порогом, равным единице.

Действительно, возьмем несмещенную оценку

$$S = \frac{1}{n_1 + n_2 - 2}\sum_{i=1}^{2}\sum_{i=1}^{n_i}(x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)' \qquad (I.7)$$

матрицы ковариаций Σ, общей для обеих совокупностей. Тогда формулу (I.4) можно представить в виде

$$R(x_0, \mathcal{K}_i^{(0)}) = \frac{1}{n_i}\sum_{j=1}^{n_i}(x_0 - x_{ij})' S^{-1}(x_0 - x_{ij}) =$$

$$= (x_0 - \bar{x}_i)' S^{-1}(x_0 - \bar{x}_i) + S_i S^{-1}.$$

Следовательно, учитывая, что

$$S_i S^{-1} \xrightarrow{p} 1 \qquad \text{при} \quad n_1, n_2 \longrightarrow \infty,$$

классификационное правило (I.5) можно записать в виде

$$\begin{aligned} &x_0 \in \mathcal{K}_1 \quad \text{, если } (x_0 - \bar{x}_1)' S^{-1}(x_0 - \bar{x}_1) > (x_0 - \bar{x}_2) S^{-1}(x_0 - \bar{x}_2), \\ &x_0 \in \mathcal{K}_2 \quad \text{- в остальных случаях.} \end{aligned} \tag{I.8}$$

Ясно, что классификационное правило (I.8) эквивалентно правилу

$$\begin{aligned} &x_0 \in \mathcal{K}_1 \quad \text{, если } p_{\bar{x}_1, S}(x_0) > p_{\bar{x}_2, S}(x_0), \\ &x_0 \in \mathcal{K}_2 \quad \text{- в остальных случаях,} \end{aligned} \tag{I.9}$$

а последнее асимптотически эквивалентно [7] правилу

$$\begin{aligned} &x_0 \in \mathcal{K}_1 \quad \text{, если } p_{\mu_1, \Sigma}(x_0) > p_{\mu_2, \Sigma}(x_0), \\ &x_0 \in \mathcal{K}_2 \quad \text{- в остальных случаях.} \end{aligned}$$

В работе [I] рассматривается потенциальная функция

$$p(x, y) = \frac{1}{1 + \alpha R^2(x, y)},$$

где $R^2(x, y)$ - квадрат евклидова расстояния между точками x и y, α - положительная постоянная. Далее определяются средние потенциалы выборок $\mathcal{K}_1^{(0)}$ и $\mathcal{K}_2^{(0)}$ в точке x_0 :

$$p(x_0, \mathcal{K}_i^{(0)}) = \frac{1}{n_i}\sum_{j=1}^{n_i}\frac{1}{1 + \alpha R^2(x_0, x_{ij})}, \quad i = 1, 2,$$

на основе которых строится классификационное правило

$x_0 \in \mathcal{K}_1$, если $P(x_0, \mathcal{K}_1^{(0)}) > P(x_0, \mathcal{K}_2^{(0)})$, (I.IO)

$x_0 \in \mathcal{K}_2$ – в остальных случаях.

Ни при каких ограничениях на законы распределений, на выбор расстояния $R(x_0, x_{ij})$ и α классификационное правило (I.IO) не сводится к правилу (I.9). Следовательно, в случае нормального распределения (наиболее часто встречающегося в приложениях) классификационное правило (I.8) не хуже, чем (I.IO).

Для выборок (I.I) из совокупностей $\mathcal{K}_1$ и $\mathcal{K}_2$, имеющих плотности (I.6), определим средний потенциал в точке $x_0 \in \mathcal{K}_1 \cup \mathcal{K}_2$ в виде

$$\Phi(x_0, \mathcal{K}_i^{(0)}) = \frac{1}{1 + \alpha_i (x_0 - \bar{x}_i)' S^{-1} (x_0 - \bar{x}_i)}, \; i = 1, 2. \quad (I.II)$$

Введем классификационное правило:

$x_0 \in \mathcal{K}_1$, если $\Phi(x_0, \mathcal{K}_1^{(0)}) > \Phi(x_0, \mathcal{K}_2^{(0)})$, (I.I2)

$x_0 \in \mathcal{K}_2$ – в остальных случаях.

Ясно, что классификационное правило (I.I2) при $\alpha_1 = \alpha_2$ эквивалентно правилу (I.8) и, следовательно, асимптотически эквивалентно оптимальному байесовскому классификационному правилу с порогом, равным единице.

В случае многомерного нормального распределения с неизвестными средними μ_1 и μ_2 известной матрицей ковариаций Σ, общей для обеих совокупностей и $\alpha_1 = \frac{n_1}{n_1 + 1}$, $\alpha_2 = \frac{n_2}{n_2 + 1}$, классификационное правило (I.I2) можно записать в виде

$x_0 \in \mathcal{K}_1$, если $\frac{n_1}{n_1+1}(x_0 - \bar{x}_1)' \Sigma^{-1} (x_0 - \bar{x}_1) < \frac{n_2}{n_2+1}(x_0 - \bar{x}_2)' \Sigma^{-1} (x_0 - \bar{x}_2)$, (I.I3)

$x_0 \in \mathcal{K}_2$ – в остальных случаях,

что совпадает с правилом классификации, полученным в работе [6] методом максимального правдоподобия с порогом, равным единице.

В работах [9 – I0] получены асимптотические разложения статистик дискриминантного анализа правил (I.8) и (I.I3) при размерностях пространства, сравнимых с объемом выборок.

II. Пусть $\mathcal{K}_i$ имеет k-мерное нормальное распределение

с плотностью $P_{\mu_i,\Sigma}(x)$ (см. (I.6)), где μ_i - неизвестны, Σ_i известны $i = I,2$. Тогда, если P_e - вероятность суммарных ошибок классификации, то

$$P_e = \int min\left[\omega_1 P_{\mu_1,\Sigma_1}(x),\ \omega_2 P_{\mu_2,\Sigma_2}(x)\right]dx \le$$
$$\sqrt{\omega_1\omega_2}\int\left[P_{\mu_1,\Sigma_1}(x)P_{\mu_2,\Sigma_2}(x)\right]^{\frac{1}{2}}dx = \delta(\mu_1,\mu_2,\Sigma_1,\Sigma_2). \qquad (2.1)$$

Пусть

$$P_{n_i}(x) = \frac{1}{n_i h_i^k}\sum_{j=1}^{n_i} K\left(\frac{x - x_{ij}}{h_i}\right),$$

$$\check{P}_{\mu_i,\Sigma_i}(x) = \frac{1}{(2\pi)^{k/2}|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2}(x-\bar{x}_i)'\Sigma_i^{-1}(x-\bar{x}_i)\right\}, \qquad (2.2)$$

$$\hat{P}_{\mu_i,\Sigma_i}(x) = \left(\frac{n_i}{n_i-1}\right)^{\frac{k}{2}}\frac{1}{(2\pi)^{k/2}|\Sigma_i|^{1/2}}\exp\left\{-\frac{n_i}{2(n_i-1)}(x-\bar{x}_i)'\Sigma_i^{-1}(x-\bar{x}_i)\right\} -$$

суть статистические оценки непараметрические [II], наибольшего правдоподобия [6] и несмещенные [7] соответственно, построенные по обучающим выборкам $\mathcal{X}_i^{(n)}$ объемов n_i, $i = I,2$.

Положив

$$K\left(\frac{x - x_{ij}}{h_i}\right) = \frac{1}{(2\pi)^{k/2}|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2h_i^2}(x-x_{ij})'\Sigma_i^{-1}(x-x_{ij})\right\}$$

и построив статистические оценки $\delta_1, \delta_2, \delta_3$ для $\delta(\mu_1,\mu_2,\Sigma_1,\Sigma_2)$, используя оценки (2.2), получим

$$\delta_1 = \frac{\sqrt{\omega_1\omega_2}\,(2h_1h_2)^{k/2}|\Sigma_1\Sigma_2|^{1/4}}{\sqrt{n_1n_2}\,|h_1^2\Sigma_1 + h_2^2\Sigma_2|^{1/2}} \times$$
$$\times \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\exp\left\{-\frac{1}{4}(x_{1i}-x_{2j})'\left(h_1^2\Sigma_1 + h_2^2\Sigma_2\right)^{-1}(x_{1i}-x_{2j})\right\}, \qquad (2.3)$$

$$\delta_2 = \frac{2^{k/2}|\Sigma_1\Sigma_2|^{1/4}}{|\Sigma_1+\Sigma_2|^{1/2}}\exp\left\{-\frac{1}{4}(\bar{x}_1-\bar{x}_2)'\left(\Sigma_1+\Sigma_2\right)^{-1}(\bar{x}_1-\bar{x}_2)\right\},$$

$$\delta_3 = \left[\frac{(n_1-1)(n_2-1)}{n_1 n_2}\right]^{\frac{k}{4}} \frac{2^{k/2}\left|\sum_1 \sum_2\right|^{\frac{1}{4}}}{\left|\frac{n_1-1}{n_1}\sum_1 + \frac{n_2-1}{n_2}\sum_2\right|^{1/2}} \times$$

$$\times \exp\left\{-\frac{1}{4}(\bar{x}_1-\bar{x}_2)'\left[\frac{n_1-1}{n_1}\sum_1 + \frac{n_2-1}{n_2}\sum_2\right]^{-1}(\bar{x}_1-\bar{x}_2)\right\}.$$

Вычислим математические ожидания статистик $\delta_1, \delta_2, \delta_3$:

$$M\delta_1 = \frac{\sqrt{n_1 n_2}\; 2^k (h_1 h_2)^{k/2}\left|\sum_1 \sum_2\right|^{\frac{1}{4}}}{\left|(2h_1^2+1)\sum_1 + (2h_2^2+1)\sum_2\right|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mu_1-\mu_2)\left[(2h_1^2+1)\sum_1 + (2h_2^2+1)\sum_2\right]^{-1} \times \right.$$
$$\left. \times (\mu_1-\mu_2)\right\},$$

$$M\delta_2 = \frac{2^{k/2}\left|\sum_1 \sum_2\right|^{\frac{1}{4}}}{\left|\frac{2n_1+1}{2n_1}\sum_1 + \frac{2n_2+1}{2n_2}\sum_2\right|^{\frac{1}{2}}} \exp\left\{-\frac{1}{4}(\mu_1-\mu_2)'\left[\frac{2n_1+1}{2n_1}\sum_1 + \frac{2n_2+1}{2n_2}\sum_2\right]^{-1} \times (2.4)\right.$$
$$\left. \times (\mu_1-\mu_2)\right\},$$

$$M\delta_3 = \left[\frac{(n_1-1)(n_2-1)}{n_1 n_2}\right]^{\frac{k}{4}} \frac{2^{k/2}\left|\sum_1 \sum_2\right|^{\frac{1}{4}}}{\left|\frac{2n_1-1}{2n_1}\sum_1 + \frac{2n_2-1}{2n_2}\sum_2\right|^{\frac{1}{2}}} \exp\left\{-\frac{1}{4}(\mu_1-\mu_2)'\left[\frac{2n_1-1}{2n_1}\sum_1 + \right.\right.$$
$$\left.\left. + \frac{2n_2-1}{2n_2}\sum_2\right]^{-1}(\mu_1-\mu_2)\right\}$$

и докажем следующую теорему.

Т е о р е м а 2. Оценки δ_2, δ_3 верхней границы δ для вероятности суммарных ошибок P_e являются состоятельными и, начиная с некоторого n_0

$$M\delta_1 > M\delta_2 > M\delta_3 .$$

Состоятельность оценок δ_2, δ_3 для δ очевидна из формулы (2.4). Далее, для простоты положим $n_1 = n_2 = n$, $h_1 = h_2 = h$ ($h_i \to 0$, $n_i h_i^k \to \infty$ при $n_i \to \infty$, ($i = 1,2$)), $\Delta = (\mu_1-\mu_2)'(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)$, получим

$$M\delta_1 = \frac{n(2h)^k\left|\sum_1 \sum_2\right|^{\frac{1}{4}}}{(2h^2+1)^{k/2}\left|\sum_1+\sum_2\right|^{1/2}} \exp\left\{-\frac{1}{2(2h^2+1)}\Delta\right\},$$

$$M\delta_2 = \left(1-\frac{1}{2n+1}\right)^{k/2} \frac{2^{k/2}\left|\sum_1 \sum_2\right|^{\frac{1}{4}}}{\left|\sum_1+\sum_2\right|^{1/2}} \exp\left\{-\frac{n}{2(2n+1)}\Delta\right\}, \qquad (2.5)$$

$$M\delta_3 = \left(\frac{2n-2}{2n-1}\right)^{k/2} \frac{2^{k/2}\left|\sum_1 \sum_2\right|^{\frac{1}{4}}}{\left|\sum_1+\sum_2\right|^{1/2}} \exp\left\{-\frac{n}{2(2n-1)}\Delta\right\}.$$

Из последних равенств видно, что начиная с некоторого $n = n_0$

$$M\delta_1 > M\delta_2 .$$

Далее,

$$\ln \frac{M\delta_3}{M\delta_2} = \frac{k}{2} \ln \left(1 - \frac{1}{2n^2 - n}\right) - \frac{n}{4n^2 - 1} \Delta$$

и

$$\ln \left(1 - \frac{1}{2n^2 - n}\right) < 0$$

для любого $n \geq 1$. Поэтому

$$\ln \frac{M\delta_3}{M\delta_2} < 0 ,$$

т.е. $M\delta_2 > M\delta_3$ для любого $n \geq 1$.
Как следствия этой теоремы отметим, что $\delta_2 > \delta_3$ для любого $n_1 = n_2 = n > 1$, что следует из неравенства

$$\ln \frac{\delta_2}{\delta_3} = \frac{1}{4(n-1)} \Delta > 0,$$

и верны следующие асимптотические неравенства

$$P_e \leq \delta_1 , \quad P_e \leq \delta_2 , \quad P_e \leq \delta_3 , \tag{2.6}$$

причем в этих неравенствах в случае неизвестных ковариационных матриц Σ_1, Σ_2 можно воспользоваться их оценками S_1 , S_2 соответственно, сохранив состоятельность δ_2 , δ_3 для δ .

Рассмотрим первое равенство из (2.3). Так как $\exp\{-z^2\} \leq \frac{1}{1+z^2}$, то получим

$$\delta_1 \leq \frac{\sqrt{\omega_1 \omega_2}\,(2h_1h_2)^{k/2} |\Sigma_1 \Sigma_2|^{\frac{1}{4}}}{\sqrt{n_1 n_2}\,|h_1^2\Sigma_1 + h_2^2\Sigma_2|^{1/2}} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \frac{1}{1 + \frac{1}{4}(x_{1i} - x_{2j})'(h_1^2\Sigma_1 + h_2^2\Sigma_2)^{-1}(x_{1i} - x_{2j})} \tag{2.7}$$

Так как величину

$$D = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \frac{1}{1 + \frac{1}{4}(x_{1i} - x_{2j})'(h_1^2\Sigma_1 + h_2^2\Sigma_2)^{-1}(x_{1i} - x_{2j})}$$

можно рассматривать как средний потенциал точек одной обучающей выборки на другой [2-4] в метрике вида (I.3), то с учетом (2.6),

(2.7) получим следующую асимптотическую связь между вероятностью суммарных ошибок P_e и потенциалом D:

$$P_e \le \delta_1 \le \frac{\sqrt{\omega_1 \omega_2 n_1 n_2}\,(2h_1h_2)^{k/2}\,|\Sigma_1\Sigma_2|^{\frac{1}{4}}}{|h_1^2\Sigma_1 + h_2^2\Sigma_2|^{1/2}}\,D\,. \qquad (2.8)$$

Наиболее просто запишется неравенство (2.8) при $h_1 = h_2 = h$, $\Sigma_1 = \Sigma_2 = I$

$$D_e \le \frac{\sqrt{\omega_1\omega_2}}{\sqrt{n_1 n_2}} \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \frac{1}{1 + \frac{1}{8h^2}(x_{1i} - x_{2j})'(x_{1i} - x_{2j})}\,, \qquad (2.9)$$

которое получено в работе [5].

Неравенства (2.6), (2.8) могут быть применены для минимизации описания, т.е. для выбора информативных признаков. Для этого нужно выбрать m -мерное подпространство ($m \le k$), минимизирующее правые части (2.6), (2.8). При этом с учетом теоремы 2 и ее следствия предпочтительнее использовать величину δ_3.
Кроме того, величины $M\delta_2$, $M\delta_3$ могут быть использованы как границы сверху для вероятности суммарных ошибок классификации нормальных совокупностей.

Литература

I. Башкиров О.А., Браверман Э.М., Мучник И.Б. Алгоритмы обучения машины распознаванию зрительных образов, основанные на методе потенциальных функций. – Автоматика и телемеханика , I964, № 5.

2. Айзерман М.Л., Браверман Э.М., Розоноэр Л.И. Теоретические основы метода потенциальных функций в задаче об обучении автоматов разделению входных ситуаций на классы. – Автоматика и телемеханика , I964, № 6.

3. Айзерман М.А., Браверман Э.М., Розоноэр Л.И. Вероятностная задача об обучении автоматов распознаванию классов и метод потенциальных функций. – Автоматика и телемеханика , I964, № 9.

4. Браверман Э.М. О методе потенциальных функций. – Автоматика и телемеханика , I965, № I2.

5. Бабу К.Ч., Калра С.Н. О применении потенциальной функции Башкирова, Бравермана и Мучнина для выделения информативных признаков при распознавании образов. – Автоматика и телеме-

ханика, 1972, № 12.

6. А н д е р с о н Т. Введение в многомерный статистический анализ. М., ГИФМЛ, 1963.

7. Л у м е л ь с к и й Я.П. Об одном способе построения асимптотически оптимальных классификационных тестов в случае многомерного нормального распределения. – Техническая кибернетика, 1972, № 2.

8. У р б а х В.Ю. Дискриминантный анализ. Основные идеи и приложения. – Сб. Статистические методы классификации. М., Изд-во МГУ, 1969, вып. 1.

9. Д е е в А.Д. Представление статистик дискриминантного анализа и асимптотические разложения при размерностях пространства, сравнимых с объемом выборок. – ДАН СССР, 1970, т. 195, № 4.

10. Д е е в А.Д. Асимптотические разложения распределений статистик дискриминантного анализа. – Сб. Статистические методы классификации. Изд-во МГУ, 1972, вып. 21.

11. Е п а н и ч н и к о в В.А. Непараметрическая оценка многомерной плотности вероятности. – Теория вероятн. и ее примен., 1969, т. XIУ, вып. 1.