УДК [519.245+519.214]: 519.237.8

М.А. ИВАНЧУК, И.В. МАЛЫК

РЕШЕНИЕ ЗАДАЧИ КЛАССИФИКАЦИИ С ИСПОЛЬЗОВАНИЕМ ε-СЕТЕЙ

Аннотация. Предложен новый метод решения задачи классификации, основанный на разделении двух множеств в пространстве R^d путем построения и разделения ε-сетей этих множеств в ранжированном пространстве относительно гиперплоскостей. Введено понятие области разделения — тех значений ε, при которых возможно разделить множества. Приведены примеры области разделения для случайных величин с разными распределениями и доказана теорема о ее сходимости. Введено понятие совокупности всех возможных ε-сетей некоторого множества и доказаны ее свойства. Доказана слабая сходимость нормированной разности эмпирической и теоретической кривых разделения к нормальному распределению, что позволяет проверять гипотезы о местонахождении теоретической кривой разделения в конкретной точке.

Ключевые слова: ε-сети, разделение множеств, размерность Вапника–Червоненкиса, классификация.

ПОСТАНОВКА ЗАДАЧИ

В 1987 г. D. Haussler и E. Welzl [1] ввели понятие ε-сетей. С этого времени данное направление стало одним из перспективных в вычислительной и комбинаторной геометрии и получило широкое применение [2–6]. Рассмотрим ранжированное пространство (X, R), где X — некоторое множество, R — совокупность подмножеств множества X. В геометрическом контексте в качестве множества X принимают евклидово пространство R^d, при этом совокупностью подмножеств R в таком пространстве может быть, например, H^d — совокупность всех подпространств; совокупность всех шаров; совокупность всех выпуклых оболочек; совокупность всех d-мерных симплексов [2]; совокупность всех прямоугольников, параллельных осям координат; совокупность всех треугольников [3]; совокупность всех клинов, образованных пересечением двух непараллельных полупространств, угол между которыми не меньше α радиан [4].

Теория ε-сетей для одного множества на данный момент хорошо развита. Так, например, B. Aronov с соавторами [5] показали, что для прямоугольников, параллельных осям координат, размер ε-сети в R^2 оценивается как $O\left(\frac{1}{\varepsilon}\log\log\frac{1}{\varepsilon}\right)$. J. Kulkarni с соавторами [4] доказали, что для клинов толщиной α в R^2 существуют ε-сети размера $O\left(\frac{\pi}{\alpha\varepsilon}\right)$. B. Gärtner [2] доказал существование ε-сетей мощностью не более $\left\lceil\frac{d\ln n}{\varepsilon}\right\rceil$, где d — размерность Вапника–Червоненкиса [3]. J. Matousek [6] в 2000 г. доказал, что если для каждого конечного множества $T \subset R^3$ в выпуклой позиции существует ε-сеть для полупространств размера $s(\varepsilon)$, то для любого конечного множества в R^3 существует ε-сеть в ранжированном пространстве (R^d, H^d) размера $3s(\varepsilon)+1$.

В настоящей статье рассматривается задача классификации двух множеств, которые сгенерированы независимыми случайными величинами. Пусть из генеральных совокупностей, сгенерированных случайными величинами ξ и η, получены выборки A и B объемами n_A, n_B. Задача состоит в нахождении разделяющей гиперплоскости L, для которой справедливо соотношение

$$P\{\xi \in L^+,\ \eta \in L^-\} = \sup_{l \in H^d} P\{\xi \in l^+, \eta \in l^-\}$$

и оценки вероятности $P\{\xi_n \in L_n^+,\ \eta_n \in L_n^-\}$, где L_n — гиперплоскость, разделяющая множества $\xi_n = A,\ \eta_n = B$.

Задачи классификации широко используются во многих отраслях, в том числе в экономике [7, 8], медицине [9–13], политологии [14–16] . Для решения этого класса задач существует большое количество методик. Сравнительный анализ использования дисперсионного анализа, кластерного анализа, вальдовской классификации для решения задач классификации при прогнозировании в медицине приведен в [17].

В настоящей статье для решения задачи классификации используется теория ε-сетей. В отличие от работ [1–6], касающихся построения ε-сетей одного множества, в данной статье рассматриваются ε-сети двух множеств в ранжированном пространстве (R^d, H^d). Описано множество возможных значений ε для двух разделимых ε-сетей, доказаны его свойства. Приведены примеры множества возможных значений ε для нормального, экспоненциального, равномерного и пуассоновского законов распределения.

ε-РАЗДЕЛИМОСТЬ ДВУХ МНОЖЕСТВ В ПРОСТРАНСТВЕ R^d

Определение 1. Множества A и B объемами n_A, n_B называются ε-разделимыми, если существуют $A_1 \subset A$, $B_1 \subset B$, для которых $\operatorname{conv}(A \setminus A_1) \cap \cap \operatorname{conv}(B \setminus B_1) = \varnothing$ и $|A_1| + |B_1| < \varepsilon(n_A + n_B)$.

Определение 2. Гиперплоскость L называется разделяющей для множеств A и B, если $\operatorname{conv}_A \subset L^+$, $\operatorname{conv}_B \subset L^-$.

Определение 3. Гиперплоскость L_ε называется ε-разделяющей для множеств A и B, если $\dfrac{|A \cap L_\varepsilon^+| + |B \cap L_\varepsilon^-|}{|A| + |B|} \geq 1 - \varepsilon$.

В [18] были доказаны необходимые и достаточные условия существования ε-сетей этих множеств.

Теорема 1. Для того чтобы множества A и B были ε-разделимыми, необходимо и достаточно, чтобы существовали ε_A, ε_B и соответствующие им ε-сети $N_A^{\varepsilon_A}$, $N_B^{\varepsilon_B}$ множеств A и B в (R^d, H^d), для которых выполнялись следующие соотношения:

$$\varepsilon_A n_A + \varepsilon_B n_B < \varepsilon(n_A + n_B), \tag{1}$$

$$\operatorname{conv} N_A^{\varepsilon_A} \cap \operatorname{conv} N_B^{\varepsilon_B} = \varnothing. \tag{2}$$

ОБЛАСТЬ РАЗДЕЛЕНИЯ И ЕЕ СВОЙСТВА

Определение 4. Множество

$$D_{A,B} = \{(\varepsilon_1, \varepsilon_2) \in (0,1)^2 : \exists\, N_A^{\varepsilon_1}, N_B^{\varepsilon_2},\ \operatorname{conv} N_A^{\varepsilon_1} \cap \operatorname{conv} N_B^{\varepsilon_2} = \varnothing\} \tag{3}$$

называется областью разделения.

Очевидно, что для любых ε_A, $\varepsilon_B \in D_{A,B}$ выполняется условие (2) теоремы 1. Следовательно, для того чтобы множества A и B были ε-разделимыми, необходимо проверить существование $(\varepsilon_A, \varepsilon_B) \in D_{A,B}$, для которых выполняется условию (1). Для этого достаточно показать, что $(\varepsilon_A^0, \varepsilon_B^0) = \underset{(\varepsilon_A, \varepsilon_B) \in D_{A,B}}{\operatorname{argmin}} \dfrac{\varepsilon_A n_A + \varepsilon_B n_B}{n_A + n_B}$ удовлетворяет условию (1). Если условие (1) не выполняется для $(\varepsilon_A^0, \varepsilon_B^0)$, то оно не будет выполняться для всех $(\varepsilon_A, \varepsilon_B) \in D_{A,B}$, т.е. множества A и B ε-неразделимы.

Предположим, что случайные величины ξ, $\eta \in R^1$ — непрерывные случайные величины с функциями распределения F_ξ, F_η.

Определение 5. Множество D_l такое, что

$$D_l := \{(x, y) \in (0,1)^2 : \exists h \in R^1,\ P\{\xi \in h_+\} \le x,\ P\{\eta \in h_-\} \le y\}, \tag{4}$$

называется областью разделения для ξ, η.

Для множества D_l имеет место следующее утверждение.

Лемма 1. Пусть существует обратная функция к F_ξ. Тогда множества D_l и $\overline{D}_l := (0,1)^2 \setminus D$ разделены кривой

$$y(x) = \min\left(F_\eta(F_\xi^{-1}(1-x)),\ 1 - F_\eta(F_\xi^{-1}(x))\right). \tag{5}$$

Доказательство. Рассмотрим два возможных случая.

1. Пусть для некоторого $h \in (-\infty, \infty)$: $F_\xi(h) > F_\eta(h)$, тогда множество D_l можно описать системой неравенств

$$\begin{cases} x \ge 1 - F_\xi(h), \\ y \ge F_\eta(h). \end{cases}$$

Тогда кривая, разделяющая множества D_l и $\overline{D}_l$, имеет вид

$$y(x) = F_\eta(F_\xi^{-1}(1-x)).$$

2. Пусть для некоторого $h \in (-\infty, \infty)$: $F_\xi(h) \le F_\eta(h)$, тогда множество D_l можно описать системой неравенств

$$\begin{cases} x \ge F_\xi(h), \\ y \ge 1 - F_\eta(h). \end{cases}$$

В этом случае кривая, разделяющая множества D_l и $\overline{D}_l$, имеет вид

$$y(x) = 1 - F_\eta(F_\xi^{-1}(x)).$$

Таким образом, в общем случае множества D_l и $\overline{D}_l$ разделяются кривой

$$y(x) = \min(F_\eta(F_\xi^{-1}(1-x)),\ 1 - F_\eta(F_\xi^{-1}(x))).$$

Лемма 1 доказана.

Замечание 1. Условие леммы может быть заменено существованием обратной функции F_η^{-1}. Тогда разделяющая кривая будет иметь вид

$$x(y) = \min\left(F_\xi(F_\eta^{-1}(1-y)),\ 1 - F_\xi(F_\eta^{-1}(y))\right).$$

Рассмотрим общий случай, когда функции распределения не обязательно имеют обратные. Воспользуемся понятием обобщенной обратной функции [19].

Определение 6. Для возрастающей функции $T: R \to R$ при $T(-\infty) = \lim_{x \downarrow -\infty} T(x)$ и $T(\infty) = \lim_{x \uparrow \infty} T(x)$ обобщенной обратной функцией называется

$$T_G^{-1}(y) = \inf \{x \in R : T(x) \geq y\}, \ y \in R, \tag{6}$$

при условии, что $\inf \varnothing = \infty$. Если $T: R \to [0,1]$ — функция распределения, то $T_G^{-1} : [0,1] \to \overline{R}$ также называется квантильной функцией T.

Тогда имеет место следующее утверждение.

Лемма 2. Множества D_l и $\overline{D_l} := (0,1)^2 \setminus D$ разделены кривой

$$y(x) = \min \left(F_\eta((F_\xi)_G^{-1}(1-x)),\ 1 - F_\eta((F_\xi)_G^{-1}(x))\right). \tag{7}$$

Доказательство. Пусть $x \in (0;1)$ — некоторая точка, для которой не существует F_ξ^{-1}. Найдем значения функции $y(x)$ (см. (7)). Тогда согласно определению обобщенной обратной функции для любого $\delta > 0$ справедлива принадлежность $y(x) + \delta \in D_l$, $y(x) - \delta \in \overline{D_l}$. Таким образом, множества D_l и $\overline{D_l}$ разделены кривой (6).

Лемма 2 доказана.

ПРИМЕРЫ ПОСТРОЕНИЯ ОБЛАСТИ РАЗДЕЛЕНИЯ

Рассмотрим примеры области разделения для наиболее часто используемых распределений.

Пример 1. Пусть случайные величины ξ и η распределены по нормальному закону с параметрами μ_ξ, σ_ξ и μ_η, σ_η. Тогда область D_l ограничена функцией

$$y(x) = \min \left(\Phi \left(\frac{\Phi_G^{-1}\left(\frac{1-x-\mu_\xi}{\sigma_\xi} \right) - \mu_\eta}{\sigma_\eta} \right), 1 - \Phi \left(\frac{\Phi_G^{-1}\left(\frac{x-\mu_\xi}{\sigma_\xi} \right) - \mu_\eta}{\sigma_\eta} \right) \right), \quad x \in (0,1).$$

На рис.1 изображен график функции $y = y(x)$ для $\mu_\xi = 3$, $\sigma_\xi = 1$ и $\mu_\eta = 5$, $\sigma_\eta = 2$.

Пример 2. Пусть случайные величины ξ и η распределены по экспоненциальному закону с параметрами μ_ξ и μ_η. Область D_l ограничена снизу функцией

$$y(x) = \min \left(1 - x^{\mu_\xi / \mu_\eta},\ (1-x)^{\mu_\xi / \mu_\eta}\right),\ x \in (0,\ 1).$$

На рис. 2 изображен график функции $y = y(x)$ для $\mu_\xi = 1$ и $\mu_\eta = 5$.

Пример 3. Пусть случайные величины ξ и η распределены по равномерному закону с параметрами a_ξ, b_ξ и a_η, b_η. Область D_l ограничена снизу функцией

$$y(x) = \min \left(\frac{b_\xi (1-x) + a_\xi x - a_\eta}{b_\eta - a_\eta},\ 1 - \frac{a_\xi (1-x) + b_\xi x - a_\eta}{b_\eta - a_\eta} \right),\ x \in (0,1).$$

На рис. 3 изображен график функции $y = y(x)$ для $a_\xi = 1$, $b_\xi = 5$ и $a_\eta = 4$, $b_\eta = 7$.

Пример 4. Пусть дискретные случайные величины ξ, η распределены по закону Пуассона с параметрами λ_ξ, λ_η. На рис. 4 изображен график функции $y = y(x)$ для $\lambda_\xi = 1$, $\lambda_\eta = 2$ при $x \in (0;1)$.

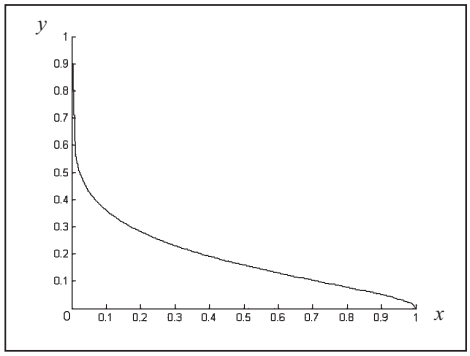

Рис. 1. График функции $y(x)$ для нормального распределения

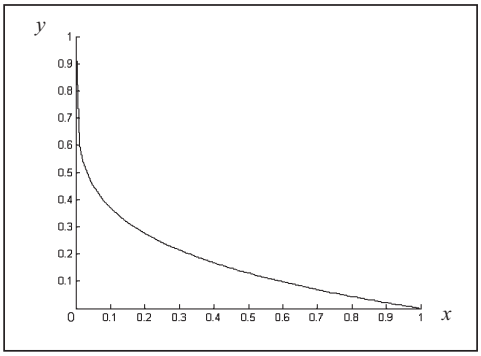

Рис. 2. График функции $y(x)$ для экспоненциального распределения

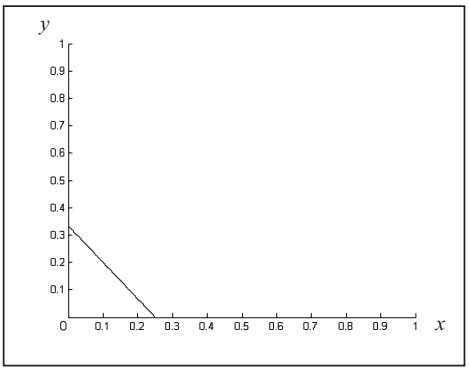

Рис. 3. График функции $y(x)$ для равномерного распределения

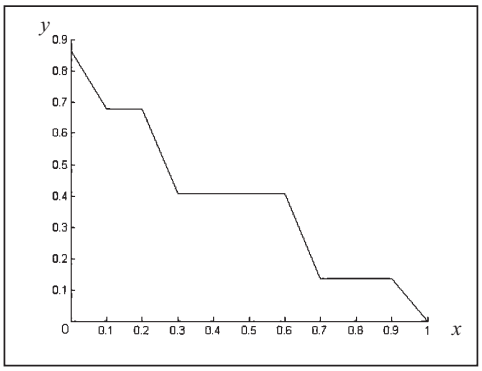

Рис. 4. График функции $y(x)$ для распределения Пуассона

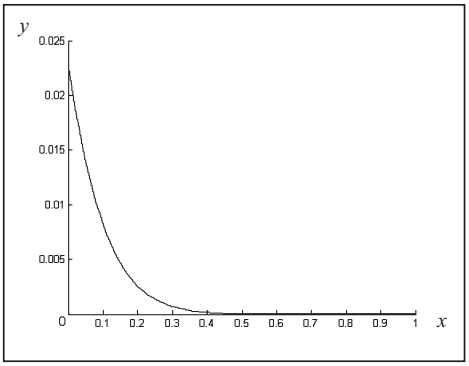

Рис. 5. График функции $y(x)$ для равномерного и нормального распределения

Пример 5. Пусть случайная величина ξ распределена по равномерному закону с параметрами a_ξ, b_ξ, а случайная величина η распределена по нормальному закону с параметрами μ_η, σ_η. Тогда область D_1 ограничена функцией

$$y(x) = \min\left(\Phi\left(\frac{b_\xi (1-x) + a_\xi x - \mu_\eta}{\sigma_\eta} \right), \right.$$

$$\left. 1 - \Phi\left(\frac{b_\xi x + a_\xi (1-x) - \mu_\eta}{\sigma_\eta} \right) \right), \quad x \in (0,1).$$

На рис. 5 изображен график функции $y = y(x)$ для $a_\xi = 1$, $b_\xi = 5$ и $\mu_\eta = 7$, $\sigma_\eta = 1$.

СОВОКУПНОСТИ ε-СЕТЕЙ И ИХ СВОЙСТВА

Обозначим $\overline{D_{A,B}} = \{(\varepsilon_1, \varepsilon_2) \in (0,1)^2 : \forall N_A^{\varepsilon_1},\ N_B^{\varepsilon_2}, \operatorname{conv} N_A^{\varepsilon_1} \cap \operatorname{conv} N_B^{\varepsilon_2} \neq \varnothing\}$.

Определение 7. Множество X называется звездным, если $\exists x_0 \in X\ \ \forall x \in X$, $\forall \alpha \in [0,1]:\ \alpha x_0 + (1-\alpha) x \in X$ [20].

Точку x_0 будем называть центральной точкой звездной области.

Докажем, что множества $D_{A,B}$ и $\overline{D_{A,B}}$ являются звездными. Для этого введем понятие совокупности всех возможных ε-сетей множества и рассмотрим его свойства.

Определение 8. Множество $\mathbf{N}_A^{\varepsilon}$, содержащее все возможные ε-сети множества A, называется совокупностью ε-сетей множества A.

Например, для $A = \{1, 2, 3\}$ при $\varepsilon = 0,2$ множество $\mathbf{N}_A^{\varepsilon}$ содержит только A; для $\varepsilon = 0,4$ имеем $\mathbf{N}_A^{\varepsilon} = \{A, \{1, 2\}, \{1, 3\}, \{2, 3\}\}$.

Лемма 3. $\mathbf{N}_A^{\varepsilon_1} \subseteq \mathbf{N}_A^{\varepsilon_2}$, если $\varepsilon_1 \le \varepsilon_2$.

Доказательство. Пусть H^+ — полупространство, порождаемое некоторой гиперплоскостью H. Согласно определению ε-сетей $\forall |A \cap H^+| \ge \varepsilon_1 |A|$ $\exists x \in A \cap H^+$: $x \in N_A^{\varepsilon_1}$. Пусть $\varepsilon_2 \ge \varepsilon_1$, тогда если $|A \cap H^+| \ge \varepsilon_2 |A|$, то $|A \cap H^+| \ge \varepsilon_1 |A|$, откуда $\forall x \in N_A^{\varepsilon_1} \Rightarrow x \in N_A^{\varepsilon_2}$, т.е. $\mathbf{N}_A^{\varepsilon_1} \subseteq \mathbf{N}_A^{\varepsilon_2}$.

Лемма 3 доказана.

Введем следующие обозначения: $A_i \subset A$ — подмножество множества A, которое содержит i точек: $A_i = \{a_1, a_2, ..., a_i\}$; $A_{i+1} \subset A$ — множество, содержащее все точки множества A_i и еще одну точку из множества A: $A_{i+1} = \{A_i \cup \{a_{i+1}\}\}$. Тогда справедлива следующая лемма.

Лемма 4. Для любого $N_{A_{i+1}}^{j/i+1}$ $\forall j = \overline{1, i}$ существует $N_{A_i}^{j/i}$, для которого выполняется соотношение

$$N_{A_{i+1}}^{j/i+1} \subseteq N_{A_i}^{j/i}.$$

Доказательство. Рассмотрим множество $\mathbf{N}_{A_i}^{j/i}$, которое является совокупностью всех возможных комбинаций точек из множества A_i по i точек, по $i-1$ точек, ..., по $i-j+1$ точек. В свою очередь, множество $\mathbf{N}_{A_{i+1}}^{j/i+1}$ — это совокупность всех возможных комбинаций точек из множества A_{i+1} по $i+1$ точек, по i точек, ..., по $i-j$ точек. Очевидно, что среди этих комбинаций точек найдется такая, что не содержит точку a_{i+1} и принадлежит множеству $N_{A_i}^{j/i} \in \mathbf{N}_{A_i}^{j/i}$.

Лемма 4 доказана.

Лемма 5. Множества $D_{A,B}$ и $\overline{D_{A,B}}$ — звездные.

Доказательство.

1. Рассмотрим точки $x = (\varepsilon_1, \varepsilon_2) \in D_{A,B}$ и $x_0 = (1; 1)$. Докажем, что $\alpha x_0 + (1-\alpha) x \in D_{A,B}$. Пусть $\alpha x_0 + (1-\alpha) x = (\varepsilon_1 + \alpha(1-\varepsilon_1), \varepsilon_2 + \alpha(1-\varepsilon_2))$.

Поскольку $\varepsilon_1 \le \varepsilon_1 + \alpha(1-\varepsilon_1)$, то согласно лемме 1 $\mathbf{N}_A^{\varepsilon_1} \subseteq \mathbf{N}_A^{\varepsilon_1+\alpha(1-\varepsilon_1)}$. Аналогично $N_B^{\varepsilon_2} \subseteq N_B^{\varepsilon_2+\alpha(1-\varepsilon_2)}$. Поскольку по определению $\exists N_A^{\varepsilon_1} \in \mathbf{N}_A^{\varepsilon_1}$, $N_B^{\varepsilon_2} \in \mathbf{N}_B^{\varepsilon_2}$: $\operatorname{conv} N_A^{\varepsilon_1} \cap \operatorname{conv} N_B^{\varepsilon_2} = \varnothing$, то существуют такие ε-сети $N_A^{\varepsilon_1} \in \mathbf{N}_A^{\varepsilon_1} \subseteq$ $\subseteq \mathbf{N}_A^{\varepsilon_1+\alpha(1-\varepsilon_1)}$, $N_B^{\varepsilon_2} \in \mathbf{N}_B^{\varepsilon_2} \subseteq \mathbf{N}_B^{\varepsilon_2+\alpha(1-\varepsilon_2)}$, для которых справедливо $\operatorname{conv} N_A^{(1-\alpha)\varepsilon_1} \cap$ $\cap \operatorname{conv} N_B^{(1-\alpha)\varepsilon_2} = \varnothing$.

Таким образом, $\alpha x_0 + (1-\alpha) x \in D_{A,B}$, т.е. множество $D_{A,B}$ — звездное.

2. Рассмотрим точки $x = (\varepsilon_1, \varepsilon_2) \in \overline{D_{A,B}}$ и $x_0 = (0; 0)$. Докажем, что $\alpha x_0 + (1-\alpha) x \in \overline{D_{A,B}}$. Пусть также $\alpha x_0 + (1-\alpha) x = ((1-\alpha)\varepsilon_1, (1-\alpha)\varepsilon_2)$.

Поскольку $\varepsilon_1 \ge (1-\alpha)\varepsilon_1$, то согласно лемме 1 $\mathbf{N}_A^{\varepsilon_1} \supseteq \mathbf{N}_A^{(1-\alpha)\varepsilon_1}$. Аналогично $\mathbf{N}_B^{\varepsilon_2} \supseteq \mathbf{N}_B^{(1-\alpha)\varepsilon_2}$. Поскольку по определению $\forall N_A^{\varepsilon_1} \in \mathbf{N}_A^{\varepsilon_1}$, $N_B^{\varepsilon_2} \in \mathbf{N}_B^{\varepsilon_2}$:

$\operatorname{conv} N_A^{\varepsilon_1} \cap \operatorname{conv} N_B^{\varepsilon_2} \neq \varnothing$, то для любых $N_A^{(1-\alpha)\varepsilon_1} \in \mathbf{N}_A^{(1-\alpha)\varepsilon_1} \supseteq \mathbf{N}_A^{\varepsilon_1}$ и $N_B^{(1-\alpha)\varepsilon_2} \in \mathbf{N}_B^{(1-\alpha)\varepsilon_2} \supseteq \mathbf{N}_B^{\varepsilon_2}$ справедливо $\operatorname{conv} N_A^{(1-\alpha)\varepsilon_1} \cap \operatorname{conv} N_B^{(1-\alpha)\varepsilon_2} \neq \varnothing$.

Таким образом, $\alpha x_0 + (1-\alpha)x \in \overline{D_{A,B}}$, т.е. множество $\overline{D_{A,B}}$ — звездное.

Лемма 5 доказана.

АСИМПТОТИКА РАЗДЕЛЯЮЩИХ КРИВЫХ

В леммах 1, 2 описана кривая, разделяющая множества $D_{\xi,\eta}$ и $\overline{D_{\xi,\eta}}$. Докажем ее сходимость с кривой, разделяющей множества $D_{A,B}$ и $\overline{D_{A,B}}$.

Теорема 2. Пусть:

1) множества A, B объемом n_A, n_B соответственно генерируются независимыми непрерывными случайными величинами ξ и η;

2) множества $D_{A,B}$ и $\overline{D_{A,B}}$ разделены кривой $y_{A,B}(x)$.

Тогда $\forall x \in (0,1)$ имеет место соотношение

$$\lim_{n_A, n_B \to \infty} y_{A,B}(x) = y(x),$$

где

$$y(x) = \min\left(F_\eta((F_\xi)_G^{-1}(1-x)),\ 1 - F_\eta((F_\xi)_G^{-1}(x))\right).$$

Доказательство. Для доказательства теоремы достаточно показать, что имеют место соотношения

$$F_{n_B}(F_{n_A}^{-1}(y)) \to F_\eta(F_\xi^{-1}(y)),\ y \in (0,1), \tag{8}$$

$$F_{n_B}(1 - F_{n_A}^{-1}(y)) \to F_\eta(1 - F_\xi^{-1}(y)),\ y \in (0,1). \tag{9}$$

Покажем справедливость соотношения (8):

$$\sup_{y\in[0,1]} |F_{n_B}(F_{n_A}^{-1}(y)) - F_\eta(F_\xi^{-1}(y))| =$$

$$= \sup_{y\in[0,1]} |F_{n_B}(F_{n_A}^{-1}(y)) + F_\eta(F_{n_A}^{-1}(y)) - F_\eta(F_{n_A}^{-1}(y)) - F_\eta(F_\xi^{-1}(y))| \le$$

$$\le \sup_{y\in[0,1]} |F_{n_B}(F_{n_A}^{-1}(y)) - F_\eta(F_{n_A}^{-1}(y))| + \sup_{y\in[0,1]} |F_\eta(F_{n_A}^{-1}(y)) - F_\eta(F_\xi^{-1}(y))| \le$$

$$\le \sup_{x\in R^1} |F_{n_B}(x) - F_\eta(x)| + \sup_{y\in[0,1]} |F_\eta(F_{n_A}^{-1}(y)) - F_\eta(F_\xi^{-1}(y))|.$$

По теореме Гливенко–Кантелли [21] для первого слагаемого справедливо

$$\sup_{x\in R^1} |F_{n_B}(x) - F_\eta(x)| \to 0,\ n_B \to \infty.$$

Предположим, что η имеет плотность $f_\eta < K$. Тогда для второго слагаемого имеем оценку

$$\sup_{y\in[0,1]} |F_\eta(F_{n_A}^{-1}(y)) - F_\eta(F_\xi^{-1}(y))| \le K \sup_{y\in[0,1]} |F_{n_A}^{-1}(y) - F_\xi^{-1}(y)|.$$

Покажем, что $\sup_{y\in[0,\,1]} |F_{n_A}^{-1}(y) - F_\xi^{-1}(y)| \to 0$.

Предположим, что $F_\xi(x_0) = y_0 \in (0,1)$ фиксированное и $|F^{-1}_{n_A}(y_0) - F^{-1}_\xi(y_0)| \not\to 0$, тогда $|F_\xi(x_0) - F_{n_A}(x_0)| \not\to 0$. Получаем противоречие.

Таким образом, имеет место соотношение (8). По аналогии соотношение (9) также справедливо.

Теорема 2 доказана.

Теорема 3. Пусть:

1) ξ_i, $i \geq 1$, имеют распределение F_ξ; η_j, $j \geq 1$, имеют распределение F_η;

2) $\xi_i, \eta_j, i, j \geq 1$, — независимые в совокупности случайные величины;

3) функции F_ξ, F_η имеют обратные и существуют ограниченные плотности f_ξ, f_η.

Тогда наблюдается слабая сходимость

$$\zeta_{n,m} = \zeta_{n,m}(y) = \sqrt{n}(F_n(F_m^{-1}(y)) - F_\eta(F_\xi^{-1}(y))) \to N(0, \sigma^2),$$

где

$$\sigma^2 = F_\eta(F_\xi^{-1}(y))(1 - F_\eta(F_\xi^{-1}(y)))$$

при $n \to \infty$, $m = O(n^\alpha), \alpha > 0$.

Доказательство. Рассмотрим математическое ожидание $\zeta_{n,m}$:

$$E\zeta_{n,m} = \sqrt{n}(EF_n(F_m^{-1}(y)) - F_\eta(F_\xi^{-1}(y))).$$

Используя определение эмпирической функции распределения, получаем

$$|E\zeta_{n,m}| = |\sqrt{n}(EI_{\eta < F_m^{-1}(y)} - F_\eta(F_\xi^{-1}(y)))| =$$

$$= |\sqrt{n}(P\{\eta < F_m^{-1}(y)\} - P\{\eta < F_\xi^{-1}(y)\})| =$$

$$|\sqrt{n}(P\{F_\xi^{-1}(y) \leq \eta < F_m^{-1}(y)\} + P\{F_m^{-1}(y) \leq \eta < F_\xi^{-1}(y)\})|.$$

Таким образом, для сходимости

$$E\zeta_{n,m} \to 0$$

достаточно показать, что

$$\sqrt{n}P\{|F_\xi^{-1}(y) - F_m^{-1}(y)| > \varepsilon\} \to 0$$

при любом ε.

Согласно определению обобщенной обратной функции получим соотношение

$$\sqrt{n}P\{|F_\xi^{-1}(y) - F_m^{-1}(y)| > \varepsilon\} \leq$$

$$\leq \sqrt{n}P\{|F_\xi(y) - F_m(y)| > \delta(\varepsilon)\} \leq \sqrt{n}2e^{-2m\delta(\varepsilon)^2} = 2e^{0{,}5\ln n - 2m\delta(\varepsilon)^2},$$

где последнее неравенство базируется на неравенстве Дворецкого–Кифера–Вольфовица [22], $\delta(\varepsilon)$ не зависит от n и m.

Таким образом, при $m = O(n^\alpha), \alpha > 0$, имеем $\lim\limits_{n\to\infty} E\zeta_{n,m} = 0$.

Рассмотрим дисперсию $\zeta_{n,m}$:

$$D\zeta_{n,m} = E\zeta_{n,m}^2 - (E\zeta_{n,m})^2.$$

Второе слагаемое стремится к нулю, поэтому важно рассмотреть первое слагаемое предыдущего равенства

$$E\zeta_{n,m}^2 = nE(F_n(F_m^{-1}(y)) - F_\eta(F_\xi^{-1}(y)))^2 =$$

$$= E(I_{\eta<F_m^{-1}(y)} - F_\eta(F_\xi^{-1}(y)))^2 +$$

$$+\frac{1}{n}\sum_{i\neq j}(I_{\eta_i<F_m^{-1}(y)} - F_\eta(F_\xi^{-1}(y)))(I_{\eta_j<F_m^{-1}(y)} - F_\eta(F_\xi^{-1}(y))).$$

Пренебрегая величинами порядка малости $O(1)$, получаем соотношение

$$D\zeta_{n,m} =$$

$$= E(I_{\eta<F_m^{-1}(y)} - F_\eta(F_\xi^{-1}(y)))^2 + \frac{1}{n}\left(\sum_{i\neq j}(EI_{\eta_i<F_m^{-1}(y)}I_{\eta_j<F_m^{-1}(y)} - p^2)\right) = A_1 + A_2,$$

где

$$p = F_\eta(F_\xi^{-1}(y)).$$

Рассмотрим каждое слагаемое отдельно. Используя теорему Лебега о мажорантной сходимости [23], получаем

$$A_1 = E(I_{\eta<F_m^{-1}(y)} - F_\eta(F_\xi^{-1}(y)))^2 \to E(I_{\eta<F_\xi^{-1}(y)} - F_\eta(F_\xi^{-1}(y)))^2 = p(1-p)$$

при $m\to\infty$.

Для второго слагаемого с учетом независимости $\eta_i, i\geq 1$, получим

$$A_2 = (n-1)(EI_{\eta_1<F_m^{-1}(y)}I_{\eta_2<F_m^{-1}(y)} - p^2) =$$

$$= (n-1)(P\{\eta_1 < F_m^{-1}(y), \eta_2 < F_m^{-1}(y)\} - p^2) =$$

$$= (n-1)(P\{\eta_1 < F_m^{-1}(y), \eta_2 < F_m^{-1}(y)\} - p^2 \pm P\{\eta_1 < F_m^{-1}(y), \eta_2 < F_\xi^{-1}(y)\}) =$$

$$= (n-1)(P\{\eta_1 < F_m^{-1}(y), \eta_2 < F_m^{-1}(y)\} - P\{\eta_1 < F_m^{-1}(y), \eta_2 < F_\xi^{-1}(y)\}) +$$

$$+(n-1)(P\{\eta_1 < F_m^{-1}(y), \eta_2 < F_\xi^{-1}(y)\} - p^2) = A_{21} + A_{22}.$$

Для первого слагаемого получим

$$|A_{21}| = (n-1)|P\{\eta_1 < F_m^{-1}(y)\}|P\{\eta_2 < F_m^{-1}(y)|\eta_1 < F_m^{-1}(y)\} -$$

$$-P\{\eta_2 < F_\xi^{-1}(y)|\eta_1 < F_m^{-1}(y)\}| \leq$$

$$\leq (n-1)|P\{\eta_2 < F_m^{-1}(y)|\eta_1 < F_m^{-1}(y)\} - P\{\eta_2 < F_\xi^{-1}(y)|\eta_1 < F_m^{-1}(y)\}|.$$

Аналогично предыдущим рассуждениям, используя неравенство Дворецкого–Кифера–Вольфовица, получаем $|A_{21}|\to 0$.

Для второго слагаемого имеем $A_{22} = (n-1)\,p(P\{\eta_1 < F_m^{-1}(y)\} - p)$, отсюда следует $|A_{22}|\to 0$.

Таким образом, $D\zeta_{n,m} \to p(1-p)$.

Последним шагом доказательства будет использование центральной граничной теоремы для асимптотически независимых случайных величин [24]

$$u_{i,m} = I_{\eta_i<F_m^{-1}(y)-p},\ i\geq 1,$$

где асимптотическую независимость следует понимать при $m \to \infty$ в следующем контексте:

$$\lim_{m\to\infty} (P\{u_{i,m} \in A, u_{j,m} \in B\} - P\{u_{i,m} \in A\}P\{u_{j,m} \in B\}) = 0$$

для $i \neq j$, $A, B \in \beta_R$.

Теорема 3 доказана.

ЗАКЛЮЧЕНИЕ

В статье введено понятие области разделения множества D_l, доказана теорема о сходимости множеств $D_{A,B}$ и D_l, а также доказано, что множества D_l и $\overline{D_l}$ являются звездными; предложен аналитический вид функции, разделяющей эти множества. Приведены примеры построения данной функции для нормального, экспоненциального, равномерного и пуассоновского распределений. Несомненнно, что эта же методика может быть рассмотрена и для многомерного случая. В теореме 3 доказана слабая сходимость нормированной разности эмпирической и теоретической кривых разделения к нормальному распределению, что позволяет проверять гипотезы о местонахождении теоретической кривой разделения в конкретной точке.

СПИСОК ЛИТЕРАТУРЫ

1. Haussler D., Welzl E. Epsilon-nets and simplex range queries // Discrete Comput. Geom. — 1987. — N 2. — P. 127–151.
2. Gärtner B., Hoffmann M. Computational Geometry. — http://www.ti.inf.ethz.ch/ew/lehre/CG12/lecture/CG%20lecture%20notes.pdf.
3. Hausler S. VC Dimension. A tutorial for the course computational intelligence. — http://www.igi.tugraz.at/lehre/CI.
4. Kulkarni J., Govindarajan S. New ε-net constructions // Canadian Conference on Computational Geometry — CCCG, 2010. — P. 159–162.
5. Aronov B., Ezra E., Sharir M. Small-size epsilon-nets for axis-parallel rectangles and boxes // Symposium on Theory of Computing, 2009. — P. 639–648.
6. Matousek J., Seidel R., Welzl E. How to net a lot with little: small ε-nets for disks and halfspaces // Sixth Annual Symposium on Computational Geometry, Berkley, CA, USA, June 07–09, 1990. — P. 16–22.
7. Vesely' A. Economic classification and regression problems and neural networks // Agricultural Economics — CZECH. — 2011. — N 57. — P. 150–157.
8. Price D., Pollock A.M., Brhlikova P. Classification problems and the dividing line between government and the market: An examination of NHS foundation trust classification in the UK // Annals of Public and Cooperative Economics. — 2011. — **82**, Issue 4. — P. 455–473.
9. Schaubel D.E., Cai J. Analysis of clustered recurrent event data with application to hospitalization rates among renal failure patients // Biostatistics. — 2005. — N 6. — P. 404–419.
10. Weatherall M., Shirtcliffe P., Travers J., Beasley R. Use of cluster analysis to define COPD phenotypes // Eur. Respir. J. — 2010. — **36.** — P. 472–474.
11. Mahr A., Katsahian S., Varet H. Revisiting the classification of clinical phenotypes of anti-neutrophil cytoplasmic antibody-associated vasculitis: A cluster analysis // Annals of the Rheumatic Diseases. — 2012. — **72**. — P. 1003–1010.
12. Mangasarian O.L., Street W.N., Wolberg W.H. Breast cancer diagnosis and prognosis via linear programming // Operation Research. — 1995. — **43**, N 4.
13. Mangasarian O.L. A simple characterization of solution sets of convex programs // Computer Sciences Technical Report #685, 1987. — P. 21–26.
14. Card D., Krueger A.B. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania // American Economic Review. — 1994. — **84**, N 4. — P. 772–793.
15. Angrist J., Victory L. Using maimonides' rule to estimate the effect of class size on student achievement // Quarterly Journal of Economics. — 1999. — **114.** — P. 533–575.
16. Lee D., Moretti E., Butler M.J. Do voters affect or elect policies? Evidence from the U.S. House // Quarterly Journal of Economics. — 2004. — **119.** — P. 807–859.

17. Ivanchuk M.A., Malyk I.V. Comparison of the methods for classification of observations in predicting complications in critically ill patients // Cybernetics and Systems Analysis. — 2015. — **51**, N 2. — P. 303–312.
18. Ivanchuk M.A., Malyk I.V. Using ε-nets for linear separation of two sets in a euclidean space R^d// Cybernetics and Systems Analysis. — 2015. — **51**, N 6. — P. 965–968.
19. Embrechts P., Hofert M. A note on generalized inverses // Mathematical Methods of Operations Research. — 2013. — **77**, N 5. — P. 423–432.
20. Smith C.R. A characterization of star-shaped sets // American Mathematical Monthly. — 1968. — **75**, N 4. — P. 386.
21. Tucker H.G. A generalization of the Glivenko–Cantelli theorem // The Annals of Mathematical Statistics. — 1959. — **30**, N 3. — P. 828–830.
22. Dvoretzky A., Kiefer J., Wolfowitz J. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator // Annals of Mathematical Statistics. — 1956. — **27**, N 3. — P. 642–669.
23. Durrett R. Probability: Theory and examples. — Forth Edition, 2013. — 386 p.
24. Усольцев Л.П. Об асимптотике и больших уклонениях в центральной предельной теореме для сумм вида $\sum f(q^n t)$ // Вестник СамГУ. Естественнонаучная серия. — 2009. — Вып. 4 (70) — С. 52–84.

Надійшла до редакції 03.12.2015

М.А. Іванчук, І.В. Малик
РОЗВ'ЯЗАННЯ ЗАДАЧІ КЛАСИФІКАЦІЇ З ВИКОРИСТАННЯМ ε-СІТОК

Анотація. Запропоновано новий метод розв'язання задачі класифікації, що базується на відокремленні двох множин в просторі R^d шляхом побудови та відокремлення ε-сіток цих множин в ранжованому просторі відносно гіперплощин. Введено поняття області поділу — тих значень ε, при яких можливо відокремити множини. Наведено приклади області поділу для випадкових величин, розподілених за найбільш вживаними законами розподілу, та доведено теорему про її збіжність. Введено поняття сукупності всіх можливих ε-сіток деякої множини та доведено деякі її властивості. Доведена слабка збіжність нормованої різниці емпіричної та теоретичної кривих відокремлення до нормального розподілу, що дозволяє перевіряти гіпотези про місцезнаходження теоретичної кривої відокремлення в конкретній точці.

Ключові слова: ε-сітки, відокремлення множин, розмірність Вапніка–Червоненкіса, класифікація.

M.A. Ivanchuk, I.V. Malyk
SOLVING THE CLASSIFICATION PROBLEM USING ε-NETS

Abstract. The new method of the solution the classification problem is proposed in the paper. The method is based on separating two sets in the space R^d by constructing and separating ε-nets of these sets in a ranked space with respect to hyperplanes. The concept of the set of possible values of ε for ε-nets of both sets is introduced in the paper. The properties of this set and the theorem of its convergence are proved. The paper contains examples of the set of possible values for the most useful distributions. The concept of the set of all possible ε-nets of the set is introduced in the paper. Weak convergence of the normalized difference of the empiric and theoretic separation curves to the normal distribution is proved. It makes possible to check the hypothesis of the place of theoretic separation curve at a specific point.

Keywords: ε-nets, sets' separation, VC-dimension, classification.

Иванчук Мария Анатольевна,
ассистент кафедры Буковинского государственного медицинского университета, Черновцы, e-mail: mgracia@ukr.net.

Малык Игорь Владимирович,
кандидат физ.-мат. наук, доцент кафедры Черновицкого национального университета имени Юрия Федьковича, e-mail: malyk.igor.v@gmail.com.