

КЛАСТЕР-АНАЛИЗ МНОГОМЕРНЫХ СЛУЧАЙНЫХ НАБЛЮДЕНИЙ ПО ГИСТОГРАММНОЙ ОЦЕНКЕ ПЛОТНОСТИ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ

The problem of nonparametric cluster analysis of multivariate observations is considered. The new clustering procedure based on histogram as estimator of mixture probability density function is proposed and analytically and experimentally investigated.

Пусть в пространстве R^N регистрируются независимые в совокупности случайные вектора-наблюдения $x_1, \dots, x_n$ из $L > 2$ классов $\{\Omega_1, \dots, \Omega_L\}$. Наблюдение x_i ($i = 1, n$) принадлежит к классу со случайным ненаблюдаемым номером $d_i^0 \in S$, $S = \{1, \dots, L\}$:

$$P\{d_i^0 = i\} = \pi_i > 0, i \in S, \quad (1)$$

где $\{\pi_i\}_{i \in S}$ – так называемые априорные вероятности классов $\{\Omega_i\}_{i \in S}$ [1, 2]: $\pi_1 + \dots + \pi_L = 1$. При фиксированном номере класса $d_i^0 = i$ наблюдение x_i описывается условной плотностью распределения вероятностей

$$q_i(x) \geq 0, x \in R^N : \int_{R^N} q_i(x) dx = 1, i \in S. \quad (2)$$

Таким образом, классы $\{\Omega_i\}_{i \in S}$ однозначно определяются своими характеристиками $\{\pi_i, q_i(\cdot)\}_{i \in S}$, которые на практике обычно неизвестны [1–3].

Задача кластер-анализа состоит в классификации выборки $X = \{x_i\}_{i=1}^n$ объема n , т. е. в построении статистической оценки $\hat{D} = \hat{D}(X) = (\hat{d}_1, \dots, \hat{d}_n)^T \in S^n$ для неизвестного (ненаблюдаемого) вектора истинной классификации $D^0 = (d_1^0, \dots, d_n^0)^T \in S^n$ (здесь T – символ транспонирования). В отличие от традиционного подхода к задаче кластер-анализа [1, 2] число классов L считается неизвестным наряду с характеристиками классов $\{\pi_i, q_i(\cdot)\}_{i \in S}$.

Предположим, что области концентрации наблюдений в R^N (кластеры [1]), соответствующие классам $\{\Omega_i\}_{i \in S}$, “сформированы” в R^N вокруг локальных мод [4, 5] безусловной плотности-смеси:

$$q(x) = \sum_{i \in S} \pi_i q_i(x), x \in R^N, \quad (3)$$

описывающей наблюдения из выборки $X = \{x_i\}_{i=1}^n$.

Определение 1. Точка $\mu_i \in R^N$ называется локальной модой (ЛМ) для $q(\cdot)$, если μ_i – локальный максимум $q(x)$ по $x \in R^N$.

Определение 2. Локальная мода μ_i называется изолированной, если $\exists \varepsilon > 0$, что окрестность $M_i^\varepsilon = \{x: |x - \mu_i| \leq \varepsilon\}$ не содержит других ло-

кальных мод, отличных от μ_i ($\|v\| = \sqrt{v^T v}$ – евклидова норма вектора $v \in R^N$).

Везде далее будем предполагать, что плотность-смесь (3) имеет L локальных мод $\{\mu_i\}_{i \in S}$, и все они изолированные.

Определение 3. Теоретическим кластером (ТК) для изолированной ЛМ $\mu_i \in R^N$ назовем такое односвязное множество V_i из R^N ($V_i \subset R^N$), для которого выполняется:

$$\forall x, y \in V_i : \begin{cases} (x - \mu_i)^T (y - \mu_i) = |x - \mu_i| |y - \mu_i| (\overline{\mu_i x} \uparrow \overline{\mu_i y}), \\ |x - \mu_i| < |y - \mu_i| \end{cases} \Rightarrow q(x) > q(y). \quad (4)$$

Множество $V_0 = R^N \setminus \bigcup_{i \in S} V_i$ (необязательно односвязное) назовем “фоновым” ТК [4] для плотности-смеси (3).

Таким образом, задача кластер-анализа выборки $X = \{x_i\}_{i=1}^n$ сводится к построению оценок для $\{\mu_i, V_i\}_{i \in S}$ по X объема n при неизвестном числе классов L .

Воспользуемся непараметрической гистограммной оценкой [5, 6] для безусловной плотности-смеси $q(\cdot)$, построенной по неклассифицированной выборке $X = \{x_i\}_{i=1}^n$ объема n :

$$\hat{q}_n(x) = \frac{1}{n \prod_{j=1}^N h_j} \sum_{t=1}^n I_{\Gamma(x, h)}(x_t), \quad x \in R^N, \quad (5)$$

где $I_A(x) = \{1, \text{если } x \in A; 0, \text{если } x \notin A\}$ – индикатор множества A ; $\Gamma(x, h)$ – ячейка “объема” $\prod_{j=1}^N h_j$ гистограммы $\hat{q}_n(x)$, в которую попала точка $x = (\tilde{x}_1, \dots, \tilde{x}_N)^T \in R^N$:

$$\Gamma(x, h) = \times_{j=1}^N \left[\left[\frac{\tilde{x}_j}{h_j} \right] h_j, \left(\left[\frac{\tilde{x}_j}{h_j} \right] + 1 \right) h_j \right),$$

здесь $[z]$ – целая часть числа $z \in R$, а h_j – коэффициент “размытости” [5, 6] по j -й компоненте N -вектора $x \in R^N$.

Оценку $\hat{q}_n(\cdot)$ из (5) можно переписать в виде:

$$\hat{q}_n(x) = \frac{1}{n \prod_{j=1}^N h_j} \sum_{j_1, \dots, j_N} I_{\Gamma_{j_1, \dots, j_N}}(x) v_{j_1, \dots, j_N}, \quad (6)$$

где $v_{j_1, \dots, j_N} = v_{j_1, \dots, j_N}(X) = \sum_{t=1}^n I_{\Gamma_{j_1, \dots, j_N}}(x_t)$ – число наблюдений из выборки

$X = \{x_i\}_{i=1}^n$, попавших в ячейку $\Gamma_{j_1, \dots, j_N} = \times_{l=1}^N [j_l h_l, (j_l + 1) h_l)$ с “номером” $(j_1, \dots, j_N)$ ($j_l \in Z$, $Z = \{0, \pm 1, \pm 2, \dots\}$, $l = 1, N$).

Представление (6) и определения 1–3 позволяют построить алгоритм оценивания L и $\{\mu_i, \hat{V}_i\}_{i \in \hat{S}}$, основанный на методе полного перебора:

- 1) по выборке X объема n строим многомерную гистограмму (6);
- 2) просматриваем все ячейки $\Gamma_{j_1, \dots, j_N}$ гистограммы, которые непусты: $v_{j_1, \dots, j_N} \neq 0$, и помечаем те из них $\Gamma_{l_1, \dots, l_N}$, все соседние к которым ниже:

$$v_{l_1, \dots, l_N} > v_{j_1, \dots, j_N} \text{ при } \sum_{s=1}^N \delta_{|j_s - l_s|, 1} \neq 0,$$

где $\delta_{k,p} = \{1, \text{ если } k = p; 0, \text{ если } k \neq p\}$ – символ Кронекера. Помеченным ячейкам $\{\Gamma_{l_1, \dots, l_N}\}$ присваиваем номера от 1 до $\hat{L}$, где $\hat{L}$ – число помеченных ячеек (оценка числа классов L);

- 3) для каждой i -й помеченной на предыдущем этапе ячейки $\hat{\Gamma}_i$ ($i = \overline{1, \hat{L}}$) формируем оценку кластера $\hat{V}_i$:

а) полагаем $\hat{V}_i := \hat{\Gamma}_i$;

б) к $\hat{V}_i$ присоединяем те ячейки, не принадлежащие к $\hat{V}_i$, которые являются “соседями” каких-либо ячеек из $\hat{V}_i$ и ниже своих “соседей” из $\hat{V}_i$. Этот процесс повторяется до тех пор, пока $\hat{V}_i$ можно пополнить.

Отметим, что оценки $\{\hat{V}_i\}_{i \in \hat{S}}$, $\hat{S} = \{1, \dots, \hat{L}\}$, для ТК $\{V_i\}_{i \in S}$ не покрывают R^N , т. е. множество $\hat{V}_0 = R^N \setminus \bigcup_{i \in \hat{S}} \hat{V}_i$, являющееся оценкой для “фонового” кластера V_0 , вообще говоря, непусто. Поэтому классификацию выборки X необходимо производить в $\hat{L}+1$ класс, т. е. строить $\hat{D} = (\hat{d}_1, \dots, \hat{d}_{\hat{L}}) \in \hat{S}_0^n$, $\hat{S}_0 = \{0\} \cup \hat{S}$:

$$\hat{d}_t = \sum_{i \in \hat{S}} i \cdot I_{\hat{V}_i}(x_t), \quad t = \overline{1, n}. \quad (7)$$

При этом наблюдения x_t , не попавшие в основные кластеры $\{\hat{V}_i\}_{i \in \hat{S}}$, относятся к “фоновому” кластеру $\hat{V}_0$ ($\hat{d}_t = 0$) и при необходимости могут быть переклассифицированы в основные кластеры, например, при помощи следующего правила:

$$\hat{d}_t = \arg \min_{i \in \hat{S}} \rho(x_t, \hat{\mu}_i), \quad (8)$$

где $\rho(x, y) \geq 0$ – какая-либо метрика в R^N ($x, y \in R^N$), в частности евклидова: $\rho(x, y) = |x - y|$.

Однако недостатком алгоритма 1)–3) является то, что он не позволяет однозначно определить оценки $\{\hat{\mu}_i\}_{i \in \hat{S}}$ для ЛМ $\{\mu_i\}_{i \in S}$: $\mu_i \in \hat{\Gamma}_i$, $i \in \hat{S}$. В качестве $\hat{\mu}_i \in \hat{\Gamma}_i$ можно выбрать, например, “центр” ячейки $\hat{\Gamma}_i$:

$$\hat{\mu}_i = \left(\left(l_1 + \frac{1}{2} \right) \cdot h_1, \dots, \left(l_N + \frac{1}{2} \right) \cdot h_N \right)^T \in \hat{\Gamma}_i, \quad \hat{\Gamma}_i := \Gamma_{l_1, \dots, l_N}, \quad i \in \hat{S}. \quad (9)$$

Исследуем статистические свойства предложенной выше процедуры кластер-анализа, состоящей из процесса 1)–3) оценивания неизвестных характеристик классов по многомерной гистограмме (6) и непосредственно процедуры принятия решений (классификации) (7)–(9).

Теорема. Если плотность $q(\cdot)$ непрерывна и ограничена на всем R^N и имеет место асимптотика:

$$h_j = h_j(n) \rightarrow 0, \quad j = \overline{1, N}, \quad \prod_{j=1}^N n_j \cdot n \rightarrow +\infty, \quad n \rightarrow +\infty, \quad (10)$$

то гистограммная оценка плотности $\hat{q}_n(\cdot)$ является состоятельной:

$$\sup_{x \in R^N} |\hat{q}_n(x) - q(x)| \xrightarrow{P} 0, \quad n \rightarrow +\infty. \quad (11)$$

Замечание. Результат теоремы останется верным, если вместо непрерывности и ограниченности плотности-смеси $q(\cdot)$ из (3) потребовать, чтобы условные плотности $\{q_i(\cdot)\}_{i \in S}$, образующие смесь (3), были непрерывны и ограничены на R^N .

Доказательство. В условиях теоремы для доказательства справедливости (11) достаточно показать, что

$$\forall x \in R^N : \hat{q}_n(x) \xrightarrow{P} q(x), \quad n \rightarrow +\infty. \quad (12)$$

Для доказательства (12) воспользуемся законом больших чисел для серий [7]. Пусть серия независимых случайных величин $\xi_{1,n}, \dots, \xi_{n,n}$, распределения вероятностей которых зависят от n , удовлетворяет условиям:

- 1) $E\{\xi_{t,n}\} = 0, \quad t = \overline{1, n};$
- 2) $F_n(\tau) = \sum_{t=1}^n E\{|\xi_{t,n}| : |\xi_{t,n}| > \tau\} \rightarrow 0, \quad n \rightarrow +\infty, \quad \forall \tau;$
- 3) $\sum_{t=1}^n E\{|\xi_{t,n}|\} \leq c < +\infty, \quad n \rightarrow +\infty,$

тогда

$$\eta_n = \sum_{t=1}^n \xi_{t,n} \xrightarrow{P} 0, \quad n \rightarrow +\infty.$$

В нашем случае, исходя из (3), положим:

$$\xi_{t,n} = (I_{\Gamma(x,h)}(x_t) - \gamma(x,h)) / (\prod_{j=1}^N h_j \cdot n),$$

где $\gamma(x,h) = \int_{\Gamma(x,h)} q(y) dy$ – вероятность попадания наблюдения x_t в ячейку

$\Gamma(x,h)$, а коэффициенты “размытости” $\{h_j\}_{j=1}^N$ зависят от n : $h_j = h_j(n)$,

$j = \overline{1, N}$. Случайная величина η_n примет вид: $\eta_n = \hat{q}_n(x) - \gamma(x,h) / \prod_{j=1}^N h_j$. Но

по теореме о среднем в асимптотике (10): $\gamma(x,h) / \prod_{j=1}^N h_j \rightarrow q(x)$, поэтому

для доказательства (12) необходимо лишь проверить условия 1)–3). Очевидно, что условие 1) выполняется: $E\{\xi_{t,n}\} = 0, \quad t = \overline{1, n}$, поскольку

$E\{I_{\Gamma(x,h)}(x_t)\} = P\{x_t \in \Gamma(x,h)\} = \gamma(x,h)$. Проверим выполнение условий 2) и

3). Отметим, что случайные величины $\{\xi_{i,n}\}_{i=1}^n$, одинаково распределены по построению, поэтому

$$\sum_{i=1}^n E\{|\xi_{i,n}|\} = n E\{|\xi_{1,n}|\} = \frac{1}{\prod_{j=1}^N h_j} E\{|I_{\Gamma(x,h)}(x_1) - \gamma(x,h)|\} \leq \frac{2\gamma(x,h)}{\prod_{j=1}^N h_j}.$$

В силу теоремы о среднем: $\sum_{i=1}^n E\{|\xi_{i,n}|\} \leq 2q(z) < +\infty$ ($z \in \Gamma(x,h)$), и условие 3) выполняется. Проверим условие 2):

$$\begin{aligned} F_n(\tau) &= \sum_{i=1}^n E\{|\xi_{i,n}| : |\xi_{i,n}| > \tau\} = n E\{|\xi_{1,n}| : |\xi_{1,n}| > \tau\} = \\ &= \frac{1}{\prod_{j=1}^N h_j} \int_{|I_{\Gamma(x,h)}(y) - \gamma(x,h)| > \prod_{j=1}^N h_j \cdot n\tau} |I_{\Gamma(x,h)}(y) - \gamma(x,h)| q(y) dy, \end{aligned}$$

расширив область интегрирования и оценив подынтегральное выражение сверху, получим:

$$\begin{aligned} F_n(\tau) &< \frac{1}{\prod_{j=1}^N h_j} \int_{I_{\Gamma(x,h)}(y) + \gamma(x,h) > \prod_{j=1}^N h_j \cdot n\tau} (1 + \gamma(x,h)) q(y) dy < \\ &\leq \frac{2}{\prod_{j=1}^N h_j} P\left\{I_{\Gamma(x,h)}(x_1) > -\gamma(x,h) + \prod_{j=1}^N h_j \cdot n\tau\right\}. \end{aligned}$$

Воспользуемся неравенством Чебышева [7] и в асимптотике (10) будем иметь:

$$\begin{aligned} P\left\{I_{\Gamma(x,h)}(x_1) > -\gamma(x,h) + \prod_{j=1}^N h_j \cdot n\tau\right\} &\leq \frac{\gamma(x,h)}{-\gamma(x,h) + \prod_{j=1}^N h_j \cdot n\tau}; \\ F(\tau) &< \frac{2\gamma(x,h)}{\prod_{j=1}^N h_j \left(\prod_{j=1}^N h_j \cdot n\tau - \gamma(x,h)\right)} \rightarrow 0, \quad n \rightarrow +\infty, \quad \forall \tau, \end{aligned}$$

что совпадает с 3). Теорема доказана.

Следствие. Пусть в условиях теоремы в области $V \subset R^n$ единственная точка максимума $\mu = \arg \max_{x \in V} q(x)$, тогда $\forall \hat{\mu} \in \left\{\arg \max_{x \in V} \hat{q}_n(x)\right\}$, где $\hat{q}_n(\cdot)$ – гистограммная оценка (5) плотности $q(\cdot)$, выполняется:

$$\hat{\mu} \xrightarrow{P} \mu, \quad n \rightarrow +\infty.$$

Доказательство непосредственно следует из (11) и известной теоремы непрерывности [6], применяемой с учетом соотношения:

$$\begin{aligned} |\hat{q}_n(\hat{\mu}) - q(\mu)| &= \left| \max_{x \in V} \hat{q}_n(x) - \max_{y \in V} q(y) \right| < \\ &\leq \max_{x \in V} |\hat{q}_n(x) - q(x)| \leq \sup_{x \in R^n} |\hat{q}_n(x) - q(x)| \xrightarrow{P} 0, \\ &n \rightarrow +\infty. \end{aligned}$$

Следствие к теореме позволяет установить в условиях асимптотики (10) состоятельность статистических оценок $\{\hat{\mu}_i\}_{i \in \hat{S}}$ для ЛМ $\{\mu_i\}_{i \in S}$ (ЛМ, по нашему предположению, изолированные), полученных по алгоритму 1)–3):

$$\hat{\mu}_i \xrightarrow{P} \mu_i, \quad n \rightarrow +\infty, \quad (13)$$

что говорит о достаточной точности кластер-процедуры, на них основанной. Более того, для классификации можно использовать не (7)–(9), а классифицировать все n наблюдений из выборки $X = \{x_i\}_{i=1}^n$ по близости к оценкам ЛМ $\{\hat{\mu}_i\}_{i \in \hat{S}}$, т. е. применять (8), (9) ко всем наблюдениям, не строя при этом оценок $\{\hat{V}_i\}_{i \in \hat{S}}$ для ТК $\{V_i\}_{i \in S}$ и избегая образования $(\bar{L}+1)$ -го “фонового” кластера.

В качестве иллюстрации полученных результатов рассмотрим часто используемую на практике модель смеси многомерных нормальных (гауссовских) распределений (модель Фишера [1]), когда условные плотности $\{q_i(\cdot)\}_{i \in S}$ из (2) многомерные гауссовские:

$$q_i(\cdot) \equiv n_N(\cdot | \mu_i^0, \Sigma), \quad i \in S; \quad (14)$$

$$n_N(x | \mu_i^0, \Sigma) = (2\pi)^{-\frac{N}{2}} (\det(\Sigma))^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu_i^0)^T \Sigma^{-1} (x - \mu_i^0)\right),$$

где $\mu_i^0 \in K^N$ – вектор математического ожидания для наблюдений из класса Ω_i (“центр” i -го класса), а Σ – общая для всех классов невырожденная ковариационная $(N \times N)$ -матрица ($\det(\Sigma) \neq 0$). В качестве метрики в (8) в данном случае целесообразно использовать “расстояние Махаланобиса” [1]:

$$\rho(x, y) = \sqrt{(x - y)^T \Sigma^{-1} (x - y)}, \quad x, y \in R^N, \quad (15)$$

заменяя неизвестную ковариационную матрицу Σ на какую-либо статистическую оценку [1–3, 6] по классифицируемой выборке $X = \{x_i\}_{i=1}^n$.

С целью экспериментального исследования эффективности кластер-процедуры 1)–3), (7)–(9) проведен вычислительный эксперимент. В модели Фишера (1), (14), (15) рассмотрен случай двух ($L=2$) равновероятных ($\pi_1 = \pi_2 = 0,5$) классов $\{\Omega_1, \Omega_2\}$ в пространстве наблюдений размерности $N=2$ с межклассовым расстоянием Махаланобиса $\Delta = \rho(\mu_1^0, \mu_2^0) = 2,56$, которому соответствует минимально возможное (байесовское) значение вероятности ошибочной классификации [1] $r_0 = \Phi\left(-\frac{\Delta}{2}\right) = 0,1$, где $\Phi(\cdot)$ – функция распределения вероятностей нормального закона с плотностью $n_1(\cdot | 0, 1)$.

Методом Монте-Карло были смоделированы $m=10$ выборок объемом $n=20$ каждая и проведена их классификация. Использовались две модификации описанного выше метода, основанного на гистограммном подходе. При 1-й модификации классификация проводилась с помощью (7)–(9), в то время как при 2-й модификации применялись сразу (8), (9) ко всем наблюдениям. Для сравнительного анализа был реализован также хорошо известный метод “L-средних” [1, 2], основанный на итерационном повторе (8) при

известном числе классов L и начальном задании приближений для “центров” классов $\{\mu_i^{(0)}\}_{i \in S}$.

В качестве характеристики эффективности использовалась доля ошибочных решений [3]:

$$\gamma_n = \frac{1}{n} \sum_{i=1}^n \left(1 - \delta_{\hat{d}_i, d_i^0} \right). \quad (16)$$

Вычислялась также средняя доля ошибочных решений по всем выборкам:

$$\bar{\gamma}_n = \frac{1}{\sum_{l=1}^m \delta_{\hat{L}^{(l)}, L}} \sum_{l=1}^m \delta_{\hat{L}^{(l)}, L} \cdot \gamma_n^{(l)}, \quad (17)$$

где $\gamma_n^{(l)}$ – показатель (16), вычисленный для l -й выборки, если число классов определено верно (при гистограммном подходе: $m^* = m - \sum_{l=1}^m \delta_{\hat{L}^{(l)}, L}$ – число выборок, для которых число классов определено неверно) или задано априори ($\hat{L}^{(l)} := L$, $l = 1, m$, – в методе “ L -средних”).

В результате численного эксперимента метод, основанный на гистограммном подходе, ошибся при определении числа классов лишь для трех выборок из десяти ($m^* = 3$), при этом средняя доля ошибочных решений (17) для 1-й модификации метода составила: $\gamma_n = 0,06$, а для 2-й: $\gamma_n = 0,00$. Средняя доля ошибочных решений в методе “ L -средних”: $\gamma_n = 0,00$, оказалась меньше, чем в методе, основанном на гистограммном подходе. Однако следует учесть, что метод “ L -средних” требует задания дополнительной априорной информации: должны быть заданы количество классов и начальные приближения для их “центров”. В методе, основанном на гистограммном подходе, эта информация не требуется: классификация производится только с использованием экспериментальных статистических данных.

Таким образом, разработанный алгоритм близок по точности классификации к известному методу “ L -средних”, однако область его практического применения значительно шире.

1. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности. М., 1989.

2. Миленький А.В. Классификация сигналов в условиях неопределенности. М., 1975.

3. Жук Е.Е., Харин Ю.С. Устойчивость в кластер-анализе многомерных наблюдений. Мн., 1998.

4. Коваленко А.П. // Автоматика и телемеханика. 1993. № 5. С. 100.

5. Жук Е.Е. // Актуальные проблемы информатики: математическое, программное и информационное обеспечение: Материалы V Межгос. науч. конф. Мн., 1996. С. 95.

6. Боровков А.А. Математическая статистика. М., 1984.

7. Боровков А.А. Теория вероятностей. М., 1986.

Поступила в редакцию 19.10.2000.

Жук Евгений Евгеньевич – доктор физико-математических наук, доцент кафедры математического моделирования и анализа данных БГУ.

Храмова Елизавета Владимировна – студентка 4-го курса ФПМИ БГУ.

