

РОБАСТНОЕ РЕШАЮЩЕЕ ПРАВИЛО L-СРЕДНИХ В СЛУЧАЕ РАВНОМЕРНОГО «ЗАСОРЕНИЯ»

The problem of robust L-means decision rule construction under uniform «contaminations» is solved using Bayesian optimality principle. The risk functional value for this decision rule is evaluated.

На практике очень часто встречается ситуация, когда области концентрации наблюдений из $L \geq 2$ классов $\{\Omega_1, \dots, \Omega_L\}$ (кластеры [1]) «равномерно окружены» в R^N наблюдениями-«засорениями». В теории кластер-анализа для решения задачи классификации при данных модельных предположениях используются методы так называемой «нечеткой» классификации [1, 2]: каждому наблюдению ставится в соответствие L-вектор вероятностей, с которыми данное наблюдение относится к тому или иному классу. Недостатком таких методов является то, что относительно наблюдения не делается однозначного вывода о принадлежности его к конкретному классу или «засорениям».

Искажения типа равномерного «засорения» удобно определять, вводя в рассмотрение дополнительный класс Ω_0 , содержащий «засорения», имеющий априорную вероятность $\epsilon (0 \leq \epsilon \leq \epsilon_+ < 1)$ и описываемый равномерной плотностью распределения вероятностей:

$$h_*(x) = \begin{cases} h \frac{1}{\mu(V)}, & x \in V, \\ 0, & x \notin V. \end{cases} \quad (1)$$

Таким образом, наблюдения-«засорения» из класса Ω_0 равномерно распределены в ограниченной односвязной области $V(\mu(\cdot))$ — мера Лебега в R^N). Класс Ω_i состоит из неискаженных наблюдений, описываемых плотностью $q(\cdot; \theta_i)$, $x \in R^N$, из параметрического семейства

$$P = \{q(x; \theta), x \in R^N; \theta \in R^N\} \quad (2)$$

с параметром сдвига $\theta_i^0 \in R^N$, равным математическому ожиданию:

$$\int_{R^N} x q(x; \theta_i^0) dx = \theta_i^0, i \in S = \{1, \dots, L\},$$

и имеет априорную вероятность $\pi_i(1 - \epsilon)$, $i \in S (\pi_1 + \dots + \pi_L = 1)$.

Через d_i^0 обозначим неизвестный номер класса, к которому принадлежит наблюдение $x_i \in R^N$, $i = \overline{1, n}$. Составной вектор «центров классов» (математических ожиданий) $\theta^0 = (\theta_1^{0T} | \dots | \theta_L^{0T})^T \in R^{LN}$ неизвестен. Задача заключается в оценивании вектора истинной классификации $D^0 = (d_1^0, \dots, d_n^0)^T$ по случайной выборке $X = (x_1 | \dots | x_n)^T$ объема n .

При отсутствии искажений ($\epsilon_+ = 0$) в кластер-анализе широко используется алгоритм L-средних [1, 3], состоящий в совместном вычислении оценок $\hat{\theta}$ и $\hat{D}$ для θ^0 и D^0 посредством решения экстремальной задачи:

$$W_n(\theta, D) \rightarrow \inf_{\theta \in R^{LN}, D \in S^n}; \quad W_n(\theta, D) = n^{-1} \sum_{i=1}^n \|x_i - \theta_{d_i}\|^2,$$

где $\|\cdot\|$ — какая-либо векторная норма. Однако при $\epsilon_+ > 0$ он теряет свои оптимальные свойства. Более того, на практике важен случай, когда $\epsilon_+ = O(1)$, т. е. наблюдений-«засорений» в выборке X объема n может быть достаточно много (их количество является величиной порядка $O(n)$).

Получим робастное (устойчивое к искажениям типа (1)) решающее правило (РП) L-средних. Пусть сначала составной вектор $\theta^0 = (\theta_1^{0T} | \dots | \theta_L^{0T})^T$ «центров классов» $\{\Omega_i\}_{i \in S}$ и область V известны априори. Рассмотрим задачу построения оптимального в смысле минимума риска (вероятности ошибочной классификации наблюдения из X) байесовского РП (БРП) [1, 4] для классификации наблюдений в $L + 1$ класс $\{\Omega_0, \Omega_1, \dots, \Omega_L\}$.

Через $U(\cdot)$ обозначим единичную функцию Хэвисайда:

$$U(z) = \begin{cases} 1, & z \geq 0, \\ 0, & z < 0. \end{cases}$$

Теорема 1. Пусть классы $\{\Omega_i\}_{i \in S}$ равновероятны: $\pi_i = L^{-1}$, $i \in S$, а параметрическое семейство (2) имеет вид:

$$P = \{q(x; \theta) = Q(\|x - \theta\|), x \in \mathbb{R}^N; \theta \in \mathbb{R}^N\}, \quad (3)$$

где функция $Q(z): [0, +\infty) \rightarrow (0, Q(0)]$ непрерывная и строго убывающая. Тогда БРП $d_*(\cdot; \theta^0; V)$ задается соотношением:

$$d_*(x; \theta^0; V) = \arg \min_{0 \leq k \leq L} g_k(x); \quad (4)$$

$$g_i(x) = \begin{cases} \|x - \theta_i^0\|, & i \in S, \\ +\infty, & i = 0, x \notin V, \\ Q^{-1}\left(\frac{\epsilon L}{(1-\epsilon)\mu(V)}\right) \cdot U\left(Q(0) - \frac{\epsilon L}{(1-\epsilon)\mu(V)}\right), & i = 0, x \in V, \end{cases}$$

($Q^{-1}(\cdot)$ — функция, обратная к $Q(\cdot)$) и обладает минимальным значением риска $R_*(\theta^0; V)$:

$$R_*(\theta^0; V) = 1 - \int_{\mathbb{R}^N} \max \left\{ \frac{1-\epsilon}{L} \cdot \max_{i \in S} q(x; \theta_i^0), \epsilon h_*(x) \right\} dx. \quad (5)$$

Д о к а з а т е л ь с т в о. Согласно [1, 4], БРП, минимизирующее вероятность ошибочной классификации, имеет вид:

$$d_*(x; \theta^0; V) = \arg \max_{0 \leq k \leq L} \bar{\pi}_k \bar{q}_k(x);$$

$$\bar{\pi}_i = \begin{cases} L^{-1}(1-\epsilon), & i \in S, \\ \epsilon, & i = 0; \end{cases}$$

$$\bar{q}_i(x) = \begin{cases} q(x; \theta_i^0), & i \in S, \\ h_*(x), & i = 0. \end{cases}$$

Воспользуемся (3), конкретным видом (1) плотности $h_*(\cdot)$, тем фактом, что строго монотонная непрерывная функция имеет обратную, и получим (4). Из [1, 4] имеем:

$$R_*(\theta^0; V) = 1 - \int_{\mathbb{R}^N} \max_{0 \leq k \leq L} \bar{\pi}_k \bar{q}_k(x) dx,$$

что с учетом конкретного вида $\{\bar{\pi}_i \bar{q}_i(\cdot)\}_{i=0}^L$ приводит к (5). Теорема доказана.

Замечание 1. Правило (4) можно применять и при $L = 1$. В этом случае решается задача удаления наблюдений-«засорений» из выборки, порожденной «гипотетическим» распределением с плотностью $q(x; \theta_1^0)$ и содержащей «примесь» из наблюдений с плотностью (1).

Предположим, что кластеры (области концентрации неискаженных наблюдений):

$$w_i = \{x: q(x; \theta_i^0) \geq v > 0, v = o(1)\}, \quad i \in S,$$

принадлежат области V , в которой равномерно распределены «засорения». Причем V является прямоугольным параллелепипедом в $\mathbb{R}^N$:

$$V = \{x: x \in [a_1, b_1] \times \dots \times [a_N, b_N], a_i < b_i, i = \overline{1, N}\};$$

$$\mu(V) = \prod_{j=1}^N (b_j - a_j),$$

где « $\times$ » — символ декартова произведения. В этом случае можно предложить следующий алгоритм, реализующий робастное РП L -средних $d_*(\cdot; \theta; \hat{V})$, получающееся из РП $d_*(\cdot; \theta^0; V)$, приведенного в (4), в

результате замены θ^0 , V на их статистические оценки $\hat{\theta}$, $\hat{V}$, вычисленные по выборке X объема n .

По X строится оценка $\mu(\hat{V})$ для $\mu(V)$:

$$\mu(\hat{V}) = \left(\frac{n+1}{n-1} \right)^N \prod_{j=1}^N \left(\max_{1 \leq k \leq n} x_{kj} - \min_{1 \leq k \leq n} x_{kj} \right),$$

получающаяся из $\mu(V)$ путем замены $(b_j - a_j)$ на статистическую оценку $(b_j \wedge a_j)$ [5]:

$$(b_j \wedge a_j) = \frac{n+1}{n-1} \cdot \left(\max_{1 \leq k \leq n} x_{kj} - \min_{1 \leq k \leq n} x_{kj} \right), j = \overline{1, N}. \quad (6)$$

Задается уровень «засорения» ϵ и проверяется условие:

$$\frac{\epsilon L}{(1-\epsilon)\mu(\hat{V})} > Q(0). \quad (7)$$

Если оно выполняется, то выборка X не классифицируется, поскольку, согласно (4), данные, из которых она состоит, не имеют ярко выраженной кластерной структуры (принимается решение: $\hat{d}_t = 0$, $t = \overline{1, n}$). В противном случае задаются начальные приближения $\{\hat{\theta}_i^{(0)}\}_{i \in S}$ для центров $\{\theta_i^0\}_{i \in S}$, и классов $\{\Omega_i\}_{i \in S}$, и строится следующий итерационный процесс:

$$\hat{d}_t^{(k)} = \arg \min_{0 \leq k \leq L} \hat{g}_i(x_t), t = \overline{1, n}; k = 1, 2, \dots; \quad (8)$$

$$\hat{g}_i(x) = \begin{cases} \|x - \hat{\theta}_i^{(k-1)}\|, & i \in S, \\ Q^{-1}\left(\frac{\epsilon L}{(1-\epsilon)\mu(\hat{V})}\right), & i = 0; \end{cases}$$

$$\hat{\theta}_i^{(k)} = \left(\sum_{t=1}^n \delta_{it}, \hat{d}_t^{(k)} \right)^{-1} \cdot \sum_{t=1}^n \delta_{it} \hat{d}_t^{(k)} x_t, i \in S. \quad (9)$$

Процесс (8), (9) завершается через конечное число шагов [1, 3] при $\hat{D}^{(k)} = \hat{D}^{(k-1)}$ ($k \geq 2$) и получается: $\hat{D} = \hat{D}^{(k)}$ — оценка для вектора истинной классификации D^0 выборки X , $\hat{\theta} = \hat{\theta}^{(k)}$ — оценка для составного вектора θ^0 .

Замечание 2. Из (6) видно, что все наблюдения, образующие выборку X , принадлежат области

$$\hat{V} = \left[\min_{1 \leq k \leq n} x_{11}, \max_{1 \leq k \leq n} x_{11} \right] \times \dots \times \left[\min_{1 \leq k \leq n} x_{1N}, \max_{1 \leq k \leq n} x_{1N} \right],$$

и случай $x_i \notin \hat{V}$ в (8) исключается.

Если $\epsilon = 0$, то алгоритм (8), (9) превращается в свой классический аналог [1, 3]:

$$\hat{d}_t^{(k)} = \arg \min_{i \in S} \|x_t - \hat{\theta}_i^{(k-1)}\|, t = \overline{1, n}; k = 1, 2, \dots;$$

$$\hat{\theta}_i^{(k)} = \left(\sum_{t=1}^n \delta_{it}, \hat{d}_t^{(k)} \right)^{-1} \cdot \sum_{t=1}^n \delta_{it} \hat{d}_t^{(k)} x_t, i \in S,$$

поскольку, согласно условиям теоремы 1, $Q^{-1}(0) = +\infty$.

Если уровень «засорения» ϵ неизвестен априори, то итерационный процесс (8), (9) можно модифицировать: задается начальное приближение $\epsilon^{(0)}$ и на k -ой итерации в (8) используется $\epsilon^{(k-1)}$ после предварительной проверки условия (7), затем, после уточнения центров классов (9), производится пересчет текущего значения уровня «засорения»:

$$\epsilon^{(k)} = n^{-1} \sum_{t=1}^n \delta_{0, \hat{d}_t^{(k)}}.$$

Рассмотрим частный случай — модель Фишера [1, 4]:

$$q(x; \theta_i^0) = n_N(x | \theta_i^0, \Sigma), i \in S; \quad (10)$$

$p_N(x|\theta_1^0, \Sigma) = (2\pi)^{-N/2} (\det(\Sigma))^{-0.5} \exp(-0.5(x - \theta_1^0)^T \Sigma^{-1}(x - \theta_1^0))$,
в которой наряду с $\{\theta_{ij}^0\}_{j \in S}$ неизвестна и ковариационная матрица Σ . Если в качестве нормы выбрать:

$$\|y\| = \sqrt{y^T \Sigma^{-1} y}, \quad y \in \mathbb{R}^N, \quad (11)$$

то функция $Q(z)$, определенная в (3), примет вид:

$$Q(z) = (2\pi)^{-N/2} (\det(\Sigma))^{-0.5} \exp(-z^2/2). \quad (12)$$

С учетом (11), (12), вычислительный алгоритм (8), (9) легко модифицируется. В качестве оценки $\hat{\Sigma}$ для Σ можно использовать любую робастную оценку [6], являющуюся положительно определенной матрицей.

Пусть $L = 1$ (смотри замечание 1), ковариационная матрица в (10) имеет специальный вид: $\Sigma = \sigma^2 I_N$ ($\sigma > 0$), а область V является N -мерным кубом, «диагонали» которого пересекаются в точке θ_1^0 , и «сторона» имеет длину, равную $2a$ ($a > 0$):

$$V = [\theta_{11}^0 - a, \theta_{11}^0 + a] \times \dots \times [\theta_{1N}^0 - a, \theta_{1N}^0 + a],$$

где θ_{1j} — j -ая компонента вектора θ_1 , $j = \overline{1, N}$.

Воспользуемся (7), (12), а также специальным видом Σ и V , получим соотношение:

$$\frac{a}{\sigma} > \sqrt{\frac{\pi}{2} \cdot \left(\frac{\epsilon}{1-\epsilon}\right)^{1/N}}, \quad (13)$$

которому должны удовлетворять a , σ , ϵ и N , чтобы задача классификации имела смысл.

Предположим, что $0 < \epsilon \ll 1 \ll 0.5$, тогда условие (13) можно усилить:

$$\frac{a}{\sigma} > \sqrt{\frac{\pi}{2}}. \quad (14)$$

Соотношением (14) удобно пользоваться для определения допустимого значения отношения $\frac{a}{\sigma}$ при неизвестном уровне «засорения», не превышающем 0.5.

Вычислим риск (5) РП (4). Обозначим через V' следующую вспомогательную область:

$$V' = \left\{ x : \frac{1-\epsilon}{\epsilon} \cdot p_N(x|\theta_1^0, \sigma^2 I_N) \geq \frac{1}{\mu(V)} \right\} = \left\{ x : (x - \theta_1^0)^T (x - \theta_1^0) \sigma^{-2} \leq p^2 \right\},$$

$$p = \sqrt{2 \cdot \left(\ln \frac{1-\epsilon}{\epsilon} + N \cdot \ln \left(\frac{a}{\sigma} \cdot \sqrt{\frac{2}{\pi}} \right) \right)}. \quad (15)$$

Пусть выполняется условие:

$$p \leq a, \quad (16)$$

тогда $V' \subset V$, и при помощи таблиц интегралов [7] имеем:

$$\begin{aligned} R_*(\theta^0; V) = & \epsilon \cdot \left(\frac{\sqrt{\pi}}{2} \right)^N \cdot \frac{(p/a)^N}{\Gamma\left(\frac{N}{2} + 1\right)} + \\ & + (1-\epsilon) \left(\left(1 - 2\Phi(-a/\sigma) \right)^N - \frac{\tilde{\Gamma}\left(\frac{N}{2}, \frac{p^2}{2}\right)}{\Gamma\left(\frac{N}{2}\right)} \right), \end{aligned} \quad (17)$$

где

$$\Phi(u) = \int_{-\infty}^u (2\pi)^{-1/2} \exp(-z^2/2) dz —$$

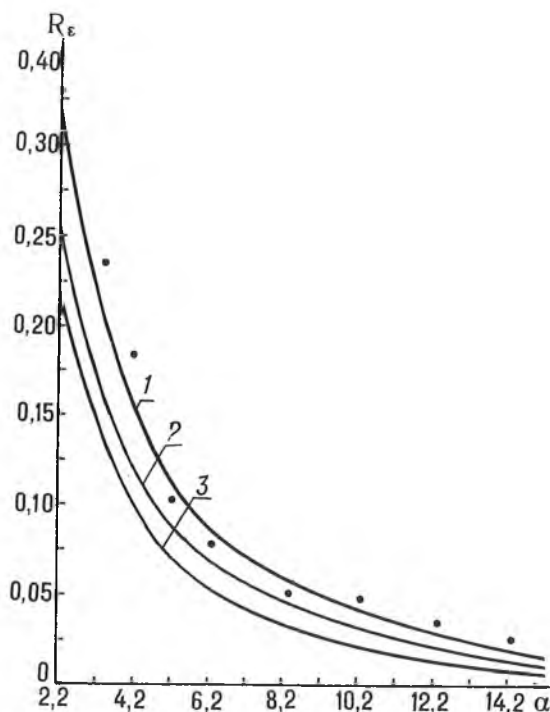
функция распределения вероятностей стандартного нормального закона;

$$\tilde{\Gamma}(\alpha, \beta) = \int_0^{\beta} e^{-z} z^{\alpha-1} dz \quad (\alpha > 0) \quad (18)$$

неполная гамма-функция ($\tilde{\Gamma}(\alpha, +\infty) = \Gamma(\alpha)$ — обычная гамма-функция). Из (15) видно, что $p = o(a)$ при $a \rightarrow +\infty$, поэтому условие (16) заведомо выполняется, а из (18) для (17) имеем:

$$R_*(\theta^0; V) \rightarrow 0, \quad a \rightarrow +\infty$$

т. е. с ростом объема области V качество классификации повышается.



Риск робастного РП L-средних в случае равномерного «засорения»:

1 — $\epsilon = 0.5$; 2 — $\epsilon = 0.3$; 3 — $\epsilon = 0.25$

На рисунке изображены графики зависимости $R_* = R_*(a)$, построенные при помощи формулы (17) для следующих конкретных значений: $N = 2$; $\sigma = 1.0$; $\epsilon = 0.25$; 0.3 ; 0.5 . Величина a изменяется на отрезке $[2.2, 15.2]$, что обеспечивает выполнение условий (14), (16). Видно, что с ростом уровня «засорения» ϵ вероятность ошибочной классификации R_* увеличивается. Символом «•» на рисунке показаны значения показателя:

$$\hat{\gamma} = \frac{1}{n \cdot M} \sum_{j=1}^M \sum_{t=1}^n \delta_{d_t}^0(j), \quad d_t(j), \quad (19)$$

подсчитанные для $\epsilon = 0.5$ по результатам вычислительных экспериментов при различных значениях a (в качестве Σ использовалась робастная оценка медианного типа [6]). Каждый такой эксперимент включал в себя моделирование $M = 10$ выборок объема $n = 20$ с нулевым вектором математического ожидания $\theta_1^0 = (0, 0)^T$, нахождение значения величины (19) по результатам классификации и нанесение его на рисунок. Как следует из рисунка, риск R_* достаточно точно характеризует величину:

$$\bar{\gamma} = E \left\{ \frac{1}{n} \sum_{t=1}^n \delta_{d_t}^0, \quad d_t \right\}$$

(значения $\hat{\gamma}$ расположены вблизи графика) и может быть использован при $n \rightarrow +\infty$ как показатель эффективности РП L-средних $d_*(\cdot; \theta; V)$.

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

2. Стирака А. И. Решение кластерной задачи большой размерности в нечеткой постановке // Кибернетика. 1991. № 1. С. 116.

3. Миленский А. В. Классификация сигналов в условиях неопределенности. М., 1975.

4. Харин Ю. С. Робастность в статистическом распознавании образов. Мн., 1992.

5. Справочник по теории вероятностей и математической статистике / Под ред. В. С. Корольюка. М., 1985.

6. Хьюбер П. Робастность в статистике. М., 1984.

7. Градштейн И. С., Рыжик И. М. Таблицы интегралов, сумм, рядов и произведений. М., 1963.

Поступила в редакцию 09.06.94.

УДК 517.948.32:517.544

О. В. ТИМОХОВИЧ

ПОСТРОЕНИЕ КАНОНИЧЕСКОЙ МАТРИЦЫ ОДНОЙ ЗАДАЧИ ЛИНЕЙНОГО СОПРЯЖЕНИЯ С ПОДСТАНОВОЧНОЙ МАТРИЦЕЙ ЧЕТВЕРТОГО ПОРЯДКА

The canonical matrix is being constructed explicitly by means of reduction of the original problem to scalar problem on Riemann surface.

Рассмотрим следующую задачу: на расширенной комплексной плоскости S с разрезом по отрезку $L = [-1, 1]$ найти кусочно-голоморфный вектор $\Phi(z) = (\Phi_1(z), \Phi_2(z), \Phi_3(z), \Phi_4(z))$, предельные значения которого удовлетворяют следующему условию:

$$\Phi^+(t) = G(t)\Phi^-(t), \quad t \in L \setminus \Lambda, \quad (1)$$

где

$$G(t) = \begin{cases} \begin{bmatrix} f_1(t) & 0 & 0 & 0 \\ 0 & 0 & 0 & f_2(t) \\ 0 & 0 & f_3(t) & 0 \\ 0 & f_4(t) & 0 & 0 \end{bmatrix}, & t \in (-1, 0), \\ \begin{bmatrix} 0 & f_1(t) & 0 & 0 \\ f_2(t) & 0 & 0 & 0 \\ 0 & 0 & 0 & f_3(t) \\ 0 & 0 & f_4(t) & 0 \end{bmatrix}, & t \in (0, 1); \end{cases}$$

$f_j(t)$, $j = \overline{1, 4}$ — гёльдеровские, нигде не обращающиеся в нуль функции, имеющие конечные односторонние пределы при $z \rightarrow -1, 0, 1$, $z \in L \setminus \Lambda$, также не обращающиеся в нуль, $\Lambda = \{-1, 0, 1\}$.

Задачу (1) будем решать в классе функций, ограниченных в окрестностях точек Λ . Если $f_j(t) \equiv 1$, $j = \overline{1, 4}$, то рассматриваемая задача эквивалентна задаче построения основных функционалов римановой поверхности, являющейся четырехлистным накрытием сферы. При этом точки $z = -1, 0, 1$ являются проекциями точек ветвления данного накрытия, а подстановки

$$\sigma_1 = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 3 & 2 \end{bmatrix}, \quad \sigma_2 = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{bmatrix}, \quad \sigma_3 = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{bmatrix}$$

описывают закон склеивания листов накрытия в окрестностях точек $-1, 0, 1$ соответственно.

Алгебраическое уравнение, соответствующее указанному накрытию, построено в [1]. Оно имеет следующий вид:

$$w^4 - 4zw^3 + \gamma zw^2 - 2(\gamma - 4)zw + \frac{1}{4}(\gamma - 4)^2 z = 0, \quad (2)$$

где $\gamma \in (-4, -3)$ — корень уравнения $\gamma^4 - 12\gamma^3 - 54\gamma^2 - 32\gamma - 64 = 0$. Род