НЕПАРАМЕТРИЧЕСКИЙ КЛАСТЕР-АНАЛИЗ СМЕСИ МНОГОМЕРНЫХ УНИМОДАЛЬНЫХ РАСПРЕДЕЛЕНИЙ

Е.Е. Жук, И.О. Гончарик

Белорусский государственный университет,
кафедра математического моделирования и анализа данных
пр. Независимости, 4, г. Минск, Беларусь
телефон: + (017) 2095530; e-mail: zhukee@mail.ru
web: www.bsu.by

Рассматривается задача кластер-анализа многомерных наблюдений, описываемых смесью унимодальных вероятностных распределений. Для классификации предлагается использовать алгоритм, основанный на непараметрическом оценивании безусловной плотности-смеси, описывающей наблюдения. В качестве оценки плотности используется многомерная гистограмма с прямоугольными ячейками, и указывается асимптотика, связывающая параметры ячеек с объемом выборки.

Ключевые слова – кластер-анализ, унимодальные распределения, плотность-смесь, гистограмма.

1 МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ВЫБОРКИ И ПОСТАНОВКА ЗАДАЧИ КЛАСТЕР-АНАЛИЗА

В пространстве R^N регистрируются независимые в совокупности случайные наблюдения $x_1,\dots,x_n$ из $L \geq 2$ классов $\{\Omega_1,\dots,\Omega_L\}$. Наблюдение $x_t \in R^N \left(t=\overline{1,n}\right)$ принадлежит к классу со случайным ненаблюдаемым номером $d_t^0 \in S$, $S=\{1,\dots,L\}$:

$$P\{d_t^0 = i\} = \pi_i > 0, \quad i \in S, \tag{1}$$

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

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

где плотности (2) считаются унимодальными: каждая имеет один локальный максимум (моду), являющийся и глобальным.

В этом случае естественно предположить, что области концентрации $\{V_i \subset R^N\}_{i\in S}$ наблюдений в R^N (кластеры [1,2]), соответствующие классам $\{\Omega_i\}_{i\in S}$, "сформированы" в R^N вокруг локальных мод $\{\mu_i \in V_i \subset R^N\}_{i\in S}$ безусловной плотности-смеси:

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

которые считаются изолированными друг от друга:

$$\begin{gathered} \exists\ \ \varepsilon > 0:\ M_i^\varepsilon \cap M_j^\varepsilon = \varnothing,\ i \neq j \in S; \\ M_i^\varepsilon = \{x : |x-\mu_i| \leq \varepsilon\}. \end{gathered} \tag{4}$$

Задача кластер-анализа [1,2] заключается в классификации выборки $X=\{x_t\}_{t=1}^n$ объема n, т.е. в построении по ней статистической оценки $\hat{D} = \hat{D}(X) = \left(\hat{d}_1,\dots,\hat{d}_n\right)^{\mathrm{T}} \in \hat{S}^n$, $\hat{S} = \{1,\dots,\hat{L}\}$, для неизвестного вектора истинной классификации $D^0 = \left(d_1^0,\dots,d_n^0\right)^{\mathrm{T}} \in S^n$ ("T" – символ транспонирования). Число классов L здесь считается неизвестным наряду с характеристиками классов $\{\pi_i, q_i(\cdot)\}_{i\in S}$.

2 МНОГОМЕРНАЯ ГИСТОГРАММА И АЛГОРИТМ КЛАСТЕР-АНАЛИЗА

Сведем задачу кластер-анализа выборки $X=\{x_t\}_{t=1}^n$ к построению по ней статистических оценок для $\{\mu_i, V_i\}_{i\in S}$ при неизвестном числе классов L. В качестве статистической оценки для безусловной плотности-смеси $q(\cdot)$ из (3) будем использовать многомерную гистограмму с прямоугольными ячейками [3], построенную по $X=\{x_t\}_{t=1}^n$ объема n:

$$\hat{q}_n(x) = \frac{1}{n \cdot \prod_{j=1}^N h_j} \sum_{j_1,\dots,j_N} \mathrm{I}_{\Gamma_{j_1,\dots,j_N}}(x)\, \nu_{j_1,\dots,j_N}, \tag{5}$$

где $\nu_{j_1,\dots,j_N} = \nu_{j_1,\dots,j_N}(X) = \sum_{t=1}^n \mathrm{I}_{\Gamma_{j_1,\dots,j_N}}(x_t)$ – число наблюдений из выборки $X=\{x_t\}_{t=1}^n$, попавших в ячейку $\Gamma_{j_1,\dots,j_N} = \underset{l=1}{\overset{N}{\times}} [j_l h_l, (j_l+1)\cdot h_l)$ с "номером" $(j_1,\dots,j_N)$ ($j_l \in Z$, $Z=\{0,\pm1,\pm2,\dots\}$, $l=\overline{1,N}$), а h_l – коэффициент

"размытости" [3] по l-й компоненте N-вектора-аргумента $x=(\tilde{x}_1,\dots,\tilde{x}_N)^{\mathrm{T}}\in R^N$, в котором оценивается значение плотности $q(x)$ из (3). Выше через $\mathrm{I}_A(x)=\{1,\text{если } x\in A;\ \ 0,\text{если } x\notin A\}$ обозначен индикатор множества A.

Алгоритм кластер-анализа, основанный на оценке (5), состоит из следующих шагов.

1) По X объема n строим гистограмму (5).

2) Среди всех непустых ячеек ($\nu_{l_1,\dots,l_N}\neq 0$) помечаем те $\Gamma_{l_1,\dots,l_N}$, все соседние к которым ниже:

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

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

3) Для каждой i-й помеченной ячейки $\hat{\Gamma}_i$ $\left(i=\overline{1,\hat{L}}\right)$ 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.

4) Строим оценки $\{\hat{\mathcal{O}}_i\}_{i\in\hat{S}}$ для локальных мод $\{\mu_i\}_{i\in S}$ как центры ячеек $\{\hat{\Gamma}_i\}_{i\in\hat{S}}$:

$$\hat{\mathcal{O}}_i=\left((l_1+\tfrac{1}{2})\cdot h_1,\dots,(l_N+\tfrac{1}{2})\cdot h_N\right)^{\mathrm{T}}\in\hat{\Gamma}_i;\quad \hat{\Gamma}_i:=\Gamma_{l_1,\dots,l_N},\ i\in\hat{S}.$$

5) Производим классификацию выборки X в $\hat{L}+1$ класс, т.е. строим $\hat{D}=\left(\hat{d}_1,\dots,\hat{d}_n\right)\in\hat{S}_0^n$, $\hat{S}_0=\{0\}\cup\hat{S}$:

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

При этом наблюдения x_t, не попавшие в основные кластеры $\{\hat{V}_i\}_{i\in\hat{S}}$, относятся к "фоновому" кластеру $\hat{V}_0$ $\left(\hat{d}_t=0\right)$ и при необходимости могут быть классифицированы в основные кластеры, например, при помощи следующего правила: $\hat{d}_t=\underset{i\in\hat{S}}{\operatorname{argmin}}\ \rho\left(x_t,\hat{\mathcal{O}}_i\right)$, где $\rho(x,y)\geq 0$ – какая-либо метрика в R^N $(x,y\in R^N)$, в частности, евклидова: $\rho(x,y)=|x-y|$.

Исследуем статистические свойства получаемых в результате применения данного алгоритма статистических оценок. На основе результатов из [3] доказана теорема.

Теорема. Если условные плотности $\{q_i(\cdot)\}_{i\in S}$ из (2) непрерывны и ограничены на всём R^N, и имеет место асимптотика:

$$h_j=h_j(n)\to 0,\ j=\overline{1,N},\quad \prod_{j=1}^{N}h_j\cdot n\to+\infty,\quad n\to+\infty, \tag{6}$$

то гистограмма $\hat{q}_n(\cdot)$ из (5) является равномерно состоятельной по вероятности оценкой для безусловной плотности-смеси $q(\cdot)$ из (3):

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

Если вдобавок ко всему локальные моды $\{\mu_i\}_{i\in S}$ плотности (3) изолированные (для них выполняется (4)), то при $n\to+\infty$ число классов оценивается верно: $\hat{L}=L$, а оценки локальных мод состоятельны по вероятности:

$$\hat{\mathcal{O}}_i\xrightarrow{P}\mathcal{O}_i,\ i\in S.$$

Асимптотика (6) при построении гистограммы (5) является практической рекомендацией по выбору коэффициентов "размытости" h_l, $l=\overline{1,N}$: их значения должны уменьшаться с ростом объема выборки n, но "не слишком быстро". При этом обеспечивается состоятельность по вероятности оценок локальных мод и точное определение числа классов. А полученные оценки мод могут быть использованы в качестве "эталонов", описывающих "типичные" значения наблюдений в классах (так называемые "центры" классов [1,2]).

Проведен вычислительный эксперимент, в котором результаты применения предложенного выше алгоритма сравнивались с известным алгоритмом L-средних [1,2], требующем точного задания числа классов. Например, для ставших уже классическими данных Фишера по ирисам (L=3, N=4, n=150) [1,2] предложенный алгоритм правильно определил число классов ($\hat{L}=3$, три типа ирисов), а те наблюдения, которые ошибочно классифицируются алгоритмом L-средних (таких наблюдений 16 из 150 при использовании метрики Евклида), также ошибочно классифицировал.

ЛИТЕРАТУРА

[1] Фукунага, К. Введение в статистическую теорию распознавания образов / К. Фукунага. – Москва: Наука, 1979. – 368 с.

[2] Жук, Е.Е. Устойчивость в кластер-анализе многомерных наблюдений / Е.Е. Жук, Ю.С. Харин. – Минск: Белгосуниверситет, 1998. – 240 с.

[3] Жук, Е.Е. Кластер-анализ многомерных случайных наблюдений по гистограммной оценке плотности распределения вероятностей / Е.Е. Жук, Е.В. Храмова // Вестн. Белорус. ун-та. Сер. 1: Физ. Мат. Информ. – 2001. – № 2. – С. 80–86.