Статистический кластер-алгоритм

Шурыгин А. М.
a.shurygin@bk.ru
Москва, МГУ им. М. В. Ломоносова, факультет ВМиК

Большинство алгоритмов кластер-анализа требуют указания числа k классов, на которое надо поделить совокупность точек наблюдения $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$. Но в приложениях величину k можно указать лишь тогда, когда кластеризация уже проведена на интуитивном уровне. Во всех алгоритмах остаётся неопределённым само понятие кластера, а обычно к кластер-анализу обращаются, когда надо разделить сгустки точек «объективно», например, в биологии разделить виды, не путая их с подвидами.

Задача явно статистическая. Против предположения о нормальности распределений в кластерах никто не возражает. К. Пирсон [1] вывел критерий проверки гипотезы о совпадении центров двух многомерных нормальных совокупностей с известными ковариациями, различающихся сдвигом. Условия очень жёсткие и для кластер-анализа непригодные.

Предположив, что кластеры являются выборками из нормальных распределений, построить «объективный» кластер-алгоритм можно следующим образом. Кластер-критерием [2] на уровне значимости α проверить гипотезу о принадлежности выборки одному нормальному распределению. Если гипотеза принимается, то выборка с вероятностью $1-\alpha$ содержит один кластер. Критерий использует свойство оценок Мешалкина [3] оценивать параметры распределения кластера, наибольшего по количеству точек. Если гипотеза отвергается, то самый большой кластер вырезается эллипсоидом на уровне значимости α и к оставшимся точкам применяется описанная процедура в цикле. Процедура выделения кластеров заканчивается, когда все точки распределены по кластерам, либо их количество невелико, так что они могут считаться засоряющими или $\alpha-$остатками от проверки гипотез.

Рассмотрим элементы этого решения.

Оценки Мешалкина [3] $\mathbf{m}_\lambda = \left(m_\lambda^{(1)}, \ldots, m_\lambda^{(p)}\right)^{\mathrm{T}}$ и $\mathbf{C}_\lambda = \{c_\lambda^{ij}\}$ параметров нормального распределения $\mathcal{N}_p(\mathbf{m}, \mathbf{C})$ удовлетворяют системе уравнений, которую удобно решать итерациями:

$$\begin{cases} \mathbf{m}_\lambda = \dfrac{\Sigma_i\, \mathbf{x}_i \exp(-\lambda q_i/2)}{\Sigma_i\, \exp(-\lambda q_i/2)}; \\ \mathbf{C}_\lambda = (1+\lambda)\dfrac{\Sigma_i\, (\mathbf{x}_i - \mathbf{m}_\lambda)(\mathbf{x}_i - \mathbf{m}_\lambda)^{\mathrm{T}} \exp(-\lambda q_i/2)}{\Sigma_i\, \exp(-\lambda q_i/2)}; \end{cases}$$

где $q_i^2 = (\mathbf{x}_i - \mathbf{m}_\lambda)^{\mathrm{T}} C_\lambda^{-1} (\mathbf{x}_i - \mathbf{m}_\lambda)$ — квадрат расстояния от точки $\mathbf{x}_i$ до оценки $\mathbf{m}_\lambda$ центра распределения $\mathbf{m}$, измеренный оценкой $\mathbf{C}_\lambda$ матрицы ковариаций $\mathbf{C}$.

Кластер-критерий. Предлагается по величине

$$K_\lambda = \frac{1}{np} \sum\nolimits_i (\mathbf{x}_i - \mathbf{m}_\lambda)^{\mathrm{T}} \mathbf{C}_\lambda^{-1} (\mathbf{x}_i - \mathbf{m}_\lambda)$$

проверять «однородность» выборки при альтернативе наличия излишнего количества точек на периферии. Доказана асимптотическая сходимость к нормальному распределению величины

$$\sqrt{np} \ln K_\lambda \xrightarrow{d} \mathcal{N}(0,\ \xi^2);$$
$$\text{где } \xi^2 = \frac{\left[2 + 4\lambda + (p+2)\lambda^2\right](1+\lambda)^{p+2}}{(1+2\lambda)^{p/2+2}} - 2.$$

Если эта величина попадает в доверительный интервал на уровне значимости $1\% \leqslant \alpha \leqslant 5\%$, то выборка считается однородной и процесс заканчивается.

Если гипотеза отвергается, производится следующая процедура: **выделение точек кластера**. Они локализуются внутри p-мерного эллипсоида, $p \geqslant 2$, с центром в точке $\mathbf{m}$, с плотностью, постоянной на поверхности эллипсоида. Для проверки гипотезы на уровне значимости α эллипсоид должен содержать $1 - \alpha$ часть распределения. Пусть соответствующая часть эллипсоида удовлетворяет по вероятности равенству

$$\mathsf{P}\left\{\sqrt{(\mathbf{x} - \mathbf{m}_\lambda)^{\mathrm{T}} \mathbf{C}_\lambda^{-1} (\mathbf{x} - \mathbf{m}_\lambda)} \leqslant a_p\right\} = 1 - \alpha.$$

Положим $\mathbf{m}_\lambda = 0$, и соответственно центрируем плотность распределения. Сожмём эллипсоид вместе с плотностью распределения по главным осям так, чтобы распределение полученного вектора $\mathbf{y}$ стало стандартным нормальным $\mathcal{N}_p(\mathbf{0}, \mathbf{I})$. Найдём радиус b_p сферы $R_p(b)$, содержащей $1 - \alpha$ часть распределения. Решим уравнение

$$\mathsf{P}\left\{\sqrt{\mathbf{y}^{\mathrm{T}}\mathbf{y}} \in R_p(b_p)\right\} = 1 - \alpha$$

относительно b_p. Тогда получим равенство

$$1 - \alpha = \frac{1}{2^{p/2-1}} \int_0^{b_p} r^{p-1} \mathbf{e}^{-r^2/2}\, dr \Big/ \Gamma(p/2),$$

которое нетрудно решить численно, увеличивая b_p от $b_p = 2$. Пренебрегая небольшими отличиями истинных параметров распределения от их оценок, можно написать приближённое равенство

$$a_p \approx b_p,$$

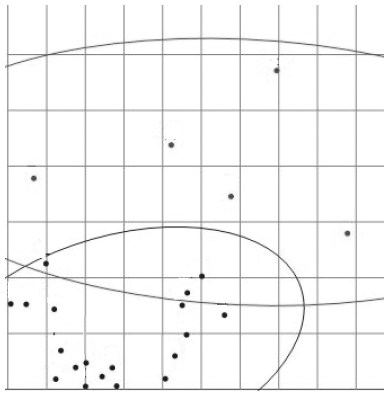

Рис. 1. Задача с пересечением двух кластеров.

решающее поставленную задачу.

О. Медведева составила программу для двумерного случая и решила ряд задач. Наиболее интересными из них были задачи со сложными наложениями и пересечениями нескольких кластеров.

Работа выполнена при поддержке РФФИ, проект № 04-01-00064.

Литература

[1] *Pearson K.* On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. // Phil. Mag. — 1900. — Vol. 50. — Pp. 157–175.

[2] *Шурыгин А. М.* Статистический кластер-критерий. // Алгоритмическое и программное обеспечение прикладного статистического анализа. — М.: Наука, 1980. — С. 360–366.

[3] *Meshalkin L. D.* Some mathematical methods for the study of non-communicable diseases. // Proc. 6-th Intern. Meeting of Uses of Epidemiol. in Planning Health Services. — Yugoslavia, Primosten. — Vol. 1. — Pp. 250–256.