# Кластеризация элементов множества на основе взаимных расстояний и близостей

***Двоенко С. Д.***
dsd@uic.tula.ru
Тула, Тульский государственный университет

Принцип несмещеной кластеризации лежит в основе известных алгоритмов кластер-анализа. Рассмотрены их модификации, когда доступна только матрица расстояний или близостей между объектами. Показана связь с алгоритмами экстремальной группировки признаков.

## Кластеризация объектов

В кластер-анализе предполагается, что объекты $\omega_i \in \Omega$, $i = 1, \ldots, N$, расположены в $n$-мерном пространстве (обычно евклидовом), где каждый представлен вектором $\mathbf{x}_i = (x_{i1}, \ldots x_{in})$, а все — матрицей данных $X(N, n)$. Каждый признак представлен своими наблюдениями $X_j = (x_{1j}, \ldots x_{Nj})^{\mathrm{T}}$, $j = 1, \ldots n$. Объекты образуют $K$ локальных сгущений (кластеры, классы, таксоны), которые следует выделить. Алгоритмы кластер-анализа (например, $K$-средних [1], семейство FOREL [2]) предполагают наличие признаков и строят «несмещенную» [3] кластеризацию: в ней все «представители» кластеров совпадают с их «центрами» $\widetilde{\mathbf{x}}_k = \overline{\mathbf{x}}_k$. Иначе центры назначаются представителями, и кластеры переопределяются. Центр кластера $\overline{\mathbf{x}}_k$ может не совпадать с его элементами $\mathbf{x}_i \in \Omega_k$.

В матрице расстояний $D(N, N)$ объект $\omega(\overline{\mathbf{x}}_k)$ не представлен. Обычно «центром» кластера выбирают объект $\overline{\omega}_k$, наименее удаленный от объектов кластера $\Omega_k$. При $\widetilde{\omega}_k = \overline{\omega}_k$ кластеризация может оказаться смещенной в пространстве, т. к. «центр» $\mathbf{x}(\overline{\omega}_k)$ не совпадет со средним $\overline{\mathbf{x}}_k$.

Относительно любого объекта $\omega_k \in \Omega$, взятого как начало координат, и любой пары объектов $\omega_i, \omega_j$; $i, j = 1, \ldots, N$, по известной теореме косинусов определяется их скалярное произведение $c_{ij} = (d_{ki}^2 + d_{kj}^2 - d_{ij}^2)/2$, где $c_{ii} = d_{ki}^2$ и $d_{pq} = d(\omega_p, \omega_q)$ — расстояние. В матрицах $C_k(N, N)$, $k = 1, \ldots, N$, их главные диагонали представляют квадраты расстояний от соответствующего начала координат $\omega_k$ до объектов $\omega_i$, $i = 1, \ldots, N$. Янгом и Хаусхолдером в задаче метрического шкалирования [4] предложено восстанавливать евклидово пространство как разложение $C_k = XX^{\mathrm{T}}$, где матрица $C_k(N-1, N-1)$ с рангом $n < N$ положительно полуопределена, $X(N-1, n)$ — матрица проекций объектов $\omega_1, \ldots \omega_{k-1}, \omega_{k+1}, \ldots \omega_N$ на $n$ ортогональных осей с началом координат $\omega_k$. В методе главных проекций Торгенсона [5] восстанавливается евклидово пространство с началом координат в центре тяжести множества объектов $\omega_i \in \Omega$, $i = 1, \ldots, N$.

Центры $\overline{\omega}_k$ кластеров $\Omega_k$, $k = 1, \ldots, K$ немедленно определяются по методу Торгенсона. Для начала координат в центре тяжести кластера $\Omega_k$ получим матрицу скалярных произведений $\overline{C}_k(N, N)$. Диагональные элементы $\overline{c}^k_{ii} = d^2(\omega_i, \overline{\omega}_k)$, $i = 1, \ldots, N$ представляют центр $\overline{\omega}_k$ кластера $\Omega_k$ квадратами расстояний до остальных объектов $\omega_i \in \Omega$, $i = 1, \ldots, N$:

$$d^2(\omega_i, \overline{\omega}_k) = \frac{1}{N_k} \sum_{p=1}^{N_k} d^2_{ip} - \frac{1}{2N_k^2} \sum_{p=1}^{N_k} \sum_{q=1}^{N_k} d^2_{pq}; \quad \omega_p, \omega_q \in \Omega_k,$$

где $N_k$ — число объектов в $\Omega_k$. Отсюда сразу определяются алгоритмы кластеризации для расстояний, например $K$-средних и FOREL.

Положительно полуопределенную матрицу $S(N, N)$ попарных близостей $s_{ij} = s(\omega_i, \omega_j) \geqslant 0$ объектов $\omega_i, \omega_j \in \Omega$ можно считать матрицей скалярных произведений векторов $\mathbf{x}_i = \mathbf{x}(\omega_i)$, $i = 1, \ldots, N$, в пространстве размерности не выше $N$. Скалярные произведения объектов $\omega_i, \omega_j$ относительно $\omega_k \in \Omega$ представлены как $s_{ij} = (d^2_{ki} + d^2_{kj} - d^2_{ij})/2$. Поскольку $s_{ii} = d^2_{ki}$, то $d^2_{ij} = s_{ii} + s_{jj} - 2s_{ij}$, и можно получить кластеризацию по расстояниям.

Пусть объекты разбиты на $K$ кластеров $\Omega_k$. Представим центр кластера $\overline{\omega}_k$ своими близостями к остальным объектам $\omega_i \in \Omega$:

$$s(\omega_i, \overline{\omega}_k) = \frac{1}{N_k} \sum_{p=1}^{N_k} s_{ip}; \quad \omega_p \in \Omega_k,$$

где $N_k$ — число объектов в $\Omega_k$. Отсюда сразу определяются алгоритмы кластеризации для близостей, например $K$-средних и FOREL.

Несмещенная кластеризация минимизирует дисперсию кластера и максимизирует среднюю близость объектов в кластере:

$$\sigma^2_k = \frac{1}{2N_k^2} \sum_{i=1}^{N_k} \sum_{j=1}^{N_k} d^2_{ij} = \frac{1}{N_k} \sum_{i=1}^{N_k} s_{ii} - \frac{1}{2N_k^2} \sum_{i=1}^{N_k} \sum_{j=1}^{N_k} s_{ij}; \quad \omega_i, \omega_j \in \Omega_k.$$

После нормировки $s'_{ij} = s_{ij}/\sqrt{s_{ii}s_{jj}}$ получим $\sigma^2_k = 1 - \delta_k$.

**Кластеризация признаков**

Группировка $n$ признаков по их корреляциям $R(n, n)$ выполняется алгоритмом $K$-средних для близостей $S(n, n)$, где $s_{ij} = r^2_{ij}$ или $s_{ij} = |r_{ij}|$, максимизируя величины ($n_k$ — число признаков $\omega_i$ в группе $\Omega_k$):

$$I_1 = \sum_{k=1}^{K} n_k \delta'_k = \sum_{k=1}^{K} \sum_{i=1}^{n_k} r^2(\omega_i, \overline{\omega}_k) \text{ и } I_2 = \sum_{k=1}^{K} n_k \delta''_k = \sum_{k=1}^{K} \sum_{i=1}^{n_k} |r(\omega_i, \overline{\omega}_k)|.$$

В алгоритмах экстремальной группировки «квадрат» и «модуль» [6] функционалы $J_1 = \sum\limits_{k=1}^{K}\sum\limits_{i=1}^{n_k} r^2(\omega_i, \pi_k)$ и $J_2 = \sum\limits_{k=1}^{K}\sum\limits_{i=1}^{n_k} |r(\omega_i, \mu_k)|$, $\omega_i \in \Omega_k$, где $\pi_k$ — главный «фактор», $\mu_k$ — центроидный «фактор» группы $\Omega_k$, характеризуют качество разбиения признаков на $K$ групп, в каждой из которой признаки наиболее сильно коррелируют со своим фактором. Факторы строятся как одновременное решение основных факторных задач: построение $K$ общих факторов и их косоугольное вращение [7].

Представим центр $\overline{\omega}_k$, главный $\pi_k$ и центроидный $\mu_k$ факторы группы $\Omega_k$ своими близостями к признакам $\omega_i \in \Omega_k$:

$$\begin{cases} s(\omega_i, \overline{\omega}_k) = (1/n_k)\sum\limits_{j=1}^{n_k} s_{ij}; \\ s(\omega_i, \pi_k) = \sum\limits_{j=1}^{n_k} \alpha_j^k s_{ij} = \lambda_k \alpha_j^k, \quad \boldsymbol{\alpha}_k = (\alpha_1^k, \ldots \alpha_{n_k}^k); \\ s(\omega_i, \mu_k) = \sum\limits_{j=1}^{n_k} s_{ij}; \end{cases}$$

где $\lambda_k$ — максимальное собственное значение, $\boldsymbol{\alpha}_k$ — соответствующий ему собственный вектор подматрицы близостей $S(n_k, n_k)$ признаков $\omega_i \in \Omega_k$.

Легко увидеть, что близости $s(\omega_i, \overline{\omega}_k)$ и $s(\omega_i, \mu_k)$ совпадают с точностью до множителя. Поэтому группировки по «модулю» — несмещенные. Группировки по «квадрату» — смещенные. Также очевидно, что $J_1 \geqslant I_1$. В итоге, оба алгоритма $K$-средних для расстояний и для близостей, примененные для кластеризации признаков, аналогичны алгоритму «модуль».

В основе алгоритма FOREL лежит процедура несмещенной кластеризации, которую можно представить как алгоритм «1-среднего». Следовательно, техника кластер-анализа, развитая для семейства FOREL, адекватно применима для группировки признаков.

Работа выполнена при поддержка РФФИ, проект № 05-01-00679 и INTAS, проект № 04-77-7347.

### Литература

[1] *Tou J. T., Gonzalez R. C.* Pattern recognition principles. — London: Addison-Wesley, 1981.

[2] *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. — Новосибирск: Изд. Ин-та матем., 1999.

[3] *Шлезингер М.И.* О самопроизвольном различении образов // Читающие автоматы, Киев: Наукова думка, 1965. — С. 38–45.

[4] *Young G., Householder A. S.* Discussion of a set of points in terms of their mutual distances // Psychometrika. — 1938. — V. 3. — P. 19–22.

[5] *Torgenson W. S.* Theory and methods of scaling. — N.Y.: J. Wiley, 1958.

[6] *Браверман Э. М.* Методы экстремальной группировки параметров и задача выделения существенных факторов // Автоматика и телемеханика. — 1970. — № 1. — С. 123–132.

[7] *Harman H. H.* Modern factor analysis. — Chicago: Univ. Chicago Press, 1976.