# Обработка многомерных данных несколькими методами кластерного анализа

Борчик Е.М.; Башаримов В.В.
Кафедра АСУ, электротехнический факультет
ГУВПО «Белорусско-Российский университет»
г. Могилев, Беларусь
e-mail: ykm@tut.by

***Аннотация*** **– Предложена процедура проверки и уточнения результатов разделения многомерных наблюдений на кластеры с использованием нескольких методов кластерного анализа. Показано, что в случае, если элементы множества $X$ представляют собой наблюдения *n* параметров множества объектов, то результат кластеризации $X$ может быть интерпретирован как матрица вероятностей принадлежности объектов определенным кластерам. Предложен критерий принадлежности объекта определенному кластеру, получена формула вычисления значений элементов обобщенной матрицы через элементы матриц вероятностей принадлежности объектов определенным кластерам.**

***Ключевые слова: кластерный анализ, K-Means, Tree Clustering, Fuzzy Relation Clustering, BelSim, STATISTICA***

## I. Введение

В имитационном моделировании возникает необходимость анализа многомерных данных, полученных при проведении имитационных экспериментов, в частности – задачи разделения множеств данных $X \subset R^n$ на непересекающиеся подмножества (кластеры). Для решения данной задачи используются методы кластерного анализа.

Пусть в ходе имитационных экспериментов получено множество наблюдений $X = \{x_i \mid x_i \in R^n, i = 1, \dots, m\}$, которое необходимо разбить на кластеры.

Для исследования выбраны методы кластеризации, являющиеся представителями основных методологических подходов к разделению исходного множества объектов на кластеры: K-Means, Tree Clustering, Fuzzy Relation Clustering (FRC) [1].

Для разбиения множества $X$ на кластеры предполагается использование нескольких методов кластеризации для проверки и уточнения результатов. При этом требуется решить задачу обобщения полученных результатов кластеризации множества наблюдений $X$ несколькими методами.

## II. Кластерный анализ данных

Кластерный анализ проводится для данных:

$$X = \{x_i \mid x_i \in R^n, i = 1, \dots, m\}. \qquad (1)$$

Пусть элементы $x_i \in X$ представляют собой измерения $n$ параметров объектов $b_r$ из множества

$$B = \{b_r \mid r = 1, \dots, |B|\}, |B| < m. \qquad (2)$$

В результате разбиения множества $X$ на кластеры каждый из $L$ применяемых методов кластеризации $M_l, l = 1, \dots, L$, ставит в соответствие номерам элементов $x_i \in X$ соответствующие им номера $j$, $j \in 1, \dots, k_l$, кластеров $K_j$, где $k_l$ – количество кластеров построенных методом $M_l$.

Для обобщения результатов кластеризации данных несколькими методами введём необходимые определения и утверждения.

Утверждение 1. Результат кластеризации множества $X$ вида (1) для каждого из методов кластеризации $M_l, l = 1, \dots, L$, может быть представлен в виде матрицы вероятностей принадлежности объектов $b_r \in B$ определенным кластерам:

$$P_l = \left\| p_{lrj} \right\|, \; r = 1, \dots, |B|, \; j \in 1, \dots, k_l. \qquad (3)$$

где $r$ – номер объекта из множества $B$; $j$ – номер кластера $K_j$; $k_l$ – количество кластеров, построенное методом $M_l$; $p_{lrj} \in [0,1]$ – вероятность принадлежности $r$-го объекта $b_r \in B$ кластеру $K_j$; $l$ – номер примененного метода кластеризации $M_l$, $l \in 1, \dots, L$.

Вероятности $p_{lrj}$ в (3) рассчитываются на основе классического определения вероятности как отношение количества случаев попадания объекта $b_r$ в кластер $K_j$ к общему количеству измерений, выполненных над объектом $b_r$.

Определение 1. Объект $b_r \in B, r = 1, \dots, |B|$, является элементом кластера $K_j$, $j = 1, \dots, k$, тогда и только тогда, когда он отнесен к данному кластеру, по крайней мере, $L^*$ методами из $L$ выбранных методов кластерного анализа, причем $L/2 < L^* \le L$, $L \ge 3$.

Утверждение 2. Пусть $P_1, P_2, \dots, P_L$ – матрицы вида (3) вероятностей принадлежности объектов $b_r, r = 1, \dots, |B|$, определенным кластерам $K_j, j = 1, \dots, k$, согласно методам кластеризации $M_1, M_2, \dots, M_L$, соответственно. Тогда значения элементов $p_{rj} \in [0,1]$ обобщенной (в смысле Определения 1) матрицы $P$ могут быть найдены суммированием коэффициентов $P_{v,L}, v = L^*, \dots, L$, производящей функции

$$\varphi(z) = \prod_{l=1}^{L} (q_{lrj} + p_{lrj} \cdot z) = \sum_{v=0}^{L} P_{v,L} \cdot z^v,$$

где $p_{lrj}$ – элементы матрицы $P_l, l = 1, \dots, L$, $p_{lrj} \in [0,1]$; $q_{lrj} = 1 - p_{lrj}$, $l = 1, \dots, L$, $r = 1, \dots, |B|$, $j = 1, \dots, k$, соответственно,

$$p_{rj} = \sum_{v=L^*}^{L} P_{v,L}. \qquad (4)$$

Замечание 1. В случае разбиения множества $X$ на $k_1$ кластер методом $M_1$, … , $k_L$ кластеров методом $M_L$, для возможности обобщения результатов кластеризации необходимо приведение матриц $P_l$, $l=1,\dots,L$, к одной размерности $k=\max\{k_1,k_2,\dots,k_L\}$.

Замечание 2. Для возможности обобщения результатов кластеризации необходимо проведение упорядочивания столбцов обобщаемых матриц $P_l$ вида (3) размерности $|B|\times k$, таким образом, что для любого $j_0 \in 1,\dots,k$ кластеры $K_{j_0}^{(l)}$, построенные методами $M_l$, $l=1,\dots,L$, будут соответствовать друг другу, то есть иметь максимально возможное количество общих элементов $x_i \in X$.

## III. Процедура определения кластеров объектов

Утверждение 3. Процедура кластеризации множества многомерных данных $X$ методами $M_l$, $l=1,\dots,L$, интерпретации результатов кластеризации и их подготовки к последующему этапу обобщения выполнима за $L^*$ итераций, $L/2 < L^* \le L$, состоящих из шагов 0-5:

Шаг 0. Вводится в рассмотрение параметр номера итерации $w$ с начальным значением $w=0$.

Шаг 1. Проводится кластеризация элементов множества $X$ методом $M_{w+1}$.

Шаг 2. По результатам кластерного анализа (Шаг 1) проводится построение в соответствии с Утверждением 1 матрицы $P_{w+1}$ вида (3) размерности $|B|\times k_{w+1}$ вероятностей принадлежности объектов $b_r \in B$, $r=1,\dots,|B|$, кластерам $K_j$, $j=1,\dots,k_{w+1}$.

Шаг 3. При $w>0$, проводится приведение матриц $P_l^{(w-1)}$, $l=1,\dots,w$ размерности $|B|\times k_{w+1}$ и матрицы $P_{w+1}$ к одной размерности $|B|\times k^{(w)}$, где

$$k^{(w)} = \max\left\{k^{(w-1)}, k_{w+1}\right\}. \qquad (5)$$

Построение матриц $P_{w+1}^{(w)} = \left\| p_{w+1rj}^{(w)} \right\|$, $P_l^{(w)} = \left\| p_{lrj}^{(w)} \right\|$, $l=1,\dots,w$, размерности $|B|\times k^{(w)}$.

Шаг 4. В случае, если $P_1^{(w)} \not\equiv P_{w+1}^{(w)}$, проводится упорядочивание столбцов матрицы $P_{w+1}^{(w)}$ по отношению к $P_1^{(w)}$.

Шаг 5. Определяется количество $S^{(w)}$ эквивалентных матриц $P_l^{(w)}$, $l=1,\dots,w+1$.

Если выполняется условие

$$(w < L-1) \wedge (S^{(w)} \le L/2), \qquad (6)$$

то $w := w+1$, шаги 1-5 повторяются.

Замечание 1 Приведение матриц $P_l$, $l=1,\dots,w$, размерности $|B|\times k^{(w-1)}$ и $P_{w+1}$ к одной размерности $k^{(w)}$ вида (5) на Шаге 3 процедуры, равносильное построению матриц $P_{w+1}^{(w)} = \left\| p_{w+1rj}^{(w)} \right\|$, $P_l^{(w)} = \left\| p_{lrj}^{(w)} \right\|$ размерности $|B|\times k^{(w)}$, проводится по формулам

$$p_{w+1rj}^{(w)} = \begin{cases} p_{w+1rj}, j \le k_{w+1}, \\ 0, k_{w+1} < j \le k^{(w)}; \end{cases} \quad p_{lrj}^{(w)} = \begin{cases} p_{lrj}^{(w-1)}, j \le k^{(w-1)}, \\ 0, k^{(w-1)} < j \le k^{(w)}. \end{cases}$$

Утверждение 4. Перестановка $I^*$ столбцов матрицы $P_{w+1}^{(w)}$ по отношению к матрице $P_1^{(w)}$ оптимальна тогда и только тогда, когда выполняется условие:

$$\rho(P_{w+1}^{(w)}(I^*), P_1^{(w)}) = \max_t\left\{\rho(P_{w+1}^{(w)}(I_t), P_1^{(w)})\right\},$$

где $t=0,\dots,k^{(w)}!-1$ – номер произведенной перестановки $I_t$, определяющей порядок следования столбцов $P_{w+1}^{(w)}$, общее количество возможных перестановок $k^{(w)}$ столбцов матрицы $P_{w+1}^{(w)}$ составляет $k^{(w)}!$; $\rho$ – метрика вида:

$$\rho(P_{w+1}^{(w)}(I_t), P_1^{(w)}) = \sum_{i=1}^{|B|}\sum_{j=1}^{k} min(p_{w+1rj}^{(w)}(I_t), p_{1rj}^{(w)}),$$

$$t=0,\dots,k^{(w)}!-1, \qquad (7).$$

Значение $\rho \ge 0$ прямо пропорционально количеству общих элементов $x_i \in X$ в соответствующих кластерах матричной пары.

Обобщение результатов кластеризации $P_1,\dots,P_L$ проводится в соответствии с Утверждением 2.

Для возможности определения кластеров объектов вводится следующее определение.

Определение 3. Объект $b_r \in B, r \in 1,\dots,|B|$ принадлежит кластеру $K_{j_0}$, $j_0 \in 1,\dots,k$ тогда и только тогда, когда вероятность принадлежности объекта кластеру в $r$-й строке обобщенной матрицы $P$ максимальна: $p_{rj_0} = \max\{p_{rj} \mid j=1,\dots,k\}$.

Предложенная процедура обобщения результатов кластеризации многомерных данных несколькими методами применяется в программно-технологическом комплексе имитации сложных систем BelSim [2].


[1] Методы и модели анализа данных: OLAP и Data Mining / А. А Барсегян и [и др.] – СПб.: БХВ – Петербург, 2004. – 336 с.: ил.

[2] Якимов, А. И. Технология имитационного моделирования систем управления промышленных предприятий : монография / А. И. Якимов. – Могилев: Белорус.-Рос. ун-т, 2010. – 304 с.: ил.