УДК 004.8

ОБОБЩЕНИЕ РЕЗУЛЬТАТОВ КЛАСТЕРИЗАЦИИ ДАННЫХ В ХОДЕ ИМИТАЦИОННЫХ ЭКСПЕРИМЕНТОВ

Е. М. БОРЧИК, В. В. БАШАРИМОВ, А. И. ЯКИМОВ
Государственное учреждение высшего профессионального образования
«БЕЛОРУССКО-РОССИЙСКИЙ УНИВЕРСИТЕТ»
Могилев, Беларусь

Пусть при имитационных экспериментах задано множество наблюдений $X = \{x_1, .., x_m\}$, $x_i \in R^n$, которое необходимо разбить на непересекающиеся подмножества (кластеры). Для решения данной задачи используются методы кластерного анализа, например, K-Means, Tree и Fuzzy Relation Clustering (FRC). Метод K-Means строит *заданное* количество кластеров, но требует охвата каждого кластера отдельным *выпуклым множеством*. Методы Tree и FRC не имеют этого ограничения, но не гарантируют построения *заданного* количества кластеров. Следует отметить, что метод FRC наиболее точен, но характеризуется трудоемкостью $O(n^4)$ от числа элементов.

Гарантированное разбиение множества X на кластеры предполагает использование нескольких методов кластеризации для проверки и уточнения результатов. Вначале разбиение производится методами K-Means и Tree. Если результаты разбиений не совпадают, то применяется метод FRC. В результате разбиения множества X на кластеры каждый из методов ставит в соответствие номерам $i = 1, ..., m$ элементов $x_i \in X$ соответствующие им номера кластеров K_j, $j \in 1, ..., k$. При этом требуется обобщение полученных результатов кластеризации X несколькими методами.

***Утверждение* 1.** Если элементы $x_i \in X$, $i = 1, ..., m$ представляют собой наблюдения n параметров N объектов $Ob = \{ob_t\}$, $t = 1, ..., N$, $N < m$, то результат кластеризации множества X можно представить в виде матрицы вероятностей принадлежности объектов $ob_t \in Ob$ определенным кластерам

$$P = \left\| p_{tj} \right\|, t = 1, ..., N, j = 1, ..., k, \quad (1)$$

где t – номер объекта, j – номер кластера, k – количество кластеров, $p_{tj} \in [0, 1]$ – вероятность принадлежности t-го объекта кластеру K_j.

Вероятности p_{tj} в (1) рассчитываются на основе определения вероятности, как отношения количества случаев попадания объекта ob_t в кластер K_j к общему количеству наблюдений, выполненных над объектом ob_t.

***Определение* 1.** Объект $ob_t \in Ob$ является элементом кластера K_j, $j \in 1, ..., k$ тогда и только тогда, когда он отнесен к данному кластеру, по крайней мере, двумя из трех выбранных методов кластеризации.

***Утверждение* 2.** Пусть P_1, P_2, P_3 – матрицы вида (1) вероятностей принадлежности объектов $ob_t, t = 1, ..., N$ определенным кластерам в соответствии с методами кластерного анализа K-Means , Tree, FRC соответственно. Тогда элементы $p_{tj} \in [0,1]$ обобщенной (в смысле Определения 1) матрицы P могут быть получены посредством применения теорем сложения/умножения вероятностей к элементам $p_{ltj}, l = 1, 2, 3$ матриц P_1, P_2, P_3:

$$p_{tj} = (1 - p_{1tj})\, p_{2tj}\, p_{3tj} + p_{1tj}\,(1 - p_{2tj})\, p_{3tj} + p_{1tj}\, p_{2tj}\,(1 - p_{3tj}) + p_{1tj}\, p_{2tj}\, p_{3tj}.$$

Обобщаемые матрицы P_1, P_2, P_3 должны иметь одну размерность. В случае разбиения множества X на разное количество кластеров, необходимо предварительно привести матрицы P_1, P_2, P_3 к одной размерности $k = \max\{k_1, k_2, k_3\}$. Приведение матрицы к необходимой размерности возможно за счет ее дополнения столбцами с нулевыми вероятностями попадания объекта в добавленные кластеры.

Проведено сравнение методов кластерного анализа на множестве X двумерных наблюдений, причем X составлено из нескольких подмножеств $X = A_1 \cup A_2 \cup A_3 \cup A_4 \cup A_5 \cup A_6$ таким образом, что кластеры выделяются визуально. Области $A_1, ..., A_6$ заполнены одинаковым количеством равномерно распределенных на интервале $[0,1]$ значений, полученных в MS Excel с использованием функции СЛЧИС() .

Метод K-Means разделил множество X на 3 кластера таким образом, что в кластер K_1 попала большая часть элементов множеств A_1, A_4; в кластер K_2 – большая часть элементов множеств A_2, A_3; к кластеру K_3 отнесено 50 % элементов множества A_5.

Методом Tree множество X разделено на 3 кластера. Кластер K_1 полностью состоит из элементов множества A_1, кластер K_2 включает элементы $A_2 - A_5$, кластер K_3 полностью состоит из элементов множества A_6.

Поскольку разбиения исходного множества X методами K-Means и Tree различны, то для уточнения результатов применяется метод кластерного анализа FRC. Алгоритм FRC при значении параметра $\alpha = 0{,}85$ разделил множество X на 3 кластера.

Анализ обобщенного результата кластеризации позволяет выделить 3 кластера: $K_1 = A_1$, $K_2 = A_2 \cup, ..., \cup A_5$ и $K_3 = A_6$. Таким образом, ожидаемый результат кластеризации исходного множества со специально заданной структурой подтвердился.