УДК 004.8

ОБРАБОТКА МНОГОМЕРНЫХ ДАННЫХ НЕСКОЛЬКИМИ МЕТОДАМИ КЛАСТЕРНОГО АНАЛИЗА

В. И. Аверченков*, А. И. Якимов, Е. М. Борчик**, В. В. Башаримов****

**Брянский государственный технический университет*
***Белорусско-Российский университет*

Поступила в редакцию 17.03.2016 г.

Аннотация. Предложена процедура проверки и уточнения результатов разделения многомерных наблюдений на кластеры несколькими методами кластерного анализа. В результате разбиения множества X на кластеры каждый из методов ставит в соответствие номерам элементов множества X соответствующие им номера кластеров. Показано, что в случае, если элементы множества X представляют собой наблюдения n параметров множества объектов, то результат кластеризации множества 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. Экспериментальные исследования проведены с использованием методов кластерного анализа K-Means, Tree Clustering, Fuzzy Relation Clustering на 4-мерных данных об ирисах, предложенных Фишером в 1936 г.

Ключевые слова: кластерный анализ, K-Means, Tree Clustering, Fuzzy Relation Clustering, многомерные данные.

Annotation. Procedure to verify and clarify the results division of multidimensional observations into clusters is offered by several methods of cluster analysis. As a result of splitting set X into clusters each of the methods puts X numbers of clusters corresponding to them in compliance to numbers of set members X. It is shown that if set members X represent observations n parameters of set objects, the result of clustering of set X can be interpreted as a. The criterion of belonging object to particular cluster is offered, we obtain a formula for computing the values of elements of the generalized matrix through elements probability matrix of objects belonging to certain clusters is received. Experimental studies are conducted with use of cluster analysis methods K-Means, Tree Clustering, Fuzzy Relation Clustering on 4-dimensional data on the irises, offered by Fischer in 1936.

Keywords: cluster analysis, K-Means, Tree Clustering, Fuzzy Relation Clustering, multidimensional data.

ВВЕДЕНИЕ

Пусть получено множество наблюдений $X = \{x_i \mid x_i \in R^n,\ i = 1, \ldots, m\}$, которое необходимо *Эксперту* – специалисту предметной области, определяющему запросы к анализу данных – разбить на непересекающиеся подмножества (кластеры) [1].

Для решения данной задачи используются методы кластерного анализа. В результате анализа существующих методов кластеризации [2, 3] разработана их классификация [4, 5], для исследования выбраны методы,

являющиеся представителями основных методологических подходов к разделению исходного множества объектов на кластеры: K-Means, Tree Clustering, Fuzzy Relation Clustering (FRC).

Метод K-Means строит *заданное* количество кластеров, но требует охвата каждого кластера отдельным *выпуклым множеством*. Методы Tree Clustering и FRC не имеют этого ограничения, но не гарантируют построения *заданного* количества кластеров. Следует отметить, что метод FRC характеризуется трудоемкостью $O\left(n^4\right)$ от числа элементов.

Для разбиения множества X на кластеры предполагается использование нескольких методов кластеризации для проверки и уточнения результатов.

Пусть для проверки и уточнения результатов кластеризации данных в качестве дополнительных начальных условий *Эксперт* определяет: методы кластеризации данных, предполагаемое количество кластеров и контрольные точки (КТ) кластеров – элементы исследуемого множества, с вероятностью 100 % (по заключению *Эксперта*) принадлежащие определённым кластерам.

Для принятия решения о разделении множества наблюдений X на кластеры *Эксперт* формулирует запрос к точности принятия решений. Задаёт пороговые нижние границы относительно необходимого и достаточного количества методов кластеризации, в соответствии с которыми принимается экспертное решение (оценка): голосование большинством голосов, принятие решений максимально возможным количеством голосов, принятие решений, по крайней мере, заданным количеством голосов и т. д.

При этом вначале разделение производится двумя методами. Если результаты разбиений не совпадают, то применяют третий и так далее методы. При этом требуется решить задачу обобщения полученных в соответствии с запросами *Эксперта* результатов кластеризации несколькими методами множества наблюдений X.

ОПИСАНИЕ КЛАСТЕРНОГО АНАЛИЗА ДАННЫХ

Кластерный анализ проводится на многомерных данных:

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

Пусть элементы $x_i \in X$, $i = 1, \ldots, m$, представляют собой измерения n параметров объектов b_r из множества

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

Тогда каждому $x_i \in X$ соответствует некоторый объект b_r и идентифицирующая информация об этом объекте.

Пусть для разделения множества (1) на кластеры используются методы кластеризации M_l, $l = 1, \ldots, L$, $L \geq 3$ (например, K-Means, Tree Clustering, FRC и так далее).

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

ОБОБЩЕНИЕ РЕЗУЛЬТАТОВ КЛАСТЕРИЗАЦИИ ДАННЫХ НЕСКОЛЬКИМИ МЕТОДАМИ

Введём необходимые определения и утверждения, позволяющие интерпретировать результаты кластеризации множества X методами M_l, $l = 1, \ldots, L$, произвести их обобщение и последующее выделение кластеров объектов $b_r \in B$.

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

$$P_l = \left\| p_{lrj} \right\|, \quad p_{lrj} \in [0, 1],$$

$$r = 1, \ldots, |B|, \quad j \in 1, \ldots, k_l, \qquad (3)$$

где r – номер объекта из множества B; j – номер кластера K_j; k_l – количество кластеров, построенное методом M_l; p_{lrj} – вероятность принадлежности r-го объекта $b_r \in B$

кластеру K_j в соответствии с методом M_l; l – номер примененного метода кластеризации M_l, $l \in 1, \ldots, L$.

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.

Замечание. Сумма строчных элементов матриц P_l вида (3) постоянна и равна единице.

Вероятности отнесения объектов к определенным кластерам одновременно L выбранными методами очень низкие, напротив – вероятности отнесения объектов к определенным кластерам, по крайней мере, одним из выбранных методов очень высоки. Поэтому рассматривается случай отнесения объекта к данному кластеру L^* методами из L выбранных методов кластерного анализа, где $L_E < L^* \le L$, L_E – точная нижняя граница определяемая *Экспертом* в зависимости от уровня значимости задачи. Например, в случае аналогии голосования большинством голосов $L_E = L/2$.

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

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

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

Доказательство. Вывод обобщающей формулы (4) приведён в [6].

Замечание 1. Для каждого из методов кластеризации свойственна собственная система нумерации кластеров K_j. Номера $j = 1, \ldots, k_l$ присваиваются кластерам K_j в соответствии с порядком их построения, который зависит от особенностей алгоритмов методов кластеризации M_l, $l = 1, \ldots, L$, основанных на использовании матриц сходства, эвристических алгоритмов перебора, идей математического программирования, на оценивании функций плотности статистического распределения и др. [2, 3].

Замечание 2. Для возможности обобщения результатов кластеризации обобщаемые матрицы $P_1, P_2, \ldots, P_L$ должны иметь одну размерность. В случае разбиения множества X на k_1 кластеров методом M_1, k_2 кластеров методом $M_2, \ldots, k_L$ кластеров методом M_L, необходимо предварительно привести матрицы $P_1, P_2, \ldots, P_L$ к одной размерности $k = \max\{k_1, k_2, \ldots, k_L\}$. Приведение матрицы к необходимой размерности возможно за счет ее дополнения столбцами с нулевыми вероятностями попадания объекта в добавленные кластеры.

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

Определение. Контрольные точки (КТ) кластеров. Пусть *Экспертом* заданы следующие дополнительные начальные условия кластеризации $X \subset R^n$:

1) ψ – предполагаемое количество кластеров K_j, $j = 1, \ldots, \psi$;

2) $Y \subset X$, $|Y| = \gamma \ge \psi$ – множество контрольных точек кластеров:

$$Y = \{y_j \mid y_j \in R^n,\ j = 1, \ldots, \gamma\}. \qquad (5)$$

При этом для контрольных точек кластеров $y_i \in Y$ выполняется условие:

$$(\forall y_i \in Y)\left(\exists! K_j\right)[y_i \in K_j],$$

$$i \in 1, ..., \gamma, \quad j \in 1, ..., \psi. \qquad (6)$$

Предлагается следующая процедура определения кластеров объектов:

Этап 1. Кластеризация множества X методами M_l, $l = 1, ..., L$; интерпретация результатов кластеризации; подготовка к этапу обобщения.

Этап 2. Обобщение результатов кластеризации; определение кластеров объектов.

Этап 1 процедуры определения кластеров объектов

Определение 2. Матрицы $P_\xi = \left\| p_{\xi rj} \right\|$, $P_\eta = \left\| p_{\eta rj} \right\|$ вида (3) размерности $\left|B\right| \times k$ эквивалентны тогда и только тогда, когда выполняется условие:

$$(P_\xi = P_\eta) \Leftrightarrow$$
$$\Leftrightarrow (\forall r \in 1, ..., \left|B\right|)(\forall j \in 1, ..., k)[\, p_{\xi rj} = p_{\eta rj}\,]. \qquad (7)$$

Определение 3. Приведение матриц $P_l = \left\| p_{lrj} \right\|$, $l = 1, ..., w$, $1 < w \le L$ размерностей $\left|B\right| \times k_l$ (множество X разделено методом M_l на k_l кластеров) к одной размерности $\left|B\right| \times k$, где $k = \max\left\{ k_1, ..., k_w \right\}$, равносильно построению матриц $P_l^{/} = \left\| p_{lrj}^{/} \right\|$ за счёт дополнения матриц P_l столбцами с нулевыми вероятностями попадания объектов в добавленные кластеры:

$$p_{lrj}^{/} = \begin{cases} p_{lrj} \mid j \le k_l, \\ 0 \quad \mid k_l < j \le k. \end{cases} \qquad (8)$$

Введём необходимые определения и утверждения, позволяющие провести упорядочивание столбцов матрицы P_T (тестируемая матрица) по отношению к матрице P_E (эталонная, согласно оценке *Эксперта*, матрица).

Определение 4. Обозначим через P_E – (эталонную) матрицу вероятностей принадлежности контрольных точек (КТ) – начальное условие кластеризации, определённое *Экспертом* – кластерам; P_T – (тестируемую) матрицу вероятностей принадлежности элементов множества X, соответствующих КТ, построенную по результатам кластеризации X методом M_l:

$$P_E = \left\| p_{Eij} \right\|, \; p_{Eij} = \begin{cases} 1, \, y_i \in K_j, \\ 0, \, y_i \notin K_j; \end{cases} \qquad (9)$$
$$P_T = \left\| p_{Tij} \right\|,$$
$$p_{Tij} = \begin{cases} 1, \left(\exists x_k \in \mathrm{X}\right)[\left(x_k = y_i\right) \wedge \left(x_k \in K_j\right)], \\ 0, \left(\exists x_k \in \mathrm{X}\right)[\left(x_k = y_i\right) \wedge \left(x_k \notin K_j\right)]. \end{cases} \qquad (10)$$

Утверждение 3. Пусть в результате кластеризации X методом M_l построено K_j, $j = 1, ..., \psi^{/}$ кластеров. Пусть выполняется условие: $\psi^{/} \ge \psi$. Возможны следующие случаи:

1) $\gamma = \psi = \psi^{/}$: предполагаемое количество кластеров совпадает с количеством построенных, для каждого из кластеров определена единственная КТ; выполняется условие:

$$\left(\left|Y\right| = \left|K_j\right|\right) \wedge \left(\forall y_i \in Y\right)\left(\exists ! K_j\right)[y_i \in K_j],$$
$$i \in 1, ..., \gamma, \;\; j \in 1, ..., \psi; \;\; \gamma = \psi = \psi^{/}; \qquad (11)$$

тогда матрица P_E вероятностей принадлежности КТ кластерам – может быть приведена к единичной матрице $P_E = E$ (квадратная симметричная матрица).

2) $\gamma > \psi = \psi^{/}$: предполагаемое количество кластеров совпадает с количеством построенных, для каждого из кластеров определено не менее одной КТ; выполняется условие:

$$\left(\left|Y\right| > \left|K_j\right|\right) \wedge \left(\forall y_i \in Y\right)\left(\exists K_j\right)[y_i \in K_j],$$
$$i \in 1, ..., \gamma, \;\; j \in 1, ..., \psi, \;\; \gamma > \psi = \psi^{/}; \qquad (12)$$

матрица P_E отличается от единичной, $P_E \ne E$, P_E – прямоугольная матрица.

3) $\gamma = \psi < \psi^{/}$: количество построенных кластеров превышает предполагаемое количество кластеров, таким образом, не для всех построенных кластеров определены КТ; выполняется условие:

$$\left(\left|Y\right| < \left|K_j\right|\right) \wedge \left(\forall y_i \in Y\right)\left(\exists K_j\right)[y_i \in K_j],$$
$$i \in 1, ..., \gamma, \;\; j \in 1, ..., \psi, \;\; \gamma = \psi < \psi^{/}. \qquad (13)$$

Возможны следующие случаи:

3а) количество построенных кластеров на один превышает количество КТ:

$$\left|Y\right| + 1 = \left|K_j\right|, \;\; \left|K_j\right| - \left|Y\right| = 1. \qquad (14)$$

3б) количество построенных кластеров превышает количество КТ более, чем на один (более одного кластера не обозначено контрольной точкой):

$$\left|Y\right| + 1 < \left|K_j\right|, \;\; \left|K_j\right| - \left|Y\right| = \delta > 1. \qquad (15)$$

Замечание. В силу прикладной специфики поставленной задачи, предлагается приведение случая (**3б**) к случаю (**3а**) за счёт объединения δ кластеров, не отмеченных контрольными точками, в один общий кластер, как прочие элементы.

Определение 5. Перестановкой из элементов конечного множества I называется всякое упорядочивание элементов этого множества.

Определение 6. Обозначим через $P_T(I_t)$ матрицы, образованные из матрицы P_T перестановками I_t её столбцов $j = 1, \dots, k$:

$$P_T(I_0) = P_T,\ P_T(I_1), \dots, P_T(I_t),\ t = 0, \dots, k!-1,$$

где t – номер произведенной перестановки I_t, определяющей порядок следования столбцов матрицы P_T:

$$I_0 = (1, 2, \dots, k),\ I_1 = (2, 1, \dots, k),\ \dots,\ I_t,$$
$$t = 0, \dots, k!-1.$$

Замечание. Общее количество возможных перестановок k столбцов матрицы P_T составляет $k!$. Могут рассматриваться перестановки только тех столбцов матрицы P_T, которые не совпадают со столбцами P_E.

Определение 7. Для матричных пар $(P_T(I_t), P_E)$, $t = 0, \dots, (k!-1)$, определяется метрика $\rho : (P_T(I_t), P_E) \to R_+ \cup \{0\}$ вида:

$$\rho(P_T(I_t), P_E) = \sum_{r=1}^{|B|} \sum_{j=1}^{k} \min(p_{Trj}(I_t), p_{Erj}),$$
$$t = 0, \dots, k!-1, \qquad (16)$$

где $p_{Trj}(I_t)$ – элементы матрицы $P_T(I_t)$ со столбцами j, следующими в порядке, соответствующем произведённой перестановке I_t столбцов матрицы P_T; p_{Erj} – элементы матрицы P_E.

Замечание. Метрика $\rho(P_T(I_t), P_E) \geq 0$ позволяет определить меру совпадения элементов соответствующих матриц. Значение ρ прямо пропорционально количеству общих элементов $x_i \in X$ в кластерах матричной пары.

Утверждение 4. Перестановка I^* столбцов матрицы P_T по отношению к матрице P_E оптимальна тогда и только тогда, когда выполняется условие

$$\rho(P_T(I^*), P_E) = \max\left\{\rho(P_T(I_t), P_E) \middle| t = 0, \dots, k!-1\right\}, \qquad (17)$$

где ρ – метрика вида (16).

Замечание 1. Возвращаясь к прежним обозначениям, результат упорядочивания столбцов P_T относительно P_E:

$$P_T = P_T(I^*).$$

Замечание 2. Возможны варианты равновероятных случаев упорядочивания столбцов P_T относительно P_E. Если перестановка I^* в соответствии с (17) не единственна, имеет место неоднозначность относительно упорядочивания столбцов P_T. В этом случае метод кластерного анализа не применим к анализируемым данным, так как не адекватно описывает реальные данные, в соответствии с оценкой (начальным условием) *Эксперта*. Результаты работы такого метода кластеризации исключаются из дальнейшего рассмотрения.

Утверждение 5. Объект $x_k \in X$, соответствующий контрольной точке $y_i \in Y$, $x_k = y_i$, принадлежит кластеру K_{j_0}, $j_0 \in 1, \dots, \psi'$ тогда и только тогда, когда вероятность принадлежности $x_k = y_i$ кластеру в i-й строке упорядоченной матрицы P_T ненулевая и максимальна.

Утверждение 5 следует из Определения 1.

Замечание 1. В соответствии с принадлежностью элементов $x_k \in X$, соответствующих КТ (начальное условие), кластерам K_{j_0}, $j_0 \in 1, \dots, \psi'$ настраивается нумерация кластеров, на которые разделяется исследуемое множество X методом M_1.

Замечание 2. Описанная выше стандартная процедура упорядочивания столбцов тестируемой матрицы P_T по отношению к столбцам эталонной матрицы P_E применима в случае $\dim(P_T) = \dim(P_E)$. В случае $\dim(P_T) \neq \dim(P_E)$ – предлагается следующая **адаптированная процедура упорядочивания**:

Шаг 1. На основании P_T строится P_T' – матрица вероятностей принадлежности элементов кластерам, помеченным контрольными точками таким образом, что $\dim(P_T') = \dim(P_E) < \dim(P_T)$.

Шаг 2. Применяется стандартная процедура упорядочивания P_T' относительно P_E.

Шаг 3. С учетом замечания к Утверждению 3 кластеры, не отмеченные контрольны-

ми точками, объединяются в один общий кластер, как прочие элементы. Размерность упорядоченной матрицы P_T' дополняется до исходной размерности $\dim(P_T)$. Для этого в матрицу P_T' добавляется строка и столбец с вероятностями попадания КТ (с учётом дополнительной КТ) в соответствующие кластеры. Матрица P_T – упорядочена.

Утверждение 6. 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, $L_E < L^* \le L$, состоящих из шагов **0–4**:

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 2а. По результатам кластерного анализа (предыдущий шаг 1) в соответствии с Определением 4 по КТ кластеров строятся тестируемая (P_T) и эталонная (P_E) матрицы. Согласно Утверждению 4 определяется оптимальная перестановка столбцов P_T относительно P_E. На основании Утверждения 5 производится настройка нумерации построенных (шаг 1) кластеров.

Шаг 2б. По результатам кластерного анализа (шаг 1) проводится построение 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 $|B|\times k^{(w)}$, при этом

$$k^{(w)} = \max\left\{ k^{(w-1)}, k_{w+1} \right\}. \qquad (18)$$

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

$$(w < L-1) \wedge (S^{(w)} \le L_E), \qquad (19)$$

то $w := w+1$, шаги 1–4 повторяются.

Доказательство

Проводится методом математической индукции.

Пусть в результате выполнения шагов итерационной процедуры на последней итерации $w^* \ge 1$ $((L>3, L^* \ge 2) \Rightarrow w^* \ge 1)$ получены приведённые к одной размерности $|B|\times k^{(w^*)}$ упорядоченные матрицы

$$P_1^{(w^*)}, \dots, P_{w^*+1}^{(w^*)},\ w^* = \min\left\{ S^{(w^*)}, L-1 \right\}.$$

Поскольку итерация с номером $w^* \ge 1$ – последняя итерация шагов 0–4 процедуры, то условие продолжения шагов процедуры (19) не выполняется, то есть выполняется условие

$$(w^* \ge L-1) \vee (S^{(w^*)} > L_E). \qquad (20)$$

По построению w^* имеет место следующее ограничение для значений w^*:

$$w^* \le L-1. \qquad (21)$$

Из (20) и (21) следует:

$$(w^* \ge L-1) \wedge (w^* \le L-1) \Leftrightarrow \Leftrightarrow (\max(w^*) = L-1). \qquad (22)$$

По построению имеет место следующее ограничение для значений $S^{(w^*)}$

$$S^{(w^*)} \le w^* + 1. \qquad (23)$$

Из (20), (22) и (23) следует:

$$L_E \overset{(20)}{<} S^{(w^*)} \overset{(23)}{\le} w^* + 1 \overset{(22)}{\le} L-1+1 = L,$$

или

$$L_E < w^* + 1 \le L. \qquad (24)$$

Поскольку начальное значение параметра нумерации итераций $w=0$, из (24) получаем общее количество выполненных итераций шагов процедуры:

$$w = w^* + 1 = L^*,\ L^* \in (L_E, L].$$

Возвращаясь к прежним обозначениям, получим матрицы $P_l = P_l^{(w^*)}$, $l=1,\dots,L$, где $L := L^* = w^* + 1$ размерности $|B|\times k$ при $k = k^{(w^*)}$. Доказательство закончено.

Замечание 1. Построение **на шаге 2б процедуры** матриц P_{w+1} вида (3) проводится в соответствии с Утверждением 1 по формуле (4).

Замечание 2. Приведение матриц P_l, $l=1,\dots,w$, размерности $|B|\times k^{(w-1)}$ и P_{w+1} к одной размерности $k^{(w)}$ (18) **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 формуле (8):

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

$$p_{lrj}^{(w)} = \begin{cases} p_{lrj}^{(w-1)} \big| \quad j \le k^{(w-1)}, \\ 0 \quad\; \big| \quad k^{(w-1)} < j \le k^{(w)}, \end{cases} \qquad (25)$$

$$l = 1, \dots, w.$$

Замечание 3. Определение **на шаге 4 процедуры** количества $S^{(w)}$ эквивалентных матриц $P_l^{(w)}$, $l = 1, \dots, w+1$ возможно посредством применения следующей формулы:

$$S^{(w)} = \max\left\{ \left(\sum\nolimits_{j=1}^{w+1} d_{ij} - 1 \right) \Big| \, d_{ij} = 1 - \Delta_{ij}, \right.$$
$$\left. i = 1, \dots, w+1 \right\} \qquad (26)$$

$$\Delta_{ij} = \begin{cases} 0 \,\big|\quad P_i^{(w)} = P_j^{(w)}, \\ 1 \,\big|\quad P_i^{(w)} \ne P_j^{(w)}. \end{cases}$$

При расчете $S^{(w)}$ элементы главной диагонали d_{ii}, $i = 1, \dots, w+1$, не учитываются.

Этап 2 процедуры определения кластеров объектов

Шаг 1. Обобщение результатов кластеризации $P_1, \dots, P_L$ в соответствии с утверждением 2.

Шаг 2. Определение кластеров объектов.

В результате кластеризации множества X вида (1) L методами, получена обобщённая матрица $P = \left\| p_{rj} \right\|$, $r = 1, \dots, |B|$, $j = 1, \dots, k$, на основе анализа значений элементов которой определяется принадлежность объекта к определенному кластеру.

Утверждение 7. 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:

$$(b_r \in K_{j_0}) \Leftrightarrow (p_{rj_0} = \max\{p_{rj} \,\big|\, j = 1, \dots, k\}). \qquad (27)$$

Утверждение 7 следует из Определения 1.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Исследование обработки многомерных данных несколькими методами кластерного анализа проводится на 4-мерных данных об ирисах, предложенных Фишером в 1936 г. Эти данные описывают длину и ширину чашелистика и лепестка для трех видов ирисов: *Setosa*, *Virginic* и *Versicol*.

Таким образом, задано множество наблюдений $X = \{x_i \,|\, x_i \in R^4, i = 1, \dots, 150\}$, которое необходимо разделить на кластеры. Элементы $x_i \in X$, $i = 1, \dots, m$, $m = 150$, представляют собой измерения n параметров, $n = 4$, объектов $b_r \in B$, $r = 1, \dots, |B|$, $|B| = 3$ – 3-х видов ирисов:

$$B = \{Setosa, Virginic, Versicol\}.$$

К данным об ирисах применена процедура определения кластеров объектов методами Tree Clustering, K-Means и Fuzzy Relation Clustering (FRC). Кластерный анализ методами K-Means и Tree Clustering проводится с использованием пакета STATISTICA 6.0. Метод кластерного анализа FRC реализован в программно-технологическом комплексе имитации сложных систем BelSim [1].

Количество применяемых методов кластеризации: $L = 3$. Пусть *Экспертом* в соответствии с уровнем значимости задачи определена точная нижняя граница L_E величины L^* ($L_E < L^* \le L$): $L_E = L/2$, соответственно $L^* = 2$.

Ожидаемое количество кластеров: $k = 3$. Пусть выполняется следующее начальное условие: заданы контрольные точки кластеров:

$$x_1 \in K_1 \subset X\,,\; x_{150} \in K_2 \subset X\,,\; x_2 \in K_3 \subset X\,.$$

Этап 1. Итерация 0.

Шаг 0. Параметру номера итерации w присваивается начальное значение $w = 0$.

Шаг 1. Проводится кластеризация элементов множества X методом M_1 – Tree Clustering.

Шаг 2. По результатам кластерного анализа (**шаг 1**) множество наблюдений X, $|X| = 150$, разделено на 3 кластера.

Необходимо упорядочить номера кластеров по контрольным точкам в соответствии с заданным начальным условием. Пусть P_E – эталонная матрица вида (3) вероятностей принадлежности 3-х контрольных точек $x_1 \in K_1$, $x_{150} \in K_2$, $x_2 \in K_3$, кластерам K_j, $j = 1, \dots, 3$; P_T – матрица вероятностей при-

надлежности рассматриваемых контрольных точек кластерам K_j в соответствии с результатом кластерного анализа (**шаг 1**, итерация 0):

$$P_E = \begin{pmatrix} 1,00 & 0,00 & 0,00 \\ 0,00 & 0,00 & 1,00 \\ 0,00 & 1,00 & 0,00 \end{pmatrix},$$

$$P_T = \begin{pmatrix} 0,00 & 1,00 & 0,00 \\ 1,00 & 0,00 & 0,00 \\ 1,00 & 0,00 & 0,00 \end{pmatrix}.$$

Необходимо упорядочивание столбцов P_T по отношению к матрице P_E.

Общее количество возможных перестановок столбцов матрицы P_T составляет $3! = 6$. Рассматриваются следующие перестановки столбцов P_T:

$$I_0 = (1,2,3),\ I_1 = (2,1,3),\ I_2 = (1,3,2),$$

$$I_3 = (3,2,1),\ I_4 = (2,3,1),\ I_5 = (3,1,2).$$

Для матричных пар $(P_T(I_t), P_E),\ t = 0, \dots, 5$ определяются значения метрики ρ:

$$\rho(P_T(I_0), P_E) = 0,\ \rho(P_T(I_1), P_E) = 2,$$

$$\rho(P_T(I_2), P_E) = 0,$$

$$\rho(P_T(I_3), P_E) = 1,\ \rho(P_T(I_4), P_E) = 2,$$

$$\rho(P_T(I_5), P_E) = 1.$$

В соответствии с утверждением 4, оптимальной по отношению к матрице P_E может быть одна из перестановок столбцов P_T: I_1, I_4, так как

$$\rho((P_T(I^*), P_E)) =$$

$$= \rho(P_T(I_1), P_E) = \rho(P_T(I_4), P_E) \overset{(17)}{=}$$

$$\overset{(17)}{=} \max\left\{\rho(P_T(I_t), P_E) \middle| t = 0, \dots, 5\right\} = 2.$$

Возвращаясь к прежним обозначениям, в соответствии с результатом упорядочивания столбцов P_T:

$$P_{T1} = \begin{pmatrix} 1,00 & 0,00 & 0,00 \\ 0,00 & 1,00 & 0,00 \\ 0,00 & 1,00 & 0,00 \end{pmatrix},$$

$$P_{T2} = \begin{pmatrix} 1,00 & 0,00 & 0,00 \\ 0,00 & 0,00 & 1,00 \\ 0,00 & 0,00 & 1,00 \end{pmatrix}.$$

В соответствии с P_{T1}: $x_1 \in K_1$, $x_{150}, x_2 \in K_2$; в соответствии с P_{T2}: $x_1 \in K_1$, $x_{150}, x_2 \in K_3$. Контрольные точки (КТ) x_{150}, x_2, заданные *Экспертом* как КТ разных кластеров, попадают в результате кластеризации множества X методом Tree Clustering в один кластер. Результаты кластеризации множества X данным методом исключаются из дальнейшего рассмотрения.

Шаги 3, 4. Пропускаются; параметр $w = 1$, осуществляется переход к следующей итерации этапа 1 процедуры кластеризации множества данных X несколькими методами.

Этап 1. Итерация 1.

Шаг 1. Проводится кластеризация элементов множества X методом M_2 – K-Means.

Шаг 2. По результатам кластерного анализа (**шаг 1**, итерация 1) множество X, $|X| = 150$, разделено на 3 кластера.

Пусть P_E – эталонная матрица вида (3) вероятностей принадлежности 3-х контрольных точек $x_1 \in K_1$, $x_{150} \in K_2$, $x_2 \in K_3$, кластерам K_j, $j = 1, \dots, 3$; P_K – матрица вероятностей принадлежности рассматриваемых контрольных точек кластерам K_j в соответствии с результатом кластерного анализа (**шаг 1**, итерация 1). Поскольку выполняется условие $P_K = P_E$, в дополнительном упорядочивании нумерации кластеров в соответствии с заданным начальным условием нет необходимости.

Проведено построение матрицы P_2 вида (3) вероятностей принадлежности 3-х видов ирисов ($|B| = 3$), кластерам K_j, $j = 1, \dots, 3$, ($k_2 = 3$), Размерность матрицы P_2: 3×3.

$$P_2 = \begin{pmatrix} 1,00 & 0,00 & 0,00 \\ 0,00 & 0,28 & 0,72 \\ 0,00 & 0,96 & 0,04 \end{pmatrix}.$$

Шаг 3. Пропускается. Размерность построенной матрицы P_2: $k^{(1)} = 3$, в соответствии с (18) вводится обозначение: $P_2^{(1)} = P_2$.

Шаг 4. Пропускается; $w = 2$, осуществляется переход к следующей итерации этапа 1 процедуры кластеризации множества данных X несколькими методами.

Этап 1. Итерация 2.

Шаг 1. Проводится кластеризация элементов множества X методом M_3 – FRC.

Шаг 2. По результатам кластерного анализа (**шаг 1**, итерация 2) (при $\alpha = 0{,}85$) множество наблюдений X, $|X| = 150$, разделено на 4 кластера.

Пусть P_E – эталонная матрица вида (3) вероятностей принадлежности 3-х контрольных точек $x_1 \in K_1$, $x_{150} \in K_2$, $x_2 \in K_3$, кластерам K_j, $j = 1, \ldots, 3$; P_F – матрица вероятностей принадлежности рассматриваемых контрольных точек кластерам K_j в соответствии с результатом кластерного анализа (**шаг 1**, итерация 2). Поскольку выполняется условие $P_F = P_E$, в дополнительном упорядочивании нумерации кластеров K_1, K_2, K_3 в соответствии с заданным начальным условием нет необходимости.

Кластер K_4, согласно методу FRC, состоит из одного элемента $K_4 = \{x_{25}\}$, $|K_4| = 1$.

Проведено построение матрицы P_3 вида (3) вероятностей принадлежности 3-х видов ирисов ($|B| = 3$), кластерам K_j, $j = 1, \ldots, 4$, ($k_3 = 4$). Размерность матрицы P_3: 3×4.

$$P_3 = \begin{pmatrix} 1{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}98 & 0{,}02 \\ 0{,}00 & 0{,}08 & 0{,}92 & 0{,}00 \end{pmatrix}.$$

Шаг 3. Размерности построенных матриц $P_2^{(1)}$ и P_3 не совпадают.

Необходимо привести матрицы к одной размерности:

$$k^{(2)} \overset{(18)}{=} \max\{k^{(1)}, k_3\} = \max\{3, 4\} = 4.$$

Приведение матриц $P_2^{(1)}$, P_3 к необходимой размерности выполняется по формулам (8) за счет их дополнения столбцами с нулевыми вероятностями попадания объекта в добавленные кластеры.

$$P_2^{(2)} = \begin{pmatrix} 1{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}28 & 0{,}72 & 0{,}00 \\ 0{,}00 & 0{,}96 & 0{,}04 & 0{,}00 \end{pmatrix},$$

$$P_3^{(2)} = P_3 = \begin{pmatrix} 1{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}98 & 0{,}02 \\ 0{,}00 & 0{,}08 & 0{,}92 & 0{,}00 \end{pmatrix}.$$

Шаг 4. При $w = 2$, $L = 3$ выполняется условие $w = L - 1$; условие (19) не выполняется; на шаге $w = 2$ этап 1 процедуры кластеризации множества данных X несколькими методами окончен.

Этап 2. Обобщение результатов кластеризации. Определение кластеров.

Пусть P_2, P_3 – матрицы вида (3) вероятностей принадлежности 3-х видов ирисов определенным кластерам K_j, $j = 1, \ldots, 4$ согласно методам кластерного анализа K-Means и Fuzzy Relation Clustering, соответственно.

$$P_2 = \begin{pmatrix} 1{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}28 & 0{,}72 & 0{,}00 \\ 0{,}00 & 0{,}96 & 0{,}04 & 0{,}00 \end{pmatrix},$$

$$P_3 = \begin{pmatrix} 1{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}98 & 0{,}02 \\ 0{,}00 & 0{,}08 & 0{,}92 & 0{,}00 \end{pmatrix},$$

$$P = \begin{pmatrix} 1{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}28 & 0{,}99 & 0{,}02 \\ 0{,}00 & 0{,}96 & 0.92 & 0{,}00 \end{pmatrix}.$$

Матрица P является обобщенной матрицей вероятностей принадлежности трёх видов ирисов *Setosa*, *Virginic* и *Versicol* определенным кластерам.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Обобщение результатов кластеризации методами K-Means, FRC и построение матрицы P проведено по формуле (4).

При $L = 2$, $L^* = 1$, $k = 4$, $|B| = 3$ и формула (4) расчета элементов обобщенной матрицы P принимает вид:

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

где p_{lrj} – элементы матриц P_2, P_3 соответственно, $l = 2, 3$, $r = 1, \ldots, 3$, $j = 1, \ldots, 4$.

Анализ обобщенного результата кластеризации по матрице P позволяет выделить 3 кластера: $K_1 = \{\text{Setosa}\}$ – с вероятностью принадлежности видов ирисов кластеру, равной 100 %; $K_2 = \{\text{Virginic}\}$, $K_3 = \{\text{Versicol}\}$ – с вероятностями принадлежности видов ирисов кластерам, равными 99 % и 96 %, соответственно.

ЛИТЕРАТУРА

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

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

3. *Паклин Н.* Алгоритмы кластеризации на службе Data Mining / Н. Паклин // Технологии анализа данных. – 2011. – Режим доступа : http: // www.basegroup.ru/library/analysis/clusterization/datamining/. – Дата доступа: 12.03.2011.

4. *Башаримов В. В.* Выбор методов кластерного анализа при решении задач оптимизации в имитационном моделировании / В. В. Башаримов, Е. М. Борчик, А. И. Якимов // Информационные технологии, энергетика и экономика (информационные технологии, математическое моделирование технологических процессов, электроника): сб. трудов 7-ой Межрег. (межд.) науч.-техн. конф. студентов и аспирантов, 8–9 апр. 2010 г. : в 3 т. – Смоленск : ф-л ГОУ ВПО МЭИ(ТУ), 2010. – Т. 2. – С. 21–26.

5. Методы, средства и технологии исследования временных последовательностей статистических данных в имитационном моделировании : отчет о НИР (заключ.) / Белорус.-Рос. ун-т ; рук. Е. А. Якимов ; исполн. : Р. В. Петров [и др.]. – Могилев, 2011. – 126 с. – Библиогр. : с. 124–126. – № ГР 20091957. – Инв. № Ф09М-171.

6. *Якимов А. И.* О совместном использовании методов кластерного анализа многомерных данных / А. И. Якимов, Е. М. Борчик, В. В. Башаримов // Доклады БГУИР. – 2011. – № 5(59). – С. 95–102.

Аверченков В. И. – д.т.н., профессор, заведующий кафедрой Компьютерные технологии и системы, Брянский государственный технический университет.
E-mail: aver@tu-bryansk.ru

Якимов А. И. – к.т.н., доцент, доцент кафедры Автоматизированные системы управления, Белорусско-Российский университет.
E-mail: ykm@tut.by

Борчик Е. М. – выпускник аспирантуры (Белорусско-Российский университет), ведущий инженер-программист Управления информационных технологий ОАО «Моготекс», Могилев, Беларусь.
E-mail: katrinb15@gmail.com

Башаримов В. В. – выпускник аспирантуры (Белорусско-Российский университет), технический директор, Частное унитарное предприятие «Авем Студио», Могилев, Беларусь.
E-mail: basharimovvv@tut.by

Averchenkov V. I. – Doctor of Technical Sciencies, Professor, Head of the Department of Computer Technologies and Systems, Bryansk State Technical University.
E-mail: aver@tu-bryansk.ru

Yakimov A. I. – Candidate of Engineering Sciences, Docent, Department of Automated control systems, Belarusian-Russian University.
E-mail: ykm@tut.by

Borchik E. M. – graduate student (Belarusian-Russian University), leading engineer-programmer Department of information technology, OJSC ” Mogoteks”.
E-mail: katrinb15@gmail.com

Basharimov V. V. – graduate student (Belarusian-Russian University), technical Director, PUE “Avem Studio”.
E-mail: basharimovvv@tut.by