УДК 519.237.8

ОСОБЕННОСТИ ПРОВЕДЕНИЯ РАЗДЕЛЯЮЩЕЙ КЛАСТЕРИЗАЦИИ ДЛЯ ИНТЕРВАЛЬНЫХ ВЕЛИЧИН

О. М. Бердникова, Т. М. Леденева

Воронежский государственный университет

Поступила в редакцию 27.11.2016 г.

Аннотация. В данной статье описаны основные особенности проведения кластерного анализа для интервальных величин. Введены определения основных числовых характеристик (внутрикластерная дисперсия, общая внутрикластерная дисперсия, общая межкластерная дисперсия) для интервальных величин и сформулирован алгоритм бинарной разделяющей кластеризации для разбиения множества объектов, характеристики которых определяются переменными интервального типа. На примере показана корректность работы алгоритма.

Ключевые слова: интервальная переменная, бинарная кластеризация, дисперсия.

Annotation. This article describes the main features of the cluster analysis for interval-valued values. Definitions of the main numerical characteristics (within-cluster variance, total within-cluster variance, between-cluster variance) for interval values are entered. An algorithm of binary divisive top-down clustering for partition the set of initial objects, which characteristics are defined by variables of interval type is formulated. Propriety of work of the algorithm is demonstrated by the example.

Keywords: interval variable, binary clustering, variance.

1. ПОСТАНОВКА ЗАДАЧИ

Пусть задано множество объектов $X = \{x_1, \ldots, x_m\}$, свойства которых характеризуются показателями $P_1, \ldots, P_n$, причем каждому показателю P_i ставится в соответствие переменная интервального типа Y_i.

Согласно [1], под *интервальной* будем понимать переменную, значением которой является промежуток одного из следующих типов: (a,b), $(a,b]$, $[a,b)$, $[a,b]$, где $a \le b$, $a, b \in R$.

В результате оценочной процедуры каждому объекту x_u будет поставлена в соответствие векторная оценка $\xi_u = (\xi_{u1}, \ldots, \xi_{un})$, а множеству объектов матрица $E = \{\xi_{ui}\}_{m \times n}$, в которой ξ_{ui} – значение интервальной переменной Y_i для объекта x_u. Требуется построить разбиение множества объектов X на группы (кластеры, классы) похожих в некотором смысле объектов. Приведенная постановка задачи является основной задачей кластерного анализа, а ее особенность заключается в интервальном представлении исходных данных, что требует разработки специальных алгоритмов для формирования кластеров.

Большинство существующих алгоритмов кластеризации основаны на определении меры сходства/несходства между объектами, что позволяет похожие объекты объединять в группы с учетом заданного порога. Меры сходства/несходства с учетом различных типов информации [1,4] формируются на основе функций расстояния [1–3], функций подобия [4], индексов сходства/несходства [6, 7]. Существует множество алгоритмов кластеризации, использующих данные меры, среди них особое внимание заслуживают алгорит-

мы иерархической кластеризации, строящие систему вложенных разбиений. Результатом работы таких алгоритмов является дерево кластеров. Выделяют следующие стратегии алгоритмов иерархической кластеризации: восходящая и нисходящая. Разделяющая кластеризация – это кластеризация с нисходящей стратегией, использующая подход «сверху вниз», при котором все исходные объекты помещаются в один кластер, а затем разделяются до тех пор, пока каждый объект не окажется в своем «индивидуальном» кластере. Примером алгоритмов такого класса является метод декомпозиционного дерева [5]. Бинарная разделяющая кластеризация – это вид разделяющей кластеризации, при котором на каждом последующем этапе кластер предыдущего уровня может быть разделен только на два новых кластера. Заметим, что при решении практических задач процесс кластеризации проводят до тех пор, пока или каждый объект не будет выделен в отдельный класс, или не будет достигнуто заранее определенное число кластеров.

Цель статьи заключается в построении алгоритма бинарной разделяющей кластеризации для объектов, векторные оценки которых имеют интервальное представление.

2. ОСНОВНЫЕ ПОНЯТИЕ И ОПРЕДЕЛЕНИЯ

Пусть задано некоторое множество X, B_X – его булеан, $\beta_X = \{X_i : X_i \subseteq X\}_{i=\overline{1,k}} \subset B_x$ – система непустых подмножеств множества X.

Систему подмножеств β_X назовем *разбиением*, если выполнятся следующие условия: $\bigcup_{i=1} X_i = X$ и $\forall i \neq j \ \left(X_i \cap X_j = \varnothing\right)$. Подмножества X_i называются *кластерами* или *классами разбиения*. Всякое разбиение множества X образует фактор-множество множества X по некоторому отношению эквивалентности. Известно, что любое отношение эквивалентности на X порождает некоторое разбиение этого множества и, обратно, каждому разбиению соответствует отношение эквивалентности.

Для всякой системы подмножеств β_X множества X можно построить ориентированный граф $G = (\beta_X, V)$ по следующему правилу:

1) каждому подмножеству $X_i \in \beta_X$, $i = \overline{1, |\beta_X|}$ поставим в соответствие вершину x_i;

2) Если $X_i \subset X_j$, то между соответствующими вершинами имеется дуга $(x_i, x_j) \in V$.

Граф назовем *иерархией*, если выполняется условие

$$\forall x_j \left(\Gamma(x_j) = \varnothing \vee \begin{Bmatrix} X_j = X_{j_1} \cup X_{j_2} \\ X_{j_1} \cap X_{j_2} = \varnothing \end{Bmatrix} \right),$$

где $\Gamma(u) = \{v : (u,v) \in V\}$, $X_{j_1} \cup X_{j_2} \in \beta_X$.

Если $\bigcup_{\{j:\, x_j \in Low\}} X_j = X$, где $Low = \{x_j : \Gamma(x_j) = \varnothing\}$, то иерархия определяет разбиение вершин заданного множества X.

Подмножества X_j, являющиеся кластерами, в дальнейшем будем обозначать C_j.

3. ОСНОВНЫЕ ПРИНЦИПЫ БИНАРНОЙ РАЗДЕЛЯЮЩЕЙ КЛАСТЕРИЗАЦИИ ДЛЯ ИНТЕРВАЛЬНЫХ ВЕЛИЧИН

Рассмотрим подробно процедуру разбиения исходного множества объектов X на кластеры C_r $\left(r = \overline{1, N}\right)$, $\bigcup_{r=1}^{N} C_r = X$. Начнем с единственного класса C_0, включающего все объекты из X, т. е. на первом шаге $C_0 = X$. Затем C_0 разбивается на два кластера C_1^1 и C_1^2, на следующем этапе один из кластеров C_1^1 или C_1^2 также разбивается на два, так что $C_1^s = C_2^{s1} \cup C_2^{s2}$, где $s = 1, 2$. После второй итерации получим разбиение исходного множества объектов на три кластера – $C_0 = \{C_1^t, C_2^{s1}, C_2^{s2}\}$, где $t, s = 1, 2$, $t \neq s$. На следующем шаге будет разбит один из указанных трех кластеров. Таким образом, будем продолжать процесс, каждый раз разбивая один из кластеров, полученных на данном шаге, на два новых, пока количество результирующих кластеров не станет равным заранее определенному значению или заданному количеству уровней в дереве кластеризации.

В рамках рассмотренной процедуры на каждом шаге необходимо ответить на следующие вопросы:

1) Какой кластер выбрать для разбиения?

2) Каким образом осуществить разбиение выбранного кластера?

Предположим, что мы находимся на уровне разбиения r, и совокупность кластеров представляет собой множество $Q_r = \{C_1,..,C_r\}$, причем $X = \bigcup_{j=1}^{r} C_j$. Для выбора кластера C_k для последующего разбиения будем использовать следующие критерии: минимизация общего количества различий в пределах одного кластера с одновременной максимизацией различий между имеющимися кластерами. Пусть выбран кластер C_k, содержащий m_k объектов, который необходимо разбить на два новых кластера, так что $C_k = C_k^1 \cup C_k^2$. Для этого вводится некоторое условие $q(\cdot)$, основанное на фактических значениях переменных и действующее таким образом, что один из новых классов будет содержать объекты, удовлетворяющие этому условию, а другой – нет, т. е.

$$\begin{aligned} C_k^1 &= \{x_u, \text{ если } q(x_u) = 1\}, \\ C_k^2 &= \{x_u, \text{ если } q(x_u) = 0\}. \end{aligned} \quad (1)$$

В данном случае при работе с интервальными величинами будем использовать правила $q(\cdot)$ следующего вида:

$$q_r(x_i) \ : \ \overline{x}_{ij} \le c_{sj},$$

где c_{si} – некоторая точка среза значений интервальной переменной Y_i; $\overline{x}_{ui} = \frac{a_{ui} + b_{ui}}{2}$ – средняя точка значения $\xi_{ui} = [a_{ui}, b_{ui}]$. Заметим, что правило $q(\cdot)$ применимо только к одной переменной Y_i. В результате получим новое разбиение $Q_{r+1} = \{C_1,..,C_k^{\ 1}, C_k^{\ 2},..,C_r\} = \{C_1,..,C_{r+1}\}$.

Для формализации критериев разбиения на кластеры введем некоторые числовые характеристики.

Пусть x_{u_1}, x_{u_2} – объекты из множества X, Y_i – интервальная переменная, соответствующая показателю P_i. Мерой несходства Ichino-Yaguchi между двумя объектами по показателю P_i называется число, которое определяется следующим образом [1]:

$$\begin{aligned} \phi_j(x_{u_1}, x_{u_2}) = &\left|\xi_{u_1 i} \oplus \xi_{u_2 i}\right| - \left|\xi_{u_1 i} \otimes \xi_{u_2 i}\right| + \\ &+\gamma\left(2\left|\xi_{u_1 i} \otimes \xi_{u_2 i}\right| - \left|\xi_{u_1 i}\right| - \left|\xi_{u_2 i}\right|\right), \end{aligned} \quad (2)$$

где $0 \le \gamma \le 0.5$ – параметр; $|A|$ – длина интервала $A = [a,b]$; операции $\oplus$ и $\otimes$ – специальные операции сложения и умножения, определяемые для интервальных величин $A_j = [a_j^{\ A}, b_j^{\ A}]$, $B_j = [a_j^{\ B}, b_j^{\ B}]$ следующим образом:

$$A_j \oplus B_j = \left[\min\left\{a_j^A, a_j^B\right\}, \max\left\{b_j^A, b_j^B\right\}\right],$$

$$A_j \otimes B_j = \left[\max\left\{a_j^A, a_j^B\right\}, \min\left\{b_j^A, b_j^B\right\}\right].$$

Частным случаем меры несходства (2) является расстояние Хаусдорфа, которое для интервальных величин определяется следующим образом:

$$\phi_i(x_{u_1}, x_{u_2}) = \max_i\left[\left|a_{u_1 i} - a_{u_2 i}\right|, \left|b_{u_1 i} - b_{u_2 i}\right|\right]. \quad (3)$$

Заметим, что формулы (2) и (3) позволяют определить несходство объектов x_{u_1}, x_{u_2} по i-му показателю в виде числа $\phi_i(x_{u_1}, x_{u_2})$, а, следовательно, каждой паре объектов будет соответствовать вектор $(\phi_1(x_{u_1}, x_{u_2}),..., \phi_n(x_{u_1}, x_{u_2}))$. Для определения обобщенной оценки несходства можно использовать функции агрегирования, в том числе, различные средние [8, 11] и порядковые операции взвешенного агрегирования [9]. В [10] предлагается нечеткая система для оценки сходства/несходства, база правил которой формируется с участием экспертов.

Для агрегирования частных оценок несходства $\phi_i(x_{u_1}, x_{u_2})$ будем использовать взвешенную среднюю квадратическую вида [11]

$$\begin{aligned} d(x_{u_1}, x_{u_2}) &= \left(\frac{\sum_{j=1}^{n} \rho_i \left(\phi_i(x_{u_1}, x_{u_2})\right)^2}{\sum_{j=1}^{n} \rho_j}\right)^{1/2} = \\ &= \left(\sum_{i=1}^{n} w_i \left(\phi_i(x_{u_1}, x_{u_2})\right)^2\right)^{1/2}, \end{aligned} \quad (4)$$

где $w_i = \frac{\rho_i}{\sum_{j=1}^{n} \rho_j}$ – нормированный вес показателя P_i.

Введем обобщенные характеристики кластеров и разбиений, основанных на этом типе средних.

Внутрикластерной дисперсией для кластера $C_k = \{x_1, .., x_{m_k}\}$ называется величина

$$CVar(C_k) = \left(\frac{\sum_{u_1=1}^{m_k} \sum_{u_2=1}^{m_k} \rho_{u_1} \rho_{u_2} \left(d\left(x_{u_1}, x_{u_2}\right)\right)^2}{\sum_{u_1=1}^{m_k} \sum_{u_2=1}^{m_k} \rho_{u_1} \rho_{u_2}} \right)^{1/2} = \left(\frac{\sum_{u_1=1}^{m_k} \sum_{u_2>u_1}^{m_k} \rho_{u_1} \rho_{u_2} \left(d\left(x_{u_1}, x_{u_2}\right)\right)^2}{\sum_{u_1=1}^{m_k} \sum_{u_2>u_1}^{m_k} \rho_{u_1} \rho_{u_2}} \right)^{1/2}, \quad (5)$$

где ρ_{u_1}, ρ_{u_2} – веса соответствующих объектов x_{u_1} и x_{u_2} (здесь возможно такое преобразование формулы, поскольку матрица расстояний для класса C_k является симметричной, так как $d\left(x_{u_1}, x_{u_2}\right) = d\left(x_{u_2}, x_{u_1}\right)$, и содержит 0 на главной диагонали).

Если ввести весовой коэффициент

$$w_{u_1,u_2} = \frac{\rho_{u_1} \rho_{u_2}}{\sum_{u_1=1}^{m_k} \sum_{u_2>u_1}^{m_k} \rho_{u_1} \rho_{u_2}},$$

то формула (5) примет более простой вид

$$CVar(C_k) = \sum_{u_1=1}^{m_k} \sum_{u_2>u_1}^{m_k} w_{u_1,u_2} d^2\left(x_{u_1}, x_{u_2}\right), \quad (6)$$

Если $p_u = 1/m$, где m – общее число объектов в X, то

$$CVar(C_k) = \frac{1}{m^2} \sum_{u_1=1}^{m_k} \sum_{u_2>u_1}^{m_k} m_k (m_k - 1) d^2\left(x_{u_1}, x_{u_2}\right). \quad (7)$$

Общей внутрикластерной дисперсией для системы кластеров $Q_r = \{C_1, .., C_r\}$ называется величина

$$TotalCVar(Q_r) = \sum_{k=1}^{r} CVar(C_k). \quad (8)$$

Общей межкластерной дисперсией для системы кластеров $Q_r = \{C_1, .., C_r\}$ называется величина

$$BetweenClas(Q_r) = TotalCVar(X) - TotalCVar(Q_r), \quad (9)$$

где $TotalCVar(X)$ – общая внутрикластерная дисперсия исходного множества объектов X.

Пусть $Q_r(k) = \{C_k^1, C_k^2\}$ – разбиение кластера C_k, тогда из всех возможных вариантов разбиения $Q_r(k)$ (всего существует $C_{m_k}^2 = m_k(m_k - 1)/2$ различных вариантов) необходимо выбрать разбиение, удовлетворяющее следующему условию:

$$\min_{L_k} \{TotalCVar(Q_r(k))\} = \min_{L_k} \{CVar(C_k^1) + CVar(C_k^2)\}, \quad (10)$$

где L_k множество всех разбиений кластера C_k.

Рассмотрим, каким образом выбрать кластер для разбиения. Пусть для выбранного кластера C_k разбиение выглядит следующим образом: $Q_{r+1} = \{C_1, .., C_k^1, C_k^2, .., C_r\}$, тогда общая внутрикластерная дисперсия будет иметь вид:

$$TotalCVar(Q_{r+1}) = \sum_{k'=1}^{r} CVar(C_{k'}) - CVar(C_k) + CVar(C_k^1) + CVar(C_k^2). \quad (11)$$

Выбор нового разбиения должен быть произведен таким образом, чтобы минимизировать общую межкластерную дисперсию или, что аналогично, максимизировать различие между полученными кластерами, то есть необходимо найти k', удовлетворяющее следующему условию:

$$k': \max_k \left\{ \underbrace{CVar(C_k) - CVar(C_k^1) - CVar(C_k^2)}_{\Delta W(C_k)} \right\} = \max_k \{\Delta W(C_k)\}. \quad (12)$$

Еще один важный момент – это признак остановки или определение значения уровня, когда процесс кластеризации может быть завершен. В соответствии с определением иерархии процесс может продолжаться до тех пор, пока каждый объект не образует кластер, однако, на практике важно получить хорошо интерпретируемые кластеры, поэтому необходима дополнительная информация.

4. АЛГОРИТМ БИНАРНОЙ РАЗДЕЛЯЮЩЕЙ КЛАСТЕРИЗАЦИИ ДЛЯ ИСХОДНОЙ ИНФОРМАЦИИ ИНТЕРВАЛЬНОГО ТИПА

1. Исходная информация: множество объектов $X=\{x_1,\ldots,x_m\}$, свойства которых характеризуются множеством показателей $P_1,\ldots,P_n$. Каждому показателю P_i ставится в соответствие переменная Y_i интервального типа. В результате оценки или наблюдения объекта x_u формируется векторная оценка $\xi_u=(\xi_{u1},\ldots,\xi_{un})$, где ξ_{ui} – значение интервальной переменной Y_i. Векторные оценки всех объектов x_u $\left(u=\overline{1,m}\right)$ целесообразно свести в матрицу $E=\{\xi_{ui}\}_{m\times n}$, где $\xi_{ui}=[a_{ui},b_{ui}]$ – значение интервальной переменной Y_j для объекта x_i. Задать желаемое количество уровней в иерархии разбиения R.

2. Для каждого значения $\xi_{ui}=[a_{ui},b_{ui}]$ $\left(u=\overline{1,m},\ i=\overline{1,n}\right)$ найдем среднюю точку $\overline{\xi}_{ui}=\dfrac{a_{ui}+b_{ui}}{2}$.

3. Для каждой переменной Y_i $\left(i=\overline{1,n}\right)$ определим точки среза. Для этого отсортируем по возрастанию величины $\overline{\xi}_{ui}$ $\left(u=\overline{1,m}\right)$ и соответствующие им объекты x_u. Найдем точки среза следующим образом:

$$c_{si}=\frac{\overline{\xi}_{si}+\overline{\xi}_{s,i+1}}{2}\quad\left(s=\overline{1,m-1},\ i=\overline{1,n}\right).$$

4. Найдем матрицу расстояний $D=\left\{d\left(x_{u_1},x_{u_2}\right)\right\}_{m\times m}$, используя формулу (4).

5. Пусть первоначальное разбиение $Q_1=C_0=X$, $r=1$, где r – количество уровней в иерархии разбиений.

Повторяем до тех пор, пока $r\le R$:

• Для каждого кластера C_k, $k=\overline{1,r}$ ищем всевозможные варианты разбиений $C_{k,s}=\left\{C_{k,s}^1,C_{k,s}^2\right\}$ и определяем значения $CVar\left(C_{k,s}^1\right)$, $CVar\left(C_{k,s}^2\right)$, $W\left(C_{k,s}\right)=CVar\left(C_{k,s}^1\right)+CVar\left(C_{k,s}^2\right)$, используя формулы (7) и (8).

• Для каждого кластера C_k $\left(k=\overline{1,r}\right)$ находим разбиение $C_{k,\min}=C_{k,s}=\left\{C_{k,s}^1,C_{k,s}^2\right\}$ со значением $\min\limits_s\left\{W\left(C_{k,s}\right)\right\}$.

• Для каждого класса C_k $\left(k=\overline{1,r}\right)$ для найденного разбиения $C_{k,\min}$ находим $\Delta W\left(C_{k,\min}\right)=CVar\left(C_k\right)-W\left(C_{k,\min}\right)$.

• Из всех разбиений $C_{k,\min}$ $\left(k=\overline{1,r}\right)$ выбираем разбиение с $\max\limits_k\left\{\Delta W\left(C_{k,\min}\right)\right\}$. Таким образом, найдены $C_{\max}$ и само разбиение $C_{k,\min}$.

• Определяем соответствующую точку среза $c_{s,\min}$ разбиения $C_{k,\min}$ для класса $C_{\max}$.

• Для найденной точки среза $c_{s,\min}$ сформулируем правило $q_r:\overline{\xi}_{ij}\le c_{s,\min}$, согласно которому осуществляется разбиение на два новых кластера в соответствии с правилом (1).

• Получили новое разбиение $Q_{r+1}=\left\{C_1,..,C_k^{\,1},C_k^{\,2},..,C_r\right\}$.

5. ИЛЛЮСТРАТИВНЫЙ ПРИМЕР

В табл. 1 представлены интервальные оценки объектов $x_1,\ldots,x_8$ по двум показателям. Для формирования кластеров будем использовать метод бинарной разделяющей кластеризации, предполагая, что желательно получить три кластера.

Сформируем матрицу расстояний (табл. 2)

Для всех значений $\xi_{ui}=[a_{ui},b_{ui}]$, $u=\overline{1,8}$, $i=1,2$ определим средние точки, отсортируем их в порядке возрастания и найдем точки среза c_{si} на основе формулы (табл. 3). На первой итерации начальное разбиение $Q_0=X$. Согласно алгоритму, необходимо для каждого класса найти всевозможные разбиения и определить для них значения $CVar\left(C_{k,s}^1\right)$, $CVar\left(C_{k,s}^2\right)$, $TotalCVar\left(C_{k,s}\right)$. Рассмотрим подробнее разбиение на примере $s=3$.

Таблица 1

Матрица интервальных оценок объектов

	x_1	x_1	x_1	x_1	x_1	x_1	x_1	x_1
P_1	[410, 460]	[410, 460]	[410, 460]	[410, 460]	[410, 460]	[410, 460]	[410, 460]	[410, 460]
P_2	[158, 175]	[158, 175]	[158, 175]	[158, 175]	[158, 175]	[158, 175]	[158, 175]	[158, 175]

Таблица 2

Матрица расстояний

	x_3	x_7	x_8	x_6	x_2	x_1	x_5	x_4
x_3	0.000	56.569	56.569	263.059	263.532	284.607	422.734	441.814
x_7	56.569	0.000	0.000	280.943	260.768	291.349	440.710	460.679
x_8	56.569	0.000	0.000	280.943	260.768	291.349	440.710	460.679
x_6	263.059	280.943	280.943	0.000	220.021	240.252	240.133	240.008
x_2	263.532	260.768	260.768	220.021	0.000	31.048	180.069	200.063
x_1	284.607	291.349	291.349	240.252	31.048	0.000	150.030	170.356
x_5	422.734	440.710	440.710	240.133	180.069	150.030	0.000	21.541
x_4	441.814	460.679	460.679	240.008	200.063	170.188	22.825	0.000

Таблица 3

Первая итерация

Объекты	ξ_{ij}	$\overline{\xi}_{ij}$	c_{js}	s	$CVar(C_s^1)$	$CVar(C_s^2)$	$TotalCVar(C_s)$	$\Delta W(C_s)$
x_3	[130,190]	160.0	165.0	1	0	28988.1	28988.1	5750.8
x_7	[170,170]	170.0	170.0	2	200.0	20522.3	20722.2	14016.6
x_8	[170,170]	170.0	240.0	3	266.7	8670.2	**8936.9**	**25801.9**
x_6	[170,450]	310.0	360.0	4	7295.6	3119.0	11214.6	23524.3
x_2	[390,430]	410.0	422.5	5	12182.9	2166.4	14349.3	20389.5
x_1	[410,460]	435.0	472.5	6	16599.4	30.8	16630.2	18108.7
x_5	[410,610]	510.0	515.0	7	26366.5	0	26366.5	8372.4
x_4	[410,630]	520.0	–	–	–	–	34738.8	–

$$C_0 = \left\{C_0^{\,1} = \{x_3, x_7, x_8\}, C_0^{\,2} = \{x_6, x_2, x_1, x_5, x_4\}\right\},$$

$$CVar\left(C_0^1\right) = \frac{1}{8\times 3}[(56,57)^2 +$$

$$+(56,57)^2 + 0^2] = 266,67,$$

$$CVar\left(C_0^2\right) = \frac{1}{8\times 5}[(220,02)^2 + (240,25)^2 + \ldots +$$

$$+(21,54)^2] = 8670,15,$$

$$TotalCVar\left(C_0\right) = CVar\left(C_0^1\right) + CVar\left(C_0^2\right) =$$

$$= 266,67 + 8670,2 = 89369,$$

$$CVar\left(C_0\right) = \frac{1}{8\times 8}[(56,57)^2 + \ldots +$$

$$+(21,54)^2] = 34738,8,$$

$$\Delta W\left(C_0\right) = CVar\left(C_0\right) - TotalCVar\left(C_0\right) =$$

$$34738,8 - 8936,9 = 25801,9.$$

На основе расчетов, приведенных в табл. 3, можно сделать вывод, что данное разбиение искомое. Соответственно, следующее разбиение будет иметь вид $Q_1 = \left\{C_1 = \{x_3, x_7, x_8\}, C_2 = \{x_6, x_2, x_1, x_5, x_4\}\right\}$. С помощью правила $q_1 : Y_1 \le 240$ происходит распределение исходных объектов по новым кластерам.

Действуя аналогичным образом на второй итерации получим разбиение $Q_2 = \left\{C_1 = \{x_3, x_7, x_8\}, C_2 = \{x_6\}, C_3 = \{x_2, x_1, x_5, x_4\}\right\}$. Распределение объектов из C_2 по новым кластерам осуществляется по правилу $q_2 : Y_1 \le 360$. На третьей итерации проверяются всевозможные варианты разбиений для классов C_1, C_2, C_3. С учетом правила $q_3 : Y_1 \le 472,5$ получается следующее разбиение $Q_3 = \left\{C_1 = \{x_3, x_7, x_8\}, C_2 = \{x_6\}, C_3 = \{x_2, x_1\}, C_4 = \{x_5, x_4\}\right\}$, которое и является окончательным. На рис. 1 представлено дерево кластеризации.

ЗАКЛЮЧЕНИЕ

Алгоритмы кластеризации широко используются для структуризации заданного множества объектов. Предложенный в ста-

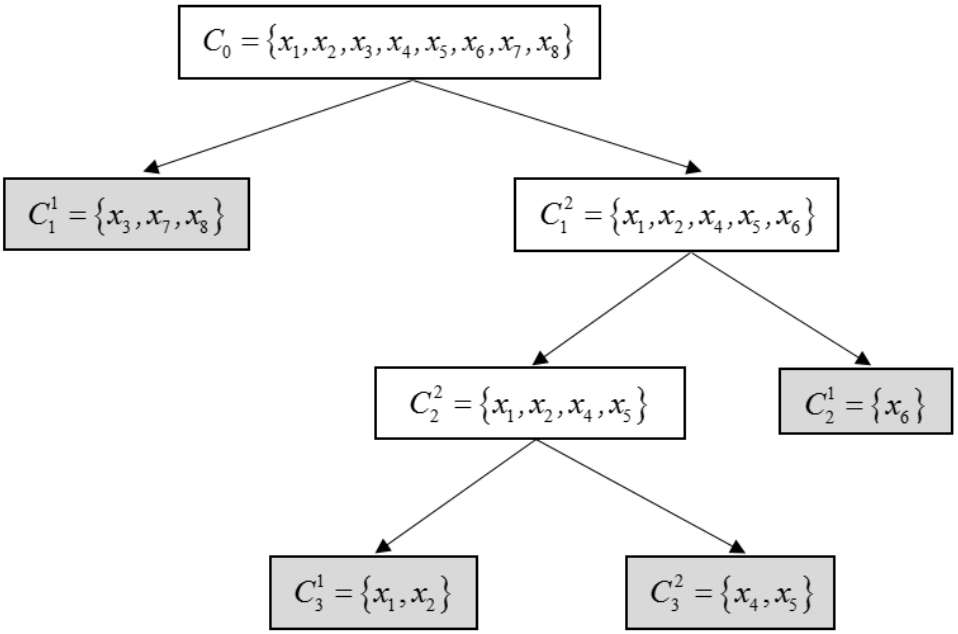

Рис. 1. Иллюстрация процесса кластеризации и окончательный результат

тье алгоритм позволяет найти не просто совокупность кластеров, а систему вложенных разбиений, так что результат представляется деревом. В качестве критериев используются два широко известных на практике – максимизация «похожести» объектов из одного класса и максимизация «различий» между классами, что позволяет сформировать достаточно компактные классы.

Важно, что метод позволяет обрабатывать приближенную информацию об объектах, заданную в интервальном виде. Очевидно, что он может быть распространен на нечеткую информацию в форме нечетких чисел, поскольку нечеткие задачи сводятся к интервальным с помощью теоремы о декомпозиции. Недостатком метода является то, что при определении оптимального разбиения какого-либо кластера необходимо исследовать всевозможные разбиения, а их может оказаться достаточно много, поэтому необходимо разработать способ приближенной оценки возможных разбиений.

СПИСОК ЛИТЕРАТУРЫ

1. *Billard L.* Symbolic Data Analysis: Conceptual Statistics and Data Mining / L. Billard, E. Diday. – England : John Wiley & Sons Ltd, 2006. – 321 p.

2. *Bertrand P.* Descriptive Statistics for Symbolic Data / P. Bertrand, F. Goupil. – Berlin : Springer-Verlag, 2000. – 245 p.

3. *Бердникова О. М.* Дискриптивная статистика для интервальных наблюдений при наличии логических правил / О. М. Бердникова, Т. М. Леденева / Вестник Воронежского государственного технического университета. – 2014. – №10.5. – С. 34–39.

4. *Леденева Т. М.* О представлении информации в задачах классификации / Т. М. Леденева, Нгуен Н.Х. // Вестник Воронежского государственного технического университета. – Воронеж, 2012. – Т. 8, № 7.1. – С. 33–38.

5. *Леденева Т. М.* О влиянии функции подобия на результаты нечеткой классификации / Т. М. Леденева, Н. Х. Нгуен // Информационные технологии. – М. : Новые технологии, 2011. – № 11. – С. 15–23.

6. *Нгуен Н. Х.* О мерах несходства для смешанных типов данных / Н. Х. Нгуен, Т. М. Леденева // Сб. трудов Междунар. конф. «Актуальные проблемы прикладной математики, информатики и механики» (Воронеж, 26–28 сентября 2011 г.). – Воронеж : ИПЦ ВГУ, 2011. – С. 242–246.

7. *Леденева Т. М.* О свойствах нечеткого отношения сходства / Т. М. Леденева, Р. К. Стрюков // Вестник Воронеж. гос. ун-та. Сер. Системный анализ и информационные технологии. – 2014. – № 4. – С. 75-79.

8. *Леденева Т. М.* Аксиоматический подход к построению функций агрегирования для оценочных систем / Т. М. Леденева, Д. А. Денисихина // Вестник Воронеж. гос. ун-та. Сер. Системный анализ и информационные технологии. – 2014. – № 3. – С. 33-39.

9. *Леденева Т. М.* Операторы агрегирования в оценочных моделях // Т. М. Леденева, Т. Н. Недикова // Информационные технологии, 2003. – № 2. – С. 2–9.

10. *Леденева Т. М.* Нечеткая система для оценки сходства многомерных объектов / Т. М. Леденева, Р. К. Стрюков, О. М. Бердникова // Системы управления и информационные технологии, 2016. – №2(64). – С. 114–120.

11. *Джини К.* Средние величины / К. Джини. – М. : Статистика, 1970. – 448 с.

Бердникова О. М. – аспирант кафедры вычислительной математики и прикладных информационных технологий, факультет прикладной математики, информатики и механики, Воронежский государственный университет.
E-mail: ms.oksana1904@mail.ru

Леденева Т. М. – д-р техн. наук, профессор, заведующий кафедрой вычислительной математики и прикладных информационных технологий, факультет прикладной математики, информатики и механики, Воронежский государственный университет.
E-mail: ledeneva-tm@yandex.ru

Berdnikova O. M. – Aspirant, Computational Mathematics and Applied Computer Science, faculty of Applied Mathematics, Computer sciences and Mechanics, Voronezh State University.
E-mail: ms.oksana1904@mail.ru

Ledeneva T. M. – Doctor of Technical Sciencies, Professor, Computational Mathematics and Applied Computer Science, faculty of Applied Mathematics, Computer sciences and Mechanics, and Mechanics, Voronezh State University.
E-mail: ledeneva-tm@yandex.ru