УДК 519.87

Е.В. Ивохин, Д.В. Апанасенко

КЛАСТЕРИЗАЦИЯ СОВОКУПНОСТИ СОСТАВНЫХ НЕЧЕТКИХ ЧИСЕЛ НА ОСНОВЕ МНОЖЕСТВ СКАЛЯРНОГО И ВЕКТОРНОГО УРОВНЕЙ

Введение

Традиционная задача группирования данных в относительно однородные группы не только остается актуальной в современных условиях, но и постоянно получает новое развитие в силу необходимости оперативной обработки огромных потоков информации и решения проблемы построения информационных сообществ [1]. Задача, возникающая в самых различных сферах науки, экономики, социума и т.д., включает в себя проблему выделения кластеров и задачу нахождения эталонных узлов в совокупности данных. При этом в реализации методов кластеризации важную роль имеют характеристики принадлежности данных к группе, способы вычисления расстояний между элементами, весовые коэффициенты, зависящие от конкретной предметной области и наблюдений, полученных в рамках этой области. Методы имеют разную вычислительную сложность и разную степень эффективности при решении конкретных прикладных задач.

Для проведения процесса кластеризации существует множество подходов, большинство из которых основывается на эвристических методах, реализующих определенные процедурные схемы исследователя и не требующих сложных статистических расчетов. Однако в случае обработки нечеткой (недостаточно определенной — fuzzy) информации их использование существенно усложняется или вообще становится невозможным из-за специфического представления нечеткости. Рассмотрим методику кластеризации неточно заданной информации, для формализации которой используются совокупности составных нечетких множеств [2].

Составные нечеткие множества как обобщение нечетких множеств Заде

Определение 1 [3]. Нечетким множеством $\tilde{A}$ в универсальном пространстве X называется совокупность пар вида $\{(x, \mu_{\tilde{A}}(x))\}$, где $x \in X$, а $\mu_{\tilde{A}}(x): X \to [0,1]$ — функция принадлежности нечеткого множества $\tilde{A}$.

Величину функции принадлежности $\mu_{\tilde{A}}(x)$ для произвольного элемента $x \in X$ называют степенью принадлежности x нечеткому множеству $\tilde{A}$. Интерпретацией степени принадлежности $\mu_{\tilde{A}}(x)$ является субъективная мера того, насколько элемент $x \in X$ соответствует понятию, смысл которого формализуется нечетким множеством $\tilde{A}$.

При этом обычное множество $L_0 = \{x \in X : \mu_{\tilde{A}}(x) > 0\}$ называется носителем нечеткого множества $\tilde{A}$, а множества $L(\alpha) = \{x \in X : \mu_{\tilde{A}}(x) \geq \alpha\}$ — множествами уровня $\alpha \in (0,1]$ нечеткого множества $\tilde{A}$.

Традиционно в качестве универсального множества (пространства) X рассматривается произвольное подмножество конечномерного пространства R^n: $X \subseteq R^n$, $x = (x_1, ..., x_n) \in R^n$.

Следует подчеркнуть, что в последнее время теория нечетких множеств превратилась в детально изученную область с широким спектром прикладных задач, в которых используется термин «нечеткая величина». Его впервые использовал Кофман [4], затем он появился в работах [3, 5].

Одним из способов описания нечетких величин является подход, предложенный Намиасом [5], в котором используется понятие пространства возможностей.

Определение 2 [6]. Нечеткая величина определяется как функция из пространства возможностей $(\theta, \mathrm{P}(\theta), \mathrm{Pos})$ на множество действительных чисел R^1, где θ — некоторое непустое множество, $\mathrm{Pos}\{A\}$ — величина меры возможности для произвольного $A \in \mathrm{P}(\theta)$, $\mathrm{P}(\theta)$ — множество всех подмножеств для θ.

С практической точки зрения данное определение недостаточно конструктивно. Однако необходимо заметить, что нечеткие величины могут быть определены разными способами. Для формализации нечетких величин на практике часто используется понятие нечеткого числа, которое эквивалентно классическому определению нечеткого множества с определенными условиями, накладываемыми на функцию принадлежности.

Пусть в качестве универсального множества X рассматривается подмножество действительных чисел, т.е. $X \subseteq R^1$. В этом случае нечеткое множество $\tilde{b}$ содержит совокупность пар, составленных из двух скалярных значений: $x \in R^1$ и $\mu_{\tilde{b}}(x)$. Будем считать, что функция принадлежности $\mu_{\tilde{b}}(x)$ имеет следующий вид:

$$\begin{gathered}\mu_{\tilde{b}}(x) = \frac{x-a}{b-a}, \ x \in [a, b]; \\ \mu_{\tilde{b}}(x) = \frac{c-x}{c-b}, \ x \in [b, c]; \\ \mu_{\tilde{b}}(x) = 0, \ x \notin [a, c].\end{gathered} \tag{1}$$

В этом случае утверждается, что задано нечеткое число $\tilde{b}$ треугольного вида, т.е. $\tilde{b} = \{(a, b, c)\}$, $a \le b \le c$, с линейными функциями принадлежности вида (1).

Не ограничивая общности, рассмотрим нечеткие числа $\tilde{b}_1, \ldots, \tilde{b}_m$ треугольного вида с линейными функциями принадлежности $\mu_{\tilde{b}_1}(x), \ldots, \mu_{\tilde{b}_m}(x)$, определенные в соответствующих универсальных множествах $X_1, \ldots, X_m$, где $X_i \subseteq R^1$, $i = \overline{1, m}$. При этом очевидно, что для носителей нечетких чисел справедливо включение $[a_i, c_i] \subseteq X_i$, $i = \overline{1, m}$. В дальнейшем для краткости будем говорить, что имеется множество нечетких чисел.

Из произвольных элементов этого множества сформируем множество вида

$$\tilde{b}^m = \{(x^1, \mu_{\tilde{b}_1}(x^1)), (x^2, \mu_{\tilde{b}_2}(x^2)), \ldots, (x^m, \mu_{\tilde{b}_m}(x^m))\}, \ x^i \in [a_i, c_i], \ i = \overline{1, m}. \tag{2}$$

Определение 3. Произвольное множество $\tilde{b}^m$ вида (2) будем называть составным нечетким числом на множестве нечетких чисел $\tilde{b}_1, \ldots, \tilde{b}_m$ треугольного вида.

Необходимо отметить, что составное нечеткое число не является в общем случае нечетким множеством, так как при этом невозможно определить универсальное множество, на котором оно задается. С другой стороны, несложно видеть, что составное нечеткое число представляет собой вектор из элементов нечетких чисел размерности m. Кроме того, этот агрегат также не следует отождествлять с

нечетким множеством $\tilde{A}^m = \{(x, \mu_{\tilde{A}^m}(x))\}$, $x = (x^1, x^2, ..., x^m) \in R^m$, задающим понятие нечеткого вектора, для построения которого используется универсальное множество вида $\underset{i=1}{\overset{m}{\times}} X_i \subseteq R^m$, с единственной функцией принадлежности $\mu_{\tilde{A}^m}(x)$ [4].

Заданную совокупность составных нечетких чисел $\tilde{b}^m$ конечного объема $N \geq 1$ будем обозначать $K(\tilde{b}^m)$, $\left|K(\tilde{b}^m)\right| = N$.

Определение 4. Множество

$$L_0^m = \{x^1 \in X_1, x^2 \in X_2, ..., x^m \in X_m : \mu_{\tilde{b}_1}(x^1) > 0,$$

$$\mu_{\tilde{b}_2}(x^2) > 0, ..., \mu_{\tilde{b}_m}(x^m) > 0\}, \; L_0^m \subseteq \underset{i=1}{\overset{m}{\times}} X_i,$$

назовем носителем составного нечеткого числа $\tilde{b}^m$.

Определение 5. Множество

$$L^m(\alpha) = \{x^1 \in X_1, x^2 \in X_2, ..., x^m \in X_m : \mu_{\tilde{b}_1}(x^1) \geq \alpha,$$

$$\mu_{\tilde{b}_2}(x^2) \geq \alpha, ..., \mu_{\tilde{b}_m}(x^m) \geq \alpha\}, \; L^m(\alpha) \subseteq L_0^m,$$

назовем множеством скалярного уровня $\alpha \in (0, 1]$ составного нечеткого числа $\tilde{b}^m$.

Определение 6. Множество

$$L^m(\alpha_1, ..., \alpha_m) = \{x^1 \in X_1, x^2 \in X_2, ..., x^m \in X_m : \mu_{\tilde{b}_1}(x^1) \geq \alpha_1,$$

$$\mu_{\tilde{b}_2}(x^2) \geq \alpha_2, ..., \mu_{\tilde{b}_m}(x^m) \geq \alpha_m\},$$

$L^m(\alpha_1, ..., \alpha_m) \subseteq L_0^m$, назовем множеством векторного уровня $(\alpha_1, ..., \alpha_m)$, $\alpha_i \in (0, 1]$, $i = \overline{1, m}$, составного нечеткого числа $\tilde{b}^m$.

Очевидно, что множества $L_{0_i} = \{x \in X_i : \mu_{\tilde{b}_i}(x) > 0\}$, $i = \overline{1, m}$, являются носителями соответствующих нечетких чисел $\tilde{b}_1, \ldots, \tilde{b}_m$, множества $L_i(\alpha) = \{x \in X_i : \mu_{\tilde{b}_i}(x) \geq \alpha\}$, $i = \overline{1, m}$, — множествами уровня $\alpha \in (0, 1]$ нечетких чисел $\tilde{b}_i$, $i = \overline{1, m}$, и справедливы следующие соотношения:

$$L_0^m = L_{0_1} \times L_{0_2} \times ... \times L_{0_m},$$

$$L^m(\alpha) = L_1(\alpha) \times L_2(\alpha) \times ... \times L_m(\alpha),$$

$$L^m(\alpha_1, ..., \alpha_m) = L_1(\alpha_1) \times L_2(\alpha_2) \times ... \times L_m(\alpha_m).$$

Определение 7. Составное нечеткое число $\tilde{b}^{m|0}$ вида $\tilde{b}^{m|0} = \{(x^1, 0), (x^2, 0), ..., (x^m, 0)\}$, $\forall x^i \in X_i$, $i = \overline{1, m}$, назовем полностью вырожденным, а составное нечеткое число $\tilde{b}^{m|j} = \{(x^1, \mu_{\tilde{b}_1}(x^1)), ..., (x^j, 0), ..., (x^m, \mu_{\tilde{b}_m}(x^m))\}$, $\mu_{\tilde{b}_i}(x^i) > 0$, $\forall x^i \in X_i$, $i = \overline{1, m}$, $i \neq j$, — вырожденным по j-й компоненте, $j = \overline{1, m}$.

Это имеет место в случае, когда нечеткое число $\tilde{b}^m$ строится из элементов $x^i \in X_i$, $i = \overline{1, m}$, в полном объеме или частично не принадлежащих носителям нечетких множеств $\tilde{b}_i$, $i = \overline{1, m}$. Очевидно, что составное нечеткое число будет

вырожденным по компонентам $j_1,..., j_k$, $j_p = \overline{1, m}$, $p = \overline{1, k}$, если оно имеет вид $\tilde{b}^{m|j_1,...,j_k} = \{(x^1, \mu_{\tilde{b}_1}(x^1)),..., (x^{j_1}, 0),..., (x^{j_k}, 0),..., (x^m, \mu_{\tilde{b}m}(x^m))\}$, $\mu_{\tilde{b}_i}(x^i) > 0$, $\forall x^i \in X_i$, $i = \overline{1, m}$, $i \neq j_1, \ldots, i \neq j_k$, и справедливо соотношение $\tilde{b}^{m|0} = \tilde{b}^{m|1,...,m}$.

Определение 8. Нечеткие числа $\tilde{b}_i$, $i = \overline{1, m}$, будем называть проекциями составных нечетких чисел, принадлежащих $K(\tilde{b}^m)$.

Определение 9. Составное нечеткое число $\tilde{b}_j^{m-1}$, построенное из элементов нечетких чисел $\tilde{b}_1, \ldots, \tilde{b}_{j-1}, \tilde{b}_{j+1}, \ldots, \tilde{b}_m$, назовем срезом совокупности составных нечетких чисел $K(\tilde{b}^m)$ по j-й компоненте, $i = \overline{1, m}$.

Пусть задано два невырожденных составных нечетких числа:

$$\tilde{u}^m = \{(x^1, \mu^u_{\tilde{b}_1}(x^1)), (x^2, \mu^u_{\tilde{b}_2}(x^2)),..., (x^m, \mu^u_{\tilde{b}_m}(x^m))\},$$

$$\tilde{v}^m = \{(y^1, \mu^v_{\tilde{b}_1}(y^1)), (y^2, \mu^v_{\tilde{b}_2}(y^2)),..., (y^m, \mu^v_{\tilde{b}_m}(y^m))\}.$$

Вычислим величину $\gamma = \min\limits_{i=\overline{1,m}} \min\{\mu^u_{\tilde{b}_i}(x^i), \mu^v_{\tilde{b}_i}(y^i)\}$, которая является минимальным значением среди значений мер принадлежности отдельных элементов обоих чисел $\tilde{u}^m$, $\tilde{v}^m$. Это значение позволяет построить два множества скалярного уровня γ в виде обычных множеств $L^m_{\tilde{u}}(\gamma)$, $L^m_{\tilde{v}}(\gamma)$, определяющие точки носителя L^m_0 составного нечеткого числа $\tilde{b}^m$, между которыми может быть вычислено расстояние $d = \| L^m_{\tilde{u}}(\gamma) - L^m_{\tilde{v}}(\gamma) \|$.

Определение 10. Нечеткая величина по Вонгу [7]

$$\tilde{W}(\tilde{b}^m) = \{(d, \gamma) : d = \| L^m(\gamma) \|, \ \gamma = \min\limits_{i=\overline{1,m}} \mu_{\tilde{b}_i}(x^i) \in (0, 1], \ x^i \in X_i, \ i = \overline{1, m}\},$$

где $\|\cdot\|$ — евклидова норма вектора в R^m, определяет метрику на множестве $K(\tilde{b}^m)$.

Таким образом, каждое составное нечеткое число $\tilde{b}^m$ «измеряется» с помощью нечеткой величины $\tilde{W}(\tilde{b}^m)$, а для определения нечеткого расстояния $\tilde{\rho}(\tilde{u}^m, \tilde{v}^m)$ между произвольными составными нечеткими числами $\tilde{u}^m$, $\tilde{v}^m$ можно использовать нечеткую величину

$$\tilde{\rho}(\tilde{u}^m, \tilde{v}^m) = \{(d, \gamma) : d = \| L^m_{\tilde{u}}(\gamma) - L^m_{\tilde{v}}(\gamma) \|, \gamma = \min\limits_{r \in \{u,v\}, i=\overline{1,m}} \mu^r_{\tilde{b}_i}(\cdot)\},$$

где $L^m_{\tilde{u}}(\gamma)$, $L^m_{\tilde{v}}(\gamma)$ — множества скалярного уровня $\gamma \in (0, 1]$ составных нечетких чисел $\tilde{u}^m$, $\tilde{v}^m$ соответственно.

В случае использования множеств векторного уровня построение такой метрики требует большей формализации. Проблема состоит в том, что в результате построения множеств векторного уровня $(\alpha_1,..., \alpha_m)$, $\alpha_i \in (0, 1]$, $i = \overline{1, m}$, для произвольных составных нечетких чисел $\tilde{u}^m$, $\tilde{v}^m$ получается два обычных множества: $L^m_{\tilde{u}}(\alpha_1,..., \alpha_m)$ и $L^m_{\tilde{v}}(\alpha_1,..., \alpha_m)$, элементы которых характеризуются разной степенью принадлежности. Естественно, существуют различные способы аксиоматизации расстояния между этими множествами.

Воспользуемся методами нахождения компромиссного решения в задачах векторной оптимизации [8]. Предположим, что для заданных чисел $\tilde{u}^m$, $\tilde{v}^m$ построены множества уровня $L^m_{\tilde{u}}(\alpha_1,...,\alpha_m)$ и $L^m_{\tilde{v}}(\alpha_1,...,\alpha_m)$. Это означает, что выполнены соотношения $\mu^u_{\tilde{b}_i}(x^i) \ge \alpha_i$, $\mu^v_{\tilde{b}_i}(y^i) \ge \alpha_i$, $i = \overline{1, m}$. Тогда можно построить составное нечеткое число вида $\tilde{D}^m = \{(|x^i - y^i|, \alpha_i), i = \overline{1, m}\}$.

Несложно заметить, что процедура формирования компромиссного решения будет состоять в решении задачи двухкритериальной оптимизации вида

$$|x^i - y^i| \to \max\limits_{i=\overline{1,m}}, \quad \alpha_i \to \max\limits_{i=\overline{1,m}}, \tag{3}$$

решение которой $d = \arg\max\limits_{i=\overline{1,m}} |x^i - y^i|$, $\gamma = \arg\max\limits_{i=\overline{1,m}} \alpha_i$, позволяет определить расстояние между множествами $L^m_{\tilde{u}}(\alpha_1,...,\alpha_m)$ и $L^m_{\tilde{v}}(\alpha_1,...,\alpha_m)$ в виде нечеткой величины по Вонгу $\tilde{W}(\tilde{D}^m) = \{(d, \gamma)\}$.

Таким образом, нечеткое расстояние $\tilde{\rho}(\tilde{u}^m, \tilde{v}^m)$ между произвольными составными нечеткими числами $\tilde{u}^m$, $\tilde{v}^m$ при использовании множеств векторного уровня можно определить как нечеткую величину $\tilde{W}(\tilde{D}^m)$, полученную в результате решения задачи (11).

Принцип сравнения расстояний между составными нечеткими числами

Сравнение нечетких расстояний при использовании множеств скалярного и векторного уровня проводится на основе способа обработки нечетких отношений, введенного С.А. Орловским [9]. Согласно этому способу, если заданы $\tilde{\rho}_1 = \tilde{\rho}(\tilde{u}^m_1, \tilde{v}^m_1)$, $\tilde{\rho}_2 = \tilde{\rho}(\tilde{u}^m_2, \tilde{v}^m_2)$ — нечеткие расстояния между составными нечеткими числами $\tilde{u}^m_1, \tilde{v}^m_1$ и $\tilde{u}^m_2, \tilde{v}^m_2$ соответственно, то можно определить нечеткое расстояние, являющееся меньшим, в понимании нечеткого отношения предпочтения «< ».

Для нахождения «меньшего» расстояния предполагается, что нечеткое отношение предпочтения $g(a, b)$ для произвольных нечетких элементов $a \in \tilde{A}, b \in \tilde{B}$ при условии $\mu_g(a, b) = \min\{\mu_{\tilde{A}}(a), \mu_{\tilde{B}}(b), \mu_{\tilde{A}\times\tilde{B}}(a, b)\} > 0$ определяет соотношение в виде $a < b$ со степенью $g(a, b) = \mu_g(a, b)$.

С его помощью можно сравнивать расстояния $\tilde{\rho}_1 = \tilde{\rho}(\tilde{u}^m_1, \tilde{v}^m_1)$ и $\tilde{\rho}_2 = \tilde{\rho}(\tilde{u}^m_2, \tilde{v}^m_2)$: выражение $g(\tilde{\rho}_1, \tilde{\rho}_2)$ означает степень того, насколько $\tilde{\rho}_1$ «меньше» $\tilde{\rho}_2$. Это также позволяет определить «ближайший» к заданному составному нечеткому числу $\tilde{z}^0$ элемент $\tilde{z}^*$, где $\tilde{z}^*$ — составное нечеткое число, значения функций принадлежности которого для каждого $x_i \in X_i$, $i = \overline{1, m}$, определяются из соотношений

$$\mu^{z^*}_{\tilde{b}_i}(x_i) = \min\limits_{x\in X_i}(1 - \mu_T(x_i, x)) = 1 - \max\limits_{x\in X_i}\mu_T(x_i, x), \quad i = \overline{1, m}.$$

Здесь T — нечеткое отношение строгого предпочтения, соответствующего $g(a, b)$, $a, b \in X_i$, $i = \overline{1, m}$,

$$\mu_T(a, b) = \begin{cases} 0, & \text{если } \mu_g(a, b) < \mu_g(b, a), \\ \mu_g(a, b) - \mu_g(b, a) & \text{в противном случае.} \end{cases}$$

Использование введенного понятия расстояния между составными нечеткими числами и способа их сравнения позволяет сформулировать алгоритмы нечеткого группирования данных, представленных в виде совокупности составных нечетких чисел из $K(\tilde{b}^m)$. Данные алгоритмы обобщают наиболее известные подходы к решению задачи кластеризации. При этом использование каждого из них имеет свои специфические особенности, о которых будет сказано ниже.

Методы кластеризации нечетких данных

Разработка конструктивных алгоритмов кластеризации нечетких данных, представленных совокупностью составных нечетких чисел из $K(\tilde{b}^m)$, включает в себя формализацию способов поиска кластерного центра множества и реализацию процедуры группирования нечетких данных в пределах заданного количества кластеров.

В литературе, посвященной проблемам кластеризации [10–12], чаще всего рассматривается три алгоритма: *C*-means, пикового группирования и разностного группирования. Эти методы представляют собой кардинально разные подходы к решению проблемы кластеризации. Как правило, первый из них требует от исследователя точного знания количества кластеров, на базе чего строится алгоритм. Второй метод требует существенных временных затрат на выполнение, осуществляя проверку достаточно большого количества нетривиальных условий принадлежности элемента к кластеру и построение самого кластера. Третий — наиболее оптимальный, но для его применения необходимо детальное знание специфических зависимостей в системе для установки значений важных параметров, без которых метод не дает желаемого результата.

Пусть задана некоторая совокупность невырожденных составных нечетких чисел $\{\tilde{A}^{m(j)},\ j=\overline{1,\ p}\}$ из $K(\tilde{b}^m)$. В дальнейшем будем считать, что используется скалярный уровень $0<\gamma\le 1$ или векторный уровень $\gamma=(\gamma_1,...,\gamma_m),\ \gamma_i\in(0,1]$, которые определяют гарантированный уровень нечеткости рассматриваемых данных.

Метод нечеткого группирования *C*-means. Предположим, что заданную совокупность можно сгруппировать в *k* кластеров [13].

Определим *k* центров кластеров в виде набора составных нечетких чисел:

$$\tilde{C}^{m(i)}=\{(x^{1(i)},\ \mu_{\tilde{A}_1}(x^{1(i)})),(x^{2(i)},\ \mu_{\tilde{A}_2}(x^{2(i)})),...,(x^{m(i)},\ \mu_{\tilde{A}_m}(x^{m(i)}))\},\ i=\overline{1,k}.$$

Исходя из того, что точные значения центров группирования неизвестны, алгоритм кластеризации должен быть итерационным. В начале процесса значения центров группирования выберем случайным образом.

Предположим, что составные нечеткие числа $\tilde{A}^{m(j)},\ j=\overline{1,q},\ q\le p$, имеют непустые множества уровня $L^m(\bar{\gamma})$, представляющие собой обычные векторы $S(\tilde{A}^{m(j)})=\{x^{1(j)},\ x^{2(j)},...,x^{m(j)}\},\ j=\overline{1,q}$. Понятно, что векторы $S(\tilde{A}^{m(j)})$ будут принадлежать к разным группам с центрами $S(\tilde{C}^{m(i)}),\ i=\overline{1,k}$, соответственно со степенью принадлежности $u_{ij},\ i=\overline{1,k},\ j=\overline{1,q}$. Начальные величины значений принадлежности также могут быть заданы случайно, с учетом очевидного условия $\sum_{i=1}^{k}u_{ij}=1$ для каждого $j=\overline{1,q}$. Без ограничения общности, $u_{ij},\ i=\overline{1,k},\ j=\overline{1,q}$, могут быть определены, например, в виде $u_{ij}=1/k,\ i=\overline{1,k},\ j=\overline{1,q}$.

Функцию ошибок, которая соответствуют такому представлению, можно определить в виде суммы частичных ошибок принадлежности центрам с учетом значений степеней принадлежности u_{ij}, $i=\overline{1,k}$, $j=\overline{1,q}$, соответственно в виде

$$E(S)=\sum_{i=1}^{k}\sum_{j=1}^{q}u_{ij}\left\|S(\tilde{C}^{m(i)}),S(\tilde{A}^{m(j)})\right\|. \tag{4}$$

Тогда задача группирования будет заключаться в нахождении минимальных значений функции ошибки $E(S)$ с p ограничениями типа $\sum_{i=1}^{k}u_{ij}=1$, $j=\overline{1,q}$. Решение этой задачи представляется в виде [13]

$$S(\tilde{C}^{m(i)})=\sum_{j=1}^{q}u_{ij}S(\tilde{A}^{m(j)})\Big/\sum_{j=1}^{q}u_{ij}\,, \tag{5}$$

$$u_{ij}=1\Big/\sum_{t=1}^{k}(\xi_{ij}^{2}/\xi_{tj}^{2})^{1/(m-1)}, \tag{6}$$

где $\xi_{ij}=\xi(S(\tilde{C}^{m(i)}),S(\tilde{A}^{m(j)}))=\left\|S(\tilde{C}^{m(i)}),S(\tilde{A}^{m(j)})\right\|$, $i=\overline{1,k}$, $j=\overline{1,q}$, — евклидово расстояние между парами векторов, определяющих центры $\tilde{C}^{m(i)}$, $i=\overline{1,k}$, и составные нечеткие числа $\tilde{A}^{m(j)}$, $j=\overline{1,q}$, соответственно.

Окончательно алгоритм формулируется следующим образом.

1. Задать вектор величин уровня $\gamma=(\gamma_1,...,\gamma_m)$, $\gamma_i\in(0,1]$.

2. Определить значения коэффициентов u_{ij}, $i=\overline{1,k}$, $j=\overline{1,q}$, с учетом условий нормирования.

3. Определить k центров кластеров в соответствии с (5).

4. Вычислить значения функции ошибки в соответствии с формулой (4). Если ее значение не превышает некоторой пороговой величины или при условии незначительного уменьшения значения ошибки относительно предыдущей итерации, то вычисления прекращаются. Текущие значения центров кластеров составляют решение задачи. В противном случае необходимо перейти к п. 4.

5. Рассчитать новые значения u_{ij}, $i=\overline{1,k}$, $j=\overline{1,q}$, по формулам (6) и перейти к п. 2.

Метод пикового группирования. В алгоритме пикового группирования, предложенного в [14], рассмотрена мера плотности размещения векторов $S(\tilde{A}^{m(j)})$, $j=\overline{1,q}$, для этого генерируются так называемые пиковые функции. При использовании p исходных векторов создается сетка, равномерно накрывающая пространство векторов $S(\tilde{A}^{m(j)})$, $j=\overline{1,q}$. Узлы этой сетки рассматриваются в качестве потенциальных нечетких центров

$$\tilde{C}^{m(i)}=\{(x^{1(i)},\,\mu_{\tilde{A}_1}(x^{1(i)})),(x^{2(i)},\,\mu_{\tilde{A}_2}(x^{2(i)})),...,(x^{m(i)},\,\mu_{\tilde{A}_m}(x^{m(i)}))\},\ i=\overline{1,k},$$

и для каждого из них рассчитывается пиковая функция $M(S(\tilde{C}^{m(i)}))$, $i=\overline{1,k}$,

$$M(S(\tilde{C}^{m(i)}))=\sum_{j=1}^{p}\exp\left(-\left\|S(\tilde{C}^{m(i)}),S(\tilde{A}^{m(j)})\right\|^{2b}/2\sigma^{2}\right),$$

где $\left\|S(\tilde{C}^{m(i)}),S(\tilde{A}^{m(j)})\right\|$, $i=\overline{1,k}$, $j=\overline{1,p}$, — евклидово расстояние между парами векторов, определяющих центры $\tilde{C}^{m(i)}$, $i=\overline{1,k}$, и составные нечеткие числа

$\tilde{A}^{m(j)}$, $j = \overline{1, p}$, соответственно, коэффициент σ — константа, которая индивидуально подбирается для каждой конкретной задачи, а b — показатель степени обобщения функции Гаусса.

Величина функции $M(S(\tilde{C}^{m(i)}))$, $i = \overline{1, k}$, рассматривается как оценка высоты пиковой функции. Она пропорциональна количеству векторов $S(\tilde{A}^{m(j)})$, $j = \overline{1, q}$, находящихся в окрестности потенциального центра $S(\tilde{C}^{m(i)})$, $i = \overline{1, k}$. Малое значение $M(S(\tilde{C}^{m(i)}))$, $i = \overline{1, k}$, свидетельствует о том, что центр $S(\tilde{C}^{m(i)})$, $i = \overline{1, k}$, размещается в области, в которой размещено небольшое количество векторов $S(\tilde{A}^{m(j)})$, $j = \overline{1, q}$. Необходимо обратить внимание на то, что коэффициент σ имеет незначительное влияние на конечные пропорции между $M(S(\tilde{C}^{m(i)}))$, $i = \overline{1, k}$, для разных значений $S(\tilde{C}^{m(i)})$, $i = \overline{1, k}$, поэтому подбор его величины не является критичным.

После вычисления значений $M(S(\tilde{C}^{m(i)}))$, $i = \overline{1, k}$, для всех потенциальных центров отбираются первые k_1 точки, имеющие наибольшие значения $M(S(\tilde{C}_1^{m(i)}))$, $i = \overline{1, k_1}$. Для выбора следующих центров необходимо, во-первых, исключить k_1 центров и узлы, размещенные в непосредственной близости от них. Это можно сделать путем переопределения пиковой функции за счет отсекания от нее значений функции Гаусса с центрами в точках $S(\tilde{C}_1^{m(l)})$, $l = \overline{1, k_1}$. Если эту вновь определенную функцию обозначить как $M_{\text{new}}(S(\tilde{C}^{m(i)}))$, $i = \overline{1, k}$, то

$$M_{\text{new}}(S(\tilde{C}^{m(i)})) = M(S(\tilde{C}^{m(i)})) - M(S(\tilde{C}_1^{m(l)}))\exp(-\| S(\tilde{C}_1^{m(l)}), S(\tilde{C}^{m(i)}) \|^{2b} / 2\sigma^2).$$

Необходимо обратить внимание на то, что значения функция $M_{\text{new}}(S(\tilde{C}^{m(i)}))$, $i = \overline{1, k}$, в точках $S(\tilde{C}_1^{m(l)})$, $l = \overline{1, k_1}$, становятся равными нулю. Отсюда следует, что последовательное отсечение центров (с максимальным значением пиковой функции) позволяет выявлять и устранять следующие центры.

Процесс поиска следующих центров $S(\tilde{C}_2^{m(l)})$, $l = \overline{k_1 + 1, k_2}$, $S(\tilde{C}_3^{m(l)})$, $l = \overline{k_2 + 1, k_3}$, …, проводится последовательно на модифицированных значениях функции M_{new} $(S(\tilde{C}^{m(i)}))$, $i = \overline{1, k}$, получающихся при удалении близкого окружения центра, обнаруженного на предыдущем этапе. Он завершается в момент локализации всех центров, использующихся в модели. Метод пикового группирования эффективен, когда размерность вектора $S(\tilde{A}^{m(j)})$, $j = \overline{1, q}$, невысока. В противном случае (при большом количестве компонент $S(\tilde{A}^{m(j)})$, $j = 1, q$) количество потенциальных центров растет достаточно быстро, процесс расчетов пиковых функций становится длительным, а сама процедура — малоэффективной.

Метод разностного группирования данных [13] — это модификация алгоритма пикового группирования, в которой векторы $S(\tilde{A}^{m(j)})$, $j = \overline{1, q}$, рассматриваются в качестве k потенциальных центров, $k = q$. Пиковая функция $D(S(\tilde{A}^{m(i)}))$, $i = \overline{1, k}$, в этом алгоритме задается в виде

$$D(S(\tilde{A}^{m(i)})) = \sum_{j=1}^{p} \exp(-\| S(\tilde{A}^{m(i)}), S(\tilde{A}^{m(j)}) \|^{2b} / (r_a / 2)^2), \quad i = \overline{1, k}.$$

Значение коэффициента r_a определяет сферу соседства. На значения $D(S(\tilde{A}^{m(i)}))$, $i = \overline{1, k}$, существенно влияют только векторы $S(\tilde{A}^{m(j)})$, $j = \overline{1, q}$, размещенные в пределах этой сферы. При большой плотности точек в окрестности $S(\tilde{A}^{m(i)})$, $i = \overline{1, k}$ (потенциального центра) значения функции $D(S(\tilde{A}^{m(i)}))$,

$i = \overline{1,k}$, велико. И наоборот, ее малые значения свидетельствуют о том, что в окрестности $S(\tilde{A}^{m(i)})$, $i = \overline{1,k}$, незначительное количество данных. Такая точка считается «неудачным» кандидатом в центры. После вычисления значений пиковой функции для каждой точки $S(\tilde{A}^{m(j)})$, $j = \overline{1,q}$, выбирается вектор $S(\tilde{A}^{m})$, для которого мера плотности $D(S(\tilde{A}^{m}))$ оказалась наибольшей. Эта точка становится первым отобранным центром $S(\tilde{C}_1^{\,m})$. Выбор следующего центра возможен после исключения предыдущего центра и всех точек, лежащих в его окрестности. Как и в методе пикового группирования, пиковая функция переопределяется в виде

$$D_{\text{new}}((S(\tilde{A}^{m(i)})) = D(S(\tilde{A}^{m(i)})) -$$

$$- D(S(\tilde{C}_1^{\,m}))\exp(-\| S(\tilde{A}^{m(i)}), S(\tilde{C}_1^{\,m})\|^{2b} / (r_a/2)^2),\ i = \overline{1,k}.$$

При переопределении функции $D(S(\tilde{A}^{m}))$ коэффициент r_b задает новое значение константы, которая очерчивает сферу соседства для следующего центра. Обычно накладывается условие $r_b \geq r_a$. Пиковая функция $D_{\text{new}}(S(\tilde{A}^{m(i)}))$, $i = \overline{1,k}$, принимает нулевое значение при $S(\tilde{A}^{m(i)}) = S(\tilde{C}_1^{\,m})$, $i = \overline{1,k}$, и близка к нулю в ближайшей окрестности этого центра.

После модификаций значений пиковой функции определяется следующая точка — $S(\tilde{A}^{m})$, для которой величина $D_{\text{new}}(S(\tilde{A}^{m}))$ максимальна. Эта точка становиться следующим центром $S(\tilde{C}_2^{\,m})$. Процесс поиска следующего центра повторяется после удаления компонентов, соответствующих построенным центрам.

Каждый из методов имеет свои собственные особенности и уточнения в зависимости от условий использования и его структурно-концептуальной реализации. Главным фактором выступает показатель наиболее оптимального метода, что на практике существенно зависит от условий конкретной задачи. Результаты практического использования методов кластеризации на заданной сетке двухэлементных составных нечетких чисел позволили сделать ряд выводов.

Если в методе *C*-means необходимо выбирать начальные коэффициенты, показывающие степень принадлежности к кластеру, исключительно случайно, то в алгоритме пиковой группировки главную роль играет покрытие плоскости равномерной сеткой с размерами ячеек, зависящими от максимального и минимального значений двухмерных данных покоординатно. Узлы этой сетки будут выступать в качестве претендентов в центры будущих кластеров и оказывается, что выбрать размеры ячеек этой сетки — задача, зависящая от конкретной ситуации. Данную задачу можно решать по правилу Рунге, которое заключается в нахождении такого разбиения области, при котором величина $\Delta_{2n} \approx \Theta | I_{2n} - I_n |$ (I_n — результат расчетов, полученных при использовании n узлов по каждой координате сетки), достигает своей наперед заданной точности. В вычислительных экспериментах для нахождения результатов полагалось $\Theta = 1/3$. Точность правила Рунге задается пользователем в зависимости от условий конкретной прикладной области.

Параметры *b* и σ, использующиеся в методе, также подбираются индивидуально для каждой задачи. При вычислении экспоненты для оценки точек сгущения вокруг кандидата для некоторых претендентов (из-за ограниченности размера представления чисел в машинной интерпретации) возникает проблема оценивания нулевого значения выражения. Поэтому перед тем, как подставлять полученные в преобразованиях значения в показатель экспоненты, их желательно разделить на некоторую константу (в рассматриваемых вычислительных примерах использовалось значение 1000). Кро-

ме этого, при реализации возникает проблема учета близости узла сетки — кандидата в центры кластера, что делает значение суммы экспоненциальных выражений очень большим, даже без учета значений этой функции для других точек. Поэтому считается необходимым ввести ограничения на степень близости возможного кандидата в центры кластера к другим точкам. Как вариант, предлагается определять минимально допустимое расстояние по правилу Рунге.

Алгоритм разностного группирования представляет собой упрощенную, но более эффективную модификацию алгоритма пикового группирования. Все замечания, сделанные для метода пикового группирования, в этой случае несущественны. Единственным нерешенным вопросом остается значение коэффициента r_a, определяющего сферу соседства. При практическом использовании это значение может также определяться по правилу Рунге с заданной точностью. Главный плюс алгоритма — то, что в качестве кандидатов в центры кластеров выступают сами входные данные.

Для анализа полученных результатов исследования проведены вычислительные эксперименты. Работа алгоритмов тестировалась на примере задачи группирования состояний нечеткой системы, функционирующей в двухмерном пространстве. Начальное размещение состояний системы проводилось случайным образом, количество точек определялось пользователем. Результаты процедуры кластеризации множества состояний нечеткой системы, состоящего из 100 точек (рис. 1, больший размер точки соответствует большей величине меры принадлежности элемента нечеткому множеству), приведены на рис. 2 (*а* — методом *C*-means (крестиком обозначены центры кластеров), $k = 7$, $\gamma \geq 0{,}3$; *б*, *в* — методом разностного группирования, $\gamma \geq 0{,}2$ и $\gamma \geq 0{,}3$; *г* — методом пикового группирования, $\gamma \geq 0{,}6$).

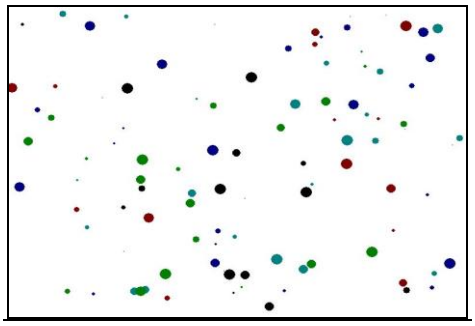
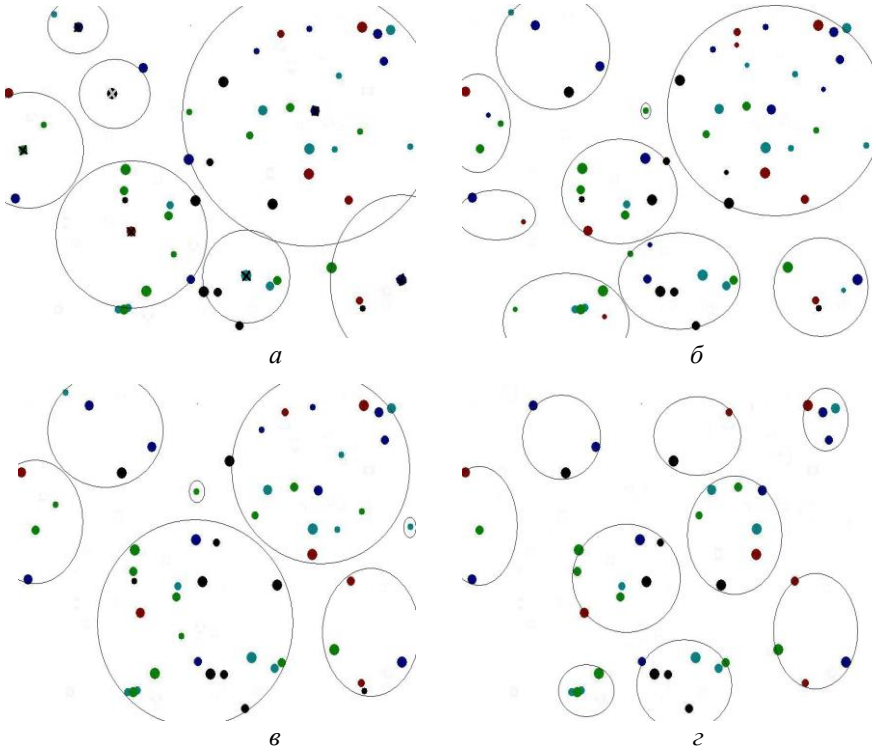


Рис. 1



а *б* *в* *г*

Рис. 2

Заключение

В данной статье предложен новый способ формализации нечеткости в виде составных нечетких чисел. Рассмотрена аксиоматика такого представления, введено понятие множеств скалярного и векторного уровней, сформулированы процедуры для вычисления расстояний между данными. На основе разработанных моделей и методов рассмотрены и решены задачи группирования состояний нечеткой системы, описанных совокупностью составных нечетких чисел, на основе множеств скалярного и векторного уровня. Основная идея разработанных алгоритмов кластеризации состоит в вычислении и использовании значений расстояния между элементами соответствующих множеств заданного скалярного или векторного уровня с учетом величины погрешности, которая определяется на прямоугольной сетке, накрывающей множество исходных данных. Предложенные алгоритмы позволяют формализовать поиск кластерных центров совокупности составных нечетких чисел и реализовать процедуры группирования данных в пределах предварительно заданного или автоматически сгенерированного по ходу алгоритма количества кластеров. Рассмотрены условия наиболее оптимального использования алгоритмов кластеризации.

Предложенный подход является модификацией существующих методов кластеризации, адаптированных к обработке нечетких данных специального вида. Приведены примеры использования данного подхода при решении практических задач, проанализированы результаты численных экспериментов.

Є.В. Івохін, Д.В. Апанасенко

КЛАСТЕРИЗАЦІЯ СУКУПНОСТІ СКЛАДЕНИХ НЕЧІТКИХ ЧИСЕЛ НА ОСНОВІ МНОЖИН СКАЛЯРНОГО ТА ВЕКТОРНОГО РІВНІВ

Запропоновано новий спосіб формалізації нечіткості у вигляді складових нечітких чисел. Розглянуто аксіоматику такого представлення, введено поняття множин скалярного і векторного рівнів, сформульовано процедури для розрахунку відстаней між даними. На основі розроблених моделей і методів розглянуто системи, що складаються з сукупності нечітких чисел, на основі множин скалярного і векторного рівня. Основна ідея розроблених алгоритмів кластеризації складається з наступних варіантів: обчислення і векторний рівень з урахуванням даних похибки, яка визначається на прямокутній сітці, що накриває множину вихідних даних. Запропоновані алгоритми дозволяють формалізувати пошук кластерних центрів сукупності складових нечітких чисел і реалізувати процедури групування даних в межах заздалегідь визначеної або автоматично згенерованої по ходу алгоритму кількості кластерів. Розглянуто умови оптимального використання алгоритмів кластеризації. Запропонований підхід є модифікацією всіх методів кластеризації, адаптованих до обробки нечітких даних спеціального виду. Наведено приклади використання даного підходу при вирішенні практичних завдань, проаналізовано результати числових експериментів.

E.V. Ivokhin, D.V. Apanasenko

CLUSTERING OF COMPOSITE FUZZY NUMBERS AGGREGATE BASED ON SETS OF SCALAR AND VECTOR LEVELS

A new method for formalizing fuzziness in the form of composite fuzzy numbers is proposed. The axiomatics of such representation is considered, the notion of sets of scalar and vector levels is introduced, procedures for calculating distances between data are formulated. On the basis of the developed models and methods, the problems of grouping the states of a fuzzy system described by a set of composite fuzzy numbers on the basis of sets of scalar and vector levels are considered and solved. The main idea of the developed clustering algorithms is to calculate and use the distance values between the elements of the corresponding sets of a given scalar or vector level, taking into account the error value, which is determined on a rectangular grid covering a set of initial data. The proposed algorithms allow us to formalize the search for cluster centers of a set of composite fuzzy numbers and implement the procedures for grouping data within a predetermined or automatically generated number of clusters during the course of the algorithm. Conditions for constructive use of clustering algorithms are considered. The proposed approach is a modification of existing clustering methods, adapted to processing of fuzzy data of a special kind. Examples of the use of this approach in solving practical problems are given, and the results of numerical experiments are analyzed.

1. *Брайчевский С.М., Ландэ Д.В.* Современные информационные потоки: актуальная проблематика // Научно-техническая информация. — 2005. — Сер. 1. — Вып. 11. — С. 21–33.
2. *Івохін Є.В., Апанасенко Д.В.* Про кластеризацію складених нечітких чисел на основі множин векторного рівня // Вісник КНУ імені Тараса Шевченка. Сер. ФМН. — 2016. — № 2. — С. 90–93.
3. *Заде Л.А.* Понятие лингвистической переменной и его применение к принятию приближенных решений. — М. : Мир, 1976. — 175 с.
4. *Кофман А.* Введение в теорию нечетких множеств. — М. : Радио и связь, 1982. — 432 с.
5. *Nahmias S.* Fuzzy variables // Fuzzy Sets and Systems. — 1978. — N 1. — P. 97–110.
6. *Лю Б.* Теория и практика неопределенного программирования. — М. : БИНОМ. Лаборатория знаний, 2005. — 416 с.
7. *Wong, M.L., Leung K.S.* Data mining using grammar based genetic programming and applications. —N.Y. : Kluwer Academics Publ., 2000. — 213 p.
8. *Волошин О.Ф., Мащенко С.О.* Моделі та методи прийняття рішень. — Київ : ВПЦ «Київський університет», 2010. — 336 с.
9. *Орловский С.А.* Проблемы принятия решения при нечеткой исходной информации. — М. : Наука, 1981. — 206 с.
10. *Jamba M.* Hierarchical cluster analysis and compliance. — M. : Finance and Statistics, 1988. — 345 с.
11. *Durand B., Odell P.* Cluster analysis. — M. : Statistics, 1977. — 128 с.
12. *Вятченин Д.А.* Нечеткие методы автоматической классификации. — Минск : Технопринт, 2004. — 219 с.
13. *Осовский С.* Нейронные сети для обработки информации. — М. : Финансы и статистика, 2002. — 344 с.
14. *Jang J.S., Sun C.T., Mizutani E.* Neuro-fuzzy and soft computing. — N.Y. : Prentice Hall, 1997. — 614 p.

Получено 14.03.2018