УДК 519.5:517.1:

Ю. Н. Минаев, д-р техн. наук
Национальный авиационный университет
(Украина, 03057, Киев, пр-т космонавта Комарова, 1,
тел. (044) 4067752, e-mail: minaev@rambler.ru),
О. Ю. Филимонова, канд. техн. наук,
Ю. И. Минаева, аспирантка
Киевский национальный университет строительства и архитектуры
(Украина, 03037, Киев, Воздухофлотский пр-т, 31,
тел. (044), 2486427, e-mail: filimonova @nm.ru;
(044) 2425462, e-mail: jumin @big-mir.net)

Иерархическая кластеризация нечетких данных

Рассмотрена кластеризация данных (построение бинарных деревьев-дендрограмм), представленных в виде нечетких переменных, моделируемых тензорами. Закодированная бинарным алфавитом дендрограмма представляет собой 2-адическое число, которое может быть использовано как ее характеристика. Сравнение иерархических кластеризаций нечетких данных и их дефадзификаций, выполненное на уровне 2-адических деревьев, позволяет сделать вывод о наличии (отсутствии) структурной близости объектов.

Розглянуто кластеризацію даних (побудову бінарних дерев-дендрограм), представлених у вигляді нечітких змінних, котрі моделюються тензорами. Закодована бінарним алфавітом дендрограма являє собою 2-адичне число, яке може бути використано як її характеристика. Порівняння ієрархічних кластеризацій нечітких даних та іхніх дефадзифікацій, виконане на рівні 2-адичних дерев, дозволяє зробити висновок про наявність (відсутність) структурної близькості об′єктів.

К л ю ч е в ы е с л о в а : нечеткая переменная, дендрограмма, кластер, тензор, 2-адическое дерево.

Проблема иерархической кластеризации (ИК) в последнее время приобрела особую актуальность в связи с извлечением знаний из данных. При этом цели кластеризации могут быть различными в зависимости от конкретной прикладной задачи [1]. Существует большой класс алгоритмов ИК, которые обеспечивают реализацию задачи обучения без учителя. В общем случае алгоритм кластеризации — это функция $f : X \to Y$, которая любому объекту $x \in X$ ставит в соответствие метку кластера $y \in Y$. Задача определения оптимального числа кластеров относительно принятого критерия качества кластеризации до настоящего времени является актуальной и сложной.

В работах [2—4] показано, что решение задачи кластеризации принципиально неоднозначно. Алгоритмы ИК в подавляющем большинстве случаев работают по так называемому агломеративному (объединительному) принципу, согласно которому на первом этапе отдельный объект считается отдельным кластером, что позволяет для одноэлементных кластеров определить функцию расстояния естественным образом: $R(\{x\}, \{x'\}) = \rho(x, x')$. Последующие шаги реализуют процесс слияний: на каждой итерации вместо пары самых близких (далеких) кластеров U и V образуется новый кластер $W = U \bigcup V$. Расстояние от нового кластера W до любого другого кластера S вычисляется на основании ранее определенных расстояний $R(U, V)$, $R(U, S)$ и $R(V, S)$ [3].

Наиболее распространенные методы определения расстояний между объектами — метод единственной связи (ЕС), $d_{AB} = \min\limits_{i \in A,\ j \in B} (d_{ij})$, и метод полной связи (ПС), $d_{AB} = \max\limits_{i \in A,\ j \in B} (d_{ij})$, где d — функция расстояния. Заметим, что расстояние d_{AB} обладает свойствами ультраметрического расстояния, т.е. все расстояния внутри кластеров меньше расстояний между ними.

Основная цель кластерного анализа (КА) [2—4] состоит в выделении групп однородных подмножеств в исходных многомерных данных. Объекты внутри этих групп схожи (в известном смысле) между собой, а объекты из разных групп — не похожи. Под похожестью понимается близость объектов в многомерном пространстве признаков. Кластерный анализ совместно с визуализацией данных в виде бинарного дерева составляет основу структурного подхода к представлению и анализу данных и обеспечивает, главным образом, получение новых знаний о природе представленного множества данных, в частности, на основании их структуры. При этом возникает новая задача — отыскать в исходном множестве объектов такое подмножество (существенно меньшей размерности), которое структурно наиболее близко к исходному, или определить, насколько соответствует представленное подмножество объектов (меньшей размерности) структуре исходного множества, хотя понятие наибольшая близость может иметь неоднозначную трактовку.

Современное состояние проблемы. В современных условиях сфера применения КА расширилась, так как бинарные деревья (дендрограммы) — результат иерархического КА (ИКА) — стали рассматривать на уровне p-адических (2-адических) чисел с применением ультраметрики. В настоящее время ИКА применяется для решения задач анализа нечетких данных (НД), которые могут быть представлены в виде нечетких множеств (НМ), интервала или некоторой совокупности данных (числовая последовательность). Проблеме решения задач КА посвящено большое число работ,

однако основная проблема — определение кластеризации, которая в наибольшей степени удовлетворяет естественной структуре, до настоящего времени не решена.

Если учесть, что даже в случае четких данных результаты кластеризации одной и той же выборки при использовании различных методов и алгоритмов могут иметь не просто существенные различия, но принципиально менять информацию об объекте, то можно предположить, какие трудности ожидают при кластеризации НД. При выборе метода кластеризации НД необходимо учитывать, что результат кластеризации существенно зависит от следующих факторов: выбора системы признаков; определения меры близости объектов; определения способа формализации представлений об эквивалентности объектов, составляющих отдельный кластер, и др.

В работе [2] показано, что недостатки алгоритмов кластеризации, приводящие к существенному несовпадению результатов или несоответствию их объективно существующим структурам, могут быть преодолены посредством формирования на основании полученных отдельных решений нового решения, наиболее согласованного с полученными ранее, и анализа оценок этого согласованного решения и отдельных решений.

Для описания различных задач кластеризации (классификации) [4] рациональным является использование теории бинарных отношений, эффективной в условиях неполной информации. Каждому отношению сопоставляют квадратную матрицу «объект-объект», элементы которой принимают значения $r_{ij} \in [0,1]$. Предложено множество способов измерения близости между отношениями (между отражающими их матрицами), при этом часто предлагаемые величины определяются аксиоматически.

Известно, что расстояние (метрика) между объектами в пространстве параметров d_{ab} есть величина, удовлетворяющая следующим аксиомам:

A_1. $d_{ab} > 0,\ d_{aa} = 0$;

A_2. $d_{ab} = d_{ba}$;

A_3. $d_{ab} + d_{bc} \geq d_{ac}$ (неравенство треугольника).

Мерой близости (сходства) называется величина ν_{ab}, имеющая предел и возрастающая с возрастанием близости объектов. Одно из определений близости основано на системе аксиом [5]:

B_1. v_{ab} — непрерывна, т.е. малому изменению положения точек в пространстве соответствует малое изменение меры;

B_2. $v_{ab} = v_{ba}$;

B_3. $0 \leq v_{ab} \leq 1,\ v_{ab} = 1 \leftrightarrow a = b.$

Условие A_3 не является конструктивным. Оно почти никогда не учитывается в расчетах, так как пригодны измерители близости, не удовлет-

воряющие неравенству треугольника, в частности, в так называемом ультраметрическом пространстве справедливо усиленное неравенство треугольника $\max(d_{ab}, d_{bc}) \geq d_{ac}$. Существует возможность перехода от расстояний к мерам близости, в частности, посредством преобразования $v = 1/(1+d)$.

Приведенные выше данные связаны с возможностью конструирования такого показателя близости между объектами, который не зависел бы от способа измерения переменных. Применение такого показателя будет давать одинаковые результаты при любых допустимых преобразованиях. В результате исследований [4] получены неконструктивные выводы относительно теоретической и практической ценности различных метрик: результаты работы алгоритмов классификации могут непредсказуемо меняться в зависимости от выбора способа измерения показателей. Конкретные особенности мер близости и рекомендации по их применению для четких данных рассмотрены в [4].

Аппроксимационный подход в КА. Все функционалы качества классификации ориентированы на решение следующей задачи: дать в явном виде представление о хорошей в целом классификации. В работе [4] показано, что таких представлений существует достаточно много, но имеются некоторые соображения универсального характера, учитывающие самые общие черты классификации. Если обозначить искомое отношение произвольного типа через Y, исходные данные через X, а оператор перехода от X и Y через P (не конкретизируя вид этих конструкций), то получим функционал, позволяющий максимально приблизить результирующее отношение к имеющимся данным: $\|Y - XP\| \to \min$, где $\| \|$ — какая-либо норма. Задачи такого типа — аппроксимация «плохо устроенного» множества X «хорошо устроенной структурой» Y — известны в математике и имеют множество приложений [5].

Базовые нотации классической кластеризации применительно к условиям неопределенности, моделируемой на уровне теории НМ (ТНМ), имеют особенности [6]. В частности, d принимают равным любой метрике в $\mathbf{R}^m$. Понятие расстояния составляет объект многих исследований. В ТНМ используются два определения расстояния [7] для НМ $\widetilde{A}$, $\widetilde{B}$ с функциями принадлежности (ФП) $\mu_{\widetilde{A}}(x_i)$, $\mu_{\widetilde{B}}(x_i) \in [0,1]$:

1) обобщенное расстояние Хеммимнга (линейное) —

$$d(\widetilde{A}, \widetilde{B}) = \sum_{i=1}^{n} \left| \mu_{\widetilde{A}}(x_i) - \mu_{\widetilde{B}}(x_i) \right|;$$

2) евклидово (квадратичное) расстояние —

$$e(\widetilde{A}, \widetilde{B}) = \left(\sum_{i=1}^{n} (\mu_{\widetilde{A}}(x_i) - \mu_{\widetilde{B}}(x_i))^2 \right)^{1/2}.$$

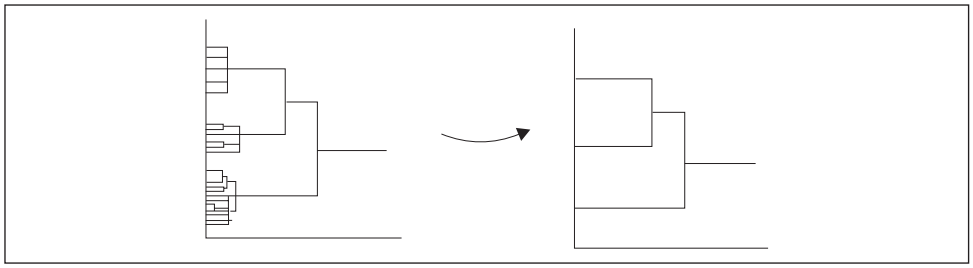


Рис. 1. Сходимость дендрограмм

При этом $d(\widetilde{A}, \widetilde{B})$ и $e(\widetilde{A}, \widetilde{B})$ — четкие величины, а НМ $\widetilde{A}, \widetilde{B}$ заданы на одном универсальном множестве (УМ).

В работе [8] упомянуто о том, что желательные свойства алгоритмов кластеризации исходят от практиков, у которых понятие «хорошая кластеризация» основано на визуальной оценке и является интуитивным. Однако это нельзя признать удовлетворительным, так как необходимо теоретическое понимание того, что должно быть разработано. Доказана теорема [8] об эквивалентности между ультраметрикой и дендрограммой, в которой дендрограмма представлена как ультраметрическое пространство. Любой метод ИК может быть рассмотрен как отображение финитного метрического пространства (МП) в финитное ультраметрическое пространство. В соответствии с [9, 10], если (X, d_X) и (Y, d_Y) — два финитных МП с метриками d_X и d_Y, а (X, u_X) и (Y, u_Y) — два финитных метрических ультраметрических пространства с метриками u_X и u_Y, соответствующими выходам, порожденным ИК по методу ЕС, то $d_{GH}((X, u_X)(Y, u_Y)) \leq$ $\leq d_{GH}((X, u_X)(Y, d_Y))$, где $d_{GH}(\,)$ — расстояние Громова— Хаусдорфа.

В работах [8, 9] рассмотрена сходимость дендрограмм (рис. 1). Это понятие формализовано посредством эквивалентного представления дендрограмм как ультраметрики и последующим определением *GH*-расстояния, которое является метрикой на ультраметрических пространствах, одинаково эффективно работающей в различных кластеризационных конструкциях. Следует заметить, что сходимость дендрограмм означает существование эквивалентности между разделением n объектов $\{O_1, ..., O_n)$ на n и m кластеров, при этом $n >> m$.

Неустойчивость метода кластеризации ПС (Complete Linkage (*CL*)) по отношению к малым возмущениям в метрике показана на рис. 2, где представлены два подобных МП. Для каждого из них справа приведена *CL*-дендрограмма. Несмотря на то, что ε достаточно мало ($\varepsilon > 0$), выходы дендрограммы — различны. Теорема 28, доказанная в [9], объясняет суть сходимости дендрограмм.

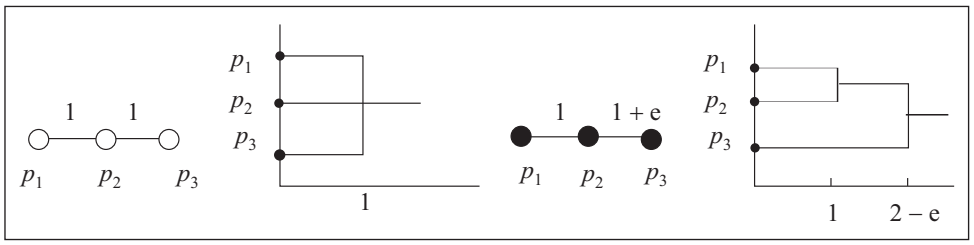


Рис. 2. Пример неустойчивости *CL*-кластеризации по отношению к возмущениям в метрике [8]

Устойчивость — важное свойство методов кластеризации, в частности НД можно рассматривать как результат влияния возмущений. В результате исследований установлено, что методы средней и полной связи кластеризации неустойчивы (в метрическом смысле), а метод ЕС отличается определенной стабильностью.

Для описания различных задач кластеризации целесообразно использование теории бинарных отношений. В работе [11] предложен алгоритм кластеризации, позволяющий выполнить разделение деревьев с различными уровнями и кластерами согласно различным композициям t_i-норм. Получено три оценки выполнения структуры с использованием композиций t-норм. Методы, используемые в кластерном анализе, основанном на нечетких отношениях (НО), можно разделить на две категории:

1 — использование объектной функции, определенной как расстояние;

2 — использование НО, которые более просты в использовании, вследствие чего требуется матрица отношений множества данных (МД). Величины нечетких переменных (НП) можно использовать для представления сходства между двумя объектами.

Четкое бинарное отношение R между двумя множествами, X и Y, определено как подмножество $X \times Y$. Это отношение $R(X, Y)$ связывается с функцией-индикатором $u_R(x, y)$, принадлежащей $\{0, 1\}$ для всех (x, y) в $X \times Y$; $u_R(x, y) = 1$, если $(x, y) \in R(X, Y)$, и $u_R(x, y) = 0$, если $(x, y) \notin R(X, Y)$. В [11] НО R между X и Y определено как НП $X \times Y$, а t-норма — как общая форма нечеткого пересечения, где она может быть использована в качестве композиции любых двух НО. Для создания иерархических структур представления оценки двух НО необходимо определить $^{\max}t$-композицию, например $^{\max}t$-композиция для НО R_1 и R_2 определена в виде

$$\mu_{R_1 \circ R_2}(x, y) = \max_{y \in X} \{t(\mu_{R_1}(x, y), \mu_{R_2}(y, z))\}, \ \forall x, z \in X,$$

где $R_1 \circ R_2$ — композиция двух отношений, R_1 и R_2, на $X \times X$. Различные композиции t_i-норм имеют следующий вид:

$$t_\omega(x,y) = \begin{cases} \min\{x,y\}, \text{ если } \max\{x,y\} = 1, \\ 0, \qquad\quad \text{если иначе;} \end{cases}$$

$$t_1(x,y) = \max\{0, x+y-1\};\ \ t_{1.5}(x,y) = xy/(2-(x+y-xy));$$

$$t_2(x,y) = xy;\ \ t_{2.5}(x,y) = xy/(x+y-xy);$$

$$t_3(x,y) = \min\{x,y\}.$$

Схемы вывода (резолюционные) для нечетких пересечений различны, порядок приведенных t-норм может быть таким: $t_\omega \le t_1 \le t_{1.5} \le t_2 \le t_{2.5} \le t_3$. Различные $^{\max}t$-композиции с различными t-нормами, очевидно, порождают различные результаты нечетких композиций. В работе [11] показано, как матрица НО R может быть декомпозирована в резолюционную форму с использованием α-срезов $(0 \le \alpha < 1)$. Нечеткое отношение R на $X \times Y$ может быть представлено в виде

$$R = \bigcup_\alpha \alpha R_\alpha = \alpha_1 R_{\alpha_1} + \alpha_2 R_{\alpha_2} + \ldots + \alpha_m R_{\alpha_m},\ 0 \le \alpha_1 \le \alpha_2 \le \ldots \le \alpha_m \le 1.$$

Используя декомпозированную резолюционную форму, соответствующую структуру можно получить из матрицы НО R. Например, матрица отношения подобия $^{\max}t_3$ из работы [11] может быть представлена в следующей резолюционной форме:

$$R = \begin{pmatrix} 1 & & & & \\ 0.7 & 1 & & & \\ 0.5 & 0.5 & 1 & & \\ 0.5 & 0.5 & 0.9 & 1 & \\ 0.5 & 0.5 & 0.5 & 0.5 & 1 \end{pmatrix} = 0.1 \begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \\ 0 & 0 & 1 & & \\ 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} + 0.9 \begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \\ 0 & 0 & 1 & & \\ 0 & 0 & 0 & 1 & \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} +$$

$$+ 0.7 \begin{pmatrix} 1 & & & & \\ 1 & 1 & & & \\ 0 & 0 & 1 & & \\ 0 & 0 & 1 & 1 & \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} + 0.5 \begin{pmatrix} 1 & & & & \\ 1 & 1 & & & \\ 1 & 1 & 1 & & \\ 1 & 1 & 1 & 1 & \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix},$$

где [ ] — матрица четкого отношения, ближайшего к нечеткому.

Отдельную группу алгоритмов ИК НД составляют алгоритмы, в которых использовано расстояние между НД (как правило, НП), основанное на методологии геометрических визуальных представлений данных, так на-

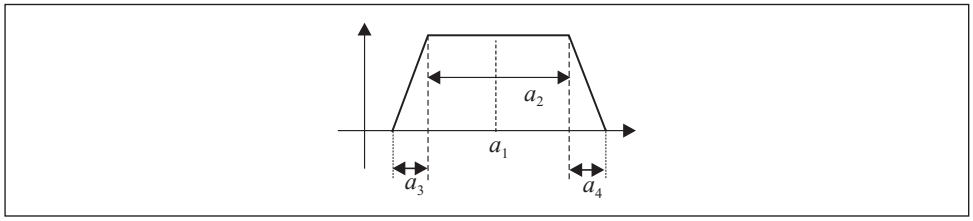


Рис. 3. Параметризация трапециевидного НД

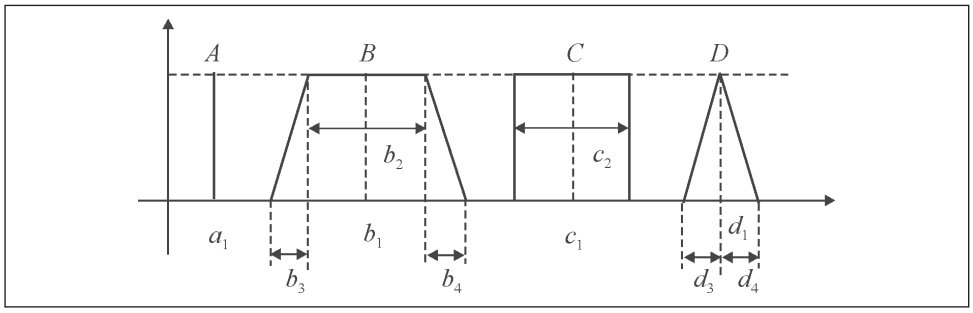


Рис. 4. Типы параметризованных НД

зываемая параметризация, при этом выполнение всех аксиом расстояния не гарантируется. В работе [12] предложен алгоритм, практически обобщающий подобную методологию решения задач кластеризации НД.

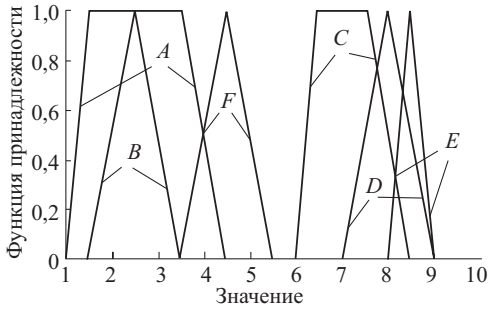
Симметричные трапециевидные НП (числа) представлены на рис. 3, 4 [12]. Параметризация трапециевидного нечеткого числа (НЧ) $\widetilde{A}$ имеет вид $\widetilde{A} = m(a_1, a_2, a_3, a_4)$, где a_1, a_2, a_3 и a_4 — центр, внутренний диаметр, левый и правый внешние радиусы. Из рис. 4 видно, что четыре типа НД можно представить единообразно. Согласно представлению:

$$\widetilde{A} = [a_1, 0, 0, 0],\ \widetilde{B} = [b_1, b_2, b_3, b_4],\ \widetilde{C} = [c_1, c_2, 0, 0],\ \widetilde{D} = [d_1, 0, d_3, d_4).$$

Пусть $\widetilde{A} = m(a_1, a_2, a_3, a_4)$, и $\widetilde{B} = m(b_1, b_2, b_3, b_4)$ — два НД. Различие между ними определяется как $d_h^2(\widetilde{A}, \widetilde{B}) = (a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2 + (a_4 - b_4)^2$.

На основании введенной меры различия в [12] предложен алгоритм кластеризации НД, приведенных на рис. 5, где НД переформулированы в соответствии с предложенной метрикой:

$$\widetilde{A} = [2.5;\ 2;\ 0.5;\ 1],\ \widetilde{B} = [2.5;\ 0;\ 1;\ 1],\ \widetilde{C} = [7;\ 1;\ 0.5;\ 1],$$
$$\widetilde{D} = [8;\ 0;\ 1;\ 1],\ \widetilde{E} = [8.5;\ 0;\ 0.5;\ 0.5],\ \widetilde{F} = [4.5;\ 0;\ 1;\ 1]\ .$$



$X = [1: 0.1: 10]$;
A = trapmf (X, [1 1.5 3.5 4.5]);
B = trimf (X, [1.5 2.5 3.5]);
C = trapmf (X, [6 6.5 7.5 8.5]);
D = trimf (X, [7 8 9]);
E = trimf (X, [8 8.5 9]);
F = trimf (X, [3.5 4.5 5.5]).

Рис. 5. Тестовое множество НП в виде $\tilde{x} = \{x / \mu\}$

Дендрограмма для объектов, полученная в [12] на основании матрицы расстояний между НД ($d_f(A, B) = 1{,}89$, $d_f(A, C) = 9{,}05$),

	A	*B*	*C*	*D*	*E*	*F*
A	0	1.89	9.05	11.4	12.06	4.31
B		0	9.17	11.00	12.00	4.00
C			0	2.08	3.09	5.20
D				0	1.06	7.0
E					0	8.01
F						0

приведена на рис. 6, *а*, на рис. 6, *б — г*, приведены дендрограммы, полученные авторами данной работы различными методами. Из рис. 6 видны особенности объектов кластеризации. В частности, влияние нечеткости на структуру данных проявляется в том, что дендрограммы для НД при использовании метода ЕС и фадзифицированных данных (см рис. 6, *а—в*) и при использовании метода ПС (см. рис. 6, *б—г*) с точки зрения ультраметрики имеют различную структуру (число уровней отличается в два раза). Сложно дать физическую интерпретацию этому факту, но несомненно, это — новое знание. В работе [12] дендрограмма названа нечеткой, однако она таковой не является, так как построена на основании вычисления четких расстояний между НД.

В работе [13] показано существование эквивалентности между иерархической кластеризацией, транзитивными нечеткими max-min-отношениями и ультраметрическим расстоянием.

Постановки основных задач и алгоритмы их решения. В работе [14] предложены тензорные модели неопределенности. Известно [7], что если $E = \{x\}$ — УМ, $x \in E$, то нечеткое подмножество $\tilde{A}$ множества E определяется как совокупность упорядоченных пар $\{(x, \mu_{\tilde{A}}(x))\}$, $\forall x \in E$, где

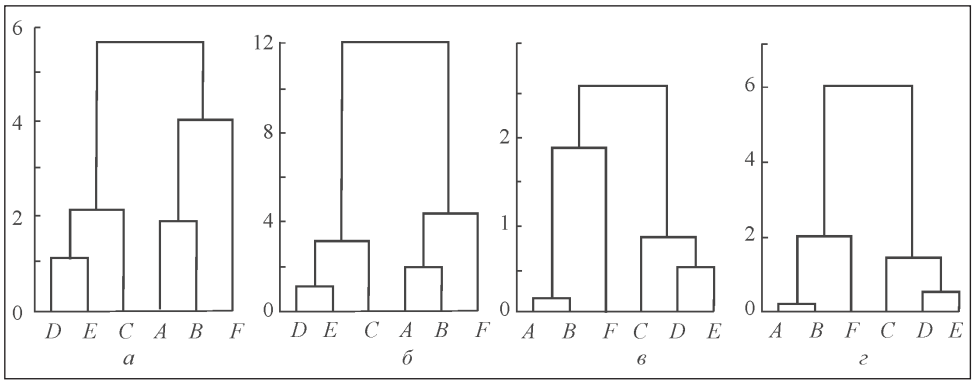


Рис. 6. Дендрограммы для НП, полученные в работе [12] (*а*), и авторами данной работы (*б* — *г*): *а* — метод ЕС; *б* — метод ПС; для дефадзифицированных НП: *в* — метод ЕС; *г* — метод ПС

$\mu_{\tilde{A}}(x)$ — ФП, $\mu_{\tilde{A}} \to [0,1]$. Введенная в работе [14] тензор-переменная (ТП), аналог НП, определена как $T_x = x \otimes \mu_x$, где $\otimes$ — операция тензорного (Кронекерова) произведения; T_x имеет матрицу размером $n \times n$.

В работе [15] предложено рассматривать многомерный массив как тензор. В соответствии с этим определением в тензорном базисе возможно представление НП — совокупности упорядоченных пар — как многомерного массива t_x с матрицей размером $2 \times n$. На рис. 7 представлена НП с выпуклой ФП и ее тензорные аналоги. С учетом тензорного представления исходных данных основные задачи в [15] сформулированы так:

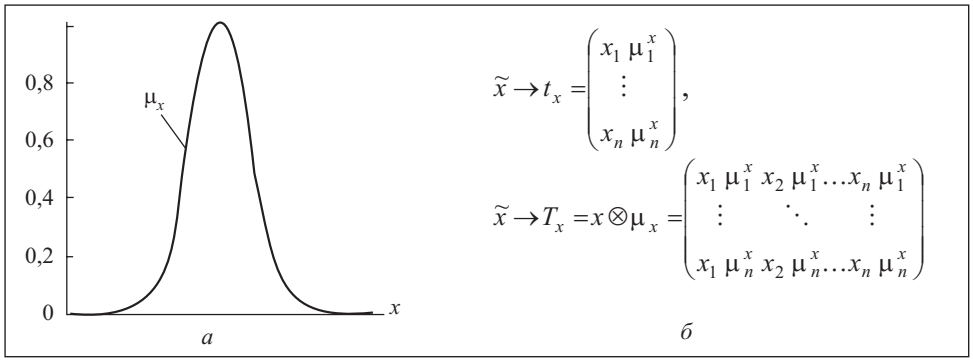
1) исследовать ИК НД, представив входные данные тензорными моделями, показать проявление влияния нечеткости на выход ИК — бинарное дерево;

2) исследовать ультраметрические свойства дендрограмм, представив входные данные тензорными моделями, определить ультраметрическую матрицу, 2-адические характеристики дендрограмм, показать возможность сравнения дендрограмм на основании вычисления 2-адической оценки.

Сопоставимость экспериментов. Нечеткие переменные в Matlab заданы в виде процедур, определяющих переменную на УМ E, например НП с треугольной ФП — $\mu^x = \text{trimf}(E,[P_1\ P_2\ P_3])$, где $P_1 < P_2 < P_3$ — параметры; НП с трапециевидной ФП — $\mu^x = \text{trapmf}(E,[P_1\ P_2\ P_3\ P_4])$, где $P_1 < P_2 < P_3 < P_4$ — параметры; НП с Гауссовой ФП — $\mu^x = \text{gaussmf}(E,[P_1\ P_2])$, где P_1 — дисперсия, P_2 — среднее.

Операции над НП выполняются в предположении, что НП представлены в виде НМ на одном универсальном множестве:

$$\underbrace{\{a_j / \mu^{a_j}\}}_{\tilde{A} \in E} *_f \underbrace{\{b_j / \mu^{b_j}\}}_{\tilde{B} \in E} \to (a_j *_f b_j) / \max(\min(\mu^{a_j}, \mu^{b_j})), (\forall j).$$



$$\widetilde{x} \to t_x = \begin{pmatrix} x_1\ \mu_1^x \\ \vdots \\ x_n\ \mu_n^x \end{pmatrix},$$

$$\widetilde{x} \to T_x = x \otimes \mu_x = \begin{pmatrix} x_1\ \mu_1^x\ x_2\ \mu_1^x \ldots x_n\ \mu_1^x \\ \vdots \qquad \ddots \qquad \vdots \\ x_1\ \mu_n^x\ x_2\ \mu_n^x \ldots x_n\ \mu_n^x \end{pmatrix}$$

Рис. 7. Моделирование НП: *а* — НП $\widetilde{x} = \{x / \mu_x\}$; *б* — t_x и T_x — первая и вторая формы тензорных аналогов НП

Следует заметить, что тензорный способ представления НМ не искажает первоначального смысла утверждений. В частности, если одно понятие описывается в ТНМ несколькими определениями, т.е. без потери общности представляется НМ с различными ФП, то это свойство сохранится и при моделировании НП в тензорном базисе. Выполненная проверка показала, что расстояния между утверждениями «примерно 5» (НП с различными ФП) в стандартном и тензорном представлении существенно меньше, чем между утверждениями «примерно 5» и «примерно 6», что свидетельствует об адекватности моделирования НП тензорными моделями. В простейшем случае для стандартной треугольной ФП получаем тензорные модели:

$$T_x = \begin{pmatrix} 0 & 0 & 0 \\ x_1 & x_2 & x_3 \\ 0 & 0 & 0 \end{pmatrix},\ t_x = \begin{pmatrix} x_1 & 0 \\ x_2 & 1 \\ x_3 & 0 \end{pmatrix}.$$

Объекты T_x и t_x имеют различные свойства, однако относительно КА они адекватны и, кроме того, имеют практически одинаковые нормы. Например для НП $\widetilde{5} \triangleq \{3/0, 5/1, 7/0\}$ существуют такие ТП: $T_x = (0\ 0\ 0; 3\ 5\ 7; 0\ 0\ 0)$ и $t_x = (3\ 0;\ 5\ 1;\ 7\ 0)$. Нормы T_x и t_x равны соответственно 9,11 и 9,13, что подтверждает их близость.

Для решения задачи ИК НД возможно применение как T_x-, так и t_x-представлений НП. Целесообразность их применения определяется типом задачи. При построении дендрограмм исходное МД должно быть упорядочено в соответствии с расстоянием, что связано с необходимостью вычислять расстояние между матрицами, моделирующими НП. Расстояние между матрицами **A** и **B** размером $n \times n$ обычно определяют как фро-

бениусовскую норму (ФН). Для $\mathbf{C} = \mathbf{A} - \mathbf{B}$ квадрат ФН имеет вид $\| \mathbf{C} \|_F^2 =$ $= \text{trace}(\mathbf{C}^T\mathbf{C})$, где $\text{trace}(\mathbf{C}) = \sum_{i=1,n} c_{ii}$. В случае разноразмерных матриц ($\mathbf{A} \to$ $\to M \times N$, $\mathbf{B} \to P \times Q$) для определения ФН используют особенности тензорной суммы $\mathbf{C} = \mathbf{A} + \mathbf{B} = \mathbf{I}_B \otimes \mathbf{A} + \mathbf{B} \otimes \boldsymbol{I}_A$, где $\mathbf{I}_A, \mathbf{I}_B$ — тензоры идентичности соответствующей размерности.

Если $\mathbf{A}$ — матрица $M \times N$, то сингулярное разложение A есть факторизация $\mathbf{A} = \mathbf{U}\sum\mathbf{V}^T$, где $\mathbf{U}$ — левосторонняя столбцово-ортонормальная $N \times r$ матрица; $\sum$ —диагональная $r \times r$ матрица собственных значений λ_i, $\lambda_1 \geq \cdots$ $\ldots \geq \lambda_r \geq 0$; V — правосторонняя собственнозначная матрица (ПСЗМ); r — ранг матрицы $\mathbf{A}$.

В работах [16, 17] показано, что для двух ПСЗМ, $\mathbf{A}_V$ и $\mathbf{B}_V$, размерностью $n \times n$, полученных SVD-разложением (Singular value decomposition) матриц $\mathbf{A}$ и $\mathbf{B}$, соответственно $\mathbf{A}_V = [a_1, \ldots, a_n]$ и $\mathbf{B}_V = [b_1, \ldots, b_n]$, а квадрат ФН для $\mathbf{C}_V = \mathbf{A}_V - \mathbf{B}_V$ имеет вид

$$\| \mathbf{C} \|_F^2 = \text{trace}(\mathbf{C}^T\mathbf{C}) = 2n - 2\sum_{i=1}^{n} \langle a_i, b_i \rangle,$$

где $\mathbf{C} = \mathbf{A} - \mathbf{B} = [a_1 - b_1, \ldots, b_n - a_n]$; a, b — ортонормальные векторы (векторизация матриц $\mathbf{A}$, $\mathbf{B}$). Эту величину можно использовать в качестве меры близости между $\mathbf{A}$ и $\mathbf{B}$.

Преимущество применения ПСЗМ для вычисления близости состоит в том, что все объекты кластеризации имеют одинаковый размер для всех НП, представленных в тензорном базисе. В общем случае объект применения ПСЗМ имеет одинаковое число (2) параметров «значение ФП» и различное число α-уровней m. Сравнение объектов (матриц) с различными размерностями в случае больших размерностей усложняется, но размер ПСЗМ в данном случае фиксирован: 2×2.

В табл. 1 и 2 приведены правосторонние матрицы сингулярных разложений и сингулярные числа ТП t-типа. Характерной особенностью ТП является то, что они имеют практически одинаковую ФН ($\sigma_{\max}$). Это объясняется тем, что ТП определена на одном универсальном множестве с большим количеством нулей. Таким образом, в данном случае применение правосторонней матрицы сингулярного разложения для НП в тензорной форме не конструктивно, так как различие в значениях настолько незначительно, что оценка близости (сходство/различие) может быть весьма приближенной, в то время как двумерный вектор сингулярных значений может быть различным.

Фрагмент программы кластеризации НД на основании вектора сингулярных величин SVD-разложения ТП, принятых в качестве признаков кластеризации приведен на рис. 8.

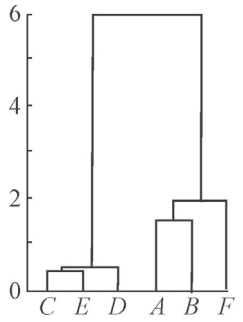
2-адические свойства бинарных деревьев иерархической кластеризации. В работах [18, 19] дан анализ полученных дендрограмм на уровне *p*-адических деревьев. Известно, что множество данных $(x_i, y_i)^T$, $i = 1, n$, представленных виде совокупности пар, можно представить в виде бинарного дерева, вычислив расстояния между каждой парой данных в метрическом базисе (например, евклидова матрица расстояний). Однако,

Таблица 1

НП	X = [1 : 0.5 : 10];	ТП, представленная в Matlab	Правосторонняя матрица сингулярного разложения
$\widetilde{A}$	trapmf (X, [1 1.5 3.5 4.5])	$tA = [X'\ A']$	$V_{tA} = \begin{pmatrix} -1.00 & 0.02 \\ -0.02 & -1.00 \end{pmatrix}$
$\widetilde{B}$	trimf (X, [1.5 2.5 3.5])	$tB = [X'\ B']$	$V_{tB} = \begin{pmatrix} -1.00 & -0.01 \\ -0.01 & 1.00 \end{pmatrix}$
$\widetilde{C}$	trapmf (X, [6 6.5 7.5 8.5])	$tC = [X'\ C']$	$V_{tC} = \begin{pmatrix} -1.00 & -0.03 \\ -0.03 & 1.00 \end{pmatrix}$
$\widetilde{D}$	trimf (X, [7 8 9])	$tD = [X'\ D']$	$V_{tD} = \begin{pmatrix} -1.00 & -0.02 \\ -0.02 & 1.00 \end{pmatrix}$
$\widetilde{E}$	trimf (X, [8 8.5 9])	$tE = [X'\ E']$	$V_{tE} = \begin{pmatrix} -1.00 & -0.01 \\ -0.01 & 1.00 \end{pmatrix}$
$\widetilde{F}$	trimf (X, [3.5 4.5 5.5])	$tF = [X'\ F']$	$V_{tF} = \begin{pmatrix} -1.00 & -0.01 \\ -0.01 & 1.00 \end{pmatrix}$

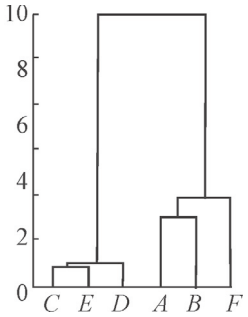
Таблица 2

Полная ТП	Сингулярное значение	Усеченная ТП	Сингулярное значение
$tA = [X'\ A']$	26.79 2.23	$tA1$= [1 0; 1.5 1; 3.5 1; 4.5 0]	6.04 1.13
$tB = [X'\ B']$	26.78 1.21	$tB1$= [1.5 0; 2.5 1; 3.5 0]	4.59 0.83
$tC = [X'\ C']$	26.80 1.54	$tC1$= [6 0; 6.5 1; 7.5 1; 8.5 0]	14.41 1.02
$tD = [X'\ D']$	26.79 1.07	$tD1$= [7 0; 8 1; 9 0]	13.94 0.82
$tE = [X'\ E']$	26.78 0.95	$tE1$= [8 0; 8.5 1; 9 0]	14.75 0.82
$tF = [X'\ F']$	26.78 1.18	$tF1$= [3.5 0; 4.5 1; 5.5 0]	7.94 0.82




```
sa=svd(tA,0); sb=svd(tB,0); sc=svd(tC,0);
sd=svd(tD,0); se=svd(tE,0); sf=svd(tF,0);
v=[sa'; sb'; sc'; sd'; se'; sf'];
figure(1)
Y1=pdist(v,'euclid');
Z1 =linkage(Y1,'single');
[H1,T1]=dendrogram(Z1,6);
set(H1,'LineWidth',2);
title('Single link vector containing the singular values')
```

а




```
figure(2)
Y2=pdist(v,'euclid');
Z2 =linkage(Y2,'complete');
[H2,T2]=dendrogram(Z2,6);
set(H2,'LineWidth',2);
title('Complete link vector containing the singular values')
```

б

Рис. 8. Дендрограммы для ТП *tA, tB*, *tC*, *tD*, *tE*, *tF*, определенных с использованием сингулярных чисел методом ЕС (*а*) и методом ПС (*б*)

если разметить ветви дендрограммы (например, 0 (левая ветвь) и 1 (правая ветвь), или соответственно – 1 и +1), полученной в метрическом базисе, то ее анализ следует выполнять в *p*-адическом (2-адическом) базисе.

Рассмотрим особенности 2-адических дендрограмм, используя результаты, изложенные в работах [18, 19]. Для анализа данных дендрограммы обычно размечают и ранжируют. Для дендрограммы, представленной на рис. 9, выполнено следующее *p*-адическое кодирование терминальных узлов, начиная с корня: $x_1 = 0 \cdot 2^7 + 0 \cdot 2^5 + 0 \cdot 2^2 + 0 \cdot 2^1$; $x_2 = 0 \cdot 2^7 + 0 \cdot 2^5 + 0 \cdot 2^2 + + 1 \cdot 2^1$; ... $x_4 = 0 \cdot 2^7 + 1 \cdot 2^5 + 0 \cdot 2^4 + 0 \cdot 2^3$; ... $x_6 = 0 \cdot 2^7 + 1 \cdot 2^5 + 1 \cdot 2^4$ и т. д. Десятичные эквиваленты *p*-адического представления терминальных узлов такие: $x_1, x_2, \ldots, x_8$ = 0, 2, 4, 32, 40, 48, 128, 192. Расстояния и норма определены соответственно так: $d_p(x, x^1) = d_p \left\| x - x^1 \right\| = 2^{-r+1}$ или $2 \cdot 2^{-r}$, где

$$x = \sum_k a_k 2^k;\ x^1 = \sum_k a_k^1 2^k;\ r = \arg\min\{a_k - a_k^1\};\ \text{норма}\ d_p(x, 0) = 2^{-1+1} = 1.$$

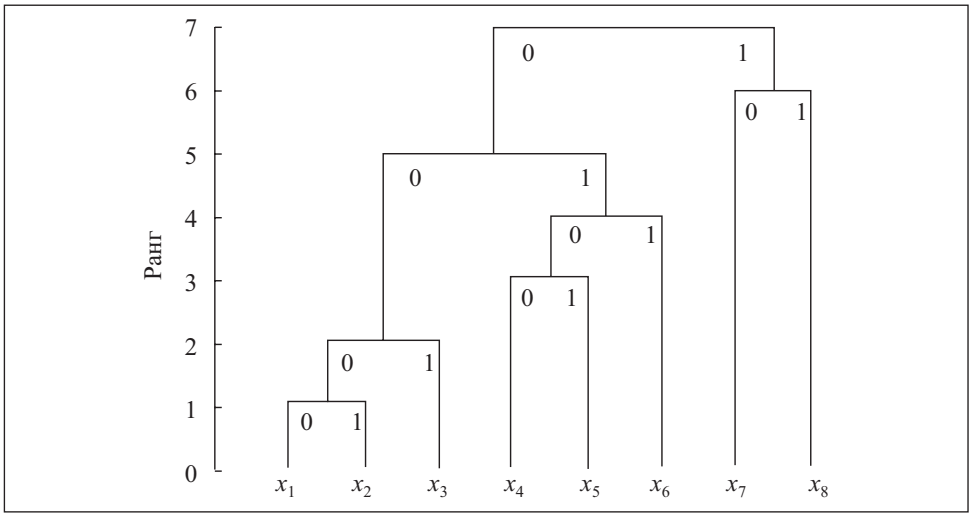


Рис. 9. Размеченная дендрограмма с восемью терминальными узлами: 0 и 1— левая и правая ветви

Для того чтобы найти *p*-адическое расстояние, рассматривают наименьший уровень *r* (если упорядочение идет от терминала к корню) (рис. 9), что идентично паре степенных рядов, которые порождают результат 2^{-r+1}. Таким образом,

$$\| x_1 - x_2 \|_2 = 2^{-2+1} = 1/2 ; \| x_1 - x_4 \|_2 = 2^{-6+1} = 1/32 ; \| x_1 - x_6 \|_2 = 2^{-6+1} = 1/32 .$$

Наименьшее *p*-адическое расстояние на рис. 9 равно 1/128, десятичным эквивалентом числа x_8 есть 208. Максимально возможный десятичный эквивалент *p*-адического числа, соответствующего восьми терминальным узлам, имеет вид $1 \cdot 2^7 + 1 \cdot 2^6 + 1 \cdot 2^5 + ... + 1 \cdot 2^4 + 1 \cdot 2^3 + 1 \cdot 2^2 + 1 \cdot 2^1 = 254$.

Следует заметить, что *p*-адическое представление, показанное на рис. 9, не инвариантно относительно дендрограммного представления. Однако, если *p*-адические представления различны для разных дендрограммных представлений, то *p*-адическая норма и *p*-адическое расстояние являются инвариантными относительно дендрограммного представления. При анализе дендрограмм используют так называемые агломеративные (накопительные) алгоритмы. Как показано в работах [18, 19], использование агломеративного алгоритма кластеризации порождает ультраметрику (т.е искусственное удовлетворение ультраметрического неравенства для заданных каких-либо трех точек) в любом множестве точек, обеспеченном парной функцией различия. Когда множество точек в пространстве данных какой-либо размерности таково, что все триплеты точек удовлетворяют ультраметрическому неравенству, это множество точек имеет естественную

иерархическую структуру, однако нельзя гарантировать, что эта иерархия является уникальной.

Закодированная дендрограмма представляет собой p-адическое число. Рассмотрим кратко основы p-адических чисел и арифметические действия над ними [20, 21]. Пусть $p \in \mathbf{N}$ — фиксированное простое число. Тогда для любого ненулевого $x \in q$ можно написать $x = p^v a / b$ для пары взаимнопростых чисел $a, b \in \mathbf{Z}$ и уникального $v \in \mathbf{Z}$ такого, что a, b не делятся на p. В общем случае целое p-адическое число для произвольного простого p представляет собой последовательность $x = (x_0, x_1, ...)$ вычетов x_n по mod p^{n+1}, удовлетворяющих условию $x_n = x_{n-1} (\mathrm{mod}\, p^{n+1})$, $n \geq 1$.

P-адическаяа норма — функция $|\cdot|_p : q \to [0, \infty)$, получаемая из равенств $|x|_p = p^{-v}$ и $|0|_p = 0$, $|\cdot|_p$, удовлетворяет усиленному неравенству треугольника (УНТ), суть которого в том, что для любых $x, y \in q$ справедливо $|x+y|_p \leq \max\{|x|_p, |y|_p\}$. Порожденная метрика $d_p(x, y) = |x - y|_p$ называется ультраметрикой и обладает рядом парадоксальных свойств [20, 21]. Относительно p-адической нормы q удовлетворяет неархимедовым свойствам, поскольку для каждого $x \in q$ никогда $|nx|_p$ не превышает $|x|_p$ для любого $n \in N$. Существует полнота поля q_p-адических чисел относительно ультраметрики d_p. Уникальное представление каждого $z \in q_p$ имеет вид

$$z = a_v p^v + ... + a_0 + a_1 p + a_2 p^2 + ...,$$

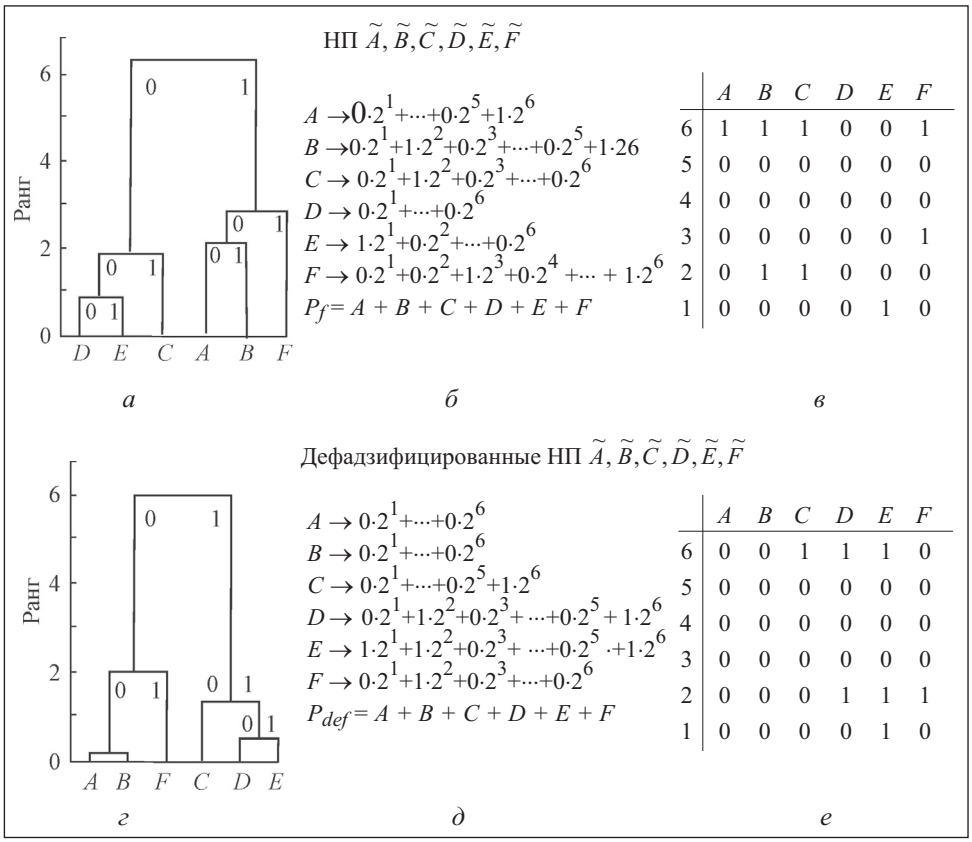
где $v \in \mathbf{Z}$ и $a_i \in \{0, 1, ..., p-1\}$ для всех $i \geq v$. Подпространство q_p — единичный шар $\mathbf{Z}_p = \{x \in q_p \,|\, |x|_p \leq 1\}$ — также может быть представлено в виде $\mathbf{Z}_p = \{a_0 + a_1 p + a_2 p^2 + ... | a_i \in \{0, 1, ..., p-1\}, \forall\, i \geq 0\}$.

Важнейшим свойством p-адических чисел является наличие иерархической структуры, в отличие от обычных чисел, располагающихся линейно. Целые p-адические числа образуют кольцо: их можно складывать, вычитать и перемножать. В нотации теории вычетов сложение и умножение целых p-адических чисел определяется формулами

$$(x+y)_n = x_n + y_n (\mathrm{mod}\, p^{n+1}),$$

$$(x\, y)_n = x_n\, y_n (\mathrm{mod}\, p^{n+1}).$$

Однако здесь отсутствует естественный порядок, понятия отрицательного и положительного числа не имеют смысла и выполняется равенство $-1 = \lim(p^n - 1)$ при $n \to \infty$, например для величины -1 в 3-адическом базисе получаем $-1_3 = .222222...$, в 2-адическом — $-1_2 = .111111$.



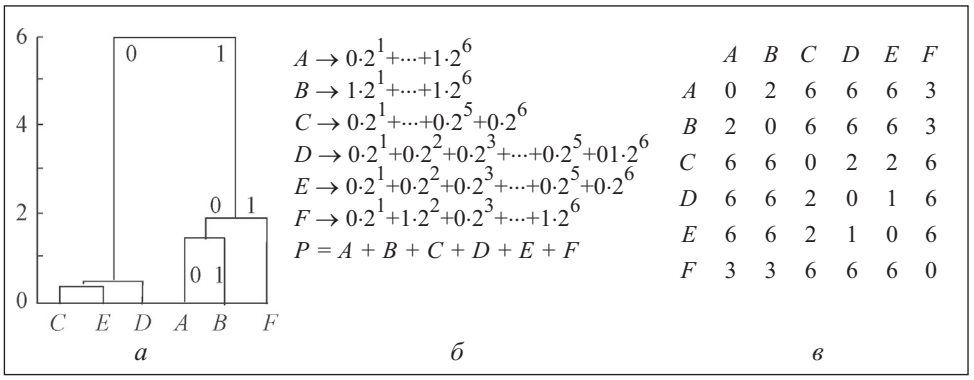
	A	B	C	D	E	F
6	1	1	1	0	0	1
5	0	0	0	0	0	0
4	0	0	0	0	0	0
3	0	0	0	0	0	1
2	0	1	1	0	0	0
1	0	0	0	0	1	0

	A	B	C	D	E	F
6	0	0	1	1	1	0
5	0	0	0	0	0	0
4	0	0	0	0	0	0
3	0	0	0	0	0	0
2	0	0	0	1	1	1
1	0	0	0	0	1	0

Рис. 10. 2-адические характеристики дендрограмм для НП и их дефадзифицированных значений: *а*, *г* — размеченные бинарные деревья; *б*, *д* — 2-адические числа, характеризующие деревья; *в*, *е*— 2-адические матрицы бинарных деревьев (тест полной связи)

Результаты экспериментальных исследований ИК НД в *p*-адическом базисе. На рис. 6 приведены результаты кластеризации НП $\widetilde{A},\widetilde{B},\widetilde{C},\widetilde{D},\widetilde{E},\widetilde{F}$ и их дефадзифицированных значений [12]. Выполним кодирование дендрограмм бинарным алфавитом и рассмотрим 2-адические матрицы НП $\widetilde{A},\widetilde{B},\widetilde{C},\widetilde{D},\widetilde{E},\widetilde{F}$, их дефадзификации и тензорные модели.

На рис. 10 представлено 2-адическое кодирование бинарных деревьев и вычислены 2-адические числа, характеризующие дендрограммы НП P_f и их дефадзификации P_{def}. Величина $\mathrm{abs}(P_f - P_{def})/\max(P_f, P_{def}) < 10\,\%$ свидетельствует о структурной близости бинарных деревьев, чего и следовало ожидать не только на основании данных, приведенных в [14], но и потому, что расстояния между НМ в соответствии с принятой парадигмой являются четкими.



	A	B	C	D	E	F
A	0	2	6	6	6	3
B	2	0	6	6	6	3
C	6	6	0	2	2	6
D	6	6	2	0	1	6
E	6	6	2	1	0	6
F	3	3	6	6	6	0

Рис. 11. 2-адические характеристики дендрограмм тензорной модели НП: *а* — размеченное бинарное дерево; *б* — 2-адическое число, характеризующее дерево (кластеризация по методу ПС); *в* — u_θ — матрица ультраметрических расстояний, связанная с иерархической кластеризацией

Сравнивая рис. 7, *а*, *б*, и рис. 10, *а*, *б*, видим, что дендрограммы для НП [12] и тензорных моделей НП достаточно близки, если брать в качестве критерия максимальные ранги дендрограмм и их вложенность. Этот вывод подтверждается также и их 2-адическими матрицами.

На рис. 11 представлены результаты кластеризации НП $\widetilde{A},\widetilde{B},\widetilde{C},\widetilde{D},\widetilde{E},\widetilde{F}$, представленных в тензорном базисе $\widetilde{A} \to tA$, $\widetilde{B} \to tB$, $\widetilde{C} \to tC$, $\widetilde{D} \to tD$, $\widetilde{E} \to tE$, $\widetilde{F} \to tF$, и 2-адические характеристики дендрограмм тензорной модели НП. Из рис. 11, *в*, видно, что $u_\theta(A, B) = 2$ является параметром, для которого A и B связаны в одном и том же кластере, аналогично A и C включены в один и тот же кластер для случая, когда $u_\theta(A,C) = 6$. Таким образом, для минимального числа факторов, принципиально определяющих результат ИК НД, в случае тензорного представления НД можно утверждать следующее:

- в качестве системы признаков в случае *t*-способа представления НД следует выбирать сингулярные числа сингулярного разложения ТП;
- в качестве мер близости объектов целесообразно использовать величину ФН между матрицами ТП;
- основным способом формализации представлений об эквивалентности объектов, составляющих отдельный кластер или их объединение (дендрограмму), может быть учет внутри- и межкластерных расстояний в ультраметрическом пространстве.

Выводы

1. Иерархическую кластеризацию НД, представленных в виде НМ, целесообразно выполнять посредством представления объектов класте-

ризации — НП — в виде ТП *T*- и *t*-типов. Тензор-переменная *T*-типа формируется как результат тензорного (кронекерова) произведения «значение ⊗ функция принадлежности» с матрицей $n \times n$, ТП *t*-типа формируется как многомерный массив с матрицей $2 \times n$ (n — число упорядоченных пар НМ).

2. При использовании тензорных моделей НД для объективности иерархической кластеризации возможно формирование тензорной модели для всего УМ. Неучет УМ приводит к НМ разной размерности, что затрудняет процедуру оценки их близости.

3. В качестве мер близости объектов (НД) целесообразно использовать величину фробениусовского расстояния между матрицами ТП.

The questions of clusterization (construction of binary trees-dendrograms) of the data, presented in the form of fuzzy variables, which are in turn simulated by tensors, are considered. A dendrogram encoded by binary alphabet is a 2-adical number, which can be used as the dendrogram characteristic. A comparison of hierarchical clusterizations of fuzzy data and their defuzzifications, performed at the level of 2-adical trees, allows us to draw a conclusion on the presence (absence) of structure nearness of the objects.

1. *Воронцов К. В.* Лекции по алгоритмам кластеризации и многомерного шкалирования// www.MachineLearning.ru.
2. *Бирюков А. С., Резанов В. В., Шмаров А. С.* Решение задач кластерного анализа коллективами алгоритмов // Ж. вычисл. матем. и матем. физ. — 2008. — **48**, № 1. — С. 176—192.
3. *Жамбю М.* Иерархический кластер-анализ и соответствия. — М. : Финансы и статистика, 1988. — 342 с.
4. *Мандель И. Д.* Кластерный анализ. — М. : Финансы и статистика, 1988. — 176 с.
5. *Тыртышников Е. Е.* Тензорные аппроксимации матриц, порожденных асимптотическими гладкими функциями//Мат. сборник. — 2003. — **194**, № 6. — С. 147—160.
6. *Zak L.* Clustering of Vaguely Defined Objects// Archivum Mathematicum (Brno). — 2002. — **38**. — P. 37—50.
7. *Кофман А.* Введение в теорию нечетких множеств: Пер. с франц. —М. : Радио и связь, 1982. — 432 с.
8. *Carlsson G. , M'emoli F.* Characterization, Stability and Convergence of Hierarchical Clustering Algorithms//Technical report. — 2009.— http://jmlr. csail. mit.edu/papers/volume11/carlsson10a/carlsson10a.pdf
9. *Carlsson G., M'emoli F.* Characterization, Stability and Convergence of Hierarchical Clustering Methods// J. of Machine Learning Research. — 2010. — № 11. — P. 1425—1470.
10. *Burago D., Burago Y., Ivanov S. A.* Course in Metric Geometry // AMS Graduate Studies in Math. American Mathematical Society. — 2001. — Vol. 33. www.math.psu.edu/petrunin/papers/alexandrov/bbi.pdf
11. *Guh Yuh-Yuan, Yang Miin-Shen, Po Rung-Wei, Lee E. S.* Establishing Performance Evaluation Structures by Fuzzy Relation-based Cluster Analysis // Computers and Mathematics with Applications. — 2008. — № 56. — P. 572—582.
12. *Gol M. G., Yazdi H. S.* A New Hierarchical Clustering Algorithm on Fuzzy Data (FHCA)// Intern. J. of Computer and Electrical Engineering. — 2010. — Vol. 2, № 1. — P. 1793— 1816.

13. *Delgado M., Gomez-Skarmeta A. F., Vila A.* Intern. J. of Approximate Reasoning.— 1996. — № 14. — P. 237—257.
14. *Минаев Ю. Н., Филимонова О. Ю.* Нечеткая математика на основе тензорных моделей неопределенности. Ч. I. Тензор-переменная в системе нечетких множеств. Ч. II. Нечеткая математика в тензорном базисе// Электрон. моделирование. — 2008. — **30**, № 1. — С. 43—59; № 2. — С. 4—21.
15. *Colda T. G., Bader B. W.* Tensor Decompositions and Applications // ACM Transactions on Mathematical Software. — 2006. — Vol. 32, № 4. — P. 635—653.
16. *Kiyoung Yang, Cyrus Shahabi* A PCA-based Similarity Measure for Multivariate Time Series —http://infolab.usc.edu/ Docs-De-mos/ mmdb04.pdf
17. *Singhal D., Seborg A.* Clustering of Multivariate Time-series Data// Proc. of the American Control Conference. Anchorage, Alaska, USA, 8-10 May, 2002. — 2002. — Vol. 5. — P. 351—358.
18. *Murtagh F.* Symmetry in Data Mining and Analysis: A Unifying View based on Hierarchy//arXiv: 50805. 2744v1 [stat.ML] 18 May 2008. — 33 p.
19. *Murtagh F., Downs G., Contreras P.* Hierarchical Clustering of Massive, High Dimensional Data Sets by Exploiting Ultrametric Embedding// SIAM J. on Scientific Computing. — 2008. — Vol. 30. — P. 707—730.
20. *Gouvea F. Q.* P-Adic Numbers: An Introduction. — Springer, 2003. — 208 p.
21. *Schikhof W. H.* Ultrametric Calculus. An Itroduction to *p*-adic Analysis. — Cambridge University Press, 1984. — 306 p.

Поступила 12.09.11;
после доработки 17.04.12

МИНАЕВ Юрий Николаевич, д-р техн. наук, профессор кафедры компьютерных систем и сетей Национального авиационного университета Украины. В 1959 г. окончил Харьковский политехнический ин-т. Область научных исследований — интеллектуальный анализ данных, применение интеллектуальных технологий в системах принятия решений.

ФИЛИМОНОВА Оксана Юрьевна, канд. техн. наук, доцент Киевского национального университета строительства и архитектуры. В 1989 г. окончила Киевский инженерно-строительный ин-т. Область научных исследований — интеллектуальный анализ данных.

МИНАЕВА Юлия Ивановна, аспирантка кафедры основ информатики Киевского национального университета строительства и архитектуры, который окончила в 2008 г. Область научных исследований — интеллектуальный анализ данных.