УДК 519.5:517.1:

***Минаев Ю.Н.**, д.т.н,
****Филимонова О.Ю.**, к.т.н.,
****Минаева Ю.И.**

ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ НЕЧЕТКИХ ДАННЫХ В ТЕНЗОРНОМ БАЗИСЕ

*** Национальный авиационный університет**
****Киевский национальный университет строительства и архитектуры**

Рассмотрены вопросы кластеризации (построение бинарных деревьев-дендрограмм) для данных, представленных в виде нечетких переменных, которые, в свою очередь, моделируются тензорами. Закодированная бинарным алфавитом дендрограмма представляет собой 2-адическое число, которое может быть использовано как характеристика дендрограммы. Сравнение иерархических кластеризаций нечеткимх данных и их дефадзификаций, выполненное на уровне 2-адических деревьев, позволяет сделать вывод о наличии (отсутствиии) структурной близости объектов.

Введение

Проблема иерархической кластеризации (ИК) в последнее время приобрела особую остроту и актуальность в связи с извлечением знаний из данных. Цели кластеризации при этом могут быть совершенно различными в зависимости от особенностей конкретной прикладной задачи [1]. В общем случае алгоритм кластеризации – это функция f: X→Y, которая любому объекту x∈X ставит в соответствие метку кластера y∈Y.

В работах [2-4] показано, что решение задачи кластеризации принципиально неоднозначно. Алгоритмы ИК в подавляющем большинстве случае работают по т.н. агломеративному (объединительному) принципу, при котором 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):

Современное состояние проблемы

КА возник как специфическая методология проведения классификации неоднород-ных статистических совокупностей, но в современных условиях сфера его применения необычайно расширилась, благодаря тому, что бинарные деревья (дендрограммы) – результат иерархической кластеризации, стали анализироваться на уровне p-адических (конкретно, 2-адических) чисел с применением *ультраметрики*. В последнее время ИКА применяется для решения задач анализа нечетких данных (НД), которые могут быть представлены в виде НМ, интервала или некоторой совокупности данных (числовая последовательность).

Для описания разнообразных задач кластеризации (классификации) в соответствии с [4] рациональным является использование теории *бинарных отношений*, которая показала новые возможности в условиях неполной информации. Все сказанное касается по сути одной проблемы: возможности конструирования такого показателя близости между объектами, который *не зависел бы* от способа измерения переменных. При наличии такого показателя его применение будет давать *одинаковые результаты* при любых допустимых преобразованиях шкал. Исследования приводят к неконструктивным

выводам относительно теоретической и практической ценности всевозможных метрик: результаты работы алгоритмов классификации могут непредсказуемо меняться в зависимости от выбора способа измерения показателей [4].

Аппроксимационный подход в кластерном анализе. Следуя работе [4], 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 в общем случае возникает естественный функционал, отражающий стремление максимально приблизить результирующее отношение к имеющимся данным: $\|Y-XP\| \to \min$, где $\|\ \|$ – какая-либо норма. Задачи такого типа – аппроксимация «плохо устроенного» множества *X* и «хорошо устроенной структурой» Y – известны в математике и имеют множество приложений [5].

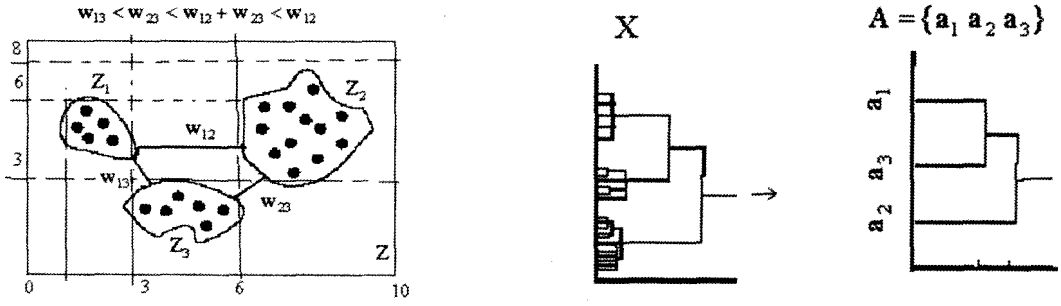
Базовые нотации классической кластеризации применительно к условиям неопределенности, моделируемой на уровне теории нечетких множеств (ТНМ) имеют особенности [6]. Понятие расстояния составляет объект исследования многих работ, в ТНМ часто используются два определения расстояния [6]: для НМ $\widetilde{A},\widetilde{B}$ с функциями при-надлежности (ФП)- $\mu_{\widetilde{A}}(x_i)$, $\mu_{\widetilde{B}}(x_i) \in [0,1]$ обобщенное расстояние Хеммимнга (линейное)- $d(\widetilde{A},\widetilde{B}) = \sum_{i=1}^{n} |\mu_{\widetilde{A}}(x_i) - \mu_{\widetilde{B}}(x_i)|$, Евклидово (квадратичное) расстояние – $e(\widetilde{A},\widetilde{B}) = \left(\sum_{i=1}^{n} (\mu_{\widetilde{A}}(x_i) - \mu_{\widetilde{B}}(x_i))^2\right)^{1/2}$. Отметим, что $d(\widetilde{A},\widetilde{B})$ и $e(\widetilde{A},\widetilde{B})$ – четкие величины, НМ $\widetilde{A},\widetilde{B}$ должны быть заданными на одном универсальном множестве (УМ). В реальных ситуациях d часто принимают равным любой метрике в $\mathbf{R}^m$.

В [7] отмечено, что желательные свойства алгоритмов кластеризации исходят от практиков, которые имеют интуитивные понятия о том, что такое *хорошая* кластеризация, оценивая ее визуально. Доказана теорема [7], которая утверждает эквивалентность между ультраметрикой и дендрограммой, представляет дендрограмму как ультраметрическое пространство, любой метод ИК может рассматриваться как отображение финитного метрического пространства (МП) в финитное ультраметрическое пространство. В соответствии с работой [8], если (X, d_X) и (Y, d_Y) – два финитных МП с метриками d_X и d_Y, (X,u_X) и (Y,u_Y) – два финитных метрических ультраметрических пространства с метриками u_X и u_Y, соответствующих выходам, порожденным ИК по методу ЕС, то:

$d_{GH}((X,u_X)\ (Y,u_Y)) \le d_{GH}((X,d_X)\ (Y,d_Y))$,

где d_{GH} () расстояние Громова-Хаусдорфа.

В работах [7, 8] поставлен вопрос *сходимости* дендрограмм, рис.1. Это понятие формализовано через эквивалентное представление дендрограмм как ультраметрики и по-следующим определением GH-расстояния между результирующими метриками. Показано, что GH-расстояние – метрика на ультраметических пространствах .В работе [8] доказана теорема 28, объясняющая сущность сходимости дендрограмм, проиллюстрированная на рис.1.



а) б) в)

Рис.1. Иллюстрация теоремы о сходимости дендрограмм [8].

В [9] предложен алгоритм, практически обобщающий некоторые подходы к решению задач кластеризации НД. Симметричные трапециевидные НП (или НЧ) представлены на рис. 2 и 3. Параметризация трапециевидного НЧ $\widetilde{A}$ обозначается $\widetilde{A} = m(a_1, a_2, a_3, a_4)$, где a_1, a_2, a_3, a_4 – центр, внутренний диаметр, левый внешний радиус и правый внешний ради-ус соответственно. Полезность этого представления, по мнению авторов работы [12], в том, что четыре типа НД (рис.7) можно представить единым образом.

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 предложено определять 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 предложен алгоритм для кластеризации НД, приведенных на рис. 4. Иллюстрация алгоритма приведена ниже:

НД переформулированы в соответствии с предложенной метрикой:

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

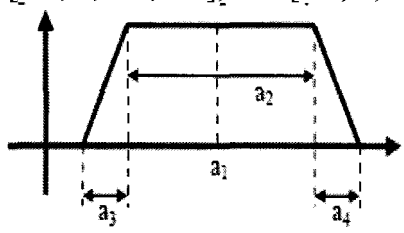


Рис.2. Параметризация трапецевидного НД

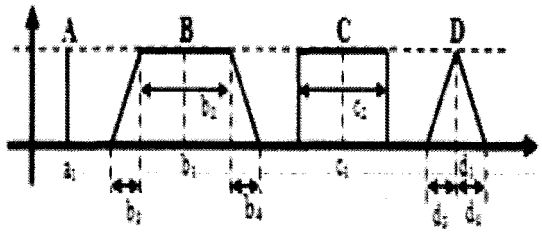
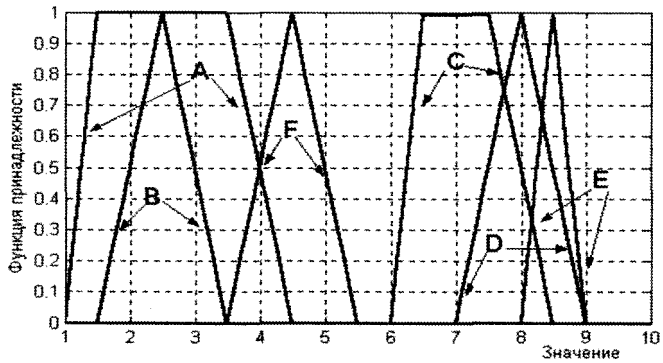


Рис. 3. 4 типа трапециевидных НД




```
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.4. Тестовое множество НП в форме $\widetilde{x} = \{x/\mu\}$

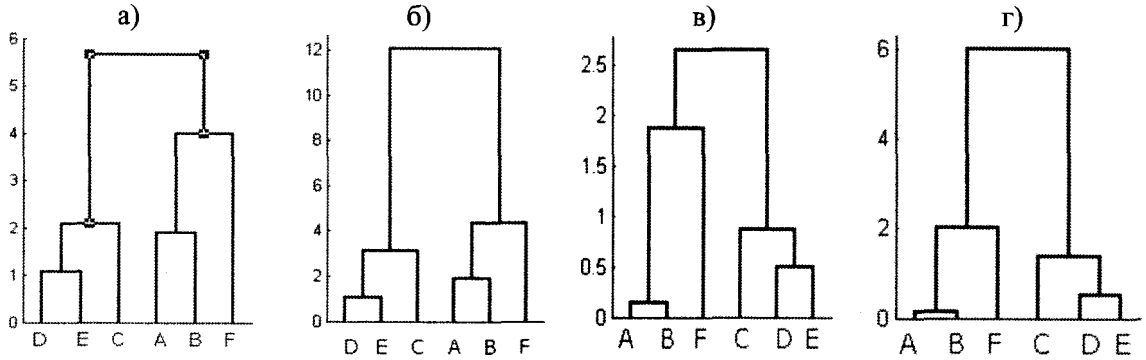


Рис. 5. Дендрограмма для **НП в метрике работы** [12]: а) метод EC, б) метод ПС; для **дефадзифицированных НП: в)** метод EC, г) метод ПС

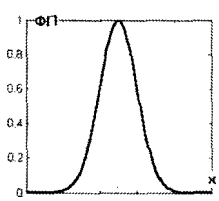
В частности, влияние нечеткости на структуру данных проявляется в данном случае в том, что дендрограммы для НД при использовании метода ЕС и фадзифицированных данных (а-в) и при использовании метода ПС (б-г) с точки зрения ультраметрики имеют различную структуру (число уровней отличается в 2 раза). Трудно дать физическую интерпретацию этому факту, но это, несомненно, новое знание.

Суммируя сказанное, любая ИК генерирует индексированную финитную последовательность гнездовых четких разделений (и наоборот), которые выполняются от начала е единичным классом равным целому множеству объектов в один одно-объектный класс.

Постановки основных задач, алгоритмы их решения

В работе [10] предложены тензорные модели неопределенности. Известно [6], что если E={x} – универсальное множество, x∈E, то нечеткое подмножество $\tilde{A}$ множества E определяется как совокупность упорядоченных пар $\{(x, \mu_{\tilde{A}}(x))\}$, $\forall x \in E$, где $\mu_{\tilde{A}}(x)$ – ФП, $\mu_{\tilde{A}} \to [0,1]$.

Введенная в работе [10] 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 × n. На рис. 6-а представлена НП с выпуклой ФП и ее тензорные аналоги (рис.6-б).



$\tilde{x} = \{x/\mu_x\}$

а)

$$\tilde{x} \to T_x = x \otimes \mu_x = \begin{pmatrix} x_1 \mu^x_1 & x_2 \mu^x_1 \cdots x_n \mu^x_1 \\ \vdots & \ddots & \vdots \\ x_1 \mu^x_n & x_2 \mu^x_n \cdots x_n \mu^x_n \end{pmatrix}$$

б)

Рис. 6. Моделирование НП: а) НП с выпуклой ФП; б) тензорный аналог НП.

С учетом тензорного представления исходных данных основные задачи работы сформулированы в следующем виде:

- исследовать ИК НД при представлении входных данных тензорными моделями, показать проявление влияния нечеткости на выход ИК – бинарное дерево;
- исследовать ультраметрические свойства дендрограмм при представлении вход-ных данных тензорными моделями, определить ультраметрическую матрицу, 2-адические характеристики дендрограмм, показать возможность сравнения дендрограмм на основании вычисления 2-адической оценки.

Сопоставимость экспериментов. НП в ПММ, например, МатЛаб, задаются в виде процедур, которые определяют переменную на УМ E, например:

НП с треугольной ФП – μ^x=trimf(E, [$\underbrace{P_1\ P_2\ P_3}_{\text{параметры}}$]), где $P_1 < P_2 < P_3$,

НП с трапециевидной ФП – μ^x=trapmf(E, [$\underbrace{P_1\ P_2\ P_3\ P_4}_{\text{параметры}}$]), где

$P_1 < P_2 < P_3 < P_4$,

НП с Гауссовой ФП – μ^x=gaussmf(E, [$\underbrace{P_1\ P_2}_{\text{параметры}}$]),

где P_1 – дисперсия, P_2 – среднее и др. Операции над НП предполагают, что НП представлены в виде НМ на одном универсальном множестве, т.е.

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

В простейшем случае для стандартной треугольной ФП имеем тензорные модели:

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

для НП $\tilde{5}^{\Delta}$= {3/0, 5/1, 7/0} имеем такие ТП:

$$T_x = \begin{pmatrix} 0 & 0 & 0 \\ 3 & 5 & 7 \\ 0 & 0 & 0 \end{pmatrix}.$$

В работе расстояние между матрицами **A** и **B** размером n × n предложено определять как норму Фробениуса (ФН), для C=A–B квадрат НФ имеет вид:

$$\|C\|_F^2 = \mathrm{trace}(C^T C),$$

где: $\mathrm{trace}(C)=\sum_{i=1,n} c_{ii}$. Если матрицы A и B векторизованы, т.е. представлены в форме одномерного массива (по столбцам), A=[a_1,…, a_n] и B=[b_1,…, b_n], то $\|C\|_F^2=\sum_{i=1}^n \langle a_i, b_i \rangle$, где $\langle a_i, b_i \rangle$ – внутреннее произведение векторов a_i и b_i. Матричная форма НП предоставляет новые дополнительные информационные возможности, в частности, в определении признаков кластеризации. Известно, что ФН матрицы A размером m × n определяется как

$$\|A\|_F = (\sum_{i=1}^m \sum_{j=1}^n (a_{ij})^2)^{1/2} = (\mathrm{trace}(A^T A))^{1/2}$$

Пусть a_i – столбцовый вектор размерности m, матрица A может быть представлена как A= =[a_1,…, a_m], соответственно НФ может быть определена из выражения

$$\|A\|_F^2 = (\mathrm{trace}(A^T A)) = \sum_{i=1}^n \langle a_i, a_i \rangle.$$

2-адические свойства бинарных деревьев иерархической кластеризации, 2-адическая кластеризация

В [11-14] рассмотрены вопросы анализа дендрограмм на уровне p-адических деревьев. *P-адическая дендрограмма* может быть получена следующим образом. 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, причом это расстояние может быть вычислено как в метрическом, так и в ультраметрическом базисе. Однако, если разметить ветви дендрограммы (например, 0 (левая ветвь) и 1 (правая ветвь), или -1 и +1 соответственно), полученной, скажем, в метрическом базисе, то ее анализ должен выполняться в *p*-адическом (конкретно, 2-адическом) базисе. Рассмотрим особенности 2-адических дендрограмм, используя результаты, изложенные в работах [11, 12].

Для анализа данных дендрограммы обычно размечают и ранжируют (рис.13). Для ранжированной дендрограммы (рис.7) создается следующее *p*-адическое кодирование терминальных узлов, проходя путь от корня:

$x_1=0\cdot2^7+0\cdot2^5+0\cdot2^2+0\cdot2^1$;

$x_2=0\cdot2^7+ +0\cdot2^5+0\cdot2^2+1\cdot2^1$;

...

$x_4=0\cdot2^7+1\cdot2^5+0\cdot2^4+0\cdot2^3$;

...

$x_6=0\cdot2^7+ +1\cdot2^5+1\cdot2^4$ и др.

Десятичные эквиваленты *p*-адического представления терминальных узлов такие: $x_1, x_1, \ldots, x_8$ = 0, 2, 4, 32, 40, 48, 128, 192. Расстояния и норма определены соответственно так: $d_p(x, x^1) = d_p\|x - x^1\| = 2^{-r+1}$ или $2\cdot2^{-r}$, где $x = \sum_k a_k 2^k$, $x^1 = \sum_k a_k^1 2^k, r = \mathrm{argmin}\{a_k - a_k^1\}$, норма – $d_p(x, 0) = 2^{-1+1}=1$. Для того, чтобы найти *p*-адическое расстояние, рассматривают наименьший уровень r (если упорядочение идет от терминала к корню, рис.7), что есть идентичным 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. 13 равно 1/128, десятичный эквивалент числа x_8 есть 208. Максимально возможный десятичный эквивалент *p*-адического числа, которое соответствует 8 терминальным узлам – $1\cdot2^7+1\cdot2^6+1\cdot2^5+ +1\cdot2^4+1\cdot2^3+1\cdot2^2+1\cdot2^1=254$.

Отметим, что *p*-адическое представление, показанное на рис.7, не есть инвариантным относительно дендрограммного представления. Однако, если *p*-адические представления разнятся для разных дендрограммных представлений, анализ показывает, что *p*-адическая норма и *p*-адическое расстояние являются *инвариантными* относительно дендрограммного представления.

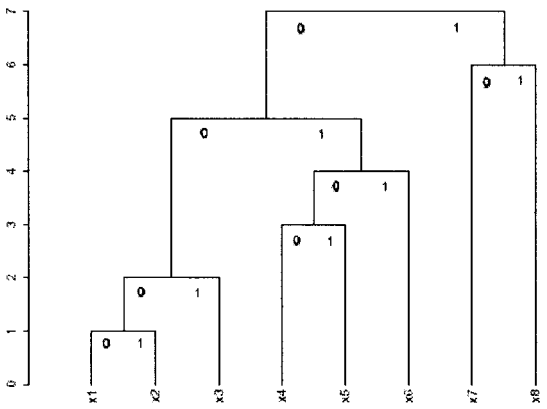


Рис.7. Размеченная и ранжированная дендрограмма с 8-ю терминальными узлами, ветви помечены как 0 (левая) и 1 (правая)

При анализе дендрограмм используют т.н. агломеративные (накопительные) алгоритмы. Как показано в роботах [11, 12], путем использования агломеративного алгоритма кластеризации можно *породить* ультраметрику (т.е получить искусственно удовлетворение ультраметрического неравеноства для заданых каких либо трех точек) в каком-либо множестве точек, обеспеченном парной функцией различия. Когда множество точек в пространстве данных какой-либо размерности таково, что все триплеты точек удовлетворяют ультраметрическому неравенству, то это множество точек имеет *естественную* иерархическую структуру.

P-адические числа (p-а.ч.). Закодированная дендрограмма представляет собой p-адическое число. Рассмотрим коротко основы p-а.ч. и арифметику над ними [3, 14]. Пусть $p\in\mathbf{N}$ будет фиксированным простым числом. Тогда для любого ненулевого $x\in q$, мы можем всегда написать $x = p^v \cdot a/b$, для пары взаимно-простых чисел $a, b \in\mathbf{Z}$ и уникального $v\in\mathbf{Z}$ так, что p не делит a, b. В общем случае целое p-а. ч. для произвольного простого *p* представляет собой последовательность $x =(x_0, x_1,...)$ вычетов x_n по mod p^{n+1}, удовлетворяющих условию $x_n= x_{n-1}(\text{mod } p^{n+1})$, $n \geq 1$.

P-adic норма – функция $|\cdot|_p$:q→[0,∞), получаемая через $|x|_p=p^{-v}$ и $|0|_p = 0$, $|\cdot|_p$ удовлетворяет *усиленному неравенству треугольника* (УНТ), состоящему в том, что для любых *x,* y∈q мы имеем $|x+y|_p \leq \max\{|x|_p, |y|_p\}$. Порожденная метрика – $d_p(x, y)=|x-y|_p$ – имеет название ультраметрика. Ультраметрика обладает целым рядом парадоксальных свойств, которые рассматриваются в 1работах [13, 14].

Относительно p-адической нормы q удовлетворяет неархимедовым свойствам, поскольку для каждого x∈ q справедливо, что $|nx|_p$ никогда не превышает $|x|_p$ для любого n∈ N. Имеет место полнота q относительно ультраметрики d_p, q_p – поле p-а. ч.. Более конкретно, существует уникальное представление каждого z $\in q_p$:

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

$a_i \in \{0, 1, \ldots, p-1\}, \forall i \geq 0\}$. Важнейшим свойством p-а.ч. является то, что они имеют *иерархическую* структуру, в отличие от обычных чисел, которые располагаются линейно. Целые *p-а.ч.* образуют кольцо: их можно складывать, вычитать и перемножать. Однако здесь отсутствует естественный порядок, понятие отрицательного и положительного числа не имеют смысла, для *p-а.ч.* выполняется $-1 = \lim(p^n - 1)$ при $n \to \infty$. Для величины -1, например, в 3-адическом базисе имеем: -1_3 =.222222..., в 2-адическом: -1_2=. 111111 …, соответствующее дерево для -1_3 имеет вид (рис.8).

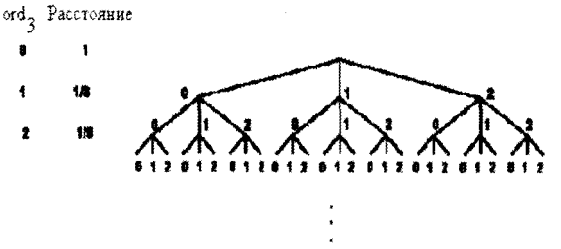


Рис.8. 3-адическое представление числа -1

В нотации теории вычетов сложение и умножение целых p-а.ч. определяется формулами:

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

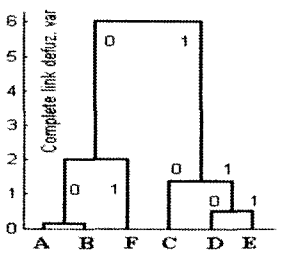
Экспериментальное исследование ИК НД в p-адическом базисе

В предыдущих разделах были приведены результаты кластеризации НД, выполненные на основании метрики, разработанной в [11].

На рис. 9 приведены результаты кластеризации НП $\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{E}, \tilde{F}$ из [12] и их дефадзифицированных значений: а), б) дендрограммы НП $\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{E}, \tilde{F}$: кластеризация по методу ЕС а) и ПС -б); в), г) – дендрограммы фадзифицированных НП, $\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{E}, \tilde{F}$: кластеризация по методу ЕС в) и ПС -г). Выполним кодирование дендрограмм бинарным алфавитом и рассмотрим 2-адические матрицы НП $\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{E}, \tilde{F}$, их дефадзификаций и тензорных моделей НП.

На рис. 9 и рис. 10 представлено 2-адическое кодирование бинарных деревьев и вычислены 2-адические числа, характеризующие дендрограммы – P_{fuzzy} и P_{defuz} соответственно. Величина $abs(P_{fuzzy}-P_{defuz})/max(P_{fuzzy}, P_{defuz})<10$ %, что действительно свидетельствует о структурной близости бинарных деревьев. Отметим, что этого следовало ожидать не только на основании работ [11], но и на том основании, что расстояния между НМ в соответствии с принятой парадигмой является четким.

Дефадзифицированные НП $\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{E}, \tilde{F}$



а)

$A \to 0 \cdot 2^1 + \cdots + 0 \cdot 2^6$

$B \to 0 \cdot 2^1 + \cdots + 0 \cdot 2^6$

$C \to 0 \cdot 2^1 + \cdots + 0 \cdot 2^5 + 1 \cdot 2^6$

$D \to 0 \cdot 2^1 + 1 \cdot 2^2 + 0 \cdot 2^3 + \cdots + 0 \cdot 2^5 + 1 \cdot 2^6$

$E \to 1 \cdot 2^1 + 1 \cdot 2^2 + 0 \cdot 2^3 + \cdots + 0 \cdot 2^5 + 1 \cdot 2^6$

$F \to 0 \cdot 2^1 + 1 \cdot 2^2 + 0 \cdot 2^3 + \cdots + 0 \cdot 2^6$

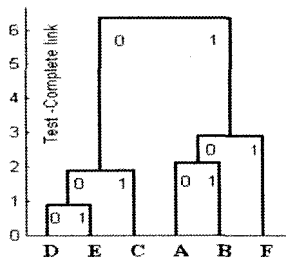
$P_{defuz} = A + B + C + D + E + F$

б)

	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

в)

Рис.9. 2-адические характеристики дендрограмм для дефадзифицированных НП: а) размеченное бинарное дерево; б) 2-адическое число, характеризующее дерево; в) 2-адическая матрица бинарного дерева



НП $\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{E}, \tilde{F}$

$A \rightarrow 0{\cdot}2^1+\cdots+0{\cdot}2^5+1{\cdot}2^6$

$B \rightarrow 0{\cdot}2^1+1{\cdot}2^2+0{\cdot}2^3+\cdots+0{\cdot}2^5+1{\cdot}26$

$C \rightarrow 0{\cdot}2^1+1{\cdot}2^2+0{\cdot}2^3+\cdots+0{\cdot}2^6$

$D \rightarrow 0{\cdot}2^1+\cdots+0{\cdot}2^6$

$E \rightarrow 1{\cdot}2^1+0{\cdot}2^2+\cdots+0{\cdot}2^6$

$F \rightarrow 0{\cdot}2^1+0{\cdot}2^2+1{\cdot}2^3+0{\cdot}2^4+\cdots+1{\cdot}2^6$

$P_{fuzzy}=A+B+C+D+E+F$

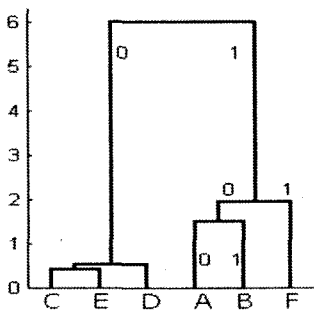
	A	B	C	D	E	F
6	1	1	1	0	0	1
5	0	0	0	0	0	0
4	0	0	0	0	0	0
3	0	0	0	0	0	1
2	0	1	1	0	0	0
1	0	0	0	0	1	0

а) б) в)

Рис.10. 2-адические характеристики дендрограмм для НП
а) размеченное бинарное дерево; б) 2-адическое число, характеризующее дерево;
в) 2-адическая матрица бинарного дерева

На рис. 11 представлены результаты кластеризации НП $\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{E}, \tilde{F}$, представленных в тензорном базисе: $\tilde{A} \rightarrow tA$, $\tilde{B} \rightarrow tB$, $\tilde{C} \rightarrow tC$, $\tilde{D} \rightarrow tD$, $\tilde{E} \rightarrow tE$, $\tilde{F} \rightarrow tF$, 2-адичес-кие характеристики дендрограмм тензорной модели НП: размеченное бинарное дерево, 2-адическое число, характеризующее дерево (кластеризация по методу ПС)

2-адические модели тензорных моделей НП $\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}, \tilde{E}, \tilde{F}$



$A \rightarrow 0{\cdot}2^1+\cdots+1{\cdot}2^6$

$B \rightarrow 1{\cdot}2^1+\cdots+1{\cdot}2^6$

$C \rightarrow 0{\cdot}2^1+\cdots+0{\cdot}2^5+0{\cdot}2^6$

$D \rightarrow 0{\cdot}2^1+0{\cdot}2^2+0{\cdot}2^3+\cdots+0{\cdot}2^5+01{\cdot}2^6$

$E \rightarrow 0{\cdot}2^1+0{\cdot}2^2+0{\cdot}2^3+\cdots+0{\cdot}2^5+0{\cdot}2^6$

$F \rightarrow 0{\cdot}2^1+1{\cdot}2^2+0{\cdot}2^3+\cdots+1{\cdot}2^6$

P=A+B+C+D+E+F

а) б)

Рис.11. 2-адические характеристики дендрограмм тензорной модели НП
а) размеченное бинарное дерево;
б) 2-адическое число, характеризующее дерево

Таким образом, относительно минимального количества факторов, принципиально определяющих результат ИК НД, в случае тензорного представления НД можно утверждать следующее:

- 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 (дендрограмма) может быть учет внутри- и межкластерных расстояний в ультраметрическом пространстве.

Выводы

1. Иерархическую кластеризацию нечетких данных, представленных в виде нечеткого множества (совокупности упорядоченных пар „значение/функция принадлежности") целесообразно выполнять путем представления объектов кластеризации – нечетких переменных – в виде тензор-переменных Т- и t-типов. Тензор-переменная Т-типа формируется как результат тензорного (Кронекерова) произведения "значение ⊗ функция принадлежности" с матрицей n×n, тензор-переменная t-типа формируется как мно-

гомерный массив с матрицей 2×n (n – количество упорядоченных пар НМ).

2. При использовании тензорных моделей нечетких данных для объективности иерархической кластеризации необходимо формирование тензорной модели для всего универсального множества, на котором определена нечеткая переменная. Неучет универсального множества приводит к нечетким множествам разной размерности, что затрудняет процедуру оценки их близости, приводит к "усеченным" тензорным моделям, что не позволят получить на одной и той же выборке одинаковые дендрограммы.

3. Общая система признаков для иерархической кластеризации нечетких данных включает объединение векторов (одномерных массивов) значений и функций принадлежности нечетких данных. Для тензорных моделей нечетких данных в качестве системы признаков следует выбирать: в случае t-способа представления НД-сингулярные числа сингулярного разложения тензор-переменной.

4. В качестве мер близости объектов (нечетких данных) целесообразно использовать величину Фробениусовского расстояния между матрицами тензор-переменных; основным способом формализации представлений об эквивалентности объектов, составляющих отдельный кластер или их объединение (дендрограмма) может быть учет внутри межкластерных расстояний в ультраметрическом пространстве.

Список литературы

1. Воронцов К. В. Лекции по алгоритмам кластеризации и многомерного шкалирования / Интернет-ресурс. – Режим доступа: 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, Т.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, т. 194, №6 . – С. 147-160

6. Кофман А. Введение в теорию нечетких множеств: Пер. с франц. – М.: Радио и связь, 1982. – 432 с.

7. Carlsson G. and M´emoli F. Characterization, stability and convergence of hierarchical clustering algorithms. Technical report, 2009.

8. Carlsson G., M´emoli F. Characterization, Stability and Convergence of Hierarchical Clustering Methods. Journal of Machine Learning Research 11 (2010) P. 1425-1470.

9. Gol M. G., Yazdi H.S. A New Hierarchical Clustering Algorithm on Fuzzy Data (FHCA).- International Journal of Computer and Electrical Engineering, Vol. 2, No. 1, February, 2010. H.1793-1816.

10. Минаев Ю.Н., Филимонова О.Ю.Нечеткая математика на основе тензорных моделей неопределенности. Часть 1 – тензор-переменная в системе нечетких множеств. Электр. моделир., № 1, т.30 , 2008. – С. 43-59; часть 2-нечеткая математика в тензорном базисе. – Электр. моделир., № 2, Т. 30, 2008. – С. 4-21.

11. Murtagh F. Symmetry in Data Mining and Analysis: A Unifying View based on Hierarchy.arXiv:50805. 2744v1 [stat.ML] 18 May 2008. – P. 33.

12. Murtagh F., Downs G., and Contreras P. Hierarchical clustering of massive, high dimensional data sets by exploiting ultrametric embedding. SIAM Journal on Scientific Computing, 2007. In press Интернет-ресурс. – Режим доступа: http://

13. Gouvea F.Q. P-Adic Numbers: An Introduction. Springer, 2003. – 208 p.

14. Schikhof W.H. Ultrametric calculus. An itroduction to p-adic analysis. Cambridge University Press, 1984. – 306 p.