e-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2018. № 3
p-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2018. № 3

УДК 004.023, 519.237

ВИКОРИСТАННЯ ДОВЖИННОЇ МІРИ ПОДІБНОСТІ В ЗАДАЧАХ КЛАСТЕРИЗАЦІЇ

Кондрук Н. Е. – канд. техн. наук, доцент, доцент кафедри кібернетики і прикладної математики Ужгородського національного університету, Ужгород, Україна.

АНОТАЦІЯ

Актуальність. Дослідження присвячено розробці гнучкого математичного апарату, який мав би досить широкий спектр засобів для групування об'єктів за різними видами мір подібності. Це дає можливість в межах розробленого підходу ефективно розв'язувати достатньо широкі класи прикладних задач із різних предметних областей та проводити розбиття об'єктів кластерами різних геометричних форм.

Метою дослідження є підвищення ефективності розв'язання прикладних задач кластеризації шляхом використання довжинної міри подібності векторних ознак об'єктів.

Методи. Описано нечітке бінарне відношення та його функцію належності, що характеризує схожість об'єктів за довжинною мірою подібності їх векторних ознак. Модифіковано метод однорівневої кластеризації, заснований на нечітких бінарних відношеннях для використання довжинної міри подібності. При цьому задаються певні величини – пороги кластеризації, що характеризують ступінь подібності об'єктів в середині кластеру. Змінюючи пороги кластеризації можна проаналізувати динаміку формування кластерів, дослідити їх структуру та взаємозв'язки між об'єктами, визначити граничні об'єкти, зробити грунтовніший аналіз отриманих результатів. Запропонований підхід не потребує попереднього визначення кількості кластерів та дозволяє проводити кластеризацію даних концентричними сферами в умовах відсутності додаткової апріорної інформації, тому може використовуватись і на етапі попереднього аналізу даних.

Результати. Розроблений підхід реалізовано у вигляді програмної системи, на основі якої розв'язано актуальну прикладну задачу дослідження інтенсивності міграційного руху населення за регіонами України.

Висновки. Проведені експериментальні дослідження показали зручність та ефективність використання довжинної міри подібності при розв'язанні прикладних задач, що потребують групування кластерами у вигляді концентричних сфер. Представлений підхід забезпечив можливість проводити нові змістовні дослідження вхідних даних. Перспективи подальших досліджень полягають у розробці системи підтримки прийняття рішень, для розв'язання задач групування об'єктів на кластери концентричними сферами, конусами, еліпсами та їх перетинами; реалізації паралельної багаторівневої кластеризації проведеної одночасно за декількома критеріями подібності об'єктів та її застосуванні; дослідженні розбиттів об'єктів різними геометричними формами кластерів для однієї вибірки вхідних даних та проведенні змістовної інтерпретації отриманих результатів.

КЛЮЧОВІ СЛОВА: нечітка кластеризація, кластер, міра подібності, автоматичне групування об'єктів, кластеризація.

АБРЕВІАТУРИ

FCM – Fuzzy C-means clustering Algorithm;
BSA – Backtracking Search optimization Algorithm;
DTW – Dynamic Time Warping distance;
SVNS – Single-valued neutrosophic sets.

НОМЕНКЛАТУРА

C – множина векторів ознак об'єктів кластеризації;

$\left|\overline{c_i}\right|$ – довжина вектора $\overline{c_i}$;

$\overline{c_i}\left(c_1^i, c_2^i, \ldots, c_n^i\right)$ – вектор ознак об'єктів кластеризації;

$\overline{c_l^*}$ – вектор-представник l-го кластера;

K^i – результуючий i-й чіткий кластер;

z – кількість утворених чітких кластерів;

$\widetilde{K}^j$ – фазифікований кластер K^j;

m – кількість об'єктів кластеризації;

n – кількість ознак об'єктів кластеризації;

O_i – i-й об'єкт кластеризації;

R – нечітке бінарне відношення;

R^D – нечітке бінарне відношення, що характеризує різницю довжин векторів ознак;

R^K – нечітке бінарне відношення, що характеризує кут між векторами ознак;

R^V – нечітке бінарне відношення, що характеризує відстань між векторами ознак;

U^l – проміжковий кластер l-го кроку;

$\|\ \|$ – потужність множини;

β – коефіцієнт розтягу;

$\widetilde{\mu}_j$ – функція належності до нечіткого кластеру $\widetilde{K}^j$;

$\mu_R(\ )$ – функція належності нечіткого бінарного відношення R;

$\mu^*_{R^D}$ – поріг кластеризації при використанні довжинної міри подібності.

ВСТУП

Кластерний аналіз є потужним інструментом інтелектуального аналізу даних, коли відсутня апріорна інформація про групування об'єктів. У зв'язку із швидкою динамікою змін в соціально-економічному, соціально-міграційному та науково-виробничому середовищі кластерний аналіз є актуальним в різних прикладних сферах і предметних областях, зокрема: при дослідженні міграційних показників; формуванні споживчого кошика; прийнятті рішення про надання

DOI 10.15588/1607-3274-2018-3-11

споживчого кредиту; виявленні потенційних хвороб пацієнтів; побудові показово-репрезентативних вибірок, тощо.

Отже, потреба в кластеризації виникає в тих областях діяльності, де є необхідність розділити об'єкти на підмножини, так щоб кожний кластер складався із схожих об'єктів за певними ознаками. Чіткий поділ на кластери можливий тільки в ідеальних умовах і при значно відмінних ознаках об'єктів кластеризації. Тому для вирішення реальних завдань все частіше застосовуються нечіткі методи, в яких розбиття об'єктів проводиться із визначенням ступеня належності об'єктів кластерам. Це дає додаткову можливість для проведення ґрунтовнішого аналізу отриманих результатів.

Крім того, більшість розроблених методів кластеризації забезпечують групування об'єктів лише за одним критерієм подібності визначеним деякою метрикою відстані. При цьому утворюються кластери тільки еліпсоїдної форми. Але існує велика кількість прикладних задач, де такий вид групування об'єктів є неадекватним поставленій меті і неефективним.

Таким чином, доцільною є розробка гнучкого математичного апарату, який мав би досить широкий спектр засобів для групування об'єктів за різними геометричними формами кластерів. Дану властивість може забезпечувати зміна міри подібності в одноріневому методі кластеризації, заснованому на нечітких бінарних відношеннях [1]. Це дає можливість в межах розробленого підходу ефективно розв'язувати достатньо широкі класи прикладних задач із різних предметних областей. Так еліпсоїдній кластеризації та використанню відстаневої міри подібності присвячена праця [1]. Кутову міру подібності, конусну кластеризацію та відповідні прикладні задачі представлено в [2–4]. Дана робота є продовженням цих досліджень і присвячена використанню довжинної міри подібності при нечіткій кластеризації, яка забезпечує групування об'єктів концентричними сферами та її застосуванню.

Отже, метою дослідження є підвищення ефективності розв'язання прикладних задач кластеризації шляхом використання довжинної міри подібності векторних ознак об'єктів.

Для досягнення мети в роботі необхідно розв'язати наступні задачі:

– описати нечітке бінарне відношення та його функцію належності, що характеризує схожість об'єктів за довжинною мірою подібності їх векторних ознак;

– модифікувати метод однорівневої кластеризації, заснований на нечітких бінарних відношеннях [1] для довжинної міри подібності та проведення кластеризації концентричними сферами;

– проілюструвати використання довжинної міри подібності для розв'язання реальної прикладної задачі.

1 ПОСТАНОВКА ЗАДАЧІ

Відсутність єдиного загальноприйнятого формулювання нечіткої модифікації задачі кластерного аналізу потребує чіткий змістовний опис досліджуваної проблеми.

Розглянемо загальну задачу нечіткого кластерного аналізу в наступній постановці.

Нехай дано деякі об'єкти $O_1,\ldots,O_m$, які характеризуються n кількісними ознаками. Кожному об'єкту $O_i, i=\overline{1,m}$ однозначно ставиться у відповідність вектор ознак $\overline{c_i}\left(c_1^i,c_2^i,\ldots,c_n^i\right),\ i=\overline{1,m}$.

Потрібно розбити задані об'єкти $O_i, i=\overline{1,m}$ на однорідні групи «схожості» (кластери) по всіх n ознаках за довжинною мірою подібності, причому визначити і міру їх належності до отриманих кластерів. Для цього, з математичної точки зору, потрібно розв'язати задачу нечіткої кластеризації векторів ознак $\overline{c_i}\left(c_i^1,c_i^2,\ldots,c_i^n\right),\ i=\overline{1,m}$.

2 ОГЛЯД ЛІТЕРАТУРИ

В наш час математичний апарат нечіткої кластеризації бурхливо розвивається і забезпечує сучасні засоби ефективного розв'язання багатьох прикладних задач. Так в [5–8] описано огляд та порівняння чітких (жорстких) та нечітких (м'яких) базових методів кластеризації. Але різна прикладна природа вхідних даних, цілей, видів кластеризації приводить до принципової неможливості побудови одного єдиного ефективного універсального методу групування. Це приводить до необхідності створення нових методів або модифікації вже існуючих при розв'язанні прикладних задач. Так в роботі [9] модифіковано класичний нечіткий c-means (FCM) алгоритм для кластеризації сенсорних вузлів бездротових сенсорних мереж та використано правило нечіткого виводу Сугено для визначення їх представників. У [10] представлено новий метод кластеризації зображень на основі комбінації FCM та BSA алгоритму.

Як зазначалось, використання відстаневих метрик (Евкліда, Махаланобіса, манхеттенської та ін.) для визначення подібності об'єктів закладено в основі більшості базових методів кластеризації. Але існує цілий ряд практичних задач де їх використання приводить до невідповідності отриманих результатів поставленим цілям та завданням кластеризації. Так в роботі [11] обґрунтовано необхідність та доцільність використання динамічної відстані часу (DTW) при кластеризації часових рядів для отримання адекватних результатів групування та запропоновано три альтернативні методи нечіткої кластеризації на її основі. В дослідженні [12] представлено модифікацію нечіткого c-means (FCM) методу із використанням ядро-індукованої відстаневої міри в задачах сегментації зображень. В [13] для вирішення проблеми вибору постачальника пропонується ієрархічний метод кластеризації, оснований на новій формулі неевклідової відстані. В [14] розроблено алгоритм кластеризації за відстанню, основаною на

DOI 10.15588/1607-3274-2018-3-11

мірі подібності між однозначними нейрософськими множинами (SVNS).

Отже, існування великої кількості таких досліджень ще раз підтверджує той факт, що специфіка прикладних задач робить неможливим автоматичне перенесення методів в іншу прикладну область без ризику свідомо отримати неякісний розв'язок. Таким чином, доцільним є розробка та розвиток математичного апарату, який передбачає можливість проведення кластеризації за якісно різними критеріями подібності об'єктів. Це, в свою чергу, дозволить проводити групування об'єктів різними геометричними формами кластерів. Зокрема, використання довжинної міри подібності забезпечить проведення кластеризації концентричними сферами та дозволить ефективно розв'язувати ширше коло прикладних задач.

3 МАТЕРІАЛИ І МЕТОДИ

В залежності від цілей кластеризації геометричні форми потрібних кластерів можуть бути різними. Крім того, одну і ту ж множину даних можна розбивати на різні види кластерів та отримувати при цьому різну змістовну інтерпретацію результатів. Дослідження [1–3] показали, що гнучким та ефективним апаратом для проведення еліптичної та конічної кластеризації є однорівневий метод п. 6 в [1]. При цьому, подібність об'єктів за деяким критерієм характеризується нечітким бінарним відношенням R на множині векторних ознак $C = \left\{\overline{c_i} \middle| i = \overline{1,m}\right\}$ із функцією належності $\mu_R\left(\overline{c_i}, \overline{c_j}\right)$, де $\mu_R : C^2 \to [0,1]$. Чим більше значення величини $\mu_R\left(\overline{c_i}, \overline{c_j}\right)$ близьке до 1, тим в більшому ступені об'єкти O_i та O_j будуть подібними за цим критерієм. Зокрема, якісна зміна виду міри подібності об'єктів призводить до зміни геометричної форми кластерів.

Так для утворення еліптично подібних кластерів зручно та ефективно користуватись мірою подібності «відстань», що описується нечітким бінарним відношенням R^V [1].

Нечітке бінарне відношення R^K [2, 3] характеризує кут відхилення між векторами ознак $\overline{c_i}$ і $\overline{c_j}$. Його використання дало можливість проводити кластеризацію конічними кластерами.

«Довжинну» міру подібності, що дозволяє розбивати вектори ознак об'єктів на кластери концентричними сферами, пропонується описати бінарним відношенням R^D із функцією належності $\mu_{R^D} : C^2 \to \left[\frac{1}{e}, 1\right]$ типу:

$$\mu_{R^D}\left(\overline{c_i}, \overline{c_j}\right) = e^{-\frac{\left|\left|\overline{c_i}\right| - \left|\overline{c_j}\right|\right|}{\Delta}}, \qquad (1)$$

де $\Delta = \max_i \left|\overline{c_i}\right| - \min_j \left|\overline{c_j}\right|$, $i = \overline{1,m}$, $j = \overline{1,m}$.

Використання експоненціальної функції належності такого виду не є випадковим. Аргументом експоненти є пронормована величина, що змінюється від 0 до 1. Тому її значеннями будуть відповідно величини від 1 до $\frac{1}{e}$. Причому, меншій різниці довжин векторів ознак об'єктів буде відповідати ближче до 1 значення μ_{R^D}. Ця властивість визначає той факт, що нечітке бінарне відношення R^D характеризує схожість векторів $\overline{c_i}$ і $\overline{c_j}$ за довжинами.

Проведемо модифікацію чіткого методу однорівневої кластеризації п. 6 в [1] для використання довжинної міри подібності об'єктів.

Нехай задана числова величина $\mu^*_{R^D} \in [0; 1]$ – поріг кластеризації. Він характеризує необхідну ступінь подібності об'єктів в межах одного кластеру. Якщо $\mu^*_{R^D} = 0$, то ступінь подібності об'єктів буде найслабшою, що приведе до формування одного кластеру сферичного виду, куди увійдуть всі об'єкти. Якщо ж $\mu^*_{R^D} = 1$, тоді, навпаки, об'єкти із різною довжиною векторів ознак сформують окремі кластери, бо ступінь подібності об'єктів буде найвищою. Отже, ближчому значенню $\mu^*_{R^D}$ до одиниці буде відповідати більша кількість сформованих кластерів.

Проведення практичних експериментів показало, що «хороша» чутливість функції типу (1) в околі свого граничного значення ($sup\,\mu_{R^D} = 1$) дозволяє проводити кластеризацію об'єктів для всіх можливих величин порогів проміжку [0; 1] із певною точністю (наприклад, із точністю 0,01). Це забезпечує можливість проводити дослідження всієї динаміки зміни кластерів та їх структури.

Приймається евристика: на основі двох «найбільш схожих» за довжинною мірою подібності незгрупованих об'єктів має формуватись новий кластер.

Далі покроково описано внесені зміни в l-у ітерацію роботи чіткого методу однорівневої кластеризації об'єктів [1] для його адаптації до використання довжинної міри подібності.

Крок 1 залишається без змін. Слід зауважити, що довжина обраного домінантного вектора-центроїда $\overline{c_l^*}$ із множини $\left\{\overline{c_i} \middle| i \in \Omega_l\right\}$ буде визначати радіус сфери, навколо якої буде формуватись l-й кластер.

При проведенні процедури центрування кластеру U^l кроків 2 та 3 довжина вектора-центроїда уточнюється, за формулою:

$$\overline{c_l^*} := \frac{\sum_{c_i \in U^l} \left|\overline{c_i}\right|}{\left\|U^l\right\|} \cdot \overline{c_l^*}. \qquad (2)$$

Якщо потрібна інформація не тільки про розподіл об'єктів по кластерам, а й про ступінь їх приналежно-

DOI 10.15588/1607-3274-2018-3-11

сті кожній із множин, то необхідно провести процедуру фазифікації.

За чітким однорівневим методом п. 6 в [1], та описаними модифікаціями проводиться кластеризація на чіткі кластери K^1, K^2, ..., K^z, $z \le m$ із відповідними представниками $\overline{c_1^*}, \overline{c_2^*}, \ldots, \overline{c_z^*}$ знайденими за формулою (2). Функції належності $\tilde{\mu}_j : C \to [0,1]$ фазифікованих кластерів $\tilde{K}^j$, $j = \overline{1,z}$ пропонується визначати за формулою:

$$\tilde{\mu}_j\left(\overline{c_i}\right) = \mu_{R^D}\left(\overline{c_i}, \overline{c_j^*}\right), \qquad (3)$$

або за формулою:

$$\tilde{\mu}_j\left(\overline{c_i}\right) = \exp\left(-\frac{\left[1-\mu_{R^D}\left(\overline{c_i}, \overline{c_j^*}\right)\right]^2}{\beta}\right). \qquad (4)$$

Зокрема, розрахований коефіцієнт β=0,0882 за правилом трьох сігм.

Використання функцій належності типу (1) та застосування формули (3) не є загальноприйнятим для фазифікації даних, так як $\tilde{\mu}_j : C \to \left[\frac{1}{e};1\right]$, тобто міра належності найвіддаленіших об'єктів до *j*-го кластеру буде не менша, як число $\frac{1}{e}$. Але, в цьому випадку, можна зробити якісний аналіз отриманих результатів кластеризації та числових значень $\tilde{\mu}_j$ згідно шкали бажаності Харрінгтона [1]. Використання функції належності гаусівського типу (4) приводить до її нормалізації, тобто $\tilde{\mu}_j : C \to (0;1]$.

4 ЕКСПЕРИМЕНТИ

Для проведення експериментів була розроблена комп'ютерна програма, що реалізує запропонований підхід при кластеризації об'єктів за довжинною мірою подібності. Вона є доповненням до вже існуючого програмного забезпечення для проведення еліпсоїдної [1] та конусної [2, 3] кластеризації і забезпечує розбиття об'єктів концентричними сферами.

Вхідною інформацією для проведення групування об'єктів є числові величини *n*, *m*, $\mu^*_{R^D}$ та координати векторів $\overline{c_i}$. Далі для чіткої кластеризації застосовується метод однорівневої кластеризації із п. 6 [1], адаптований до використання довжинної міри подібності, що описана формулою (1). Фазифікація кластерів проводиться за формулами (3) та (4).

Враховуючи те, що якість машинної кластеризації визначається її відповідністю класифікації, що проведена людиною, верифікацію розробленого підходу буде проведено на прикладній задачі кластеризації у двовимірному просторі. Це дасть додаткову візуальну можливість оцінити отриманий результат.

Все більш глобальними стають проблеми зростаючих масштабів міграції населення України та її регулювання, модернізації міграційної політики держави. Вони потребують високий та сучасний рівень наукових засобів вивчення міграційних процесів населення в сучасних економічних, суспільних та політичних умовах. Зокрема, системний підхід до аналізу міграційних явищ передбачає дослідження інтенсивності міграційного руху населення.

Отже, пропонується, розглянути актуальну задачу дослідження інтенсивності міграційного руху населення регіонів України, наприклад, за січень-листопад 2017 року. Вхідні дані отримані із офіційного сайту Держкомстату України, представлені в наступній таблиці.

Таблиця 1 – Міграційний рух населення у січні-листопаді 2017 року

№	Регіони України	Кількість прибулих осіб	Кількість вибулих осіб
1	Вінницька	5454	9816
2	Волинська	7608	8405
3	Дніпропетровська	48663	24741
4	Донецька	6686	30039
5	Житомирська	12857	13649
6	Закарпатська	5271	5451
7	Запорізька	6411	9215
8	Івано-Франківська	13963	12588
9	Київська	50611	22098
10	Кіровоградська	9847	11395
11	Луганська	2203	20918
12	Львівська	26483	24031
13	Миколаївська	7215	8939
14	Одеська	20712	17430
15	Полтавська	18137	18962
16	Рівненська	13909	15591
17	Сумська	15207	15957
18	Тернопільська	5889	7384
19	Харківська	52007	40497
20	Херсонська	4091	6673
21	Хмельницька	6368	9060
22	Черкаська	14720	15093
23	Чернівецька	4828	4787
24	Чернігівська	7485	9499
25	м.Київ	32363	28881

Вхідною інформацією для експериментальних досліджень групування регіонів України були пронормовані дані табл. 1 та різні пороги кластеризації.

5 РЕЗУЛЬТАТИ

Згідно шкали бажаності Харрінгтона та проведених попередніх практичних досліджень [1] найбільш

DOI 10.15588/1607-3274-2018-3-11

значимими та змістовними при розв'язанні практичних задач виявились величини порогів кластеризації близькі до 0,8. Отримані фрагментарні результати чіткої кластеризаці за довжинною мірою подібності із функцією належності (1) представлено в табл. 2 та рис. 1.

Таблиця 2 – Фрагментарні результати чіткої кластеризації розглядуваної сукупності

Числові значення порогу кластеризації	Результати кластеризації
$\mu^*_{R^D} \in [0{,}64;\ 0{,}84]$	Кластер 1: об'єкти з номерами 1, 2, 5, 6–8, 10, 11, 13, 16–18, 20–24;
	Кластер 2: об'єкти з номерами 4, 12, 14, 15;
	Кластер 3: об'єкти з номерами 3, 9, 19, 25.
$\mu^*_{R^D} \in [0{,}85;\ 0{,}86]$	Кластер 1: об'єкти з номерами 1, 2, 5, 6-8, 10, 11, 13, 16, 18, 20-24;
	Кластер 2: об'єкти з номерами 4, 12, 14, 15, 17;
	Кластер 3: об'єкт з номером 25;
	Кластер 4: об'єкти з номерами 3, 9;
	Кластер 5: об'єкт з номером 19.

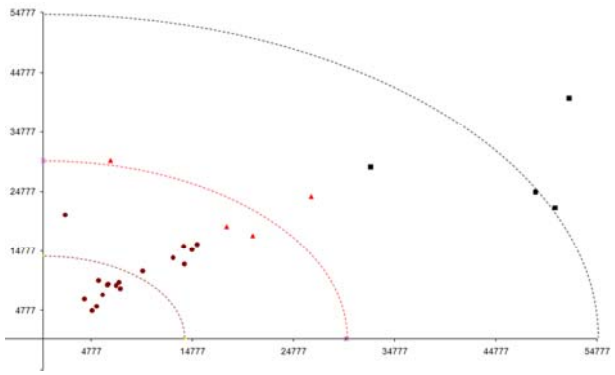

Рисунок 1 – Геометрична інтерпретація результатів чіткої кластеризації за довжинною мірою подібності при

$$\mu^*_{R^D} \in [0{,}64;\ 0{,}84]$$

На рис. 1 пунктирними лініями позначено дуги концентричних кіл, що відповідають представникам відповідних кластерів.

Для інтерпретації нечітких результатів кластеризації було використано матриці нечіткого розподілу об'єктів по кластерах (табл. 3). Проаналізуємо, наприклад, матрицю відповідного нечіткого розбиття (табл. 3) при $\mu^*_{R^D} \in [0{,}64;\ 0{,}84]$ та його лінійні діаграми (рис. 1–2).

Для візуального представлення нечітких результатів побудована лінійна діаграма нечітких розбиттів. По осі ординат діаграми відкладаються значення ступенів належності, а по осі абсцис – номери об'єктів. Належність об'єктів кластеру визначається точкою перетину ліній, що відповідає номеру об'єкта та ступеня належності об'єкта кластеру. Номер кластера вказується поряд з точкою.

Таблиця 3 – Матриці нечіткого розбиття досліджуваної сукупності об'єктів при $\mu^*_{R^D} \in [0{,}64;\ 0{,}84]$

№ об'єкта	Міри належності виду (3)			Міри належності виду (4)		
	Кластер 1	Кластер 2	Кластер 3	Кластер 1	Кластер 2	Кластер 3
1	0,98	0,43	0,05	0,95	0,73	0,48
2	0,98	0,44	0,05	0,96	0,73	0,48
3	0,06	0,27	1,00	0,50	0,66	1,00
4	0,50	1,00	0,28	0,75	0,99	0,67
5	0,93	0,71	0,09	0,92	0,83	0,54
6	0,89	0,32	0,03	0,90	0,68	0,45
7	0,98	0,43	0,05	0,95	0,73	0,48
8	0,93	0,71	0,09	0,92	0,83	0,54
9	0,06	0,25	1,00	0,50	0,65	0,99
10	1,00	0,57	0,07	0,98	0,78	0,51
11	0,87	0,80	0,12	0,89	0,86	0,57
12	0,34	0,91	0,42	0,69	0,91	0,72
13	0,98	0,44	0,05	0,96	0,73	0,48
14	0,64	0,97	0,20	0,80	0,95	0,63
15	0,67	0,96	0,19	0,81	0,94	0,62
16	0,87	0,80	0,12	0,89	0,86	0,56
17	0,83	0,84	0,13	0,87	0,87	0,57
18	0,94	0,38	0,04	0,93	0,71	0,46
19	0,02	0,09	0,72	0,42	0,54	0,83
20	0,89	0,33	0,03	0,90	0,69	0,45
21	0,97	0,43	0,05	0,95	0,73	0,48
22	0,86	0,80	0,12	0,89	0,86	0,57
23	0,86	0,30	0,03	0,89	0,68	0,44
24	0,99	0,46	0,05	0,97	0,74	0,49
25	0,18	0,63	0,71	0,61	0,80	0,82

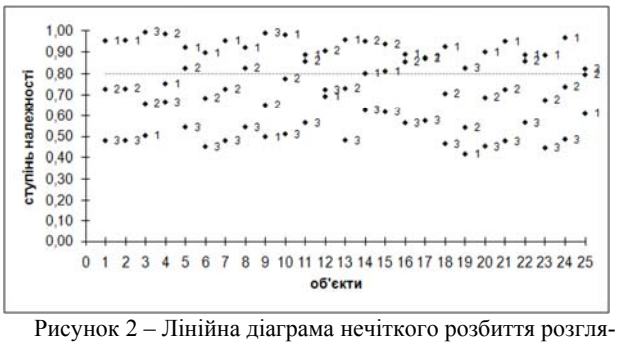

Рисунок 2 – Лінійна діаграма нечіткого розбиття розглядуваної сукупності із мірою належності, визначеною за формулою (3)

На рис. 2 пунктирною лінією представлено умовну межу ступеня значної (сильної) подібності об'єктів в межах одного кластеру. З діаграми рис. 2 видно, що об'єкт із номером 17 є граничним для кластерів 1 та 2. Тому при повторній кластеризації із збільшенням порогу

DOI 10.15588/1607-3274-2018-3-11

він може бути перехідним між цими кластерами (див. табл. 2). Об'єкти із номерами 5, 8, 11, 16, 22 хоча і віднесені до 1-го кластеру при чіткій кластеризації також мають високу ступінь подібності до об'єктів 2-го кластеру. Об'єкт із номером 25 3-го кластеру має сильну подібність до об'єктів 2-го кластеру, а 14 та 15 об'єкти із 2-го кластеру подібні до об'єктів 1-го кластеру.

Далі представлено отриману лінійну діаграму (рис. 3) при використанні формули нормування (4).

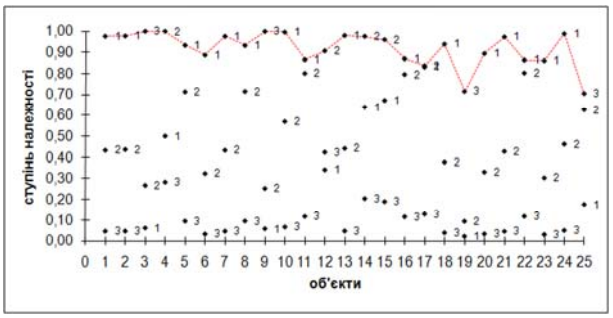

Рисунок 3 – Лінійна діаграма нечіткого розбиття розглядуваної сукупності із нормованою мірою належності, визначеною за формулою (4)

Як видно з діаграми формула (4) не змінює картину розбиття, а лише уточнює її. Пунктирною лінією виділено всі точки із максимальними значеннями функції належності. Об'єкти з номерами 25 та 19 мають найменші міри подібності до своїх чітких кластерів серед всіх виділених (подібність цих об'єктів є «найслабшою» в межах відповідного кластеру), тому при збільшенні порогу кластеризації об'єктів вони будуть ключовими при формуванні нових кластерів (див. табл. 2).

Отже, фазифікація чітких кластерів дає можливість провести додатковий аналіз взаємозв'язків між об'єктами, уточнює їх ступінь подібності та дає змогу визначити граничні (перехідні) об'єкти.

Згідно отриманих результатів можна зробити наступну змістовну інтерпретацію поставленої задачі:

– до регіонів України із пасивним міграційним рухом населення за січень-листопад 2017 року можна віднести: Вінницьку, Волинську, Житомирську, Закарпатську, Запорізьку, Івано-Франківську, Київську, Кіровоградську, Луганську, Миколаївську, Рівненську, Тернопільську, Херсонську, Хмельницьку, Черкаську, Чернівецьку, Чернігівську області;

– в Сумській області спостерігається граничний пасивно-посередній міграційний рух населення за січень-листопад 2017 року;

– до регіонів України із посереднім міграційним рухом населення за січень-листопад 2017 року можна віднести: Донецьку, Львівську, Одеську та Полтавську області;

– до регіонів України із посиленим міграційним рухом населення за січень-листопад 2017 року можна віднести: Дніпропетровську, Київську, Харківську області та м. Київ.

Перехідна динаміка міграційного руху спостерігається в:

– Сумській, Луганській, Рівненській, Черкаській, Житомирській та Івано-Франківській області від пасивного до посереднього темпу;

– Одеській та Полтавській області від посереднього до пасивного темпу;

– м. Київ від посиленого до посереднього темпу.

6 ОБГОВОРЕННЯ

Порівняння методів кластерного аналізу не є зовсім коректним бо не існує єдиного критерію оптимальності оцінки результатів кластеризації. Кожен із них має свої недоліки та переваги і може бути ефективнішим при розв'язанні певного класу задач.

Зокрема, проведені експериментальні дослідження показали зручність та ефективність методу одноріневої кластеризації п. 6 в [1] адаптованого до використання довжинної міри подібності для розв'язання деяких класів прикладних задач, коли відстаневі метрики не є коректними. При цьому можна визначити основні переваги запропонованого підходу:

– дає можливість проводити кластеризацію концентричними сферами та отримувати якісно нові змістовні результати;

– фазифікація чітких кластерів дозволяє визначати ступінь подібності об'єктів, виявляти граничні об'єкти, робити ґрунтовніший аналіз отриманих результатів;

– вибір різних порогів кластеризації дає додаткову можливість спостерігати за динамікою формування кластерів, зміною їх структури та виявляти приховані взаємозв'язки між об'єктами;

– може бути використаний як для попереднього аналізу даних, так і для проведення самої процедури кластеризації.

Використання нечіткого бінарного відношення R^D в методі одноріневої послідовної кластеризації [1] забезпечило можливість проводити нові змістовні дослідження вхідних даних.

Дана праця є продовженням та розвитком досліджень [1–3]. В подальшому передбачається розроблений підхід використати для:

– реалізації паралельної багаторівневої кластеризації одночасно проведеної за декількома критеріями подібності та її застосування;

– розробки системи підтримки прийняття рішень, що забезпечить групування об'єктів на кластери концентричними сферами, конусами, еліпсами та їх перетинами;

– дослідження використання кластеризації об'єктів за різними геометричними формами кластерів для однієї вибірки вхідних даних та проведення змістовної інтерпретації отриманих результатів.

ВИСНОВКИ

Вирішується проблема розвитку методів кластеризації, основаних на нечітких бінарних відношеннях для проведення розбиття об'єктів концентричними сферами.

DOI 10.15588/1607-3274-2018-3-11

Наукова новизна отриманих результатів полягає в тому, що описано нечітке бінарне відношення R^D та його функцію належності, які характеризують довжинну міру подібності векторних ознак об'єктів. Модифіковано метод однорівневої кластеризації [1] для використання довжинної міри подібності об'єктів. При цьому задаються певні величини – пороги кластеризації, що характеризують ступінь схожості об'єктів в середині кластеру. Змінюючи пороги кластеризації можна проаналізувати динаміку формування кластерів, дослідити їх структуру та взаємозв'язки між об'єктами. Запропонований підхід дозволяє проводити кластеризацію об'єктів концентричними сферами в умовах відсутності додаткової апріорної інформації, тому може використовуватись і на етапі попереднього аналізу даних.

Практичне значення отриманих результатів полягає в розробленому програмному забезпеченні, що реалізує представлений підхід. Проведення експериментів показало його ефективність при розв'язанні певних класів прикладних задач кластерного аналізу. Проілюстровано роботу нечіткого однорівневого методу, основаного на довжинній мірі подібності на реальній задачі дослідження інтенсивності міграційного руху населення регіонів України. Проведено аналіз та змістовну інтерпретацію отриманих результатів.

ПОДЯКИ

Роботу виконано в рамках держбюджетної науково-дослідної теми Ужгородського національного університету «Розробка математичних моделей і методів для оброблення інформації та інтелектуального аналізу даних» (номер державної реєстрації 0115U004630).

ЛІТЕРАТУРА/ ЛИТЕРАТУРА

1. Kondruk N. Clustering method based on fuzzy binary relation / N. Kondruk // Eastern-European Journal of Enterprise Technologies. – 2017. – No. 2(4). – P. 10–16. DOI: 10.15587/1729–4061.2017.94961
2. Кондрук Н. Е. Алгоритм кластеризації критеріального простору для задач вибору / Н. Е. Кондрук, М. М. Маляр //Вісник Київського університету. Серія: фіз.-мат. наук. – 2006. – Вип. 3. – С. 225–229.
3. Кондрук Н. Е. Деякі методи автоматичного групування об'єктів / Н. Е. Кондрук // Південно-Європейський журнал передових технологій. – 2014. – Т. 2, № 4 (68). – С. 20–24.
4. Кондрук Н. Е. Системи підтримки прийняття рішень для автоматизованого складання дієт / Н. Е. Кондрук // Управління розвитком складних систем. – 2015. – Вип. 23(1). – С. 110–114.
5. Peters G. Soft clustering–fuzzy and rough approaches and their extensions and derivatives / G. Peters // International Journal of Approximate Reasoning. – 2013. – Vol. 54, № 2. – P. 307–322. DOI: 10.1016/j.ijar.2012.10.003
6. Banu P. K. N. Performance analysis of hard and soft clustering approaches for gene expression data / P. K. N. Banu, S. Andrews //International Journal of Rough Sets and Data Analysis (IJRSDA). – 2015. – Vol. 2, № 1. – P. 58–69. DOI: 10.4018/ijrsda.2015010104
7. Bora D. J. Comparative study between fuzzy clustering algorithm and hard clustering algorithm / D. J. Bora, D. Gupta, A. A. Kumar // International Journal of Computer Trends and Technology (IJCTT). – 2014. – Vol. 10(2). – C. 108–113. DOI: 10.14445/22312803/IJCTT-V10P119
8. Jipkate B. R. A comparative analysis of fuzzy c-means clustering and k means clustering algorithms / B. R. Jipkate, V. V. Gohokar // International Journal Of Computational Engineering Research. – 2012. – Vol. 2. – № 3. – P. 737–739.
9. Shokouhifar M. Optimized sugeno fuzzy clustering algorithm for wireless sensor networks / M. Shokouhifar, A. Jalali //Engineering applications of artificial intelligence. – 2017. – Vol. 60. – P. 16–25. DOI: 10.1016/j.engappai.2017.01.007
10. Toz G. A Fuzzy Image Clustering Method Based on an Improved Backtracking Search Optimization Algorithm with an Inertia Weight Parameter / G. Toz, İ. Yücedağ, P. Erdoğmuş // Journal of King Saud University–Computer and Information Sciences. – 2018. In press. DOI: 10.1016/j.jksuci.2018.02.011
11. Izakian H. Fuzzy clustering of time series data using dynamic time warping distance / H. Izakian, W. Pedrycz, I. Jamal // Engineering Applications of Artificial Intelligence. – 2015. – Vol. 39. – P. 235–244. DOI: 10.1016/j.engappai.2014.12.015
12. Chen, S. Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure / S. Chen, D. Zhang //IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). – 2004. – Vol. 34, № 4. – P. 1907–1916. DOI: 10.1109/TSMCB.2004.831165
13. Heidarzade A. Supplier selection using a clustering method based on a new distance for interval type-2 fuzzy sets: A case study / A. Heidarzade, I. Mahdavi, N. Mahdavi-Amiri // Applied Soft Computing. – 2016. – Vol. 38. – P. 213–231. DOI: 10.1016/j.asoc.2015.09.029
14. Ye J. Clustering methods using distance-based similarity measures of single-valued neutrosophic sets / J. Ye // Journal of Intelligent Systems. – 2014. – Vol. 23, № 4. – P. 379–389. DOI: 10.1515/jisys-2013-0091

Стаття надійшла до редакції 18.02.2018.
Після доробки 02.04.2018.

УДК 004.023, 519.237

ИСПОЛЬЗОВАНИЕ ДЛИННОВОЙ МЕРЫ СХОДСТВА В ЗАДАЧАХ КЛАСТЕРИЗАЦИИ

Кондрук Н. Э. – канд. техн. наук, доцент, доцент кафедры кибернетики и прикладной математики Ужгородского национального университета, Ужгород, Украина.

АННОТАЦИЯ

Актуальность. Исследование посвящено разработке гибкого математического аппарата, который имеет достаточно широкий спектр средств для группировки объектов по различным видам мер сходства. Это даст возможность в рамках разработанного подхода эффективно решать достаточно широкие классы прикладных задач из разных предметных областей и проводить кластеризацию кластерами различных геометрических форм.

Целью исследования является повышение эффективности решения прикладных задач кластеризации путем использования длинновой меры сходства векторных признаков объектов.

Методы. Описано нечеткое бинарное отношение и его функцию принадлежности, характеризирующие подобие объектов по длинновой мере сходства их векторных признаков. Модифицировано метод одноуровневой кластеризации, основанный на нечетких бинарных отношениях для использования длинновой меры сходства. При этом задаются определенные величины - пороги кластеризации, характеризующие степень подобия объектов внутри кластера. Изменяя пороги кластеризации можно проанализировать динамику формирования кластеров, исследовать их структуру и взаимосвязи между объектами, определить предельные объекты, провести более глубокий анализ полученных результатов. Предложенный подход не требует предварительного определения коли-

DOI 10.15588/1607-3274-2018-3-11

чества кластеров и позволяет проводить кластеризацию данных концентрическими сферами в условиях отсутствия дополнительной априорной информации, поэтому может использоваться и на этапе предварительного анализа данных.

Результаты. Разработанный подход реализован в виде программной системы на основании которой решена актуальная прикладная задача исследования интенсивности миграционного движения населения по регионам Украины.

Выводы. Проведенные экспериментальные исследования показали удобство и эффективность использования длинновой меры сходства при решении прикладных задач, требующих группировки кластерами в виде концентрических сфер. Представленный подход обеспечил возможность проводить новые содержательные исследования входных данных. Перспективы дальнейших исследований заключаются в разработке системы поддержки принятия решений для решения задач группировки объектов на кластеры концентрическими сферами, конусами и эллипсами и их пересечениями; реализации параллельной многоуровневой кластеризации по различным критериям и ее применении; исследовании разбиения объектов разными геометрическими формами кластеров для одной выборки входных данных и проведении содержательной интерпретации полученных результатов.

КЛЮЧЕВЫЕ СЛОВА: нечеткая кластеризация, кластер, мера сходства, автоматическая группировка объектов, кластеризация.

UDC 004.023, 519.237

USE OF LENGTH-BASED SIMILARITY MEASURE IN CLUSTERING PROBLEMS

Kondruk N. E. – PhD, Associate Professor, Associate Professor of Department of Cybernetics and Applied Mathematics, Uzhgorod National University, Uzhgorod, Ukraine.

ABSTRACT

Context. The study is devoted to the development of a flexible mathematical apparatus, which should have a sufficiently wide range of means for grouping objects into different types of similarity measures. This makes it possible, within the framework of the developed approach, to efficiently solve sufficiently broad classes of applied problems from different subject areas and to partition objects with clusters of different geometric forms.

Objective. The aim of the study is improvement of the efficiency of solving cluster problems by applying a similar measure of the vector characteristics of objects.

Method. A fuzzy binary relation and its membership function describing the similarity of objects according to the level of similarity of their vector attributes are described. The method of single-level clustering, based on fuzzy binary relations for the use of a similarity measure, is modified. In this case, certain values are set – the thresholds of clusterization that characterize the similarity degree of objects within the cluster. By changing the thresholds of clusterization, one can analyze the dynamics of cluster formation, investigate their structure and interrelationships between objects, determine the ultimate objects, and make a thorough analysis of the obtained results. The proposed approach does not require a preliminary determination of the number of clusters and allows clustering of data in concentric spheres in the absence of additional a priori information, so it can be used at the stage of preliminary data analysis.

Results. The developed approach is implemented in the form of a software system on the basis of which the actual applied problem of investigating the intensity of population migration by regions of Ukraine is solved.

Conclusions. The conducted experimental researches show the convenience and efficiency of using the similarity measure for solving applied problems requiring clustering in the form of concentric spheres. The presented approach provides an opportunity to conduct new meaningful studies of input data. Prospects for further research are development of a decision support system, to solve the problems of grouping objects into clusters by concentric spheres, cones, ellipses and their intersections; implementation of parallel multi-level clustering carried out simultaneously by several criteria of similarity of objects and their application; study of the partitioning of objects by different geometric forms of clusters for a single sample of input data and carrying out a meaningful interpretation of the obtained results.

KEYWORDS: fuzzy clustering, cluster, measure of similarity, automatic grouping of objects, clustering.

REFERENCES

1. Kondruk N. Clustering method based on fuzzy binary relation, *Eastern-European Journal of Enterprise Technologies*, 2017, No. 2(4), pp. 10–16. DOI: 10.15587/1729-4061.2017.94961
2. Kondruk N. E., Malyar M. M. Algorytm klasteryzacii' kryterial'nogo prostoru dlja zadach vyboru, *Visnyk Kyi'vs'kogo universytetu*, 2006, Issue. 3, pp. 225–229.
3. Kondruk, N. E. Dejaki metody avtomatychnogo grupuvannja ob'jektiv, *Eastern-European Journal of Enterprise Technologies*, 2014, Vol. 2, No. 4 (68), pp. 20–24.
4. Kondruk N. E. Systemy pidtrymky pryjnjattja rishen' dlja avtomatyzovanogo skladannja dijet, *Management of Development of Complex Systems*, 2015, Issue. 23(1), pp. 110–114.
5. Peters, G. Soft clustering-fuzzy and rough approaches and their extensions and derivatives, *International Journal of Approximate Reasoning*, 2013, Vol. 54, No. 2, pp. 307–322. DOI: 10.1016/j.ijar.2012.10.003
6. Banu P. K. N., Andrews S. Performance analysis of hard and soft clustering approaches for gene expression data, *International Journal of Rough Sets and Data Analysis (IJRSDA)*, 2015, Vol. 2, No. 1, pp. 58–69. DOI: 10.4018/ijrsda.2015010104
7. Bora D. J., Gupta D., Kumar A. A. Comparative study between fuzzy clustering algorithm and hard clustering algorithm, *International Journal of Computer Trends and Technology (IJCTT)*, 2014, Vol. 10(2), pp. 108–113. DOI: 10.14445/22312803/IJCTT-V10P119
8. Jipkate B. R., Gohokar V. V. A comparative analysis of fuzzy c-means clustering and k means clustering algorithms, *International Journal Of Computational Engineering Research*, 2012, Vol. 2, No. 3, pp. 737–739.
9. Shokouhifar M., Jalali A. Optimized sugeno fuzzy clustering algorithm for wireless sensor networks, *Engineering applications of artificial intelligence*, 2017, Vol. 60, pp. 16–25. DOI: 10.1016/j.engappai.2017.01.007
10. Toz G., Yücedağ İ., Erdoğmuş P. A Fuzzy Image Clustering Method Based on an Improved Backtracking Search Optimization Algorithm with an Inertia Weight Parameter, *Journal of King Saud University-Computer and Information Sciences*, 2018. In press. DOI: 10.1016/j.jksuci.2018.02.011
11. Izakian H., Pedrycz W., Jamal I. Fuzzy clustering of time series data using dynamic time warping distance, *Engineering Applications of Artificial Intelligence*, 2015, Vol. 39, pp. 235–244. DOI: 10.1016/j.engappai.2014.12.015
12. Chen S., Zhang D. Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2004, Vol. 34, No. 4, pp. 1907–1916. DOI: 10.1109/TSMCB.2004.831165
13. Heidarzade A., Mahdavi I., Mahdavi-Amiri N. Supplier selection using a clustering method based on a new distance for interval type-2 fuzzy sets: A case study, *Applied Soft Computing*, 2016, Vol. 38, pp. 213–231. DOI: 10.1016/j.asoc.2015.09.029
14. Ye J. Clustering methods using distance-based similarity measures of single-valued neutrosophic sets, *Journal of Intelligent Systems*, 2014, Vol. 23, No. 4, pp. 379–389. DOI: 10.1515/jisys-2013-0091

DOI 10.15588/1607-3274-2018-3-11