УДК 004.8

ВИКОРИСТАННЯ ІНТЕРВАЛЬНИХ ФУНКЦІЙ НАЛЕЖНОСТІ В ЗАДАЧАХ КЛАСТЕРИЗАЦІЇ ДАНИХ СОЦІАЛЬНОГО ХАРАКТЕРУ

Н.Р. КОНДРАТЕНКО, О.О. СНІГУР

Розглянуто вплив рівня нечіткості на результати нечіткого кластерного аналізу. Запропоновано підхід до розв'язання задачі кластеризації на основі інтервальних нечітких множин типу 2 із застосуванням індексу вірогідності Квона. Роботу методу продемонстровано на прикладі кластеризації країн світу за рівнем розвитку.

ВСТУП

Зараз у сучасній Україні та в усьому світі посилюється значення наукового аналізу проблем соціального характеру, зокрема співвідношення рівня життя різних верств населення, питання гендерної нерівності, диференціації країн та регіонів на основі технічного, соціально-економічного, інтелектуального, природного факторів тощо.

Багатомірність явищ, які розглядаються, ставить особливі вимоги до математичних методів розв'язання цих задач. Передумовою побудови достовірних математико-статистичних моделей у таких умовах є виявлення в даних компактних однорідних сукупностей, існування яких можна приписати об'єктивно існуючим суспільним закономірностям. Одним із методів, що дозволяють виявляти такі сукупності, використовуючи широке коло показників, є кластерний аналіз. Він є найпотужнішим інструментом для проведення багатомірних досліджень. Його застосування в таких задачах є цілком виправданим, оскільки вперше кластерний аналіз застосували саме в соціології [1]. Для здійснення процедури кластеризації не потрібно апріорних знань про розподіл генеральної сукупності. Велика її перевага полягає в тому, що вона дозволяє робити розбиття об'єктів не за одним параметром, а за цілим набором ознак. Крім того, кластерный аналіз, на відміну від більшості математико-статистичних методів, не накладає жодних обмежень на вид об'єктів, що розглядаються, і дозволяє оперувати множиною вихідних даних практично довільної природи [2]. Це дає змогу говорити про можливість створення методів кластеризації, придатних для розв'язання практично будь-яких соціально-економічних задач, а не лише задач певного класу.

Про актуальність розв'язання задач кластеризації, орієнтованих на соціальні дані, свідчить велика кількість праць із цієї тематики. Зокрема, у ро-

ботах [1, 3] здійснено спроби розв'язання задач регіонального районування та соціально-економічного прогнозування. Проте математичні методи, що лежать в основі цих досліджень, суттєво обмежені припущенням, що вхідні дані є абсолютно точними, правдивими та незашумленими. Метод, який запропоновано в роботі [4], попри високі оптимізаційні властивості, ставить аналогічну вимогу. Відомо, що на практиці такі умови трапляються вкрай рідко, особливо в галузі соціології, усі показники якої ґрунтуються на результатах соціологічних опитувань та офіційних даних, що надані різного роду урядовими організаціями. Стовідсоткової достовірності таких даних гарантувати ніхто не може, тому ця задача вимагає методів кластерного аналізу, стійких до викидів та шуму. Один із таких методів — метод PCM (Possibilistic C-Means — можливісних С-середніх) — запропоновано в роботі [5]. Він надзвичайно стійкий до шумів у вхідних показниках, але ґрунтується на нечітких множинах типу 1. Це не дає змогу дати повністю адекватну оцінку досліджуваній множині даних, оскільки крім точок, що вносять шум, у характеристиках кожної точки закладено певну невизначеність, яка не може не перенестись на результат кластеризації. При цьому характеризувати ступінь належності точки до кластеру одним числом недостатньо. Унаслідок дії невизначеностей саме це число також трансформується в нечітку множину, що веде до необхідності оперування нечіткими множинами типу 2. Ідея нечіткої множини типу 2 як поглиблення та узагальнення множини першого типу належить Л. Заде [6]. Узагальнена нечітка множина вимагає задання великої кількості параметрів, що не завжди має практичний сенс. Тому часто обмежуються використанням інтервальних функцій належності [7, 8, 9]. На сьогодні такий підхід застосовується у великій кількості різних задач: класифікації образів [10], моделювання та класифікації мультимедійного трафіку [11], керування мобільними роботами [12], прийнятті рішень [13], прогнозуванні часових послідовностей [14, 15, 7], апроксимації функцій [16] тощо.

Беручи до уваги позитивні результати цих та інших досліджень, видається можливим застосувати математичний апарат нечітких множин типу 2 і в задачі кластеризації, зокрема такої, що орієнтована на множини даних соціального характеру.

Мета роботи — розробка методу кластерного аналізу даних соціального спрямування на основі інтервальних нечітких множин типу 2.

ВИХІДНІ ПЕРЕДУМОВИ ТА ПОСТАНОВКА ЗАДАЧІ

Існує велика кількість методів кластеризації, які можна класифікувати на чіткі та нечіткі. Чіткі методи кластеризації розбивають вихідну множину об'єктів $\mathbf{X}$ на декілька підмножин, що не перетинаються. При цьому будь-який об'єкт із $\mathbf{X}$ належить лише одному кластеру. Нечіткі методи кластеризації дозволяють одному й тому самому об'єкту належати одночасно до декількох (або навіть до всіх) кластерів, але з різним ступенем. Єдиною відмінністю є те, що у випадку нечіткого розбиття ступінь належності об'єкта до кластера приймає значення з інтервалу $[0,1]$, а при чіткому — з двохелементної множини $\{0,1\}$. Нечітка кластеризація в багатьох ситуаціях адекват-

ніше описує характер вихідної множини, наприклад, для об'єктів, розташованих на межі кластерів [2].

Основою переважної більшості сучасних методів нечіткого кластерного аналізу є алгоритм FCM (Fuzzy C-Means) Дж. Беждека [17].

Проте якість знайдених центрів суттєво залежить від попереднього вибору як значень μ_{ij}, так і центрів c_i. Крім того, FCM використовує обмеження, подібне до того, що накладає на шуканий розв'язок теорія ймовірностей: сума ступенів належності i-ї точки до всіх кластерів $j=\overline{1,N}$ становить 1: $\sum_{i=1}^{c}\mu_{ij}=1$ для всіх j [18]. Таке обмеження має на меті уникнути тривіального розв'язку, коли всі ступені належності виявляються рівними нулю, і дає змістовні результати в тих прикладних застосуваннях, де припущення про «імовірнісну» природу ступенів належності має практичний сенс.

Але, оскільки ступені належності, отримані за такого обмеження, відносні, вони непридатні в тих задачах, в яких ступінь належності точки до кластера має відображати її типовість, характерність саме для цього кластера. Це повністю узгоджується з теорією нечітких множин Заде, адже ступінь належності точки до класичної нечіткої множини є абсолютною величиною, незалежною від ступенів належності цієї ж точки до інших нечітких множин, визначених на тій самій універсальній множині. Таке формулювання доцільніше для більшості задач кластеризації, оскільки ступінь належності точки до кластера є мірою того, наскільки ця точка є носієм спільних характеристик кластера, її типовості; і ступінь належності не повинен залежати від того, як вона розташована відносно інших кластерів.

Виходячи з цього, у роботі [5] було переглянуто цільову функцію методу FCM таким чином, щоб за досягнення її мінімуму ступені належності для репрезентативних точок кластерів були високими, а для не репрезентативних — низькими, незалежно від взаємного положення точок та кластерів. Результуючий функціонал має вигляд:

$$E=\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}^{m}d_{ij}^{2}+\sum_{i=1}^{c}\eta_i\sum_{j=1}^{N}(1-\mu_{ij})^{m}, \qquad (1)$$

де η_i — додатне число.

Значення η_i визначає відстань від центра кластера, на якій значення ступеня належності точки до кластера стає рівним 0,5.

За такої цільової функції відповідним чином змінюються також і формули для перерахунку змінних величин методу:

$$\mu_{ij}=\frac{1}{1+\left(\frac{d_{ij}^{2}}{\eta_j}\right)^{\frac{1}{m-1}}};\ \eta_{ij}=\frac{\sum_{j=1}^{N}\mu_{ij}^{m}d_{ij}^{2}}{\sum_{j=1}^{N}\mu_{ij}^{m}}.$$

Співвідношення, що використовується для перерахунку координат центрів кластерів, порівняно з FCM залишається без змін:

$$c_i = \frac{\sum_{j=1}^{p} \mu_{ij}^m x_j}{\sum_{j=1}^{p} \mu_{ij}^m}.$$

Розв'язки, які отримано при такому підході, більше відповідають дійсності. Таке розуміння ступенів належності має ще один позитивний момент: воно дає змогу легко відфільтрувати точки, що вносять шум, оскільки вони за такого формулювання матимуть низькі ступені належності до всіх без винятку кластерів.

Не зважаючи на таке вдосконалення, одна проблема залишається спільною для FCM та PCM: обидва методи в усіх обчисленнях спираються на параметр m, що задає рівень нечіткості кластерів.

Випадок $m = 1$ відповідає чіткій кластеризації. Зі зростанням m ступені належності всіх без винятку точок до всіх кластерів наближаються до 0,5, як показано на рис. 1 (для випадку двох кластерів). Кожна крива зображає зміну ступеня належності точки до одного з кластерів.

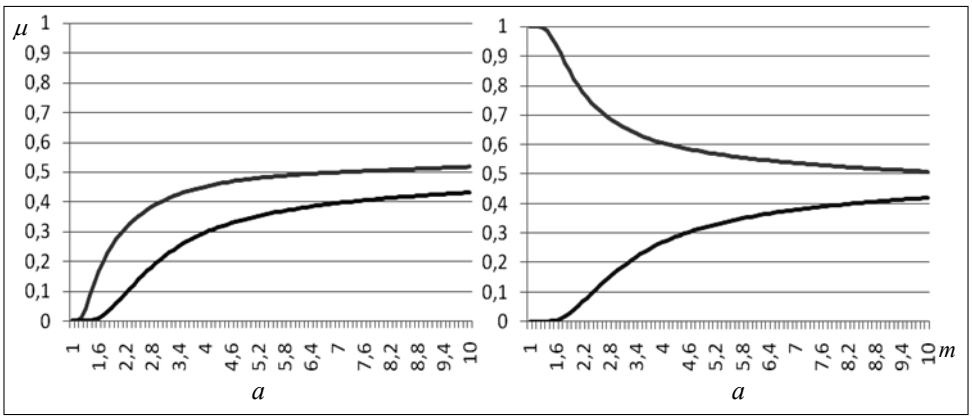

Рис. 1. Зміна ступенів належності точки до кластерів при зміні рівня нечіткості: *а* — точка нерепрезентативна; *б* — точка репрезентативна

На рис. 1 видно, що в усіх випадках m змінюється монотонно, обрати на такій кривій одну оптимальну точку неможливо. Тому закономірно, що строго обґрунтованих механізмів визначення m не існує.

Параметр m, як правило, задається емпірично дослідником, при цьому доводиться повністю покладатися на це заздалегідь задане значення без жодних гарантій його правильності. З цим пов'язана невизначеність, яку неможливо врахувати, коли отримане значення міри належності точки до кластера є єдиним числом. Тому для того, щоб убезпечити себе від помилкового результату, що пов'язаний із неправильним вибором значення m, доцільно використовувати інтервальні функції належності типу 2. Такий підхід найчастіше застосовується тоді, коли точний характер розподілу ступенів належності другого типу в області між границями інтервалу невідомий. Саме такий випадок являє собою задача кластеризації: невідомо, чи піддається

виділенню та математичному опису закономірність, за якою розподілені ступені належності другого типу, та чи має дослідження цієї закономірності практичний сенс. З іншого боку, інформація про верхню та нижню функції належності, що описують кожен кластер залежно від значення параметра m, має виняткову цінність, оскільки інтервал (його ширина та розташування відносно нуля та одиниці) несе значно більше інформації про міру належності точки до кластера, ніж єдине число. Наприклад, ширина інтервалу може свідчити про ступінь точності отриманого розв'язку. Тому пропонується модифікувати алгоритм кластеризації, який наведено в [5], для роботи з інтервальними ступенями належності. Цим буде досягнуто повне врахування невизначеності, пов'язаної з різними можливими значеннями рівня нечіткості, для подальшого аналізу результатів кластеризації.

Нехай є N об'єктів $x = \{x_1, x_2, \ldots, x_N\}$. Необхідно розбити їх на c кластерів та визначити місця розташування центрів кластерів c_i, $i = \overline{1,c}$, а також ступені належності μ_{ij} кожної з точок x_i до кластера c_i. Виходячи з визначення ступеня належності як міри типовості заданої точки для відповідного кластера, знайти такі значення шуканих параметрів, які ведуть до мінімуму функціонала (1). Враховуючи властивості рівня нечіткості m та його вплив на результати кластерного аналізу, представити ступені належності у вигляді інтервалів, ліва та права границі яких лежать у межах $[0,1]$.

МЕТОДИКА ДОСЛІДЖЕННЯ

Для розв'язання поставленої задачі пропонуємо модифікацію алгоритму кластерного аналізу РСМ [5]. Окрім нетрадиційного трактування ступенів належності та стійкості до шуму він володіє ще однією властивістю. Йдеться про те, що, оскільки міри належності однієї й тієї самої точки до різних кластерів незалежні одна від одної, ступінь належності точки до одного з них можна змінити без обов'язкової процедури перерахунку ступенів її належності до всіх інших кластерів. Ця властивість є надзвичайно корисною, оскільки вона дає змогу «розтягти» ступінь належності точки до кластера з чіткого значення в інтервал, і це не ставить під загрозу виконання обмеження на суму значень ступенів належності точки до всіх наявних кластерів.

Не зважаючи на всі переваги, у класичному алгоритмі РСМ не вдалося уникнути спільного для переважної більшості методів кластеризації недоліку: він передбачає апріорне задання числа кластерів до початку виконання обчислень. Найпростіший шлях уникнути цієї проблеми — виконувати розбиття при різній можливій кількості кластерів та порівнювати результати за певним критерієм оптимальності. У роботі [19] наведено декілька функціоналів, які називаються індексами достовірності та цілком відповідають вимогам, що висуває ця задача до критеріїв такого роду. Скористаємося індексом Квона, зокрема, для визначення оптимального числа кластерів для заданого рівня нечіткості m:

$$V_k(c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^m \left\|x_j - v_i\right\|^2 + \frac{1}{c}\sum_{i=1}^{c}\left\|v_i - \bar{v}\right\|^2}{\min\limits_{i\neq j}\left\|v_i - v_j\right\|^2},$$

де μ_{ij} — ступінь належності точки j до кластера i; v_i — центр j-го кластера; $\bar{v}$ — середнє значення центрів кластерів; m — рівень нечіткості; c — кількість кластерів; N — кількість точок.

Що менше значення має V_k, то кращим вважається розбиття.

Проте визначення кількості кластерів — не єдине застосування цього показника. У межах цього підходу пропонується використовувати його також для визначення меж інтервалу розтягу ступеня належності. Межі інтервалу визначимо, керуючись поведінкою індекса Квона на заданому інтервалі зміни параметра m (рис. 2, *а*). Практичний інтерес викликає лише перший його локальний мінімум, який спостерігається за оптимального значення m [5] (рис. 2, *б*). Тому за межі інтервалу приймемо праву та ліву точки перегину кривої, найближчі до розглядуваного локального мінімуму.

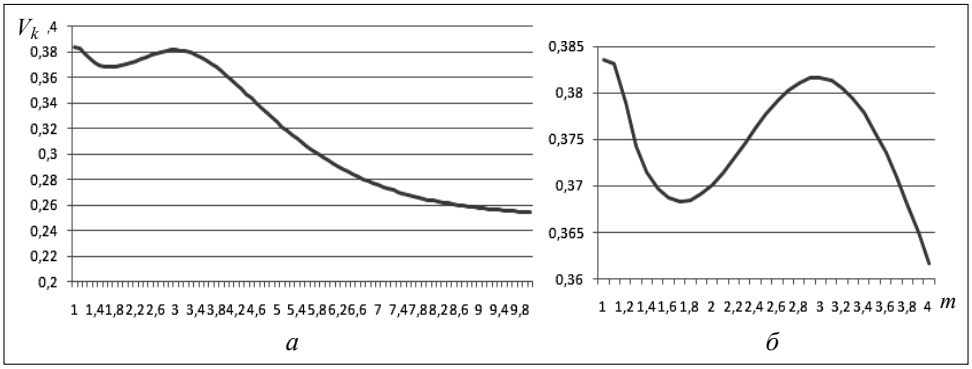

Рис. 2. Зміна індексу Квона залежно від зміни рівня нечіткості: *а* — в межах від 1 до 10; *б* — збільшений фрагмент: від 1 до 4

При такому підході отриманий нечіткий кластер матиме вигляд, як показано на рис. 3. Для його повного опису достатньо визначити лише верхню та нижню функції належності.

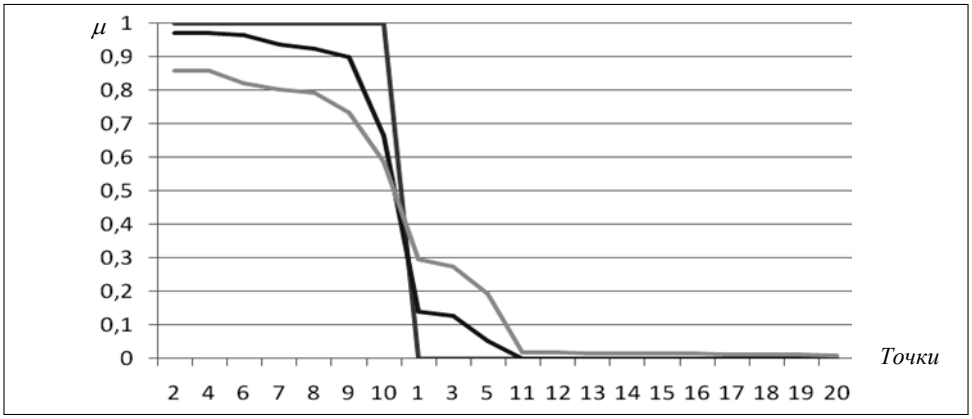

Рис. 3. Інтервальні функції належності точок до кластеру

Для початкової ініціалізації центрів кластерів використаємо звичайний метод FCM. Він збігається за лічені ітерації, тому якнайкраще підходить для цього завдання, адже воно вимагає грубого наближеного розв'язку.

Отже, сформулюємо покроковий алгоритм розв'язання задачі кластерного аналізу в заданій постановці.

1. Глобальний індекс Квона ініціалізувати максимально можливим значенням.

2. Задати початкову кількість кластерів $c = 2$.

3. Визначити приблизні місця розташування центрів кластерів за допомогою алгоритму FCM.

4. Оцінити значення η для результату роботи FCM.

5. Сформувати матрицю D як матрицю Евклідових відстаней від кожної точки з вихідної множини до центра кожного з кластерів.

6. Задати початкове значення рівня нечіткості $m = 1$.

7. Розрахувати початкове значення локального індексу Квона.

8. Розрахувати функцію належності для кожної з пар (точка, кластер), користуючись відповідним співвідношенням із методу PCM.

9. Перерахувати положення центрів кластерів за формулою, спільною для обох методів.

10. Перерахувати матрицю відстаней D.

11. Розрахувати цільову функцію PCM при заданих значеннях ступенів належності, координат центрів кластерів, елементів матриці D та вектора η.

12. Якщо розраховане значення цільової функції менше за отримане на попередній ітерації, повернутись до кроку 8.

13. Розрахувати значення локального індексу Квона при заданому m. Якщо воно менше за попереднє значення, то збільшити m та повернутись до кроку 8.

14. Перерахувати значення глобального індексу Квона. Зберегти проміжні результати обчислень для поточного $m = m_k$.

15. Якщо кількість кластерів менша за кількість точок у вихідній множині, збільшити c та перейти до кроку 3.

16. Серед усіх проміжних результатів обрати варіант розбиття з мінімальним значенням глобального індексу Квона. Подати ступені належності у вигляді інтервалів, обмежених їхніми значеннями при $m = 1$ та отриманим на кроці 14 $m = m_k > 1$.

КОМП'ЮТЕРНИЙ ЕКСПЕРИМЕНТ ЗА ПОКАЗНИКАМИ ЛЮДСЬКОГО РОЗВИТКУ В КРАЇНАХ СВІТУ

Для аналізу було взято дані зі щорічного звіту ООН з людського розвитку за 2010 р. [20] для всіх незалежних держав світу за такими показниками:

- середня очікувана тривалість життя при народженні;
- середня тривалість освітньої підготовки громадян;
- ВНП на душу населення;

- індекс гендерної нерівності в країні.

У результаті у вхідних даних було виділено 3 компактних кластери (табл. 1 та 2).

Таблиця 1. Координати центрів кластерів

Показники	Кластер 1	Кластер 2	Кластер 3
Індекс гендерної нерівності	0,615842	0,290623	0,73791
ВВП на душу населення	0,096877	0,3814	0,018852
Тривалість життя	0,732709	0,920551	0,415003
Кількість років освітньої підготовки	0,532329	0,837381	0,253012

Таблиця 2. Інтервальні ступені належності країн до кластерів

Країни	Кластер 1		Кластер 2		Кластер 3	
Algeria	0,976295	1	0,00087	0,097652	0,000931	0,090729
Australia	1,98E-05	0,018657	0,514417	0,962005	8,02E-07	0,005172
Austria	0,000102	0,03422	0,791044	0,995485	2,54E-06	0,008165
Bangladesh	0,015304	0,112614	2,66E-05	0,0229	0,663903	0,765039
Belgium	5,08E-05	0,026589	0,801127	0,999869	1,61E-06	0,006799
Benin	0,000528	0,039172	8,02E-06	0,013866	0,884219	0,99949
Brazil	0,945761	0,999997	0,000745	0,090964	0,001058	0,09548
…	…	…	…	…	…	…
Togo	0,006406	0,093062	2,20E-05	0,021245	0,946151	0,968477
Tunisia	0,62551	0,935993	0,000945	0,098948	0,000628	0,075782
Turkey	0,696438	0,9928	0,000536	0,078341	0,002025	0,121107
United Kingdom	0,000259	0,049088	0,832883	0,978633	4,25E-06	0,010059
Venezuela	0,588932	0,950364	0,000741	0,088684	0,000924	0,088121
Zimbabwe	2,17E-05	0,015335	4,13E-06	0,010372	0,596237	0,974833

Отримані кластери мають вигляд нечітких множин типу 2 (рис. 4). Значна ширина інтервалу деяких конкретних значень ступенів належності дає змогу судити про наявність шумів у вхідних даних.

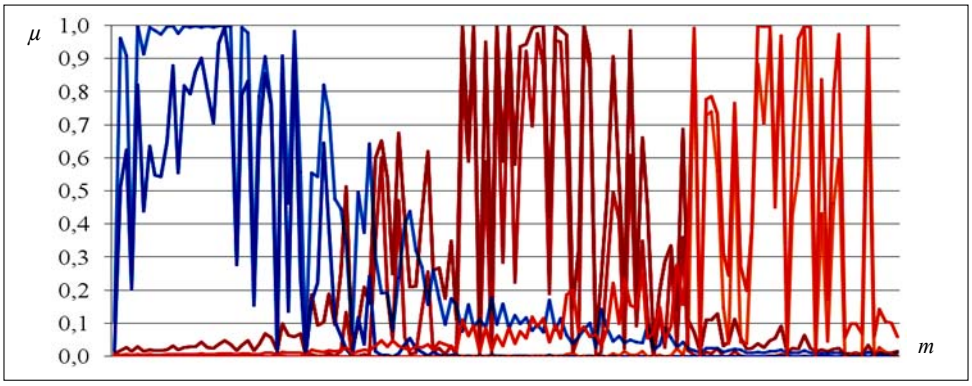

Рис. 4. Графічне подання результатів кластеризації

Інтервальні значення ступенів належності було обчислено, виходячи зі значень рівня нечіткості $m = [1;1{,}65]$. На рис. 5 показано характер зміни індексу Квона залежно від значення m.

На рис. 5, *б* можна побачити локальний мінімум цієї залежності в точці $m = 1{,}65$. Цю точку і прийнято за праву межу інтервалу зміни параметра m.

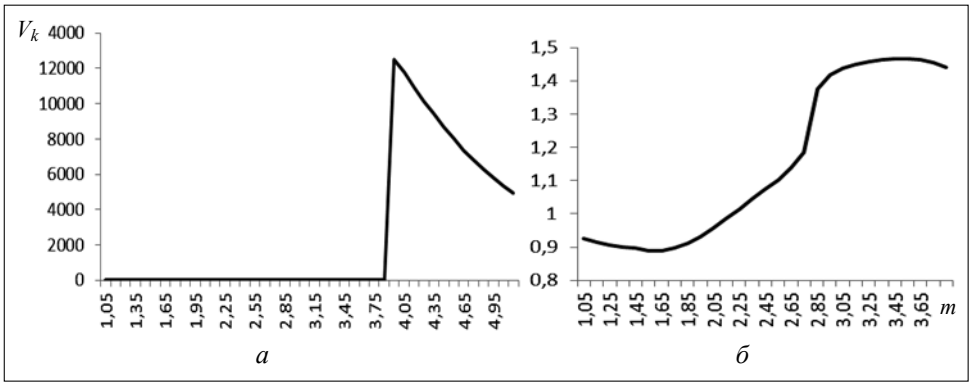

Рис. 5. Характер зміни індексу Квона залежно від значення рівня нечіткості (*а* — для $m = [1;5]$; *б* — для $m = [1;3{,}7]$)

При цьому варто скористатися такими рекомендаціями:

- Метод дає змогу отримати інтервал, в якому знаходиться оптимальне значення рівня нечіткості. Інтервал (його ширина та розташування відносно нуля та одиниці) несе значно більше інформації про міру належності точки до кластера, ніж єдине число.
- У разі значної ширини інтервалу слід залучити експерта для прийняття остаточного рішення. Експерт, у свою чергу, може прийняти рішення про повторення експерименту у зв'язку зі значною невизначеністю у вхідних даних.

ВИСНОВКИ

Запропоновано метод кластеризації на основі інтервальних функцій належності типу 2 з використанням індексу вірогідності Квона для визначення оптимального числа кластерів та меж інтервальних значень ступенів належності.

Значна ширина інтервалу деяких конкретних значень ступенів належності дає змогу зробити висновок про наявність шумів у вхідних даних. Зважаючи на це та на високу складність поставленої прикладної задачі, для прийняття будь-якого остаточного рішення доцільно залучати експертів з цієї галузі.

Метод випробувано в прикладній задачі соціального характеру та отримано змістовні результати, що свідчить про перспективність використання запропонованого підходу в задачах такого роду.

ЛІТЕРАТУРА

1. *Котова Е.С.* Кластерный анализ в задачах социально-экономического прогнозирования. — http://vuzlib.net/beta3/ html/1/4055/4081/.

2. *Мандель И.Д.* Кластерный анализ. — М.: Статистика, 1988. — 176 с.
3. *Серебрякова Л.А.* Методы оценки уровня социально-экономического развития регионов // Вестн. СевКавГТУ. Сер. Экономика. — 2003. — № 3 (11). — http://science.ncstu.ru/articles/econom/11/02.pdf.
4. *Захарченко С.М., Кондратенко Н.Р., Манаєва О.О.* Дослідження можливостей генетичного алгоритму в задачі кластеризації користувачів мережі Internet // Інформаційні технології та комп'ютерна інженерія. — 2010. — № 2 (18). — С. 67–72.
5. *Krishnapuram R., Keller J.M.* A possibilistic approach to clustering // IEEE Transactions on Fuzzy Systems. — 1993. — № 1(2). — P. 98–110.
6. *Zadeh L.A.* Fuzzy sets as a basis for a theory of possibility // Fuzzy sets and systems 100 Supplement. — 1999. — P. 9–34.
7. *Liang Q., Mendel J.M.* Interval type-2 fuzzy logic systems: theory and design // IEEE Transactions on Fuzzy Systems. — 2000. — **8**. — P. 535–550.
8. *Mendel J.M., John R.I., Liu F.* Interval Type-2 Fuzzy logic systems made simple // IEEE Transactions on Fuzzy Systems. — 2006. — **14**. — № 6. — P. 808–821.
9. *Mendel J.M., John R.I.* Interval Type-2 fuzzy sets made simple // IEEE Transactions on Fuzzy Systems. — 2002. — **10**. — № 2. — P. 117–127.
10. *Zeng J., Liu Z.Q.* Type-2 Fuzzy sets for pattern classification: A review // Proceedings of the IEEE Symposium on Foundations of computational intelligence. — 2007. — P. 193–200.
11. *Liang Q., Mendel J M.* MPEG MBR Video traffic modeling and classification using fuzzy technique // IEEE Transactions on Fuzzy Systems. — 2001. — **9**. — № 1. — P. 183–193.
12. *Wu K.C.* Fuzzy interval control of mobile robots // Computers and Electrical Engineering. — 1996. — **22**. — P. 211–229.
13. *Yager R.R.* Fuzzy subsets of type II in decisions // Cyber Journals. — 1980. — **10**. — P. 137–159.
14. *Karnik N.N., Mende J.M.* Applications of type-2 fuzzy logic systems to forecasting of time series // Information Sciences. — 1999. — **120**. — P. 89–111.
15. *Mendel J.M.* Uncertainty, fuzzy logic, and signal processing // Signal Processing Journal. — 2000. — **80**. — P. 913–933.
16. *Karnik N.N., Mendel J.M.* An introduction to type-2 fuzzy logic systems. — Los Angeles, CA. — http://sipi.usc. edu/~mendel/report.
17. *Bezdek J.C.* Pattern recognition with fuzzy objective function algorithms. — NY: Plenum Press, 1981. — 256 p.
18. *Зайченко Ю.П.* Нечеткие модели и методы в интеллектуальных системах. — К.: Издат. дом «Слово», 2008. — 344 с.
19. *Oliveira J.V., Pedrycz W.* Advances in fuzzy clustering and its applications. — Chichester: John Wiley & Sons Ltd., 2007. — 435 p.
20. *The Real* Wealth of Nations: pathways to human development. Human development report 2010: 20-th anniversary edition . — UNDP, 2010. — 227 p.

Надійшла 07.06.2011