УДК 004.8

Н.Р. Кондратенко, О.О. Снігур

ІНТЕРВАЛЬНА НЕЧІТКА КЛАСТЕРИЗАЦІЯ НА ОСНОВІ АЛЬТЕРНАТИВНИХ КРИТЕРІЇВ ЯКОСТІ

The paper studies several clustering validity indices (Kwon index, Xie-Beni index, partition index) in view of the fuzzy parameter. We reveal the pattern of change in indices being researched against the fuzzy parameter change. We introduce an interval type-2 fuzzy clustering method based on combination of three validity indices. The membership values are presented as intervals. It allows preserving completeness of information on a set of their possible values, as well as reducing the influence of each specific index on uncertainty reflected in the result. The latter is achieved by detecting the intersection area of intervals of fuzzy parameter values based on every studied index. The solutions tolerance of results' abnormal observations is achieved by using the PCM robust clustering method. We analyze the widths of intervals of membership values obtained using the proposed approach in the case of noisy data or data containing abnormalities. Using the proposed approach the countries of the world are clustered relying on their human development characteristics.

Вступ

Кластерний аналіз – одна із задач інтелектуального аналізу даних, що полягає в розподілі множини об'єктів, кожен із яких характеризується певним набором ознак, на класи таким чином, що об'єкти з одного класу вважаються якісно подібними, а об'єкти з різних класів – відмінними.

Традиційні (чіткі) методи кластеризації ставлять у відповідність кожній точці строго один кластер. Методи нечіткої кластеризації базуються на теорії нечітких множин Заде [1] та допускають належність кожної точки до всіх кластерів із різними ступенями належності. Найпоширенішим із них є метод Fuzzy C-Means (FCM) [2]. Однією з найважливіших проблем практичного застосування методів нечіткої кластеризації є нестійкість їх розв'язків до наявності аномальних спостережень у досліджуваній сукупності даних. Під аномаліями, або аномальними спостереженнями, слід розуміти спостереження, що належать до кластерів, число представників яких у досліджуваній сукупності є істотно малим порівняно з числом представників основних кластерів. Для розв'язання задачі нечіткої кластеризації за умов наявності в досліджуваній сукупності аномальних спостережень розроблено спеціальні оптимізаційні методи, які прийнято називати робастними. Серед найвідоміших робастних методів нечіткої кластеризації – метод Possibilistic C-Means (PCM) [3].

Оскільки процес кластеризації являє собою навчання без учителя, всі без винятку наближені методи кластерного аналізу чутливі до вибору значень параметрів методу. Для розглядуваного методу PCM параметрами є:

- початкові значення центроїдів;
- число кластерів;
- рівень нечіткості кластерів.

У даній статті приділено увагу визначенню оптимального розв'язку задачі кластеризації відносно рівня нечіткості. Рівень нечіткості – це параметр, який кількісно характеризує, наскільки нечіткими, розмитими є кластери, отримані на виході методу, та який задається, як правило, емпірично дослідником [4].

Для визначення параметрів кластеризації існує ряд критеріїв якості: коефіцієнт розбиття Беждека, критерії Квона, Хіе-Бені тощо [5–9]. Кластеризація відповідним методом виконується кілька разів за різних комбінацій значень досліджуваних параметрів, після чого на основі одного з наведених вище критеріїв вибираються оптимальні значення.

Зважаючи на те, що жоден з існуючих на сьогодні критеріїв не є універсальним та не дає змоги отримати єдино правильне значення параметра, пропонується розв'язувати задачу кластеризації в інтервальній формі [10, 11], щоб убезпечити себе від помилкового результату, пов'язаного з неправильним вибором значення рівня нечіткості. Такий підхід дасть змогу зберегти повноту інформації про множину можливих значень рівня нечіткості та, відповідно, ступенів належності об'єктів до кластерів [12]. Тому доцільно розв'язати задачу кластерного аналізу з використанням математичного апарату нечітких множин типу 2.

Постановка задачі

Нехай є N об'єктів $x = \{x_1, x_2, \ldots, x_N\}$. Необхідно розбити їх на c кластерів і 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_j. Виходячи з визначення ступеня належності як міри типовості заданої точки для відповідного кластера, потрібно знайти такі значення шуканих параметрів, які ведуть до мінімуму цільового функціонала методу РСМ. Враховуючи властивості рівня нечіткості m та його вплив на результати кластерного аналізу, необхідно відобразити ступені належності у вигляді інтервалів, ліва та права границі яких лежать у межах [0, 1].

Метою роботи є розширення можливостей методу РСМ через введення інтервальних ступенів належності, що дає змогу враховувати та моделювати невизначеності, наявні у вихідних даних.

Методика дослідження

Для визначення оптимального значення рівня нечіткості пропонується використовувати так звані критерії якості (validity indices [9]) кластеризації. Їх пряме призначення – кількісна оцінка якості результатів кластеризації як поєднання умов подібності об'єктів у межах одного кластера та відмінності об'єктів із різних кластерів. Найбільш традиційне їх застосування – визначення оптимального числа кластерів. У даній статті число кластерів вважатимемо вихідним параметром, заданим заздалегідь. Критерії ж якості пропонується використовувати для визначення оптимального значення рівня нечіткості.

Розглянемо найпоширеніші критерії якості.

Індекс розбиття (Partition Index):

$$SC(c,m) = \sum_{i=1}^{c} \frac{\sum_{k=1}^{N} (\mu_{i,k})^m \,||\, x_k - v_i \,||^2}{\sum_{k=1}^{N} \mu_{i,k} \sum_{j=1}^{c} ||\, v_j - v_i \,||^2},$$

де μ_{ij} – ступінь належності точки j до кластера i; v_j – центр j-го кластера; m – рівень нечіткості; c – кількість кластерів; N – кількість точок.

Критерії Хіе-Бені та Квона беруть до уваги геометричні властивості кластерів, а не лише відстані об'єктів до їх центрів.

Критерій Квона:

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

де $\bar{v}$ – середнє значення центрів кластерів.

Критерій Хіе-Бені:

$$XB(c,m) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N} (\mu_{ij})^m \,||\, x_j - v_i \,||^2}{N \min\limits_{i,j} ||\, v_i - x_j \,||^2}.$$

Менші значення наведених критеріїв відповідають кращим варіантам розбиття.

Для визначення характерних властивостей досліджуваних критеріїв візьмемо метод можливісної кластеризації РСМ, оскільки він є стійким до шумів та аномальних явищ у вихідних наборах даних [13]. Метод РСМ допускає, що досліджуваній сукупності можуть належати спостереження неосновних кластерів, що веде до ослаблення обмеження на суму ступенів належності:

$$\sum_{i=1}^{c} \mu_{ij} \leq 1,$$

де c – число основних кластерів, що задається до початку процедури кластеризації [3].

Цільовий функціонал методу РСМ можна подати таким чином:

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

де d_{ij}^2 – квадрат Евклідової відстані між об'єктами i та j; η_i – додатне число. η_i визначає відстань від центра кластера, на якій значення ступеня належності точки до кластера стає рівним 0,5.

За такої цільової функції змінні величини методу визначаються за такими співвідношеннями:

$$\mu_{ij} = \frac{1}{1 + \left(\frac{d_{ij}^2}{\eta_j} \right)^{\frac{1}{m-1}}}; \quad \eta_{ij} = \frac{\sum\limits_{j=1}^{N} \mu_{ij}^m d_{ij}^2}{\sum\limits_{j=1}^{N} \mu_{ij}^m}; \quad c_i = \frac{\sum\limits_{j=1}^{c} \mu_{ij}^m x_j}{\sum\limits_{j=1}^{c} \mu_{ij}^m}.$$

Для початкової ініціалізації центрів кластерів і ступенів належності будемо використовувати метод FCM Дж. Беждека [2].

Дослідимо поведінку наведених вище індексів залежно від зміни рівня нечіткості m на прикладі тестових наборів даних, зображених на рис. 1. Випадок 1, *б* відрізняється від 1, *а* наявністю аномальних об'єктів (шуму).

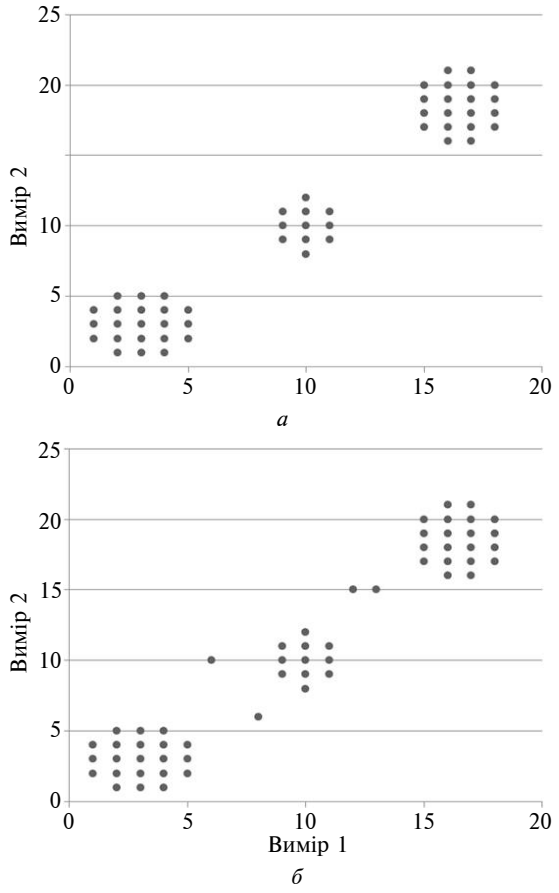


Рис. 1. Тестові набори даних: *а* – ідеальний (без аномальних спостережень); *б* – зашумлений (з аномальними спостереженнями)

Розв'яжемо поставлену задачу методом РСМ за різних значень параметра m та обчислимо значення критеріїв якості за наведеними вище співвідношеннями. Залежності значення відповідного критерію від значення рівня нечіткості m наведено на рис. 2.

Випадки *а*, *в* і *д* на рис. 2 відповідно демонструють поведінку критеріїв Квона, Хіе-Бені та індексу розбиття на наборі даних 1, *а*, в якому відсутні аномальні спостереження. Випадки *б*, *г* і *е* відповідають зашумленому набору даних 1, *б*.

Якщо говорити про самі результати кластеризації (ступені належності точок до кластерів), то кожен із наведених критеріїв має глобальний або локальний мінімум, що відповідає рекомендованому значенню m. Поклавши це значення за оптимальне, можна побудувати розв'язок, подібний до зображеного на рис. 3.

На рис. 3, *а* наведено розв'язок задачі кластеризації за значення $m = 2$, що відповідає мінімуму кривої зміни індексу Хіе-Бені на наборі даних 1, *а*. Результати кластеризації подано у вигляді нечітких множин: для кожного об'єкта на вісі абсцис по вісі ординат відкладено його ступені належності до кожного з кластерів. Набору даних 1, *б* відповідає розв'язок 3, *б*. На рис. 3, *в* зображено класичні нечіткі множини типу 1.

Отже, поклавши значення рівня нечіткості $m = 2$, ми отримали деякий розв'язок задачі кластеризації для заданого набору даних. Але, проаналізувавши криві на рис. 2, можна помітити, що вони мають мінімуми за різних значень m. Так, крім уже згаданого $m = 2$ у випадку 2, *в*, маємо результати $m = 1{,}9$ для 2, *а*, $m = 2{,}1$ для 2, *д*, $m = 1{,}7$ для 2, *б* і *г*, $m = 1{,}6$ для 2, *е*. Таким чином, залежно від способу розв'язання задачі (критерію якості) та якості вхідних даних (наявності чи відсутності аномалій) нами отримано п'ять варіантів розв'язку, кожен із яких претендує на оптимальність. Шостим претендентом можна вважати значення $m = 1{,}5$, рекомендоване в [14] для даної задачі. Остаточно визначити, який із розв'язків-претендентів є правильним і чи є серед них правильний розв'язок, не видається можливим. Тому для того щоб убезпечити себе від помилкового результату, пов'язаного з неправильним вибором значення m, доцільно використовувати нечіткі ступені належності типу 2. При цьому сам ступінь належності точки до кластера являє собою нечітку множину типу 1.

У праці [4] поставлену задачу розв'язано в інтервальній формі з використанням особливостей поведінки критерію Квона. Взявши за границі інтервалу зміни рівня нечіткості точки перегину кривих на рис. 2, *а* і *б*, отримаємо кластери у вигляді нечітких множин типу 2, як показано на рис. 4. Випадок 4, *а* відповідає набору 1, *а*, випадок 4, *б* – набору 1, *б*.

Інтервал зміни рівня нечіткості, отриманий з використанням критерію Квона – [1,3; 3,7] – має доволі значну ширину. Це відображається і на значеннях ступенів належності. Хоча в результатах у цілому прослідковується тенденція до утворення компактних кластерів, деякі об'єкти мають істотну невизначеність у ступенях належності. Серед таких об'єктів, наприклад, точка 20 ($\mu_{20,2} = [0{,}52; 0{,}98]$). Ця невизначеність є природною особливістю, характерною для будь-якої емпіричної оцінки.

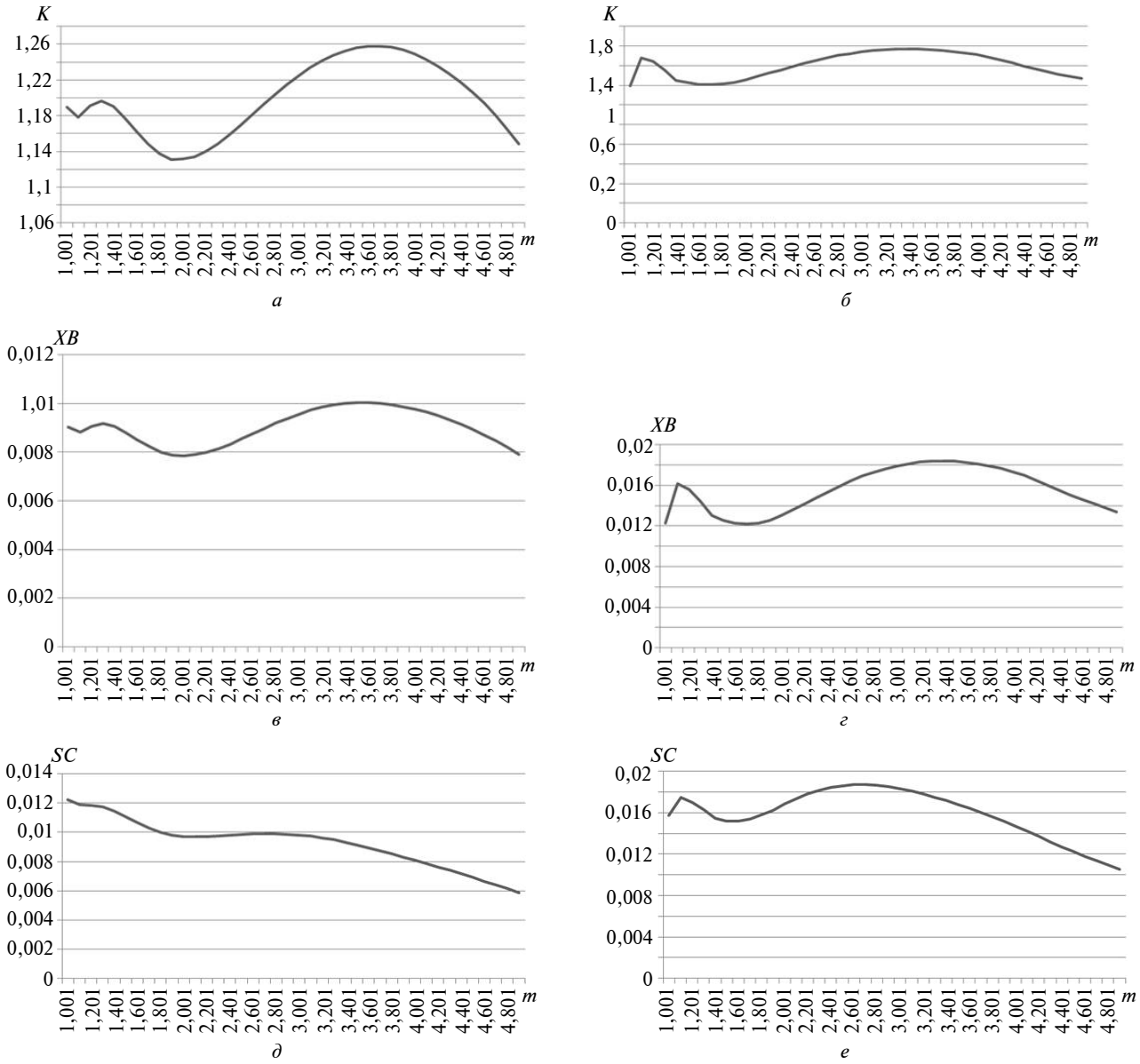


Рис. 2. Поведінка критеріїв якості залежно від зміни рівня нечіткості: *а* – критерій Квона, набір 1, *а*; *б* – критерій Квона, набір 1, *б*; *в* – критерій Хіе-Бені, набір 1, *а*; *г* – критерій Хіе-Бені, набір 1, *б*; *д* – індекс розбиття, набір 1, *а*; *е* – індекс розбиття, набір 1, *б*

Аномальні спостереження, наявні в наборі 1, *б*, також вносять невизначеність у результати кластеризації. Але найбільшим недоліком цього методу є те, що сам критерій якості, що використовується, вносить певну невизначеність у результати та генерує ступені належності з широким інтервалом навіть для "чистого" набору даних, апріорі вільного від аномальних спостережень.

Для того щоб зменшити вплив внутрішніх особливостей одного конкретного критерію якості на кінцевий результат, пропонується залучити до формування інтервалу інші критерії. Усі три критерії, що розглядалися в цій роботі, демонструють схожу поведінку на множині значень параметра *m*. Для кожного з них можна виокремити інтервал, обмежений двома найближчими до першого локального мінімуму точками перегину. Остаточний інтервал пропонується визначати як область перетину інтервалів, отриманих за критеріями Квона, Хіе-Бені та індексом розбиття.

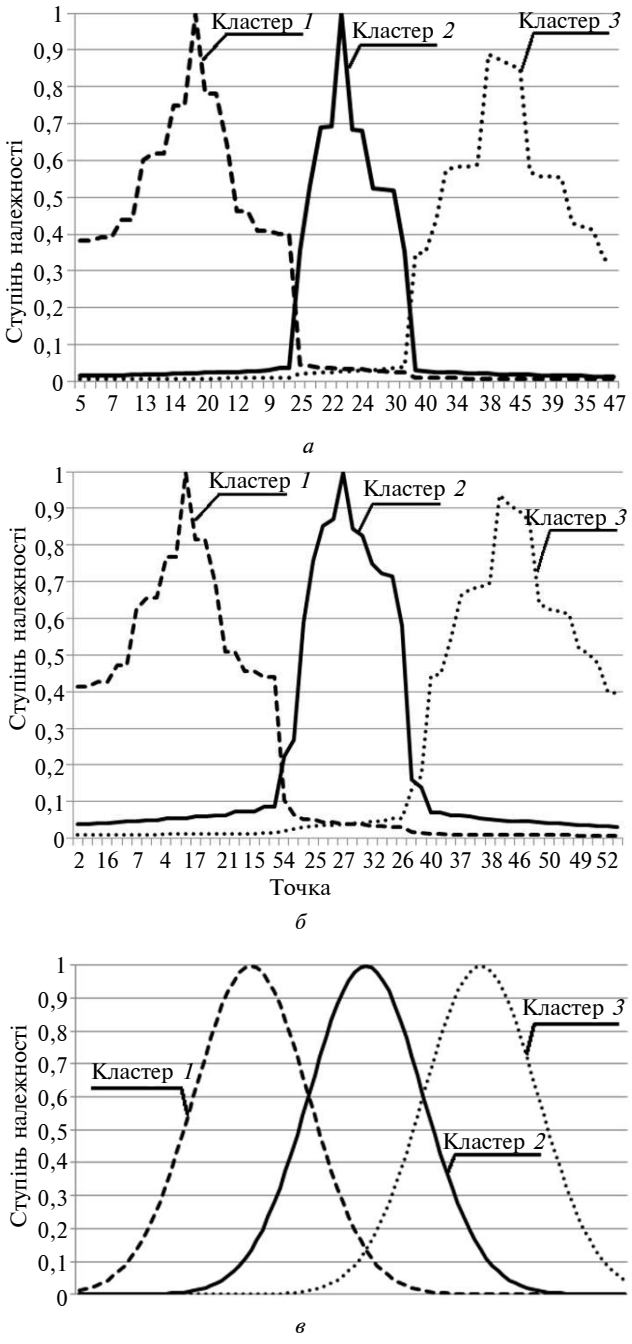


Рис. 3. Результати кластеризації як нечіткі множини типу 1: *а* – ступені належності при $m = 2$ для 1, *а*; *б* – ступені належності при $m = 2$ для набору 1, *б*

Так, у наведеному тестовому прикладі для набору 1, *а* маємо

$$\tilde{m}_a = \tilde{m}_{K_a} \cap \tilde{m}_{XB_a} \cap \tilde{m}_{SC_a} =$$

$$= [1,3;3,7] \cap [1,3;3,5] \cap [1,3;2,7] = [1,3;2,7].$$

Для набору 1, *б*:

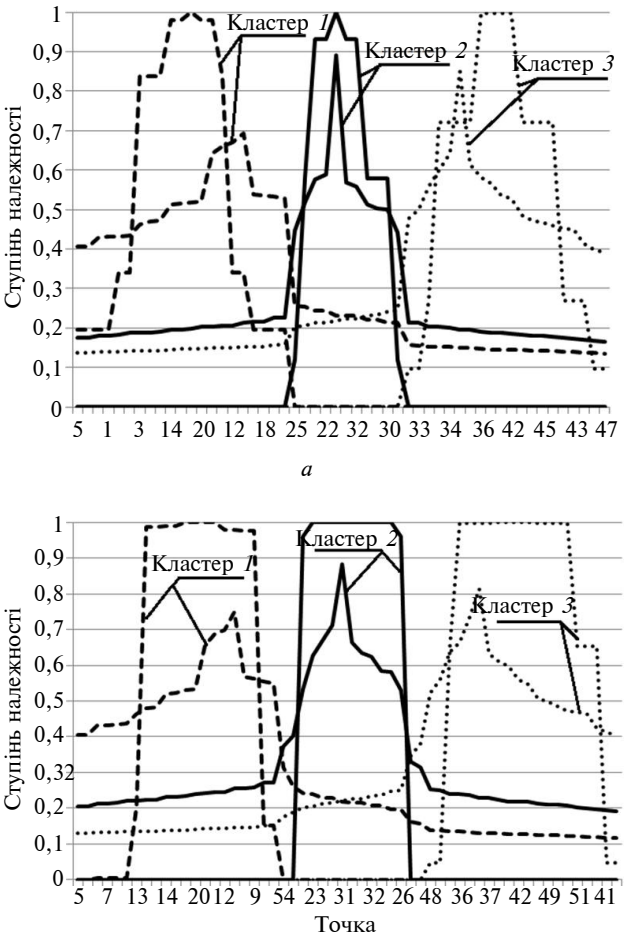


Рис. 4. Результати інтервальної нечіткої кластеризації на основі критерію Квона: *а* – для набору даних 1, *а*; *б* – для набору даних 1, *б*

$$\tilde{m}_б = \tilde{m}_{K_б} \cap \tilde{m}_{XB_б} \cap \tilde{m}_{SC_б} =$$

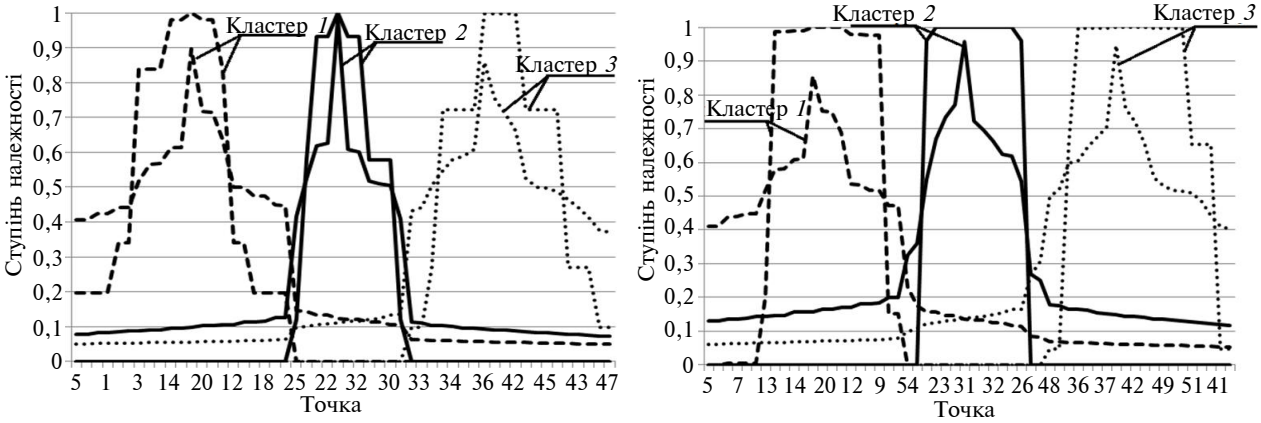
$$= [1,1;3,4] \cap [1,1;3,4] \cap [1,1;2,7] = [1,1;2,7].$$

Інтервальні нечіткі кластери, отримані за таких значень рівня нечіткості, наведено на рис. 5.

Отже, за допомогою комбінування трьох критеріїв якості та визначення області перетину інтервалів, отриманих за кожним із них, отримано:

а) інтервал меншої ширини для позбавленого аномалій набору даних 1, *а* порівняно з інтервалом на основі єдиного критерію;

б) ширший інтервал зміни рівня нечіткості та, відповідно, більш розмиті кластери для зашумлених даних набору 1, *б*, що відповідає очікуваному результату.



а *б*

Рис. 5. Результати інтервальної нечіткої кластеризації на основі комбінації критеріїв Квона, Хіе-Бені та індексу розбиття: *а* – для набору даних 1, *а*; *б* – для набору даних 1, *б*

Комп'ютерний експеримент

Перевіримо роботу запропонованого методу на реальному наборі даних. Для аналізу візьмемо дані зі щорічного звіту ООН за 2010 р. [15] для країн світу за такими показниками:

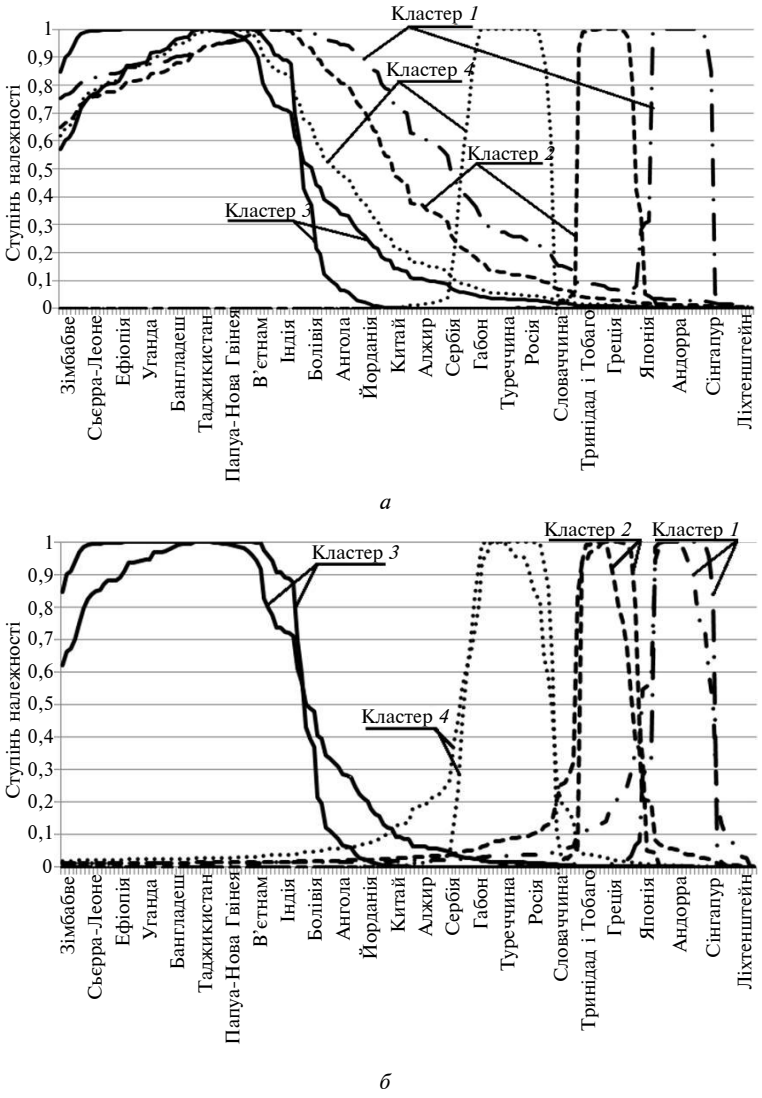
- середня очікувана тривалість життя;
- середня тривалість освітньої підготовки громадян;
- ВНП на душу населення.

Використання запропонованого підходу дало змогу виділити у множині вихідних даних 4 кластери з центроїдами, наведеними в таблиці.

Розподіл інтервальних нечітких ступенів належності наведено на рис. 6. На рис. 6, *а* значення границь інтервалу зміни рівня нечіткості отримано лише за критерієм Квона; вони становлять [1,3; 2]. Бачимо, що на верхній границі інтервалу вся досліджувана множина вироджується в один суцільний кластер, що практично не піддається аналізу на предмет наявності в ньому інших закономірностей. У той самий час використання комбінації з трьох критеріїв дає рівень нечіткості в межах [1,3; 1,8] та ступені належності, як на рис. 6, *б*.

На рис. 6, *б* можна виокремити 4 кластери у вигляді інтервальних нечітких множин та водночас виявити об'єкти, що несуть у собі невизначеність.

Що стосується складу кластерів, то до першого з них потрапили країни з



а

б

Рис. 6. Результати кластеризації. Ступені належності: *а* – за критерієм Квона; *б* – за критеріями Квона, Хіе-Бені та індексом розбиття

Таблиця. Результати кластеризації. Центроїди

Показник	Кластер *1*	Кластер *2*	Кластер *3*	Кластер *4*
Тривалість життя	80,80059	75,51261	59,14754	72,26426
Кількість років освіти	10,86078	9,764072	4,67611	8,832988
ВВП на душу населення	36114,24	23564,78	1788,912	12876,31

найвищим рівнем життя, а саме – Швеція, Данія, Німеччина, Франція, Фінляндія, Японія та ін. Другий кластер сформували переважно країни Центральної Європи, а також Саудівська Аравія з інтервалом ступеня належності [0,8; 1] та Нова Зеландія, ступінь належності якої також має широкий інтервал – [0,68; 0,99]. Третій, найбільший, кластер утворили країни Азії та Африки з економікою, що розвивається, та постсоціалістичні країни. Найхарактерніші країни цього кластера – Камерун, Нігерія, Лаос, Камбоджа, Таджикистан. До цього ж кластера потрапила і Україна зі ступенем належності [0,58; 0,88]. У складі кластера 4 – країни Латинської Америки (Бразилія, Панама, Венесуела) та європейські країни з порівняно невисокими показниками людського розвитку (Румунія, Болгарія, Чорногорія).

Висновки

Існуючі методи нечіткої кластеризації мають ряд параметрів, вибір яких впливає на результат їх роботи. Для методу робастної кластеризації РСМ такими параметрами є початкові значення центроїдів, число кластерів та рівень нечіткості. У статті приділено увагу визначенню оптимального розв'язку задачі кластеризації відносно рівня нечіткості. Для цього використано критерії якості Квона, Хіе-Бені та індекс розбиття. Жоден із цих критеріїв поодинці не є універсальним та не гарантує правильного розв'язку в кожній конкретній прикладній задачі. Тому поставлену задачу розв'язано в інтервальній формі із застосуванням математичного апарату нечітких множин типу 2. Поведінку трьох названих вище критеріїв залежно від зміни рівня нечіткості досліджено на тестових наборах даних, один із яких містить штучно внесені аномалії. Виявлено граничні значення інтервалу для рівня нечіткості за кожним із критеріїв. Запропоновано визначати остаточне інтервальне значення за областю перетину інтервалів, отриманих за трьома критеріями. Такий підхід дає можливість враховувати та моделювати невизначеності, наявні у вихідних даних. Проведено експеримент із використання запропонованого підходу для кластеризації країн світу за показниками людського розвитку та отримано змістовні результати.

Подальші дослідження доцільно присвятити способам автоматичного визначення інших параметрів кластеризації, таких як початкові значення центроїдів і число основних кластерів.

1. *L.A. Zadeh*, "Fuzzy sets as a basis for a theory of possibility", Fuzzy sets and systems, vol. 100, Sup. 1, pp. 9–34, 1999.
2. *J.C. Bezdek*, Pattern recognition with fuzzy objective function algorithms. New York: Plenum Press, 1981, 256 pp.
3. *Залесская К.М.* Анализ устойчивости методов нечеткой кластеризации к выбору их параметров // Искусственный интеллект. – 2010. – № 4. – С. 359–369.
4. *Кондратенко Н.Р., Манаєва О.О.* Нечітка кластеризація з урахуванням індексу вірогідності в задачах соціального спрямування // Системний аналіз та інформаційні технології: Матер. Міжнар. науково-технічної конф. SAIT 2011. – К.: ННК "ІПСА" НТУУ "КПІ", 2011. – С. 265.
5. *M. Halkidi et al.*, "On Clustering Validation Techniques", J. of Intelligent Inform. Syst., 17:2/3, pp. 107–145, 2001.
6. *Y. Liu et al*, "Understanding of Internal Clustering Validation Measures", in 2010 IEEE Int. Conf. Data Mining, Sydney, NSW (Australia), 2010, pp. 911–916.
7. *M. Ramze Rezaee et al.*, "A new cluster validity index for the fuzzy c-means", Pattern Recognitio, Let. 19, pp. 237–246, 1998.
8. *F. Kovacs et al.*, Cluster validity measurement techniques. London, UK: Academic Press, 2004, pp. 388–393.
9. *Advances* in Fuzzy Clustering and Its Applications, *J.*V. Oliveira and W. Pedrycz (Eds.). Chichester, UK: John Wiley & Sons Ltd., 2007, 435 pp.
10. *Q. Liang and J.M. Mendel*, "Interval type-2 fuzzy logic systems: Theory and design", IEEE Trans. Fuzzy Syst., vol. 8, pp. 535–550, 2000.
11. *Кондратенко Н.Р., Чеборака О.В., Куземко С.М.* Прогнозування часових послідовностей з використанням різновходових нечітких моделей на основі інтервальних функцій належності // Наукові вісті НТУУ "КПІ". – 2007. – № 4 – С. 62–68.
12. *Бардачев Ю.Н., Кричковский В.В., Маломуж Т.В.* Методологическая предпочтительность интервальных

экспертных оценок при принятии решений в условиях неопределенности // Вісник Харк. нац. ун-ту. – 2010. – № 890 – С. 18–28.

13. *R. Krishnapuram and J.M. Keller*, "A Possibilistic Approach to Clustering", IEEE Trans. Fuzzy Syst., vol. 1, no. 2, pp. 98–110, 1993.

14. *J.C. Bezdek*, "Fuzzy Mathematics in Pattern Classification", in PhD Thesis, Cornell University, Ithaca, New York, 1973.

15. *The Real* Wealth of Nations: Pathways to Human Development, Human Development Report 2010: 20th Anniversary Edition, UNDP, 2010, 227 pp.

Рекомендована Радою
Навчально-наукового комплексу
"Інститут прикладного системного
аналізу" НТУУ "КПІ"

Надійшла до редакції
18 травня 2012 року