УДК 004.8

Н.Р. Кондратенко, к.т.н., доцент; О.О. Манаєва, магістрантка

ВИКОРИСТАННЯ ІНТЕРВАЛЬНИХ ФУНКЦІЙ НАЛЕЖНОСТІ В ЗАДАЧАХ КЛАСТЕРИЗАЦІЇ ДАНИХ СОЦІАЛЬНОГО ХАРАКТЕРУ

Вступ

У сучасній Україні та 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 при цьому розглядаються, ставить особливі вимоги до математичних методів розв’язання цих задач. Передумовою побудови достовірних математико-статистичних моделей в 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 найбільш потужним інструментом для проведення багатомірних досліджень. Його застосування в таких задачах є цілком виправданим, оскільки перше застосування кластерний аналіз знайшов 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 ніяких обмежень на вид об'єктів, що розглядаються, і дозволяє оперувати множиною вихідних даних практично довільної природи [2]. Це дає змогу говорити про можливість створення універсальних 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 даної тематики. Зокрема, в роботах [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, запропонований в роботі [4], попри високі оптимізаційні властивості, ставить аналогічну вимогу. Відомо, що на практиці такі умови трапляються вкрай рідко, особливо в галузі соціології, всі показники якої ґрунтуються на результатах соціологічних опитувань та офіційних даних, наданих різного роду урядовими організаціями. Стовідсоткової достовірності таких даних гарантувати не може ніхто, тому дана задача вимагає методів кластерного аналізу, стійких до викидів та шуму. Один із таких методів – PCM (Possibilistic C-Means) – 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 закладена певна невизначеність, яка не може не перенестись на результат кластеризації. При цьому характеризувати ступінь належності точки до кластеру одним числом недостатньо. Внаслідок дії невизначеностей самé 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 сьогоднішній день 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] та ін.

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.

Особливості реалізації моделей кластерного аналізу

Існує велика кількість методів кластеризації, які можна класифікувати на чіткі та нечіткі. Чіткі методи кластеризації розбивають вихідну множину об'єктів **X** на декілька підмножин, що не перетинаються. При цьому будь-який об'єкт із **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 при нечіткому розбитті 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 є "природнішою" за чітку, наприклад, для об'єктів, розташованих на межі кластерів [2, 17].

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

Проте якість знайдених центрів суттєво залежить від попереднього вибору як значень μ_{ij}, так і центрів c_j. Крім того, FCM використовує обмеження, подібне до того, що накладає на шуканий розв'язок теорія ймовірностей: сума ступенів належності і-тої точки до всіх кластерів $j = \overline{1, N}$ становить 1: $\sum_{i=1}^{c} \mu_{ij} = 1$ для всіх j. Таке обмеження має на меті уникнути тривіального розв'язку, коли всі ступені належності виявляються рівними 0, і дає змістовні результати в тих прикладних застосуваннях, де припущення про «ймовірнісну» природу ступенів належності має практичний сенс.

Але, оскільки ступені належності, отримані при такому обмеженні, відносні, вони непридатні в тих задачах, у яких ступінь належності точки до кластеру повинен відображати її типовість, характерність саме для цього кластеру. 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 є більш придатним для більшості задач кластеризації, оскільки ступінь належності точки до кластеру є мірою того, наскільки ця точка є носієм спільних характеристик кластеру, її типовості, і не повинен залежати від того, як вона розташована відносно інших кластерів.

Виходячи з цього, в роботі [5] було переглянуто цільову функцію методу FCM таким чином, щоб при досягненні 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:

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

де η_i – додатне число.

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

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

$$\mu_{ij} = \frac{1}{1 + \left(\frac{d_{ij}^2}{\eta_i} \right)^{\frac{1}{m-1}}} \qquad \eta_i = \frac{\sum_{j=1}^{N} \mu_{ij}^{m} d_{ij}^{2}}{\sum_{j=1}^{N} \mu_{ij}^{m}}$$

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, отримані при такому підході, більше відповідають дійсності та інтуїтивному уявленню про природу кластерів. Таке розуміння ступенів належності має ще один позитивний момент: воно дає змогу легко відфільтрувати точки, що вносять шум, оскільки вони при такому формулюванні 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. 2 (для випадку двох кластерів). Кожна крива зображає зміну ступеня належності точки до одного з кластерів.

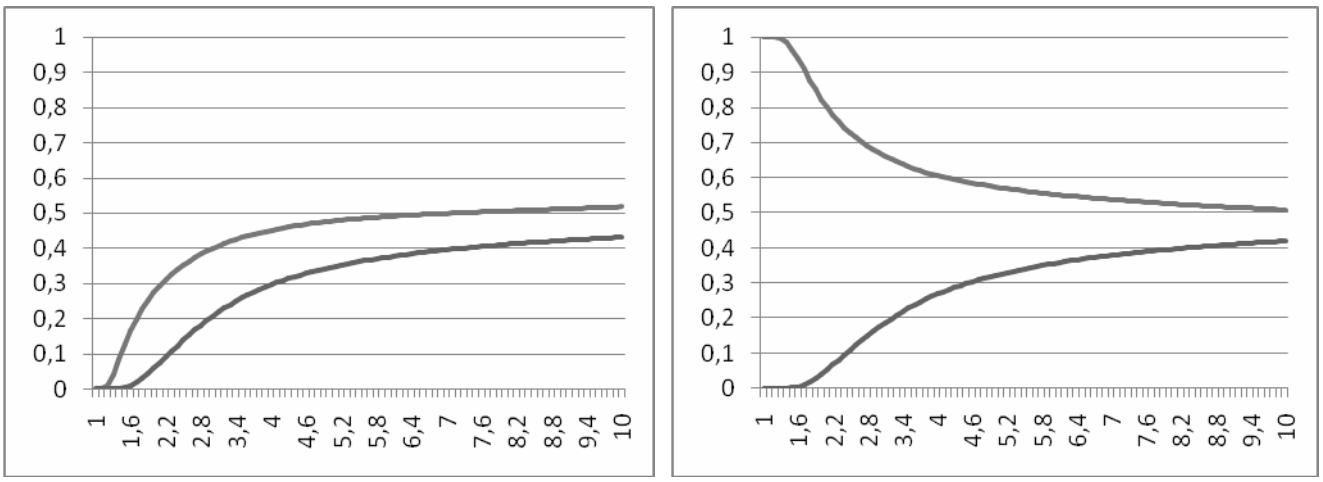


Рисунок 2 – Зміна ступенів належності точки до кластерів при зміні рівня нечіткості:
а) точка нерепрезентативна;
б) точка репрезентативна.

На рис. 2 видно, що в усіх випадках m змінюється монотонно, вибрати 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 кластеру являє собою єдине число. Тому для того, щоб убезпечити себе від помилкового результату, пов’язаного з 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 параметру m, має виняткову цінність, оскільки інтервал (його ширина та розташування відносно нуля та одиниці) несе значно більше інформації про міру належності точки до кластеру, ніж єдине число. Наприклад, ширина інтервалу може свідчити про ступінь точності отриманого розв’язку. Тому пропонується модифікувати алгоритм кластеризації, наведений 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.

Постановка задачі та методика дослідження

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 кластеру c_j . Виходячи з визначення ступеня належності як міри типовості заданої точки для відповідного кластеру, знайти такі значення шуканих параметрів, які ведуть до мінімуму функціоналу (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].

В основі пропонованого методу лежить алгоритм кластеризації 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. Дана властивість є надзвичайно корисною, оскільки вона дає змогу «розтягти» ступінь належності точки до кластеру 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, в 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 обійти цю проблему – виконувати розбиття при різній можливій кількості кластерів та порівнювати результати за певним критерієм оптимальності. В роботі [19] наведено декілька функціоналів, які називаються індексами достовірності та цілком відповідають вимогам, що висуває дана 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 кластеру i;

v_j - центр j-того кластеру;

$\bar{v}$ - середнє значення центрів кластерів;

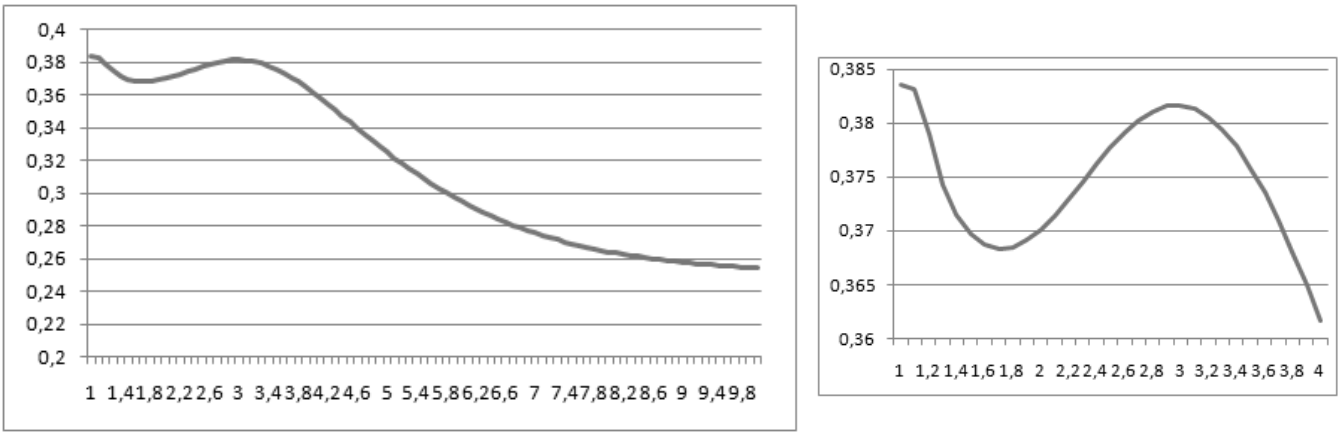
m – рівень нечіткості;

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. В рамках даного підходу пропонується використовувати його також для визначення меж інтервалу розтягу ступеня належності. За одну з меж інтервалу логічно прийняти значення ступеня належності точки до кластеру при m = 1, що відповідає випадку чіткої кластеризації. Іншою ж його межею буде значення ступеня належності при певній величині m > 1, зумовленій поведінкою індекса Квона на заданому інтервалі зміни параметру m (рис. 4, а). Практичний інтерес викликає лише перший його локальний мінімум, який спостерігається при оптимальному значенні m (рис. 4, б). Після його досягнення функція $V_k(c)$ веде себе по-різному; в даній задачі це несуттєво.



а) б)

Рисунок 4 – Зміна індексу Квона залежно від зміни рівня нечіткості:
а) в межах від 1 до 10;
б) збільшений фрагмент: від 1 до 4.

В роботі [19] стверджується, що перший локальний мінімум $V_k(c)$ відповідає оптимальному значенню m. Отже, другу межу зміни інтервалу ступеня належності буде задаватися саме цією величиною.

Таким чином, ступінь належності точки до кожного з кластерів буде лежати в межах $[\mu_{ijL}, \mu_{ijU}]$, де границями інтервалу є відповідні значення μ_{ij} при m = 1 та в точці першого локального мінімуму індексу Квона. Тобто отриманий розв'язок з одного боку обмежується випадком чіткої кластеризації, що дає змогу з першого ж погляду однозначно віднести точку до того чи іншого кластеру, а з іншого – певним нечітким значенням, за яким можна судити про те, наскільки типовою (нетиповою) є точка для даного кластеру.

При такому підході отриманий нечіткий кластер матиме вигляд, як показано на рис. 5. Для його повного опису достатньо визначити лише верхню та нижню функції належності.

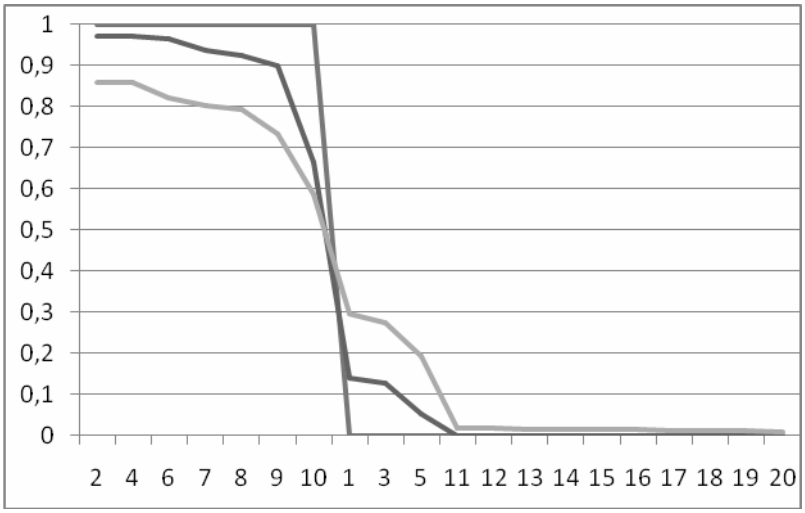


Рисунок 5 – Інтервальні функції належності точок до кластеру

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.

Виходячи з усього сказаного вище, 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 центру 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 вектору 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, збільшити m та повернутись до кроку 8.
14. Перерахувати значення глобального індексу Квона. Зберегти проміжні результати обчислень для m = 1 та поточного $m = m_k > 1$.

15. Якщо кількість кластерів менша за кількість точок у вихідній множині, збільшити c та перейти до кроку 3.
16. Серед усіх проміжних результатів вибрати 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 $m = m_k > 1$.

Кластерний аналіз країн світу за рівнем розвитку

Для аналізу було взято дані зі щорічного звіту ООН за 2010 рік [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.

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

Таблиця 1 – Координати центрів кластерів

	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 освіти	0,532329	0,837381	0,253012

Таблиця 2 – Інтервальні ступені належності країн до кластерів

	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. 6). Значна ширина інтервалу деяких конкретних значень ступенів належності дає змогу судити про наявність шумів у вхідних даних.

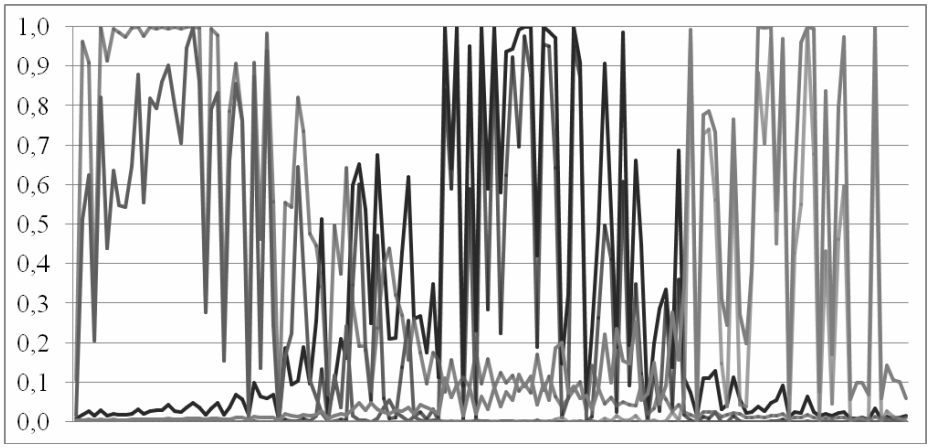


Рисунок 6 – Графічне представлення результатів кластеризації

Інтервальні значення ступенів належності було обчислено, виходячи зі значень рівня нечіткості m = [1; 1,65]. На рис. 6 показано характер зміни індексу Квона залежно від значення m.

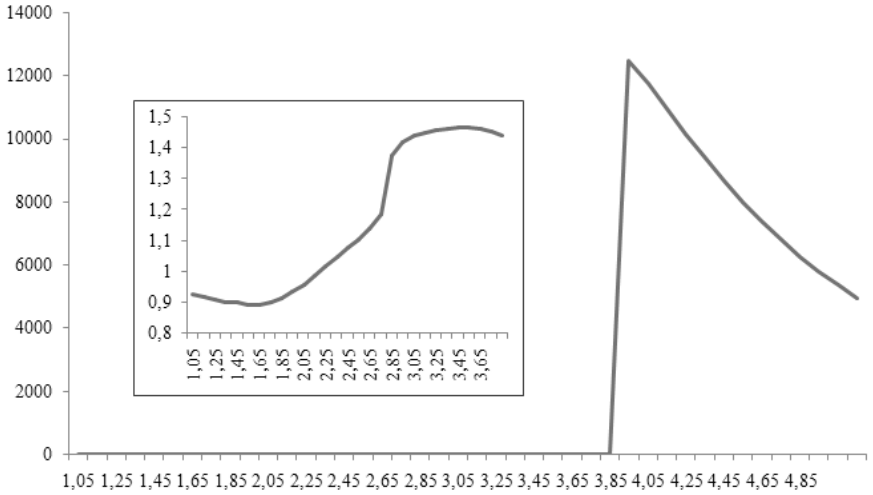


Рисунок 6 – Характер зміни індексу Квона залежно від значення рівня нечіткості

Отримані результати свідчать про те, що нечіткість другого порядку дає змогу повною мірою врахувати невизначеності, характерні для даної задачі, а саме такі, що пов'язані з неможливістю точно задати рівень нечіткості m. Значна ширина інтервалу деяких конкретних значень ступенів належності дає змогу судити про наявність шумів у вхідних даних. Зважаючи на це та високу складність поставленої прикладної задачі, для прийняття будь-якого остаточного рішення доцільно залучати експертів у даній галузі.

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. Таким чином, нечіткість другого порядку дає змогу повною мірою врахувати невизначеності, характерні для даної задачі, а саме такі, що пов'язані з неможливістю точно задати рівень нечіткості m.

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

СПИСОК ЛІТЕРАТУРИ

1. Котова Е. С. Кластерный анализ в задачах социально-экономического прогнозирования [Електронний ресурс] / Е. С. Котова. – Режим доступу з: 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 // Вестник СКГТУ. Серия «Экономика». №3 (11), 2003.
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. R. Krishnapuram, J.M. Keller. A Possibilistic Approach to Clustering. IEEE Transactions on Fuzzy Systems, 1(2):98-110, 1993.
6. L.A. Zadeh. Fuzzy sets as a basis for a theory of possibility. Fuzzy sets and systems 100 Supplement, pp. 9-34, 1999.
7. Q. Liang, J. M. Mendel . Interval type-2 fuzzy logic systems: Theory and design, IEEE Trans. Fuzzy Syst., vol. 8, pp. 535–550, 2000.
8. J. M. Mendel, R. I. John, F. Liu. Interval Type-2 Fuzzy logic systems made simple, IEEE Transactions on fuzzy systems, vol. 14, No. 6, pp. 808-821, 2006.
9. J. M. Mendel, R. I. John. Interval Type-2 Fuzzy sets made simple, IEEE Transactions on fuzzy systems, vol. 10, No. 2, pp. 117-127, 2002.
10. J. Zeng, Z. Q. Liu. Type-2 Fuzzy sets for pattern classification: A review, Proceedings of the IEEE Symposium on Foundations of computational intelligence, pp. 193-200, 2007.
11. Q. Liang, J. M. Mendel. MPEG MBR Video traffic modeling and classification using fuzzy technique, IEEE Transactions on fuzzy systems, vol. 9, No. 1, pp. 183-193, 2001.
12. K. C. Wu, Fuzzy interval control of mobile robots, Comput. Elect Eng., vol. 22, pp. 211–229, 1996.
13. R. R. Yager, Fuzzy subsets of type II in decisions, J. Cybern., vol. 10 pp. 137–159, 1980.
14. N. N. Karnik, J. M. Mendel, Applications of type-2 fuzzy logic systems to forecasting of time series, Information Sciences, vol. 120, pp. 89–111, 1999.
15. J. M. Mendel, Uncertainty, fuzzy logic, and signal processing, Signal Proc. J., vol. 80, pp. 913–933, 2000.
16. N. N. Karnik and J. M. Mendel. An Introduction to Type-2 Fuzzy Logic Systems. Univ. of Southern Calif., Los Angeles, CA. [Електронний ресурс]. – Режим доступу з:
http://sipi.usc.edu/~mendel/report
17. Штовба С.Д. Введение в теорию нечетких множеств и нечеткую логику [Електронний ресурс] / С.Д. Штовба. – Режим доступу з:
http://matlab.exponenta.ru/fuzzylogic/book2/index.php.
18. Зайченко Ю.П. Нечеткие модели и методы в интеллектуальных системах. – К.: «Издательский дом «Слово», 2008. – 344 с.
19. J.V. Oliveira, W. Pedrycz. Advances in Fuzzy Clustering and Its Applications. John Wiley & Sons Ltd., 2007. – 435 pp.
20. The Real Wealth of Nations: Pathways to Human Development. Human Development Report 2010: 20th Anniversary Edition . – UNDP, 2010. – 227 pp.

Кондратенко Наталія Романівна – к.т.н., доцент, професор кафедри захисту інформації.
Вінницький національний технічний університет.

Манаєва Ольга Олексіївна – магістрантка.
Вінницький національний технічний університет.