УДК 519.237.8

Н.Е. КОНДРУК, М.М. МАЛЯР

АНАЛІЗ КЛАСТЕРНИХ СТРУКТУР ЗА РІЗНИМИ МІРАМИ ПОДІБНОСТІ

Анотація. Наведено аналіз кластерних утворень, що використовують в практичних задачах. У різних дослідженнях сегментацію даних зазвичай виконують лише одною формою кластерів. Запропоновано здійснювати кластеризацію за різними мірами подібності одних і тих самих досліджуваних даних та виявляти різні види взаємозв'язків між ними. Це дає змогу проводити більш повний, різнобічний та системний аналіз утворених сегментів у прикладних задачах. Верифікацію цього підходу реалізовано на практичній задачі аналізу демографічних процесів у низці європейських країн.

Ключові слова: кластеризація, аналіз кластерів, інтерпретація кластерів, демографічні процеси.

ВСТУП

Останнім часом технологію кластеризації використовують для аналізу даних різної природи, зокрема за відсутності попередньої інформації про кількість кластерів та навчальну вибірку. Спектр застосування кластерного аналізу дуже широкий та представлений в багатьох роботах [1–12]. Це дослідження присвячене підходам до кластеризації, за допомогою яких потрібно визначити не лише кластерну структуру об'єктів (як у задачах розпізнавання образів [1]), а й змістовно інтерпретувати отримані сегменти (наприклад, ринку, цільової аудиторії, у маркетингових дослідженнях тощо) [2–4]. При цьому обґрунтованість висновків інколи не є достатньо переконливою, оскільки можливо сформувати різні види кластерів на основі тих самих даних залежно від методу або способу проведення аналізу. Важливе значення для отримання коректних результатів має вибір міри подібності, яка не спотворює взаємозв'язків між об'єктами у разі, якщо ці взаємозв'язки потребують змістовної інтерпретації [1]. Це є першим кроком до визначення валідності кластерів, який відбувається ще до етапу аналізу. В одному методі кластеризації реалізують тільки одну міру подібності. У багатьох програмних пакетах і алгоритмах зазвичай використовують евклідову відстань, що зумовлює утворення еліпсоїдних кластерів. Тому пропонується проводити аналіз даних за різними мірами подібності. Придатним інструментарієм для цього є метод кластеризації, що базується на нечітких бінарних відношеннях [5] та здійснює кластеризацію еліпсоїдну, конусну та концентричними сферами.

ОГЛЯД СУЧАСНИХ ДОСЛІДЖЕНЬ З ПРИКЛАДНОЇ КЛАСТЕРИЗАЦІЇ

Інтерпретація кластерної структури наборів даних є одним із важливих етапів прикладної кластеризації. Метою проведення кластерного аналізу в таких задачах є аналіз утвореної сегментації. Це дає можливість отримувати нову необхідну інформацію для прийняття рішень у багатьох сферах діяльності.

Зокрема, в [6, 7] здійснено кластерний аналіз поведінки користувачів електроенергії під час використання розумних лічильників для планування роботи розподільних мереж. Було сформовано десять різних груп поведінки клієнтів. У праці [8] визначено групи ризику набуття надмірної ваги та виникнення ожиріння серед дітей і підлітків на основі показників фізичної активності та харчових звичок. Поведінку користувачів за даними потоків кліків у реальних соціальних мережах

проаналізовано в [9]. Виконано групування школярів [10] та офісних робітників [11] на основі їхнього сприйняття комфорту і важливості факторів навколишнього середовища з подальшою інтерпретацією утворених груп. Використанню кластерного аналізу в маркетингових дослідженнях туристичної галузі, просуванню мобільних застосунків та ринку електромобілів присвячено роботи [2–4]. У цих дослідженнях на базі змістовної інтерпретації утворених сегментів цільових користувачів побудовано їхній профіль. Також у [12] проведено аналіз сегментів, утворених на основі інтенсивності міграційного руху населення в регіонах України. Однак в усіх згаданих застосуваннях кластерного аналізу реалізовано лише одну міру подібності даних. Для забезпечення системного підходу до аналізу сегментацій пропонується використовувати три міри подібності на базі методу кластеризації, який ґрунтується на нечітких бінарних відношеннях [5], що дає змогу здійснювати більш повний, різнобічний аналіз утворених кластерів.

ПОСТАНОВКА ЗАДАЧІ. МЕТОДИ ПРИКЛАДНОЇ КЛАСТЕРИЗАЦІЇ

Нехай об'єкти (дані) $O_1, \dots, O_m$ характеризуються n кількісними ознаками (атрибутами). Кожному об'єкту O_i, $i = \overline{1, m}$, ставиться у відповідність вектор ознак $\bar{c}_i (c_1^i, c_2^i, \dots, c_n^i)$, $i = \overline{1, m}$. Згрупувавши вектори ознак, сегментуємо також об'єкти.

Ставиться задача виконати кластеризацію цих об'єктів за різними мірами подібності та змістовно інтерпретувати отримані сегменти.

Для реалізації процесу кластеризації вибрано метод, що базується на нечітких бінарних відношеннях [5, п. 6]. Він дає можливість групувати об'єкти кластерами різних геометричних форм, змінюючи лише міри подібності об'єктів. Для визначення кількості кластерів задають певні величини — пороги кластеризації μ_R^*, що характеризують ступінь подібності об'єктів у середині кластера. Змінюючи пороги кластеризації, можна проаналізувати динаміку формування кластерів, дослідити їхню структуру та взаємозв'язки між об'єктами.

Подібність об'єктів за деяким критерієм характеризується нечітким бінарним відношенням R на множині векторних ознак $C = \{\bar{c}_i \mid i = \overline{1, m}\}$ із функцією належності $\mu_R (\bar{c}_i, \bar{c}_j)$, де $\mu_R : C^2 \to [0, 1]$. Види функції μ_R можуть бути різними, але чим її значення ближче до 1, тим більшою мірою об'єкти O_i та O_j мають бути подібними за певним критерієм. Для дослідження використано три типи подібності об'єктів.

Міру подібності «відстань» описують нечітким бінарним відношенням R^{dis} та функцією належності вигляду

$$\mu_{R^{\text{dis}}} (\bar{c}_i, \bar{c}_j) = \exp \left\{ - \frac{| \bar{c}_i - \bar{c}_j |}{\max\limits_{p, l = \overline{1, m}} | \bar{c}_p - \bar{c}_l |} \right\}. \qquad (1)$$

Тобто, чим ближчими будуть точки C_i та C_j, тим ближче значення $\mu_{R^{\text{dis}}}$ буде до 1. Використання однорівневого методу кластеризації, основаного на нечітких бінарних відношеннях за мірою подібності R^{dis}, забезпечує утворення еліпсоподібних кластерів.

Міру подібності «довжина» описують бінарним відношенням R^{len} із функцією належності такого вигляду:

$$\mu_{R^{\text{len}}} (\bar{c}_i, \bar{c}_j) = \exp \left\{ - \frac{| \, | \bar{c}_i \, | - | \bar{c}_j \, | \, |}{\max\limits_{p = \overline{1, m}} | \bar{c}_p | - \min\limits_{p = \overline{1, m}} | \bar{c}_p |} \right\}. \qquad (2)$$

Ця міра визначає подібність векторів-ознак за довжиною та зумовлює утворення кластерів у формі концентричних сфер.

«Кутове» нечітке бінарне відношення R^{ang} характеризує кут між векторами ознак $\bar{c}_i$ та $\bar{c}_j$. Його використання дає змогу здійснювати кластеризацію конічними кластерами. Функцію належності цього відношення визначають формулою

$$\mu_{R^{\text{ang}}}(\bar{c}_i, \bar{c}_j) = \exp\left\{-\frac{1-\dfrac{\bar{c}_i \cdot \bar{c}_j}{|\bar{c}_i||\bar{c}_j|}}{2}\right\}. \tag{3}$$

У кожному із випадків (1)–(3) аргументом експоненти є нормовані величини, що змінюються від 0 до 1. Тому її значеннями будуть відповідно величини від 1 до $1/e$. Крім того, проведення практичних експериментів показало, що «хороша» чутливість експоненти в околі свого граничного значення 1 дає змогу виконувати кластеризацію об'єктів для всіх можливих порогових величин μ_{R^*} проміжку [0; 1] із певною точністю (наприклад, із точністю 0.01).

КЛАСТЕРИЗАЦІЯ НИЗКИ ЄВРОПЕЙСЬКИХ КРАЇН ЗА ПРИРОДНИМ І МІГРАЦІЙНИМ ПРИРОСТОМ НАСЕЛЕННЯ

Для сегментування даних у практичних задачах різними геометричними формами кластерів розроблено програмну систему, що реалізує кластеризацію, основану на нечітких бінарних відношеннях за мірами подібності (1)–(3), та вибрано двовимірні дані, оскільки це надає додаткову можливість верифікувати отримані результати дослідження. У наш час у багатьох країнах світу відбуваються швидкі зміни в демографічних процесах, що має величезний вплив на економіку. Тому для проведення аналізу використано реальні дані із офіційного сайту Євростату [13] про природний та міграційний приріст (скорочення) населення в 29 європейських країнах за 2019 р. Наведемо перелік цих країн (об'єктів): 1 — Бельгія, 2 — Болгарія, 3 — Чехія, 4 — Данія, 5 — Німеччина, 6 — Естонія, 7 — Ірландія, 8 — Греція, 9 — Іспанія, 10 — Франція, 11 — Хорватія, 12 — Італія, 13 — Латвія, 14 — Литва, 15 — Угорщина, 16 — Мальта, 17 — Нідерланди, 18 — Австрія, 19 — Польща, 20 — Португалія, 21 — Румунія, 22 — Словаччина, 23 — Фінляндія, 24 — Швеція, 25 — Великобританія, 26 — Ісландія, 27 — Ліхтенштейн, 28 — Норвегія, 29 — Швейцарія.

Для порівняння отриманих результатів кластеризації за різними типами подібності вибрано порогові значення μ^*_R, що відповідають структурі із трьох кластерів.

У разі використання порогових даних $\mu^*_{R^{\text{dis}}} = 0.75$ за мірою подібності «відстань» (1) отримано такі групи об'єктів:

кластер 1 — об'єкти за номерами 1, 2, 3, 4, 6, 7, 8, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 26, 27, 28, 29;

кластер 2 — об'єкти за номерами 5, 9, 12;

кластер 3 — об'єкти за номерами 10, 25.

На рис. 1 наведено графічну інтерпретацію кластеризації на основі міри подібності R^{dis}, де x_1 — природний приріст населення, x_2 — міграційний.

Результати групування за мірою подібності R^{len} (2) на три кластери відповідають пороговому значенню $\mu^*_{R^{\text{len}}} = 0.84$:

кластер 1 — об'єкти за номерами 1, 2, 3, 4, 6, 7, 8, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 26, 27, 28, 29;

кластер 2 — об'єкти за номерами 10, 12;

кластер 3 — об'єкти за номерами 5, 9, 25.

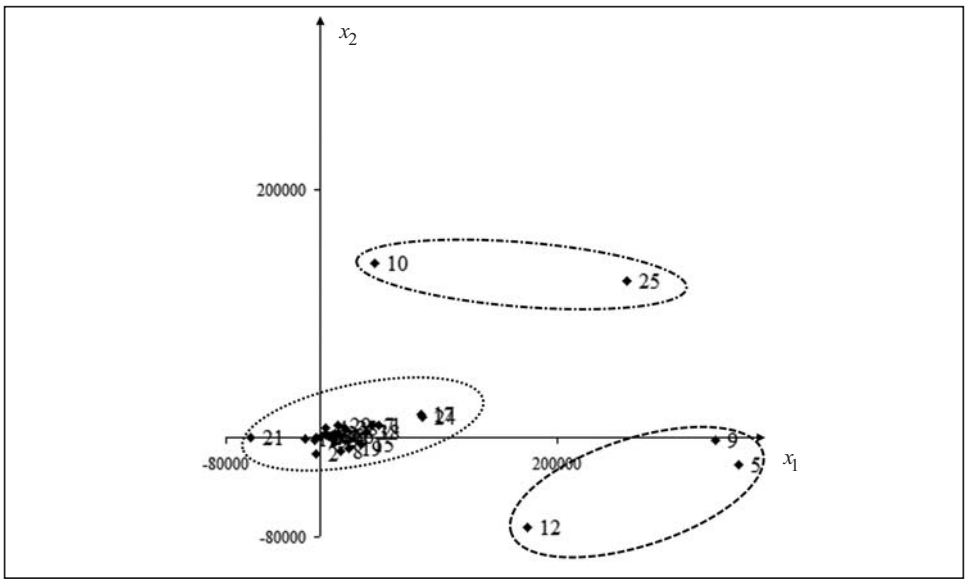

Рис. 1

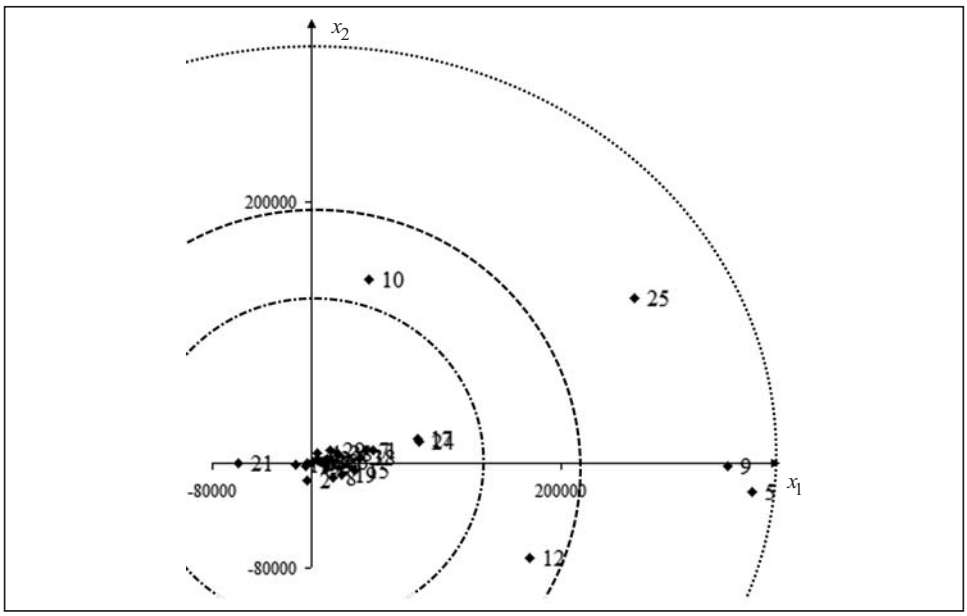

Рис. 2

На рис. 2 візуалізовано кластери, утворені за мірою подібності R^{len}.

За мірою подібності R^{ang} та значенням порогового коефіцієнта $\mu^{*}_{R^{\text{ang}}} = 0.84$ сформовано три кластери:

кластер 1 — об'єкти за номерами 2, 11, 13, 14, 21;

кластер 2 — об'єкти за номерами 8, 12, 15, 19, 20, 5;

кластер 3 — об'єкти за номерами 1, 3, 4, 6, 7, 9, 10, 16, 17, 18, 22, 23, 24, 25, 26, 27, 28, 29.

Для унаочнення всі вектори ознак переведено в орти, оскільки функція належності конусної міри подібності $\mu_{R^{\text{ang}}}$ враховує лише кут між векторами і не залежить від їхньої довжини (рис. 3).

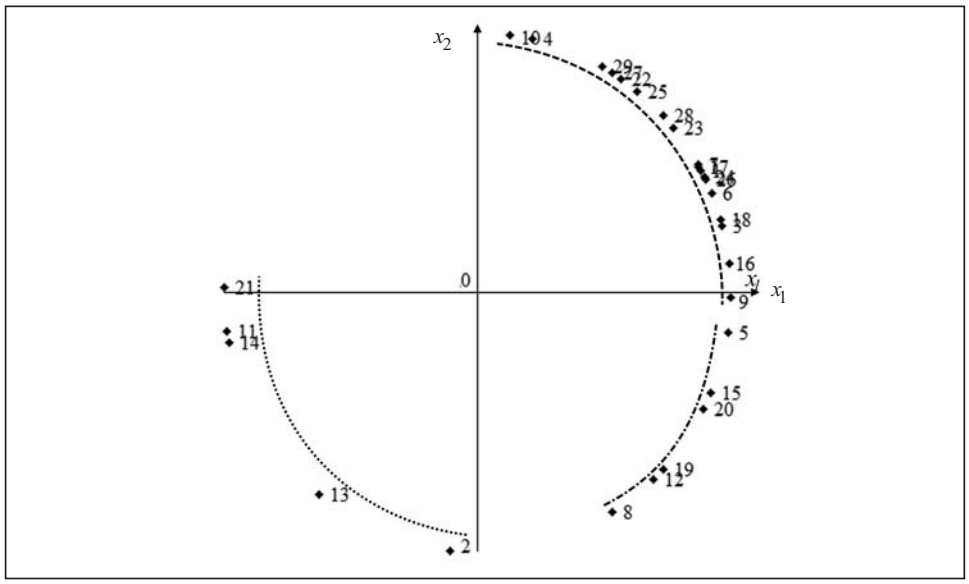

Рис. 3

Із одержаних результатів випливає, що кластеризація, проведена за мірами подібності (1)–(3) для одних і тих самих наборів даних є різною. Далі наведемо одну із можливих змістовних інтерпретацій отриманих груп.

Використання міри подібності «відстань» зумовлює утворення еліптичних кластерів. Сформовану у такий спосіб структуру можна інтерпретувати як таку, що групує країни за числовою величиною міграційного та природного руху населення. Кластер 1 містить країни із низькими показниками міграційного та природного приросту чи скорочення населення; кластер 2 (Німеччина, Іспанія, Італія) — із високим міграційним приростом, але природним скороченням населення; кластер 3 (Франція, Великобританія) — із високим природним та міграційним приростом. Цікавим є той факт, що для формування кластера 1 ключовими були абсолютні величини обох ознак, для кластера 2 групування здійснювалося за величиною міграційного приросту (за ознакою природного скорочення враховувався лише його від'ємний знак).

У разі групування об'єктів кластерами у формі концентричних сфер знаки числових величин ознак до уваги не беруть. Кластерна структура при цьому може бути інтерпретована як така, що визначає інтенсивність прояву ознак. Кластер 1 і для цієї структури не змінився та містить 24 країни із слабо вираженою інтенсивністю приросту населення, кластер 2 (Італія, Франція) — країни із середньою інтенсивністю міграційного та природного приросту населення, а кластер 3 (Німеччина, Іспанія, Великобританія) — із значною інтенсивністю цих ознак.

Кутова міра подібності (3) формує кластери конічної форми. Ця структура визначає країни із схожим характером ознак. До кластера 1 (Болгарія, Хорватія, Латвія, Литва, Румунія) належать країни із скороченням населення за обома атрибутами (зокрема, Румунія має близький до 0 природний приріст). Кластер 2 (Греція, Італія, Угорщина, Польща, Португалія, Німеччина) можна інтерпретувати як такий, що містить країни із міграційним приростом та природним скороченням населення, кластер 3 — із приростом населення за цими самими ознаками. Це твердження є не зовсім коректним щодо об'єкта за номером 9 (Іспанія) кластера 3, де спостерігається «від'ємний» природний приріст, але він близький до 0. Слід зауважити, що для всіх граничних об'єктів кластерів загальний харак-

тер поведінки ознак у кожному сегменті може дещо відрізнятись, але метод, що ґрунтується на нечітких бінарних відношеннях [5, п. 6], також має інструменти визначення граничних об'єктів для кожного кластера.

Це дослідження є продовженням робіт [5, 12, 14, 15] стосовно використання інструментів кластерного аналізу для аналітики прикладних даних. У перспективі передбачається розробити узагальнений індекс для визначення оцінки якості кластерної структури, утвореної на основі мір подібності (2), (3); побудувати автоматизовану інтелектуальну систему, яка б надавала змогу реалізовувати кластерний аналіз даних різної природи за різними критеріями подібності.

ВИСНОВКИ

Розглянуто проблему використання системного підходу в кластерному аналізі до дослідження наборів даних на існування різних видів зв'язків. Вперше до одного і того самого набору даних застосовано кластеризацію на основі нечітких бінарних відношень [5] за різними мірами подібності (1)–(3). Показано, що використання мір подібності за відстанню, довжиною та кутом дає можливість виявити різні кластерні структури, що характеризують різні змістовні інтерпретації зв'язків об'єктів у кожному окремому сегменті. Такий підхід сприяє більш повному розумінню даних у разі розв'язання задачі визначення індикації змісту отриманого кластера. Також розроблено програмну систему, яка реалізує метод кластеризації, що ґрунтується на нечітких бінарних відношеннях R^{dis}, R^{len}, R^{ang} [5], та дає змогу утворювати кластери різних геометричних форм (еліпсоїдні, конусоподібні, концентричних сфер). Виконано апробацію системи на актуальній прикладній задачі кластеризації країн Європи за міграційним та природним приростом (скороченням) за 2019 рік.

СПИСОК ЛІТЕРАТУРИ

1. Hulianytskyi L.F., Riasna I.I. Automatic classification method based on a fuzzy similarity relation. *Cybernetics and Systems Analysis.* 2016. Vol. 52, N 1. P. 30–37. https://doi.org/10.1007/s10559-016-9796-3.
2. Lascu D.-N., Manrai L.A., Manrai A.K., Gan A. A cluster analysis of tourist attractions in Spain: Natural and cultural traits and implications for global tourism. *European Journal of Management and Business Economics*. 2018. Vol. 27, N 3. P. 218–230. https://doi.org/10.1108/EJMBE-08-2017-0008.
3. Sanders I., Short C.E., Bogomolova S., Stanford T., Plotnikoff R., Vandelanotte C., Olds T., Edney S., Ryan J., Curtis R.G., Maher C. Characteristics of adopters of an online social networking physical activity mobile phone app: Cluster analysis. *JMIR Mhealth Uhealth*. 2019. Vol. 7, N 6. P. 1–11. https://doi.org/10.2196/12484.
4. Morton C., Anable J., Nelson J.D. Consumer structure in the emerging market for electric vehicles: Identifying market segments using cluster analysis. *International Journal of Sustainable Transportation*. 2017. Vol. 11, N 6. P. 443–459. https://doi.org/10.1080/15568318.2016.1266533.
5. Kondruk N. Clustering method based on fuzzy binary relation. *Eastern-European Journal of Enterprise Technologies.* 2017. Vol. 2, N 4. P. 10–16. https://doi.org/10.15587/1729-4061.2017.94961.
6. Haben S., Singleton C., Grindrod P. Analysis and clustering of residential customers energy behavioral demand using smart meter data. *IEEE Trans. on Smart Grid*. 2015. Vol. 7, N 1. P. 136–144. https://doi.org/10.1109/TSG.2015.2409786.
7. Geng D., Xia X., Fu X. A knowledge discovery method for characteristics extraction of power consumption based on cluster analysis. *Chinese Automation Congress* (22–24 Nov 2019, Hangzhou, China). Hangzhou, China, 2019. P. 1288–1292. https://doi.org/10.1109/CAC48633.2019.8996438.

8. Leech R.M., McNaughton S.A., Timperio A. The clustering of diet, physical activity and sedentary behavior in children and adolescents: a review. *International Journal of Behavioral Nutrition and Physical Activity.* 2014. Vol. 11, N 4. P. 1–9. https://doi.org/10.1186/1479-5868-11-4.
9. Wang G., Zhang X., Tang S., Zheng H., Zhao B. Unsupervised clickstream clustering for user behavior analysis. *CHI Conf. on Human Factors in Computing Systems* (6–12 May 2016, New York). New York, 2016. P. 225–236. https://doi.org/10.1145/2858036.2858107.
10. Zhang D., Ortiz M.A., Bluyssen P.M. Clustering of Dutch school children based on their preferences and needs of the IEQ in classrooms. *Building and Environment.* 2019. Vol. 147. P. 258–266. https://doi.org/10.1016/j.buildenv.2018.10.014.
11. Kim D.H., Bluyssen P.M. Clustering of office workers from the OFFICAIR study in The Netherlands based on self-reported health and comfort. *Building and Environment.* 2020. Vol. 176. P. 1–19. https://doi.org/10.1016/j.buildenv.2020.106860.
12. Kondruk N.E. Use of length-based similarity measure in clustering problems. *Radio Electronics. Computer Science. Control*. 2018. N 3. P. 98–105. https://doi.org/10.15588/1607-3274-2018-3-11.
13. Eurostat. European Statistical System. URL: https://ec.europa.eu/.
14. Kondruk N.E. A comparative study of cluster validity indices. *Radio Electronics. Computer Science. Control.* 2019. N 4. P. 59–67. https://doi.org/10.15588/1607-3274-2019-4-6.
15. Маляр М.М., Кондрук Н.Е. Структурування критеріального простору за кутовою мірою подібності. *Наук. вісн. Ужгород. ун-ту. Серія «Математика і інформатика»*. 2020. Вип. № 1 (36). С. 85–91. https://doi.org/10.24144/2616-7700.2020.1(36).85-91.

Надійшла до редакції 16.09.2020

Н.Э. Кондрук, Н.Н. Маляр
АНАЛИЗ КЛАСТЕРНЫХ СТРУКТУР ПО РАЗНЫМ МЕРАМ СХОДСТВА

Аннотация. Приведен анализ кластерных образований, используемых в практических задачах. В различных исследованиях сегментацию данных обычно выполняют только одной формой кластеров. Предложено осуществлять кластеризацию разными мерами сходства одних и тех же исследуемых данных и выявлять различные виды взаимосвязей между ними. Это позволяет проводить более полный, разносторонний и системный анализ образованных сегментов в прикладных задачах. Верификация такого подхода реализована на практической задаче анализа демографических процессов в некоторых европейских странах.

Ключевые слова: кластеризация, анализ кластеров, интерпретация кластеров, демографические процессы.

N.E. Kondruk, M.M. Malyar
ANALYSIS OF CLUSTER STRUCTURES BY DIFFERENT SIMILARITY MEASURES

Abstract. The cluster analysis formations used in practical tasks is presented. In various studies, data segmentation is usually performed with only one type of clusters. It is proposed to carry out clustering by various similarity measures to the same investigated data and to identify different types of relationships between them. This allows for a more complete, versatile, and systematic analysis of the formed segments in applied problems. This approach is verified using a practical problem of analyzing demographic processes in some European countries.

Keywords: clustering, cluster analysis, cluster interpretation, demographic processes.

Кондрук Наталія Емерихівна,
кандидатка техн. наук, доцентка кафедри Державного вищого навчального закладу «Ужгородський національний університет», e-mail: natalia.kondruk@uzhnu.edu.ua.

Маляр Микола Миколайович,
доктор техн. наук, професор кафедри Державного вищого навчального закладу «Ужгородський національний університет», e-mail: mykola.malyar@uzhnu.edu.ua.