УДК 519.237.8
DOI https://doi.org/10.24144/2616-7700.2024.44(1).106-113

І. В. Горват[1], Н. Е. Кондрук[2], Є. Б. Кондрук[3], В. А. Нерода[4]

[1] ДВНЗ «Ужгородський національний університет»,
аспірант кафедри кібернетики і прикладної математики
inna.horvat@uzhnu.edu.ua
ORCID: https://orcid.org/0009-0009-3686-3565

[2] ДВНЗ «Ужгородський національний університет»,
доцент кафедри кібернетики і прикладної математики,
кандидат технічних наук
natalia.kondruk@uzhnu.edu.ua
ORCID: https://orcid.org/0000-0002-9277-5131

[3] ДВНЗ «Ужгородський національний університет»,
аспірант кафедри кібернетики і прикладної математики
yevhen.kondruk@uzhnu.edu.ua
ORCID: https://orcid.org/0009-0002-8555-3351

[4] ДВНЗ «Ужгородський національний університет»,
аспірант кафедри кібернетики і прикладної математики
vladyslav.neroda@uzhnu.edu.ua
ORCID: https://orcid.org/0009-0009-2037-6631

СЕГМЕНТУВАННЯ КРАЇН ЄВРОСОЮЗУ ЗА ФІНАНСОВОЮ ДОПОМОГОЮ СТУДЕНТАМ

Використання технік кластеризації та їх порівняльного аналізу є невід'ємною складовою сучасних наукових досліджень через їх потенціал у виявленні структур та патернів у складних наборах даних. Ці техніки дозволяють класифікувати об'єкти за схожістю та групувати їх у кластери, що сприяє розумінню прихованих зв'язків та виявленню нових знань. Дослідження присвячено вивченню практичних аспектів використання технік кластеризації у задачі сегментування країн Євросоюзу за фінансовою допомогою студентам і включає в себе порівняльний аналіз методів кластеризації (k-Means, ієрархічної кластеризації), забезпечуючи цим об'єктивність та точність отриманих результатів. Використано різні індекси для визначення оптимальної кількості кластерів, такі як метод ліктя, метод силуету, метод Девіса-Болдіна та індекс Калінскі-Харабаса. Отримано чотири ідентичні кластери за обома методами, отже дані мають виражену структуру, яка однозначно інтерпретується як чотири різні категорії. Такий результат свідчить про консистентність та надійність знайдених кластерів, що додатково підтверджує значущість проведеної змістовної інтерпретації.

Ключові слова: кластерний аналіз, k-Means, ієрархічна кластеризація.

1. Вступ. Обробка та аналіз даних стає все більш важливим завданням у багатьох галузях, від науки до бізнесу. Одним із ключових етапів цього процесу є кластеризація — метод, що дозволяє групувати схожі об'єкти для подальшого вивчення їхніх властивостей та взаємодій. В цьому контексті, методи кластеризації виявляються вирішальними для отримання структурованої, неочевидної інформації з наборів даних. Однак вибір оптимального підходу групування для конкретних задач може бути нетривіальним, оскільки різні техніки мають власні особливості та обмеження. Тому порівняння методів кластерного аналізу виникає з необхідності вибору оптимального підходу до обробки даних в залежності від конкретного об'єкта дослідження та доменної області.

Фінансова допомога студентам є важливим аспектом освітнього процесу в багатьох країнах світу. З метою ефективного розподілу ресурсів та забезпечення максимальної користі виникає потреба в її аналізі.

Одними з потужних інструментів аналізу даних є методи кластеризації, такі як k-Means та ієрархічна кластеризація. Вони є популярними методами машинного навчання, які дозволяють автоматично групувати схожі об'єкти в окремі кластери та дозволяють виявити приховані закономірності, структури, що можуть бути важливими для подальшого аналізу та прийняття рішень.

Дослідження присвячено вивченню практичних аспектів використання техніки кластеризації у реальних задачах. Це може стати потужним інструментом для аналізу та управління фінансовою допомогою студентам і сприяти більш ефективному використанню ресурсів та забезпеченню максимальної користі для студентської громадськості.

Мета дослідження полягає в сегментуванні країн Євросоюзу за фінансовою допомогою студентам. Отримані результати можуть бути корисним для грантових організацій та урядових установ у прийнятті рішень щодо розподілу фінансових ресурсів в освітній сфері.

2. Огляд літератури. В [1] автори провели порівняльний аналіз методів k-Means та ієрархічної кластеризації і встановили, що k-Means ефективний у кластеризації великих наборів даних, і його продуктивність покращується зі збільшенням кількості кластерів. Для категоріальних даних було застосовано ієрархічний алгоритм, і відповідно до його складності використано новий підхід для надання рангових значень кожному категоріальному атрибуту. Встановлено, що центроїдний алгоритм ефективніший, ніж алгоритм ієрархічної кластеризації. Під час кластеризації певних (зашумлених) даних обидва методи мають певну неоднозначність.

Основний підхід авторів [2] полягав у розгляді даних з одним і двома кластерами відповідно до еталонної моделі очікуваних кластерів (включаючи ядро, перехідні та віддалені області) і застосуванні різних методів агломеративної кластеризації до цих даних.

У статті [3] автори основну увагу звернули на популярний алгоритм k-Means і проблеми ініціалізації та неможливості обробки даних із змішаними типами ознак. Експериментальний аналіз показав, що не існує універсального рішення для задач алгоритму k-means.

У [4] порівняльні дослідження авторів базувались на часі виконання та обсязі пам'яті, яку використовували техніки кластеризації. Щодо часу виконання метод k-середніх виявився кращим, для агломеративної ієрархічної кластеризації зі збільшенням розмірності наборів даних час виконання швидко зростав. Щодо використання пам'яті, темп збільшення використання пам'яті був вищий для агломеративної ієрархічної кластеризації, однак k-Means використовував більше пам'яті.

Дана праця присвячена дослідженню ефективності застосування інструментів кластерного аналізу для аналізу фінансової допомоги студентам в країнах Європейського союзу з урахуванням їхнього рівня освіти.

3. Постановка проблеми. Для досягнення поставленої мети необхідно вирішити наступні завдання:

– створити набір даних «Фінансова допомога студентам»;

– розробити інформаційно-аналітичну систему, яка реалізує класичні методи кластеризації k-середніх та ієрархічну, засоби визначення кількості кластерів;
– 3D-візуалізувати структуру даних, визначити оптимальну кількість кластерів;
– провести порівняльний аналіз отриманих результатів та їх інтерпретувати.

4. Методи дослідження. Метод k-Means [8] — це алгоритм кластеризації, який використовується для групування даних в k кластерів.

Він розділяє набір даних на попередньо задану кількість кластерів (k) так, щоб об'єкти всередині одного кластера були більш схожі один на одного, ніж на об'єкти з інших кластерів. Алгоритм включає ініціалізацію центроїдів, призначення об'єктів до найближчих центроїдів, перерахунок центроїдів та оцінку якості кластеризації.

Ієрархічна кластеризація [8] — це метод групування даних, де починаючи з кожного об'єкта як окремої групи, алгоритм поступово об'єднує інші кластери в залежності від їх схожості. Цей процес може бути агломеративним, коли кластери поступово об'єднуються, або дивізійним, коли відбувається розділення кластерів. В результаті отримуємо деревоподібну структуру, яку можна візуалізувати у вигляді дендрограми.

Індекс силуету (Silhouette Index) [5, 6] визначає кількість кластерів за допомогою внутрішнього індексу. Внутрішній індекс — це спосіб визначити валідність кластерів без зовнішньої інформації. Двома основними даними, які використовуються для визначення внутрішнього індексу є згуртованість і поділ кластерів. Згуртованість вимірює наскільки тісно дані пов'язані один з одним в одному кластері. Поділ кластерів показує наскільки кожен кластер відокремлений від іншого. Інструмент, який використовується для вимірювання згуртованості і поділу кластерів, полягає у вимірюванні їх відстані. Відстань за замовчуванням є Евклідовою, але є також інші: Манхеттенська, Мінковського, тощо.

Індекс ліктя (Elbow Index) [7] використовується для визначення кількості кластерів у наборі даних за допомогою візуалізації з обчислення суми квадратів квадратів відстаней між кожною точкою даних і центроїдом свого кластера. Цей показник дозволяє ідентифікувати точку на графіку, де зміна суми квадратів відстаней спадає раптово, утворюючи вигин «ліктя». Кількість кластерів визначається як точка, де спостерігається цей раптовий спад, що вказує на оптимальну кількість груп, яка найкраще відображає структуру даних.

Індекс Девіда-Болдіна (Davies_bouldin Index) [8] вимагає ініціалізації значень діапазону k. Для кожного значення k розраховується значення індексу Девіда-Болдіна і зберігається для відображення, коли значення k досягне максимуму, тоді виконується візуалізація і те значення яке найближче до 0 і є найкращим.

Індекс Калінські-Харабаса (Calinski_harabasz Index) [9] розраховується як відношення міжкластерної дисперсії до внутрішньокластерної дисперсії, з урахуванням кількості кластерів. Значення k, яке відповідає найбільшому числу, вибирається як оптимальна кількість кластерів для заданого набору даних.

5. Експерименти. У центрі уваги цього дослідження — застосування методів k-Means та ієрархічної кластеризації до задачі аналізу фінансової допомоги

студентам за рівнем освіти.

Збір даних про фінансову допомогу студентам за рівнем освіти — у відсотках від загальних державних видатків проводився із офіційного веб-сайту Eurostat (https://ec.europa.eu/eurostat). Дані складаються з окремих звітів по фінансовій допомозі студентам за різні роки. Крім того, на етапі попередньої обробки вони були усереднені за роками 2012–2014, 2015–2017, 2018–2020 і переведені у формат csv.

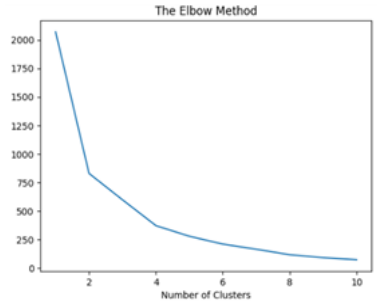
Для кластеризації за допомогою алгоритму k-Means необхідно спочатку визначити кількість кластерів. Для розв'язання цієї проблеми немає єдиного універсального способу. У даному дослідженні було використано кілька індексів визначення оптимальної кількості кластерів, а саме індекс ліктя, індекс силуету, індекс Девіса-Болдіна та індекс Калінскі-Харабаса. Кожен із них визначає найкраще значення кількості кластерів, вибір робиться на основі агрегованих показників більшості із них.

Індекс ліктя показав дві ліктьові точки, що визначило оптимальну кількість кластерів, як показано на рис. 1(a) на рівні 2 та 4.

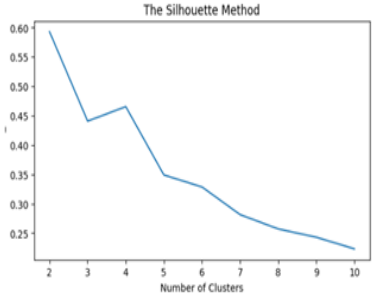
Індекс силуету оцінює якість кластеризації, враховуючи як внутрішню однорідність кластерів, так і різницю між ними. Оптимальна кількість кластерів визначається чотирма групами, де значення силуету досягає максимуму (рис. 1(b)).

Індекс Девіса-Болдіна оцінює якість кластеризації, враховуючи середню відстань між кластерними центрами та розміри кластерів. Оптимальна кількість кластерів визначилась чотирма — мінімумом індексу Девіса-Болдіна (рис. 1(c)).

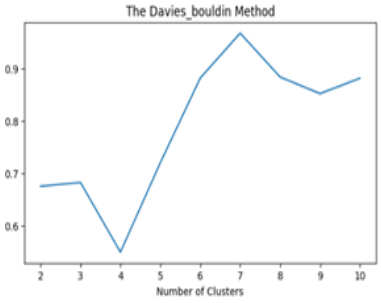
Індекс Калінскі-Харабаса оцінює якість кластеризації, враховуючи внутрішню однорідність кластерів та відстань між кластерними центрами. Оптимальна кількість кластерів — 4: найбільше значення індексу Калінскі-Харабаса рис. 1(d).



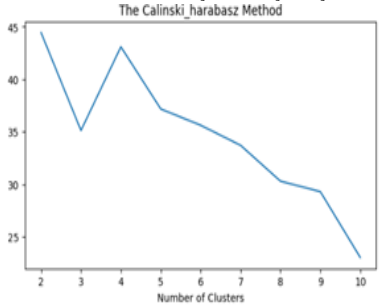
a. Індекс ліктя.



b. Індексу силуету.



c. Індекс Девіса-Болдіна.



d. Індекс Калінскі-Харабаса.

Рис. 1. Моделі визначення оптимальної кількості кластерів.

Провівши аналіз отриманих результатів обрано оптимальну кількість кластерів — 4. Наступний крок — візуалізація кластеризації. Для кращого розуміння результатів використаємо 3D графік. Останній крок — це інтерпретація результатів.

Для візуалізації результатів ієрархічної кластеризації використано деревоподібну структуру — дендрограму (рис. 2). На ній кожен лист відповідає окремому об'єкту, а висота кожної гілки показує відстань між об'єднаними кластерами. За дендрограмою визначено оптимальний рівень розподілу, що також відповідає 4 кластерам і найкраще відображає структуру даних шляхом обрізання дерева на певному рівні (червона горизонтальна лінія).

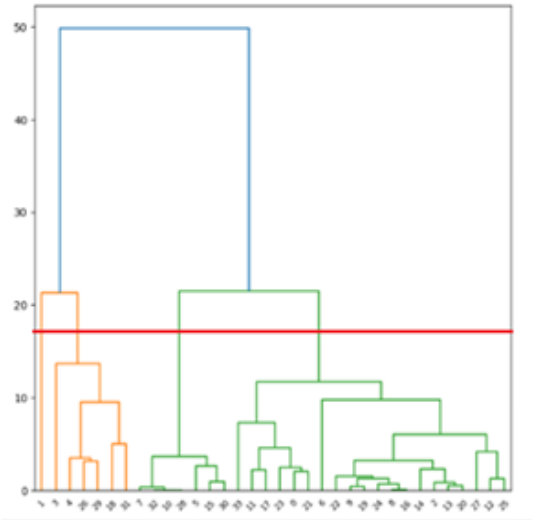


Рис. 2. Дендрограма ієрархічної кластеризації.

6. Результати дослідження. При кластеризації набору даних методом k-Means та методом ієрархічної кластеризації було отримано 4 ідентичних кластери (рис. 3):

Кластер 1: Естонія, Греція, Хорватія, Люксембург, Ліхтенштейн, Швейцарія, Сербія.

Кластер 2: Бельгія, Чехія, Ірландія, Іспанія, Франція, Італія, Кіпр, Латвія, Литва, Угорщина, Мальта, Австрія, Польща, Португалія, Румунія, Словенія, Словаччина, Фінляндія, Ісландія, Туреччина.

Кластер 3: Данія, Німеччина, Нідерланди, Швеція, Норвегія, Велика Британія.

Кластер 4: Болгарія.

7. Обговорення. Методи кластеризації k-Means і ієрархічна кластеризація — це два поширених підходи до розв'язання задачі кластеризації у машинному навчанні і аналізі даних. Обидва методи мають свої переваги і недоліки.

У контексті вихідних даних, модель k-середніх призначає кожному об'єкту конкретний кластер, визначаючи його приналежність до певної групи, тоді як ієрархічна кластеризація дозволяє отримати інформацію про всю структуру кластерів на різних рівнях ієрархії, що дає змогу аналізувати дані на різних рівнях деталізації.

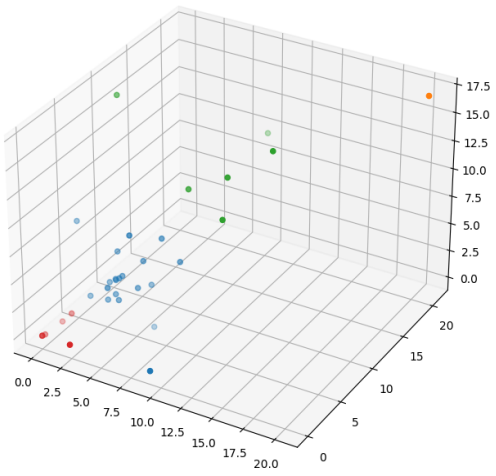


Рис. 3. Візуалізація кластерної структури.

Отримання чотирьох ідентичних кластерів як центроїдним методом, так і методом ієрархічної кластеризації вказує на те, що дані мають виражену структуру, яка однозначно інтерпретується як чотири різні категорії. Такий результат свідчить про консистентність та надійність знайдених кластерів, що додатково підтверджує їхню значущість у виявленні внутрішніх патернів та структур в досліджуваних даних.

Кластер 1 можна інтерпретувати як країни, які надають мінімальну фінансову допомогу студентам. Кластер 2 містить країни, які надають допомогу нижчу за середній рівень. Кластер 3 характеризує країни, які надають допомогу на рівні вище за середній. Кластер 4 відображає країну, яка забезпечує найвищий рівень фінансової допомоги студентам.

8. Висновки та перспективи подальших досліджень. У даній статті було проведено комплексне дослідження яке є розвитком напрямку прикладного аналізу даних [10–12] і спрямоване на аналіз фінансової допомоги студентам в залежності від країни Євросоюзу. На початковому етапі роботи було створено набір даних, що включав в себе відомості про фінансову підтримку студентів країн Євросоюзу з 2012 по 2020 роки відображену трьома усередненими періодами. Далі розроблено інформаційно-аналітичну систему, що реалізує класичні методи кластеризації, такі як k-Means та ієрархічну кластеризацію та ряд індексів визначення оптимальної кількості кластерів. Проведено структурування датасету та виділено чотири групи схожості.

Представлена 3D-модель структури даних, що сприяло кращому розумінню даних. Проведений порівняльний аналіз отриманих результатів показав їхній збіг, що свідчить про стабільність виявленої структури даних та підтверджує відповідність отриманих кластерів реальним зв'язкам у досліджуваному наборі даних. Наведена змістовна інтерпретація може бути корисна для грантових організацій та урядових установ у прийнятті рішень щодо розподілу фінансових ресурсів в освітній сфері.

Список використаної літератури

1. Gupta A., Sharma H., Akhtar A. A comparative analysis of k-means and hierarchical clustering. *EPRA International Journal of Multidisciplinary Research (IJMR).* 2021. Vol. 7, No. 8. P. 412–418. DOI: https://doi.org/10.36713/epra8308
2. Tokuda E. K., Comin C. H., Costa L. D. F. Revisiting agglomerative clustering. *Physica A: Statistical mechanics and its applications.* 2022. Vol. 585, No. 126433. P. 1–17. DOI: https://doi.org/10.1016/j.physa.2021.126433
3. Ahmed M., Seraj R., Islam S. M. S. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics.* 2020. Vol. 9, No. 8. P. 1–12. DOI: https://doi.org/10.3390/electronics9081295
4. Karthikeyan B., Dipu Jo George, G. Manikandan, Tony T. A comparative study on k-means clustering and agglomerative hierarchical clustering. *International Journal of Emerging Trends in Engineering Research.* 2020. Vol. 8, No 5. P. 1600–1604. DOI: https://doi.org/10.30534/ijeter/2020/20852020
5. Saputra D., Saputra M., Oswari L. Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method. *Sriwijaya international conference on information technology and its applications.* 2019. Vol. 172. P. 341–346. DOI: https://doi.org/10.2991/aisr.k.200424.051
6. Kondruk N. E. A comparative study of cluster validity indices. *Radio Electronics. Computer Science. Control.* 2019. No 4. P. 59–67. DOI: https://doi.org/10.15588/1607-3274-2019-4-6
7. Ashari I., Dwi Nugroho E., Baraku R., Novri Yanda I., and Liwardana R. Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta. *Journal of Applied Informatics and Computing.* 2023. Vol. 7, No. 1. P. 95–103. DOI: https://doi.org/10.30871/jaic.v7i1.4947
8. Hassan I., Abdullahi M., Ali Y. Analysis of Techniques for Selecting Appropriate Number of Clusters in K-means Clustering Algorithm. International Conference on Computing and Advances in Information Technology. 2021. P. 90–96.
9. Rachwał A. et al. Determining the quality of a dataset in clustering terms. Applied Sciences. 2023. Vol. 13, No. 5. P. 1–20. DOI: https://doi.org/10.3390/app13052942
10. Kondruk N. E. Use of length-based similarity measure in clustering problems. *Radio Electronics. Computer Science. Control.* 2018. Vol. 3, No. 46. P. 98–105. DOI: https://doi.org/10.15588/1607-3274-2018-3-11
11. Kondruk N. E., Malyar M. M. Analysis of Cluster Structures by Different Similarity Measures. *Cybernetics and Systems Analysis.* 2021. Vol. 57. P. 436–441. DOI: https://doi.org/10.1007/s10559-021-00368-4
12. Кондрук Н. Е. Моделі багатофакторного прогнозування. *Науковий вісник Ужгородського університету. Серія : Математика і інформатика.* 2022. Т. 40, № 1. С. 168–174. DOI: https://doi.org/10.24144/2616-7700.2022.40(1).168-174

Horvat I. V., Kondruk N. E., Kondruk Y. B., Neroda V. A. Segmentation of European Union countries by financial aid to students.

The use of clustering techniques and their comparative analysis is an integral part of modern scientific research due to their potential to reveal structures and patterns in complex data sets. These techniques allow classifying objects by similarity and grouping them into clusters, which helps to understand hidden relationships and discover new knowledge. The research is devoted to the study of practical aspects of using clustering techniques in the task of segmenting EU countries by student financial aid and includes a comparative analysis of clustering methods (k-Means, hierarchical clustering), thus ensuring the objectivity and accuracy of the results obtained. Various indices were used to determine the optimal number of clusters, such as the elbow method, the silhouette method, the Davis-Boldin method, and the Kalinski-Harabas index. Four identical clusters were obtained using both methods, so the data has a distinct structure that can be unambiguously interpreted as four different categories. This result indicates the consistency and reliability of the found clusters, which further confirms the significance of the meaningful interpretation.

Keywords: cluster analysis, k-Means, hierarchical clustering.

References

1. Gupta, A., Sharma, H., & Akhtar, A. (2021). A comparative analysis of k-means and hierarchical clustering. *EPRA International Journal of Multidisciplinary Research (IJMR)*, 7(8). https://doi.org/10.36713/epra8308
2. Tokuda, E. K., Comin, C. H., & Costa, L. D. F. (2022). Revisiting agglomerative clustering. *Physica A: Statistical mechanics and its applications,* 585(126433), 1–17. https://doi.org/10.1016/j.physa.2021.126433
3. Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics,* 9(8), 1–12. https://doi.org/10.3390/electronics9081295
4. Karthikeyan, B., George, D. J., Manikandan, G., & Thomas, T. (2020). A comparative study on k-means clustering and agglomerative hierarchical clustering. *International Journal of Emerging Trends in Engineering Research,* 8(5), 1600–1604. https://doi.org/10.30534/ijeter/2020/20852020
5. Saputra, D. M., Saputra, D., & Oswari, L. D. (2020, May). Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method. In *Sriwijaya international conference oninformation technology and its applications (SICONIAN 2019).* Atlantis Press. https://doi.org/10.2991/aisr.k.200424.051
6. Kondruk, N. E. (2019). A comparative study of cluster validity indices. *Radio Electronics. Computer Science. Control.* 4, 59–67. https://doi.org/10.15588/1607-3274-2019-4-6
7. Ashari, I. F., Nugroho, E. D., Baraku, R., Yanda, I. N., & Liwardana, R. (2023). Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta. *Journal of Applied Informatics and Computing,* 7(1), 95–103. https://doi.org/10.30871/jaic.v7i1.4947
8. Hassan, I. H., Abdullahi, M., & Ali, Y. S. (2021). Analysis of Techniques for Selecting Appropriate Number of Clusters in K-means Clustering Algorithm. *International Conference on Computing and Advances in Information Technology.* 90–96.
9. Rachwał, A. et al. (2023). Determining the quality of a dataset in clustering terms. *Applied Science,* 13(5), 1–20.
10. Kondruk, N. E. (2018). Use of length-based similarity measure in clustering problems. *Radio Electronics. Computer Science. Control,* 3(46), 98–105. https://doi.org/10.15588/1607-3274-2018-3-11
11. Kondruk, N. E., & Malyar, M. M. (2021). Analysis of Cluster Structures by Different Similarity Measures. *Cybern. Syst. Anal.,* 57, 436–441. https://doi.org/10.1007/s10559-021-00368-4
12. Kondruk, N. E. (2022). Models of multivariate forecasting. Scientific Bulletin of Uzhhorod University. Series of Mathematics and Informatics, 40(1), 168–174. https://doi.org/10.24144/2616-7700.2022.40(1).168-174

Одержано 29.04.2024