УДК 004.85:519.237.8
DOI https://doi.org/10.24144/2616-7700.2026.48(1).146-152

Н. Е. Кондрук

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

ВАЛІДАЦІЯ ЕФЕКТИВНОСТІ ЕКСПЕРТНО-ОРІЄНТОВАНОГО КОДУВАННЯ ДЛЯ АНАЛІЗУ СХОЖОСТІ ОРДИНАЛЬНИХ ДАНИХ

У роботі розв'язується задача підвищення ефективності кластерного аналізу об'єктів, що характеризуються категоріальними впорядкованими (ординальними) ознаками. Досліджено ефективність нової метрики відстані, яка, на відміну від традиційних підходів (SMC, коефіцієнт Жаккарда), враховує рангову природу атрибутів та величину інтервалів між ними. На основі експериментального дослідження з використанням набору даних UCI «Car Evaluation». Показано, що інтеграція експертних знань через механізм нерівномірного ранжування призводить до збільшення дисперсії попарних відстаней та суттєвого покращення сепарабельності кластерів. Ефективність запропонованого підходу підтверджено зростанням індексу Adjusted Rand Index та зниженням індексу Девіса-Болдіна порівняно з метрикою Говера та стандартними методами.

Ключові слова: категоріальні дані, ординальні ознаки, кластерний аналіз, ранжування, експертні оцінки.

1. Вступ. Розробка ефективних математичних інструментів для визначення подібності між об'єктами, описаними нечисловими значеннями, є актуальною задачею в аналізі даних, машинному навчанні та теорії прийняття рішень [1]. Значна частина властивостей реальних об'єктів характеризується саме категоріальними ознаками.

Ключова проблема полягає у тому, що традиційні метрики для категоріальних даних, такі як коефіцієнт простого збігу (Simple Matching Coefficient, SMC) або відстань Жаккарда (Jaccard Index) [2, 3], часто трактують ординальні ознаки як номінальні. Це призводить до ігнорування внутрішнього природного порядку значень (наприклад, {«низький», «середній», «високий»}), що, у свою чергу, генерує неточні або неінтуїтивні оцінки подібності [4].

Інша поширена помилка при роботі з ординальними даними полягає у їхньому наївному числовому кодуванні послідовними цілими числами. Таке кодування неявно припускає рівні інтервали (відстані) між суміжними категоріями, що є математично необґрунтованим, оскільки, за визначенням ординальних шкал, інтервали між категоріями невідомі або нерівні [4, 5, 6].
Запропонований підхід для вирішення описаних проблем [1] ґрунтується на двох взаємопов'язаних інструментах: метриці відстані ($d(\cdot)$), похідній від зваженої Манхеттенської відстані, та мірі подібності ($\mu_{\mathbb{R}^{\text{ORD}}}$).

На відміну від багатьох сучасних мір подібності, які є моделями типу "чорна скринька" і досягають високої точності за рахунок низької зрозумілості [1], запропонована методика $d(\cdot)$ є експертно-орієнтованою. Її архітектура свідомо

покладається на експертне визначення системи числових рангів для категорій. Це дозволяє фахівцям, які глибоко розуміють предметну область, кодувати свої знання та переваги безпосередньо у метрику. Така прозорість та обґрунтованість підвищують довіру користувача до аналітичної системи [1].

Метою даного дослідження є емпірична валідація та кількісна оцінка ефективності запропонованої в [1] метрики відстані для аналізу схожості об'єктів, описаних категоріальними впорядкованими ознаками.

Для досягнення поставленої мети необхідно розв'язати наступні завдання:

- підбір готового датасету, що містить ординальні дані із мітками класів для точнішої валідації;
- розробка системи ранжування даних на основі знань предметної області;
- обчислення матриці відстаней та формування порівняльного тестового простору на базі еталонних метрик (Simple Matching Coefficient та Gower's Distance);
- проведення агломеративної кластеризації із фіксованою кількістю кластерів відповідно до істинної розмітки даних;
- аналіз емпіричних результатів.

2. Методи. Запропонований підхід в [1] вимагає кодування вектора значень ознак $\overline{a_i}$ відповідним вектором числових показників рангів $\overline{r_i}\,(r_1^i, r_2^i, \ldots, r_n^i)$. Ранги повинні відображати строгий лінійний порядок за значимістю або перевагою, заданий для кожної ознаки A_k.

Відстань між двома об'єктами O_i та O_j визначається як різновид зваженої Манхеттенської метрики [1]:

$$d\,(O_i, O_j) = \sum_{k=1}^{n} \frac{\left|r_k^i - r_k^j\right|}{\Delta_k}, \tag{1}$$

де $\left|r_k^i - r_k^j\right|$ є абсолютною різницею рангів об'єктів O_i та O_j за k-ю ознакою. Ключовим елементом є знаменник $\Delta_k = |r_{k1} - r_{ks_k}|$, який характеризує розмах рангової шкали для k-ї ознаки [1].

Цей механізм Δ_k виконує функцію самоналагоджувального вагового коефіцієнта та нормалізації. Кожен доданок у сумі $\frac{\left|r_k^i - r_k^j\right|}{\Delta_k}$ є нормованою величиною, яка приймає значення з проміжку $[0; 1]$. Це гарантує, що "вклади" кожної ознаки у загальну відстань $d\,(\cdot)$ є співрозмірними. Таке нормування усуває ризик того, що ознака з більшим чисельним діапазоном рангів (наприклад, від 1 до 100) буде чисельно домінувати над ознакою з вужчим діапазоном (наприклад, від 1 до 5), навіть якщо обидві ознаки мають однакову важливість у предметній області. Кожна складова формули (1) фактично характеризує відсоток відмінності об'єктів за відповідною ознакою [1].

3. Вибір набору даних. Для експериментальної валідації обрано датасет «Car Evaluation» з репозиторію UCI [7]. Він містить 1728 екземплярів і шість вхідних ознак, які є виключно категоріальними і мають чіткий ординальний порядок [7, 8]. Дані походять від простої ієрархічної моделі прийняття рішень (DEX), яка оцінює прийнятність автомобіля (CAR) на основі складових, таких як PRICE та COMFORT [7, 9]. Така природа даних підтверджує, що ефективна міра подібності має інтегрувати знання про відносну важливість (ранг) цих

ознак. Наявність відомої цільової ознаки (class: unacc, acc, good, vgood — 4 класи) дозволяє використовувати зовнішні індекси валідації (ARI, NMI). Це критично важливо для об'єктивного кількісного порівняння якості кластеризації з істинною структурою даних [10].

4. Проектування системи рангів. Центральним елементом валідації є дослідження того, як експертно визначена система рангів впливає на кінцеву роздільність кластерів. Було розроблено дві системи кодування, що демонструють контрастні підходи.

Система A — послідовне (наївне) кодування. Ранги присвоюються як послідовні натуральні числа. Цей підхід імітує ситуацію, коли ординальність врахована, але специфічні знання про нерівномірну інтенсивність переходів між категоріями ігноруються.

Система B — експертне кодування. Ранги присвоюються з нерівномірними інтервалами ("скачками"). Як зазначалося в [1], якщо рівень ознаки має суттєво впливати на розв'язок, рекомендується підсилювати його системою рангів зі скачками. У цьому датасеті ознаки buying (ціна), maint (обслуговування) та safety (безпека) є найважливішими для визначення загальної прийнятності автомобіля. Тому нерівномірні інтервали застосовуються для збільшення ваги відмінностей у критичній зоні (між med, high та vhigh).

Таблиця 1.

Фрагмент експертного проектування рангових Систем A та B

Ознака	***Мітка значення***	***Послідовний ранг (A)***	***Ранг зі скачком (B)***
buying	vhigh	4	10
	high	3	7
	med	2	3
	low	1	1
safety	high	3	5
	med	2	2
	low	1	1

Система рангів для ознаки maint спроектована відповідно до системи рангів ознаки buying, а для всіх інших ознак відповідно до системи ознаки safety (табл. 1).

Аналіз впливу рангового кодування показує, що у Системі A будь-який перехід між суміжними категоріями дає однаковий внесок у відстань. Наприклад, для buying, перехід від 'low' до 'med' дає різницю 1, що після нормалізації становить $1/\Delta_k$. Натомість, у Системі B, перехід від 'med' до 'high' дає різницю 4, що, за нормалізації 1/9, становить 4/9. Це підсилення відмінностей у критичних зонах передбачає, що система B краще моделює когнітивну дистанцію та ієрархічні переваги, які лежать в основі прийняття рішень про прийнятність автомобіля.

5. Методологія експерименту. Для проведення експерименту та валідації запропонованих метрик було обрано метод ієрархічної агломеративної кластеризації (Agglomerative Hierarchical Clustering). Вибір саме цього алгоритму

зумовлений його детермінованою природою (результат не залежить від випадкової ініціалізації центрів, як у k-means або k-medoids) та гнучкістю в роботі з довільними мірами подібності.

Оскільки цільовий датасет має чотири відомих класи, кількість кластерів була фіксована на рівні $K = 4$ для безпосереднього порівняння отриманої структури з істинною розміткою [7].

Проведено порівняння чотирьох метрик для оцінки ефективності запропонованого підходу. Перша та друга базується на запропонованій метриці (1) для Систем A і B. Третьою метрикою є відстань Говера (Gower's Distance) [11] — універсальна метрика для змішаних типів даних, зокрема ординальних ознак, що слугує еталоном завдяки поєднанню ранжування й нормалізації. Четвертою виступає коефіцієнт простого співпадіння (SMC) або Hamming Distance — класична метрика для номінальних даних [2], яка в даному дослідженні використовується як контрольна, оскільки повністю ігнорує внутрішній порядок ординальних ознак.

Якість кластеризації оцінювалася за допомогою набору зовнішніх та внутрішніх індексів валідації [10]. До зовнішніх індексів належать ті, що використовуються за наявності істинних міток класів і дають змогу оцінити, наскільки отримана кластеризація відповідає реальній структурі даних. Серед них застосовували коригований індекс Ранда (ARI) [10], який вимірює схожість між прогнозованою та істинною кластеризацією з поправкою на випадковість і прямує до 1 для повної згоди, а також нормалізовану взаємну інформацію (NMI) [10], що відображає спільний обсяг інформації між двома розподілами міток і також наближається до 1 для якісного збігу. До внутрішніх індексів, які оцінюють геометричні властивості кластерної структури незалежно від істинних міток, належать Silhouette Score [12] — показник того, наскільки кожен об'єкт подібний до елементів власного кластера порівняно з елементами найближчого сусіднього кластера, де значення, близькі до 1, свідчать про чітко відокремлені та щільні кластери, — та індекс Девіса–Боулдена (DBI) [10], який оцінює співвідношення внутрішньокластерної дисперсії до міжкластерної відстані, причому нижчі значення DBI означають кращу якість кластеризації.

6. Результати. Обчислення відстаней $d(O_i, O_j)$ для всіх пар об'єктів у датасеті «Car Evaluation» показало суттєву різницю в розподілі між Системами A та B.

Таблиця 2.

Таблиця середніх значень та дисперсій розподілів відстані (1) Систем A та B

Показник	**Система A**	**Система B**
Середнє	2.58	2.61
Дисперсія	0.72	0.82

Агломеративна кластеризація ($K = 4$) була виконана для всіх чотирьох тестованих метрик. Результати, представлені в Таблиці 3, демонструють кількісну оцінку якості кластерних рішень.

7. Обговорення. Первинний статистичний аналіз отриманих матриць відстаней виявив помітну різницю в дисперсії розподілів попарних відстаней (табл. 2). Для Системи A, що використовує рівномірні інтервали між рангами,

Таблиця 3.

Порівняння ефективності кластеризації K-Medoids за різними метриками

Метрика відстані	**ARI↑**	**NMI↑**	**Silhouette Score↑**	**DBI↓**
Simple Matching Coeff.	0.013	0.060	0.220	4.162
Gower's Distance	0.055	0.129	0.099	2.464
$d(\cdot)$ System A	0.027	0.060	0.095	2.404
$d(\cdot)$ System B	0.119	0.195	0.279	1.797

дисперсія склала $\sigma^2 \approx 0.72$. Водночас застосування Системи B, яка базується на нелінійних "експертних" інтервалах для критичних ознак (безпека, ціна), призвело до зростання дисперсії до $\sigma^2 \approx 0.82$. Це збільшення на 14% свідчить про "розтягування" метричного простору, що дозволило підвищити контрастність між об'єктами різних класів та зменшити схожість між об'єктами, що мають критичні відмінності. Ефективність такого перетворення простору була підтверджена результатами кластеризації методом агломеративної кластеризації. Зведена таблиця метрик якості (табл. 3) демонструє явну перевагу запропонованого підходу з експертним налаштуванням над традиційними методами.

Натомість використання Системи B забезпечило найкращі показники за всіма критеріями. ARI зріс до 0.119, що, хоч і не є абсолютним ідеалом, суттєво перевищує результати інших методів і вказує на наявність кореляції зі справжніми мітками класів. Особливо показовим є індекс DBI, який для Системи B досяг мінімального значення (1.797). Оскільки менше значення DBI свідчить про кращу відокремленість кластерів, це підтверджує гіпотезу, що введення нелінійних ваг дозволило сформувати більш компактні та віддалені одна від одної групи об'єктів. Також метрика Silhouette (0.279) є найвищою для Системи B, що свідчить про вищу щільність об'єктів усередині сформованих кластерів.

Акцент на суттєвому зниженні індексу Девіса-Болдіна є критично важливим, оскільки він слугує об'єктивним критерієм якості внутрішньої геометричної структури сформованого простору ознак, незалежним від зовнішньої розмітки класів. Така динаміка свідчить про те, що інтеграція експертних знань через запропоновану систему нелінійного ранжування дозволила не лише мінімізувати внутрішньокласову дисперсію, забезпечивши високу компактність груп, але й суттєво збільшити відстань між центроїдами кластерів. Це підтверджує гіпотезу, що запропонована метрика $d(\cdot)$ ефективно трансформує простір ординальних даних, усуваючи проблему "розмитості"меж між класами, яка є характерною для рівномірних шкал та традиційних метрик, і формує чітко сепаровані кластери, що відповідають природній логіці предметної області.

8. Висновки та перспективи подальших досліджень. У дослідженні вирішується актуальна задача розвитку методів кластерного аналізу для об'єктів [1, 13], що описуються категоріальними впорядкованими (ординальними) ознаками. На основі проведеного теоретичного та експериментального дослідження отримано наступні результати: обґрунтовано ефективність метрики відстані (1), яка, на відміну від традиційних коефіцієнтів (SMC, Jaccard), враховує рангову природу даних; експериментально доведено перевагу запропонованого підходу на прикладі датасету UCI «Car Evaluation». Порівняльний

аналіз показав, що використання експертно-орієнтованої системи ваг (Система В) дозволяє суттєво покращити якість кластеризації порівняно з метрикою Говера та простим лінійним кодуванням, зокрема, досягнуто зростання індексу Adjusted Rand Index (ARI) до 0.119, що свідчить про кращу відповідність виявлених кластерів реальній структурі даних. Встановлено, що введення нелінійних інтервалів між рангами призводить до збільшення дисперсії розподілу попарних відстаней на 14%. Це «розтягування» метричного простору дозволило мінімізувати індекс Девіса-Болдіна, що підтверджує формування більш компактних та добре сепарованих кластерів.

Перспективи подальших досліджень вбачаються у інтеграції розробленої міри подібності в алгоритм, що заснований на нечітких бінарних ідношеннях, а також у розробці методів автоматичного визначення оптимальних рангових інтервалів без залучення експерта.

Конфлікт інтересів

Кондрук Наталія Емерихівна, членкиня редакційної колегії, є авторкою цієї статті та не брала участі в редакційному розгляді й ухваленні рішення щодо рукопису. Опрацювання рукопису здійснювалося незалежним редактором.

Фінансування

Дослідження здійснено в рамках кафедральної науково-дослідної роботи «Методи обчислювального інтелекту для обробки і аналізу даних» (державний реєстраційний номер 0121U109279).

Доступність даних

Усі дані доступні в цифровій або графічній формі в основному тексті рукопису.

Використання штучного інтелекту

Авторка підтверджує, що при створенні даної роботи вона не використовувала технології штучного інтелекту.

Список використаної літератури

1. Kondruk, N. E. (2023). Methods for determining similarity of categorical ordered data. *Radio Electronics, Computer Science, Control,* (2), 31. https://doi.org/10.15588/1607-3274-2023-2-4 [in Ukrainian].
2. Suárez, J., García, S., & Herrera, F. (2021). A tutorial on distance metric learning: Mathemati-

cal foundations, algorithms, experimental analysis, prospects and challenges. *Neurocomputing,* 425, 300–322. https://doi.org/10.1016/j.neucom.2020.08.017
3. Mathisen, B., Aamodt, A., Bach, K., & Langseth, H. (2019). Learning similarity measures from data. *Progress in Artificial Intelligence,* 9, 129–143. https://doi.org/10.1007/s13748-019-00201-2
4. Desai, A., Singh, H., Pudi, V., & Gopalan, S. (2011). DISC: Data-Intensive similarity measure for categorical data. *Advances in Knowledge Discovery and Data Mining,* 6635, 469–481. https://doi.org/10.1007/978-3-642-20847-8_39
5. Cunningham, P. (2009). A taxonomy of similarity mechanisms for case-based reasoning. *IEEE Transactions on Knowledge and Data Engineering,* 21, 1532–1543. https://doi.org/10.1109/TKDE.2008.227
6. Nikpour, N., Aamodt, A., & Bach, K. (2018). Bayesian-supported retrieval in BNCreek: A knowledge-intensive case-based reasoning system. *Case-Based Reasoning Research and Development, 11156,* 323–338. https://doi.org/10.1007/978-3-030-01081-2_22
7. Bohanec, M. (1988). Car Evaluation [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5JP48.
Retrieved from: https://archive.ics.uci.edu/dataset/19/car+evaluation
8. Dyussenbayev, A. (2017). Age periods of human life. *Advances in Social Sciences Research Journal,* 4(6), 258–263. https://doi.org/10.14738/assrj.46.2924
9. Kondruk, N. (2017). Clustering method based on fuzzy binary relation. *Eastern-European Journal of Enterprise Technologies,* 2(4), 10–16. https://doi.org/10.15587/1729-4061.2017.94961
10. Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis.* John Wiley & Sons, 355 pp. https://doi.org/10.1002/9780470316801
11. Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics,* 27(4), 857–871. https://doi.org/10.2307/2528823
12. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics,* 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7
13. Kondruk, N. E., & Malyar, M. M. (2021). Analysis of cluster structures by different similarity measures. *Cybernetics and Systems Analysis,* 57, 436–441. https://doi.org/10.1007/s10559-021-00368-4

Kondruk N. E. Validation of the effectiveness of expert-oriented encoding for similarity analysis of ordinal data.

The paper addresses the problem of improving the effectiveness of cluster analysis for objects characterized by categorical ordered (ordinal) features. The effectiveness of a new distance metric is investigated, which, unlike traditional approaches (SMC, Jaccard coefficient), takes into account the rank-based nature of attributes and the magnitude of intervals between them. An experimental study was conducted using the UCI Car Evaluation dataset. It is shown that the integration of expert knowledge through a non-uniform ranking mechanism leads to an increase in the variance of pairwise distances and a significant improvement in cluster separability. The effectiveness of the proposed approach is confirmed by an increase in the Adjusted Rand Index and a decrease in the Davies–Bouldin index compared to the Gower metric and standard methods.

Keywords: categorical data, ordinal features, cluster analysis, ranking, expert assessments.

Отримано: 01.12.2025 Прийнято: 19.12.2025 Опубліковано: 29.01.2026