критеріїв одночасно, а за високої розмірності – поєднувати їх зі зниженням розмірності або використовувати спеціалізовані підходи [1, 5]; остаточний вибір k варто узгоджувати з предметною інтерпретацією результатів.

Список використаних джерел

1. Zhu X., Li L. Estimating the Number of Clusters in High-Dimensional Large Datasets. International Journal of Data Warehousing and Mining. 2023. Vol. 19. DOI: https://doi.org/10.4018/IJDWM.316142.

2. Rykov A., De Amorim R. C., Makarenkov V., Mirkin B. Inertia-Based Indices to Determine the Number of Clusters in K-Means: An Experimental Evaluation. IEEE Access. 2024. Vol. 12. P. 11761–11773. DOAJ: https://doaj.org/article/ffed5e38461848999f79c3575b6b85e9.

3. Abushilah S. F., Abbas R. H. Performance Evaluation of Some Clustering Algorithms under Different Validity Indices. Mathematical Modelling of Engineering Problems. 2023. Vol. 10, No. 4. P. 1271–1280. DOI: https://doi.org/10.18280/mmep.100420.

4. Chicco D., Campagner A., Spagnolo A., Ciucci D., Jurman G. The Silhouette coefficient and the Davies-Bouldin index are more informative than Dunn index, Calinski-Harabasz index, Shannon entropy, and Gap statistic for unsupervised clustering internal evaluation of two convex clusters. PeerJ Computer Science. 2025. Vol. 11. Article e3309. DOI: https://doi.org/10.7717/peerj-cs.3309.

5. Bai Y., Chu L. A Graph-based Approach to Estimating the Number of Clusters. arXiv preprint. 2024. arXiv:2402.15600. URL: https://arxiv.org/abs/2402.15600.

УДК 004.85

ВПЛИВ МЕТРИКИ ВІДСТАНІ НА ЕФЕКТИВНІСТЬ МЕТОДУ K-НАЙБЛИЖЧИХ СУСІДІВ: ЕКСПЕРИМЕНТ НА НАБОРІ ДАНИХ WINE

Артем Курилов

Миколаївський національний аграрний університет, м. Миколаїв, Україна.

Науковий керівник: Т. С. Кучмійова,
кандидат економічних наук, доцент,
Миколаївський національний аграрний університет

Метод k найближчих сусідів (k-NN) – один із найпростіших алгоритмів класифікації, що не потребує припущень про розподіл даних: Т. Кавер і П. Харт довели, що похибка правила найближчого сусіда не перевищує подвоєної байєсівської похибки [1]. Метод застосовують у різних прикладних галузях, зокрема в прогнозуванні захворювань [2]. Як метод «лінивого» навчання k-NN виконує всі обчислення на етапі прогнозування, є чутливим до нерелевантних ознак, а числові ознаки перед обчисленням відстаней, як правило, нормалізують [3]. Якість класифікації визначається двома гіперпараметрами – метрикою відстані та кількістю сусідів k [4]. Єдиної найкращої метрики не існує: в експерименті на 33 наборах даних UCI навіть запропонована автором відстань виявилась найточнішою лише у 9 випадках [4], а огляд [5] фіксує значні розбіжності в точності між різними відстанями. Тому метрику необхідно обирати емпірично для кожної задачі.

Мета – кількісно оцінити вплив п’яти метрик відстані (евклідової, манхеттенської, Чебишева, Мінковського при p = 3 та косинусної) на точність і швидкодію класифікатора k-NN та з’ясувати роль стандартизації ознак.

Використано набір даних Wine з репозиторію UCI – 178 зразків вина трьох сортів, описаних 13 хімічними ознаками [6]; пропущених значень і дублікатів немає. Діапазони ознак відрізняються майже у 1000 разів (наприклад, Hue – 0,48–1,71, Proline – 278–1680), тому застосовано z-стандартизацію, параметри якої обчислювали лише за навчальною вибіркою, щоб уникнути витоку даних. Вибірку стратифіковано розділено на навчальну (70 %, 124 зразки) і тестову (30 %, 54 зразки) частини, random_state = 42. Для кожної метрики оптимальне k у діапазоні від 1 до 30 підбирали за допомогою GridSearchCV із 5-кратною стратифікованою крос-валідацією; класифікатор реалізовано

засобами бібліотеки scikit-learn (клас K Neighbors Classifier) [7]. Час навчання та прогнозування усереднено за 50 запусків.

Для усунення впливу різного масштабу ознак застосовано z-стандартизацію:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (1)$$

де x_{ij} – значення j-ї ознаки i-го зразка; μ_j та σ_j – середнє значення й стандартне відхилення j-ї ознаки, обчислені за навчальною вибіркою. Відстань між стандартизованими зразками x і z у m-вимірному просторі ознак (m = 13) для метрик Мінковського задається формулою

$$d_p(x,z) = \left(\sum_{j=1}^{m} |x_j - z_j|^p\right)^{\frac{1}{p}} \quad (2)$$

де p = 1 відповідає манхеттенській, p = 2 – евклідовій метриці, а p = 3 – метриці Мінковського, дослідженій у роботі. Відстань Чебишева та косинусна відстань визначаються як

$$d_\infty(x,z) = \max_{1 \le j \le m} |x_j - z_j| \quad (3)$$

$$d_{\cos}(x,z) = 1 - \frac{\sum_{j=1}^{m} x_j z_j}{\sqrt{\sum_{j=1}^{m} {x_j}^2}\sqrt{\sum_{j=1}^{m} {z_j}^2}} \quad (4)$$

Зведені показники на тестовій вибірці наведено в табл. 1. Найкращий результат показала манхеттенська метрика при k = 5 (точність 98,15 %, F1-міра 0,9811, одна помилка на 54 зразки), найгірший – метрика Чебишева при k = 1 (87,04 %). Різниця становить 11,11 відсоткового пункту, що підтверджує істотний вплив вибору метрики. Порядок «манхеттенська – евклідова – Чебишева» узгоджується з даними [4], де середня точність за 33 наборами UCI становила 0,834, 0,823 та 0,786 відповідно. Імовірне пояснення переваги манхеттенської метрики – її менша чутливість до викидів, оскільки відхилення не підносяться до квадрата; у роботі це припущення окремо не перевірялося.

Таблиця 1 - Порівняння метрик відстані (тестова вибірка, стандартизовані дані)

Метрика	**Опт. k**	**Точність, %**	**F1 (macro)**	**Час прогнозування, мс**
Манхеттенська	5	98,15	0,9811	1,15
Евклідова	14	96,30	0,9636	1,26
Мінковського (p = 3)	3	94,44	0,9445	2,34
Косинусна	6	94,44	0,9445	1,42
Чебишева	1	87,04	0,8653	1,09

Стандартизація ознак підвищила точність усіх п’яти метрик на 14,82–29,63 відсоткового пункту (найбільше – для метрики Мінковського при p = 3). Час навчання практично однаковий для всіх метрик (0,26–0,45 мс), тоді як час прогнозування метрики Мінковського при p = 3 (2,34 мс) приблизно в 1,6–2 рази більший, ніж у решти метрик (1,09–1,42 мс). Обмеженням є невеликий обсяг тестової вибірки: одна помилка змінює точність на 1,85 відсоткового пункту, тому різницю між сусідніми метриками слід тлумачити обережно. Крім того, в огляді [5] для набору Wine найвищу точність (98,5 %) за іншої методики та k = 1 показала відстань Канберра, яку в цій роботі не досліджували.

Вплив стандартизації наочно подано на рис. 1: без неї точність класифікації становила лише від 64,81 % (метрика Мінковського при p = 3) до 79,63 % (манхеттенська метрика), а після стандартизації – від 87,04 % до 98,15 %. Отже, нестандартизовані ознаки з великим діапазоном (насамперед Proline) домінують у відстані й фактично «приглушують» внесок решти ознак, а різниця між метриками на сирих даних має інший характер, ніж на стандартизованих.

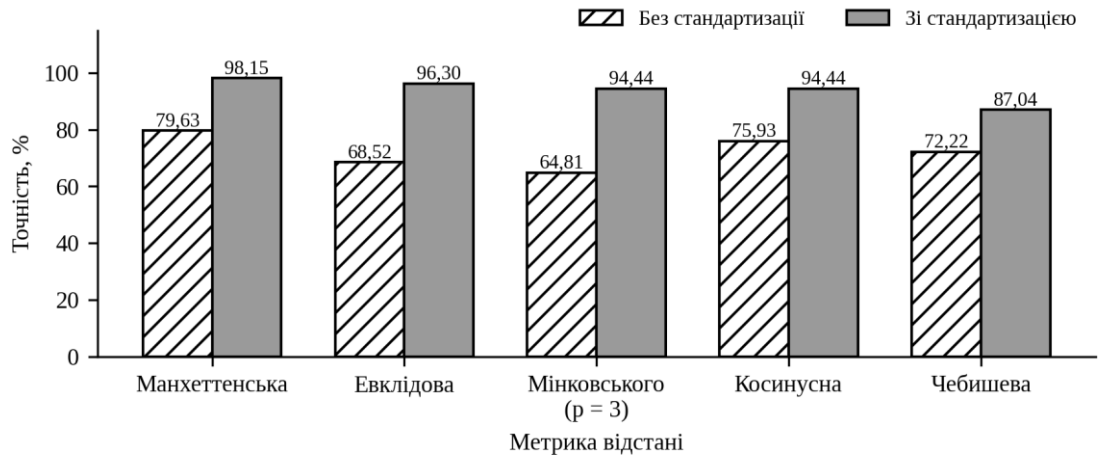


Рис. 1. Точність класифікації k-NN на тестовій вибірці до та після z-стандартизації ознак

Практичне значення результатів полягає в обґрунтуванні простої процедури налаштування k-NN для табличних числових даних: стандартизація ознак за параметрами навчальної вибірки, порівняння кількох метрик і підбір k за стратифікованою крос-валідацією. Така процедура не потребує складних обчислень і може бути перенесена на інші набори даних, зокрема на агрономічні вибірки з ознаками різного масштабу. Водночас результати отримано на одному невеликому наборі даних, тому їх не слід вважати універсальними.

Висновки. 1) Вибір метрики відстані істотно впливає на ефективність k-NN: на наборі Wine точність коливається від 87,04 % до 98,15 %. 2) Стандартизація ознак є обов'язковим кроком перед застосуванням метричних методів – вона підвищила точність усіх досліджених метрик. 3) Для табличних числових даних із ознаками різного масштабу манхеттенську метрику доцільно розглядати як пріоритетного кандидата поряд з евклідовою, а остаточний вибір метрики та k – підтверджувати крос-валідацією. Перспективи подальших досліджень – розширення набору метрик (Канберра, Махаланобіса) і наборів даних та перенесення методики на агрономічні задачі класифікації за багатовимірними сенсорними даними.

Список використаних джерел

1. Cover T. M., Hart P. E. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967. Vol. 13, No. 1. P. 21–27. DOI: 10.1109/TIT.1967.1053964. URL: https://doi.org/10.1109/TIT.1967.1053964 (дата звернення: 15.10.2026).

2. Uddin S., Haque I., Lu H., Moni M. A., Gide E. Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Scientific Reports. 2022. Vol. 12, Art. 6256. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC9012855 (дата звернення: 15.10.2026).

3. Cunningham P., Delany S. J. k-Nearest neighbour classifiers: 2nd edition (with Python examples). arXiv:2004.04523. 2020. URL: https://arxiv.org/abs/2004.04523 (дата звернення: 15.10.2026).

4. Rodrigues É. O. Combining Minkowski and Chebyshev: New distance proposal and survey of distance metrics using k-nearest neighbours classifier. Pattern Recognition Letters. 2018. Vol. 110. P. 66–71. URL: https://arxiv.org/abs/2112.12549 (дата звернення: 15.10.2026).

5. Prasath V. B. S., Abu Alfeilat H. A., Lasassmeh O. et al. Distance and similarity measures effect on the performance of K-nearest neighbor classifier – a review. arXiv:1708.04321. 2017. URL: https://arxiv.org/abs/1708.04321 (дата звернення: 15.10.2026).

6. Aeberhard S., Forina M. Wine [Dataset]. UCI Machine Learning Repository. 1991. DOI: 10.24432/C5PC7J. URL: https://archive.ics.uci.edu/dataset/109/wine (дата звернення: 15.10.2026).

7. KNeighborsClassifier : документація scikit-learn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html (дата звернення: 15.10.2026).