УДК 004.85:519.237.8

ПОРІВНЯННЯ МЕТОДІВ ВИЗНАЧЕННЯ ОПТИМАЛЬНОЇ КІЛЬКОСТІ КЛАСТЕРІВ

Іван Гребенюк

Миколаївський національний аграрний університет. м. Миколаїв, Україна

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

Кластерний аналіз є одним із базових інструментів навчання без учителя та широко застосовується в інтелектуальному аналізі даних, біоінформатиці, обробці зображень, сегментації клієнтів і моніторингу агропромислових систем. Більшість поширених алгоритмів, насамперед k-means, потребують попереднього задання кількості кластерів k, яка в реальних задачах, як правило, невідома [1]. Надто мале значення k призводить до штучного злиття природних груп, надто велике – до дроблення однорідних об’єктів, і в обох випадках подальші висновки спотворюються. Для оцінювання k запропоновано десятки критеріїв, проте, як свідчать порівняльні дослідження, жоден із них не є універсальним: вони по-різному реагують на форму кластерів, їх перекриття, розмір вибірки та розмірність простору ознак [2, 3]. Тому актуальним завданням є зіставлення найпоширеніших методів на даних із відомою структурою. Мета роботи – порівняти шість підходів до визначення оптимальної кількості кластерів і сформулювати практичні рекомендації щодо їх вибору.

У роботі [2] виконано експериментальне порівняння індексів, побудованих на інерції k-means, з використанням індексу силуету як еталона; автори запропонували варіант індексу «ліктя», що використовує значення k із кроком у два–три, і дійшли висновку, що загального переможця немає, хоча окремі результати достатньо переконливі, щоб слугувати орієнтиром. У дослідженні [3] ієрархічні алгоритми з різними типами зв’язку оцінювалися за індексами KL, Calinski–Harabasz (CH), силуету та gap на симульованих даних: найкращу відповідність істинним кластерам (понад 90 %) забезпечило поєднання методу середнього зв’язку з індексом CH, тоді як gap-статистика у багатьох сценаріях повертала мінімальне значення k = 2. Автори [4] порівняли силует, індекс Дейвіса–Болдіна, індекс Данна, CH, ентропію Шеннона та gap-статистику для k-means при k = 2 на опуклих штучних даних і п’яти медичних наборах та встановили, що силует і індекс Дейвіса–Болдіна є інформативнішими й надійнішими для опуклих неперекривних кластерів в евклідовому просторі.

Окремий напрям досліджень пов’язаний із даними великої розмірності. Для них метод ліктя стає обчислювально дорогим і не завжди дає чітке «коліно» [1], а відстані між об’єктами всередині кластерів і між кластерами при зростанні розмірності майже вирівнюються, через що глобальні дистанційні критерії втрачають розрізнювальну здатність [5]. У роботі [1] запропоновано спосіб оцінювання k для великих багатовимірних наборів, а в [5] – непараметричну статистику на основі графів подібності з доведеною консистентністю, яка за результатами симуляцій перевершує класичні критерії саме у високорозмірному випадку.

Для кожного k = 2…10 виконувалася кластеризація методом k-means, після чого оцінювалися шість критеріїв. Метод ліктя: будується залежність сумарного внутрішньокластерного розкиду

$$W_k = \sum_{j=1}^{k} \sum_{x \in C_j} \|x - \mu_j\|^2 \qquad (1)$$

від k, а оптимальним вважається значення, після якого зменшення W_k різко сповільнюється. Для автоматизації використано правило найбільшої відстані точки кривої до хорди, що з’єднує її крайні точки. Коефіцієнт силуету: для об’єкта i

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \qquad (2)$$

де $a(i)$ – середня відстань до об’єктів власного кластера, $b(i)$ – середня відстань до найближчого іншого кластера; обирається k з найбільшим середнім значенням s. Індекс Calinski–Harabasz (відношення міжкластерної й внутрішньокластерної дисперсій, максимум):

$$CH(k) = \frac{\frac{tr(B_k)}{k-1}}{\frac{tr(W_k)}{n-k}} \quad (3)$$

Індекс Davies–Bouldin (середня подібність кластерів, мінімум), де σ_i – середній розкид i-го кластера, d – відстань між центроїдами:

$$DB(k) = \frac{1}{k}\sum_{i=1}^{k} max_{j \neq i} \frac{\sigma_i + \sigma_j}{d(\mu_i, \mu_j)} \quad (4)$$

Gap-статистика порівнює логарифм розкиду з його очікуваним значенням для еталонного розподілу без кластерної структури; обирається найменше k, для якого $Gap(k) \geq Gap(k+1) - s_{k+1}$:

$$Gap(k) = E^*[lnW_k] - lnW_k \quad (5)$$

Байєсівський інформаційний критерій (BIC) для гауссівських сумішей (GMM), мінімум, де p – кількість параметрів моделі, n – обсяг вибірки, $\hat{L}$ – максимум правдоподібності:

$$BIC = p \cdot ln\, n - 2 \cdot ln\, \hat{L} \quad (6)$$

Обчислення виконано мовою Python 3 із бібліотекою scikit-learn. Використано п’ять наборів даних (рис. 1): три синтетичні – A (чотири добре розділені гауссівські групи, 600 точок), B (шість груп із помітним перекриттям, 900 точок) і C (два «півмісяці», 600 точок, неопукла структура), а також два реальні навчальні набори – Iris (150 об’єктів, 4 ознаки, 3 види) і Wine (178 об’єктів, 13 ознак, 3 сорти), попередньо стандартизовані. Для k-means застосовано 10 випадкових ініціалізацій, для GMM – повні матриці коваріації та 3 ініціалізації; для gap-статистики – 10 еталонних вибірок із рівномірного розподілу та правило однієї стандартної похибки. Якість методу оцінювалася за збігом оцінки $\hat{k}$ з відомим числом груп k* та за середньою абсолютною похибкою MAE = середнє $|\hat{k} - k^*|$ по всіх наборах.

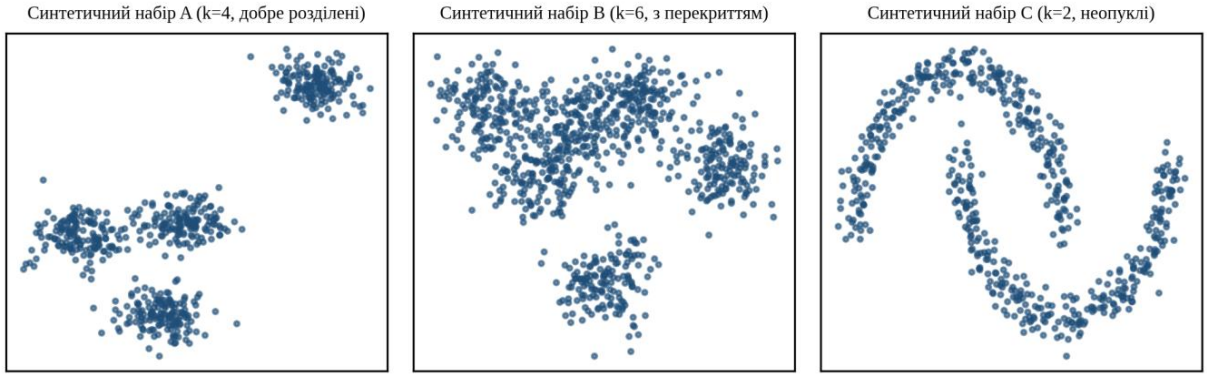


Рис. 1. Синтетичні набори даних A, B та C

Оцінки кількості кластерів, отримані кожним методом, наведено в таблиці 1. На наборі A, де групи чітко відокремлені, правильне значення k = 4 визначили метод ліктя, CH, gap-статистика та BIC. Силует і індекс Дейвіса–Болдіна обрали k = 2, але значення силуету для k = 2 (0,722) лише на 0,017 перевищує значення для k = 4 (0,705): практично нівельована різниця свідчить про те, що ці критерії схильні надавати перевагу грубшій структурі даних.

На наборі B із перекриттям груп (рис. 2) лише індекс CH точно визначив k = 6. Силует, індекс Дейвіса–Болдіна та BIC дали k = 5, тобто недооцінили кількість груп на одиницю, а метод ліктя (k = 4) через відсутність виразного перегину кривої виявився менш надійним. Показово, що gap-статистика має максимум саме при k = 6, проте правило однієї стандартної похибки, орієнтоване на мінімальну складність моделі, обрало k = 3. Отже, результат залежить не лише від самого критерію, а й від правила вибору k на його основі.

Таблиця 1. Оцінки кількості кластерів різними методами (k - відоме число груп; виділено збіги)*

Набір даних	k*	Лікоть	Силует	CH	DB	Gap	BIC (GMM)
A (blobs, чіткі групи)	4	4	2	4	2	4	4
B (blobs, перекриття)	6	4	5	6	5	3	5
C (півмісяці)	2	4	8	10	8	4	8
D (Iris)	3	5	2	2	2	3	2
E (Wine)	3	3	3	3	3	3	2
Кількість збігів з k* (із 5)		2	1	3	1	3	1
MAE = середнє $\lvert\hat{k} - k^*\rvert$		1,2	2,0	1,8	2,0	1,0	1,8

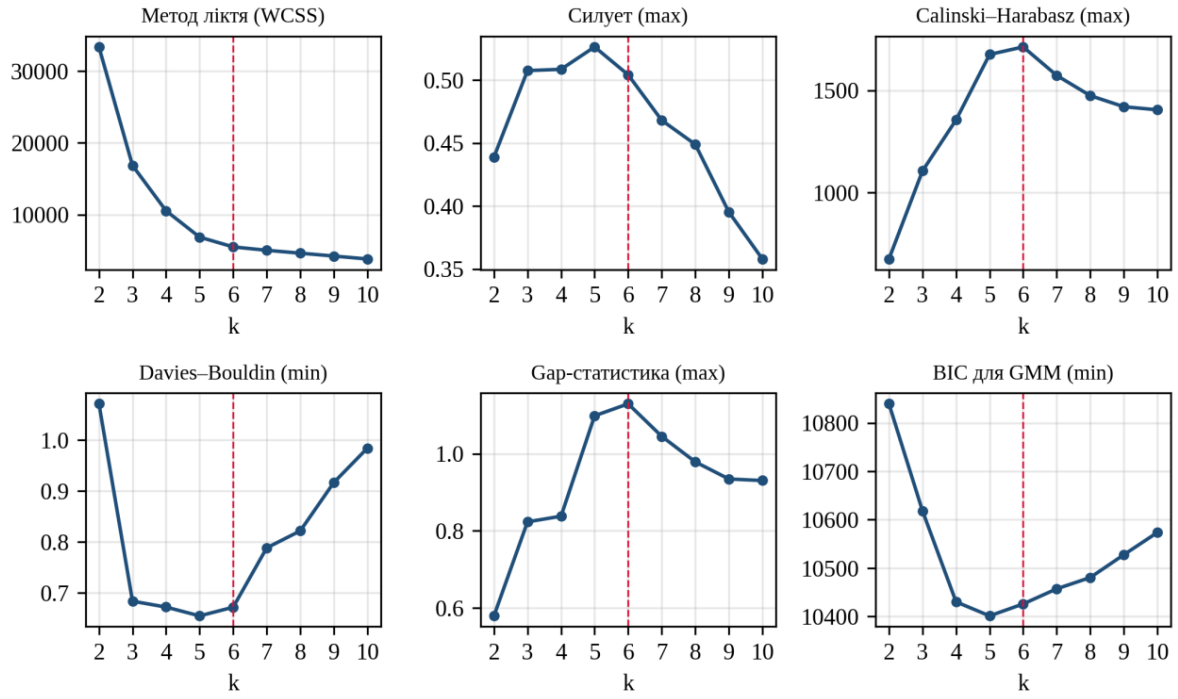


Рис. 2. Значення критеріїв залежно від k для набору B (істинне k = 6 позначено пунктиром)*

Набір C (півмісяці) продемонстрував принципове обмеження всіх розглянутих методів: усі вони спираються на опуклість кластерів, тому дали завищені оцінки (4 – для методу ліктя та gap-статистики, 8 – для силуету, індексу Дейвіса–Болдіна та BIC, 10 – для CH, тобто на межі діапазону пошуку). Це узгоджується з висновками [4] про застосовність силуету та індексу Дейвіса–Болдіна для опуклих кластерів в евклідовому просторі. Для таких даних доцільно використовувати алгоритми, що не припускають опуклості (спектральні, щільнісні), та відповідні їм критерії.

На реальних наборах картина неоднозначна (рис. 3). Для Iris силует, CH, індекс Дейвіса–Болдіна та BIC вказали k = 2: два види з трьох частково перекриваються, тому «відомі» три класи не збігаються з двома природно відокремленими групами, і це не є помилкою критеріїв у строгому сенсі. Gap-статистика дала k = 3, метод ліктя – k = 5. Для Wine усі критерії на основі k-means узгоджено вказали k = 3, тоді як BIC обрав k = 2. Водночас максимальний силует для Wine становить лише близько 0,29, що свідчить про слабко виражену кластерну структуру та підтверджує необхідність оцінювати не лише оптимальне k, а й якість самого розбиття.

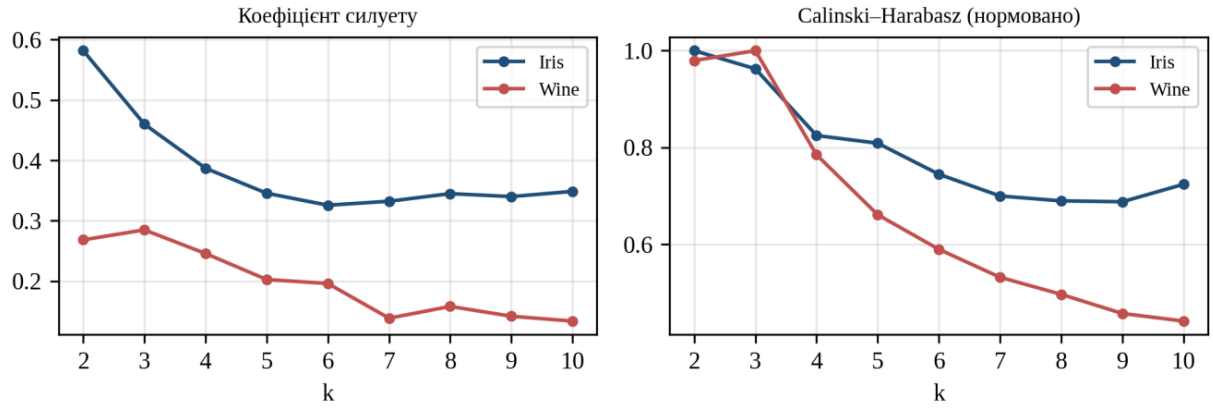


Рис. 3. Коефіцієнт силуету та нормований індекс CH для наборів Iris і Wine

Узагальнені підсумки (таблиця 1) такі: найбільшу кількість збігів (по 3 з 5) показали індекс CH та gap-статистика, причому gap-статистика мала й найменшу середню похибку (MAE = 1,0); силует, індекс Дейвіса–Болдіна та BIC дали по одному збігу. Порівняльну характеристику методів наведено в таблиці 2. Слід підкреслити, що експеримент є ілюстративним: використано п’ять наборів і одну реалізацію випадкових даних, тому не можна стверджувати про переваги методів у загальному випадку, що узгоджується з висновком [2] про відсутність загального переможця.

Таблиця 2. Порівняльна характеристика методів визначення кількості кластерів

Метод	Що оцінює	Переваги	Обмеження
Лікоть (WCSS)	Компактність кластерів	Простота, наочність; результат безпосередньо випливає з k-means	Суб’єктивність; «коліно» відсутнє або не єдине; втрата чіткості при зростанні розмірності [1, 5]
Силует	Компактність і відокремленість	Обмежений діапазон [−1; 1], індивідуальна оцінка об’єктів; висока надійність для опуклих кластерів [4]	Складність $O(n^2)$; схильність до грубшої структури; погано працює для неопуклих форм
Calinski–Harabasz	Відношення міжкластерної до внутрішньокластерної дисперсії	Швидко обчислюється; у [3] найкраще узгоджується з істинними кластерами	Орієнтований на опуклі кластери та евклідову метрику; при великому k часто зростає
Davies–Bouldin	Подібність кластерів	Проста інтерпретація (менше — краще) [4]	Орієнтований на опуклі кластери; чутливий до викидів
Gap-статистика	Відхилення від нульового розподілу	Формальна статистична основа; може вказати відсутність структури	Ресурсомісткий (моделювання еталонів); нестійкий для нерівномірних даних та великої розмірності [5]
BIC для GMM	Правдоподібність зі штрафом за складність	Імовірнісна інтерпретація; враховує форму кластерів	Припущення про вид розподілу; перепараметризація при великій розмірності [5]

Висновки. 1) Жоден із розглянутих методів не є універсальним: результат залежить від форми кластерів, їх перекриття, розмірності даних і правила вибору k. 2) У проведеному експерименті найстабільніші результати показали індекс Calinski–Harabasz та gap-статистика; силует, індекс Дейвіса–Болдіна та BIC частіше обирали грубшу структуру (меншу кількість кластерів). 3) Для неопуклих кластерів усі методи, орієнтовані на k-means, дають суттєво завищені оцінки, тому слід спершу перевіряти припущення про форму кластерів. 4) На практиці доцільно застосовувати кілька

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