УДК 004.048 + 004.94

О.Р. ЯРЕМА
Львівський національний університет імені Івана Франка
Д.О. СЕНЧИШЕН
Херсонський державний університет
С.А. БАБІЧЕВ
Університет Яна Євангеліста Пуркінє, Усті-над-Лабем, Чехія;
Херсонський державний університет

ГІБРИДНА МОДЕЛЬ ОЦІНЮВАННЯ ЕФЕКТИВНОСТІ МЕТРИК БЛИЗЬКОСТІ ДЛЯ ДАНИХ ЕКСПРЕСІЇ ГЕНІВ

У статті представлено розробку та застосування гібридної моделі для оцінювання ефективності метрик близькості у високорозмірних даних експресії генів, яка інтегрує методи інтелектуального аналізу даних та машинного навчання в єдиній комплексній структурі. Основна увага приділяється порівняльному аналізу кореляційної відстані, метрик, заснованих на взаємній інформації, та метрики Васерштейна, з метою визначення їхньої ефективності для завдань кластеризації та класифікації. Отримані результати демонструють, що кореляційна відстань і метрика Васерштейна забезпечують високу точність та стабільність, що робить їх придатними для інтеграції з діагностичними системами. Для підвищення надійності класифікації реалізовано стекінг-модель, яка компенсує можливі помилки кластеризації та забезпечує стабільну продуктивність незалежно від використаної метрики та структури кластерів. Запропонований конвеєр обробки даних забезпечує автоматизований, стандартизований і масштабований аналіз великих масивів даних експресії генів, що узгоджується із принципами персоналізованої медицини. Завдяки можливості ранньої діагностики захворювань та підтримці розроблення індивідуальних стратегій лікування отримані результати мають велике значення для вдосконалення сучасних діагностичних систем у межах концепції прецизійної медицини. Застосування гібридних підходів дозволяє ефективно поєднувати переваги різних методів оцінювання подібності та класифікації, що сприяє підвищенню точності прогнозування на основі даних експресії генів. Запропонований методологічний підхід може бути використаний у біоінформатиці для аналізу складних біологічних систем, оптимізації процесів обробки генетичних даних і розроблення нових алгоритмічних рішень у сфері медичної діагностики. Такий підхід сприяє адаптації сучасних інформаційних технологій для виявлення біомаркерів захворювань, що забезпечує інтеграцію отриманих результатів у клінічну практику та підтримує розвиток персоналізованих методів лікування.

Ключові слова: *дані експресії генів, метрики близькості, гібридна модель, кластеризація, класифікація, персоналізована медицина.*

O.R. YAREMA
Ivan Franko National University of Lviv
D.O. SENCHYSHEN
Kherson State University
S.A. BABICHEV
Jan Evangelista Purkyně University, Ústi nad Labem, Czech Republic;
Kherson State University

HYBRID MODEL FOR EVALUATING THE EFFECTIVENESS OF PROXIMITY METRICS FOR GENE EXPRESSION DATA

This paper presents the development and application of a hybrid model for evaluating the effectiveness of proximity metrics in high-dimensional gene expression data, integrating data mining and machine learning methods within a unified comprehensive framework. The study focuses on a comparative analysis of correlation distance, mutual information-based metrics, and the Wasserstein metric to assess their efficiency for clustering and classification tasks. The obtained results demonstrate that correlation distance and the Wasserstein metric provide high accuracy and stability, making them suitable for integration into diagnostic systems. To enhance classification reliability, a stacking model has been implemented to compensate for potential clustering errors and ensure consistent performance regardless of the applied metric and cluster structure. The proposed data processing pipeline enables automated, standardized, and scalable analysis of large-scale gene expression datasets, aligning with the principles of personalized medicine. By facilitating early disease diagnosis and supporting the development of individualized treatment strategies, the findings contribute significantly to improving modern diagnostic systems within the framework of precision medicine. The application of hybrid

https://doi.org/10.32782/mathematical-modelling/2025-8-1-2**8**

approaches allows for an effective combination of the advantages of different similarity evaluation and classification methods, enhancing prediction accuracy based on gene expression data. The proposed methodological approach can be utilized in bioinformatics for analyzing complex biological systems, optimizing genetic data processing workflows, and developing new algorithmic solutions in medical diagnostics. This approach promotes the adaptation of modern information technologies for identifying disease biomarkers, ensuring the integration of obtained results into clinical practice, and supporting the advancement of personalized treatment strategies.

Key words: *gene expression data, proximity metrics, hybrid model, clustering, classification, personalized medicine.*

Постановка проблеми

Персоналізована медицина потребує ефективного аналізу даних експресії генів для ранньої діагностики захворювань. Однак вибір оптимальних метрик близькості для кластеризації та класифікації залишається відкритою проблемою. Традиційні кореляційні метрики добре працюють для лінійних залежностей, але можуть бути не досить ефективними за складних взаємозв'язків. Метрики на основі взаємної інформації враховують нелінійні залежності, але мають високу обчислювальну складність. Відстань Васерштейна пропонує альтернативний підхід, що дозволяє порівнювати розподіли даних, проте потребує ретельної перевірки ефективності.

Актуальною є також необхідність інтеграції різних методів для підвищення точності та стабільності моделей. Гібридні моделі поєднують алгоритми машинного навчання та інтелектуального аналізу даних, що дозволяє краще обробляти високорозмірні біологічні дані. Одним із викликів є автоматизація підбору гіперпараметрів, що традиційно виконується вручну. Використання Баєсової оптимізації може значно підвищити ефективність цього процесу.

Ще однією проблемою є стабільність отриманих результатів. Використання стекінг-моделі дозволяє компенсувати помилки кластеризації та забезпечити узгодженість прогнозів. Розв'язання цих проблем сприятиме створенню надійної системи аналізу генетичних даних для персоналізованої медицини.

Аналіз останніх досліджень та публікацій

Оцінювання метрик близькості для аналізу високорозмірних генетичних даних є предметом численних досліджень, оскільки від вибору метрики залежить ефективність кластеризації та класифікації. Серед основних підходів виділяються метрики, засновані на взаємній інформації, кореляційні метрики та відстань Васерштейна.

Метрики, засновані на взаємній інформації, широко застосовуються для оцінювання залежностей між змінними та використовуються для відбору ознак, кластеризації та аналізу генетичних мереж [1–3]. Вони добре виявляють нелінійні зв'язки, але мають високу обчислювальну складність, що обмежує їх використання у великих наборах даних. Низка досліджень показала, що поєднання взаємної інформації з іншими методами, як-от графові моделі або підходи Баєса, дозволяє підвищити точність оцінки подібності між профілями експресії генів [2; 4].

Кореляційні метрики є традиційними інструментами для аналізу схожості профілів експресії [5; 6]. Вони ефективні для виявлення лінійних зв'язків, проте менш чутливі до складних нелінійних залежностей. Кореляційна відстань активно застосовується в біологічних дослідженнях, зокрема для аналізу експресії генів і класифікації біологічних станів. Використання цих метрик у поєднанні з алгоритмами кластеризації дозволяє формувати структуровані підгрупи зразків, що важливо для діагностики та прогнозування захворювань.

Відстань Васерштейна (Wasserstein distance), що базується на теорії оптимального транспорту, демонструє високу ефективність у завданнях кластеризації та зменшення розмірності даних [7; 8]. Ця метрика враховує розподіл значень і є менш чутливою до локальних шумів, що робить її привабливою для аналізу генетичних даних [9; 10]. У низці досліджень вона використовується для виявлення підтипів раку й інших складних біологічних систем, де традиційні евклідові метрики можуть бути не досить ефективними [7].

https://doi.org/10.32782/mathematical-modelling/2025-8-1-28

Гібридні підходи до обробки генетичних даних активно розвиваються та дозволяють поєднувати переваги різних методів [11; 12]. Наприклад, інтеграція алгоритмів машинного навчання з методами інтелектуального аналізу даних дозволяє досягти більшої точності у визначенні значущих генів і класифікації зразків [13; 14]. Одним із перспективних напрямів є застосування стекінг-моделей, що об'єднують різні класифікатори для покращення стабільності результатів [15].

Окрім вибору метрики, ключовим викликом є автоматизація процесу обробки даних і оптимізація параметрів моделей. Традиційні методи потребують ручного налаштування параметрів, що значно знижує ефективність аналізу великих масивів даних. Використання Баєсової оптимізації для автоматичного підбору параметрів дозволяє покращити результати класифікації та зменшити час обчислень [12].

Отже, останні дослідження підтверджують важливість вибору метрики для аналізу генетичних даних і ефективність гібридних підходів у підвищенні точності класифікації. Проте питання щодо оптимального поєднання цих методів та їх автоматизації залишаються відкритими й потребують подальших досліджень.

Мета дослідження

Метою дослідження є розроблення та впровадження гібридної моделі для оцінювання ефективності метрик близькості у високорозмірних даних експресії генів з інтегруванням методів інтелектуального аналізу даних і машинного навчання.

Виклад основного матеріалу дослідження

Блок-схема моделі, що використовувалася у процесі моделювання для оцінювання ефективності метрик близькості профілів експресії генів, зображена на рисунку 1.

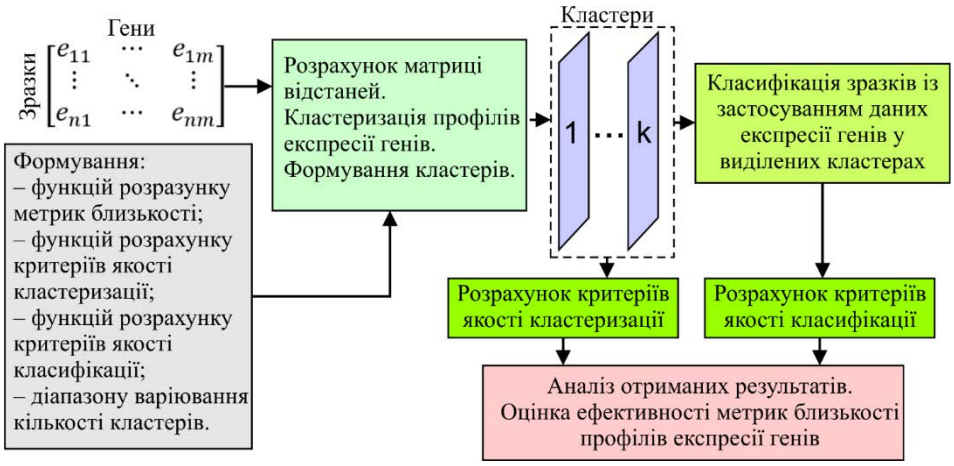

Рис. 1. Блок-схема моделі оцінювання ефективності метрик близькості профілів експресії генів на основі комплексного застосування методів кластеризації та класифікації зразків на основі даних експресії генів у виділених кластерах

Практична реалізація моделі передбачає наявність таких етапів:

1. Формування експериментальних даних експресії генів у формі матриці, де рядки – зразки, що досліджуються (або умови проведення експерименту), а стовпчики – гени, значення експресії яких визначає рівень активності окремої функції біологічного організму.

2. Формування функцій розрахунку метрик близькості, функції розрахунку критерію якості кластеризації, критеріїв якості класифікації. Ініціалізація інтервалу зміни кількості кластерів у процесі застосування алгоритму кластеризації: $k = \overline{2, k_{max}}$.

https://doi.org/10.32782/mathematical-modelling/2025-8-1-28

3. Вибір першої метрики зі списку сформованих функції розрахунків метрик $it_{metr} = 1$.

4. Розрахунок матриці відстаней між усіма парами профілів експресії генів із застосуванням поточної метрики.

5. Ініціалізація першого значення зі списку кількості кластерів: $k = 2$.

6. Застосування алгоритму кластеризації до профілів експресії генів. Формування кластерів. Розрахунок критеріїв якості кластеризації.

7. Застосування класифікатора до даних експресії генів у сформованих кластерах. Розрахунок критеріїв якості класифікації.

8. Якщо $k < k_{max}$, збільшення значення *k* на 1 та перехід на крок 6 цієї процедури. У протилежному випадку, якщо дотримано умову $it_{metr} < it_{metr}^{max}$, вибір функції розрахунку наступної метрики та перехід на крок 4. Якщо остання умова не дотримана, перехід на наступний крок моделювання.

9. Обробка отриманих результатів шляхом створення діаграм залежності значень критеріїв кластеризації та класифікації від кількості кластерів для кожної метрики, діаграм часових витрат за застосування різних метрик у процесі кластеризації профілів експресії генів.

10. Аналіз отриманих результатів. Вибір оптимальної метрики як за часом обробки даних, так і за критеріями якості кластеризації та класифікації, з урахуванням їхньої узгодженості під час вибору оптимальної кластерної структури.

У рамках поточних досліджень вивчалися такі метрики близькості профілів експресії генів:

1. Метрики на основі взаємної інформації. Взаємна інформація (Mutual Information (далі – MI)) є мірою, що кількісно визначає ступінь залежності між двома випадковими величинами. Якщо дві величини цілком незалежні, взаємна інформація між ними дорівнюватиме нулю, а якщо вони цілком визначені одна одною, MI буде максимальною. Формально, взаємна інформація між двома дискретними змінними *X* і *Y* визначається як:

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right), \tag{1}$$

де: $p(x,y)$ – спільна ймовірність значень *x* і *y*; $p(x)$ та $p(y)$ – відповідні марґінальні ймовірності.

Взаємна інформація може бути також визначена через ентропію Шеннона:

$$MI(X,Y) = H(X) + H(Y) - H(X,Y), \tag{2}$$

де: $H(X)$ і $H(Y)$ – ентропія Шеннона векторів значень *X* та *Y* відповідно; $H(X,Y)$ – спільна ентропія *X* і *Y*.

Взаємна інформація вказує, наскільки інформація про одну змінну зменшує невизначеність щодо іншої змінної. Якщо *X* і *Y* незалежні, тоді спільна ймовірність *p (x,y)* дорівнює добутку марґінальних імовірностей *p (x), p (y)*, $M(X,Y) = 0$. Якщо одна змінна цілком визначає іншу, тоді спільна ймовірність *p (x, y)* дорівнює ймовірності однієї змінної, взаємна інформація буде максимальною.

Формула (2) виражає залежність між двома випадковими змінними через їхні ентропії. Загальновідомо, що ентропія Шеннона являє собою міру невизначеності або розмаїття значень відповідного вектора *X* або *Y*. Якщо, наприклад, вектор *X* приймає різні значення з рівними ймовірностями, ентропія цього вектора буде максимальною. Аналогічно для вектора *Y*. Спільна ентропія *H (X, Y)* відображає невизначеність щодо спільної появи значень *X* і *Y*. Що більший ступінь незалежності *X* і *Y*, то вища спільна ентропія. Якщо *X* і *Y* цілком незалежні, їхня спільна ентропія дорівнюватиме сумі ентропій кожної зі змінних:

https://doi.org/10.32782/mathematical-modelling/2025-8-1-28

$$H(X,Y) = H(X) + H(Y), \tag{3}$$

взаємна інформація в такому разі дорівнює нулю: $MI(X,Y) = 0$. Це означає, що знання значення одного вектора змінних не дає жодної інформації про значення іншого вектора. Якщо вектори змінних *X* і *Y* є цілком залежними (тобто за знання однієї змінної можна точно передбачити іншу), тоді спільна ентропія *H (X, Y)* дорівнюватиме ентропії однієї із цих змінних (оскільки невизначеність залишатиметься лише в одній змінній). У такому разі:

$$H(X,Y) = H(X) = H(Y). \tag{4}$$

Отже, у такому разі взаємна інформація буде максимальною і дорівнюватиме ентропії будь-якої змінної:

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) = H(X) = H(Y). \tag{5}$$

Тому в разі максимальної схожості між векторами взаємна інформація досягає свого максимального значення. Варто відзначити, що взаємна інформація сама собою є мірою залежності між двома змінними, але вона не є відстанню у класичному розумінні, оскільки не задовольняє властивостей метрики (наприклад, вона не є симетричною і не завжди виконує нерівність трикутника). Однак існують різні способи перетворення взаємної інформації на метрики, які дозволяють використовувати її для вимірювання відстані між об'єктами. У рамках поточних досліджень використовувалася інформаційна метрична відстань (Information Metric Distance), яка є одним зі способів перетворення взаємної інформації на відстань:

$$D_{IM}(X,Y) = \sqrt{2(H(X|Y) + H(Y \mid X))} = \sqrt{2\left(H(X) + H(Y) - 2MI(X,Y)\right)}. \tag{6}$$

Варто зазначити, що ця метрика симетрична й задовольняє нерівність трикутника, тому за критеріями метричного простору її можна вважати справжньою метрикою.

Загальновідомо, що основною мірою в інформаційній теорії є ентропія Шеннона для дискретного випадкового вектора *X*, що визначається як:

$$H(X) = -\sum_{x\in\chi} P(x)\log_2\left(P(x)\right), \tag{7}$$

де $P(x)$ – імовірність того, що змінна вектора *X* приймає значення *x*, а χ – простір можливих значень вектора *X*. Ця формула відображає кількість інформації або невизначеності, яка міститься в розподілі вектора *X*.

Для двох векторів змінних *X* і *Y* спільна ентропія $H(X,Y)$ визначається як:

$$H(X,Y) = -\sum_{x\in\chi}\sum_{y\in\psi} P(x,y)\log_2\left(P(x,y)\right), \tag{8}$$

де $P(x,y) = \frac{n_{xy}}{N}$ – спільна ймовірність значень векторів *X* і *Y*, а ψ – простір можливих значень вектора *Y*.

https://doi.org/10.32782/mathematical-modelling/2025-8-1-28

Наявні методи оцінювання взаємної інформації розрізняються залежно від способу оцінювання ймовірності виникнення відповідних подій. До найбільш поширених можна віднести такі:

– Метод на основі максимальної правдоподібності (Empirical Mutual Information (далі – EMI)).

– Метод Jeffreys. Метод використовують, коли немає апріорної інформації про ймовірності. Він дає перевагу уникнення занадто сильного впливу нульових частот, але не надає вагомого апріорного перекосу.

– Метод Laplace. Застосування методу передбачає додавання фіктивного спостереження для кожної категорії. Це класичний підхід регуляризації, який також називають корекцією Лапласа. Він працює краще для малих вибірок, коли деякі категорії можуть не з'являтися.

– Метод Schurmann-Grassberger (далі – SG). Метод призначений для малих вибірок.

– Метод Minimax. Метод намагається мінімізувати максимальну похибку.

– Оцінка взаємної інформації за методом Джеймса та Стейна (Shrinkage Estimate of Mutual Information). Метод використовує технологію «стискання» (shrinkage), щоб поєднати емпіричні ймовірності з регуляризаційним параметром для уникнення переобчислення на малих вибірках. Це допомагає зменшити вплив рідкісних подій і стабілізувати оцінки ймовірностей.

2. **Кореляційна метрика.** Кореляційна метрика, зокрема кореляція Пірсона, є мірою лінійної залежності між двома змінними або векторами. У контексті профілів експресії генів вона дозволяє оцінити схожість між рівнями експресії генів у двох профілях. Коефіцієнт кореляції Пірсона вимірює силу й напрямок лінійного зв'язку між двома профілями експресії генів:

$$r(X,Y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\sum_{i=1}^{n}(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}, \tag{9}$$

де: x_i і y_i – значення експресії генів у двох профілях X і Y відповідно; $\bar{x}$ і $\bar{y}$ – середні значення рівнів експресії у профілях X і Y; n – кількість вимірів експресії для кожного зразка, що досліджується.

Загальновідомо, що значення коефіцієнта кореляції Пірсона змінюється в межах від −1 (ідеальна негативна лінійна залежність) до 1 (ідеальна позитивна лінійна залежність). За $r=0$ кореляція між профілями відсутня. Відстань між профілями експресії генів у такому разі може бути розрахована за формулою:

$$d(X,Y)=\sqrt{1-r(X,Y)}. \tag{10}$$

Якщо $r(X,Y)=1$, то $d(X,Y)=0$ (ідеальна кореляція), за $r(X,Y)=0$ $d(X,Y)=1$, за $r(X,Y)=-1$ $d(X,Y)=\sqrt{2}$. Отже, ця трансформація задовольняє вимогу, що метрика відстані повинна бути невід'ємною та симетричною. Використання квадратного кореня дозволяє «згладити» результат і зробити його пропорційним стандартним уявленням відстані.

3. **Метрика на основі відстані Васерштейна.** Відстань Васерштейна, також відома як «землеробна відстань» (*earth mover's distance*), є мірою відстані між двома ймовірнісними розподілами [10]. Ця метрика описує мінімальну кількість «роботи», необхідну для того, щоб перетворити один розподіл на інший. Відстань Васерштейна широко використовується для порівняння розподілів у багатьох сферах, зокрема в біоінформатиці для аналізу профілів експресії генів. Для двох одномірних імовірнісних розподілів P і Q відстань Васерштейна першого порядку (випадок загальної відстані Канторовича) визначається як:

$$W(P,Q)=\inf_{\gamma\in\Gamma(P,Q)}\int|x-y|\,d\gamma(x,y), \tag{11}$$

https://doi.org/10.32782/mathematical-modelling/2025-8-1-28

де: $\Gamma(P,Q)$ – множина всіх можливих сполучень (транспортувань) між розподілами P і Q; $d\gamma(x,y)$ – порція інформації (маса), яку потрібно перемістити з точки x у P до точки y у Q; $|x-y|$ – відстань між двома точками в метричному просторі.

Отже, відстань Васерштейна вимірює мінімальну роботу, необхідну для того, щоб перетворити один профіль експресії на інший, з урахуванням різниці в рівнях експресії окремих генів.

Оцінювання ефективності метрик здійснювалось із застосуванням алгоритму кластеризації k-Medoids, який використовує як центри кластерів реальні точки даних (на відміну від K-Means, який використовує середні значення центрів кластерів). K-Medoids мінімізує суму відстаней між точками у кластерах і центрами цих кластерів (медоїдами). Вибір цього алгоритму визначається тим фактом, що він може бути реалізований на основі матриці відстаней (dist), яка містить відстані між кожною парою точок, що в рамках поточних досліджень є визначальним.

До переваг алгоритму варто віднести той факт, що алгоритм k-Medoids може працювати з будь-якими мірами відстаней, зокрема й неевклідовими метриками, як-от відстань на основі взаємної інформації, кореляційна відстань або відстань Васерштейна. Це є критично важливим під час роботи із профілями експресії генів, де не завжди доречна евклідова або Мангеттена відстані. Навіть більше, оскільки медоїд завжди є реальною точкою з набору даних, він не змінюється на «нереальні» значення, як це може статися в k-Means, де центр кластера може бути середнім значенням кількох точок, що сприяє підвищенню впливу викидів на результат кластеризації. До переваг алгоритму можна також віднести те, що k-Medoids не потребує обчислення середніх значень у кожній ітерації, що може бути важливим для наборів даних із великою кількістю вимірів, як у разі з експресією генів, де кількість генів може перевищувати десятки тисяч, а кількість зразків – тисячі.

Вочевидь, в оцінюванні якості кластерної структури критерій якості кластеризації має бути комплексним і містити компоненту, що враховує як характер розподілу профілів експресії генів в окремих кластерах, так і характер розподілу кластерів у просторі ознак. У рамках поточних досліджень перша компонента критерію якості кластеризації розраховувалася як середнє значення відстані від профілів до медоїда кластера, до якого належить цей профіль:

$$QC_{in} = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{N_i}\sum_{j=1}^{N_i} dist\left(x_{ij}, medoid_i\right), \tag{12}$$

де: k – кількість кластерів; N_i – кількість профілів експресії генів в i-му кластері; x_{ij} – j-й профіль, що в i-му кластері; $medoid_i$ – медоїд i-го кластера.

Друга компонента критерію розраховувалася як середнє значення відстані між медоїдами всіх кластерів, що становлять кластерну структуру:

$$QC_{out} = \frac{2}{k(k-1)}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} dist\left(medoid_i, medoid_j\right). \tag{13}$$

Загалом, краща кластеризація відповідає більш високій щільності розташування профілів експресії генів в окремих кластерах і меншій щільності розташування медоїдів у просторі ознак. У такому разі критерій якості кластеризації можна розрахувати так:

$$QC = \frac{QC_{in}}{QC_{out}}. \tag{14}$$

Менше значення критерію (14) відповідає кращій кластерній структурі.

https://doi.org/10.32782/mathematical-modelling/2025-8-1-28

Ефективність як запропонованих метрик, так і критеріїв якості кластеризації оцінювалася шляхом застосування перехресного контролю, якість кластеризації оцінювалася шляхом розрахунку як критерію кластеризації, так і критеріїв якості класифікації зразків, що містили як атрибути профілі експресії генів, виділені у сформовані кластери. Критерії якості класифікації розраховувалися на основі оцінки похибок першого та другого роду із застосуванням стандартних методів.

Експериментальні дані, що використовувалися у процесі моделювання, являють собою дані експресії генів, отриманих із застосуванням методу секвенування молекул РНК (RNA-Sequencing). Вони доступні на сайті проєкту TCGA (The Cancer Genome Atlas), спрямованого на дослідження генетичних змін, пов'язаних з онкологічними захворюваннями, шляхом геномного аналізу.

Для оцінювання ефективності запропонованого критерію якості кластеризації на основі різних метрик близькості у процесі моделювання розраховувалися також класичні внутрішні критерії якості кластеризації: Calinski – Harabasz, Silhouette та Davies – Bouldin. Оптимальна кластеризація відповідає максимальним значенням критеріїв Calinski – Harabasz і Silhouette та мінімальному значенню критерію Davies – Bouldin. Результати моделювання представлені в таблицях 1–3.

Таблиця 1

Результат кластеризації профілів експресії генів
за застосування кореляційної метрики близькості

Clusters	Quality	Calinski – Harabasz	Silhouette	Davies – Bouldin
2	0,740	618,810	0,037	4,847
3	0,732	448,261	−0,013	7,474
4	**0,720**	288,856	−0,030	7,065
5	0,721	262,884	−0,043	6,765
6	0,724	263,003	−0,062	6,094
7	0,731	**955,972**	−0,118	**4,541**
8	0,739	798,652	−0,143	5,534
9	0,737	818,621	−0,184	5,111
10	0,738	813,036	−0,184	4,525

Таблиця 2

Результат кластеризації профілів експресії генів
за застосування метрики близькості Васерштейна

Clusters	Quality	Calinski – Harabasz	Silhouette	Davies – Bouldin
2	0,201	**27 895,5989**	**0,527**	**0,684**
3	0,190	22 074,838	0,373	1,024
4	0,180	17 513,103	0,278	1,303
5	0,161	14 596,655	0,222	1,516
6	0,149	12 503,164	0,185	1,679
7	0,138	10 788,453	0,151	1,979
8	0,132	9 469,112	0,124	2,294
9	0,131	8 388,892	0,099	2,599
10	0,123	7 663,755	0,095	2,628

https://doi.org/10.32782/mathematical-modelling/2025-8-1-28

Таблиця 3

Результат кластеризації профілів експресії генів за застосування різних типів метрик близькості на основі оцінки взаємної інформації

Clusters	Quality	Calinski – Harabasz	Silhouette	Davies – Bouldin
Метод на основі максимальної правдоподібності				
2	1,125	**663,654**	**0,416**	**0,874**
3	**1,096**	419,2474	0,144	1,618
Метод Jeffreys				
2	1,115	741,757	**0,410**	**0,886**
3	**1,106**	**840,249**	0,132	2,231
Метод Laplace				
2	1,128	**2 458,667**	**0,147**	**2,169**
3	1,101	1 665,504	0,047	3,319
4	**1,083**	1 359,880	0,084	2,776
Метод Джеймса та Стейна				
2	1,126	692,4446	0,413	0,880

Аналіз отриманих результатів дозволив зробити висновок, що тільки за застосування кореляційної метрики та метрики на основі відстані Васерштейна профілі експресії генів розподіляються по кластерах від 2 до 10 приблизно рівномірно. У результаті застосування метрик на основі оцінки взаємної інформації кількість кластерів варіюється від двох, за застосування методу Джеймса та Стейна, до чотирьох, за застосування методу Лапласа. У разі збільшення кількості кластерів додаткові кластери містили один профіль, що не є припустимим. Із цієї причини кількість кластерів було обмежено. Аналіз отриманих результатів також дозволяє зробити висновок щодо розбіжності значень внутрішніх критеріїв під час оцінювання якості кластерної структури. Так, у разі застосування кореляційної метрики за запропонованим критерієм якості оптимальною є чотирикластерна структура (відповідає мінімальному значенню критерію), а за критеріями Calinski – Harabasz і Davies – Bouldin оптимальною є семікластерна структура. Критерій Silhouette у такому разі не є ефективним. За застосування метрики на основі відстані Васерштейна за всіма критеріями оптимальною є двокластерна структура. За збільшення кількості кластерів якість кластеризації за всіма критеріями погіршується. За застосування метрик на основі оцінки взаємної інформації результати кластеризації за вибраними критеріями також не узгоджуються один з одним, за оцінювання взаємної інформації за методом Джеймса та Стейна ідентифікована тільки двокластерна структура. Але варто зазначити, що за значеннями всіх критеріїв якості кластерної структури, що використовувалися у процесі моделювання, метрика на основі відстані Васерштейна є найбільш привабливою.

Подальшим кроком моделювання є застосування класифікатора до зразків, що містять як атрибути профілі експресії генів у виділених кластерах. Класифікація зразків на основі виділених у кластери даних експресії генів здійснювалася із застосуванням алгоритму Random Forest із 5-fold крос-валідацією. Цей вибір зумовлений результатами попередніх досліджень стосовно порівняльного аналізу різних типів класифікаторів під час ідентифікації об’єктів на основі даних експресії генів. Алгоритм Random Forest має декілька ключових переваг, які роблять його оптимальним вибором для завдання класифікації зразків на основі даних експресії генів. Результати моделювання для двокластерної та трикластерної структур зображені на рис. 2 і 3. Аналогічні результати отримані для інших кластерних структур. Аналіз отриманих результатів підтверджує зроблений у попередньому підрозділі висновок щодо більш високої ефективності кореляційної метрики та метрики на основі відстані Васерштейна порівняно з метриками на основі оцінки взаємної інформації. У результаті застосування метрик на основі взаємної інформації кластери профілів експресії генів формуються нерівномірно, що призводить до великої розбіжності в точності класифікації зразків на основі даних експресії генів у виділених кластерах за всіма критеріями, що використовувалися у процесі моделювання.

https://doi.org/10.32782/mathematical-modelling/2025-8-1-28

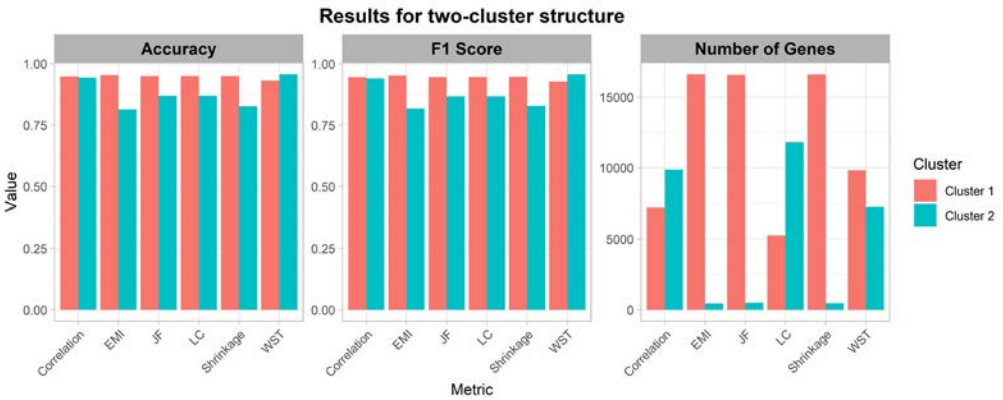

Рис. 2. Результати порівняльного аналізу ефективності метрик близькості для профілів експресії генів у двокластерній структурі

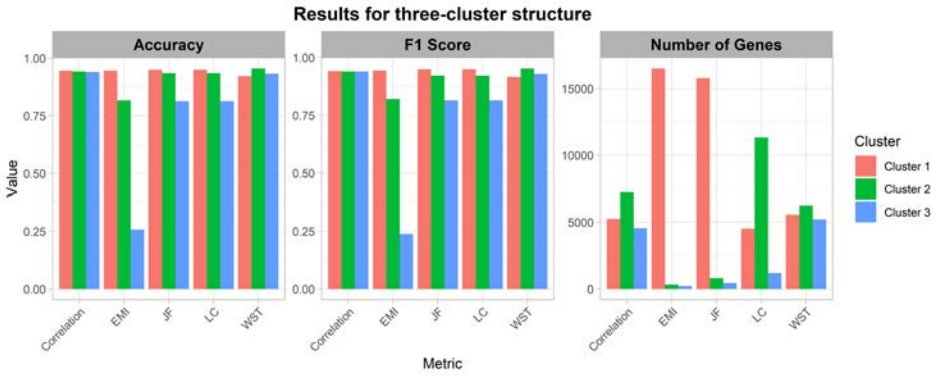

Рис. 3. Результати порівняльного аналізу ефективності метрик близькості для профілів експресії генів у трикластерній структурі

Навіть більше, метрики на основі взаємної інформації характеризуються невиправдано малою кількістю генів у деяких кластерах, що може свідчити про їхню нестабільність у процесі кластеризації. Аналіз отриманих результатів також дозволяє зробити висновок, що метрика Васерштейна демонструє стабільно високі значення точності класифікації (Accuracy) та F1-міри для структур із 2 до 10 кластерів. Вона перевершує кореляційну метрику, особливо в багатокластерних структурах (6–10 кластерів), де показує краще співвідношення між якістю класифікації (за відповідними критеріями) та рівномірністю кластеризації (кількість генів у кластерах). Метрика Васерштейна забезпечує більш збалансований розподіл генів між кластерами, що є ключовим чинником для забезпечення коректного аналізу кластерів у завданнях класифікації на основі даних експресії генів. Отже, результати поточного моделювання дозволяють зробити висновок, що метрики на основі взаємної інформації не забезпечують належного балансу між якістю класифікації та кількістю генів у кластерах, що робить їх менш ефективними у завданнях аналізу й обробки даних експресії генів. Водночас метрика Васерштейна, завдяки своїм високим показникам точності, F1-міри та збалансованому розподілу генів, є більш привабливим вибором порівняно з кореляційною метрикою та метриками на основі взаємної інформації.

Висновки

У статті запропоновано гібридну модель оцінювання ефективності метрик близькості для високорозмірних даних експресії генів, яка поєднує методи інтелектуального аналізу даних і машинного навчання. Запропонована модель дозволяє комплексно оцінити ефективність різних

https://doi.org/10.32782/mathematical-modelling/2025-8-1-28

підходів до вимірювання подібності профілів експресії генів, зокрема й кореляційну метрику, метрики на основі взаємної інформації та відстань Васерштейна. Аналіз отриманих результатів показав, що кореляційна метрика та відстань Васерштейна забезпечують вищу точність та стабільність кластеризації порівняно з метриками, заснованими на взаємній інформації. Відстань Васерштейна демонструє найбільш збалансований розподіл генів між кластерами та стабільно високі значення точності класифікації незалежно від кількості кластерів. Кореляційна метрика також показує високу ефективність, особливо щодо меншої кількості кластерів. Метрики, засновані на взаємній інформації, виявилися менш стабільними та не забезпечують рівномірного розподілу профілів, що призводить до значних розбіжностей у точності класифікації.

Запропонований підхід дозволяє автоматизувати процес вибору оптимальної метрики й адаптувати методи аналізу до конкретних наборів генетичних даних. Включення стекінг-моделі сприяє підвищенню точності класифікації, знижує вплив помилок кластеризації. Застосування розробленої моделі може сприяти вдосконаленню діагностичних систем у межах концепції персоналізованої медицини, забезпечити більш точну та стабільну ідентифікацію біомаркерів захворювань. Отримані результати можуть бути використані для розроблення нових алгоритмічних рішень у біоінформатиці, оптимізації процесів обробки генетичних даних і покращення систем підтримки ухвалення рішень у медичній діагностиці.

Список використаної літератури

1. Chen Y., Ye J., Li J. Aggregated wasserstein distance and state registration for hidden markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2020. Vol. 42 (9). P. 2133–2147.
2. Anh C.-T., Kwon Y.-K. Mutual information based on multiple level discretization network inference from time series gene expression profiles. *Applied Sciences (Switzerland)*. 2023. Vol. 13 (21). Art. 11902.
3. Barman S., Kwon Y.-K. A novel mutual information-based boolean network inference method from time-series gene expression data. *PLoS ONE*. 2017. Vol. 12 (2). Art. e0171097.
4. Pan X., Sun J., Yu H., Xue Y. Feature selection using non-dominant features-guided search for gene expression profile data. *Complex and Intelligent Systems*. 2023. Vol. 9 (6). P. 6139–6153.
5. Rezapour M., Walker S., Ornelles D., et al. A comparative analysis of rna-seq and nanostring technologies in deciphering viral infection response in upper airway lung organoids. *Frontiers in Genetics*. 2024. Vol. 15. Art. 1327984.
6. Bakry K., Emeish W., Embark H., et al. Expression profiles of four nile tilapia innate immune genes during early stages of aeromonas veronii infection. *Journal of Aquatic Animal Health*. 2024. Vol. 36 (2). P. 164–180.
7. Cao Q., Zhao J., Wang H., Guan Q., Zheng C. An integrated method based on Wasserstein distance and graph for cancer subtype discovery. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2023. Vol. 20 (6). P. 3499–3510.
8. Ocal K., Grima R., Sanguinetti G. Wasserstein distances for estimating parameters in stochastic reaction networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2019. Vol. 11773. P. 347–351.
9. Zhou K., Yin Z., Gu J., Zeng Z. A feature selection method based on graph theory for cancer Classification. *Combinatorial Chemistry and High Throughput Screening*. 2024. Vol. 27 (5). P. 650–660.
10. Zhang H. Feature selection using approximate conditional entropy based on fuzzy information granule for gene expression data classification. *Frontiers in Genetics*. 2021. Vol. 12. Art. 631505.
11. Almugren N., Alshamlan H. A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. *IEEE Access*. 2019. Vol. 7. P. 78533–78548.

https://doi.org/10.32782/mathematical-modelling/2025-8-1-28

12. Guelib B., Bounab R., Aliouane S., et al. Optimizing gene selection for alzheimer's disease classification: A Bayesian approach to filter and embedded techniques. *Applied Soft Computing*. 2024. Vol. 167. Art. 112307.
13. Wang T., Jia L., Xu J., et al. A hybrid intelligent optimization algorithm to select discriminative genes from large-scale medical data. *International Journal of Machine Learning and Cybernetics*. 2024. Vol. 15 (12). P. 5921–5948.
14. Yaqoob A., Verma N., Aziz R., Shah M. Rna-seq analysis for breast cancer detection: a study on paired tissue samples using hybrid optimization and deep learning techniques. *Journal of Cancer Research and Clinical Oncology*. 2024. Vol. 150 (10). Art. 455.
15. Esfandiari A., Nasiri N. Gene selection and cancer classification using interaction-based feature clustering and improved-binary bat algorithm. *Computers in Biology and Medicine*. 2024. Vol. 181. Art. 109071.

References

1. Chen, Y., Ye, J., & Li, J. (2020). Aggregated wasserstein distance and state registration for hidden markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 42 (9), 2133–2147 [in English].
2. Anh, C.-T., & Kwon, Y.-K. (2023). Mutual information based on multiple level discretization network inference from time series gene expression profiles. *Applied Sciences (Switzerland)*, 13 (21), Art. 11902 [in English].
3. Barman, S., & Kwon, Y.-K. (2017). A novel mutual information-based boolean network inference method from time-series gene expression data. *PLoS ONE*, 12 (2), Art. e0171097 [in English].
4. Pan, X., Sun, J., Yu, H., & Xue, Y. (2023). Feature selection using non-dominant features-guided search for gene expression profile data. *Complex and Intelligent Systems,* 9 (6), 6139–6153 [in English].
5. Rezapour, M., Walker, S., & Ornelles, D., et al. (2024). A comparative analysis of rna-seq and nanostring technologies in deciphering viral infection response in upper airway lung organoids. *Frontiers in Genetics*. 15. Art. 1327984 [in English].
6. Bakry, K., Emeish, W., & Embark, H. et al. (2024). Expression profiles of four nile tilapia innate immune genes during early stages of aeromonas veronii infection. *Journal of Aquatic Animal Health*, 36 (2), 164–180 [in English].
7. Cao, Q., Zhao, J., Wang, H., Guan, Q., & Zheng, C. (2023). An integrated method based on Wasserstein distance and graph for cancer subtype discovery. *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* 20 (6), 3499–3510 [in English].
8. Ocal, K., Grima, R., & Sanguinetti, G. (2019). Wasserstein distances for estimating parameters in stochastic reaction networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11773, 347–351 [in English].
9. Zhou, K., Yin, Z., Gu, J., & Zeng, Z. (2024). A feature selection method based on graph theory for cancer Classification. *Combinatorial Chemistry and High Throughput Screening,* 27 (5), 650–660 [in English].
10. Zhang, H. (2021). Feature selection using approximate conditional entropy based on fuzzy information granule for gene expression data classification. *Frontiers in Genetics*, 12, Art. 631505 [in English].
11. Almugren, N., & Alshamlan, H. (2019). A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. *IEEE Access*, **7**, 78533–78548 [in English].
12. Guelib, B., Bounab, R., & Aliouane, S., et al. (2024). Optimizing gene selection for alzheimer's disease classification: A Bayesian approach to filter and embedded techniques. *Applied Soft Computing,* 167, Art. 112307 [in English].

https://doi.org/10.32782/mathematical-modelling/2025-8-1-28

13. Wang, T., Jia, L., & Xu, J., et al. (2024). A hybrid intelligent optimization algorithm to select discriminative genes from large-scale medical data. *International Journal of Machine Learning and Cybernetics,* 15 (12), 5921–5948 [in English].
14. Yaqoob, A., Verma, N., Aziz, R., & Shah, M. (2024). Rna-seq analysis for breast cancer detection: a study on paired tissue samples using hybrid optimization and deep learning techniques. *Journal of Cancer Research and Clinical Oncology*, 150 (10), Art. 455 [in English].
15. Esfandiari, A., & Nasiri, N. (2024). Gene selection and cancer classification using interaction-based feature clustering and improved-binary bat algorithm. *Computers in Biology and Medicine*, 181, Art. 109071 [in English].

Ярема Олег Романович – к.т.н., доцент, доцент кафедри цифрової економіки та бізнес-аналітики Львівського національного університету імені Івана Франка. E-mail: oleh.yarema@lnu.edu.ua, ORCID: 0000-0003-3736-4820.

Сенчишен Денис Олександрович – аспірант кафедри комп'ютерних наук та програмної інженерії Херсонського державного університету. E-mail: dsenchishen@ksu.ks.ua, ORCID: 0000-0002-4311-7095.

Бабічев Сергій Анатолійович – д.т.н., професор, професор кафедри інформатики університету Яна Євангеліста Пуркінє в Усті-над-Лабем, Чехія; професор кафедри фізики Херсонського державного університету. E-mail: sergii.babichev@ujep.cz, sbabichev@ksu.ks.ua, ORCID: 0000-0001-6797-1467.

Yarema Oleh Romanovych – Candidate of Technical Sciences, Associate Professor, Associate Professor at the Department of Digital Economy and Business Analytics of the Ivan Franko National University of Lviv. E-mail: oleh.yarema@lnu.edu.ua, ORCID: 0000-0003-3736-4820.

Senchyshen Denys Oleksandrovych – Postgraduate Student at the Department of Computer Science and Software Engineering of the Kherson State University. E-mail: dsenchishen@ksu.ks.ua, ORCID: 0000-0002-4311-7095.

Babichev Sergii Anatoliiovych – Doctor of Technical Sciences, Professor, Professor at the Department of Informatics at Jan Evangelista Purkyně University in Ústí nad Labem, Czech Republic; Professor at the Department of Physics of the Kherson State University. E-mail: sergii.babichev@ujep.cz, sbabichev@ksu.ks.ua, ORCID: 0000-0001-6797-1467.

https://doi.org/10.32782/mathematical-modelling/2025-8-1-28