УДК 004.043
Сергій Бабічев, Тетяна Гончаренко
Херсонський державний університет, Херсон, Україна
ORCID 0000-0001-6797-1467
ORCID 0000-0002-2021-9320

ЗАСТОСУВАННЯ БІКЛАСТЕРНОГО АНАЛІЗУ ДЛЯ ФОРМУВАННЯ ПІДМНОЖИН КОЕРЕНТНИХ ДАНИХ

DOI 10.14308/ite000780

У статті запропоновано новий підхід до аналізу даних за допомогою бікластерного аналізу, який суттєво відрізняється від традиційної кластеризації. Проаналізовано наявні наукові праці, у яких схарактеризовано метод бікластерного аналізу та особливості його застосування. Автори зосереджуються на виявленні коерентних підмножин у складних даних, що виходять за межі звичайних даних, наприклад, експресії генів. Основну увагу приділено дослідженню, як бікластерний аналіз може виявляти приховані зв'язки у даних, що часто залишаються непоміченими при використанні традиційних методів. Сформовано критерії якості бікластеризації даних експресії генів та оцінено ефективність внутрішніх критеріїв Автори детально розглядають якість бікластерів, використовуючи середньоквадратичну помилку (MSE) та взаємну інформацію, забезпечуючи в такий спосіб достовірність та об'єктивність результатів. Особливість бікластерного аналізу полягає у його здатності виявляти бікластери різних розмірів та форм, що є важливим для розуміння складних та неоднорідних даних. Це дозволяє не тільки виділяти локальні шаблони в підмножинах даних, але й розкривати більш складні взаємозв'язки. Стаття також акцентує на оптимізації гіперпараметрів і використанні критеріїв якості для досягнення найбільш точних результатів. Мета дослідження полягає не тільки в ідентифікації коерентних підмножин даних, але й у глибшому розумінні структурних особливостей і взаємозв'язків, відкритих завдяки бікластерному аналізу. Ця робота відкриває нові перспективи для аналізу складних даних, пропонуючи більш глибоке розуміння їх структури та динаміки. Особливо цінним є здатність методу виявляти перетинні бікластери, що сприяє виявленню складніших і глибших залежностей у даних.

***Ключові слова**: бікластерний аналіз, штучні бікластери, критерії якості бікластеризації, взаємна інформація, середньоквадратична помилка*

1. Вступ

Постановка проблеми. Сучасний аналіз даних набуває нових горизонтів із використанням методів бікластерного аналізу, які принципово відрізняються від традиційної кластеризації. У традиційній кластеризації основну увагу зосереджено на групуванні об'єктів (рядків або стовпців) на основі їх подібності, нехтуючи можливою важливістю взаємодій між різними осями даних. Натомість бікластерний аналіз дозволяє одночасно групувати як рядки, так і стовпці, виділяючи взаємно корельовані підмножини даних, що є надзвичайно важливим для аналізу складних даних.

Головною метою нашого дослідження є розробка методик бікластерного аналізу, спрямованих на виявлення коерентних підмножин у складних даних, альтернативних до класичних даних експресії генів. Ми прагнемо дослідити, яким чином різні підходи до бікластеризації можуть виявити приховані зв'язки у даних, які можуть бути неочевидними при використанні традиційних методів. Особливу увагу ми приділяємо оцінці якості бікластерів за допомогою середньоквадратичної помилки (MSE) та взаємної інформації, щоб забезпечити достовірність та об'єктивність отриманих результатів.

Бікластерний аналіз відкриває можливості для детального розуміння структурних особливостей даних. Цей метод дозволяє не тільки виділяти локальні шаблони в підмножинах даних, але й розкриває різноманітні взаємозв'язки між різними компонентами даних. Однією з ключових особливостей бікластерного аналізу є його здатність виявляти бікластери різних розмірів та форм, що є особливо важливим для розуміння складних та неоднорідних даних. Крім того, бікластери можуть перетинатися, дозволяючи одному рядку чи стовпцю входити до декількох бікластерів одночасно, що відкриває шлях для виявлення більш складних і глибоких залежностей у даних.

У контексті цього дослідження ми плануємо детально вивчити та визначити кращі практики та методи бікластерного аналізу, зосередившись на оптимізації гіперпараметрів і використанні критеріїв якості для забезпечення надійності та точності отриманих результатів. Наша мета полягає не лише в ідентифікації коерентних підмножин даних, але й у глибшому розумінні взаємозв'язків і структурних особливостей, що відкриваються завдяки бікластерному аналізу.

Аналіз сучасних літературних джерел. Застосуванню бікластерного аналізу для обробки складних даних у наш час присвячено велику кількість наукових праць. Так, у [1] представлено огляд метаевристичних підходів до вирішення проблем бікластеризації, які ефективно вирішують складні оптимізаційні завдання в обмежений час обчислень і адаптуються до різних формулювань проблем. Особливу увагу приділено оптимізаційним методам і ключовим елементам пошуку: представленню, цільовій функції та операторам варіацій, з дискусією одно- та багатоцільових підходів і висвітленням нових напрямів досліджень. У [2] досліджується прихована блокова структура в гетерогенній моделі панельних даних, виходячи з припущення, що коефіцієнти регресії мають групові структури серед індивідів та структурні зміни у часі, де точки зміни можуть впливати на групові структури, а структурні зміни можуть варіюватися між підгрупами. Для відновлення прихованої блокової структури автори пропонують робастний бікластерний підхід, який використовує M-оцінювання та увігнуті штрафи об'єднання, а також розробляють алгоритм на основі локальної квадратичної апроксимації для оптимізації цільової функції, який є більш компактним та ефективним порівняно з алгоритмом ADMM. Крім того, встановлено оракульну властивість для штрафованих M-оцінювачів і доведено, що запропонований оцінювач із високою ймовірністю відновлює приховану блокову структуру, що також підтверджується добрими результатами на практиці за допомогою симуляційних досліджень на кількох наборах даних.

У [3] для покращення якості бікластеризації та вилучення модулів використовується комбінація методів, заснованих на теорії адаптивного резонансу (ART), – бікластеризація ARTMAP (BARTMAP) та топологічний ART (ToposART), що разом формують ToposBARTMAP, який успадковує здатність виявляти топологічні асоціації під час зменшення обсягу даних. Метод ToposBARTMAP був протестований на 35 реальних наборах даних про рак і порівняний з іншими методами (бі)кластеризації, показавши статистично значне покращення порівняно з іншими оціненими методами в експериментах з упорядкованими та перемішаними даними, а також продемонстрував кращі результати при ідентифікації бікластерів постійного, масштабного, зміщеного та зміщеного масштабного типів у експериментах з 12 синтетичними наборами даних. Графічне представлення було удосконалено для відображення асоціацій генних бікластерів і оцінено на наборі даних NCBI GSE89116, що містить рівні експресії 39,326 зондів, відібраних протягом 38 спостережень. У [4] запропоновано новий алгоритм бікластеризації для бінарних даних, названий Алгоритм Бінарної Бікластеризації на Основі Матриці Різниць Суміжності (AMBB), який покращує баланс між часом виконання та ефективністю. Алгоритм AMBB будує матрицю суміжності на основі значень різниць суміжності, а отриману підматрицю, яка оновлюється за допомогою матриці різниць суміжності, називають бікластером, що дозволяє групувати гени, які проявляють схожі реакції в різних умовах, важливих для подальшого аналізу генів, а експерименти на синтетичних та реальних наборах даних візуально демонструють високу практичність алгоритму AMBB.

Ураховуючи певні досягнення у сфері бікластерного аналізу для обробки складних даних, існують невирішені проблеми, до яких можна віднести відсутність ефективних методів оптимізації гіперпараметрів відповідного алгоритму. Це особливо актуально в контексті нових підходів, таких як комбінація методів на основі теорії адаптивного резонансу для бікластеризації, що вимагають точного налаштування гіперпараметрів для ефективної роботи. Також виокремлено проблему балансування між часом виконання та ефективністю алгоритмів, особливо в ситуаціях з бінарними даними, де потрібна розробка нових стратегій оптимізації, що забезпечать швидку та точну обробку даних.

2. Формування критеріїв якості бікластеризації даних експресії генів

У загальному випадку, критерії оцінки якості бікластерної структури можуть бути розділені на три групи:

1. *Внутрішні критерії*. Ці критерії дозволяють оцінити окремі бікластери у кластерній структурі без їх порівняння з еталонними бікластерами, формування яких у випадку застосування даних експресії генів є проблематичним. Найбільш розповсюдженими внутрішніми критеріями можна вважати наступні метрики:

●Середній квадратичний залишок (Mean Squared Residue – MSR) [5, 6]. Вимірює коерентність бікластера або ступінь узгодженості чи однорідності значень всередині бікластера. В ідеальному випадку, бікластер повинен містити дуже схожі значення, які формують певний шаблон або структуру. Ця однорідність може бути визначена як сталість, адитивність або множникова однорідність. Константна коерентність передбачає, що всі значення у бікластері є приблизно однаковими. При адитивній коерентності різниці між значеннями в рядках чи стовпцях є приблизно однаковими. Множникова коерентність передбачає, що значення в рядках чи/та стовпцях є приблизно однаковими після множення на певний множник. Для бікластера з I рядків і J стовпців формула розрахунку MSR критерія має вигляд:

$$MSR = \frac{1}{|I|\times|J|}\sum_{i\in I, j\in J}\left(x_{ij} - \overline{x}_i - \overline{x}_j + \overline{x}\right)^2 \quad (1)$$

де: x_{ij} – елемент матриці; $\overline{x}_i$, $\overline{x}_j$ та $\overline{x}$ – середні значення рядка i, стовпця j та загальний середній показник бікластера відповідно.

Менше значення критерія (1) свідчить про підвищену коерентність бікластера.

●Об’єм (Volume). Розраховується як добуток кількості рядків та стовпців. Більше значення цього критерія вказую на те, що бікластер відображає більш загальну структуру даних.

●Варіабельність (Variability). Вимірює розкид чи варіабельність значень усередині бікластера і може бути розрахованим за класичною формулою розрахунку дисперсії:

$$V = \frac{1}{|I|\times|J|}\sum_{i\in I, j\in J}\left(x_{ij} - \overline{x}\right)^2 \quad (2)$$

Менше значення варіабельності зазвичай вказує на більш стійкий шаблон у бікластері.

●Зв’язність (Connectivity). Вимірює кількість зв’язків або відносин між елементами у бікластері і може бути розрахована за формулою:

$$Connectivity = \frac{\text{Кількість зв язків між елементами у бікластері}}{\text{Максимальна можлива кількість зв'язків}} \quad (3)$$

Слід зазначити, що розрахунок значення цього критерія передбачає реконструкцію генної регуляторної мережі на попередньому кроці шляхом застосування відповідного алгоритму, при цьому одним із ключових гіперпараметрів є коефіцієнт трешолдінгу, значення якого суттєво впливає на кількість реальних зв’язків між елементами бікластера. При цьому вища зв’язність вказує на сильніші відношення між елементами у бікластері.

●Стійкість (Robustness). Оцінює, наскільки стійкими залишаються бікластери при незначних змінах у вхідних даних, наприклад, при додаванні гаусівського “білого” шуму.

2. *Зовнішні критерії*. Зовнішні критерії якості бікластеризації базуються на порівнянні результатів бікластеризації з певним стандартом або зовнішньою еталонною структурою.

Зазвичай такі критерії використовуються в тих випадках, коли доступні додаткові (зовнішні) дані про структуру даних або про справжній розподіл бікластерів. Найбільш розповсюджуваними на цей час зовнішніми критеріями якості бікластеризації є Індекс Ранду (Rand Index – RI) та Індекс Жаккара (Jaccard Index – JI).

●*Rand Index*. Визначає подібність між двома бікластеризаціями і заснований на похибках першого та другого роду [7]:

$$RI = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

де: *TP* – кількість пар об'єктів, правильно віднесених до одного бікластера; *TN* – кількість пар об'єктів, правильно віднесених до різних бікластерів; *FP* – кількість пар об'єктів, помилково віднесених до одного бікластера; *FN* – кількість пар об'єктів, які помилково вважаються належними до різних бікластерів.

●Jaccard Index. Розраховується як відношення спільних об'єктів у двох бікластерах до загальної кількості об'єктів у бікластерах (їх об'єднання) [8]:

$$JI = \lfloor \frac{B1 \cap B2}{B1 \cup B2} \rfloor \quad (5)$$

де B1 і B2 – два бікластера.

За наявності кількості бікластерів у бікластеризаціях більшої за 2 і за відсутності перетину між бікластерами формула розрахунку індексу Жаккара набуває вигляду:

$$JI(BC1,\ BC2) = \frac{1}{n_1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \lfloor \frac{B_i(BC1) \cap B_j(BC2)}{B_i(BC1) \cup B_j(BC2)} \rfloor \quad (6)$$

де: n_1 та n_2 – кількість бікластерів у бікластеризаціях $BC1$ і $BC2$ відповідно.

За наявності перетину між бікластерами формула (6) уточнюється так:

$$JI_{corrected}(BC1,\ BC2) = \frac{JI(BC1, BC2)}{max(JI(BC1,BC1), JI(BC2,BC2))} \quad (7)$$

Слід зазначити, що за відсутності перетину між бікластерами формула (7) трансформується у формулу (6), оскільки знаменник у формулі (7) дорівнюватиме одиниці. Зовнішні критерії, на відміну від внутрішніх, вимагають додаткової інформації щодо структури даних або правильний (еталонний) розподіл бікластерів, що у випадку застосування даних експресії генів, як експериментальних даних, є проблематичним. У більшості випадків вони використовуються для валідації різних алгоритмів бікластеризації.

2.1. *Внутрішній критерій якості бікластеризації на основі оцінки взаємної інформації*

Як було зазначено вище, бікластеризація являє собою процес одночасної кластеризації рядків та стовпців матриці. У контексті аналізу даних експресії генів, експериментальні дані представлені як матриця, де рядки є гени, а стовпці – умови проведення експерименту або навпаки, і значення у матриці відображають рівень вираження гену під певною умовою, тобто його експресію. Бікластер у цьому випадку визначає підмножину генів, які мають подібні профілі експресії в підмножині умов. Одним із способів оцінити якість бікластера є застосування аналізу взаємної інформації (BI) між рядками та стовпцями. BI може вказувати на те, наскільки інформація в рядках і стовпцях залежить одна від одної, і тому велике значення BI може вказувати на високу коерентність бікластера. До найбільш розповсюджених методів оцінки взаємної інформації слід віднести такі [9]:

●Взаємна інформація (Mutual Information – MI):

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (8)$$

де: X, Y є вектори, між яким здійснюється оцінка BI; $p(x,y)$ – спільний розподіл імовірностей X та Y; $p(x)\ i\ p(y)$ – маргінальні розподіли ймовірностей.

●Нормалізована взаємна інформація – визначається як відношення взаємної інформації до геометричного середнього ентропій двох векторів:

$$NMI(X,Y) = \frac{MI(X,Y)}{\sqrt{H(X),H(Y)}} \quad (9)$$

де $H(X)$ та $H(Y)$ – ентропії векторів X та Y відповідно.

●Відносна ентропія або Кульбак-Лейблерова дивергенція [10] – це міра відстані між двома розподілами ймовірностей:

$$D_{KL}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)} \quad (10)$$

де $P(i)$ – це ймовірність розподілу P, а $Q(i)$ – ймовірність розподілу Q.

Слід зазначити, що ця відстань не є симетричною, тобто $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, тому для підвищення об'єктивності варто обчислювати двосторонню дивергенцію з подальшим усередненням двох дивергенцій:

$$D_{KL}(P,Q) = \frac{D_{KL}(P||Q) + D_{KL}(Q||P)}{2} \quad (11)$$

Взаємна інформація є мірою спільної інформації між двома векторами випадкових величин, але вона сама по собі не є метрикою відстані. Трансформування значення BI у відстань може бути здійснено різними способами. У межах дисертаційних досліджень застосовується метрика на основі ентропії Шеннона:

$$d(X,Y) = H(X) + H(Y) - 2MI(X,Y) \quad (12)$$

де $H(X)$ та $H(Y)$ – значення ентропій Шеннона векторів X та Y відповідно, $MI(X,Y)$ – взаємна інформація між векторами X та Y. У цьому випадку, якщо розглядати два ідентичних розподіли даних, то $H(X) = H(Y) = MI(X,Y)$ і $d(X,Y) = 0$. При збільшенні різниці між розподілами даних значення взаємної інформації зменшується, що призводить до збільшення відстані між цими векторами.

Розрахунок значення внутрішнього критерія для оцінки коерентності бікластера передбачає оцінку середньої відстані як між рядками, так і між стовпцями бікластера. Покрокова процедура розрахунку цього критерія передбачає наявність таких кроків:

1. Розрахунок середньої відстані між усіма парами рядків бікластера:

$$QC_{row} = \frac{2}{nrow \times (nrow-1)} \sum_{i=1}^{nrow-1} \sum_{j=i+1}^{nrow} d(X_i, X_j) \quad (13)$$

2. Розрахунок середньої відстані між усіма парами стовпців бікластера:

$$QC_{col} = \frac{2}{ncol \times (ncol-1)} \sum_{i=1}^{ncol-1} \sum_{j=i+1}^{ncol} d(Y_i, Y_j) \quad (14)$$

3. Розрахунок середнього значення критеріїв (4.13) і (4.14):

$$QC = \frac{QC_{row} + QC_{col}}{2} \quad (15)$$

Мінімальне значення критерія (15) відповідає максимальному рівню коерентності бікластера. При цьому слід зазначити, що при застосуванні будь-якого алгоритму кластеризації до даних експресії генів, відмінною рисою яких є великий обсяг даних, можливе виникнення досить великої кількості бікластерів з низьким значенням коерентності, які не дозволяють однозначно ідентифікувати клас зразків, що досліджуються. Більш того, архітектура бікластеризації великою мірою визначається параметрами відповідного алгоритму, що використовується для формування кластерної структури. Тому виникає також проблема оптимізації параметрів алгоритму, для вирішення якої в межах поточних досліджень використовується алгоритм оптимізації Байєса, застосування якого передбачає такі етапи:

Етап I. Визначення цільової функції.

1.1. Вибір алгоритму бікластеризації, який приймає на вхід значення параметрів цільової функції. Застосування алгоритму до даних експресії генів. Формування бікластерної структури.

1.2. Вибір бікластера та оцінка його коерентності за формулами (8) – (15).

Етап II. Визначення діапазону зміни параметрів.

2.1. Для кожного параметру визначення діапазону варіювання його значень.

Етап III. Вибір моделі та запуск алгоритму оптимізації.

3.1. Вибір моделі алгоритму оптимізації Байєса. У межах досліджень використовувалася модель на основі гаусівських процесів.

3.2. Застосування алгоритму оптимізації Байєса з використанням обраної моделі. Формування найкращої комбінації гіперпараметрів за сформованою цільовою функцією.

Етап IV. Перевірка результату та формування компромісного рішення щодо оптимальної комбінації гіперпараметрів.

4.1. Застосування вищезазначеної процедури до перших п’яти (кількість бікластерів може варіюватися у процесі моделювання) бікластерів із подальшим аналізом отриманих результатів із метою формування компромісного рішення щодо оптимальної комбінації параметрів алгоритму.

Етап V. Застосування алгоритму бікластеризації до даних експресії генів. Формування бікластерної структури. Оцінка коерентності виділених бікластерів та формування підмножини бікластерів із високим значенням коерентності для подальших досліджень.

3. Оцінка ефективності внутрішніх критеріїв якості бікластеризації із застосуванням штучних бікластерів

Оцінка ефективності внутрішніх критеріїв якості бікластеризації здійснювалася із застосуванням штучних даних, які містили п’ять однакових за розміром коерентних (з різним ступенем коерентності) бікластерів, що не перетинаються один з одним (рис. 1). Як можна бачити з рисунку, синтетичні дані містять п’ять бікластерів, які можуть бути ідентифіковані шляхом застосування відповідного алгоритму бікластерного аналізу даних. Однак слід зазначити, що порівняння бікластеризацій, отриманих при застосуванні алгоритму бікластеризації із досконалою бікластеризацією шляхом розрахунку відповідних зовнішніх критеріїв якості бікластеризації у цьому випадку не є об’єктивним, оскільки при цьому можливе виділення великої кількості малих бікластерів (результати моделювання це підтвердило), які можуть суттєво вплинути на значення зовнішнього критерія.

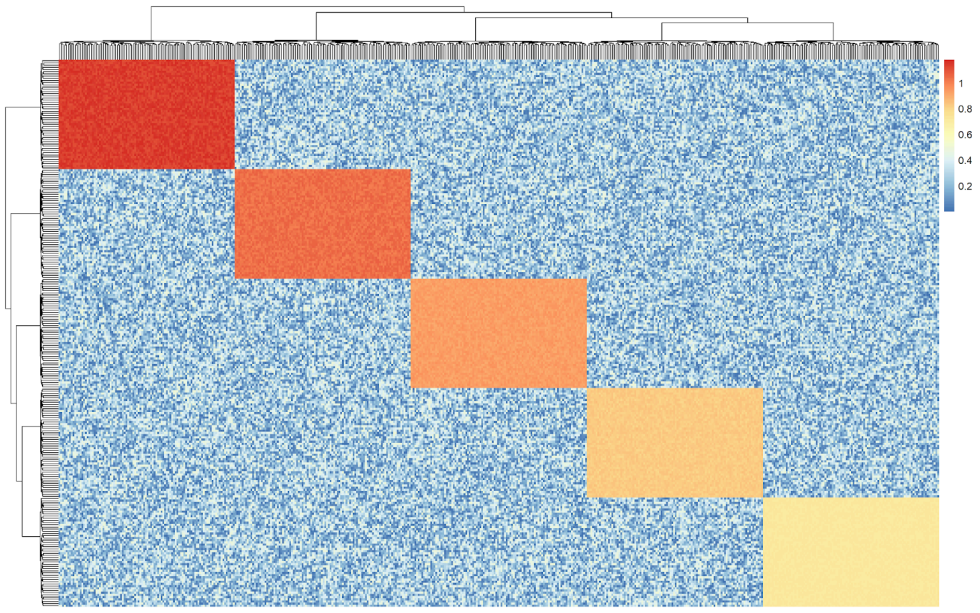

Рис. 1. Теплова карта розподілу бікластерів у синтетичних даних

Ураховуючи зазначене, оцінка ефективності відповідних внутрішніх критеріїв якості бікластеризації синтетичних даних, що представлені на рис. 1, здійснювалася шляхом порівняння значень критерії, розрахованих на п’яти перших бікластерах, зі значеннями цих

критеріїв, розрахованих на сформованих бікластерах, при цьому розраховувалися відносні відхилення значень відповідних критеріїв:

$$QC_{rel} = \frac{|QC_{exp} - QC_{perf}|}{QC_{perf}} \quad (16)$$

де: QC_{exp} – значення відповідного критерія якості бікластеризації (MSR або MI), що отримано в процесі застосування алгоритму бікластерного аналізу до перших п'яти бікластерів; QC_{perf} – значення критеріїв, розрахованих для досконалої бікластеризації, що зображена на рис. 1.

Процес моделювання здійснювався у програмному середовищі R [11] із застосуванням пакету *biclust* [12], що містить функції застосування різних алгоритмів бікластерного аналізу. Ураховуючи дослідження, що представлені у [13], у межах поточного моделювання процес бікластеризації здійснювався із застосуванням алгоритму *ensemble* [14], ефективність якого за результатами моделювання, що представлені у [13], є суттєво вищою у порівнянням із застосуванням інших алгоритмів бікластерного аналізу. Результат роботи алгоритму *ensemble* визначається двома параметрами: коефіцієнтом трешолдінгу (*thr*) та приблизним відношенням кількості рядків та стовпців у бікластерах (*simthr*). Процес моделювання передбачав зміну значень цих параметрів у заздалегідь визначеному діапазоні з певним кроком із розрахунком абсолютних значень критеріїв MSR та MI та їх відносних значень, визначених за формулою 16. Оцінка ефективності критеріїв визначалася шляхом аналізу збіжностей між абсолютними та відносними значеннями.

Практична реалізація алгоритму передбачає такі етапи:

Етап I. Налаштування моделі.

1.1. Ініціалізація інтервалу та кроку зміни значень гіперпараметрів *thr* та *simthr* алгоритму бікластеризації *ensemble*.

1.2. Розрахунок значень критеріїв MSR та MI за формулами (1) і (12) для досконалої бікластеризації, при цьому на першому етапі розраховувалися значення критеріїв для кожного бікластера, а на другому – формувалося узагальнене значення кожного критерія як середнє арифметичне всіх значень відповідного критерія.

Етап II. Визначення оптимального значення параметру *thr*.

2.1. Фіксація значення параметру *simthr*, ураховуючи приблизне відношення рядків та стовпців у бікластерах (при поточному моделюванні це значення було вибране на рівні 0.3).

2.2. Ініціалізація початкового значення *thr* параметру (thr_{min}).

2.3. Застосування алгоритму бікластеризації *ensemble* з поточними параметрами до синтетичних даних.

2.4. Виділення п'яти перших бікластерів і розрахунок для кожного бікластера рівня його коерентності за формулами (13) – (15) із застосуванням метрик близькості на основі MSR та MI. Розрахунок середнього значення кожного критерія як середнє арифметичне значень критеріїв, що розраховані для кожного виділеного бікластера.

2.5. Розрахунок відносного критерія для кожної метрики за формулою (16).

2.6. Якщо виконується умова $thr < thr_{max}$, збільшення значення *thr* на крок його зміни та перехід на крок 2.3 цієї процедури. У протилежному випадку, аналіз отриманих результатів та фіксація оптимального значення *thr*, що відповідає мінімумам критеріїв якості бікластеризації, що використовуються.

Етап III. Визначення оптимального значення параметру *simthr*.

3.1. Ініціалізація початкового значення *simthr* параметру ($simthr_{min}$).

3.2. Реалізація кроків 2.3 – 2.5 цієї процедури.

3.3. Якщо виконується умова $simthr < simthr_{max}$, збільшення значення *simthr* на крок його зміни та перехід на крок 3.2 цієї процедури. У протилежному випадку, аналіз отриманих результатів та фіксація оптимального значення *simthr*, що відповідає мінімальним значенням критеріїв якості бікластеризації.

Етап IV. Формування бікластерної структури та аналіз отриманих результатів.

4.1. Застосування алгоритму бікластеризації *ensemble* з оптимальними параметрами до синтетичних даних.

4.2. Формування бікластерної структури та аналіз отриманих результатів.

На рис. 2. зображено результати моделювання щодо застосування вищезазначеного алгоритму до синтетичних даних для визначення оптимального значення параметру *thr* алгоритму бікластеризації *ensemble*. Аналіз отриманих результатів дозволяє зробити висновок, що значення відносних та абсолютних критеріїв якості бікластеризації змінюються узгоджено, що свідчить про адекватність метрик на основі MSR і MI для формування бікластерної структури. Оптимальне значення параметру *thr*, що відповідає мінімумам критеріїв, дорівнює 0.37.

Подальші результати моделювання дозволили зробити висновок, що для синтетичних даних значення параметру *simthr* при фіксованому значенні *thr* не впливає на результат бікластеризації. На рис. 3 зображено результат застосування алгоритму бікластеризації ensemble з оптимальними параметрами (*thr* = 0.37, *simthr* = 0.3) до синтетичних даних, зображених на рис. 1. Як можна бачити, кластерна структура повністю повторює кластерну структуру, що зображена на рис. 1. Цей факт свідчить про адекватність запропонованих критеріїв якості бікластеризації.

4. Висновки

Отримані у статті результати дослідження демонструють важливість бікластерного аналізу в сучасному аналізі даних, особливо у виявленні коерентних підмножин у складних даних. Представлений метод застосування бікластерного аналізу, що зосереджений на оптимізації гіперпараметрів і використанні критеріїв якості, дозволив нам не тільки ідентифікувати коерентні підмножини, але й глибше зрозуміти структурні особливості та взаємозв'язки в даних. Отримані результати моделювання підтвердили адекватність метрик на основі середньоквадратичної помилки (MSE) та взаємної інформації (MI) для формування бікластерної структури, де оптимальне значення параметра *thr* встановлено на рівні 0.37.

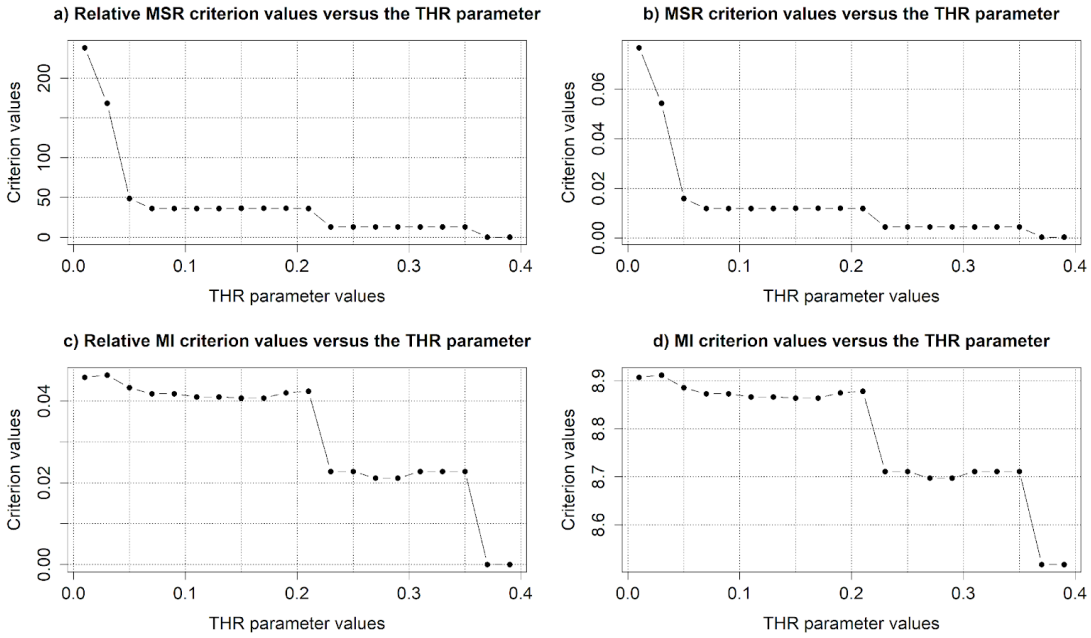

Рис. 2. Результати моделювання щодо визначення оптимального значення параметру thr алгоритму бікластеризації ensemble

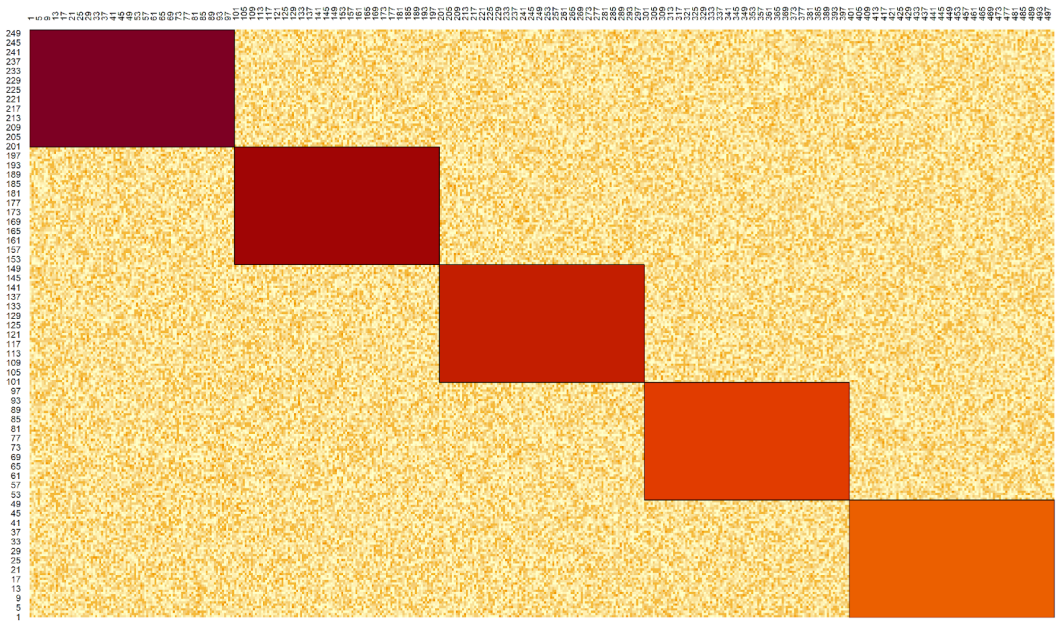

Рис. 3. Результат застосування алгоритму бікластеризації ensemble з оптимальними параметрами до синтетичних даних

Результати досліджень показали, що для синтетичних даних значення параметру *simthr* при фіксованому значенні *thr* не впливає на результат бікластеризації, що свідчить про стабільність нашого підходу. Застосування алгоритму бікластеризації *ensemble* з оптимальними параметрами (*thr* = 0.37, *simthr* = 0.3) до синтетичних даних продемонструвало повну відповідність між кластерною структурою модельних та реальних даних, що є підтвердженням адекватності запропонованих нами критеріїв якості для бікластеризації. Ці результати відкривають нові перспективи для застосування бікластерного аналізу в різних галузях, де аналіз складних даних є ключовим.

Подальшою перспективою досліджень авторів є застосування запропонованого методу до обробки більш складних даних експресії генів на основі комплексного застосування бікластерного аналізу та аналізу генної онтології.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. José-García, A., Jacques, J., Sobanski, V., Dhaenens, C. (2023). Metaheuristic Biclustering Algorithms: From State-of-the-art to Future Opportunities. *ACM Computing Surveys*, 56 (3), art. no. 69. DOI: 10.1145/3617590
2. Cui, W., Li, Y. (2023). Bicluster Analysis of Heterogeneous Panel Data via M-Estimation. *Mathematics*, 11 (10), art. no. 2333. DOI: 10.3390/math11102333
3. Yelugam, R., Brito da Silva, L. E., Wunsch, D. C. (2023). Topological biclustering ARTMAP for identifying within bicluster relationships. *Neural Networks*, 160, 34–49. DOI: 10.1016/j.neunet.2022.12.010
4. Chu, H.-M., Liu, J.-X., Zhang, K., Zheng, C.-H., Wang, J., Kong, X.-Z. (2022). A binary biclustering algorithm based on the adjacency difference matrix for gene expression data analysis. *BMC Bioinformatics*, 23 (1), art. no. 381. DOI: 10.1186/s12859-022-04842-4
5. Dutta, S., Hore, M., Ahmad, F., Saba, A., Kumar, M., Das C. (2019). SBI-MSREimpute: A sequential biclustering technique based on mean squared residue and euclidean distance to predict missing values in microarray gene expression data. *Advances in Intelligent Systems and Computing*, 813, 673–685. DOI: 10.1007/978-981-13-1498-8_59
6. Yin, L., Qiu, J., Gao, S. (2018). Biclustering of Gene Expression Data Using Cuckoo Search and Genetic Algorithm. *International Journal of Pattern Recognition and Artificial Intelligence*, 32(11), art. no. 1850039. DOI: 10.1142/S0218001418500398

7. Gates, A. J., Ahn, Y.-Y. (2017). The impact of random models on clustering similarity. *Journal of Machine Learning Research*, vol. 18, 1–28.
8. Babichev, S., Osypenko, V., Lytvynenko, V., Voronenko, M., Korobchynskyi, M. (2018). Comparison Analysis of Biclustering Algorithms with the use of Artificial Data and Gene Expression Profiles. I*EEE 38th International Conference on Electronics and Nanotechnology, ELNANO 2018 – Proceedings*, art. no. 8477439, 298–304.
9. Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27 (3), 379–423 & 623–656.
10. Kullback, S., Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22 (1), 79–86.
11. The R Project for Statistical Computing. Available on: https://www.r-project.org/
12. Kaiser S., Santamaria R., Khamiakova T, et al. Biclust: Bicluster Algorithms. Available on: https://cran.r-project.org/web/packages/biclust/index.html
13. Babichev, S., Osypenko, V., Lytvynenko, V., Voronenko, M., Korobchynskyi, M. (2018). Comparison Analysis of Biclustering Algorithms with the use of Artificial Data and Gene Expression Profiles. *IEEE 38th International Conference on Electronics and Nanotechnology, ELNANO 2018 – Proceedings*, art. no. 8477439, 298–304. DOI: 10.1109/ELNANO.2018.8477439
14. Kaiser, S. (2011). *Biclustering: Methods, Software and Application*. Dissertation, LMU München: Fakultät für Mathematik, Informatik and Statistik.

Sergii Babichev, Tetiana Goncharenko
Kherson State University, Kherson, Ukraine

APPLICATION OF BICLUSTERING ANALYSIS FOR FORMING COHERENT DATA SUBSETS

This paper introduces a new approach to data analysis using biclustering, which significantly differs from traditional clustering methods. Available scientific works were analyzed, which characterized the method of bicluster analysis and the features of its application. The authors focus on identifying coherent subsets within complex data, extending beyond typical data such as gene expression. They emphasize exploring how biclustering analysis can uncover hidden connections in data, often overlooked by conventional methods. Quality criteria for biclustering of gene expression data were formed and the effectiveness of internal criteria was evaluated. The quality of biclusters is thoroughly examined using mean squared error (MSE) and mutual information, ensuring the reliability and objectivity of the results. A distinctive feature of biclustering analysis is its ability to identify biclusters of various sizes and shapes, crucial for understanding complex and heterogeneous data. This approach not only highlights local patterns in data subsets but also reveals more intricate interrelations. The article also stresses the importance of optimizing hyperparameters and using quality criteria to achieve the most accurate results. The research aims not only to identify coherent data subsets but also to gain a deeper understanding of structural features and interconnections revealed by biclustering analysis. This work opens new prospects for analyzing complex data, offering a deeper insight into their structure and dynamics. Particularly valuable is the method's ability to detect overlapping biclusters, aiding in uncovering more complex and profound dependencies in the data.

Keywords: Biclustering analysis, artificial biclusters, quality criteria for biclustering, mutual information, mean squared error

Стаття надійшла до редакції 26.11.2023
The article was received 26 November 2023