

УДК 577.215.2:519.237.62:68.05

ГІБРИДНА МОДЕЛЬ ІДЕНТИФІКАЦІЇ ЗРАЗКІВ ДАНИХ ЕКСПРЕСІЇ ГЕНІВ НА ОСНОВІ КЛАСТЕР-БІКЛАСТЕРНОГО АНАЛІЗУ ІЗ ВИКОРИСТАННЯМ ЗГОРТКОВОЇ НЕЙРОННОЇ МЕРЕЖІ

І. М. Лях¹, Б. В. Дурняк²

¹ДВНЗ «Ужгородський національний університет»,
пл. Народна, 3, Ужгород, 88000, Україна

²Українська академія друкарства,
вул. Під Голоском, 19, Львів, 79020, Україна

У сучасній біоінформації аналіз даних експресії генів відіграє важливу роль у дослідженні біологічних процесів та механізмів регуляції генів. Зростаючі обсяги даних, що генеруються за допомогою мікроматриць ДНК та РНК-секвенування, роблять їх аналіз складним завданням. Традиційні методи аналізу даних, такі як кластерний аналіз, можуть не забезпечити достатньої точності та інформативності для виявлення біологічно значущих закономірностей.

Існують нові підходи, які поєднують декілька методів аналізу даних, що дають змогу отримати більш глибокий та всебічний аналіз. Одним з таких методів є кластер-бікластерний аналіз, який поєднує кластерний аналіз з бікластерним аналізом. Кластерний аналіз дає змогу групувати гени за подібністю їхніх профілів експресії, а бікластерний аналіз — виявляти групи генів, які експресуються одночасно в різних умовах. Тому досліджується ефективність застосування покрової процедури кластеризації та бікластеризації даних експресії генів з використанням аналізу генної онтології.

Ключові слова: експресія генів, кластеризація, бікластеризація, генна онтологія, аналіз даних.

Постановка проблеми. Аналіз даних експресії генів відіграє важливу роль у сучасній біології та медицині. Він дає змогу досліджувати механізми регуляції генів, виявляти біомаркери захворювань, розробляти нові методи лікування та діагностики. Однак складність та багатовимірність даних експресії генів робить їх аналіз складним завданням. Традиційні методи аналізу даних, такі як мікроматриці ДНК та РНК-секвенування, генерують великі обсяги даних, які важко інтерпретувати вручну.

Аналіз останніх досліджень та публікацій. Виконані дослідження з цієї тематики присвячені гібридним моделям ідентифікації зразків даних експресії генів на основі аналізу GO, а також кластерному та бікластерному аналізу [1–3]. Водночас недостатньо уваги приділено розробці нових гібридних моделей, які поєднують різні методи аналізу даних.

Мета статті — представлення гібридної моделі для ідентифікації зразків даних експресії генів, яка поєднує аналіз генної онтології (GO), спектральну кластеризацію, бікластерний аналіз та згорткову нейронну мережу. Ця модель має на меті покращити точність та ефективність ідентифікації біологічно значущих закономірностей та зв'язків між генами даних експресії генів.

Виклад основного матеріалу дослідження. Формування підмножин інформативних даних експресії генів може здійснюватися як із застосуванням кластерного аналізу, так і бікластерного аналізу на етапі передобробки даних. Водночас комплексне застосування бікластерного аналізу та аналізу генної онтології забезпечує більш глибокий і всебічний аналіз. Якщо бікластерний аналіз дає змогу виявити підмножини коерентних генів та зразків, враховуючи умови проведення експерименту, то аналіз генної онтології дає змогу сформувати список інформативних ідентифікаторів генів, враховуючи тип та стан біологічного об'єкта, що досліджується. Досліджується ефективність і доцільність застосування покрокової процедури кластеризації та бікластеризації даних експресії генів із використанням аналізу генної онтології як на рівні передобробки даних експресії генів, так і після реалізації процесу бікластеризації. Доцільність реалізації процедури кластер-бікластерного аналізу визначається такими причинами: реалізація етапу кластеризації дає змогу сформувати підмножини профілів експресії генів на основі їх подібності за обраною метрикою. Формування та оцінка кластерної структури здійснювалася із застосуванням метрики на основі оцінки взаємної інформації профілів експресії генів. Реалізація процедури бікластеризації на сформованих кластерах дає змогу виділити підмножини генів, які мають високу експресію для певних зразків або при певних умовах проведення експерименту. Таким чином, комбінація цих двох методів дає змогу забезпечити більш глибокий і всебічний аналіз. Кластеризація дає змогу виявити загальні тенденції в даних, а бікластеризація дозволяє виявити специфічні взаємозв'язки, які можуть бути важливими для ідентифікації стану об'єкта в системах діагностики [4].

На етапі реалізації бікластерного аналізу в процесі моделювання застосовувалися дані експресії генів, що отримані під час застосування алгоритму спектральної кластеризації. Аналіз результатів моделювання дав змогу сформувати два кластери даних експресії генів, які містили 6150 і 8466 генів, відповідно. Кожний кластер містив 3021 зразків, які відповідали шести типам раку. Одна група зразків відповідала об'єктам, у яких ракова хвороба не виявлена. На першому етапі визначалися оптимальні значення гіперпараметрів алгоритму бікластеризації *ensemble* (*thr* та *simthr*), які відіграють ключову роль у процесі формування бікластерів, із застосуванням алгоритму оптимізації Байєса. Якщо перший параметр визначає порогове значення, за якого елементи (гени або зразки) включаються у бікластер, то другий параметр визначає поріг подібності бікластерів. Більше значення цього параметра означатиме, що тільки дуже схожі бікластери будуть об'єднані разом. Тоді як нижчий поріг дозволить більшу різноманітність в ансамблі. Як цільова об'єктивна функція при застосуванні алгоритму оптимізації Байєса середнє значення відстані на основі оцінки взаємної інформації, що розраховувалася для рядків і стовпців п'яти перших бікластерів за формулами (1) – (3).

1. Розрахунок середньої відстані між усіма парами рядків бікластера:

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

2. Розрахунок середньої відстані між усіма парами стовпців бікластера:

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

3. Розрахунок середнього значення критеріїв (1) і (2):

$$QC = \frac{QC_{row} + QC_{col}}{2}. \quad (3)$$

На основі моделювання були визначені такі параметри алгоритму бікластеризації ensemble:

- для даних експресії генів першого кластеру: thr = 0.529, simthr = 0.248;
- для даних експресії генів другого кластеру: thr = 0.518, simthr = 0.267.

Результати моделювання щодо розподілу зразків і генів у виділених бікластерах зображені на рис. 1.

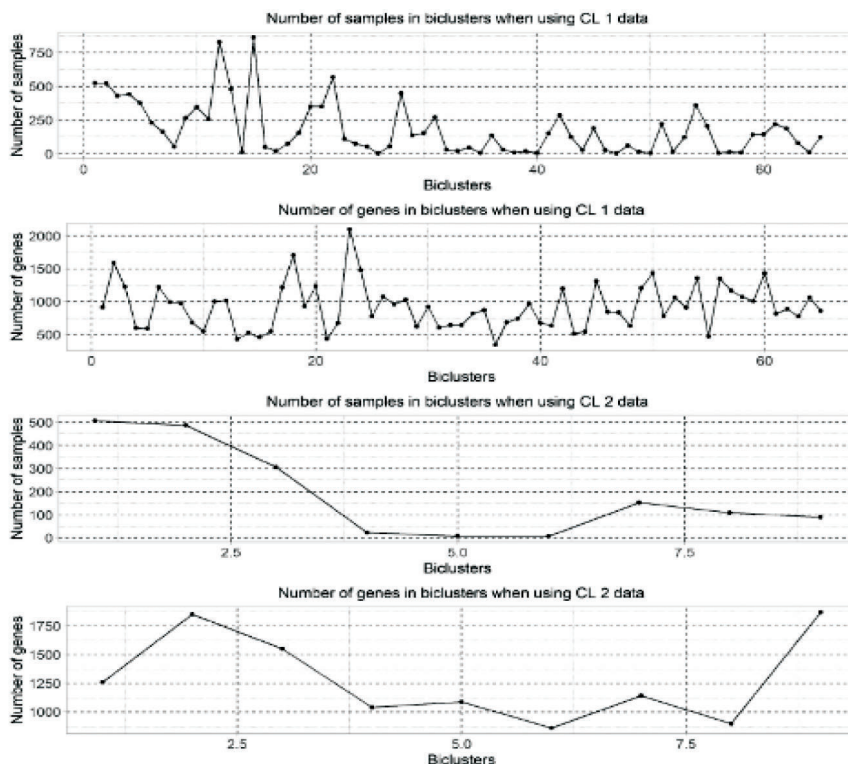


Рис. 1. Результати моделювання щодо застосування бікластерного аналізу до даних експресії генів у виділених кластерах

Як бачимо, при майже однакових параметрах алгоритму бікластеризації результати бікластеризації суттєво відрізняються між собою. Так, під час застосування

даних експресії генів першого кластера (6150 генів) отримано 65 бікластерів, водночас спостерігається велика кількість малих бікластерів, які містять відносно невелику кількість як зразків, так і генів. Результати бікластеризації на даних другого кластера (8466 генів) є більш привабливими як за кількістю кластерів (9), так і за наповненістю бікластерів [5].

На наступному етапі формувалася список унікальних ідентифікаторів генів шляхом застосування аналізу генної онтології до даних у виділених бікластерах кожного кластера на першому кроці та об'єднання виділених підмножин значущих генів з подальшим формуванням підмножин значущих профілів експресії генів на основі даних кожного кластера. За результатом реалізації цього кроку сформовані нові підмножини даних експресії генів, які містили 4442 і 4520 значущих генів, що відповідають даним першого та другого кластера, відповідно. Наступним кроком є застосування ЗНМ до сформованих даних відповідно до методики. У табл. 1 зображено результати застосування алгоритму оптимізації Байєса для визначення оптимальних гіперпараметрів ЗНМ у випадку застосування сформованих даних експресії генів у результаті кластер-бікластерного аналізу. Результат навчання та валідації моделей сформованих нейронних мереж із застосуванням 5-Fold крос-валідації зображені на рис. 2 і рис. 3. Аналіз отриманих діаграм дає змогу зробити висновок, що як і у випадку застосування даних експресії генів, отриманих шляхом застосування тільки кластерного аналізу, при застосуванні даних, отриманих при використанні кластер-бікластерного аналізу, перенавчання моделі не спостерігається, оскільки значення точності класифікації та функції втрат, отриманих на даних для навчання й валідації моделі, змінюються узгоджено в рамках допустимої похибки [6].

Таблиця 1

Результати моделювання щодо визначення оптимальних гіперпараметрів ЗНМ на основі даних, отриманих шляхом застосування кластер-бікластерного аналізу

Оптимальні гіперпараметри ЗНМ при застосуванні даних першого кластера							
Перший згортковий шар			Другий згортковий шар			Значення dropout	Кількість нейронів щільнісного шару
Кількість фільтрів	Розмір вікна фільтрації	Розмір вікна згортки	Кількість фільтрів	Розмір вікна фільтрації	Розмір вікна згортки		
52	6	4	15	3	3	0.47	159
Оптимальні гіперпараметри ЗНМ при застосуванні даних другого кластера							
Перший згортковий шар			Другий згортковий шар			Значення dropout	Кількість нейронів щільнісного шару
Кількість фільтрів	Розмір вікна фільтрації	Розмір вікна згортки	Кількість фільтрів	Розмір вікна фільтрації	Розмір вікна згортки		
39	7	2	59	10	3	0.29	64

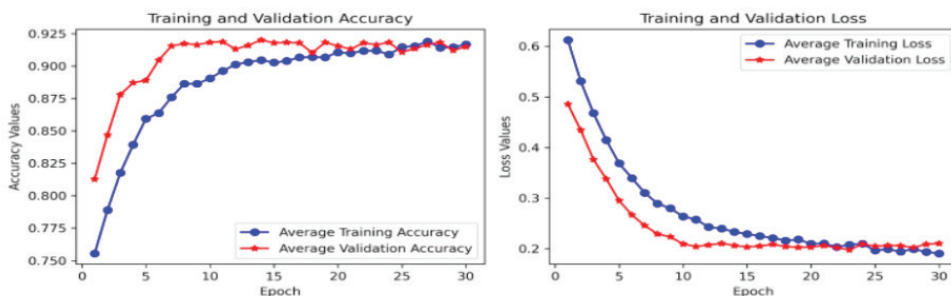


Рис. 2. Результати моделювання щодо навчання та валідації ЗНМ із застосуванням кластер-бікластерного аналізу при використанні даних експресії генів першого кластера

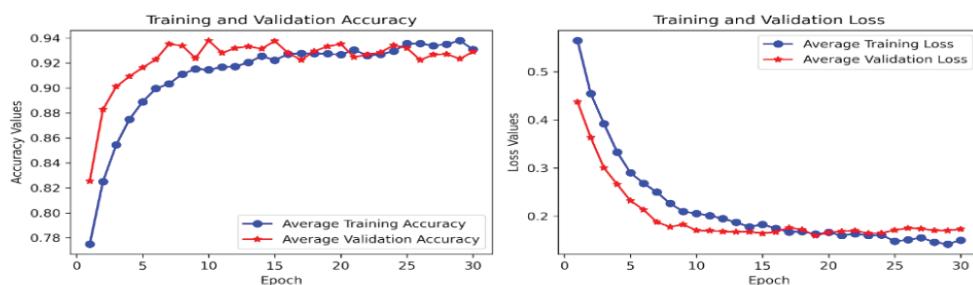


Рис. 3. Результати моделювання щодо навчання та валідації ЗНМ із застосуванням кластер-бікластерного аналізу при використанні даних експресії генів другого кластера

На рис. 4 і рис. 5 зображені результати застосування навчених моделей ЗНМ до тестових даних експресії генів. Аналіз отриманих результатів свідчить про більш високу ефективність моделей при застосуванні даних експресії генів, отриманих шляхом кластер-бікластерного аналізу. Цей факт підтверджує доцільність комплексного застосування кластерного аналізу, бікластеризації даних виділених кластерів та аналізу генної онтології для формування підмножин значущих генів, враховуючи тип об'єкта, що досліджується.

Застосування бікластерного аналізу до даних експресії генів відповідних кластерів на першому кроці та аналізу генної онтології на другому кроці призвів до зменшення кількості значущих генів з 6150 до 4442 при використанні даних першого кластера, та з 8466 до 4526 при використанні даних другого кластера. При цьому точність ідентифікації типу ракового захворювання у багатьох випадках підвищується, порівняно з результатами, що отримані на попередньому етапі моделювання при використанні даних, отриманих шляхом застосування тільки кластерного аналізу. Результати порівняльного аналізу точності класифікації зразків при застосуванні даних експресії генів, отриманих при використанні кластерного та кластер-бікластерного аналізу і аналізу генної онтології, зображені на рис. 6. Аналіз отриманих результатів підтверджує зроблений раніше висновок щодо низької точності класифікації зразків пацієнтів, у яких ракова хвороба не виявлена. Але

точність класифікації цих зразків також є вищою при застосуванні другого типу даних, порівняно з даними, отриманими як результат кластерного аналізу [7].

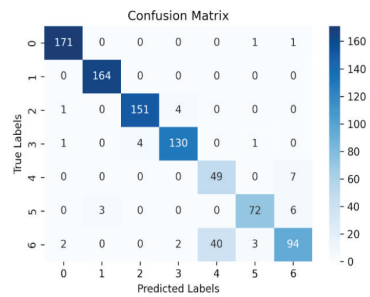


Рис. 4. Матриця збіжності, що сформована за результатом застосування моделі ЗНМ до даних експресії генів першого кластера, отриманих шляхом застосування кластер-бікластерного аналізу та аналізу генної онтології

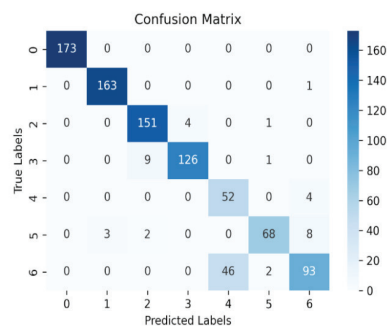


Рис. 5. Матриця збіжності, що сформована за результатом застосування моделі ЗНМ до даних експресії генів другого кластера, отриманих шляхом застосування кластер-бікластерного аналізу та аналізу генної онтології

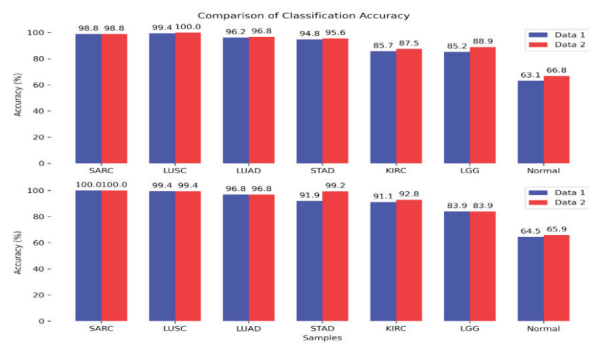


Рис. 6. Порівняльний аналіз результатів точності класифікації зразків на основі даних експресії генів, отриманих шляхом застосування кластерного аналізу та кластер-бікластерного аналізу і аналізу генної онтології: у верхньому рядку наведені результати аналізу даних першого кластера, а у другому — другого кластера

Точність класифікації типу раку зразків у більшості випадків також є вищою при застосуванні даних експресії генів, отриманих при використанні кластер-бі-кластерного аналізу та аналізу генної онтології, що свідчить про ефективність застосування запропонованої покрокової процедури обробки даних експресії генів та доцільність її використання в системах діагностики об'єктів на основі даних експресії генів.

Висновки. У цьому дослідженні було наведено гібридну модель для ідентифікації зразків даних експресії генів, яка поєднує в собі аналіз GO, алгоритм спектральної кластеризації, бікластерний аналіз та згорткову нейронну мережу (CNN). Ця модель продемонструвала значну перевагу над традиційними методами, покращивши точність та ефективність ідентифікації біологічно значущих закономірностей та зв'язків між генами. Завдяки цій моделі було виявлено нові біомаркери захворювань та закладено фундамент для розробки нових методів лікування і діагностики ракових захворювань.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Paul D. Thomas. The Gene Ontology and the meaning of biological function. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6438694/>.
2. Pietro H. Guzzi. Ontology in Bioinformatics. URL: <https://www.sciencedirect.com/science/article/abs/pii/B9780128096338204901>.
3. Yihong Gong, Wei Xu. Machine Learning for Multimedia Content Analysis. Pp. 37–70. URL: https://link.springer.com/chapter/10.1007/978-0-387-69942-4_3.
4. Babichev S., Yasinska-Damri L., Liakh I., Škvor J. Hybrid Inductive Model of Differentially and Co-Expressed Gene Expression Profile Extraction Based on the Joint Use of Clustering Technique and Convolutional Neural Network. *Applied Sciences* (Switzerland). 2022. Vol. 12 (22). Art. № 11795.
5. Babichev S., Yasinska-Damri L., Liakh I. A Hybrid Model of Cancer Diseases Diagnosis Based on Gene Expression Data with Joint Use of Data Mining Methods and Machine Learning Techniques (2023). *Applied Sciences* (Switzerland). Vol. 13 (10). Art. № 6022.
6. Yasinska-Damri L., Babichev S., Durnyak B., Goncharenko T. Application of Convolutional Neural Network for Gene Expression Data Classification. *Lecture Notes on Data Engineering and Communications Technologies*. 2023. Vol. 149. Pp. 3–24.
7. Babichev S., Durnyak B., Sharko O., Sharko A. Technique of metals strength properties diagnostics based on the complex use of fuzzy inference system and hybrid neural network. *Communications in Computer and Information Science*. 2020. Vol. 1158. Pp. 114–126.

REFERENCES

1. Paul, D. Thomas. The Gene Ontology and the meaning of biological function. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6438694/> (in English).
2. Pietro H. Guzzi. Ontology in Bioinformatics. Retrieved from <https://www.sciencedirect.com/science/article/abs/pii/B9780128096338204901> (in English).
3. Yihong, Gong, & Wei, Xu. Machine Learning for Multimedia Content Analysis, 37–70. Retrieved from https://link.springer.com/chapter/10.1007/978-0-387-69942-4_3 (in English).

4. Babichev, S., Yasinska-Damri, L., Liakh, I., & Škvor, J. (2022). Hybrid Inductive Model of Differentially and Co-Expressed Gene Expression Profile Extraction Based on the Joint Use of Clustering Technique and Convolutional Neural Network: Applied Sciences (Switzerland), 12 (22). Art. № 11795 (in English).
5. Babichev, S., Yasinska-Damri, L., & Liakh, I. A Hybrid Model of Cancer Diseases Diagnosis Based on Gene Expression Data with Joint Use of Data Mining Methods and Machine Learning Techniques (2023): Applied Sciences (Switzerland), 13 (10). Art. № 6022 (in English).
6. Yasinska-Damri, L., Babichev, S., Durnyak, B., & Goncharenko, T. (2023). Application of Convolutional Neural Network for Gene Expression Data Classification: Lecture Notes on Data Engineering and Communications Technologies, 149, 3–24 (in English).
7. Babichev, S., Durnyak, B., Sharko, O., & Sharko, A. (2020). Technique of metals strength properties diagnostics based on the complex use of fuzzy inference system and hybrid neural network: Communications in Computer and Information Science, 1158, 114–126 (in English).

doi: 10.32403/1998-6912-2024-1-68-136-144

HYBRID MODEL FOR IDENTIFYING GENE EXPRESSION DATA PATTERNS BASED ON CLUSTER-BICLUSTER ANALYSIS USING CONVOLUTIONAL NEURAL NETWORKS

I. M. Liakh¹, B. V. Durnyak²

¹*Uzhhorod National University,
3, Narodna Square, Uzhhorod, 88000, Ukraine
igor.lyah@uzhnu.edu.ua*

²*Ukrainian Academy of Printing,
19, Pid Holoskom St., Lviv, 79020, Ukraine*

In modern bioinformatics, gene expression data analysis plays a crucial role in studying biological processes and gene regulation mechanisms. The increasing volume of data generated by DNA microarrays and RNA sequencing makes their analysis a challenging task. Traditional data analysis methods, such as cluster analysis, may not provide sufficient accuracy and information for identifying biologically significant patterns.

New approaches that combine multiple data analysis methods offer the potential for deeper and more comprehensive analysis. One such method is cluster-bicluster analysis, which combines cluster analysis with bicluster analysis. Cluster analysis allows genes to be grouped based on the similarity of their expression profiles, while bicluster analysis identifies groups of genes that are co-expressed under different conditions.

In this study, a Bayesian optimization algorithm is used to determine the optimal hyperparameters of the convolutional neural network for the application of gene expression data generated as a result of cluster-bicluster analysis. The results of training and

validating the models formed by the neural network are used with 5-fold cross-validation.

The results of the comparative analysis of cancer type classification accuracy confirm the feasibility of applying the proposed step-by-step gene expression data processing procedure, which includes cluster-bicluster analysis. This highlights its potential for use in gene expression data-based diagnostic systems. Therefore, the effectiveness of applying a step-by-step procedure for clustering and biclustering gene expression data using gene ontology analysis is studied.

Keywords: *gene expression, clustering, biclustering, gene ontology, data analysis.*

Стаття надійшла до редакції 12.04.2024.

Received 12.04.2024.