УДК 004.048

ГІБРИДНА ІНДУКТИВНА МОДЕЛЬ КЛАСТЕРИЗАЦІЇ ПРОФІЛІВ ЕКСПРЕСІЙ ГЕНІВ НА ОСНОВІ АЛГОРИТМУ SOTA

Л. М. Ясінська-Дамрі[1], І. М. Лях[2], Б. В. Дурняк [1], С. А. Бабічев[3]

[1]Українська академія друкарства,
вул. Під Голоском, 19, Львів, 79020, Україна

[2]Ужгородський національний університет,
пл. Народна, 3, Ужгород, 88000, Україна

[3]Херсонський державний університет,
вул. Університетська, 27, Херсон, 73000, Україна

Подано результати дослідження щодо розробки гібридної індуктивної моделі кластеризації профілів експресій генів на основі комплексного застосування алгоритму кластеризації SOTA (Self-Organizing Tree Algorithm) та згорткової нейронної мережі. Модель зображена у вигляді структурної блок-схеми покрокової процедури реалізації процедури кластеризації у рамках індуктивної технології об'єктивної кластеризації на першому кроці і застосування згорткової нейронної мережі до даних експресій генів у сформованих кластерах на другому кроці. Формування проміжних кластеризацій здійснювалося на основі аналізу значень критерію балансу, який містив як компоненти внутрішні та зовнішні критерії якості кластеризації. Остаточний вибір оптимальної кластеризації відповідав максимальному значенню точності класифікації об'єктів під час застосування згорткової нейронної мережі.

Ключові слова: *алгоритм кластеризації SOTA, згорткова нейронна мережа, дані експресій генів, кластеризація профілів експресій генів, індуктивна технологія об'єктивної кластеризації, класифікація даних, точність класифікації.*

Постановка проблеми. Розробка гібридних моделей кластеризації профілів експресій генів на основі комплексного застосування методів інтелектуального аналізу даних та машинного навчання є одним із актуальних напрямів сучасної біоінформатики. Особливістю даних експресій генів є велика кількість генів, що визначають стан біологічного об'єкта, що досліджується, велика розмірність профілю експресій генів, яка визначається кількістю об'єктів, що досліджувалися, при формуванні експериментальних даних і відсутність інформації щодо кількості кластерів, на які необхідно розділити експериментальні дані. Традиційні алгоритми кластеризації у більшості випадків орієнтовані на низькорозмірні (об'єкти характеризуються невеликою кількістю атрибутів) дані. Крім того, сучасним алгоритмам кластеризації властива похибка відтворюваності, сутність якої полягає у

тому, що результати кластеризації, що отримані на одних даних, не повторюються при застосуванні інших аналогічних даних. Вирішення цієї проблеми можливо шляхом застосування індуктивної технології об'єктивної кластеризації (ІТОК), ідея якої наведена у працях [1] і отримала подальший розвиток у публікаціях [2, 3]. Однак потрібно зазначити, що застосування ІТОК дає змогу сформувати підмножину проміжних кластеризацій, що відповідають локальним максимумам критерію балансу, що містить як компоненти внутрішні та зовнішні критерії якості кластеризації даних. Остаточне рішення щодо формування оптимальної кластерної структури зазвичай ухвалюється, враховуючи мету поставленої задачі, що має великий відсоток суб'єктивізму.

Тому виникає проблема підвищення об'єктивності формування кластерної структури профілів експресій генів завдяки комплексному застосуванню індуктивних методів об'єктивної кластеризації та методів класифікації об'єктів на основі даних у сформованих кластерах.

Аналіз останніх досліджень та публікації. На сьогодні розроблено та реалізовано велику кількість алгоритмів кластерного аналізу. Їхні ключові відмінності один від одного полягають у виборі способу формування кластерної структури, завданні метрики оцінювання ступеня близькості, наявності або відсутності самоорганізації тощо. У працях [4, 5] наведено структурну схему покрокової процедури формування кластерної структури та структурну блок-схему способів формування кластерної структури і найбільш розповсюджених на сьогодні алгоритмів кластеризації даних. Вибір відповідного алгоритму визначається типом даних, що досліджуються та бажаним способом формування кластерної структури.

Дослідження, що присвячені вирішенню проблеми класифікації великих даних шляхом застосування згорткових нейронних мереж (ЗНМ), наведені у працях [6–9]. У цих публікаціях автори досліджують різні структури ЗНМ та особливості їх застосування для класифікації різних типів зображень, часових рядів тощо. У публікаціях [10, 11] автори запропонували модель класифікації об'єктів на основі даних експресій генів на основі ЗНМ. Аналіз отриманих результатів свідчить про високу ефективність моделей класифікації об'єктів, що містять як атрибути великі обсяги даних, на основі ЗНМ. Але потрібно зазначити, що, незважаючи на певні досягнення у цій предметній галузі, проблема створення оптимальної моделі кластеризації профілів експресій генів на сьогодні не має однозначного вирішення.

Мета статті — розробка гібридної індуктивної моделі кластеризації профілів експресій генів на основі комплексного застосування алгоритму кластеризації SOTA та згорткової нейронної мережі.

Виклад основного матеріалу дослідження. Самоорганізуючий алгоритм кластеризації SOTA (Self-Organizing Tree Algorithm) [13] є логічним продовженням самоорганізуючих нейронних мереж на основі карт Кохонена [14]. Результатом застосування алгоритму SOTA є бінарне топологічне дерево, що формується відповідно до алгоритму вирощування просторової клітинної структури Fritzke [15]. Застосування алгоритму Fritzke призводить до збільшення кількості вузлів дерева в області більшої щільності об'єктів (профілів експресій генів), при цьому концентрація

об'єктів в області з меншою щільністю не змінюється. Процес формування клітинної структури дерева, що лежить в основі алгоритму кластеризації SOTA, проілюстровано на рис. 1. Як вхідні дані розглядається множина профілів експресій генів. Як можна бачити, у початковому стані структура системи має вигляд бінарного дерева, що складається з трьох вузлів: дві клітини об'єднані через зовнішній материнський вузол (рис. 1 (а)). У початковому стані кожна клітина та вузол характеризується вектором, довжина якого дорівнює довжині профілю експресій гена (вектор даних, що подається на вхід мережі), а значення лежать в інтервалі варіювання відповідних значень експресій генів вхідних даних. Практична реалізація алгоритму кластеризації SOTA передбачає наявність таких етапів:

Етап I. Ініціалізація. Реалізація цього етапу передбачає присвоєння клітинам та вузлу векторів ваг, довжина яких дорівнює довжині векторів профілів експресій генів, що подаються на вхід системи, а значення лежать в інтервалі варіювання експресій відповідних генів. На цьому етапі задаються також параметри для корекції ваг клітин і вузлів у процесі подавання на вхід мережі профілів експресій генів з огляду на умову: $\alpha_w > \alpha_p > \alpha_s$, де α_w, α_p та α_s — параметри для корекції ваг клітини переможця (winner), кореневого (parent) вузла та сусідньої клітини (sister) відповідно. Автори алгоритму SOTA [12] запропонували таке співвідношення параметрів для корекції ваг: $\alpha_w = 2\alpha_p$, $\alpha_p = 5\alpha_s$. У процесі моделювання було досліджено різні комбінації цих параметрів. Як результат, був прийнятий авторський варіант співвідношень цих параметрів як найбільш оптимальний. На етапі ініціалізації також задається граничне значення відносного коефіцієнта варіації ваг усіх зовнішніх клітин E для двох послідовних ітерацій, яке визначає одну із основних умов зупинки алгоритму.

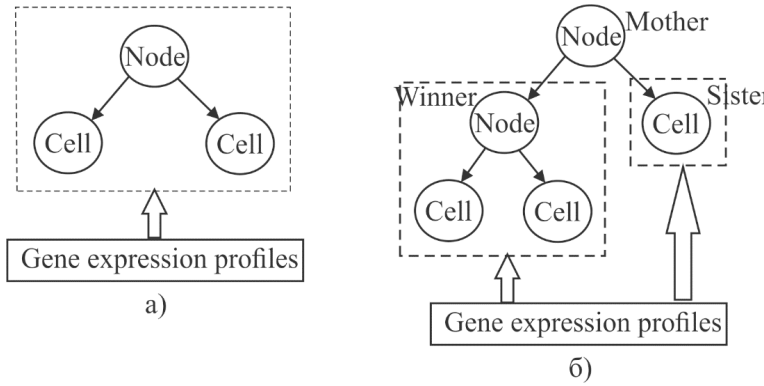

Рис. 1. Ілюстрація процесу формування дерева (формування клітинної структури) у процесі роботи алгоритму кластеризації SOTA:
а) — початковий стан системи;
б) — стан системи після одного циклу

Етап II. Адаптація. На цьому етапі на вхід усіх зовнішніх клітин поступово подаються профілі експресій генів, що складають вхідні дані. Розраховується ступінь

близькості між профілями експресій генів і вектором ваг відповідної клітини. На цьому кроці виділяється клітина переможець (winner) за мінімальною відстанню між профілем експресій гена та вектором ваг і, відповідно до принципу «переможець забирає усе», профіль експресій гена поміщається у клітину-переможець. На наступному кроці підлаштовуються ваги клітини переможця, сусідньої клітини та кореневого вузла відповідно до формули:

$$W_i(\tau+1) = W_i(\tau) + \eta \cdot \left(P_j - W_i(\tau)\right), \quad (1)$$

де $W_i(\tau)$ й $W_i(\tau+1)$ — ваги i-ї клітини на ітераціях τ й $\tau + 1$ відповідно; P_j — j-й профіль експресій гена, що подається на вхід мережі; η — параметр, що визначає швидкість підлаштування ваг відповідних клітин і розраховується за формулою:

$$\eta_\tau = \alpha_i \cdot \frac{1-\tau}{N} \cdot (1 - bt), \quad (2)$$

де η_τ — значення η-параметра на ітерації τ; $\alpha_з$ — параметр для корекції ваг i-ї клітини; N — максимальна кількість профілів експресій генів, що подаються на вхід мережі; t — максимальна кількість операцій в одному циклі; b — коефіцієнт, що визначає швидкість зміни параметра η.

Підлаштування ваг клітин та кореневого вузла здійснюється відповідно до такого принципу: якщо сусідня з клітиною переможцем клітина має нащадки, то підлаштовуються тільки ваги клітини переможця. У протилежному випадку підлаштовуються ваги клітини переможця, сусідньої клітини та кореневого вузла.

Етап ІІІ. Формування структури дерева (мережі). На першому кроці цього етапу розраховується коефіцієнт варіації профілів експресій генів у кожній клітині:

$$R_i = \frac{\sum_{j=1}^{K} d\left(P_j, W_i\right)}{K}, \quad (3)$$

де W_i — вектор ваг у i-й клітині; P_j — j-й профіль експресій генів; K — кількість профілів експресій генів у i-й клітині.

На другому кроці розраховується значення сумарного коефіцієнта варіації як сума коефіцієнтів варіації усіх зовнішніх клітин:

$$\varepsilon = \sum_{i=1}^{p} R_i, \quad (4)$$

Критерієм оцінки збіжності алгоритму (або зупинки) є відносна зміна сумарного коефіцієнта варіації на двох послідовних ітераціях:

$$E = \frac{\left|\varepsilon_\tau - \varepsilon_{\tau-1}\right|}{\varepsilon_{\tau-1}} \le E_{гр}, \quad (5)$$

де $E_{гр}$ — граничне значення відносної зміни коефіцієнта варіації.

Якщо виконується умова (5), алгоритм зупиняється з фіксацією структури дерева і відповідних кластерів, що містять профілі експресій генів у зовнішніх клітинах. У протилежному випадку здійснюється зростання дерева. Клітина, що має

найбільше значення коефіцієнта варіації, ділиться на дві частини і стає вузлом (рис. 1 (б)), при цьому усім клітинам і вузлу присвоюються ідентичні значення ваг. Далі повторюються етапи II і III. Умовою зупинки алгоритму є або виконання умови (5), або досягнення максимальної кількості ітерацій, що задається на етапі ініціалізації початкових параметрів алгоритму.

На основі аналізу вищенаведеного процесу можна зробити висновок, що результат робити алгоритму (структура сформованого дерева) визначається граничним значенням відносної зміни коефіцієнта варіації або кількістю ітерацій та значенням параметра для корекції ваг сусідньої клітини (параметри для корекції інших клітин визначаються з огляду на умову: $\alpha_w = 2\alpha_p$, $\alpha_p = 5\alpha_s$). У рамках дослідження граничне значення відносної зміни коефіцієнта варіації було прийнято за 0. За цієї умови зупинка алгоритму визначалась або повторюванням характеру розподілу профілів експресій генів у кластери на двох послідовних ітераціях, або досягненням граничної кількості ітерацій. Визначення оптимального значення параметра для корекції ваг сусідньої клітини α_s здійснювалося із застосуванням індуктивних методів аналізу складних систем (індуктивної технології об'єктивної кластеризації). Структурна блок-схема гібридної індуктивної моделі формування підмножин диференційно експресованих та взаємно корельованих профілів експресій генів на основі алгоритму кластеризації SOTA зображена на рис. 2. Реалізація алгоритму передбачає такі етапи:

Етап I. Формування даних та ініціалізація алгоритму кластеризації SOTA.

1.1. Формування матриці профілів експресій генів, де рядки — об'єкти, що досліджуються, а стовпці — гени, значення експресії яких визначають стан відповідного об'єкта.

1.2. Видалення неінформативних генів за критеріями: абсолютне значення експресії гена, дисперсія та ентропія Шеннона профілів експресій генів.

1.3. Формування метрики оцінки ступеня близькості профілів експресій генів і відповідних кластерів. У дослідженнях використовувалася гібридна модифікована метрика на основі комплексного застосування гібридної метрики максимізації взаємної інформації та критерію узгодженості Пірсона.

1.4. Формування двох еквівалентних підмножин профілів експресій генів.

1.5. Формування критеріїв якості кластеризації для оцінки якості кластерної структури. Як внутрішні критерії для оцінки характеру групування профілів експресій генів у кластерах були використані WB-індекс та PBM критерій, зовнішні критерії розраховувалися як нормалізована різниця відповідних внутрішніх критеріїв, а критерій балансу розраховувався на основі функції бажаності Харрінгтона.

1.6. Ініціалізація алгоритму. Завдання граничного значення відносної зміни коефіцієнта варіації $E_{гр} = 0$ та інтервалу і кроку зміни параметра для корекції ваг сусідньої клітини α_s. Відповідно до рекомендації авторів алгоритму SOTA, параметри для корекції ваг інших клітин визначаються з огляду на умову: $\alpha_w = 2\alpha_p$, $\alpha_p = 5\alpha_s$.

Етап II. Кластеризація профілів експресій генів та визначення оптимальних параметрів алгоритму SOTA.

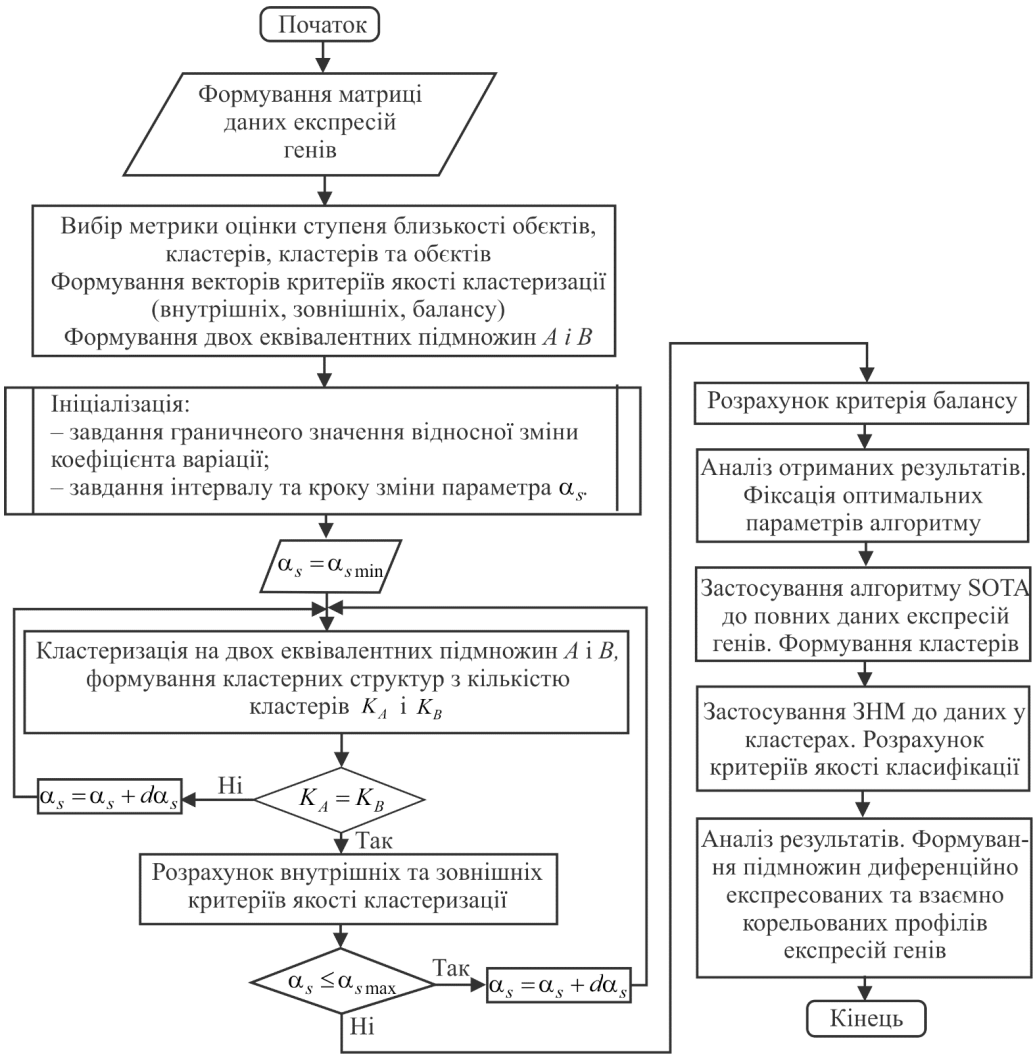

Рис. 2. Алгоритм гібридної індуктивної моделі формування підмножин диференційно експресованих та взаємно корельованих профілів експресій генів на основі алгоритму кластеризації SOTA

2.1. Ініціалізація першого значення параметра для корекції ваг сусідньої клітини $\alpha_s = \alpha_{s\min}$.

2.2. Застосування алгоритму кластеризації SOTA до профілів експресій генів, що містяться в еквівалентних підмножинах A і B. Формування кластерів.

2.3. Якщо кількість кластерів в еквівалентних підмножинах однакова, розрахунок внутрішніх та зовнішніх критеріїв якості кластеризації та збільшення значення параметра для корекції ваг сусідньої клітини: $\alpha_s = \alpha_s + d\alpha_s$. У протилежному випадку збільшення параметра для корекції ваг сусідньої клітини без розрахунку критеріїв.

2.4. Якщо $\alpha_s \le \alpha_{s\max}$, перехід на крок 2.2 цієї процедури. У протилежному випадку — розрахунок критерію балансу.

2.5. Аналіз отриманих результатів. Фіксація значень параметра для корекції ваг сусідньої клітини, що відповідають максимальним значенням критерію балансу.

2.6. Застосування алгоритму кластеризації SOTA з визначеними оптимальними значеннями параметра для корекції ваг сусідньої клітини до повної сформованої на кроці 1.2 множини профілів експресій генів. Формування кластерної структури у кожному випадку.

Етап ІІІ. Застосування згорткової нейронної мережі до даних експресій генів у сформованих кластерах.

3.1. Передобробка даних експресій генів у сформованих кластерах шляхом додавання профілів з нульовою експресією для отримання необхідної кількості профілів для коректного застосування згорткової нейронної мережі.

3.2. Застосування ЗНМ до даних експресій генів у виділених кластерах. Розрахунок критеріїв якості класифікації даних.

3.3. Аналіз отриманих результатів. Формування підмножин диференційно експресованих та взаємно корельованих підмножин профілів експресій генів, які відповідають максимальним значенням як критерію балансу, так і критерію якості класифікації об’єктів, що досліджуються.

Результати моделювання щодо визначення оптимального значення параметра для корекції ваг сусідньої клітини α_s наведені на рис. 3.

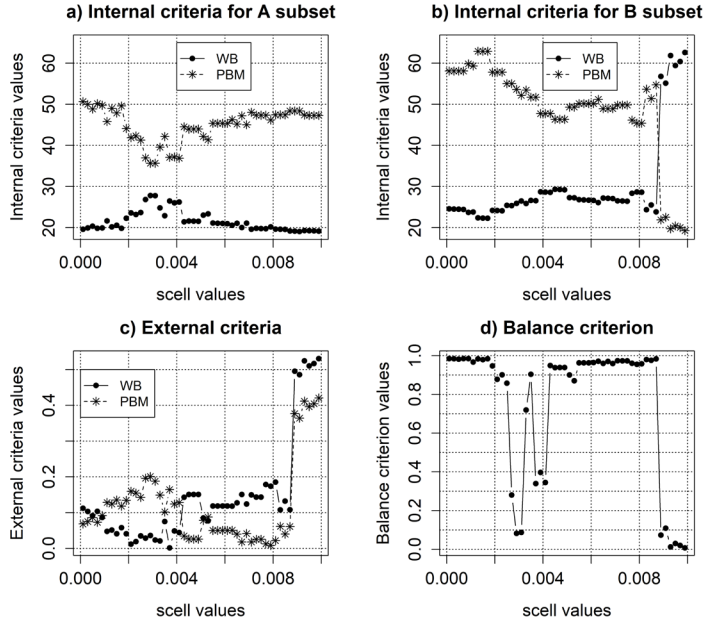

Рис. 3. Результати моделювання щодо визначення оптимальних параметрів алгоритму кластеризації SOTA на основі застосування індуктивних методів аналізу складних систем

Значення параметра α_s при цьому змінювалося в інтервалі від 0.0001 до 0.01 з кроком 0.0002. Моделювання здійснювалося у програмному середовищі R із застосуванням функцій пакета *clValid* [17]. Функція *sota()* передбачає можливість реалізації алгоритму SOTA із застосуванням евклідової та кореляційної метрик. Як експериментальні дані, застосовувалися 156 ДНК-мікрочипів пацієнтів, що досліджувалися на рак легенів, з яких 65 пацієнтів були ідентифіковані як здорові, а у 91 пацієнта була ідентифікована ракова пухлина. У початковому стані кожний мікрочип містив 54675 генів. На етапі передобробки кількість генів була скорочена до 10000 шляхом застосування процесу редукції на основі аналізу значень статистичних критеріїв та ентропії Шеннона. Враховуючи високу розмірність профілів експресій генів, формування кластерної структури при роботі алгоритму здійснювалося із застосування кореляційної метрики, а розрахунок внутрішніх критеріїв якості кластеризації проводився із застосуванням модифікованої гібридної метрики на основі ентропійних критеріїв та критерію узгодженості Пірсона. Відповідно до першої частини алгоритму, структурну блок-схему, якого зображено на рис. 2, на першому етапі підмножина профілів експресій генів (1000 профілів) розділялася на дві еквівалентні підмножини A і B, що містили однакову кількість попарно близьких об'єктів. Далі на кожному кроці реалізації алгоритму SOTA на еквівалентних підмножинах профілів експресій генів розраховувалися внутрішні критерії якості кластеризації. На рис. 3 a, b зображені діаграми залежності значень цих критеріїв від параметра алгоритму SOTA α_s, водночас за цими критеріями оптимальна кластеризація відповідає мінімальному значенню WB-індексу і максимальному значенню PBM критерію. Як можна бачити на отриманих діаграмах, значення цих критеріїв, розрахованих на підмножинах A і B, деякою мірою суперечать один одному. На рис. 3 c зображені діаграми зовнішніх критеріїв, розрахованих із застосуванням відповідних внутрішніх критеріїв. Аналіз цих діаграм також не дає змоги однозначно визначити оптимальне значення параметра α_s (за мінімальним значенням). На рис. 3 d зображено діаграму залежності критерію балансу від значення параметра алгоритму, аналіз якої дає змогу виділити інтервали варіювання значень параметра α_s, що відповідають максимумам критерію з одного боку і стійкій кластеризації з іншого боку. Більш ретельний аналіз отриманих результатів показав, що максимального значення (0.985) критерій балансу досягає на четвертому кроці реалізації цієї процедури, при цьому аналіз значень внутрішніх критеріїв також підтверджує той факт, що на цьому кроці якість групування профілів експресій генів у кластери є досить висока. Потрібно зазначити, що у всіх випадках (на кожному ітераційному кроці) дані розділялися на два кластери. Алгоритм зупинявся при повторюванні конфігурації розподілу профілів у кластери на двох послідовних ітераціях. Отже, значення $\alpha_s = 0.0007$ було обрано як оптимальне для подальших досліджень.

На другому етапі реалізації алгоритму, наведеному на рис. 2, алгоритм кластеризації SOTA з визначеними параметрами застосовувався до повного набору профілів експресій генів (10000), у результаті чого отримано два кластера: перший містив 6020 профілів, а другий — 3980. Для коректного застосування згорткової

нейронної мережі перший кластер доповнювався профілями з нульовою експресією для отримання загальної кількості 6050, при цьому застосовувався фільтр (50×121) на першому згортковому шарі та (25×242) на другому. Другий кластер доповнювався до загальної кількості 4000 профілів, при цьому фільтри мали розміри (50×80) на першому шарі та (25×160) на другому. Результати моделювання щодо застосування згорткових нейронних мереж до даних експресій генів, що містяться у сформованих кластерах, зображені на рис. 4 і рис. 5.

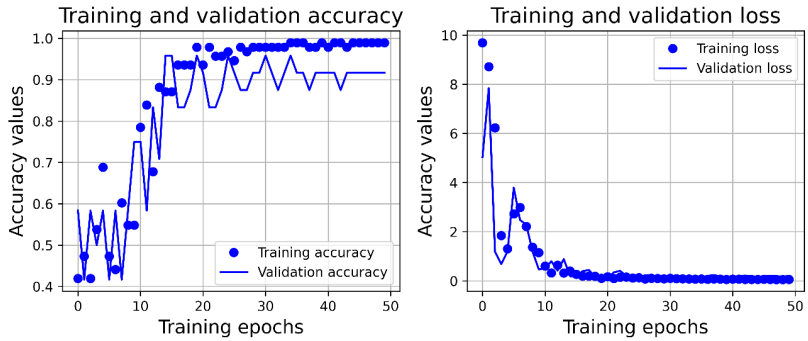

Рис. 4. Діаграми зміни точності класифікації та значення функції втрат у процесі навчання одновимірної двошарової нейронної мережі на даних у першому кластері

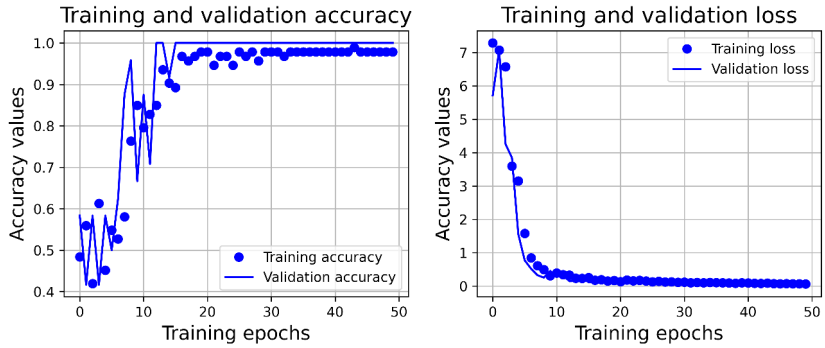

Рис. 5. Діаграми зміни точності класифікації та значення функції втрат у процесі навчання одновимірної двошарової нейронної мережі на даних у другому кластері

Точність класифікації об’єктів відповідних тестових підмножин даних становила при цьому 95 % для першого кластера і 97 % для об’єктів другого кластера, при цьому із об’єктів було коректно ідентифіковано 37 і 38 у першому та у другому випадках відповідно.

На наступному етапі до профілів експресій генів у відповідних кластерах покроково застосовувався алгоритм SOTA з подальшою класифікацією об’єктів, що

містяться у відповідних кластерах шляхом застосування одновимірної двошарової ЗНМ. Результати моделювання наведені у табл.

Таблиця

Результати моделювання щодо покрокового застосування гібридної моделі кластеризації профілів експресій генів на основі алгоритму SOTA

Етап	Кількість генів	Розмір фільтра	F-індекс		Точність класифікації, %	Втрати
			Кластер 1	Кластер 2		
1	6020	50×121 25×242	0.93	0.96	95	0.152
	3980	50×80 25×160	**0.96**	**0.98**	**97**	**0.146**
2	**3011**	50×61 50×122	**0.96**	**0.98**	**97**	**0.197**
	3009	50×61 50×122	**0.96**	**0.98**	**97**	**0.165**
	1934	50×39 25×78	0.92	0.96	95	0.227
	2046	50×41 20×82	0.89	0.94	92	0.205
3	**1639**	40×41 20×82	**0.96**	**0.98**	**97**	**0.156**
	1372	40×35 20×70	0.92	0.96	95	0.199
	1519	40×38 20×76	**0.96**	**0.98**	**97**	**0.123**
	1490	30×50 15×100	0.89	0.94	92	0.193

На рис. 6 зображено дендрограму розподілу початкової кількості профілів експресій генів на кластери з відповідною точністю класифікації об’єктів. Аналіз отриманих результатів дає змогу зробити висновок, що комплексне застосування алгоритму кластеризації SOTA і згорткової нейронної мережі дозволяє виділити кластери профілів експресій генів, які дають змогу з високою точністю ідентифікувати об’єкти, що містять як атрибути значення експресій генів, які локалізовані у відповідних кластерах.

Як можна бачити, на першому етапі 10000 профілів експресій генів були розділені на два кластера, при цьому для другого (меншого кластера) досягалася висока точність класифікації на тестовій підмножині даних, із 39 об’єктів 38 ідентифіковано коректно. Для більшого кластера точність класифікації була нижчою за різними параметрами, при цьому два об’єкти із 39 були ідентифіковані помилково.

Але потрібно зазначити, що подальше розділення профілів експресій генів, що містилися у меншому кластері, на підмножини, що становили 1934 та 2046 профілів експресій генів, погіршило точність класифікації, тобто розділення меншого кластера на підкластери у цьому випадку не є доцільним. Дійти іншого висновку можна у випадку розділення більшого кластера (6020 профілів експресій генів) на підкластери. Розділення цього кластера на дві підмножини (3011 і 3009 профілів експресій генів) підвищує точність класифікації об'єктів, що містять як атрибути значення експресій генів цих кластерів. Подальше їхнє розділення на більш дрібні підмножини дає змогу на кожній гілці виділити підмножину більш інформативних за критеріями класифікації профілів експресій генів.

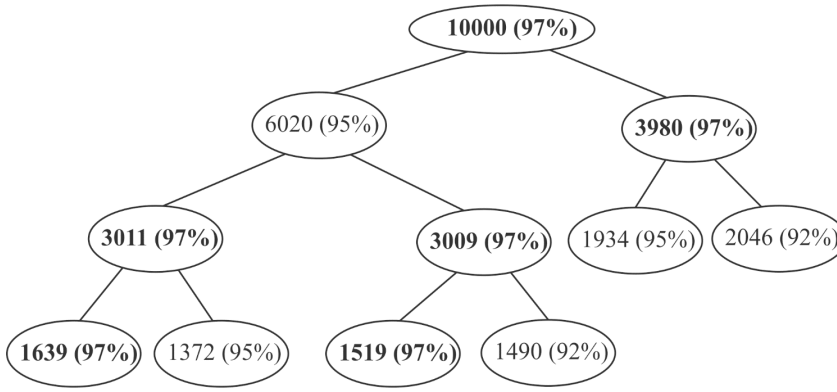

Рис. 6. Дендрограма розподілу профілів експресій генів у кластери з точністю класифікації об'єктів, що містять як атрибути значення експресій генів у відповідних кластерах

Висновки. Наведено результати досліджень щодо створення гібридної індуктивної моделі кластеризації профілів експресій генів на основі комплексного застосування алгоритму кластеризації SOTA та одновимірної згорткової нейронної мережі для класифікації об'єктів на основі даних експресій генів у виділених кластерах. Як експериментальні дані, застосовувалися 156 ДНК-мікрочипів пацієнтів, що досліджувалися на рак легенів, з яких 65 пацієнтів були ідентифіковані як здорові, а у 91 пацієнта була ідентифікована ракова пухлина. У початковому стані кожний мікрочип містив 54675 генів. На першому етапі кількість генів була скорочена до 10000 шляхом застосування процесу редукції на основі аналізу значень статистичних критеріїв та ентропії Шеннона. На другому етапі до профілів експресій генів застосовувався алгоритм кластеризації SOTA, реалізований в рамках індуктивної технології об'єктивної кластеризації, що дало змогу сформувати проміжні кластеризації, які відповідають максимальному значенню критерію балансу. На останньому етапі до даних експресій генів у виділених кластерах застосовувалася згорткова нейронна мережі з розрахунком критеріїв класифікації об'єктів, що містять як атрибути дані експресій генів у виділених кластерах. Оптимальні кластери при цьому відповідають максимальним значенням критеріїв якості класифікації об'єктів.

Проведенні дослідження створюють умови для підвищення об'єктивності ідентифікації об'єктів шляхом розпаралелювання процесу обробки інформації, виділення найбільш інформативних профілів експресій генів за критеріями якості класифікації і прийняття компромісного рішення шляхом аналізу результатів класифікації об'єктів, що містять як атрибути тільки профілі експресій найбільш інформативних генів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Madala H. R., Ivakhnenko A. G. Inductive Learning Algorithms for Complex Systems Modeling. CRC Press, 1994. 365 p.
2. Babichev S., Taif M. A., Lytvynenko V., Korobchinskyi M. Objective clustering inductive technology of gene expression sequences features. Communications in Computer and Information Science. In the book «Beyond Databases, Architectures and Structures», edited by S. Kozelski and D. Mrozek, 2017. Pp. 359–372.
3. Babichev S., Gozhyj A., Kornelyuk A., Lytvynenko V. Objective clustering inductive technology of gene expression profiles based on SOTA clustering algorithm. *Biopolymers and Cell.* Kiev : National Academy of Science Ukraine, 2017. Vol. 33 (5). Pp. 379–392.
4. Soni N., Ganatra A. Categorization of Several Clustering Algorithms from Different Perspective: A Review. *International Journal of Advanced Research in Computer Science and Software Engineering*. 2012. Vol. 2 (8). Pp. 63–68.
5. Xu R., Wunsch D.C. Survey of Clustering Algorithms. *IEEE Transactions on neural Networks.* 2005. Vol. 16. Pp. 645–678.
6. Chuang Y.-H., Huang S.-H., Hung T.-M. et al. Convolutional neural network for human cancer types prediction by integrating protein interaction networks and omics data. *Scientific Reports*. 2021. Vol. 11 (1), art. no. 20691.
7. Busaleh M., Hussain M., Aboalsamh H. A. Breast mass classification using diverse contextual information and convolutional neural network. *Biosensors*. 2021. Vol. 11 (11), art. no. 419.
8. Li J., Sun W., Feng X. et al. A dense connection encoding–decoding convolutional neural network structure for semantic segmentation of thymoma. *Neurocomputing*. 2021. Vol. 451. Pp. 1–11.
9. Cao X., Pan J.-S., Wang Z. et al. Application of generated mask method based on Mask R-CNN in classification and detection of melanoma. *Computer Methods and Programs in Biomedicine*. 2021. Vol. 207, art. no. 106174.
10. Mostavi M, Chiu Y-C., Huang Y., Chen Y. Convolutional neural network models for cancer type prediction based on gene expression. *BMC Medical Genomics*. 2020. Vol. 13 (5), art. no. 44.
11. Ramires R., Chiu Y., Horerra A. et al. Classification of cancer types using graph convolutional neural networks. *Frontiers in Physics*. 2020. Vol. 8, art. no. 203.
12. Dorazo J., Carazo J. M. Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. *Journal of Molecular Evolution*. 1997. Vol. 44 (2). Pp. 226–234.
13. Kohonen T. Self-Organizing Maps (Third Extended Edition), New York, 2001.
14. Fritzke B. Growing Cell Structures. A Self-Organizing Network for Unsupervised and Supervised Learning. *Neural Networks*. 1994. Vol. 7 (9). Pp. 1441–1420.

15. Brock G., Pihur V., Datta S., Datta S. clValid: An R Package for Cluster Validation. *Journal of Statistical Software*. 2008. Vol. 25 (4). Pp. 1–22.

REFERENCES

1. Madala, H. R., & Ivakhnenko, A. G. (1994). Inductive Learning Algorithms for Complex Systems Modeling. CRC Press (in English).
2. Babichev, S., Taif, M. A., Lytvynenko, V., & Korobchinskyi, M. (2017). Objective clustering inductive technology of gene expression sequences features. Communications in Computer and Information Science. In the book «Beyond Databases, Architectures and Structures», edited by S. Kozelski and D. Mrozek, 359–372 (in English).
3. Babichev, S., Gozhyj, A., Kornelyuk, A., & Lytvynenko, V. (2017). Objective clustering inductive technology of gene expression profiles based on SOTA clustering algorithm: Biopolymers and Cell. Kiev : National Academy of Science Ukraine, 33 (5), 379–392 (in English).
4. Soni, N., & Ganatra, A. (2012). Categorization of Several Clustering Algorithms from Different Perspective: A Review: International Journal of Advanced Research in Computer Science and Software Engineering, 2 (8), 63–68 (in English).
5. Xu R., Wunsch D.C. (2005). Survey of Clustering Algorithms. IEEE Transactions on neural Networks, 16, 645–678 (in English).
6. Chuang, Y.-H., Huang, S.-H., & Hung, T.-M. et al. (2021). Convolutional neural network for human cancer types prediction by integrating protein interaction networks and omics data: *Scientific Reports,* 11 (1), art. no. 20691 (in English).
7. Busaleh, M., Hussain, M., & Aboalsamh, H. A. (2021). Breast mass classification using diverse contextual information and convolutional neural network: *Biosensors*, 11 (11), art. no. 419 (in English).
8. Li, J., Sun, W., & Feng, X. et al. (2021). A dense connection encoding–decoding convolutional neural network structure for semantic segmentation of thymoma: *Neurocomputing*, 451, 1–11 (in English).
9. Cao, X., Pan, J.-S., & Wang, Z. et al. (2021). Application of generated mask method based on Mask R-CNN in classification and detection of melanoma: *Computer Methods and Programs in Biomedicine*, 207, art. no. 106174 (in English).
10. Mostavi, M, Chiu, Y-C., Huang, Y., & Chen, Y. (2020). Convolutional neural network models for cancer type prediction based on gene expression: BMC Medical Genomics, 13 (5), art. no. 44 (in English).
11. Ramires, R., Chiu, Y., & Horerra, A. et al. (2020). Classification of cancer types using graph convolutional neural networks: Frontiers in Physics, 8, art. no. 203 (in English).
12. Dorazo, J., & Carazo, J. M. (1997). Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree: Journal of Molecular Evolution, 44 (2), 226–234 (in English).
13. Kohonen, T. (2001). Self-Organizing Maps (Third Extended Edition), New York (in English).
14. Fritzke, B. (1994). Growing Cell Structures. A Self-Organizing Network for Unsupervised and Supervised Learning: Neural Networks, 7 (9), 1441–1420 (in English).
15. Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). clValid: An R Package for Cluster Validation: Journal of Statistical Software, 25 (4), 1–22 (in English).

doi: 10.32403/1998-6912-2022-1-64-48-62

HYBRID INDUCTIVE MODEL OF GENE EXPRESSION PROFILES CLUSTERING BASED ON SOTA ALGORITHM

L. M. Yasinska-Damri[1], I. M. Liakh[2], B. V. Durnyak[1], S. A. Babichev[3]

1Ukrainian Academy of Printing,
19, Pid Holoskom St., Lviv, 79020, Ukraine
Lm.yasinska@gmail.com, durnyak@uad.lviv.ua

2Uzhhorod National University,
3, Narodna Square, Uzhhorod, 88000, Ukraine
ihor.lyah@uzhnu.edu.ua

3Kherson State University,
27, Universytetska St., Kherson, 73000, Ukraine
sbabichev@ksu.ks.ua

The results of the research regarding the development of a hybrid inductive model of gene expression profiles clustering based on the joint application of the SOTA clustering algorithm (Self-Organizing Tree Algorithm) and the convolutional neural network are presented in the paper. The model is presented as a structural block chart of a stepwise procedure for implementing the clustering algorithm within the framework of objective clustering inductive technology in the first step and the application of a convolutional neural network to gene expression data in the formed clusters in the second step. As an experimental data, the authors used gene expression data of patients studied on lung cancer. 156 patients were studied in total, of which 65 were identified as healthy and 91 patients were diagnosed with cancer. Each of the studied objects contained 54,675 genes. In the first stage, 10,000 of the most informative genes in terms of statistical criteria and Shannon entropy were allocated. The formation of intermediate clustering was carried out on the basis of the analysis of the balance criterion values, which contained, as the components, both the internal and external clustering quality criteria. The final choice of the optimal clustering corresponded to the maximum value of the objects classification accuracy when using a convolutional neural network.

The performed research creates the conditions for improving the objectivity of the object identification by parallelizing the information processing, carefully selecting the most informative gene expression profiles according to the classification quality criteria and making a compromise decision by analyzing the results of the classification of the object containing only the most informative gene expression profiles.

A further perspective of the author's research is the practical implementation of the proposed technique using various current gene expression data.

Keywords: *SOTA clustering algorithm, convolutional neural network, gene expression data, clustering of gene expression profiles, objective clustering inductive technology, data classification, classification accuracy.*

Стаття надійшла до редакції 18.03.2022.
Received 18.03.2022.