

Л. М. Ясінська-Дамрі, С. А. Бабічев,  
Б. В. Дурняк, М. Ф. Ясінський

**МОДЕЛІ ТА МЕТОДИ ОБРОБКИ ДАНИХ ЕКСПРЕСІЇ ГЕНІВ  
У СИСТЕМАХ ДІАГНОСТИКИ СКЛАДНИХ ОБ'ЄКТІВ**

Львів — 2025

УДК 004.048; 004.94

Затверджено вченою радою Херсонського державного університету  
(протокол № 16 від 26.05.2025 р.)

Рецензенти:

**Лях І. М.** – д-р техн. наук, професор  
(Ужгородський державний університет, м. Ужгород)

**Гожий О. П.** – д-р техн. наук, професор  
(Чорноморський національний університет ім. Петра Могили, м. Миколаїв)

**Сеньківський В. М.** – д-р техн. наук, професор  
(Національний університет «Львівська політехніка», м. Львів)

Л. М. Ясінська-Дармі, С. А. Бабичев, Б. В. Дурняк, М. Ф. Ясінський  
Моделі та методи бробки даних експресії генів у системах діагностики складних  
об'єктів. Монографія. ДП «Всеукраїнське спеціалізоване видавництво «Світ»,  
2025. 249 с.

ISBN 978-966-441-821-5

Монографія присвячена теоретичним і практичним аспектам розробки моделей та методів обробки даних експресії генів для задач діагностики складних об'єктів. Представлено концепцію інформаційної технології, що поєднує інтелектуальні методи аналізу даних, кластеризацію, методи машинного навчання та аналіз генних онтологій. Розглянуто сучасні підходи до попередньої обробки даних експресії, виділення інформативних ознак та побудови діагностичних моделей. Окрему увагу приділено створенню гібридних моделей, що інтегрують статистичні та інтелектуальні методи для підвищення ефективності діагностики. Практичну частину присвячено застосуванню запропонованих методів для аналізу даних у біоінформатиці та медичних дослідженнях. Монографія буде корисною для науковців, аспірантів, магістрантів та практиків, що працюють у галузях біоінформатики, медичної інформатики та машинного навчання.

УДК 004.048; 004.94

ISBN 978-966-441-821-5

© Ясінська-Дармі Л. М., Бабічев С. А., Дурняк Б. В., Ясінський М. Ф., 2025  
© ЛА Піраміда, 2025

## ЗМІСТ

|                                                                                                                                                                     |    |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Вступ.....                                                                                                                                                          | 7  |
| Розділ 1. Аналіз сучасного стану методів обробки даних експресії генів .....                                                                                        | 9  |
| 1.1. Біологічні засади передачі та трансформації генної інформації<br>у біологічному організмі.....                                                                 | 9  |
| 1.2. Загальна процедура формування та обробки даних для створення<br>діагностичних моделей на основі даних експресії генів.....                                     | 10 |
| 1.2.1. Технології формування експериментальних даних для визначення<br>ступеня активності (експресії) генів, які визначають стан об'єкта, що<br>досліджується ..... | 11 |
| 1.2.2. Методи формування матриці експресії генів та видалення генів<br>з низьким значенням експресії для всіх об'єктів, що досліджуються.....                       | 13 |
| 1.2.2.1. Експерименти із застосуванням ДНК-мікрочіпів .....                                                                                                         | 13 |
| 1.2.2.2. Методи обробки результатів експерименту на основі<br>застосування технології секвенування молекул мРНК .....                                               | 19 |
| 1.2.3. Методи статистичного аналізу та редукції профілів експресії генів.....                                                                                       | 21 |
| 1.2.3.1. Огляд сучасних методів бікластерного аналізу.....                                                                                                          | 22 |
| 1.2.3.2. Огляд сучасних методів та алгоритмів кластеризації даних<br>експресії генів.....                                                                           | 25 |
| 1.3. Аналіз сучасних методів та інформаційних технологій редукції профілів<br>експресії генів.....                                                                  | 34 |
| 1.4. Огляд сучасних методів машинного навчання для обробки даних<br>експресії генів.....                                                                            | 42 |
| 1.5. Постановка мети та завдань дослідження .....                                                                                                                   | 48 |
| Розділ 2. Моделі, методи та алгоритми формування підмножин інформативних<br>профілів експресії генів .....                                                          | 51 |
| 2.1. Порівняльний аналіз метрик оцінки ступеня близькості профілів<br>експресії генів.....                                                                          | 51 |
| 2.1.1. Метод максимізації взаємної інформації .....                                                                                                                 | 52 |
| 2.1.1.1. Оцінка ефективності методу максимізації взаємної інформації<br>при застосуванні різних методів розрахунку ентропії Шеннона .....                           | 55 |
| 2.1.1.2. Модифікований метод максимізації взаємної інформації.....                                                                                                  | 63 |
| 2.1.2. Метод оцінки ступеня близькості профілів експресії генів на основі<br>критерія узгодженості Пірсона ( $\chi^2$ -статистика).....                             | 67 |

|                                                                                                                                                                          |     |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 2.1.3. Метод оцінки ступеня близькості профілів експресії генів на основі кореляційної відстані .....                                                                    | 68  |
| 2.1.4. Порівняльний аналіз ефективності метрик оцінки ступеня близькості профілів експресії генів .....                                                                  | 69  |
| 2.1.5. Метод розрахунку комплексного критерія для формування підмножини найбільш інформативних профілів експресії генів .....                                            | 73  |
| 2.2. Теоретичні дослідження щодо оцінки ефективності застосування гібридної метрики близькості профілів експресії генів у моделях кластеризації даних .....              | 75  |
| 2.2.1. Аналіз ефективності метрик оцінки ступеня близькості профілів експресії генів при застосуванні однокластерної структури даних .....                               | 76  |
| 2.2.2. Розрахунок внутрішнього та зовнішнього критеріїв якості групування профілів експресії генів на основі застосування мультикластерної структури .....               | 80  |
| 2.2.2.1. Щільнісний алгоритм формування кластерної структури OPTICS .....                                                                                                | 81  |
| 2.2.2.2. Формування кластерної структури на основі принципів індуктивної технології об'єктивної кластеризації .....                                                      | 83  |
| Розділ 3. Теоретичні та практичні засади індуктивної гібридної моделі кластеризації профілів експресії генів.....                                                        | 91  |
| 3.1. Концепція та особливості застосування індуктивних методів аналізу складних процесів при створенні індуктивних моделей кластеризації даних.....                      | 91  |
| 3.1.1. Архітектура та концептуальний опис покрокової процедури формування кластерної структури на основі принципів індуктивної технології об'єктивної кластеризації..... | 95  |
| 3.2. Індуктивна гібридна модель кластеризації даних на основі щільнісних алгоритмів.....                                                                                 | 98  |
| 3.2.1. Формування векторів внутрішніх та зовнішніх критеріїв якості кластеризації залежно від типу експериментальних даних .....                                         | 102 |
| 3.2.2. Формування функцій розрахунку критерію балансу на основі виділених внутрішніх критеріїв якості кластеризації, враховуючи тип експериментальних даних .....        | 120 |
| 3.2.2.1. Алгоритм формування функції розрахунку критерію балансу для даних «Compound» .....                                                                              | 120 |
| 3.2.2.2. Алгоритм формування функції розрахунку критерію балансу для даних «Jain» .....                                                                                  | 121 |
| 3.2.2.3. Алгоритм формування функції розрахунку критерію балансу для даних «Seeds».....                                                                                  | 121 |
| 3.2.2.4. Алгоритм формування функції розрахунку критерію балансу для даних профілів експресії генів пацієнтів, що досліджувалися на рак легенів.....                     | 122 |
| 3.2.3. Результати моделювання щодо практичної реалізації гібридної індуктивної моделі щільнісного алгоритму кластеризації DBSCAN.....                                    | 123 |
| 3.2.3.1. Практична реалізація гібридної індуктивної моделі щільнісного алгоритму кластеризації DBSCAN на даних «Compound» .....                                          | 123 |

|                                                                                                                                                                                                           |     |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 3.2.3.2. Практична реалізація гібридної індуктивної моделі щільнісного алгоритму кластеризації DBSCAN на даних «Jain».....                                                                                | 125 |
| 3.2.3.3. Практична реалізація гібридної індуктивної моделі щільнісного алгоритму кластеризації DBSCAN на даних «Seeds».....                                                                               | 128 |
| 3.2.3.4. Практична реалізація гібридної індуктивної моделі щільнісного алгоритму кластеризації DBSCAN на даних профілів експресії генів.....                                                              | 131 |
| 3.2.4. Результати моделювання щодо практичної реалізації гібридної індуктивної моделі щільнісного алгоритму кластеризації OPTICS .....                                                                    | 133 |
| 3.2.4.1. Практична реалізація гібридної індуктивної моделі щільнісного алгоритму кластеризації OPTICS на даних «Compound».....                                                                            | 134 |
| 3.2.4.2. Практична реалізація гібридної індуктивної моделі щільнісного алгоритму кластеризації OPTICS на даних «Jain» .....                                                                               | 136 |
| 3.2.4.3. Практична реалізація гібридної індуктивної моделі щільнісного алгоритму кластеризації OPTICS на даних «seeds» .....                                                                              | 138 |
| 3.2.4.4. Практична реалізація гібридної індуктивної моделі щільнісного алгоритму кластеризації OPTICS на даних експресії генів.....                                                                       | 140 |
| Розділ 4. Гібридні моделі формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів.....                                                                           | 144 |
| 4.1. Аналіз ефективності згорткових нейронних мереж для класифікації об'єктів на основі даних експресії генів .....                                                                                       | 145 |
| 4.1.1. Архітектура, будова та модель згорткової нейронної мережі.....                                                                                                                                     | 145 |
| 4.1.2. Моделювання щодо застосування згорткової нейронної мережі у задачах класифікації даних експресії генів.....                                                                                        | 150 |
| 4.1.2.1. Експериментальні дані профілів експресії генів.....                                                                                                                                              | 150 |
| 4.1.2.2. Модель одновимірної згорткової нейронної мережі .....                                                                                                                                            | 151 |
| 4.1.2.3. Модель одновимірної двошарової згорткової нейронної мережі .....                                                                                                                                 | 154 |
| 4.1.2.4. Модель двовимірної згорткової нейронної мережі.....                                                                                                                                              | 156 |
| 4.1.2.5. Модель двовимірної тришарової згорткової нейронної мережі .....                                                                                                                                  | 158 |
| 4.1.3. Оцінка робастності згорткової нейронної мережі при різних рівнях шумової компоненти.....                                                                                                           | 159 |
| 4.2. Концептуальний опис гібридної моделі формування диференційно експресованих та взаємно корельованих профілів експресії генів на основі комплексного застосування алгоритмів кластеризації та ЗНМ..... | 161 |
| 4.2.1. Гібридна індуктивна модель кластеризації профілів експресії генів на основі алгоритму кластеризації SOTA .....                                                                                     | 164 |
| 4.2.2. Практична реалізація гібридної індуктивної моделі кластеризації профілів експресії генів на основі алгоритму кластеризації SOTA .....                                                              | 169 |
| 4.2.3. Практична реалізація гібридної індуктивної моделі кластеризації профілів експресії генів на основі алгоритму спектральної кластеризації....                                                        | 174 |

|                                                                                                                                                                                                                                   |     |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 4.2.3.1. Покрокова процедура реалізації алгоритму спектральної кластеризації.....                                                                                                                                                 | 176 |
| 4.2.3.2. Гібридна індуктивна модель кластеризації профілів експресії генів на основі алгоритму спектральної кластеризації .....                                                                                                   | 178 |
| 4.2.3.3. Результати моделювання щодо практичної реалізації гібридної індуктивної моделі кластеризації профілів експресії генів на основі алгоритму спектральної кластеризації.....                                                | 180 |
| 4.3. Концептуальна схема інформаційної технології діагностики складних об'єктів на основі комплексного застосування алгоритму спектральної кластеризації, ЗНМ та методу альтернативного голосування.....                          | 182 |
| Розділ 5. Застосування розроблених моделей, методів та алгоритмів для створення системи діагностики ракових захворювань .....                                                                                                     | 185 |
| 5.1. Формування та передобробка експериментальних даних .....                                                                                                                                                                     | 186 |
| 5.1.1. Формування та аналіз даних експресії генів із застосуванням функцій та модулів пакетів TCGAblolinks і Bioconductor .....                                                                                                   | 186 |
| 5.1.2. Видалення неінформативних генів за статистичними критеріями та ентропією Шеннона.....                                                                                                                                      | 189 |
| 5.2. Застосування індуктивної моделі об'єктивної кластеризації на основі алгоритму спектральної кластеризації для формування підмножин диференційно експресованих та взаємно корельованих підмножин профілів експресії генів..... | 195 |
| 5.3. Застосування згорткової нейронної мережі до даних експресії генів у сформованих кластерах.....                                                                                                                               | 198 |
| 5.4. Застосування системи альтернативного голосування для прийняття остаточного рішення щодо стану пацієнтів, що досліджуються .....                                                                                              | 208 |
| 5.4.1. Практична реалізація моделі бінарної класифікації стану здоров'я пацієнтів на основі мультикластерної структури даних експресії генів із застосуванням методу альтернативного голосування.....                             | 211 |
| 5.4.2. Практична реалізація мультикласової моделі ідентифікації стану здоров'я пацієнтів на основі мультикластерної структури даних експресії генів із застосуванням методу альтернативного голосування .....                     | 214 |
| 5.4.2.1. Результати діагностики стану пацієнтів на основі застосування алгоритму Random Forest.....                                                                                                                               | 216 |
| 5.4.2.2. Результати діагностики стану пацієнтів на основі застосування згорткової нейронної мережі .....                                                                                                                          | 219 |
| Висновки .....                                                                                                                                                                                                                    | 223 |
| Conclusions .....                                                                                                                                                                                                                 | 227 |
| Список використаних джерел .....                                                                                                                                                                                                  | 227 |

## Вступ

Сучасні штучні системи обробки інформації, що використовуються у різноманітних галузях інтелектуального аналізу даних та машинного навчання, здебільшого ґрунтуються на аналогії функціонування відповідних процесів у біологічних організмах. До таких процесів слід віднести функціонування генної мережі, імунні процеси, функціонування нейронних мереж тощо. Особливістю таких систем є великий рівень складності, здатність навчатися, розпаралелювання процесу обробки інформації, великий рівень захисту, здатність розпізнавати та приймати адекватні рішення. У цьому контексті, розробка сучасних штучних моделей обробки великих даних можлива із застосуванням системного підходу, який передбачає комплексне застосування знань та методів з різних практичних галузей, а саме: молекулярна біологія, математика, інформатика, фізика, хімія. Такий підхід створює умови для підвищення об'єктивності обробки великих даних у реальному часі за рахунок застосування ансамблів методів, гібридних моделей та ефективних критеріїв якості оцінки результатів на відповідному етапі реалізації процесу обробки інформації.

Однією з актуальних проблем сучасної біоінформатики є обробка даних експресії генів, отриманих шляхом реалізації ДНК-мікрочіпових експериментів або більш сучасним методом секвенування молекул РНК. До особливостей експериментальних даних слід віднести велику кількість генів, що визначають стан відповідного біологічного організму. Як відомо, кількість експресованих (активних) генів, що складають геном людини, приблизно дорівнює 25000. При цьому кількість кожного типу гена варіюється в інтервалі від 0 до сотень тисяч. Створення систем медичної діагностики або моделей реконструкції генних регуляторних мереж на повних даних не є ефективним, оскільки процес інтерпретації отриманих результатів у цьому випадку є проблематичним. На цей час існують і активно застосовуються для обробки великих даних моделі глибинного навчання, а саме: глибинні нейронні мережі, згорткові нейронні мережі, штучні глибинні мережі, які дозволяють отримати задовільну точність діагностики на великих даних, але при цьому виникає проблема часу навчання, чутливості мережі та верифікації отриманих моделей на інших, аналогічних даних. Вирішення вищезазначеної проблеми можливе шляхом розробки нових та удосконалення існуючих методів, моделей та алгоритмів обробки великих даних. Реалізація цього процесу передбачає формування якісних експериментальних даних шляхом застосування адекватних методів нормалізації значень експресії генів та редукції низько-експресованих генів на першому етапі. На другому етапі виникає проблема формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів шляхом застосування гібридних моделей кластеризації високовимірних даних. Третім етапом є розробка моделі діагностики складних захворювань на основі сформованих кластерів профілів експресії генів із застосуванням методів глибинного навчання з відповідною верифікацією отриманих результатів.

Слід зазначити, що незважаючи на певні досягнення у цій предметній галузі, проблема обробки даних експресії генів з метою створення системи медичної діагностики, що дозволяє з високою точністю ідентифікувати стан пацієнта на ранньому етапі виникнення хвороби не має однозначного вирішення. Сучасна інформаційна технологія

обробки даних експресії генів для діагностики складних захворювань (різні типу раку, хвороби Паркінсона та Альцгеймера тощо) або реконструкції генної регуляторної мережі має забезпечувати:

- обґрунтований вибір методів формування та передобробки експериментальних даних для формування масиву експресії генів для об'єктів, що досліджуються;
- застосування ефективних метрик оцінки ступеня близькості між профілями експресії генів на основі оцінки взаємної інформації між відповідними профілями;
- обґрунтоване формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів із застосуванням відповідних внутрішніх та зовнішніх критеріїв якості кластеризації даних;
- створення гібридних моделей діагностики стану об'єкта, що досліджується, із застосуванням індуктивних моделей об'єктивної кластеризації та глибинних нейронних мереж;
- обґрунтований вибір методів валідації моделей діагностики складних захворювань із застосуванням відповідних критеріїв якості оцінки ефективності моделі на тестовій підмножині даних експресії генів.

Аналіз сучасних літературних джерел у галузі створення систем діагностики складних об'єктів на основі даних експресії генів дозволяє зробити висновок, що сьогодні не існує цілісної та закінченої теорії обробки даних експресії генів для формування адекватних підмножин диференційно експресованих та взаємно корельованих профілів експресії генів, що дозволяють з високою точністю ідентифікувати стан пацієнта, хоча дослідження у цій предметній галузі створюють передумови для її створення і подальшого удосконалення.

Іншим актуальним завданням при створенні інформаційної технології обробки даних експресії генів є розробка гібридної моделі класифікації об'єктів на основі комплексного застосування як алгоритмів кластеризації, що реалізовані в рамках індуктивної технології об'єктивної кластеризації, так і глибинних нейронних мереж, при цьому для підвищення об'єктивності прийняття рішення щодо стану об'єкта, що досліджується, на етапі кластеризації профілів експресії генів здійснюється розпаралелювання потоку обробки інформації.

Автори висловлюють щире подяку рецензентам книги: д-ру техн. наук, професору Ляху І. М., д-ру техн. наук, професору Гожому О. П., д-ру техн. наук, професору Сеньківському В. М. за поради та цінні вказівки при підготовці видання.

## Розділ 1. Аналіз сучасного стану методів обробки даних експресії генів

У цьому розділі зроблено аналіз сучасного стану методів обробки даних експресії генів для створення систем діагностики складних захворювань або реконструкції генних регуляторних мереж (ГРМ) на основі експериментальних даних, отриманих шляхом застосування ДНК мікрочіпів або методів секвенування молекул РНК. У обох випадках експериментальні дані подані у вигляді високовимірної матриці експресії генів, де рядки представляють гени (тип генів), а стовпці – біологічні об'єкти, що досліджуються, або навпаки. Як було зазначено у вступі, складність вирішення проблеми створення системи діагностики захворювань або реконструкції ГРМ на основі повного набору даних визначається великою кількістю генів. Як показує аналіз експериментальних даних [32], геном людини містить приблизно 25000 активних генів. Стільки ж генів знаходяться у неактивному стані (нульова експресія). Які гени є активними у даний час залежить від особливостей біологічного організму і його сучасного стану. Тому передобробка даних експресії генів для вибору генів, що дозволяють з максимальною роздільною здатністю розрізнити класи об'єктів, що досліджуються, є необхідним етапом для створення ефективної системи діагностики складних захворювань на основі даних експресії генів. У розділі проаналізовано основні етапи передобробки даних експресії генів для вибору найбільш інформативних, враховуючи мету задачі, що розв'язується, визначені недоліки існуючих методів для удосконалення цього процесу. Розглянуто сучасні підходи з ідентифікації об'єктів на основі даних експресії генів, сформульовано концептуальні засади інформаційної технології обробки профілів експресії генів для створення системи діагностики захворювань на основі даних експресії генів.

### 1.1. Біологічні засади передачі та трансформації генної інформації у біологічному організмі

У 1958 р. Ф. Крик сформулював правило, що визначає напрям передачі генетичної інформації у живих організмів від нуклеїнових кислот до білків [33]. Сформульована гіпотеза у 1970 р. згідно з даними, що були накопичені на той час, і трансформована у центральну догму молекулярної біології [34], яка постулює, що генетична інформація зберігається у ДНК і передається до інформаційної РНК (іРНК), а далі від молекул іРНК до рибосом, які у свою чергу синтезують тип білка, що відповідає даному гену. Таким чином можна зробити висновок, що кількість генів пропорційна кількості білка, синтезованого цим геном. Рис. 1.1 ілюструє цей процес [35].



Рис. 1.1. Центральна догма молекулярної біології

Під геном тут розуміємо послідовність нуклеотидів, які містять інформацію щодо структури білка, що синтезується даним геном. Експресія генів визначає ступінь його активності [36, 37]. Значення експресії є пропорційним кількості генів, що відповідають даному типу білка. Якщо ген не експресований (значення експресії 0), він може бути видалений з даних на даному етапі дослідження без втрати інформації про стан досліджуваного об'єкта.

Генна мережа біологічного організму являє собою сукупність молекулярних регуляторів, що взаємодіють між собою та з іншими речовинами у біологічній клітині та визначають рівень експресії генів молекул матричної рибонуклеїнової кислоти, які у свою чергу синтезують відповідний тип білка [38]. Генна регуляторна мережа являє собою мережу, яка реконструйована із застосуванням даних експресії генів, отриманих проведенням відповідних експериментів [39]. Вочевидь, що якісно реконструйована ГРМ має відображати характер взаємодії генів та/або білків у біологічному об'єкті, що досліджується. З іншого боку, значення експресії відповідних генів для клітин здорового і хворого організмів повинні суттєво відрізнятися. До того ж, експресії генів для клітин організмів, що відповідають різним хворобам, повинні також відрізнятися між собою. Таким чином, на етапі передобробки даних необхідно виділити взаємно-експресовані гени, профілі яких найбільш корельовані між собою. Під профілем експресії гена тут розуміємо вектор значень експресії гена, що визначені для різних об'єктів. Довжина вектора профілю у цьому випадку дорівнює кількості об'єктів, що досліджувалися в процесі проведення експерименту.

## 1.2. Загальна процедура формування та обробки даних для створення діагностичних моделей на основі даних експресії генів

Покрокову процедуру формування даних та обробки результатів експерименту для реконструкції та моделювання ГРМ та/або створення моделей діагностики складних захворювань зображено на рис. 1.2 [1, 4].

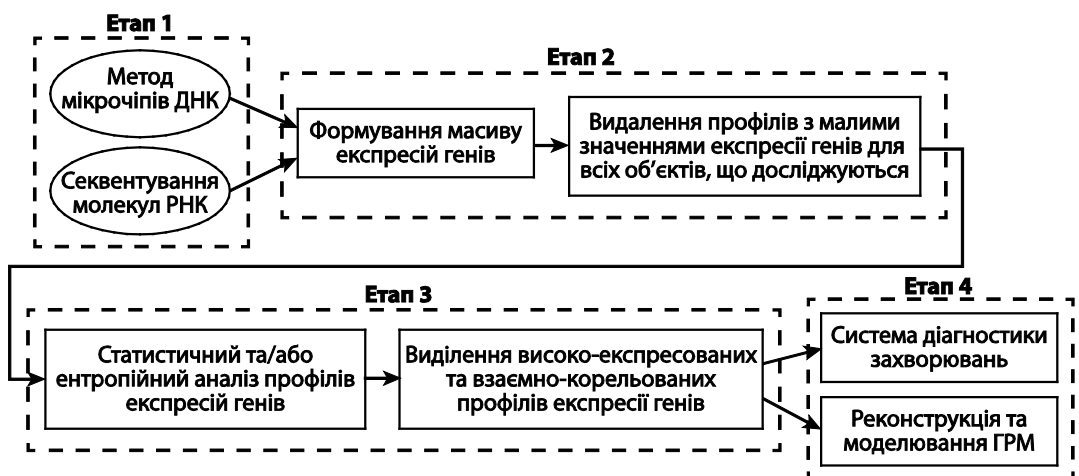


Рис. 1.2. Блок-схема покрокової процедури формування та обробки експериментальних даних експресії генів

Реалізація цього процесу передбачає чотири етапи: безпосереднє проведення експерименту; формування масиву експресії генів та видалення неекспресованих та генів з малим значенням експресії для всіх об'єктів, що досліджуються; статистичний та ентропійний аналізи отриманих даних для виділення взаємно корельованих генів, які дозволяють розрізнити об'єкти, що досліджуються з високою роздільною здатністю; реконструкція та моделювання ГРМ або створення системи діагностики хвороби.

### 1.2.1. Технології формування експериментальних даних для визначення ступеня активності (експресії) генів, які визначають стан об'єкта, що досліджується

Сьогодні для одночасного визначення експресії великої кількості генів актуальними є два типи експерименту: на основі ДНК-мікрочіпів [40] і метод секвенування молекул РНК [41, 42]. Схема покрокового процесу проведення експерименту із застосуванням ДНК-мікрочіпу зображена на рис. 1.3 [43]. Загалом заготовка ДНК-мікрочіпу являє собою тверду плоску поверхню, на якій упорядковано (у заздалегідь відомому порядку) розташовані олігонуклеотидні послідовності. Поверхня мікрочіпу при цьому ділиться на сотні тисяч ділянок, у кожній з яких містяться проби з відомою послідовністю нуклеотидів (кожна послідовність відповідає конкретному гену), що здатні зв'язати певну ділянку молекули мРНК. Практична реалізація цього процесу передбачає наявність двох типів проб: контрольна проба, що містить гени з відомою концентрацією (експресією) і експериментальна проба, яка містить гени, експресію яких необхідно оцінити. На відповідні проби заздалегідь наносяться флуоресцентні мітки: червоного (СУ5) та зеленого (СУ3) кольорів на контрольні та експериментальні проби, відповідно.

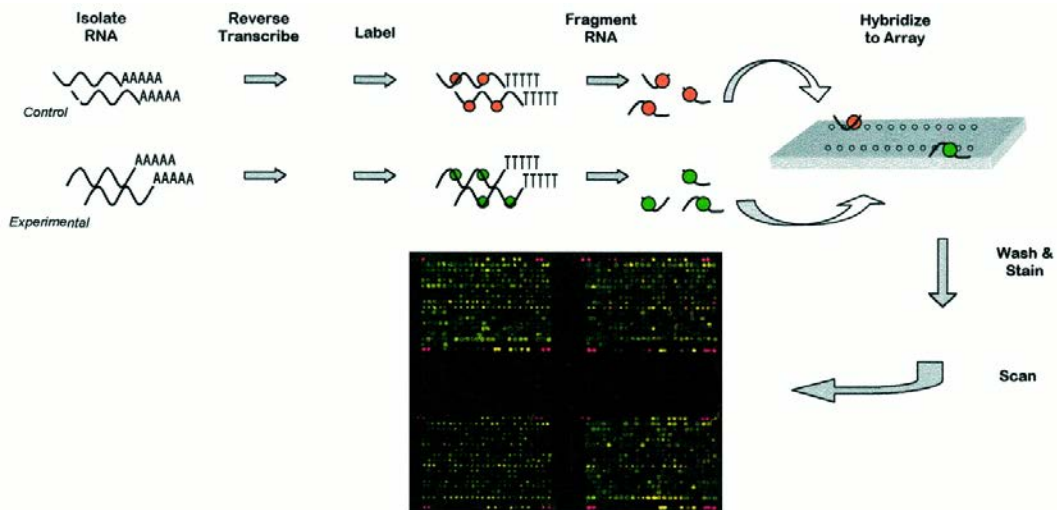


Рис. 1.3. Схема покрокового процесу проведення експерименту із застосуванням ДНК-мікрочіпу

У процесі проведення експерименту молекули мРНК, що містяться у контрольній та експериментальній пробах, вступають у реакцію з відповідними послідовностями на мікрочіпі, відповідно до принципу комплементарності, тобто здійснюється процес гібридизації [44]. Інтенсивність світла флуоресцентної речовини у відповідній точці мікрочіпу визначається відносною кількістю проб контрольної та експериментальної речовин, що прореагували з цією послідовністю нуклеотидів. Після процесу лазерного сканування формується матриця інтенсивностей світла:

$$I = \log_2 \frac{I(Cy5)}{I(Cy3)}. \quad (1.1)$$

До недоліків методу ДНК-мікрочіпів слід віднести високу похибку проведення експерименту, яка визначається як біологічними, так і технологічними факторами [44]. До біологічних факторів слід віднести крос-гібридизацію – явище, викликане тим, що в процесі гібридизації деякі послідовності молекул мРНК у контрольній або експериментальній пробах можуть бути частково комплементарними відповідними послідовностям нуклеотидів на мікрочіпі. Дані молекули не відповідають послідовностям на мікрочіпі, але присутніх зав'язків може бути достатньо для вступу у реакцію. Як наслідок, виникає «біологічний шум». Для зменшення цього шуму на мікрочіп, окрім послідовностей, що повністю комплементарні відповідним послідовностям молекул мРНК (РМ – perfectmatch) наносять послідовності, у яких середній нуклеотид замінюється на комплементарний до нього (ММ – mismatch). На етапі трансформації значень інтенсивностей світла у значення експресії відповідних генів здійснюється РМ корекція, що дозволяє частково компенсувати вплив явища крос-гібридизації. До технологічних факторів слід віднести похибку сканування та обробки значень інтенсивностей світла. Як наслідок, виникає фоновий оптичний шум.

Трансформація інтенсивностей світла у відповідні значення експресії генів здійснюється у чотири етапи: фонові корекція, нормалізація, РМ корекція та сумарізація [45]. Кожний етап передбачає можливість застосування різних методів. Тому вибір оптимальної комбінації методів також впливає на якість матриці експресії генів.

До переваг методу ДНК-мікрочіпів слід віднести низьку вартість, порівняно з методом секвенування молекул мРНК. Крім того, його застосування дозволяє одночасно оцінити експресію десятків тисяч генів. Один мікрочіп містить послідовності нуклеотидів, які відповідають усім генам геному біологічного організму, що досліджується.

Альтернативою експерименту із застосуванням ДНК-мікрочіпів є метод секвенування молекул мРНК [41, 42]. Структурна схема покрокового процесу реалізації даного методу зображена на рис. 1.4 [46].

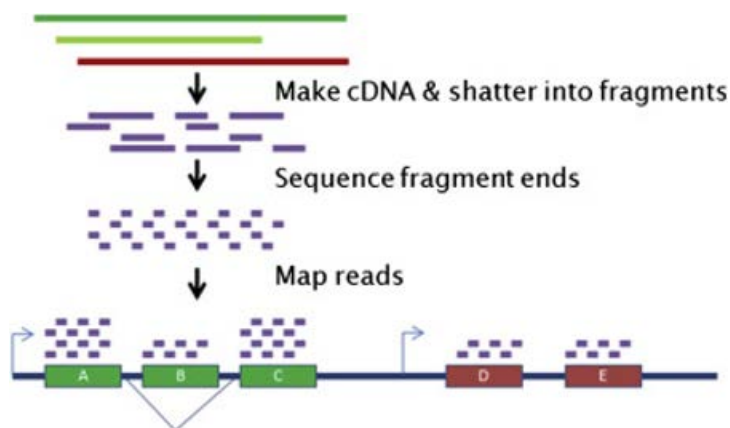


Рис. 1.4. Схема покрокового процесу реалізації методу секвенування молекул мРНК

Практична реалізація цього процесу дозволяє з високою точністю встановити послідовність нуклеотидів у молекулах генів, що досліджуються. Як наслідок, отримується матриця кількостей різних типів генів для досліджуваних об'єктів. Перевагою методу секвенування молекул мРНК, порівняно з методом ДНК-мікрочіпів, є суттєво більш висока точність оцінки значень експресії генів. До недоліків слід віднести високу вартість проведення експерименту.

Однак слід відзначити, що результатом реалізації як одного, так і другого типу експерименту є високовимірна матриця експресії генів, де рядки являють собою гени, а стовпці – об'єкти, що досліджуються. Кількість генів, включаючи неекспресовані, досягає при цьому 50000.

### 1.2.2. Методи формування матриці експресії генів та видалення генів з низьким значенням експресії для всіх об'єктів, що досліджуються

Як було зазначено раніше, результатом експериментів із застосуванням ДНК-мікрочіпів або методу секвенування молекул мРНК є високовимірна матриця інтенсивності світла у першому випадку і матриця абсолютної кількості генів для об'єктів, що досліджуються, у другому випадку. Наступним кроком реалізації наведеної на рис. 1.2 покрокової процедури обробки експериментальних даних є формування масиву експресії генів з подальшим видаленням неекспресованих або генів з малими значеннями експресії для всіх об'єктів, що досліджуються.

#### 1.2.2.1. Експерименти із застосуванням ДНК-мікрочіпів

Вирішенню проблеми трансформації масиву інтенсивностей світла у масив експресії генів сьогодні присвячено велику кількість робіт [45, 47–50]. Реалізація цього процесу передбачає чотири етапи: фонові корекція, нормалізація, РМ корекція та сумаризація. У таблиці 1.1 наведено методи, що можуть бути застосовані для реалізації кожного з етапів. Методи фонові корекції РМ-ММ проб (Idealismismatch)

та MAS-5 запропоновано компанією Affymetrix [51]. Перший метод передбачає комплексне застосування як РМ (Perfectmatch), так і ММ (Mismatch) проб.

Таблиця 1.1

*Етапи та відповідні їм методи трансформації матриці інтенсивностей світла ДНК-мікрочіпів у матрицю експресії генів*

|        | Етапи              |                   |                 |                        |
|--------|--------------------|-------------------|-----------------|------------------------|
|        | Фонові корекції    | Нормалізація      | РМ корекція     | Сумаризація            |
| Методи | Idealmismatch [60] | Quantiles [59,61] | Subtractum [60] | TukeyBi-weight [60,70] |
|        | RMA [61]           | Constant [59,60]  | MAS [60]        | Avgdiff [60]           |
|        | MAS-5 [60]         | Loess [64]        | pmonly [60]     | Medianpolish [61]      |
|        | DFCM [62]          | Contrast [65]     | –               | Li-Wong [66]           |
|        | –                  | Qspline [67]      | –               | Playerout [60]         |
|        | –                  | Invariantset [66] | –               | –                      |

Основна ідея застосування даного методу передбачає розрахунок зваженого середнього значення Тьюк і WMT по множині логарифмів відношень значень РМ та ММ для відповідних пар проб (що відповідають одному гену). На наступному кроці розраховується фонові поправки відповідно до формули:

$$\Delta I_{ik} = \begin{cases} MM_{ik}, \text{ якщо } MM_{ik} \leq PM_{ik} \\ \frac{PM_{ik}}{2^{WMT_k}}, \text{ якщо } MM_{ik} \geq PM_{ik}, WMT_k > \tau_c \\ \frac{PM_{ik}}{2^{\frac{\tau_c}{1 + (\tau_c - WMT_{ik})/\tau_s}}}, \text{ якщо } MM_{ik} \geq PM_{ik}, WMT_k \leq \tau_c, \end{cases} \quad (1.2)$$

де  $i$  та  $k$  – це номер проби та гену, відповідно;  $\tau_c$  та  $\tau_s$  – стали відмінності та масштабованості, які визначаються в процесі обробки даних емпіричним шляхом. Остаточне значення інтенсивності світла, що відповідає гену, визначається як різниця вихідної інтенсивності та фонові поправки (1.2).

Метод фонові корекції MAS-5 (Microarray Affymetrix Suite 5.0) [51] передбачає розділення поверхні мікрочіпу на однакові прямокутні клітини. Далі для кожної клітини розраховується границя фонові складові якнайменше 2% значення від інтенсивностей світла у кожній клітині ( $b_k$ ). Значення фонові поправки для кожної проби на мікрочіпі розраховується за формулою:

$$\Delta I_{ik} = \frac{\sum_{i=1}^k \omega_{ik} b_k}{\sum_{i=1}^k \omega_{ik}}, \quad (1.3)$$

де  $\omega_{ik}$  – ваговий коефіцієнт проби  $i$  у клітині  $k$ :

$$\omega_{ik} = \frac{1}{d_{ik}^2 + s_0}. \quad (1.4)$$

$d_{ik}$  у формулі (1.4) розраховується як евклідова відстань від проби  $i$  до центру  $k$ -ї клітини,  $s_0$  – згладжуючий коефіцієнт, значення якого залежить від кількості клітин

на мікрочіпі і визначається емпіричним шляхом у процесі моделювання. Скориго- ване значення інтенсивності світла проби  $i$  визначається так:

$$I_{ik} = \max(\max(I_{0ik}, 0.5) - b_k, 0.5), \quad (1.5)$$

де  $I_{0ik}$  – вихідне значення інтенсивності світла  $i$ -ї проби у  $k$ -й клітині.

RMA (Robust Multichip Average) метод фонові корекції [52] передбачає застосування тільки РМ проб. Передбачається, що інтенсивність світла відповідної проби є комбінацією корисної та шумової компонент, при цьому інтенсивність світла корисної складової розподілена за експоненціальним законом, а шуму – за нормальним. Фонова корекція у цьому випадку здійснюється видаленням шумової складової.

При застосуванні DFCM (Distribution Free Convolution Model) методу фонові корекції [53] інтенсивність світла також являє собою комбінацію інтенсивностей світла корисної компоненти та шумової складової, але при цьому не робиться ніяких припущень щодо закону розподілу даних компонент. Застосування цього методу передбачає наявність РМ та ММ проб.

Необхідність етапу нормалізації визначається тим, що умови проведення експерименту для різних мікрочіпів (об'єктів, що досліджуються) можуть розрізнятися як за часовими, так і технологічними параметрами. На етапі нормалізації дані мікро- чіпу приводяться до однакового розподілу, що дозволяє зменшити похибку при порівняльному аналізі профілів експресії генів, значення яких визначені при різних умовах проведення експерименту. Як показує аналіз літературних джерел у цьому напрямі дослідження [50–58], існуючі методи нормалізації даних ДНК мікрочіпів можна розділити на дві групи. До першої групи відносяться методи, що вико- ристовують опорну підмножину експериментальних даних. Друга група методів використовує повний набір даних, що досліджуються. Методи нормалізації, що складають першу групу, поділяються у свою чергу на лінійні та нелінійні. Цю групу складають лінійні методи (Constant), метод інваріантної множини (Invariantset) і метод сплайнів (Qspline). Другу групу складають метод контрастів (Contrast), метод циклічної локальної регресії (Loess) та квантильна нормалізація (Quantiles).

Застосування алгоритму лінійної нормалізації [50, 51] передбачає розрахунок усіченого середнього  $\bar{x}$  на першому кроці. При цьому 2% мінімальних та макси- мальних значень відкидаються. Далі обирається опорний вектор і розраховується середнє значення інтенсивностей світла у ньому  $\bar{x}_{op}$ . Нормалізація даних здійсню- ється відповідно до формули:

$$x'_{ik} = \frac{\bar{x}_{op}}{\bar{x}} x_{ik}, \quad (1.6)$$

де  $i$  та  $k$  ідентифікують ген та об'єкт, відповідно.

Застосування методу циклічної локальної регресії [55] передбачає формування векторів  $M$  та  $A$ , елементи яких розраховуються так:

$$M_i = \log_2 \left( \frac{x_{ik}}{x_{in}} \right), A_i = \frac{1}{2} \log_2 (x_{ik} \cdot x_{in}), \quad (1.7)$$

де  $x_{ik}$  та  $x_{in}$  – значення експресії  $i$ -го гену для зразків  $k$  та  $n$  відповідно.

Передбачається, що між векторами  $M$  та  $A$  існує лінійна регресійна залежність. На другому кроці розраховується нормуюча поправка, як різниця між фактичним значенням елемента  $i$  вектора  $M$  та значенням регресійної функції, що відповідає  $i$ -му гену:

$$\delta M_i = M_i - \hat{M}_i. \quad (1.8)$$

Нормалізовані значення інтенсивностей світла, що відповідають  $i$ -му гену для зразків  $k$  та  $n$  розраховуються відповідно до рівняння:

$$x'_{ik} = 2^{(A_i + \frac{\delta M_i}{2})}, x'_{in} = 2^{(A_i - \frac{\delta M_i}{2})}. \quad (1.9)$$

Метод контрастів [56] також передбачає формування векторів  $M$  та  $A$  на першому кроці. Однак на відмінність між методом циклічної локальної регресії, на другому кроці обирається опорний вектор і всі подальші розрахунки виконуються по відношенню до опорного вектора.

При застосуванні методу квантільної нормалізації [50, 52] передбачається, що дані у  $k$ -векторах розподілені за нормальним законом. При цьому графік квантілів у  $k$ -вимірному просторі може бути представлений прямою лінією, яка лежить уздовж діагоналі, що задана вектором:  $(\frac{1}{\sqrt{k}}, \dots, \frac{1}{\sqrt{k}})$ . Нормалізація даних у цьому випадку передбачає знаходження проекції точок точкової  $k$ -вимірної діаграми розсіювання на цю діагональ. Таким чином, вихідні значення інтенсивностей світла трансформуються у значення квантілів. При цьому діапазон зміни значень інтенсивностей світла буде однаковий для всіх об'єктів, що досліджуються.

При застосуванні методу інваріантної множини [57] на першому кроці формується опорна множина РМ проб з якомога меншою дисперсією значень інтенсивностей світла у кожному зразку. Опорний вектор формується як середнє значення інтенсивностей світла усіх векторів опорної множини. Наступний крок передбачає знаходження залежності між значеннями інтенсивності світла у опорному векторі та у векторах відповідних зразків даних, що досліджуються з використанням лінійної або нелінійної функції. Для знаходження параметрів функції застосовується значення інтенсивностей світла у опорному векторі та у відповідних векторах зразків. Отримується система рівнянь, яка дозволяє визначити параметри функції, на підставі яких виконується нормалізація значень інтенсивностей світла, що відповідають даним, що досліджуються.

Застосування методу інтерполяції сплайнами [58] передбачає використання кубічних сплайнів для згладжування неоднорідностей при створенні діаграм розподілу інтенсивностей світла. На першому етапі інтенсивності світла розподіляються на групи квантілів з однаковою кількістю значень інтенсивності світла у кожній групі. Нормалізація даних у цьому випадку передбачає масштабування інтенсивностей світла  $k$ -го квантіля масиву так, щоб воно дорівнювало інтенсивності світла  $k$ -го квантіля опорного масиву. Цей процес повторюється для кожного масиву даних. Інтенсивності світла, що лежать між квантілями, корегуються шляхом інтерполяції сусідніх квантілів.

Етап РМ корекції дозволяє зменшити вплив ефекту неспецифічної гібридизації шляхом урахування інтенсивностей світла ММ проб [51]. При застосуванні методу РМ корекції «subtractmm» значення інтенсивності світла на РМ пробах зменшується на величину інтенсивностей світла у відповідних ММ пробах. Метод «MAS» РМ корекції є удосконалення методу «subtractmm». Застосування цього методу враховує той факт, що інтенсивність світла у ММ пробі може бути вищою за інтенсивність світла у відповідній РМ пробі, що може призвести до негативного значення інтенсивності світла. При застосуванні методу РМ корекції «rmonly» використовуються тільки РМ проби, а ММ проби не використовуються. Тобто при застосуванні цього методу ефект крос-гібридизації не враховується.

Необхідність етапу сумаризації визначається тим фактом, що для підвищення об'єктивності експерименту, на мікрочіпі знаходяться декілька нуклеотидних послідовностей, що відповідають одному типу генів. Тому на цьому етапі формується остаточне значення експресії генів, що досліджуються. Алгоритм «avgdiff» є найпростішим алгоритмом цієї серії [51]. Значення експресії генів у цьому випадку розраховується як середнє арифметичне значення інтенсивностей світла проб, що відповідають цьому гену. При застосуванні алгоритму сумаризації «Tukey Bi-Weight» [61] значення експресії генів розраховується як зважене середнє Тьюкі інтенсивностей світла, що відповідають цьому гену. Даний алгоритм входить в комплекс методів обробки даних мікрочіпових експериментів MAS-5, запропонований компанією Affymetrix [51]. Метод сумаризації «liwong» [58] заснований на припущенні, що між різницею значень інтенсивностей світла РМ та ММ проб відповідного гену і його експресією існує лінійна залежність. Експресія цього гена розраховується як зважене середнє значення різниць РМ-ММ проб, що відповідають цьому гену. Алгоритм сумаризації «medianpolish» [52, 59] також враховує характер взаємодії РМ та ММ проб відповідного гену із значенням його експресії. При цьому за рівень експресії гена приймається концентрація інтенсивностей проб на мікрочіпі, що відповідають цьому гену. При застосуванні методу сумаризації «playerout» [60] оцінюються вагові коефіцієнти інтенсивностей світла проб на мікрочіпі, що відповідають цьому гену, значення експресії якого розраховується як середнє значення отриманих ваг.

Очевидно, що якість обробки даних ДНК-мікрочіпу суттєво залежить від обраної комбінації методів реалізації кожного з етапів. Кожний метод має свої переваги та недоліки і об'єктивний вибір комбінації методів може бути здійснений тільки за умови застосування відповідних критеріїв якості оцінки отриманих даних експресії генів. У працях [62, 63], автори запропонували покроковий алгоритм об'єктивного вибору комбінації методів обробки ДНК-мікрочіпу на основі інформаційної ентропії Шеннона, що розраховувалася із застосуванням методу Джеймса та Стейна (James-Stein shrinkage estimator) [64]. Передбачалося, що оптимальна комбінація методів відповідає мінімальному значенню критерію ентропії Шеннона. Блок-схему запропонованого алгоритму зображено на рис. 1.5.

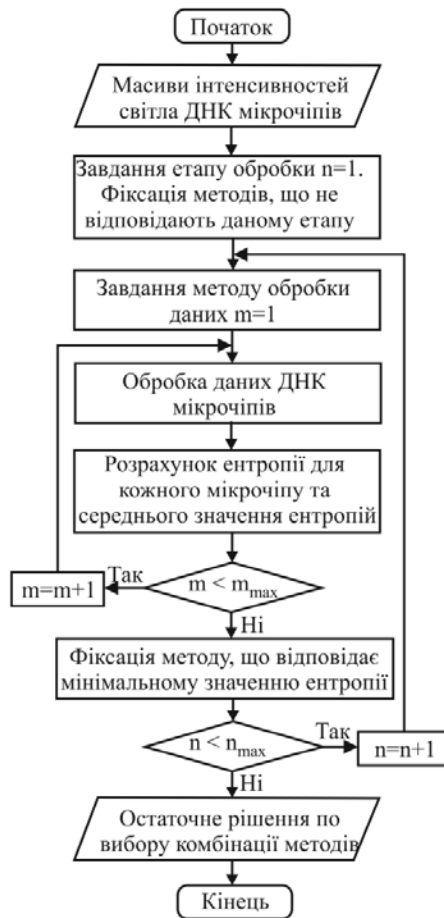


Рис. 1.5. Блок-схема алгоритму з вибору оптимальної комбінації методів обробки даних ДНК мікрочіпів на основі критерію ентропія Шеннона

Практична реалізація цього алгоритму передбачала фіксацію довільної комбінації методів на етапах від другого до четвертого на першому кроці. Далі по чергово застосовувалися всі методи, що відповідають першому етапу. Оптимальний метод відповідав мініимальному значенню критерію якості. На наступних етапах ця покрокова процедура застосовувалася до інших комбінацій методів.

Недоліком вищеописаного підходу є те, що критерій ентропії Шеннона у цьому випадку є умовно об'єктивним. Автори показали, що вибір типу розрахунку ентропії не має суттєвого впливу на результат, що отримується, але об'єктивність рішення з вибору оптимальної комбінації методів може бути підвищена тільки при застосуванні кількісних критеріїв, що враховують стан об'єктів, які досліджувалися при виготовленні мікрочіпів. Тобто ця модель потребує удосконалення.

### 1.2.2.2. Методи обробки результатів експерименту на основі застосування технології секвенування молекул мРНК

Експерименти зі секвенування молекул мРНК мають перевагу перед експериментами із застосуванням ДНК-мікрочіпів, яка полягає у більш високій точності оцінки значення експресії генів. До недоліків цього типу експериментів можна віднести високу вартість, але з кожним роком їх вартість зменшується, що дає підставу вважати, що з часом цей тип експерименту повністю замінить експеримент із застосуванням ДНК-мікрочіпів.

Методам формування та обробки даних експерименту по секвенуванню молекул мРНК у останні роки присвячено велику кількість наукових праць [65–73]. Запропоновані методи відрізняються між собою за типами та порогоми, які використовуються для ідентифікації генів і оцінки їх кількості та за алгоритмами, що використовуються на етапі формування експериментальних даних. Слід зазначити, що вибір алгоритму має дуже великий вплив на точність оцінки ряду молекул мРНК у зразках, що секвенуються. З цієї причини дослідження різних інструментів для обробки даних експериментів з секвенування молекул мРНК може сприяти підвищенню якості експериментальних даних, що отримуються.

Як було зазначено раніше, одним із наслідком реалізації експерименту щодо секвенування молекул мРНК є матриця кількості генів для об'єктів, що досліджуються. На рис. 1.6 зображено блок-схему покрокової процедури обробки експериментальних даних, отриманих при секвенуванні молекул мРНК.

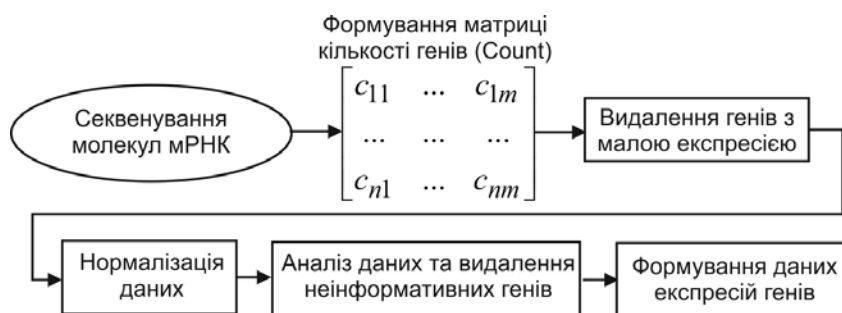


Рис. 1.6. Покрокова процедура формування експериментальних даних при застосуванні експерименту щодо секвенування молекул мРНК

Як показує аналіз експериментальних досліджень, наведений у роботі [74], у вихідних даних кількість генів дорівнює приблизно 55000, при цьому половина генів мають нульову експресію для усіх об'єктів, що досліджуються. Очевидно, що дані гени можуть бути видалені з даних без втрати інформації. До того ж, існують гени, кількість яких для усіх об'єктів, що досліджуються, є досить малою, порівняно з кількістю високо- та середньо експресованих генів. Ці гени також можуть бути видалені з даних без суттєвої втрати інформації. При цьому слід зазначити, що вибір порогового значення, за яким гени поділяються на інформативні та неінформативні, є досить умовним і визначається емпіричним шляхом у процесі моделювання з

урахуванням кількості генів, які необхідно залишити для подальшого дослідження. Кількість генів зазвичай варіюється в інтервалі від 0 (неекспресовані гени) до десятків мільйонів. Тому наступний крок передбачає трансформацію абсолютних значень кількостей генів у більш привабливий діапазон їх змін. У модулі Bioconductor [75, 76], який реалізований у різних програмних середовищах (Matlab, R, python) і містить велику кількість функцій для формування та обробки даних експериментів з визначення експресії генів, цей крок виконується таким чином: на першому етапі абсолютні значення кількості генів трансформуються у відносні відповідно до рівняння:

$$x'_{ij} = \frac{x_{ij}}{\sum_{i=1}^n x_{ij}} \cdot 10^6, \quad (1.10)$$

де  $i = \overline{1, n}$  та  $j$  гени та досліджуваний об'єкт. Порогове значення відносної кількості генів, що дозволяє розділити гени на низько експресовані та високо експресовані приймається у цьому випадку за 1 ( $\log_2 1 = 0$ ). Тобто, якщо  $x'_{ij} \leq 1$  для всіх об'єктів, що досліджуються, ген видаляється з даних як неекспресований.

Методи нормалізації, що застосовуються до даних експресії генів, отриманих шляхом секвенування молекул мРНК [77–79]. У найпростішому випадку, до трансформованих відповідно до формули (1.10) значень застосовується функція  $\log_2$ . TMM (Trimmed Mean of M values) метод нормалізації даних експресії генів заснований на припущенні, що гени, значення експресії яких для різних зразків суттєво не відрізняється одне від одного, ідентифікуються як недиференційно експресовані. На наступному кроці розраховується лінійний індекс масштабування для кожної проби як  $\log_2$  зважених середніх значень експресії генів щодо їх абсолютних значень. TMMwsp (TMM with singleton pairing) є модифікація TMM методу. При застосуванні цього методу гени, що мають нульові значення експресії для окремих зразків, ігноруються. В процесі масштабування використовуються тільки гени з позитивним значенням експресії. Застосування RLE (Relative Log Expression) методу нормалізації передбачає розрахунок медіани середніх геометричних значень експресії генів для усіх зразків, що досліджуються. Відношення медіани кожної проби до медіани усієї вибірки використовується як коефіцієнт масштабування. UQS (Upper Quartile Scaling) метод нормалізації передбачає розрахунок коефіцієнта масштабування із застосуванням 75% квантилів після видалення генів з нульовою експресією для усіх зразків.

У статті [74] описано порівняльний аналіз вищенаведених методів нормалізації при застосуванні даних експресії генів пацієнтів, що досліджувалися на нейроблас-тому головного мозку при отриманні різних доз  $\text{MnCl}_2$  (0,1,5,10,50,100  $\mu\text{M}$  протягом 5 годин). Дані отримані шляхом секвенування молекул мРНК. Для оцінки ефективності відповідного методу, на першому етапі зразки були розбиті на шість груп (кластерів) залежно від отриманої дози. Як критерій якості був використаний критерій якості кластеризації, що розраховувався як мультиплікативна комбінація критерію Calinski-Harabasz та WB-index [80–82]:

$$QC = \frac{N_c(N_c-1) \cdot QCW^2}{(N_s-N_c) \cdot QCB^2}, \quad (1.11)$$

де  $N_c$  є кількість кластерів;  $N_s$  – кількість зразків (об’єктів, що досліджуються);  $QCW$  – середня кореляційна відстань від зразків до центру кластера, де дані зразки розташовані;  $QCB$  – середня кореляційна відстань між центрами кластерів. Мінімальне значення критерію (1.11) відповідало оптимальному методу нормалізації для даного типу даних. Результати моделювання показали що UQS метод нормалізації за даним критерієм дозволяє отримати найбільш якісні дані.

Однак слід зазначити, що запропонований у роботі [74] метод досить великий відсоток суб’єктивності і потребує доопрацювання. Підвищення об’єктивності методу можливе за рахунок більш ретельного вибору кількісних критеріїв якості обробки даних експресії генів, застосування сучасних методів інтелектуального аналізу даних та машинного навчання та верифікації моделі із застосуванням різних даних експресії генів.

### 1.2.3. Методи статистичного аналізу та редукції профілів експресії генів

Як було зазначено у розділі 1.2.2, геном людини містить приблизно 25000 активних генів. Після видалення генів з малим значенням експресії кількість генів зменшується приблизно до 10000 генів. Подальшим кроком є видалення неінформативних або вибір найбільш інформативних профілів експресії генів, які дозволяють ідентифікувати об’єкти, що досліджуються, враховуючи клас, до якого вони належать з найкращою роздільною здатністю. Сьогодні актуальними є дві групи методів, що дозволяють на основі статистичного аналізу вибрати взаємно корельовані профілі експресії генів, які дозволяють розрізнити класи об’єктів, що досліджуються: методи кластеризації та бікластеризації. Різницю між кластерним та бікластерним аналізами зображено на рис. 1.7.

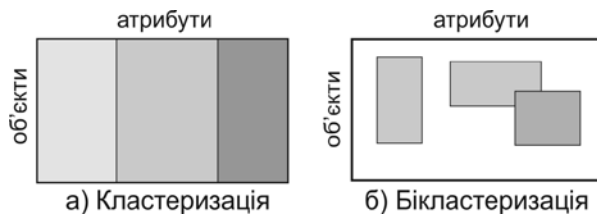


Рис. 1.7. Різниця між результатами кластерного та бікластерного аналізів

Як можна бачити, методи кластерного аналізу дозволяють розділити об’єкти або атрибути, що характеризують об’єкти, на кластери залежно від ступеня їх близькості відповідно до обраної метрики. Результатом бікластерного аналізу є кластери взаємно корельованих рядків та стовпців, при цьому можливі випадки, коли один і той самий об’єкт або/та атрибут належать до різних кластерів. На цей час існує велика кількість алгоритмів як кластерного, так і бікластерного аналізу, огляд яких наведено у наступних підрозділах.

### 1.2.3.1. Огляд сучасних методів бікластерного аналізу

Бікластерному аналізу даних експресії генів на даний час присвячено велику кількість наукових праць [83–90]. Під бікластером у цьому випадку розуміють підмножину взаємно корельованих рядків та стовпців [91]. Нехай початкова матриця експресії генів містить  $n$  рядків (об'єктів) та  $m$  стовпчиків (генів):

$$A = \{e_{ij}\}, i = \overline{1, n}, j = \overline{1, m}, \quad (1.12)$$

де  $e_{ij}$  – значення експресії  $j$ -го гену, визначене для  $i$ -го зразка (об'єкта).

Нехай  $X = \{x_1, x_2, \dots, x_n\}$  є підмножина рядків а  $Y = \{y_1, y_2, \dots, y_m\}$  – множина стовпців матриці (1.12). Тоді матрицю  $A$  можна представити так:  $A = (X, Y)$ . Бікластер у цьому випадку являє собою підмножину матриці  $A$ :  $A_{IJ} = (I, J)$ , де  $I = \{i_1, i_2, \dots, i_p\}$  і  $J = \{j_1, j_2, \dots, j_s\}$  є підмножини взаємно корельованих рядків і стовпців, відповідно.

У роботах [92–94] автори провели порівняльний аналіз наявних алгоритмів бікластерного аналізу з виділенням їхніх переваг та недоліків. У статті [95] запропоновано спектральні алгоритм бікластерного аналізу, зроблено оцінку ефективності даного алгоритму на синтетичних бікластерах та профілях експресії генів, виділені переваги та недоліки цього алгоритму. У роботі [96] авторами наведено модельну класифікацію бікластерів залежно від характеру взаємодії між рядками та стовпцями. Нехай бікластер  $A_{IJ} = (I, J)$  містить  $p$  рядків та  $s$  стовпців, а  $\delta$  є константа, значення якої належить діапазону змін значень матриці  $A$ .

Найпростішим типом бікластера є бікластер зі сталими значеннями:  $e_{ij} = \delta$ . У реальних даних цей тип бікластеру найчастіше являє собою шумову компоненту і видаляється в процесі фільтрації даних.

До другого типу бікластерів відносять бікластери із сталими значеннями рядків або стовпців. Алгоритми, що застосовуються для виділення таких типів бікластерів, наведено у працях [97–99]. Основна ідея полягає в наступному. На першому етапі дані нормалізуються за рядками або стовпцями для приведення їх до єдиного діапазону. На другому етапі розраховується евклідова відстань між відповідними рядками та стовпцями і виділяються бікластери, які відповідають мінімальному значенню евклідової відстані між відповідними елементами.

У роботі [100] один із перших алгоритмів бікластерного аналізу даних експресії генів. Реалізація цього алгоритму передбачає фіксацію порогового значення середньоквадратичної похибки під час роботи алгоритму. За цей трешолдінговий коефіцієнт приймається позитивне значення дельта-параметру  $\delta \geq 0$ , який визначає структуру бікластерів та їхню кількість. Класифікацію алгоритмів бікластерного аналізу даних експресії генів та їх порівняльний аналіз з виділенням переваг та недоліків наведено у роботі [92].

У праці [101] представлено результати досліджень щодо порівняльного аналізу алгоритмів бікластеризації із застосуванням синтетичних бікластерів та даних експресії генів пацієнтів, що досліджувалися на рак легенів [102]. Було досліджено наступні алгоритми: 1) Bimax [100]; 2) Cheng and Church (CC) алгоритм бікластеризації [103]; 3) спектральна бікластеризація [95]; 4) ensemble метод бікластеризації [104].

Синтетичні бікластери, що використовувалися у процесі моделювання, наведено на рис. 1.8. Як зовнішній критерій, що дозволяє порівняти результати бікластеризації відповідних даних при застосуванні різних алгоритмів бікластерного аналізу, було використано індекс Жаккара:

$$JI(BC(res_1), BC(res_2)) = \frac{1}{n_1} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left( \frac{BC_i(BC(res_1)) \cap BC_j(BC(res_2))}{BC_i(BC(res_1)) \cup BC_j(BC(res_2))} \right), \quad (1.13)$$

де  $n_1$  і  $n_2$  – кількість бікластерів у бікластеризаціях 1 та 2, відповідно.

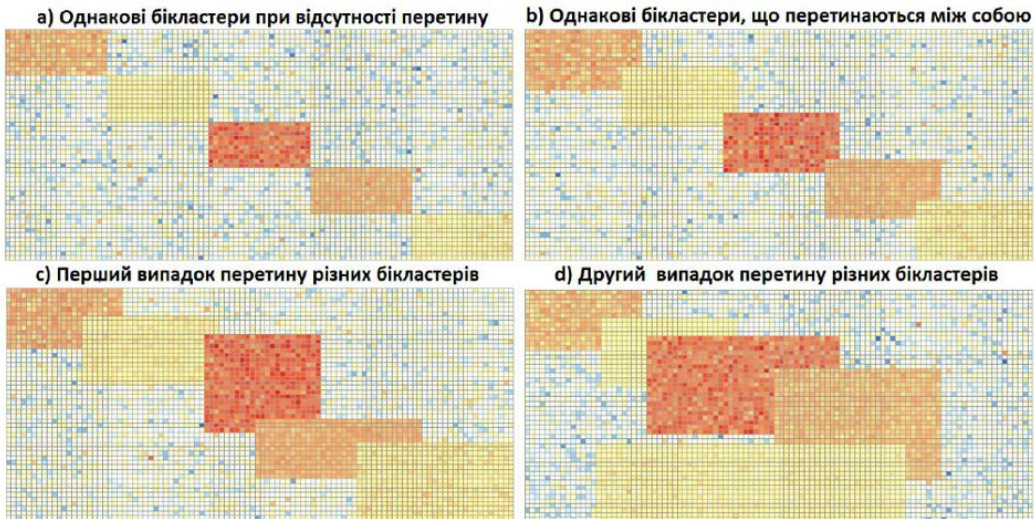


Рис. 1.8. Синтетичні бікластери, що використовувалися для оцінки алгоритмів бікластеризації даних

У випадку наявності перетину бікластерів формула (1.13) коригується таким чином:

$$JI_1(BC(res_1), BC(res_2)) = \frac{JI(BC(res_1), BC(res_2))}{\max(JI(BC(res_1), BC(res_1)), JI(BC(res_2), BC(res_2)))}. \quad (1.14)$$

Слід зазначити, що у випадку відсутності перетину бікластерів формула (1.14) трансформується у (1.13).

Внутрішній критерій оцінки якості відповідної бікластеризації розраховувався як зважене середнє евклідових відстаней між усіма рядками та стовпчиками даних відповідного бікластера, при цьому на останньому кроці розраховувався комплексний внутрішній критерій якості як середнє значення внутрішніх критеріїв, що отримані для усіх бікластерів даної бікластеризації. Мінімальне значення комплексного критерію якості відповідало найкращій бікластеризації.

Результати моделювання, що представлені у праці [101], показали високу ефективність алгоритму «ensemble» порівняно з іншими алгоритмами бікластерного

аналізу. У випадку використання синтетичних бікластерів застосування цього алгоритму дозволяє отримати адекватну бікластеризацію, порівняно з досконалою, що отримана штучним методом. Окрім того, отримані результати моделювання показали високу ефективність внутрішнього критерію якості бікластеризації, оскільки його глобальний мінімум відповідав глобальному максимуму індексу Жаккара, і зміни значень даних критеріїв показали високий рівень взаємної кореляції.

Як результат досліджень автори запропонували інформаційну технологію бікластеризації даних на основі алгоритму бікластерного аналізу «ensemble», яку було застосовано до аналізу даних експресії генів (виділення бікластерів). У результаті моделювання на першому кроці були визначені оптимальні параметри алгоритму бікластеризації «ensemble», які відповідали мінімальному значенню внутрішнього критерію якості. На другому кроці цей алгоритм був застосований до даних експресії генів. Було видалено 27 бікластерів. На рис. 1.9 зображені профілі експресії генів для дев'яти найбільших бікластерів. На основі проведених досліджень авторами було запропоновано гібридну модель кластер-бікластерного аналізу, яка дозволяє виділити бікластери для подальшого аналізу. Блок-схему даної моделі зображено на рис. 1.10, практична реалізація якої передбачає застосування щільнісного алгоритму DBSCAN для видалення профілів експресії генів, які ідентифіковані як шум, на першому кроці. На другому кроці, до виділених профілів експресії генів застосовувався алгоритм кластеризації SOTA з кореляційною метрикою. При цьому множину профілів експресії генів було розділено на дві приблизно однакові підмножини. На останньому кроці до отриманих підмножин було застосовано бікластерний аналіз.

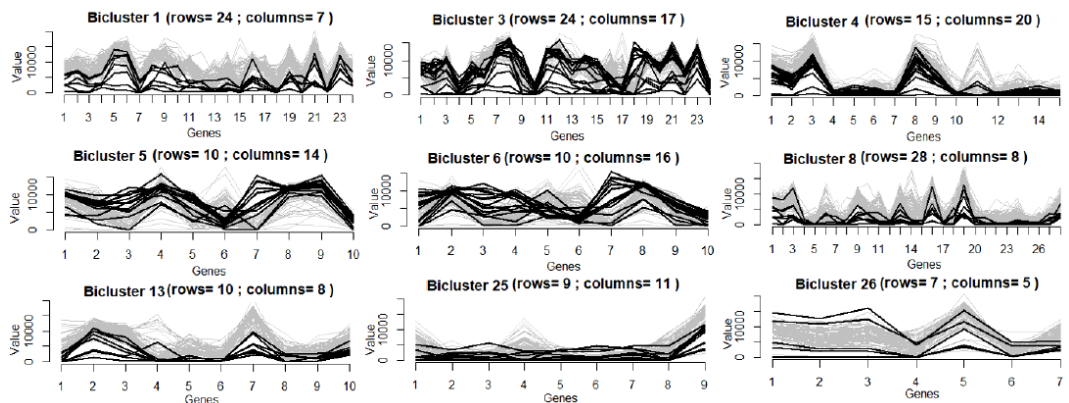


Рис. 1.9. Результати моделювання з бікластеризації профілів експресії генів

До основних проблем, що обмежують успішне застосування бікластерного аналізу, слід віднести велику кількість малих бікластерів та великий обсяг даних, що не використовуються, оскільки вони не попадають у бікластери. Окрім того, більшість (якщо не всі) бікластери не містять усі зразки, що досліджуються, що обмежує їх застосування у системах діагностики захворювань та ускладнює процес інтерпретації реконструйованих моделей генних регуляторних мереж у процесі їх

подальшого моделювання. Таким чином, можна зробити висновок, що алгоритми бікластерного аналізу потребують удосконалення за рахунок створення нових методів формування бікластерів, що відповідають поставленим цілям. Актуальним є також науковий напрям розробки гібридних моделей на основі комплексного використання алгоритмів кластерного та бікластерного аналізів.

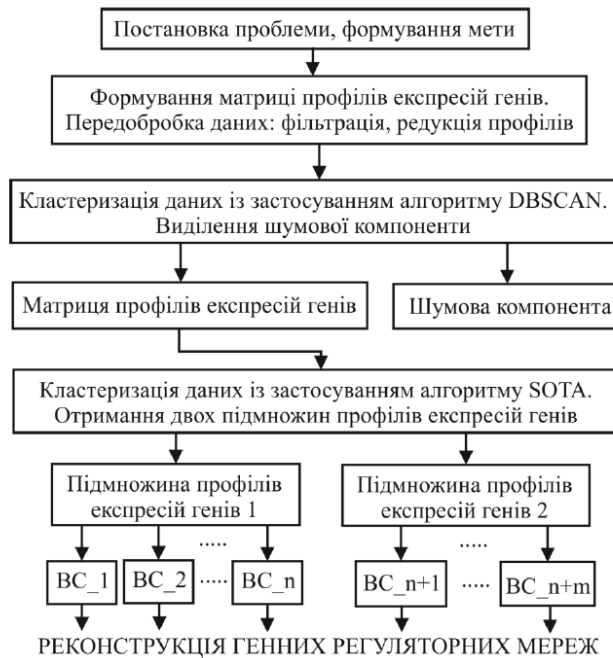


Рис. 1.10. Блок-схема гібридної моделі покрокового кластер-бікластерного аналізу

### 1.2.3.2. Огляд сучасних методів та алгоритмів кластеризації даних експресії генів

Задача кластеризації даних експресії генів зводиться до розбиття множини профілів експресії генів на непусті підмножини кластерів, які у ідеальному випадку не перетинаються, при цьому профілі експресії генів у окремих кластерах мають найбільший рівень взаємної кореляції. На відмінність від бікластеризації, у випадку застосування кластерного аналізу кластери містять усі об'єкти, що досліджуються. Цей факт створює умови для розробки систем діагностики складних захворювань шляхом створення класифікатора хвороби на основі кластерів, які містять взаємно корельовані профілі експресії генів, що дозволяють розрізнити стан здоров'я пацієнтів.

Загалом задача кластеризації може бути подана таким чином. Нехай початкові дані представлені у вигляді матриці експресії генів:

$$A = \{e_{ij}\}, i = \overline{1, n}, j = \overline{1, m}, \quad (1.15)$$

де  $n$  і  $m$  – кількість рядків (об'єктів) та стовпців (генів), відповідно. Результатом кластеризації є множина кластерів  $k$ :

$$K = \{K_s\}, s = \overline{1, k}, K_1 \cup K_2 \cup \dots \cup K_k = A;$$

$$K_s \cap K_p = \emptyset; s \neq p; s, p = \overline{1, k}. \quad (1.16)$$

Сьогодні питанням кластеризації даних різної природи присвячена велика кількість наукових праць [105–115]. У загальному випадку реалізація процесу групування об'єктів передбачає такі кроки:

1. Постановка задачі. Аналіз та формування даних, оцінка (якщо можливо) цільової кількості кластерів.
2. Визначення метрики, що визначає ступінь близькості між об'єктами, кластерами, об'єктами та кластерами.
3. Визначення критеріїв оцінки якості цільової кластерної структури.
4. Вибір алгоритму кластеризації, визначення параметрів його функціонування.
5. Реалізація процесу кластеризації даних, формування кластерної структури.
6. Розрахунок критеріїв якості, валідація моделі, інтерпретація та аналіз отриманих результатів.

На цей час 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 огляді [108] наведено структурну схему покрокової процедури формування кластерної структури [116] (рис. 1.11) та блок-схему способів формування кластерної структури і особливо поширених сьогодні алгоритмів кластеризації даних (рис. 1.12). Вибір відповідного алгоритму визначається типом даних, що досліджуються та бажаним способом формування кластерної структури.

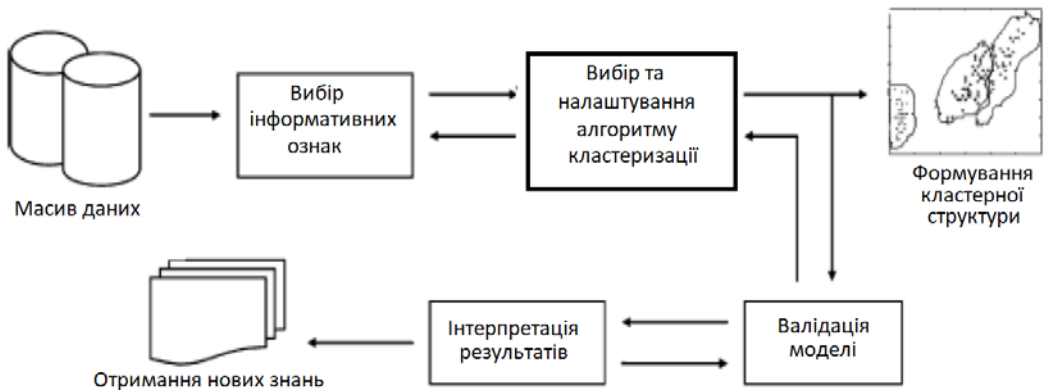


Рис. 1.11. Структурна схема покрокової процедури формування кластерної структури [77]

Результатом реалізації ієрархічних алгоритмів кластеризації є дендрограма, тобто дерево рішень на різних етапах групування об'єктів. Основною проблемою у цьому випадку є вибір оптимального ієрархічного рівня групування (агломеративний алгоритм) або розділення (дивізійний алгоритм) об'єктів, що досліджуються. Окрім того, різні методи реалізації одного алгоритму можуть генерувати

різні структури дендрограм, що може впливати на об'єктивність прийняття остаточного рішення з формування кластерної структури.

Самоорганізуючі карти Кохонена [117–119] та їх логічне продовження алгоритм кластеризації SOTA (Self Organizing Tree Algorithm) [12, 121] у деякій мірі вирішують проблему організації кластерної структури, але результат їх роботи залежить від вибору параметрів відповідного алгоритму, що потребує досліджень з оцінки якості відповідної кластеризації на основі аналізу кількісних критеріїв якості кластеризації, що ускладнює їхнє успішне застосування.

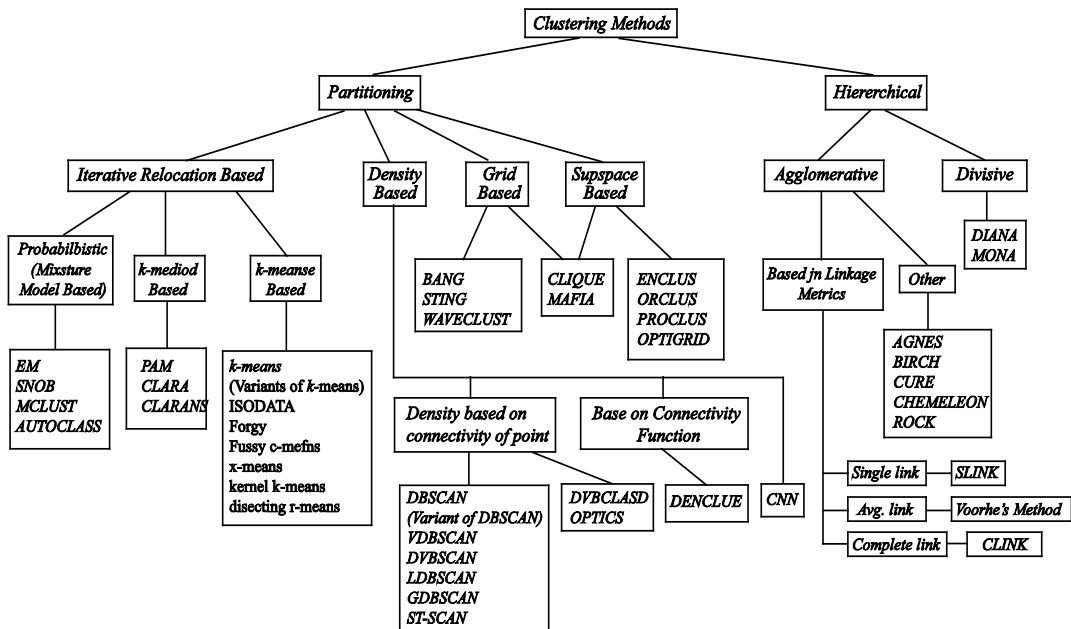


Рис. 1.12. Блок-схема способів та найбільш розповсюджених алгоритмів формування кластерної структури

Реалізація ітеративних алгоритмів формування кластерної структури передбачає покрокову (ітеративну) процедуру перерозподілу об'єктів між кластерами до досягнення умови зупинки відповідного алгоритму. До цього типу алгоритмів належать алгоритми k-середніх (k-means і x-means) [122, 123] та с-середніх (fuzzy c-means) [124, 125], щільнісні алгоритми DBSCAN та OPTICS [126–128], еволюційні та генетичні алгоритми [129–131], ґратчасті алгоритми [132] та алгоритми на основі формувань підпросторів вихідного простору розташування об'єктів [133–135]. Простота реалізації та прозорість є головними перевагами алгоритмів на основі ітераційного розподілу об'єктів у заздалегідь визначені кластери. До основних недоліків слід віднести високу чутливість до вибору параметрів алгоритмів, які визначають кластерну структуру. До того ж, більшість алгоритмів цього типу потребують ініціалізацію кількості кластерів на початку роботи алгоритму, що не завжди є можливим. У випадку кластеризації великої кількості профілів експресії генів оптимальна кількість кластерів може бути визначена тільки на основі застосу-

вання кількісних критеріїв якості кластеризації тільки після розподілу генів у кластери. Еволюційні генетичні алгоритми використовують поняття «популяція індивідуумів». Практична реалізація цього типу алгоритмів передбачає застосування одного або декількох генетичних операторів, що враховують генетичну мінливість об'єкта, що досліджується. Логічним продовженням цього типу алгоритмів є еволюційні алгоритми на основі генетичного програмування, що містять у якості окремих індивідуумів генетичні програми (код). Основна відмінність класичних еволюційних алгоритмів від еволюційних алгоритмів на основі генетичного програмування полягає у природі індивідуумів. Якщо у еволюційних генетичних алгоритмах індивідууми являють собою рядки символів встановленої довжини (хромосоми), то у еволюційних алгоритмах на основі генетичного програмування індивідууми наведено у вигляді нелінійних функцій або генетичних програм. Налаштування даних алгоритмів передбачає кодування окремих індивідуумів у рядки відповідних символів, які в процесі роботи алгоритму експресуються у нелінійні об'єкти різних форм, що відповідають послідовності символів цього індивідуума. До основних недоліків еволюційних генетичних алгоритмів слід віднести те, що функціональні можливості індивідуумів визначає повна структура хромосоми, що обмежує їх роздільну здатність при розпізнаванні об'єктів. Цей недолік деякою мірою компенсують еволюційні алгоритми на основі генетичного програмування. Нелінійні об'єкти, які створюються в процесі реалізації коду символів окремих індивідуумів, дозволяють підвищити чутливість алгоритму, але сьогодні вони потребують доопрацювання з метою підвищення стабільності і точності. Природним розвитком еволюційних генетичних та на основі генетичного програмування алгоритмів є алгоритми на основі програмування експресії генів. У цьому типі алгоритмів код послідовності символів індивідуумів являє собою значення експресії генів, які здатні репрезентувати нелінійний об'єкт будь-якої форми. Структура хромосом у цьому випадку здатна представити об'єкт будь-якої форми. Основним недоліком цього типу алгоритму є те, що у випадку застосування об'єктів високої розмірності, якими є профілі експресії генів, застосування алгоритму програмування експресії генів потребує великої кількості комбінацій значень експресії генів у відповідних послідовностях, що суттєво ускладнює процес кодування інформації.

Щільнісні ітераційні алгоритми кластеризації формують кластери відповідно до щільності розподілу об'єктів у просторі ознак. Особливістю цього типу алгоритмів є те, що їх застосування дозволяє виділити об'єкти, що мають низьку щільність розподілу порівняно з основною групою об'єктів. Здебільшого ці об'єкти ідентифікуються як шум. Ця особливість може бути використана для фільтрації даних на етапі попередньої обробки. З іншого боку, об'єкти з високою щільністю групуються у відповідні кластери. Однак слід відзначити, що щільнісні алгоритми мають велику чутливість до параметрів їх налаштування, які визначають характер групування об'єктів у кластери. Тому ефективність використання цього типу алгоритмів залежить від попередніх досліджень з визначення параметрів алгоритму з використанням критеріїв оцінки якості кластеризації даних, що досліджуються.

Ітераційні гранчасті алгоритми являють собою простір ознак у вигляді багатовимірної ґратки довільної форми. Подальші кроки алгоритмів цього типу передбачає кластеризацію комірок відповідної ґратки залежно від густини розподілу об'єктів у відповідних комірках. Основним недоліком цього типу алгоритмів є те, що результат їхньої роботи суттєво залежить від розміру комірок ґратки, що обмежує їх використання для даних високої розмірності.

Наведений огляд алгоритмів дозволяє зробити висновок, кожний з них має свої переваги та недоліки і орієнтований на відповідний тип даних. До того ж результат роботи відповідного алгоритму суттєво залежить від його вихідних параметрів, оптимальне значення яких може бути визначене на основі застосування критеріїв якості кластеризації, що враховують як розподіл об'єктів у кластерах, так і розподіл кластерів у просторі ознак. Критерії якості кластеризації своєю чергою поділяються на внутрішні та зовнішні. Внутрішні критерії дозволяють оцінити якість конкретної кластеризації залежно від характеру розподілу об'єктів і кластерів. Зовнішні критерії дозволяють оцінити та порівняти характер розподілу об'єктів у різних кластерних структурах.

У таблиці 1.2 наведені внутрішні критерії, що сьогодні застосовуються для оптимізації кластерної структури та правила їх інтерпретації. Правила *max* та *min* свідчать про те, що оптимальна кластерна структура відповідає максимальному або мінімальному значенню внутрішнього критерію, відповідно. Правила *max diff* та *min diff* своєю чергою дозволяють визначити оптимальну кластерну структуру на основі оцінки максимальної або мінімальної зміни значення цього критерію у процесі перебору кластеризацій.

Таблиця 1.2

*Внутрішні критерії оцінки якості кластерної структури*

| №  | Критерій                | Правило         | №  | Критерій                    | Правило         |
|----|-------------------------|-----------------|----|-----------------------------|-----------------|
| 1  | Ball-Hall [136]         | <i>max diff</i> | 14 | McClain-Rao [148]           | <i>min</i>      |
| 2  | Banfeld-Raftery [137]   | <i>min</i>      | 15 | PBM [149]                   | <i>max</i>      |
| 3  | C index [138]           | <i>min</i>      | 16 | Point biserial [150]        | <i>max</i>      |
| 4  | Calinski Harabasz [139] | <i>max</i>      | 17 | Ratkowsky Lance [151]       | <i>max</i>      |
| 5  | Davies Bouldin [140]    | <i>min</i>      | 18 | Ray Turi [152]              | <i>min</i>      |
| 6  | $ T / W $ [141]         | <i>min diff</i> | 19 | Scott-Symons [141]          | <i>min</i>      |
| 7  | Dunn [142]              | <i>max</i>      | 20 | SD [153]                    | <i>min</i>      |
| 8  | Dunn generalized [143]  | <i>max</i>      | 21 | S_Dbw [154]                 | <i>min</i>      |
| 9  | Gamma [144]             | <i>max</i>      | 22 | Silhouette [155]            | <i>max</i>      |
| 10 | G + [145]               | <i>min</i>      | 23 | Tr(W) [156]                 | <i>max</i>      |
| 11 | $k^2 W $ [146]          | <i>max diff</i> | 24 | Tr(W <sup>-1</sup> B) [156] | <i>max diff</i> |
| 12 | $\log T / W $ [147]     | <i>min diff</i> | 25 | Wemmert-Gancarski [157]     | <i>max</i>      |
| 13 | $\log(BGSS=WGSS)$ [148] | <i>min diff</i> | 26 | Xie-Beni [158]              | <i>min</i>      |

Вибір відповідного критерію визначається типом даних, що досліджуються, і повинен у кожному випадку на етапи попередньої обробки даних. Однак слід зазначити, що задовільна кластеризація для одних даних за відповідними внутрішніми критеріями не гарантує аналогічних результатів для інших аналогічних даних.

Алгоритмам кластеризації властива похибка відтвореності. Іншими словами, задовільні результати кластеризації, що отримані на одних даних, не завжди повторюються при застосуванні інших аналогічних даних. Вирішення даної проблеми можливе за рахунок ретельної валідації моделі, при цьому до одних даних застосовуються різні алгоритми кластеризації з подальшим порівнянням результатів із застосуванням зовнішніх критеріїв якості кластеризації даних, які враховують лише розподіл відповідних об'єктів у різних кластеризаціях без оцінки характеру цього розподілу.

Здебільшого існуючі зовнішні критерії якості кластеризації даних засновані на матриці повторюваності, що представляє кількість пар об'єктів, які належать або не належать відповідним кластерам у різних кластеризаціях  $C_1$  та  $C_2$ . У цьому випадку існує чотири можливості:

- *yy* (*yes yes*): два об'єкти належать до одного кластера у кластеризаціях  $C_1$  та  $C_2$ ;
- *yn* (*yes no*): об'єкта належать до одного кластера у кластеризації  $C_1$ , але до різних кластерів у кластеризації  $C_2$ ;
- *ny* (*no yes*): два об'єкта належать до одного кластера у кластеризації  $C_2$ , але до різних кластерів у кластеризації  $C_1$ ;
- *nn* (*no no*): два об'єкта належать до різних кластерів у кластеризаціях  $C_1$ , та  $C_2$ .

Загальна кількість пар об'єктів у цьому випадку може бути визначена таким чином:

$$N_T = \frac{N(N-1)}{2} = yy + yn + ny + nn. \quad (1.17)$$

Формалізація зовнішніх критеріїв у цьому випадку може бути зроблена на закладі похибок першого та другого роду. Припустимо, що кластеризація  $C_1$  є еталонною. Точність розподілу (*precision*) у цьому випадку визначається як частка точок, які розподілені у кластеризації  $C_2$  коректно щодо кластеризації  $C_1$ :

$$Pr = \frac{yy}{yy+ny}. \quad (1.18)$$

Відтворюваність у цьому випадку (*recall coefficient*) визначається як частина пар об'єктів, згрупованих у кластеризаціях  $C_1$ , та  $C_2$  однаково:

$$Rc = \frac{yy}{yy+yn}. \quad (1.19)$$

Особливо поширені зовнішні критерії оцінки якості кластеризації даних, що засновані на визначеннях (1.18) та (1.19), є такі [159]:

- *The Czekanowski-Dice index* визначається як гармонічне середнє точності та відтворюваності:

$$C = \frac{2 \cdot Pr \cdot Rc}{Pr + Rc} = \frac{2yy}{2yy + yn + ny}. \quad (1.20)$$

- *The Folkes-Mallows index* визначається як геометричне середнє точності та відтворюваності:

$$C = \sqrt{Pr \cdot Rc} = \frac{yy}{\sqrt{(yy+yn) \cdot (yy+ny)}}. \quad (1.21)$$

— *The Kulczynski index* визначається як арифметичне середнє точності та відтворюваності:

$$C = \frac{Pr+Rc}{2} = \frac{1}{2} \left( \frac{yy}{yy+ny} + \frac{yy}{yy+yn} \right). \quad (1.22)$$

— *The Hubert index*:

$$C = \frac{N_T \times yy - (yy+yn)(yy+ny)}{\sqrt{(yy+yn)(yy+ny)(nn+yn)(nn+ny)}}. \quad (1.23)$$

— *The Jaccard index*:

$$C = \frac{yy}{yy+yn+ny}. \quad (1.24)$$

— *The MCNemar index*:

$$C = \frac{nn-ny}{\sqrt{nn+ny}}. \quad (1.25)$$

— *The Phi index*:

$$C = \frac{yy \times nn - yn \times ny}{(yy+yn)(yy+ny)(yn+nn)(ny+nn)}. \quad (1.26)$$

— *The Rand index*:

$$C = \frac{yy+nn}{N_T}. \quad (1.27)$$

— *The Rogers-Tanimoto index*:

$$C = \frac{yy+nn}{yy+nn+2(yn+ny)}. \quad (1.28)$$

— *The Russel-Rao index*:

$$C = \frac{yy}{N_T}. \quad (1.29)$$

— *The Sokal-Sneath indexes*:

$$\begin{aligned} C_1 &= \frac{yy}{yy+2(yn+ny)}, \\ C_2 &= \frac{yy+nn}{yy+nn+\frac{1}{2}(yn+ny)}. \end{aligned} \quad (1.30)$$

Головний недолік наведених вище зовнішніх критеріїв якості кластеризації полягає у тому, що їх застосування передбачає використання одних і тих самих даних для різних кластеризацій. При порівняльному аналізі ефективності різних алгоритмів кластерного аналізу це виправдовується, але при валідації моделі, що заснована на конкретному алгоритмі використовуються зазвичай інші дані і застосування критеріїв (1.30) – (1.30) є проблематичним.

Вирішення цієї проблеми можливе на основі застосування зовнішніх критеріїв, які не передбачають використання однакових даних. У працях [160–163] автори запропонували модель вирішення цієї проблеми на основі застосування методів індуктивного моделювання, реалізація якої передбачає послідовний перебір кластеризацій у межах припустимого діапазону з метою вибору найкращої на основі комплексного застосування внутрішніх та зовнішніх критеріїв якості кластеризації.

Якщо припустити, що  $W$  – множина усіх можливих кластеризацій, то оптимальною за відповідними критеріями якості є кластеризація, для якої:

$$K_{opt} = \operatorname{argmin}_{K \in W} QC(K) \text{ або } K_{opt} = \operatorname{argmax}_{K \in W} QC(K), \quad (1.31)$$

де  $QC(K)$  – критерій якості кластеризації, що використовуються для оцінки якості відповідних кластерних структур. У праці [163] авторами введено поняття «критерій балансу» та «об’єктивна кластеризація». Об’єктивною у цьому випадку називається кластеризація, яка за кількістю кластерів, характером розподілу об’єктів або ознак у відповідних кластерах та кількістю невідповідностей найменше відрізняється від експертної. Ця ідея отримала подальший розвиток у роботах [164–179].

В основу методології індуктивного моделювання складних систем покладені три фундаментальні принципи, що запозичені з різних наукових напрямків і які дозволили створити цілісну, органічно взаємопов’язану теорію:

1. Принцип евристичної самоорганізації, тобто послідовного перебору різноманітних ускладнюючих моделей-претендентів для вибору найкращих за доцільно визначеним зовнішнім критерієм або групою критеріїв оцінки якості кластеризації даних.

2. Принцип зовнішнього доповнення, основна ідея якого полягає у необхідності використання «свіжої інформації» для об’єктивної верифікації моделі.

3. Принцип неостаточності рішень, ідея якого полягає у генерації не одного, а певної множини проміжних результатів з подальшим вибором із них найкращих варіантів.

Реалізація принципу зовнішнього доповнення в рамках індуктивної технології об’єктивної кластеризації передбачає наявність зовнішнього критерію оцінки якості кластеризації на двох рівнопотужних підмножинах (містять однакову кількість попарно близьких об’єктів), який може бути комплексним і враховувати як розподіл кластерів в отриманих кластеризаціях, так і розташування об’єктів у відповідних кластерах у різних кластеризаціях. Окрім того, при аналізі складних даних можлива ситуація, коли внутрішні та зовнішні критерії суперечать один одному. Наприклад, кластеризації, що отримані на рівнопотужних підмножинах є однаково незадовільними, при цьому значення зовнішнього критерію досягає локального екстремуму. У цьому випадку виникає необхідність у розрахунку комплексного критерію балансу, який містить у якості компонент внутрішні та зовнішні критерії оцінки якості кластеризації.

Реалізація принципу неостаточності рішень стосовно до індуктивної технології об’єктивної кластеризації передбачає фіксацію кластеризацій, що відповідають локальним екстремумам критерію балансу. Кожний локальний екстремум відповідає об’єктивній кластеризації при певній мірі деталізації. Остаточний вибір і, як наслідок, фіксація кластеризації визначається на підставі поставленої мети задачі, що розв’язується.

На рис. 1.13 зображено окремі модулі моделі об’єктивної кластеризації та характер їх взаємодії [167–169]. Реалізація даної моделі передбачає такі етапи:



Рис. 1.13. Характер взаємодії окремих модулів у індуктивній моделі об'єктивної кластеризації

- визначення функції близькості об'єктів, що досліджуються, тобто метрики, що визначає ступінь подібності профілів експресії генів;
- розбиття вихідної множини об'єктів, що досліджуються, на дві рівнопотужні підмножини. Під рівнопотужними у цьому випадку розуміються підмножини, що містять однакові кількості попарно близьких об'єктів;
- завдання способу утворення кластерів (сортування, перегрупування, об'єднання, поділ тощо);
- визначення критеріїв оцінки якості кластеризації QC, як міри схожості кластерів у різних кластеризаціях;
- організація руху до максимуму або мінімуму значення критерію QC оцінки якості кластеризації;
- визначення способу фіксації об'єктивної кластеризації, що відповідає екстремуму критерію балансу.

У праці [172] авторами наведено порівняльний аналіз різних метрик афінності для оцінки ступеня близькості профілів експресії генів щодо центрів відповідних кластерів. Було доведено, що у випадку застосування профілів експресії генів кореляційна метрика має суттєво більш високу роздільну здатність, порівняно з іншими метриками (евклідовою та манхеттенською).

Як внутрішній критерій якості кластеризації, авторами запропонований комплексний критерій, що розраховувався як мультиплікативна комбінація критеріїв Calinski-Harabasz [139] та WB-index [173]. Зовнішній критерій розраховувався як нормалізована різниця між значеннями внутрішніх критеріїв, отриманих на відповідному рівні кластеризації даних на рівнопотужних підмножинах. Комплексний критерій балансу розраховувався на основі застосування функції бажаності Харрінгтона [174], максимальне значення якої відповідало оптимальній кластеризації за обраними критеріями.

Запропонована модель об'єктивної кластеризації була практично реалізована на основі алгоритмів кластеризації *k-means* [170], агломеративного ієрархічного та густотних алгоритмів DBSCAN та OPTICS [171] та самоорганізуючого алгоритму SOTA [166]. Проведені дослідження дозволили авторам запропонувати гібридну модель кластер-бікластерного аналізу для вибору найбільш інформативних профілів експресії генів для подальшої реконструкції генної регуляторної мережі [174]. Застосування даної моделі передбачає покрокову двохетапну кластеризацію даних із застосуванням алгоритму DBSCAN на першому етапі для видалення генів, що ідентифікуються як шум. Далі до профілів експресії генів застосовується алгоритм кластеризації SOTA з кореляційною метрикою. Як результат, отримуються дві приблизно рівних підмножини профілів експресії генів, до яких на остаточному етапі застосовуються бікластерним аналіз, результатом якого є бікластери взаємно корельованих профілів експресії генів та об'єктів. Оптимальні параметри відповідних алгоритмів кластеризації визначалися на підставі індуктивної моделі об'єктивної кластеризації.

Однак, слід зазначити, що незважаючи на певні досягнення у даному напрямку, запропоновані моделі не є досконалими і потребують доопрацювання. Передусім слід зазначити високий відсоток суб'єктивізму у виборі критеріїв оцінки якості формування кластерної структури. Окрім того, окремі алгоритми при їх оптимальних параметрах за обраними критеріями можуть давати різні результати. У цьому випадку є доцільним застосування ансамблів алгоритмів з більш ретельною верифікацією отриманих моделей.

### **1.3. Аналіз сучасних методів та інформаційних технологій редукції профілів експресії генів**

Як було зазначено у попередніх двох підрозділах, методи кластеризації та бікластеризації дозволяють виділити взаємно корельовані профілі експресії генів, але їх застосування супроводжується високим відсотком суб'єктивізму за рахунок недосконалості відповідних критеріїв якості. До того ж при цьому може бути втрачена корисна інформація при видаленні або незастосуванні профілів експресії генів, що несуть значущу інформацію щодо стану об'єкта, що досліджується. У цьому випадку доцільним є використання ансамблю методів інтелектуального аналізу даних та машинного навчання для аналізу та виділення інформативних профілів експресії генів.

Останнім часом проблемі оцінки рівня інформативності профілів експресії генів з метою їх редукції присвячено велику кількість наукових праць. Так, у праці [175] представлено результати досліджень щодо виділення профілів експресії генів молекул міРНК. Шляхом парного порівняння було проведено скринінг диференційовано експресованих міRNA. В процесі моделювання було виділено у середньому з кожної бібліотеки 2744989 необроблених зчитувань із 9888123. Всього було виявлено 2565 міРНК. У праці [176] автори зробили порівняльний аналіз різних методів класифікації даних експресії генів для вибору найбільш інформативних профілів експресії генів із застосуванням похибок першого та другого роду. До

основних недоліків цього підходу є те, що у реальних умовах ми не можемо напевно знати клас, до якого належить той або інший ген, тобто ця методика має великий відсоток суб'єктивізму.

У огляді [177] авторами проведено порівняльний аналіз сучасних гібридних моделей вибору інформативних профілів експресії генів для вирішення проблеми подальшої класифікації об'єктів, що досліджено на різні типи раку. Виконано аналіз ефективності різних моделей редукції генів від фільтрації шляхом оцінки рівня інформативності відповідного гена на основі статистичного аналізу до комплексного застосування кластерного аналізу та методів класифікації даних. Основним критерієм в усіх випадках було застосовано критерії оцінки якості класифікатора. При цьому було проаналізовано різні комбінації методів інтелектуального аналізу даних та машинного навчання.

Гібридні методи вибору інформативних генів на основі генетичного алгоритму представлені у [178–182]. Так, у роботі [178] представлено результати досліджень з розробки гібридного алгоритму вибору ознак на основі комплексного застосування методу максимізації взаємної інформації (Mutual Information Maximization (MIM)) та адаптивного генетичного алгоритму (AGA). Запропонований гібридний метод був названий MIMAGA-Selection. На першому етапі обиралися профілі експресії генів з високим значенням взаємної кореляції шляхом застосування MIM методу. Максимальна кількість генів при цьому була 300. На другому етапі до обраних генів застосовувався адаптивний генетичний алгоритм AGA. Для тестування запропонованого алгоритму було використано шість наборів даних про експресію мульти- та бінарних генів раку. Класифікація даних виконувалася нейронною мережею прямого поширення (МЕН – машина екстремального навчання). У процесі моделювання було проведено 30 повторень процесу класифікації. Для порівняння точності алгоритму MIMAGA-Selection автори застосували три наявні алгоритми, а саме: послідовний прямий відбір (SFS), ReliefF та MIM з класифікатором ELM на одному наборі даних з однаковим номером цільового гена. Результати моделювання показали більш високу ефективність запропонованого MIMAGA методу, порівняно з іншими методами, що використовувалися. Окрім того, автори провели класифікацію генів у рамках MIMAGA-Selection методу із застосуванням чотирьох різних класифікаторів; нейронну мережу зворотного поширення (BP), машину опорних векторів (SVM), ELM та регуляризовану машину екстремального навчання (RELM). Всі чотири класифікатори досягали точності вище ніж 80%.

У роботі [179] запропоновано гібридну модель вибору профілів експресії генів на основі комплексного застосування генетичного алгоритму із встановленням динамічних параметрів (GADP) та статистичного  $\chi^2$ -тесту на однорідність. Запропонований підхід дозволяє автоматизувати вибір кількості профілів експресії генів без втручання людини. Оцінка ефективності моделі виконувалася шляхом застосування класифікатора SVM. Для вимірювання ефективності запропонованого методу, порівняно з наявними методами, було використано шість наборів даних про рак. Результати показали високу точність запропонованої моделі. Для деяких даних точність сягала 100% при меншій кількості генів, що застосовувалися в процесі класифікації об'єктів.

У роботі [180] запропоновано гібридну модель відбору генів, що заснована на комплексному використанні кореляційного аналізу (Correlation-based Feature Selection (CFS)) та генетичного алгоритму (Taguchi-Genetic Algorithm (TGA)). Реалізація процесу відбору генів передбачала два етапи. На першому етапі видалялися профілі експресії генів з малим значенням взаємної кореляції, порівняно з іншими генами із застосуванням CFS методу. На наступному кроці до генів, що залишилися, застосовувався алгоритм TGA з відповідними функціями, які були сформовані на етапі фільтрації. Класифікатор К-найближчого сусіда (KNN) використовувався для оцінки ефективності запропонованого методу з точки зору точності класифікації. Було оцінено одинадцять наборів даних експресії генів, що досліджувалися на рак. Результати моделювання показали високу ефективність запропонованого підходу, оскільки висока точність класифікації була досягнута у 10 наборах даних з 11, при цьому для шістьох наборів даних точність досягла 100%.

У роботі [181] представлено результати досліджень з розробки гібридної моделі відбору генів з даних, що отримані шляхом ДНК-мікрочіпового експерименту, на основі комплексного застосування генетичного алгоритму GA та методу інтелектуального динамічного генетичного алгоритму (Intelligent Dynamic Genetic Algorithm (IDGA)). Реалізація моделі передбачала два етапи. На першому етапі для відбору 500 найкращих генів незалежно один від другого застосовувалися методи фільтрації Лапласіана та Фішера. На другому етапі було застосовано метод IDGA. Для оцінки ефективності запропонованої моделі було використано п'ять наборів даних пацієнтів, що досліджувалися на рак. Як класифікатори використовували машину опорних векторів (SVM), наївний байесівський класифікатор (NB) та метод  $k$ -найближчих сусідів (KNN). IDGA алгоритм було застосовано сім разів. На основі аналізу результатів моделювання запропонований метод дозволив отримати точність 100% на чотирьох наборах даних. До того ж застосування IDGA з методом фільтрації Фішера дозволив отримати кращі результати, порівняно з методом фільтрації Лапласіана, на чотирьох наборах даних.

Гібридний метод вибору профілів експресії генів на основі комплексного застосування інформаційний приріст (Information Gain (IG)) та стандартного генетичного алгоритму (SGA), запропоновано у роботі [182]. Цей метод отримав назву IG/SGA. На першому кроці для зменшення об'єктів застосовувався IG метод приросту інформації. На другому етапі до отриманих даних застосовувався генетичний алгоритм. Класифікацію проводили за допомогою методу генетичного програмування GP. Було використано сім наборів даних мікрочіпів пацієнтів, що досліджувалися на рак. Результати показали, що використання запропонованого методу дозволило досягнути 100% точності класифікації для двох наборів даних.

Гібридні моделі на основі застосування алгоритму оптимізації мурашиної колонії (Ant Colony Optimization (ACO)) представлені у працях [183, 184]. Метод, наведений у праці [183], поєднує в собі автомати клітинного навчання та оптимізацію колонії мурашок (CLA-ACO). Практична реалізація моделі передбачає три етапи. На першому кроці дані фільтруються за методом Фішера. На другому етапі формувалася підмножина профілів експресії генів з оптимальним набором

функцій комплексним застосуванням автоматів клітинного навчання та мурашкового алгоритму. Третій етап полягав у остаточному формуванні підмножини профілів експресії генів на основі розрахунку та аналізу похибок першого та другого роду із застосуванням ROC аналізу. Оцінка ефективності запропонованої моделі проводилася двома етапами. На першому етапі проводився статистичний аналіз профілів експресії генів із застосуванням різних методів з подальшим ранжуванням даних та відбору найбільш інформативних. На другому етапі до даних застосовувалися метод CLA-ACO з подальшим розрахунком кількісних критеріїв якості відповідної моделі. Подальше тестування показало високу ефективність запропонованої моделі для класифікації об'єктів, що досліджувалися.

У роботі [184] запропоновано гібридну модель відбору профілів експресії генів на основі комплексного застосування мурашкових алгоритмів та адаптивної моделі оптимізації стовбурових клітин (Adaptive Stem Cell Optimization (ASCO)). Застосування цієї моделі передбачало фільтрацію профілів експресії генів на основі рівня їх взаємної кореляції. Оцінка ефективності моделі проводилася шляхом застосування різних класифікаторів. Для тестування були обрані дані експресії генів, що отримані шляхом проведення ДНК-мікрочіпових експериментів. Результати тестування також показали високу ефективність запропонованої гібридної моделі.

Порівняльний аналіз сучасних гібридних методів та моделей формування підмножин інформативних профілів експресії генів для подальшої реконструкції генних регуляторних мереж або створення систем діагностики захворювань наведений у таблиці 1.3 [177].

*Таблиця 1.3*

*Порівняльний аналіз гібридних моделей формування підмножин інформативних профілів експресії генів*

| По-сил. | Метод фільтрації                          | Метод групування       | Класифікатор | Дані            | Точність, % | Кількість генів |
|---------|-------------------------------------------|------------------------|--------------|-----------------|-------------|-----------------|
| 1       | 2                                         | 3                      | 4            | 5               | 6           | 7               |
| [178]   | Mutual Information Maximization (MIM)     | Genetic Algorithm (GA) | SVM          | Colon [196]     | 83.41       | 202             |
| [179]   | $\chi^2$ -test                            | Genetic Algorithm (GA) | SVM          | Colon [196]     | 100         | 8               |
|         |                                           |                        |              | DLBCL [197]     | 100         | 6               |
|         |                                           |                        |              | SRBCT [198]     | 100         | 8               |
|         |                                           |                        |              | Leukemia1 [199] | 100         | 5               |
| [180]   | Correlation-based Feature Selection (CFS) | Genetic Algorithm (GA) | KNN          | SRBCT [198]     | 100         | 29              |
|         |                                           |                        |              | Prostate [201]  | 99,2        | 24              |
|         |                                           |                        |              | Lung [200]      | 98,42       | 195             |

Продовження табл. 1.3

| 1     | 2                                           | 3                                 | 4                | 5               | 6     | 7    |
|-------|---------------------------------------------|-----------------------------------|------------------|-----------------|-------|------|
| [181] | Laplacian and Fisherscore                   | Genetic Algorithm (GA)            | SVM              | SRBCT [198]     | 100   | 18   |
|       |                                             |                                   |                  | Leukemia1 [199] | 100   | 15   |
|       |                                             |                                   |                  | Prostate [201]  | 96,3  | 14   |
|       |                                             |                                   |                  | Breast [202]    | 100   | 2    |
|       |                                             |                                   |                  | DLBCL [158]     | 100   | 9    |
|       |                                             |                                   | KNN              | SRBCT [197]     | 91,6  | NA   |
|       |                                             |                                   |                  | Leukemia1 [199] | 97,2  | NA   |
|       |                                             |                                   |                  | Prostate [201]  | 95,6  | NA   |
|       |                                             |                                   |                  | Breast [202]    | 95,5  | NA   |
|       |                                             |                                   |                  | DLBCL [197]     | 97,9  | NA   |
|       |                                             |                                   | NB               | SRBCT [198]     | 98,2  | NA   |
|       |                                             |                                   |                  | Leukemia1 [199] | 93,1  | NA   |
|       |                                             |                                   |                  | Prostate [201]  | 93,4  | NA   |
|       |                                             |                                   |                  | Breast [202]    | 100   | NA   |
|       |                                             |                                   |                  | DLBCL [197]     | 95,8  | NA   |
| [182] | Information Gain (IG)                       | Genetic Algorithm (GA)            | GP               | Colon [196]     | 85,48 | 60   |
|       |                                             |                                   |                  | Leukemia1 [199] | 97,06 | 3    |
|       |                                             |                                   |                  | Lung [200]      | 100   | 9    |
|       |                                             |                                   |                  | Prostate [201]  | 100   | 26   |
| [183] | Fisher Criterion                            | Ant Colony Optimization (ACO)     | SVM              | Leukemia1 [199] | 95,5  | 3202 |
|       |                                             |                                   |                  | Prostate [201]  | 98,35 | 14   |
|       |                                             |                                   | KNN              | Leukemia1 [199] | 94,30 | 3    |
|       |                                             |                                   |                  | Prostate [201]  | 99,25 | 15   |
|       |                                             |                                   | NB               | Leukemia1 [201] | 99,95 | 4    |
|       |                                             |                                   |                  | Prostate [201]  | 99,40 | 10   |
| [184] | Mutual Information (MI)                     | Ant Colony Optimization (ACO)     | Fuzzycl assifier | Colon [196]     | 100   | NA   |
|       |                                             |                                   |                  | Leukemia1 [199] | 100   | NA   |
|       |                                             |                                   |                  | Prostate [201]  | 90,85 | NA   |
| [186] | Fisher Criterion                            | Bat Algorithm (BA)                | SVM              | SRBCT [198]     | 85    | 6    |
|       |                                             |                                   |                  | Prostate [201]  | 94,1  | 6    |
|       |                                             |                                   | KNN              | SRBCT [198]     | 100   | 6    |
|       |                                             |                                   |                  | Prostate [201]  | 97,1  | 6    |
|       |                                             |                                   | NB               | SRBCT [198]     | 100   | 6    |
|       |                                             |                                   |                  | Prostate [201]  | 97,1  | 6    |
| [186] | Independent Component Analysis (ICA)        | Artificial Bee Colony (ABC)       | NB               | Colon [157]     | 98,14 | 16   |
|       |                                             |                                   |                  | Leukemia1 [160] | 98,68 | 12   |
|       |                                             |                                   |                  | Leukemia2 [164] | 97,33 | 15   |
|       |                                             |                                   |                  | Lung [161]      | 92,45 | 24   |
| [187] | Minimum redundancy Maximum Relevance (mRMR) | Artificial Bee Colony (ABC)       | SVM              | Colon [196]     | 96,77 | 15   |
|       |                                             |                                   |                  | SRBCT [198]     | 100   | 10   |
|       |                                             |                                   |                  | Leukemia1 [199] | 100   | 14   |
|       |                                             |                                   |                  | Leukemia2 [203] | 100   | 20   |
|       |                                             |                                   |                  | Lung [200]      | 100   | 8    |
|       |                                             |                                   |                  | Lymphoma [204]  | 100   | 5    |
| [188] | Probabilistic Random Function               | Particle Swarm Optimization (PSO) | KNN              | Colon [196]     | 84,38 | 60   |
|       |                                             |                                   |                  | SR207 [198]     | 89,28 | 100  |
|       |                                             |                                   |                  | Lymphoma [204]  | 87,71 | 50   |

Продовження табл. 1.3

| 1     | 2                                           | 3                                                 | 4                  | 5               | 6     | 7    |
|-------|---------------------------------------------|---------------------------------------------------|--------------------|-----------------|-------|------|
| [189] | Correlation-based Feature Selection (CFS)   | Particle Swarm Optimization (PSO)                 | NB                 | Colon [196]     | 94,89 | 4    |
|       |                                             |                                                   |                    | SRBCT [198]     | 100   | 34   |
|       |                                             |                                                   |                    | Leukemia1 [199] | 100   | 4    |
|       |                                             |                                                   |                    | Leukemia2 [203] | 100   | 6    |
|       |                                             |                                                   |                    | Lymphoma [204]  | 100   | 24   |
|       |                                             |                                                   |                    | MLL [205]       | 100   | 30   |
|       |                                             |                                                   |                    | Breast [202]    | 100   | 10   |
| [190] | Random Forest Ranking (RFR)                 | Black Hole Algorithm (BHA)                        | Bagging Classifier | Colon [196]     | 91,93 | 3    |
|       |                                             |                                                   |                    | MLL [205]       | 98,61 | 5    |
| [191] | Fisher-Markov selector                      | Biogeography algorithm                            | SVM                | SRBCT [198]     | 100   | 6    |
|       |                                             |                                                   |                    | Prostate [201]  | 98,3  | 12   |
|       |                                             |                                                   |                    | Lung [200]      | 98,4  | 16   |
| [192] | Symmetrical Uncertainty (SU)                | Harmony Search Algorithm (HSA)                    | NB                 | Colon [196]     | 87,53 | 9    |
|       |                                             |                                                   |                    | SRBCT [198]     | 99,89 | 37   |
|       |                                             |                                                   |                    | Leukemia1 [199] | 100   | 26   |
|       |                                             |                                                   |                    | Leukemia2 [203] | 100   | 24   |
|       |                                             |                                                   |                    | Lymphoma [204]  | 100   | 10   |
|       |                                             |                                                   |                    | MLL [205]       | 98,97 | 10   |
| [193] | Logarithmic Transformation                  | Grasshopper-Optimization-Algorithm                | NB                 | Colon [196]     | 95    | NA   |
|       |                                             |                                                   |                    | Leukemia1 [199] | 94    | NA   |
| [194] | Fast Correlation-Based Filter (FCBF)        | Particle Swarm Optimization and Genetic Algorithm | SVM                | Colon [196]     | 96,30 | 1000 |
|       |                                             |                                                   |                    | DLBCL [197]     | 100   | 3204 |
| [195] | Minimum redundancy Maximum Relevance (mRMR) | Genetic Bee Colony and Genetic Algorithm          | SVM                | Colon [196]     | 98,38 | 10   |
|       |                                             |                                                   |                    | SRBCT [198]     | 100   | 6    |
|       |                                             |                                                   |                    | Leukemia1 [199] | 100   | 4    |
|       |                                             |                                                   |                    | Leukemia2 [203] | 100   | 8    |
|       |                                             |                                                   |                    | Lung [200]      | 100   | 4    |
|       |                                             |                                                   |                    | Lymphoma [204]  | 100   | 4    |

На основі аналізу літературних джерел та даних таблиці 1.3 можна зробити висновок, що проблема об'єктивного вибору інформаційних з точки зору роздільної здатності генів, які дозволяють розпізнати об'єкти, що досліджуються не має однозначного вирішення. Здебільшого висока точність класифікації досягається при застосуванні малої кількості найбільш інформативних генів. Як було зазначено вище, вихідна кількість активних генів у геномі людини дорівнює приблизно 25000. Видалення генів з малою експресією дозволяє скоротити цю кількість до 7–10 тис. Як видно з таблиці 1.3, тільки у декількох випадках виділена кількість генів перевищує 100 при високій точності класифікатора. Мала кількість генів не дозволяє реконструювати якісну генну регуляторну мережу, що дозволить дослідити характер взаємодії елементів системи при різних її станах для визначення групи генів, які визначають відповідну хворобу на першому етапі, визначення характеру їх взаємодій один з одним та з іншими елементами системи на другому етапі, і

визначення характеру впливу зміни експресії даних генів на інші гени мережі в процесі подальшого моделювання мережі на третьому етапі. До суттєвих недоліків методів, наведених у таблиці 1.3 слід віднести також те, що у більшості випадків параметри відповідних алгоритмів визначалися емпіричним шляхом у процесі моделювання, тобто моделі не є самоорганізуючими. Цей факт вносить великий відсоток суб'єктивізму в процес прийняття остаточного рішення з виділення генів. Вирішення цієї проблеми можливо шляхом застосування ансамблю відповідних методів з подальшим прийняттям рішення на основі аналізу кількісних критеріїв якості, що відповідають цьому етапу.

Часткове вирішення цієї проблеми представлено у працях [206–208]. Автори запропонували гібридну модель виділення профілів експресії генів на основі комплексного застосування кількісних статистичних критеріїв та інформаційної ентропії Шеннона, алгоритму кластеризації SOTA та ансамблю бінарних класифікаторів. При цьому остаточне рішення приймалося на основі аналізу результатів системи нечіткого логічного висновку. Оптимальні параметри алгоритму кластеризації SOTA були визначені заздалегідь на основі запропонованої авторами індуктивної моделі об'єктивної кластеризації [166–172]. Блок-схему запропонованої моделі зображено на рис. 1.14. Практична реалізація цієї моделі передбачає такі етапи:

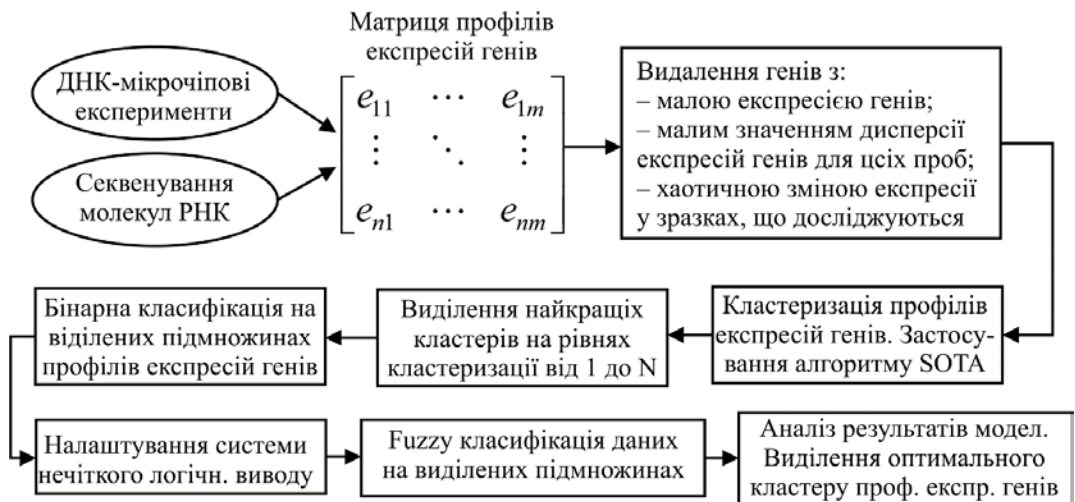


Рис. 1.14. Блок-схема покрокової процедури виділення підмножини інформативних взаємно корельованих генів

1. Видалення неінформативних генів за статистичними критеріями та ентропією Шеннона. Передбачалося, що якщо середнє абсолютне значення експресії профіля гену та значення дисперсії менші а значення інформаційної ентропії Шеннона більші відповідних граничних значень, даний ген видалявся з бази даних як неінформативний. Граничні значення відповідних параметрів визначалися емпіричним шляхом у процесі моделювання. Для цього множина об'єктів, що досліджувалася, розбивалася на групи залежно від класу, до якого належить об'єкт

(хворий чи ні або стан хвороби). На кожному кроці аналізу та видалення генів розраховувався критерій якості групування об'єктів, який враховував як щільність розподілу об'єктів у окремих групах, так і щільність розподілу центрів окремих груп у просторі ознак. Граничні значення відповідали екстремуму даного критерія якості.

2. Покрокова кластеризація профілів експресії генів на ієрархічних рівнях від 1 до  $N$  із застосуванням алгоритму кластеризації SOTA з кореляційною метрикою оцінки відстані між профілями експресії генів. Результати моделювання показали, що на одному кроці застосування цього алгоритму дані розділяються на дві приблизно однакові підмножини. Тобто кількість кластерів змінюється при цьому від 2-х до  $2^N$  на першому та  $N$ -му ієрархічних рівнях, відповідно. Розрахунок критерію якості групування об'єктів у кластерах на кожному ієрархічному рівні.

3. Виділення найкращих підмножин профілів експресії генів на кожному ієрархічному рівні, які відповідають екстремуму критерію якості групування даних у кластерах.

4. Застосування ансамблю класифікаторів до виділених кластерів: GLM, SVM, CART та RF. Розрахунок кількісних оцінок ефективності кожного із класифікаторів (Accuracy (AC), F-measure (F), Matthews correlation coefficient (MCC)). Формування проміжних рішень щодо результатів класифікації об'єктів на кожному ієрархічному рівні для кожного методу класифікації.

5. Налаштування системи нечіткого логічного висновку, вибір алгоритму виводу, формування функцій належності нечітких множин та бази правил.

6. Нечітка класифікація об'єктів із розрахунком кількісних показників якості. Аналіз отриманих результатів.

Моделювання проводилося на даних ДНК-мікрочіпових експериментів пацієнтів, що досліджувалися на ранній стадії раку легенів [209]. Початкові дані містили 156 об'єктів (мікрочіпів) та 54675 генів. Застосування моделі видалення неінформативних генів дозволило зменшити кількість генів до 21431 (було видалено 33244 генів). У результаті застосування дальших кроків моделі було виділено 401 генів, що дозволяють отримати максимальну точність класифікації за групою методів (96,5%). За думкою авторів, дані профілі експресії генів можуть у подальшому бути використані для реконструкції генної мережі для дослідження характеру взаємодії генів залежно від стану здоров'я пацієнта.

Однак слід зазначити, що основним недоліком запропонованої моделі є те, що дані містять тільки два стани (класи), що обмежує її застосування для мульти-класових даних. Більш того, модель потребує валідації із застосуванням інших типів даних експресії генів. Перший етап моделі також має відсоток суб'єктивізму, оскільки розбиття об'єктів на заздалегідь відомі кластери не дозволяє об'єктивно оцінити результат класифікації об'єктів у реальних умовах.

#### 1.4. Огляд сучасних методів машинного навчання для обробки даних експресії генів

Як зазначалося вище, до особливостей даних на основі профілів експресії генів слід віднести високу розмірність, яка визначається кількістю активних генів, що дозволяють ідентифікувати стан відповідного об'єкта. Створення системи діагностики на основі застосування великої кількості генів не є доцільним, оскільки початкові дані містять досить велику шумову компоненту (гени, експресія яких не є визначальною для ідентифікації відповідної хвороби), що може вплинути на точність і чутливість моделі діагностики. З іншого боку, застосування повної матриці експресії генів потребує великої кількості комп'ютерних та часових ресурсів для обробки інформації в процесі налаштування моделі. У монографії розглянуті та запропоновані нові методи виділення інформативних профілів експресії генів, враховуючи хворобу, що досліджується, але у будь-якому випадку розмірність даних експресії генів залишається високою, що обмежує застосування наявних класифікаторів для ідентифікації об'єктів на основі даних експресії генів.

Як добре відомо, задача класифікації являє собою підгалузь машинного навчання, реалізація якої передбачає навчання комп'ютерів без явного програмування (навчання з учителем) [210]. Існує велика кількість різних класичних методів класифікації даних, а саме: метод найближчого сусіда (K-Nearest-Neighbor (KNN)), метод опорних векторів (Support Vector Machine (SVM)), дерева рішень (Decision Tree (DT)), випадковий ліс (Random Forest (RF)), класифікатори на основі методу Байєса (Bayes), нейронні мережі (Neural Network (NN)) тощо. Загалом методи машинного навчання можна розділити на три основні класи: методи на основі контрольованого навчання, навчання без контролю та навчання з підкріпленням. Контрольоване навчання передбачає застосування підмножини даних із заздалегідь відомими класами, при цьому зіставляються вхідні дані (атрибути об'єктів, що досліджуються) з відповідними виходами (класами) для отримання функціональної залежності між входами та виходами, яка найкращим чином відображає залежність між входом та виходом.

Неконтрольоване навчання не передбачає маркування даних. Як зазначалося у попередніх підрозділах монографії, алгоритми на основі неконтрольованого навчання (алгоритми кластеризації) передбачають виявлення скритої інформації у даних з подальшим групуванням даних на підкласи на основі відповідних мір подібності. При цьому виникає необхідність вибору мір подібності між об'єктами, об'єктами і кластерами, та кластерами. Більш того, при наявності великих даних виникає проблема вибору кількості кластерів, що вирішується із застосуванням критеріїв якості кластеризації, які є чутливими до типу експериментальних даних.

Навчання з підкріпленням зазвичай застосовується для обох типів даних (маркованих і немаркованих). Реалізація алгоритмів навчання з підкріпленням передбачає послідовне підлаштування моделі, при цьому на кожному кроці вихід залежить від поточного входу, а вхід на наступній ітерації визначається вихід на попередній ітерації. Моделі на основі навчання з підкріпленням зазвичай засновані на максимізації поняття кумулятивної винагороди.

Традиційні підходи щодо оцінки ефективності класифікатора при розв'язанні мультикласових задач із застосуванням даних експресії генів передбачають розділення початкової кількості об'єктів як мінімум на дві групи у приблизному відношенні 70/30%. При цьому перша множина об'єктів застосовується для навчання моделі, а друга – для тестування. Розрахунок критеріїв якості ідентифікації об'єктів, а саме точність роботи моделі (Accuracy (ACC)), точність ідентифікації позитивних випадків (Precision), чутливість (Sensitivity or Recall), F-міри (F-measure) тощо, зазвичай здійснюється на тестовому наборі даних, які не використовувалися при навчанні моделі. Точність класифікації даних розраховується як відношення коректно ідентифікованих об'єктів до загальної кількості об'єктів у тестовій базі даних [211]:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}, \quad (1.32)$$

де  $TP$  (True Positive) і  $FP$  (False Positive) є коректно та хибно ідентифіковані позитивні випадки (наявність відповідної хвороби);  $TN$  (True Negative) і  $FN$  (False Negative) є коректно та хибно ідентифіковані негативні випадки (відсутність відповідної хвороби).

Влучність (Precision (PR)) визначається як точність ідентифікації позитивних випадків або середня ймовірність ідентифікації відповідної хвороби (релевантного пошуку) [211]:

$$PR = \frac{TP}{TP + FP}. \quad (1.33)$$

Повнота (Recall or Sensitivity) визначає середню ймовірність коректно ідентифікованих істинно позитивних випадків:

$$RC = \frac{TP}{TP + FN}. \quad (1.34)$$

F-міра у контексті оцінки точності класифікатора визначається як середнє гармонічне влучності та повноти [211]:

$$F = \frac{2 \cdot PR \cdot RC}{PR + RC}. \quad (1.35)$$

Значення як точності, так і F-міри змінюється в інтервалі від 0 (абсолютно незадовільна класифікація) до 1 (всі об'єкти ідентифіковані коректно).

У роботі [212] автори запропонували для класифікації даних експресії генів метод екстремального навчання (extremal learning machine (ELM)), при цьому для вибору інформативних профілів експресії генів застосовувався кореляційний аналіз. Автори отримали результати точності бінарної класифікації на рівні 79%. У дослідженні було використано 60 пухлин центральної нервової системи (при цьому дані містили також інші типи пухлин). Однак слід зазначити, що при експерименті використовувався невеликий набір даних, що може впливати на адекватність отриманих результатів. Крім того, ефективність запропонованої моделі оцінювалася на основі лише одного набору даних, що також не сприяє адекватності отриманих

результатів. І точність отриманих результатів також не є високою, що, без сумніву, є одним із недоліків отриманих результатів досліджень.

У роботі [213] представлено результати досліджень стосовно застосування п'яти методів машинного навчання для класифікації даних експресії генів, а саме: алгоритм випадкового лісу, метод опорних векторів, наївний байєсівський класифікатор, дерева рішень C4.5 та метод  $k$ -найближчого сусіда. У цьому дослідженні автори застосували дані експресії генів з різним рівнем соматичних мутацій 358 пацієнтів, які були отримані з бази даних TCGA — *The Cancer Genome Atlas* (Атлас Геному Раку) і використані для прогнозування раку молочної залози. Найвища точність, що була досягнута в цьому дослідженні, становила 70% при використанні алгоритму випадкового лісу, тоді як інші методи машинного навчання, використані в дослідженні, були менш точними.

У роботі [214] для класифікації різних підтипів раку легенів було застосовано наївний байєсівський класифікатор та два методи відбору ознак (Relief/Limma). У цьому дослідженні використовувався набір даних експресії генів GDS3257 і дані метилування ДНК з бази даних TCGA. 19 генів при цьому були пов'язані з аденокарциномою, плоскоклітинним раком і карциномою легенів. Відібрані гени були використані як атрибути для прогнозування підтипів раку легенів. Точність запропонованої моделі становила 89% для обох типів даних (метилування ДНК GDS3257 і дані з бази TCGA). Однак слід зазначити, що отримані результати також не є об'єктивними, оскільки вони не підтверджені на інших аналогічних даних.

Результати досліджень стосовно застосування чотирьох добре відомих методів машинного навчання до чотирьох різних типів даних експресії генів для прогнозування раку легенів представлені у роботі [215]. Набори даних були отримані від Гарвардської медичної школи і склалися з 203 зразків, що містили експресію 12600 генів; Мічиганський університет (192 зразки та 7129 генів); Університет Торонто (63 зразки та 2880 генів) і Brigham and Women's Hospital (181 зразок та 12533 генів). Аналіз результатів досліджень показав, що максимальна точність класифікації зразків варіювалася в інтервалі від 88% до 94% при застосуванні набору даних Brigham and Women's Hospital і методу опорних векторів. Точність класифікації об'єктів при застосуванні інших наборів даних і інших методів класифікації була значно нижчою, що також свідчить про обмеженість запропонованих методів.

У роботі [216] автори застосували класифікатори на основі алгоритму випадкового лісу та методу опорних векторів для класифікації двох підтипів раку легенів: (аденокарциноми (AC) і плоскоклітинний рак (SCC)). У цьому дослідженні вони також застосували методи Монте-Карло (MCSF) і інкрементальний відбір ознак для ідентифікації інформативних генів. Масиви даних експресії генів компанії Affymetrix U133 (20502 генів), що містили 77 зразків раку легенів AC та 73 зразків раку легенів SCC з Gene Expression Omnibus бази даних (GEO GSE43580) були використані в процесі моделювання. Аналіз результатів моделювання дозволив зробити висновок, що при відборі 1100 оптимальних ознак (генів) для класифікації за допомогою класифікатора SVM була досягнута більша точність, порівняно з

використанням 43 інформативних ознак (генів), отриманих методом MCSF. Точність класифікації при цьому зменшилася з 0,96 до 0,86 при застосуванні методу опорних векторів та з 0,93 до 0,88 при застосуванні методу випадкового лісу. Цей факт свідчить про важливість кроку вибору інформативних профілів експресії генів.

Результати досліджень щодо обробки даних експресії генів, отриманих шляхом застосування методу ДНК мікрочіпів із зразків аденокарциноми легенів (86 зразків пухлин і 10 зразків без пухлин, зібраних із репозиторію біомедичних даних Кент-Ріджа, доступного з [217] з 7129 генами), представлені у [218]. Для формування підмножини інформативних генів було застосовано метод оцінки взаємної інформації, при цьому 70% зразків використовувалися для навчання моделі, а 30% зразків для тестування. У цьому дослідженні було застосовано три методи бінарної класифікації для ідентифікації наявності пухлини після вибору генів-кандидатів, які мали безпосереднє відношення до раку легенів. Кілька відібраних генів були оцінені на предмет біологічної значущості в патології раку легенів. За результатами досліджень було виділено шість генів з високими показниками приросту інформації, які можуть бути пов'язані з раком легенів. Окрім того, у запропонованій моделі ці гени використовувалися як гени біомаркери для класифікації раку легенів за допомогою трьох методів класифікації (багатошаровий перцептрон (MLP), випадковий підпростір (RSS) і (послідовна мінімальна оптимізація (SMO)). Досягнута точність класифікації становила 87%, 68% і 92% при застосуванні класифікаторів. MLP, RSS і SMO, відповідно. Слід зазначити, що отримана точність класифікації зразків у цьому випадку також не є задовільною, що можна пояснити недосконалістю моделі формування підмножини інформативних профілів експресії генів.

У роботі [219] автори запропонували класифікатор багатозернистий каскадний ліс (gcForest) і стратегію вибору інформативних ознак для прогнозування чотирьох підтипів раку молочної залози. У процесі моделювання були використані дані експресії генів RNA-Seq з атласу геному раку TCGA, при цьому запропоновано метод вибору інформативних генів, що використовувалися для підвищення точності класифікації та скорочення часу навчання. Ефективність запропонованого класифікатора оцінювалася шляхом порівняння результатів класифікації даних при застосуванні класифікатору gcForest із трьома іншими добре відомими класифікаторами (KNN, SVM та MLP). При цьому на попередньому етапі були відібрані 30 найбільш інформативних генів. Аналіз результатів моделювання показав більш високу ефективність запропонованого алгоритму gcForest, порівняно з іншими класифікаторами. У цьому дослідженні була досягнута точність 0,92. Однак слід зазначити, що суттєвим недоліком алгоритму gcForest є те, що він працює з дискретними даними, тобто значення експресії генів необхідно дискретизувати, що призводить до втрати інформації і як наслідок, до зменшення точності класифікації даних. Крім того, представлені дослідження не використовували зовнішні дані для оцінки адекватності запропонованої моделі.

Останнім часом із розширенням комп'ютерних ресурсів та можливостей їх застосування для обробки даних експресії генів широке розповсюдження отримали методи штучного інтелекту на основі глибинного навчання (Deep Learning (DL)).

DL можна визначити як галузь машинного навчання, яка працює шляхом генерації структури, що має кілька рівнів, у яких наступний рівень вводу даних є виходом попереднього рівня (рис. 1.15) [220].

Структура мережі DL має на меті імітувати механізми людського мозку шляхом інтерпретації різних типів даних, включаючи звук, текст і зображення [221]. DL використовує принципи, що подібні до принципів лінійної регресії, де кожен нейрон має зважене значення, оновлене шляхом застосування алгоритму градієнтного спуску через зворотне поширення для зменшення функції глобальної втрати [222]. Підходи DL сприяли подоланню труднощів прогнозування раку шляхом прискорення аналізу при високій точності ідентифікації хвороби. Найпоширенішими архітектурами мереж глибокого навчання є згорткові нейронні мережі (Convolutional Neural Networks (CNN)), рекурентні нейронні мережі (Recurrent Neural Networks (RNN)) і штучні нейронні мережі (Artificial Neural Networks (ANN)).

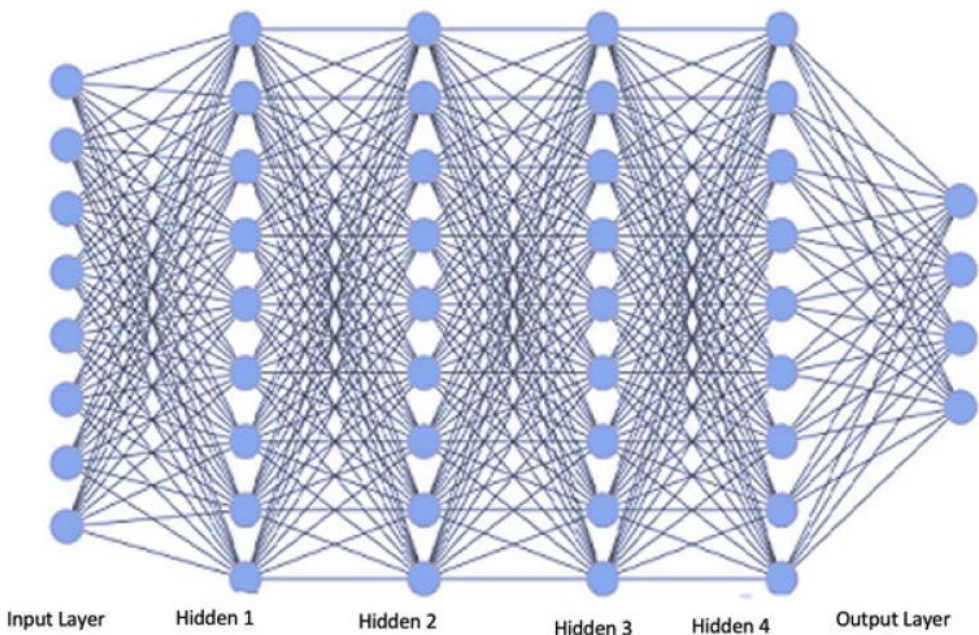


Рис. 1.15. Блок-схема нейронної мережі на основі глибокого навчання

Результати досліджень щодо раннього діагностування раку із застосування глибокої нейронної мережі (Deep Neural Network (DNN)) наведено у роботі [223]. Запропонована модель була протестована на 37 різних типах раку. В процесі експерименту були застосовані дані з атласу геному раку TCGA. Набір даних складався з 10663 зразків (9807 пухлин і 856 нормальних зразків) для 37 типів раку, при цьому вони містили експресію 10000 генів. В процесі моделювання були досліджені моделі глибоких нейронних мереж із трьома різними структурами: 3NN, 5NN та 9NN, при цьому здійснювався порівняльний аналіз з моделлю на основі застосування методу опорних векторів шляхом розрахунку точності класифікації

зразків, що досліджувалися. У таблиці 1.4 наведено результати, що отримані авторами в процесі моделювання.

Таким чином, найбільшу точність діагностики досягнуто при застосуванні моделі 5NN, але слід зазначити, що точність ідентифікації хвороби не є високою, до того ж, як показав аналіз результатів, отриманих авторами, чутливість моделі за типом раку також не є високою, тобто запропонована модель потребує доопрацювання.

*Таблиця 1.4*

*Результати моделювання щодо застосування різних типів класифікаторів для діагностики різних типів ракових захворювань*

| Методи | Точність | Влучність | Повнота | F-міра |
|--------|----------|-----------|---------|--------|
| SVM    | 82%      | 79%       | 83%     | 82%    |
| 3NN    | 84%      | 80%       | 89%     | 84%    |
| 5NN    | 90%      | 90%       | 90%     | 90%    |
| 9NN    | 83%      | 80%       | 88%     | 83%    |

У роботі [224] автори представили результати досліджень щодо застосування ансамблю моделей, заснованих на методах глибокого навчання, для діагностики трьох типів ракових захворювань: раку легенів, молочної залози та шлунку. Авторі також в процесі моделювання використали дані експресії генів з атласу геному раку TCGA, було застосовано 162 зразка раку легенів, 878 зразків з раком молочної залози і 271 з раком шлунку. Точність діагностики хвороби для всіх наборів даних становила 98%. Однак слід зазначити, що незважаючи на вражаючи результати запропонована модель має ряд суттєвих недоліків. Передусім, запропонована модель має велику складність і потребує великих часових затрат у процесі навчання і великих комп'ютерних ресурсів. До другого суттєвого недоліку слід віднести той факт, що модель на основі ансамблю методів глибинного навчання важко інтерпретувати.

У роботі [225] автори дослідили кілька моделей згорткових нейронних мереж (CNN) для діагностики різних типів ракових захворювань із застосування неструктурованих даних експресії генів. У процесі моделювання було досліджено три моделі згорткових нейронних мереж з різними комбінаціями гіперпараметрів (рис. 1.16): одновимірна модель нейронної мережі 1D-CNN (на вхід мережі подаються одновимірні профілі експресії генів); двовимірна згорткова нейронна мережа 2D-Vanilla-CNN (вхідні дані формуються як матриця даних експресії генів); гібридна згорткова нейронна мережа 2D-Hybrid-CNN (як вхідні дані, містить двовимірну матрицю та одновимірний вектор експресії генів). Моделі навчалися та тестувалися даних експресії генів, що містили 10340 зразків 33 типів ракових захворювань і 713 зразків, взятих з нормальних тканин (пухлина не виявлена). Дані містили 7100 генів, експресія яких використовувалася як атрибути. Дані також були взяті з Атласу геному раку (TCGA). Результати моделювання показали високу ефективність запропонованих моделей. Для діагностики 34 класів (33 з раковими пухлинами та

один нормальний) точність прогнозування на тестовій підмножині даних варіювалася в інтервалі від 93.9% до 95%.

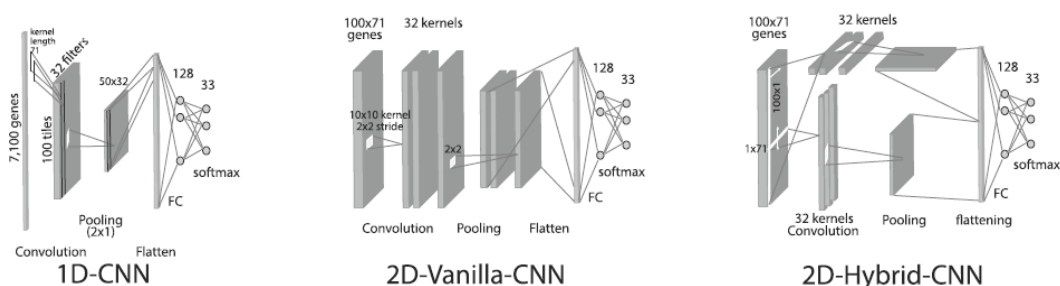


Рис. 1.16. Блок-схема моделей згорткових нейронних мереж [225]

У цьому дослідженні автори дослідили також стійкість згорткових мереж до шумової складової, що дозволило адаптувати процедуру застосування моделі згорткової нейронної мережі до необхідної розмірності вхідних даних шляхом додавання профілів генів з нульовою експресією без втрати точності прогнозування мережі. Однак слід зазначити, що навчання запропонованих моделей також потребує багато часу. Крім того, автори застосовували неструктуровані дані, що також може впливати на точність та чутливість запропонованих моделей.

## 1.5. Постановка мети та завдань дослідження

Наведений аналіз сучасного стану досліджень щодо створення та обробки даних експресії з метою розробки та удосконалення систем діагностики складних захворювань дозволяє зробити висновок, що на цей час не існує ефективної та закінченої інформаційної технології обробки даних експресії генів з метою діагностики складних захворювань [2]. Наявні моделі систем діагностики захворювань на основі даних експресії генів дозволяють з певною точністю ідентифікувати відповідну хворобу, але вони мають певні недоліки, які пов'язані з недосконалістю моделей обробки великих даних для вибору інформативних профілів експресії генів, низькою об'єктивністю формування та подальшої обробки профілів експресії генів, невисокою об'єктивністю прийняття рішень щодо наявності або відсутності відповідної хвороби тощо. Вирішення проблеми підвищення об'єктивності формування підмножин профілів експресії генів та подальшої їх обробки можливе за рахунок удосконалення існуючих та розробки нових ефективних критеріїв оцінки якості сформованих даних експресії генів та розробки гібридних моделей групування підмножин профілів експресії генів та діагностики захворювань, які передбачають розпаралелювання процесу обробки даних та прийняття компромісного рішення на основі отриманих результатів обробки даних на сформованих підмножинах профілів експресії генів. Таким чином, можна зробити висновок, що на цей час існує ряд невирішених проблем, які безпосередньо впливають на ефективність процесу діагностики складних захворювань на основі даних експресії генів, а саме:

- методи оцінки ступеня близькості високовимірних профілів експресії генів не є досконалыми і потребують доопрацювання шляхом розробки гібридних метрик на основі оцінки взаємної інформації профілів, що досліджуються з метою підвищення об'єктивності формування підмножин інформативних даних експресії генів;
- немає ефективних методів редукції профілів експресії генів для зменшення кількості профілів експресії генів видаленням неінформативних за групою критеріїв із застосуванням гібридних метрик близькості профілів експресії генів;
- немає ефективної інформаційної технології об'єктивної кластеризації високовимірних профілів експресії генів, що дозволяє сформувати підмножини диференційно експресованих та взаємно корельованих профілів експресії генів для подальшого їх застосування у гібридних моделях діагностики складних захворювань;
- немає ефективних методів та моделей об'єктивної класифікації об'єктів на основі даних експресії генів із застосуванням ансамблю класифікаторів для підвищення об'єктивності прийняття рішення щодо наявності або відсутності відповідної хвороби;
- немає ефективної інформаційної технології обробки даних експресії генів для створення системи діагностики складних захворювань, відмінною рисою якої є розпаралелювання потоку обробки та аналізу даних і застосування гібридних глибинних моделей, що дозволяє у реальному часі підвищити об'єктивність отримання рішення щодо стану об'єкта, що досліджується.

Мета роботи полягає у розробці теоретичних та практичних засад інформаційної технології обробки даних експресії генів для діагностики складних захворювань, відмінною рисою якої є більш висока, порівняно з наявними методами об'єктивність формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів та більш висока об'єктивність прийняття рішення щодо стану об'єкта за рахунок розпаралелювання потоку обробки інформації та застосування гібридних критеріїв оцінки якості реалізації відповідного етапу та застосування гібридних глибинних моделей інтелектуального аналізу даних для ідентифікації стану відповідного об'єкта.

Для досягнення поставленої мети у монографії розв'язуються такі задачі:

1. Розробка методології обробки даних експресії генів з метою формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів та подальшого застосування сформованих підмножин для діагностики складних хвороб.

2. Розробка гібридної моделі формування підмножини диференційно експресованих та взаємно корельованих профілів експресії генів на основі комплексного застосування ансамблю методів оцінки взаємної інформації між профілями експресії генів.

3. Розробка гібридної моделі об'єктивної кластеризації профілів експресії генів, застосування якої передбачає паралельну оцінку кластерної структури, отриманої на

еквівалентних підмножинах даних із розрахунком внутрішніх та зовнішніх критеріїв якості кластеризації, при цьому остаточне рішення щодо формування ансамблю внутрішніх критеріїв та кластерної структури здійснюється на основі оцінки точності класифікації об'єктів, які містять виділені у кластери профілі експресії генів як атрибути.

4. Розробка гібридної моделі діагностики складних захворювань на основі даних експресії генів, реалізація якої передбачає комплексне застосування методів кластеризації та класифікації даних експресії генів, відмінною рисою якої є висока об'єктивність прийняття рішень щодо стану об'єкта, що досліджується за рахунок ретельного формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів, розпаралелювання потоку обробки інформації та обґрунтованого вибору методів кластеризації та класифікації даних із застосуванням відповідних критеріїв якості реалізації відповідного етапу.

5. Практична реалізація розроблених моделей, методів та алгоритмів із застосуванням мультикласових експериментальних даних експресії генів, отриманих шляхом секвенування молекул РНК з оцінкою точності як бінарної класифікації об'єктів (наявність або відсутність хвороби), так і мультікласової класифікації об'єктів, що досліджуються (ідентифікація типу хвороби) з порівнянням отриманих результатів з аналогічними дослідженнями у цій предметній галузі.

## Розділ 2. Моделі, методи та алгоритми формування підмножин інформативних профілів експресії генів

У цьому розділі представлені теоретичні дослідження щодо створення гібридних моделей обробки даних експресії генів на основі комплексного застосування різного типу кількісних критеріїв оцінки ступеня близькості профілів експресії генів та рівня їх інформативності, методів кластерного аналізу та класифікації об'єктів, що досліджуються, для вибору оптимальної підмножини (або оптимальних підмножин) профілів експресії генів з точки зору роздільної здатності даних пацієнтів при різних стадіях відповідної хвороби [2, 3, 5, 8, 15, 26]. Як видно з розділу 1, сьогодні здебільшого для вибору профілів експресії генів застосовуються гібридні моделі, які передбачають комплексне застосування методів інтелектуального аналізу даних (статистичний аналіз і алгоритми кластеризації) та машинного навчання (методи класифікації) для прийняття остаточного рішення з вибору генів із застосуванням похибок першого та другого роду. При цьому можливе застосування різних методів оцінки рівня інформативності профілів експресії генів, кластеризації та класифікації, які можуть суперечити один одному і у деяких випадках давати різні результати. Підвищення об'єктивності вибору профілів експресії генів у цьому випадку можливе при використанні ансамблю методів реалізації відповідного етапу обробки даних експресії генів. Реалізація цієї концепції передбачає дослідження за оцінкою ефективності відповідних методів для вибору групи методів, що дозволяють отримати узгоджені між собою результати, при цьому основним критерієм, на наш погляд, повинна бути точність класифікації відповідних об'єктів із застосуванням виділених профілів експресії генів. У другому розділі цієї монографії представлені результати досліджень у цій предметній галузі.

### 2.1. Порівняльний аналіз метрик оцінки ступеня близькості профілів експресії генів

Як було зазначено у розділі 1, вибір підмножини інформативних профілів експресії генів для подальшого створення системи підтримки прийняття рішень при діагностиці різних захворювань або реконструкції генної регуляторної мережі передбачає оцінку інформативності та ступеня близькості профілів експресії генів шляхом застосування окремих метрик, методів або групи методів визначення міри близькості профілів експресії генів, що досліджуються. Як показав аналіз таблиці 1.3, на цей час для оцінки ступеня близькості профілів експресії генів для їх фільтрації можуть бути використані різні методи. Тому перед групуванням або видаленням профілів експресії генів необхідно виконати аналіз існуючих методів оцінки інформативності та взаємної близькості профілів експресії генів. Нехай вихідні дані експресії генів подано у вигляді матриці:

$$G = \{e_{ij}\}, i = \overline{1, n}; j = \overline{1, m},$$

де  $n$  – кількість генів, що визначають стан відповідного об'єкта;  $m$  – кількість об'єктів або зразків, що досліджуються.

У цьому випадку критерієм формування підмножин профілів експресії генів може бути цільова функція:

$$C(e_s, e_p) = \min f(e_s, e_p) \text{ або } C(e_s, e_p) = \max f(e_s, e_p), \quad (2.1)$$

де  $e_s, e_p$  – профілі експресії генів  $s$  і  $p$ , відповідно;  $f(\cdot)$  – функція подібності, що використовується для оцінки ступеня близькості профілів експресії генів  $e_s$  і  $e_p$ .

Очевидно, що у цьому випадку вибір методу визначається функцією подібності, яка властива цьому методу.

### 2.1.1. Метод максимізації взаємної інформації

Формально взаємну інформацію двох векторів дискретних змінних  $e_s$  і  $e_p$  можна оцінити таким чином [226]:

$$I(e_s, e_p) = \sum_{x \in e_s} \sum_{y \in e_p} p(x, y) \log_2 \left( \frac{p(x, y)}{p(x)p(y)} \right), \quad (2.2)$$

де  $p(x, y)$  – функція спільного розподілу ймовірностей у векторах  $e_s$  і  $e_p$ ;  $p(x)$  і  $p(y)$  – функції розподілу ймовірностей векторів  $e_s$  і  $e_p$  відповідно. При застосуванні поняття ентропії взаємна інформація може бути виражена так:

$$I(e_s, e_p) = H(e_s) + H(e_p) - H(e_s, e_p), \quad (2.3)$$

де  $H(e_s)$ ,  $H(e_p)$  і  $H(e_s, e_p)$  – ентропії векторів  $e_s, e_p$  і спільна ентропія даних векторів, відповідно. Ентропія у цьому випадку визначається як міра невизначеності відповідного стану системи і розраховується за формулою Шеннона [227, 228]:

$$H(e) = - \sum_{i=1}^m p_i(e) \log_2 p_i(e), \quad (2.4)$$

де  $m$  – кількість досліджуваних зразків (довжина профілю експресії відповідного гену);  $p_i(e)$  – ймовірність реалізації відповідного дискретного значення експресії  $i$ -ї змінної. У цьому випадку спільна ентропія Шеннона розраховується за формулою:

$$H(e_s, e_p) = - \sum_{i_s=1}^m \sum_{i_p=1}^m p_i(e_s, e_p) \log_2 p_i(e_s, e_p), \quad (2.5)$$

де  $p_i(e_s, e_p)$  – спільна ймовірність появи  $i$ -го значення у профілях експресії генів  $e_s$  і  $e_p$ .

Як показує аналіз літературних джерел [228, 229], наявні методи розрахунку ентропії Шеннона розрізняються способом розрахунку ймовірності реалізації відповідного стану системи і загалом можуть бути поділені на дві групи. Перша група методів заснована на оцінці частот виникнення відповідного стану системи. Методи другої групи передбачають розрахунок ентропії вектора змінних безпосередньо без застосування частот виникнення відповідних станів системи. Блок-схему особливо поширених методів оцінки ентропії Шеннона зображено на рис. 2.1.

Ідея методу максимальної правдоподібності (Maximum Likelihood method (ML)) заснована на припущенні, що величини вектора експресії відповідного гена заздалегідь дискретизуються так, що кожна змінна може приймати  $n$  значень. Тоді ймовірність реалізації  $i$ -го стану може бути визначена стандартним способом через

частоту виникнення відповідної події  $p_i = n_i/n$  і формула розрахунку ентропії Шеннона у цьому випадку набуває вигляду:

$$H^{ML}(e) = -\sum_{i=1}^n \frac{n_i}{n} \log_2 \frac{n_i}{n}. \quad (2.6)$$

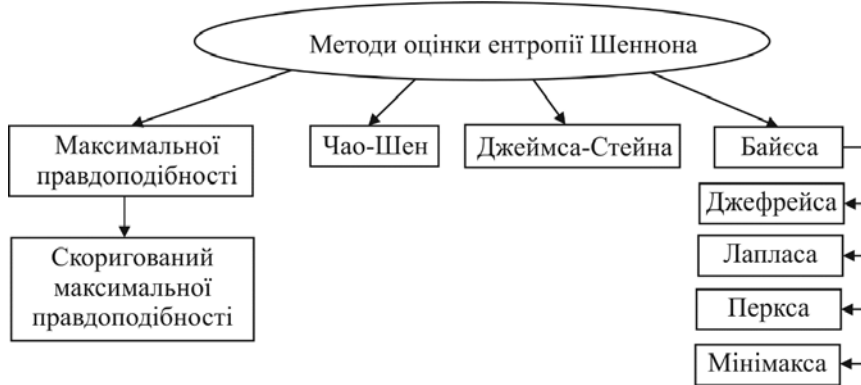


Рис. 2.1. Блок-схема методів оцінки ентропії Шеннона

Скоригований метод максимальної правдоподібності (Bias Correction Maximum Likelihood Method (BCML)) враховує зміщення значення ентропії Шеннона за рахунок наявності нульових ймовірностей [229, 230]. Формула розрахунку ентропії Шеннона у цьому випадку набуває вигляду:

$$H^{BCML}(e) = H^{ML}(e) + \frac{n_{>0}-1}{2n}, \quad (2.7)$$

де  $n_{>0}$  є кількість відліків з ненульовою частотою виникнення відповідної події.

Чао-Шен метод розрахунку ентропії Шеннона (Chao-Shen (CS)) заснований на комплексному застосуванні методів Horvitz-Thompson [231] і Good Turing [232]. Ймовірність виникнення значення відповідної ознаки у цьому випадку розраховується так:

$$p_i^{GT} = \left(1 - \frac{m_1}{n}\right) p_i^{ML}, \quad (2.8)$$

де  $m_1$  – кількість клітин з  $n_i = 1$ .

Остаточне значення ентропії Шеннона у цьому випадку розраховується за формулою:

$$H^{CS}(e) = -\sum_{i=1}^m \frac{p_i^{GT} \log_2 p_i^{GT}}{1 - (1 - p_i^{GT})^n}. \quad (2.9)$$

Метод Джеймса та Стейна розрахунку ентропії Шеннона (James-Stein Shrinkage Estimator (JS)) [229] передбачає комплексне використання двох моделей: високорозмірної моделі (вектор даних містить велику кількість атрибутів) з високою дисперсією та малим зміщенням і модель з низькою розмірністю, що характеризується малим значенням дисперсії і високим зміщенням. Застосування методу передбачає розділ вектора ознак на  $k$  однакових клітин, при цьому ймовірність у  $i$ -й клітині розраховується за формулою [229]:

$$p_i^{JS} = \lambda p_i + (1 - \lambda) p_i^{ML}, \quad (2.10)$$

де  $p_i^{ML}$  – ймовірність, що розрахована за методом максимальної правдоподібності (2.5);  $p_i = 1/n_i$  – цільова ймовірність у клітині  $i$  (якщо всі ознаки у даній клітині мають різні значення),  $n_i$  – кількість ознак у цій клітині;  $\lambda$  – параметр інтенсивності, значення якого визначається так:

$$\lambda = \frac{1 - \sum_{i=1}^k (p_i^{ML})^2}{(n-1) \sum_{i=1}^k (p_i - p_i^{ML})^2}, \quad (2.11)$$

де параметр  $n$  визначає кількість ознак у векторі даних, що досліджується.

При розрахунку ймовірності розподілу даних у виділених клітинах за формулами (2.9) і (2.10), ентропія Шеннона визначається за стандартною формулою (2.4):

$$H^{JS}(e) = - \sum_{i=1}^k p_i^{JS} \log_2 p_i^{JS}. \quad (2.12)$$

Зрозуміло, що у цьому випадку значення ентропії залежить від кількості клітин, на які розбивається вектор даних. Тому на попередньому етапі застосування цієї моделі необхідно провести дослідження за оцінкою чутливості моделі до кількості клітин, на які може бути розділений вихідний вектор даних.

Застосування методу Байєса для розрахунку ентропії Шеннона також передбачає розбивку вектора даних на  $k$  клітин з оцінкою ймовірностей розподілу величин досліджуваного вектора  $\pi = \{\pi_i\}_{i=1}^k$ , ( $\sum_{i=1}^k \pi_i = 1$ ) і апріорне формування вектора ймовірностей  $\pi$  з подальшим визначенням умовних ймовірностей  $p(H|\pi)$  і  $p(\pi|x)$  [233]:

$$H(x) = \int H(\pi) p(H|\pi) p(\pi|x) d\pi, \quad (2.13)$$

де  $p(\pi|x)$  означає значення ймовірності у відповідній клітині за умовою, що ознака  $x$  належить цій клітині:

$$p(\pi_i|x) = \prod_{j=1}^k p(\pi_i|x_j), \text{ де } p(\pi_i|x_j) = \pi_i \text{ за умови } x_j \in i. \quad (2.14)$$

Умовна ймовірність  $p(H|\pi)$  визначається функцією Дірака відповідно до умови:

$$p(H|\pi) = \delta(H - \sum_{i=1}^k \pi_i \log_2 \pi_i). \quad (2.15)$$

У цьому випадку характер розподілу ймовірностей  $\pi_i$  відповідає розподілу Діріхле з параметром концентрації  $\alpha$ :

$$p(\pi) \propto \prod_{i=1}^k \pi_i^{\alpha-1}. \quad (2.16)$$

Остаточний розрахунок ймовірності у  $i$ -й клітині за методом Байєса при відповідному параметрі концентрації у цій клітині  $\alpha_i$  виконується відповідно до формули:

$$p_i^{BS} = \frac{n_i + \alpha_i}{n + \sum_{i=1}^k \alpha_i}. \quad (2.17)$$

Параметр концентрації  $\alpha$  у цьому випадку визначається заздалегідь відповідно до даних таблиці 2.1. Значення ентропії Шеннона розраховується за стандартною формулою (2.4).

Метод максимізації взаємної інформації передбачає максимізацію функції (2.3) у випадку максимальної подібності векторів експресії генів  $e_i$  і  $e_p$ .

#### 2.1.1.1. Оцінка ефективності методу максимізації взаємної інформації при застосуванні різних методів розрахунку ентропії Шеннона

Основна ідея оцінки ефективності відповідного методу розрахунку ентропії Шеннона при застосуванні методу максимізації взаємної інформації (МВІ) заснована на припущенні, що вектори профілів експресії генів, виділені за критерієм МВІ, мають з високою точністю розрізняти об'єкти, атрибутами яких є ці профілі експресії генів. Точність класифікатора у цьому випадку має досягати максимального значення.

Таблиця 2.1

*Різновиди методу Байєса для оцінки ймовірностей у клітинах при різних значеннях параметра концентрації  $\alpha$*

| Значення параметра концентрації $\alpha$ | Метод оцінки ймовірності   |
|------------------------------------------|----------------------------|
| 0                                        | Maximum Likelihood (ML)    |
| 1/2                                      | Jeffreys JF (Джефрейса)    |
| 1                                        | Laplace LP (Лапласа)       |
| 1/n                                      | Schurmann-Grassberger (SG) |
| $\sqrt{\sum_{i=1}^n x_i} / n$            | Minimax MnM (мінімаксний)  |

Таким чином, ефективність методів розрахунку ентропії Шеннона може бути оцінена за критеріями класифікації об'єктів, атрибутами яких є профілі експресії генів, що виділені застосуванням методу МВІ із відповідним методом розрахунку ентропії Шеннона. Блок-схему алгоритму реалізації даної процедури зображено на рис. 2.2. Практична реалізація запропонованого алгоритму передбачає такі етапи:

**Етап I.** Формування матриці профілів експресії генів та вектору методів розрахунку ентропії Шеннона.

1.1. Аналіз даних ДНК-мікрочіпових експериментів або методу секвенування молекул РНК та формування матриці профілів експресії генів, де рядки відповідають об'єктам, що досліджуються, а стовпці – генам, які характеризують відповідні об'єкти.

1.2. Формування вектора методів розрахунку ентропії Шеннона для подальшої оцінки взаємної інформації профілів експресії генів.

**Етап II.** Формування матриці відстаней між профілями експресії генів за функцією взаємної інформації.

2.1. Вибір першого методу із сформованого вектора методів оцінки ентропії Шеннона ( $n=1$ ).

2.2. Розрахунок функції взаємної інформації для усіх пар профілів експресії генів, що складають матрицю даних експресії генів, за формулою (2.3). Формування матриці відстаней між профілями експресії генів.

2.3. Вибір  $k_{max}$  кількості найближчих профілів експресії генів, що відповідають критерію максимізації взаємної інформації, значення якого встановлюється емпіричним шляхом у процесі моделювання.



Рис. 2.2. Блок-схема алгоритму оцінки ефективності методу МВІ при застосуванні різних методів розрахунку ентропії Шеннона

**Етап III.** Класифікація об'єктів, що досліджуються та визначення критеріїв якості класифікації даних.

3.1. Вибір двох найближчих за обраною метрикою профілів експресії генів ( $k = 2$ ).

3.2. Ініціалізація етапу проведення класифікації об'єктів, що містять як атрибути експресії виділених генів ( $p = 1$ ). Для підвищення об'єктивності класифікації

дана процедура повторювалася 10 разів ( $p_{max} = 10$ ) з перерозподілом об'єктів у навчальній та тестовій підмножинах даних.

3.3. Реалізація процедури класифікації даних на відповідному етапі  $p$ .

3.4. Розрахунок критеріїв якості класифікації даних.

3.5. Збільшення значення  $p$  на одиницю ( $p = p + 1$ ). Якщо  $p \leq p_{max}$ , перехід на крок 3.3. У протилежному випадку, виконується розрахунок середнього значення критеріїв якості класифікації для відповідного методу розрахунку ентропії Шеннона.

3.6. Збільшення кількості найближчих профілів експресії генів на 1 ( $k = k + 1$ ). Якщо  $k \leq k_{max}$ , перехід на крок 3.2 цієї процедури. У протилежному випадку формування матриці критеріїв якості класифікації об'єктів для різної кількості виділених за методом MBI профілів експресії генів.

3.7. Вибір наступного методу розрахунку ентропії Шеннона ( $n = n + 1$ ). Якщо  $n \leq n_{max}$ , перехід на крок 2.2. У протилежному випадку формування остаточного масиву критеріїв якості класифікації даних для методів оцінки ентропії Шеннона, що використовувалися у процесі моделювання.

**Етап IV.** Аналіз отриманих результатів.

4.1. Порівняльний аналіз отриманих результатів з вибором оптимальних методів оцінки ентропії Шеннона при застосуванні методу MBI.

Практична реалізація цього алгоритму проводилася на основі програмного забезпечення R [234] із застосуванням даних профілів експресії генів GSE19188 пацієнтів, що досліджувалися на ранній стадії раку легенів [235]. Дані були отримані використанням ДНК-мікрочіпів і містили 156 мікрочіпів. Як показав аналіз анотації даних, 65 мікрочіпів містили дані експресії генів здорових пацієнтів, а на 91 мікрочіпі були ідентифіковані експресії генів пацієнтів, хворих на рак легенів (легка форма). У роботі [207] авторами наведені результати дослідження щодо комплексного застосування методів ієрархічної кластеризації та бінарної класифікації для виділення найбільш інформативних профілів експресії генів, що дозволяють з високою точністю ідентифікувати об'єкти, що досліджуються. Із 54675 генів було виділено 401 при точності класифікації 93,5%. Ця підмножина профілів експресії генів була використана в процесі моделювання. На рис. 2.3 зображено теплову карту розподілу значень експресії генів, що досліджуються. Як видно з дендрограми об'єктів, залежно від стану здоров'я, вони можуть бути адекватно розділені на дві групи у приблизному відношенні 65/91. При цьому хворі пацієнти можуть бути також розділені на підгрупи залежно від стану хвороби.

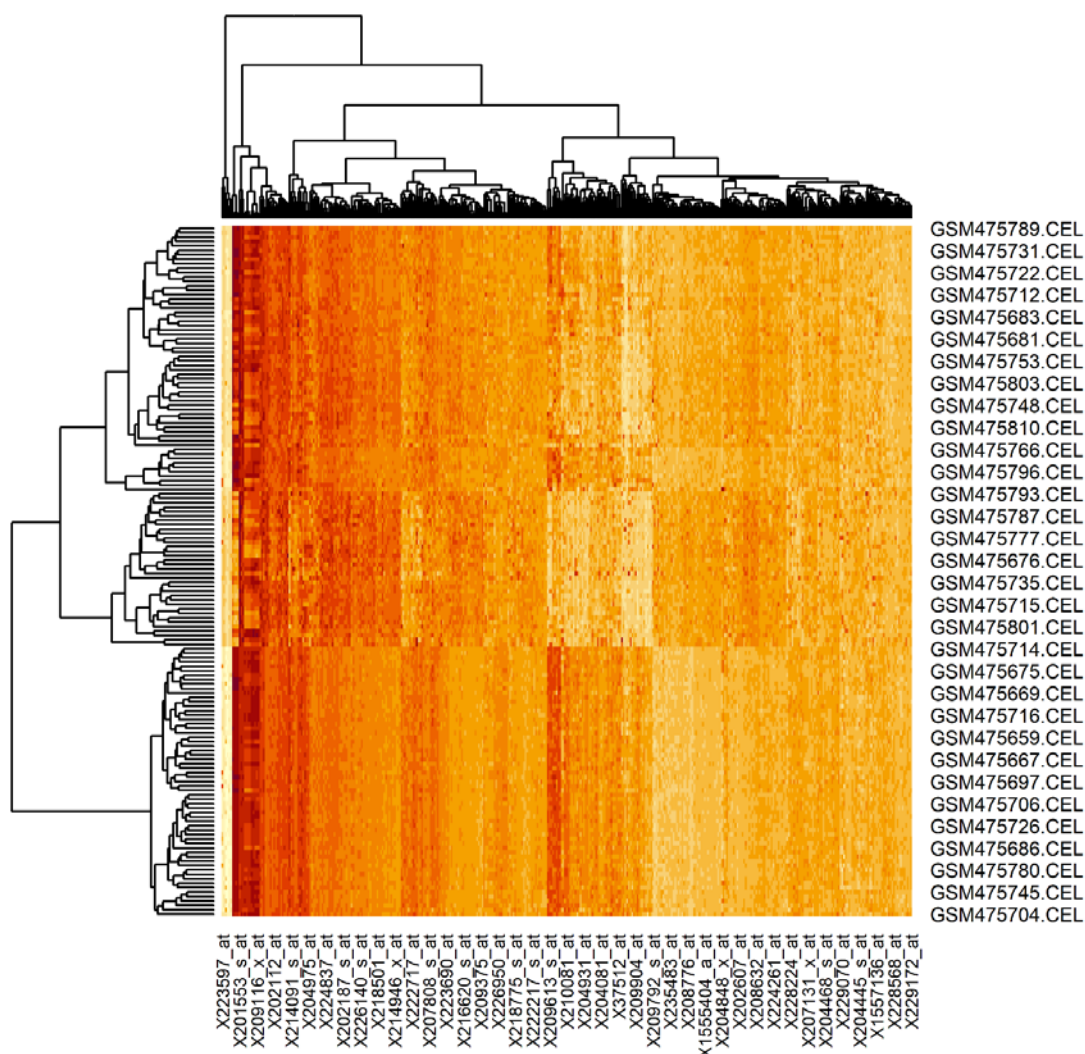


Рис. 2.3. Теплова карта розподілу значень експресії генів пацієнтів, що досліджувалися на рак легенів

Оскільки мета моделювання полягає у порівняльному аналізі різних методів оцінки ентропії Шеннона при застосуванні методу MBI, то на першому етапі випадково обирався один профіль експресії генів, до якого додавався найближчий до нього профіль за методом MBI. На другому етапі отримана підмножина профілів експресії генів збільшувалася на один профіль шляхом додавання наступного найближчого до першого гена профіля. Далі на кожному кроці кількість генів у підмножині поступово збільшувалася на одиницю. На кожному етапі проводилася бінарна класифікація об'єктів шляхом застосування класифікатора «Випадковий ліс» (Random Forest (RF)) [236], ефективність якого для бінарної класифікації профілів експресії генів доведена у [207].

Якість класифікації даних при застосуванні відповідних методів оцінки ентропії Шеннона здійснювалася із використанням критеріїв, що містять як компоненти похибки першого та другого роду [4, 6, 9, 24, 26]. Таблиця збіжності (confusionaltable), яка містить як похибки першого, так і другого роду, наведено у таблиці 2.2.

Таблиця 2.2

*Таблиця збіжності при діагностиці та класифікації пухлини у пацієнтів, що діагностувалися на рак легенів*

| Реальний стан пацієнта за результатами діагностування | Результат класифікації об'єктів           |                                             |
|-------------------------------------------------------|-------------------------------------------|---------------------------------------------|
|                                                       | Хворий (пухлина)                          | Здоровий                                    |
| Пухлина                                               | Правильно прийнята (True positives TP)    | Неправильно знехтувана (False negatives FN) |
| Здоровий                                              | Неправильно прийнята (False positives FP) | Правильно знехтувана (True negatives TN)    |

Оцінка результатів класифікації об'єктів здійснювалася на основі застосування таких критеріїв:

— точність (Accuracy(AC)) – визначається як загальна ймовірність прогнозування класифікатором правильних результатів:

$$AC = \frac{TP + TN}{TP + FP + TN + FN} . \quad (2.18)$$

— F-міра (F-measure (F)) – визначається як гармонічне середнє влучності (Precision (PR) – positive predicted values) та повноти (Recall or Sensitivity (RC)) [246]:

$$F = \frac{2 \cdot PR \cdot RC}{PR + RC} . \quad (2.19)$$

де влучність (PR) та повнота (RC) визначаються таким чином:

$$PR = \frac{TP}{TP + FP}; RC = \frac{TP}{TP + FN} .$$

Коефіцієнт кореляції Меттьюза (Matthews correlation coefficient (MCC)) – використовується у машинному навчанні як міра ефективності бінарного класифікатора [238]:

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} . \quad (2.20)$$

Особливо високі значення критеріїв (2.18)–(2.20) відповідають більш високій ефективності класифікатора.

На рис. 2.4 зображено діаграми розподілу значень критерія взаємної інформації, що розраховувалися для всіх комбінацій пар профілів експресії генів (усі комбінації з профілем першого гену) із застосуванням різних методів розрахунку ентропії Шеннона.

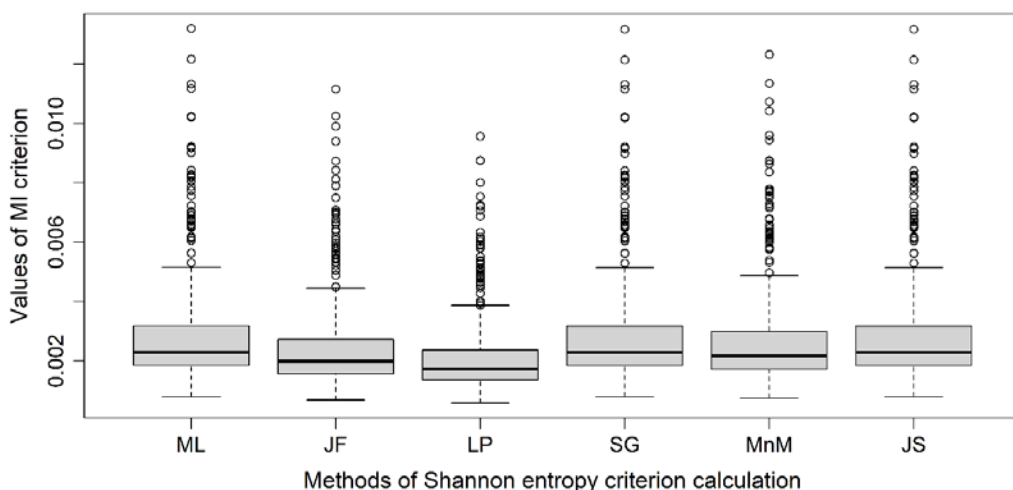


Рис. 2.4. Візуалізація характеру розподілу значень критерію взаємної інформації (Mutual Information), розрахованого при застосуванні різних методів розрахунку ентропії Шеннона

Аналіз отриманих діаграм дозволяє зробити висновок, що при застосуванні усіх методів розрахунку ентропії Шеннона існують викиди, які відповідають найбільш близьким об'єктам за критерієм взаємної інформації, але значення цього критерію відрізняються між собою при застосуванні різних методів розрахунку ентропії Шеннона, що, безумовно, може впливати на результати класифікації відповідних об'єктів.

На рисунках 2.5–2.7 представлені діаграми залежності критеріїв якості класифікації об'єктів, розрахованих за формулами (2.18)–(2.20) при збільшенні кількості профілів експресії генів від 2 до 100, які сортувалися в міру зменшення відстані до першого профіля експресії гена за критерієм взаємної інформації. Аналіз отриманих діаграм дозволяє зробити висновок, що при застосуванні різних методів розрахунку ентропії Шеннона при розрахунку критерію взаємної інформації отримуються результати класифікації, що не завжди узгоджуються між собою. При цьому діапазон зміни значень критеріїв в усіх випадках відповідає високій точності класифікації даних. Однак слід зазначити, що збільшення об'єктивності прийняття рішень при виборі найбільш інформативних профілів експресії генів на основі максимізації взаємної інформації можливо при використанні ансамблю методів оцінки ентропії Шеннона, при цьому виникає необхідність розрахунку комплексного критерію взаємної інформації, компонентами якого мають бути значення ентропії Шеннона, що розраховані на основі застосування різних методів реалізації цієї процедури.

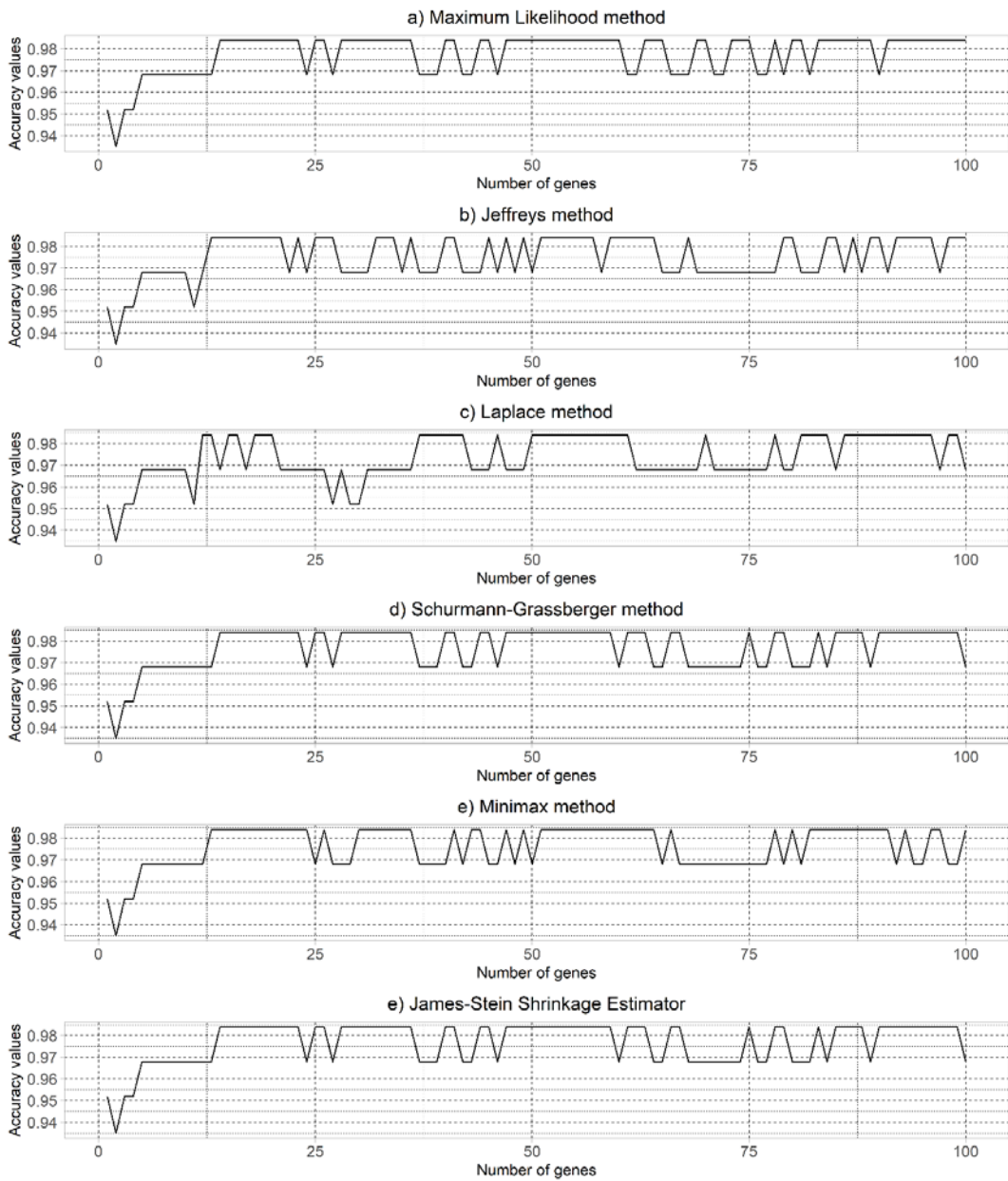


Рис. 2.5. Діаграми розподілу критерія точності класифікації (Ассурасу) об'єктів, що містять гени, виділені шляхом застосування критерія максимізації взаємної інформації при різних методах оцінки ентропії Шеннона

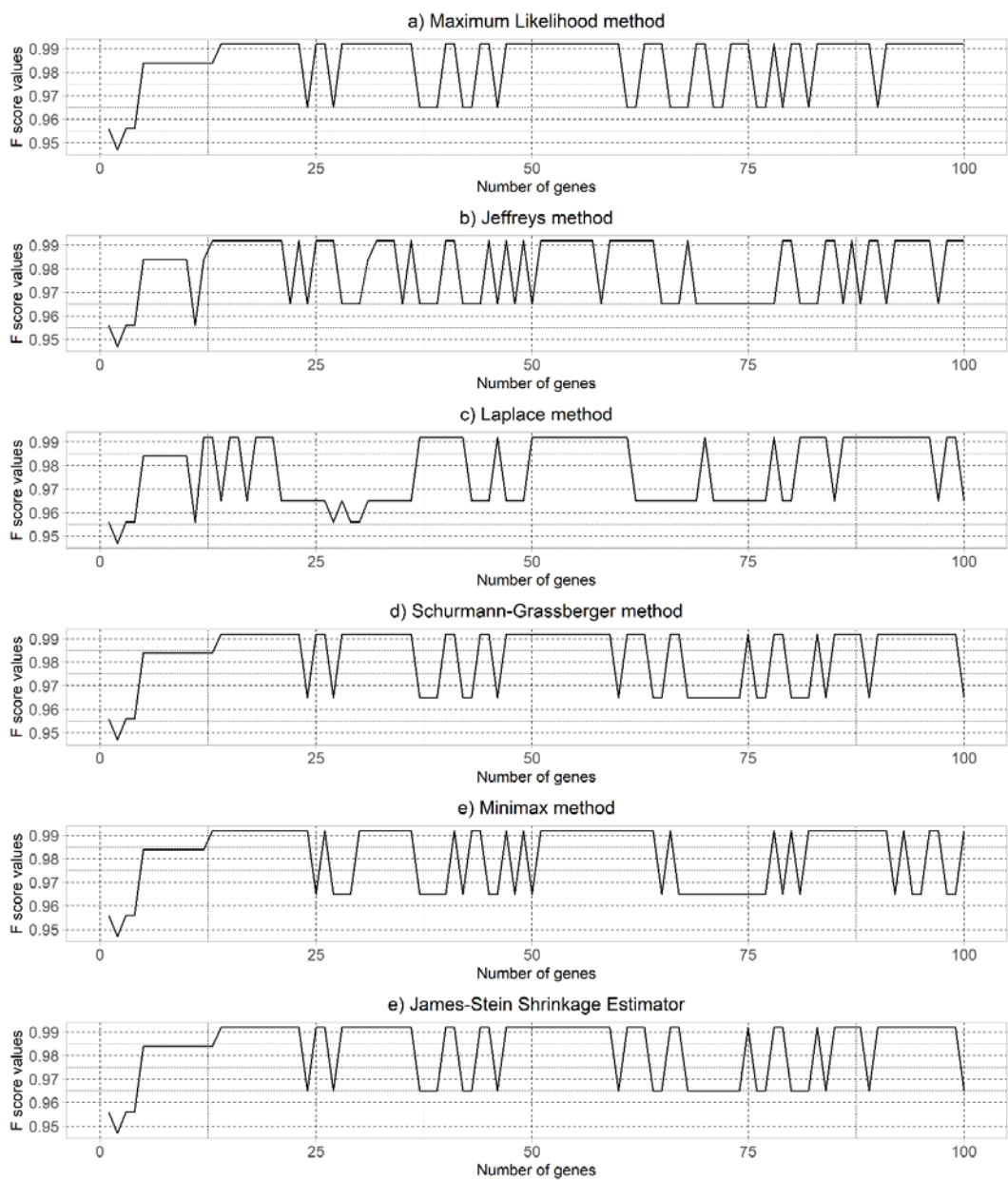


Рис. 2.6. Діаграми розподілу F-міри (F-score) на основі критерію взаємної інформації при застосуванні різних методів оцінки ентропії Шеннона

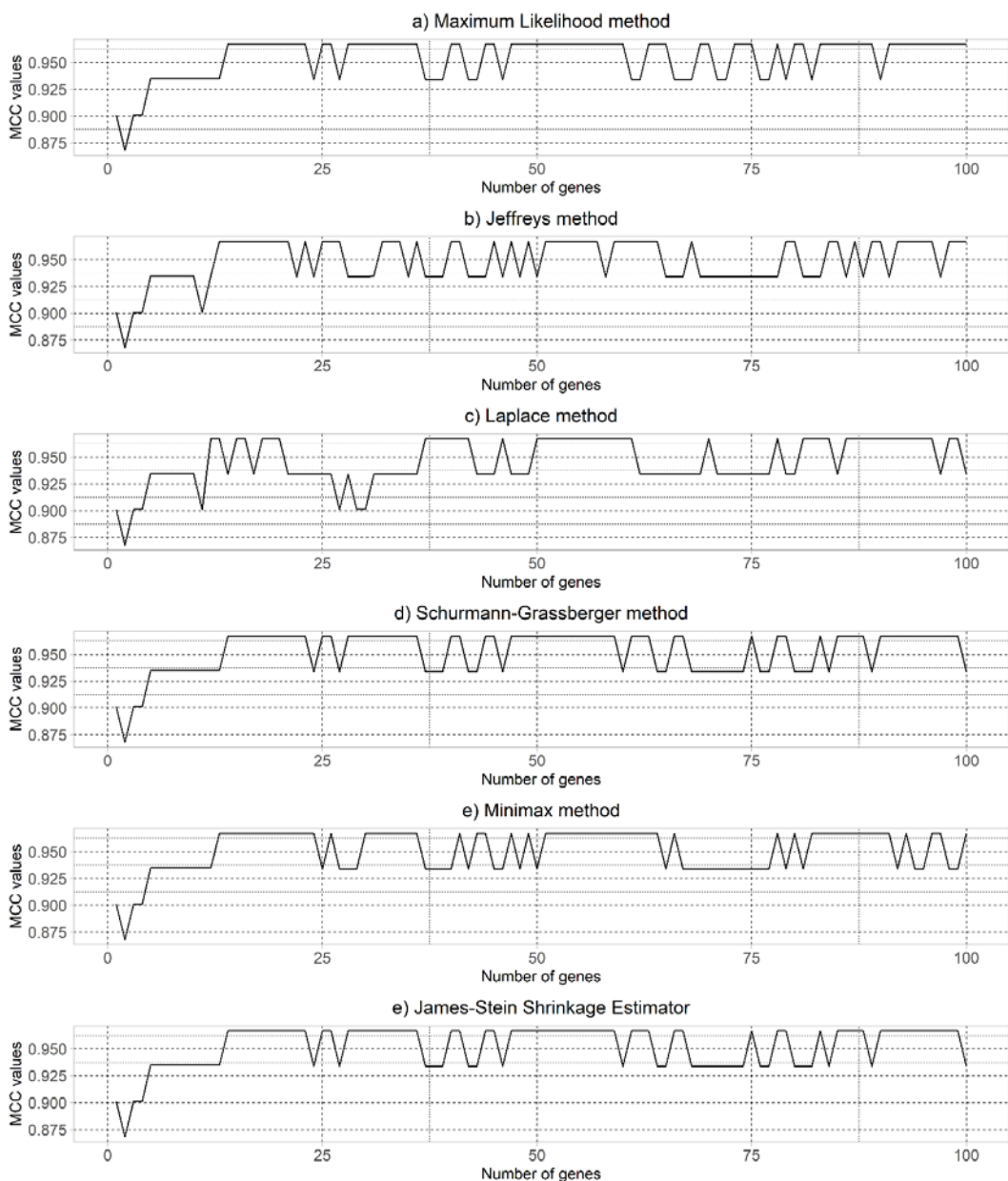


Рис. 2.7. Діаграми розподілу коефіцієнта кореляції Меттьюза (Matthews correlation coefficient (MCC)) на основі критерію взаємної інформації при застосуванні різних методів оцінки ентропії Шеннона

#### 2.1.1.2. Модифікований метод максимізації взаємної інформації

У цьому підрозділі задача підвищення об'єктивності критерію максимізації взаємної інформації при застосуванні ансамблю методів розрахунку ентропії Шеннона розв'язується застосуванням функції бажаності Харрінгтона [239], яка сьогодні успішно використовується у різних галузях наукових досліджень для

розв'язання багатокритеріальних задач. Функція бажаності Харрінгтона задається таким рівнянням:

$$d = \exp(-\exp(-Y)), \quad (2.21)$$

де значення безрозмірного параметра  $Y$  змінюється у діапазоні від -2 ( $d = 0$ ) до 5 ( $d = 1$ ). Тобто, значення бажаності ( $d$ ) при застосуванні цього методу змінюється в інтервалі від 0 до 1.

На рисунку 2.8 зображено графік функції бажаності Харрінгтона. Точки перетину функції на графіку  $0.63 = 1 - 1/e$   $0.37 = 1/e$  відповідають граничним значенням бажаності всередині інтервалу зміни функції. Значення бажаності 0,37 зазвичай відповідає межі допустимих значень. Значення від 0,37 до 0,63 відповідають задовільній бажаності, від 0,63 до 0,8 – добрій бажаності, вище ніж 0,8 – відмінній бажаності.

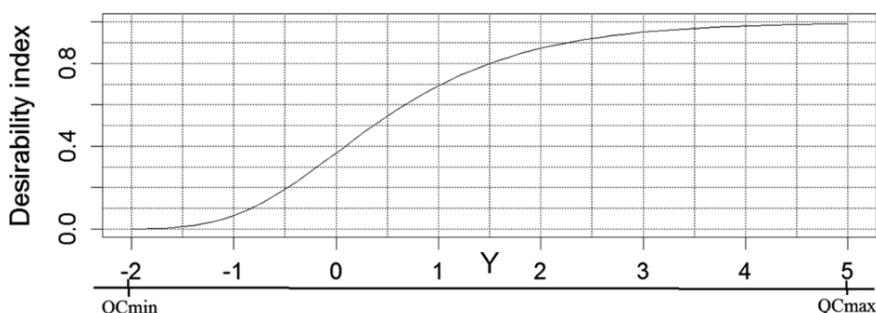


Рис. 2.8. Функція бажаності Харрінгтона

Алгоритм розрахунку комплексного критерію взаємної інформації на основі функції бажаності Харрінгтона передбачає такі кроки:

1. Трансформація шкал значень відповідних метрик, що застосовуються в рамках досліджень, у шкалу безрозмірного показника  $Y$ , значення якого змінюється у інтервалі від -2 до 5, відповідно до лінійного рівняння:

$$Y = a + b \cdot MI, \quad (2.22)$$

де  $a$  і  $b$  – коефіцієнти, що визначаються емпіричним шляхом із застосуванням граничних значень даних, що досліджуються:

$$\begin{aligned} Y_{min} &= a + b \cdot MI_{min} \\ Y_{max} &= a + b \cdot MI_{max}. \end{aligned} \quad (2.23)$$

2. Розрахунок параметру  $Y$  для кожного значення відповідного критерію:

$$Y_i = a + b \cdot MI_i. \quad (2.24)$$

3. Розрахунок приватних бажаностей для кожного значення відповідних критеріїв:

$$d_i = \exp(-\exp(-Y_i)). \quad (2.25)$$

4. Розрахунок узагальненого індексу як середнє геометричне усіх приватних бажаностей критеріїв якості, що використовуються у процесі розв'язання цієї задачі:

$$MMI = \sqrt[n]{\prod_{i=1}^n d_i}. \quad (2.26)$$

Найвище значення узагальненого індексу бажаності відповідає максимальному значення взаємної інформації за групою методів оцінки ентропії Шеннона, що використовуються у процесі моделювання. На рис. 2.9 зображено блок-схему алгоритму вибору групи інформативних профілів експресії на основі комплексного застосування методу класифікації об'єктів, що досліджуються, і модифікованого критерію взаємної інформації, що розраховується за формулами (2.22)–(2.26). Практична реалізація алгоритму передбачає такі етапи:

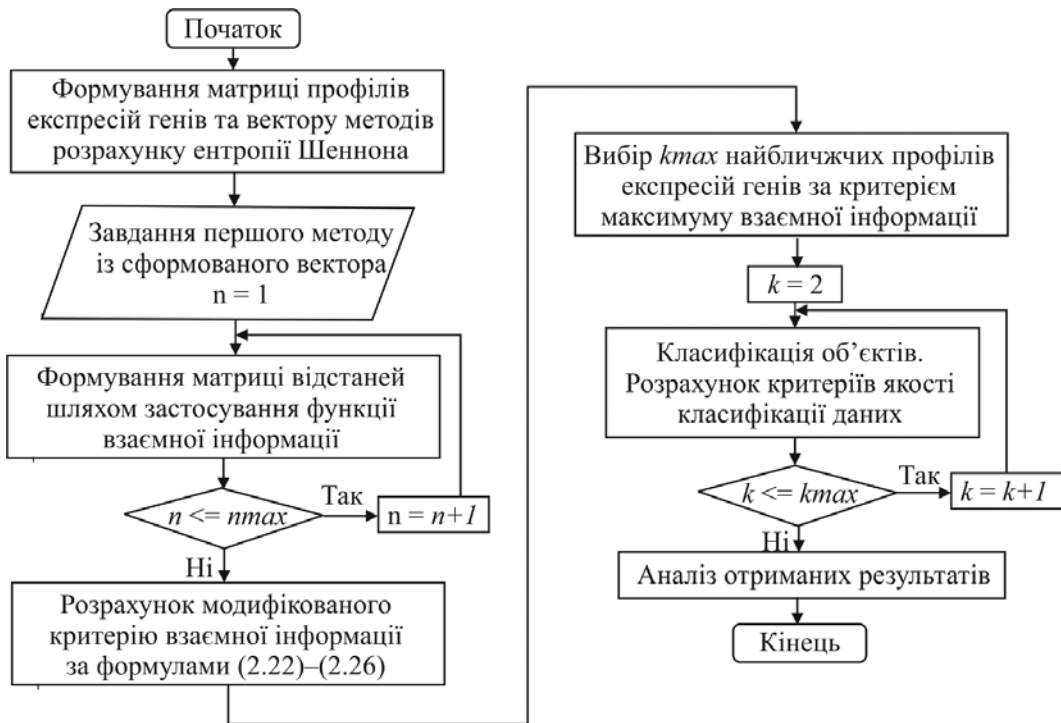


Рис. 2.9. Алгоритм формування матриці найбільш інформативних профілів експресії генів за модифікованим критерієм максимізації взаємної інформації

**Етап І.** Формування вихідних даних профілів експресії генів та вектора методів розрахунку ентропії Шеннона.

1.1. Формування матриці профілів експресії генів, де рядки – об'єкти, що досліджуються, а стовпці – гени з відповідними значеннями експресії, що формують відповідний профіль.

1.2. Формування вектора методів розрахунку ентропії Шеннона, які використовуються у процесі розрахунку критерія взаємної інформації. Вибір першого методу.

**Етап II.** Формування таблиці значень критеріїв взаємної інформації, які розраховані при застосуванні різних методів оцінки ентропії Шеннона.

2.1. Вибір першого профілю експресії генів із вихідної таблиці даних.

2.2. Розрахунок значення взаємної інформації між першим та наступними профілями експресії генів при послідовному застосуванні усіх методів розрахунків ентропії Шеннона.

**Етап III.** Розрахунок модифікованого узагальненого значення взаємної інформації.

3.1. Формування трансформованої шкали безрозмірного показника  $Y$  шляхом розрахунку коефіцієнтів  $a$  і  $b$  відповідно до системи рівнянь (2.23).

3.2. Розрахунок показника  $Y$  для усіх значень взаємної інформації, що складають таблицю даних, отриманих на етапі 2 за формулою (2.24).

3.3. Розрахунок приватних бажаностей для кожного методу розрахунку ентропії Шеннона за формулою (2.25).

3.4. Формування вектора узагальненого модифікованого критерія взаємної інформації за формулою (2.26).

**Етап IV.** Класифікація даних. Розрахунок критеріїв якості класифікації даних.

4.1. Вибір двох найближчих за критерієм (2.26) профілів експресії генів.

4.2. Застосування алгоритму класифікації даних. Розрахунок критеріїв якості класифікації даних на відповідному етапі.

4.3. Збільшення кількості найближчих до першого профілів експресії генів на одиницю. Якщо кількість профілів експресії генів менша за максимальну кількість, перехід на крок 4.2 даної процедури, у протилежному випадку – перехід на наступний етап.

**Етап V.** Аналіз отриманих результатів.

5.1. Побудова діаграм залежності критеріїв якості класифікації об'єктів від кількості генів, виділених за модифікованим критерієм максимуму взаємної інформації.

5.2. Аналіз отриманих результатів.

На рис. 2.10 зображені діаграми залежності відповідних критеріїв якості класифікації об'єктів від кількості профілів експресії генів, виділених за модифікованим критерієм максимізації взаємної інформації. Кількість генів у процесі моделювання змінювалася від 2 до 100.

На основі аналізу отриманих результатів можна зробити висновок, що запропонований метод дозволяє об'єктивно виділити групи профілів експресії генів, що дозволяють адекватно з високою точністю (99%) класифікувати об'єкти, що досліджуються. Підвищення об'єктивності у цьому випадку обумовлене коректним використанням ансамблю методів розрахунку ентропії Шеннона, значення яких застосовується при оцінці взаємної інформації відповідних профілів експресії генів.

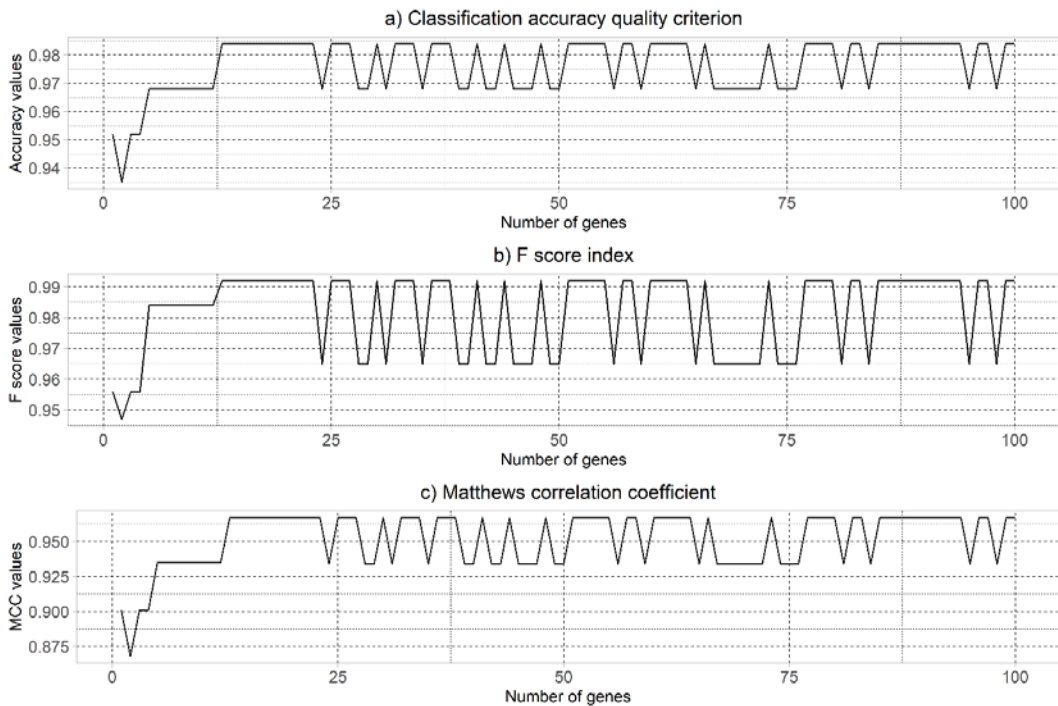


Рис. 2.10. Діаграми розподілу значень критеріїв якості класифікації даних при різних кількостях профілів експресії генів, виділених за модифікованим критерієм максимізації взаємної інформації

### 2.1.2. Метод оцінки ступеня близькості профілів експресії генів на основі критерія узгодженості Пірсона ( $\chi^2$ -статистика)

Статистичний критерій узгодженості Пірсона  $\chi^2$  дозволяє перевірити гіпотезу, що значення профілів експресії генів розподілені за однаковим законом. Нехай профіль  $i$ -го гена представлений вектором значень експресії  $e_i = (e_{ij}), j = \overline{1, m}$ , де  $m$  визначає кількість об'єктів або умов проведення експерименту щодо визначення експресії відповідного гену. Якщо діапазон зміни значень експресії цього гену  $[e_i^{min}, e_i^{max}]$  розділити на кінтервалів, що не перетинаються між собою  $[e_{is}^r, e_{ip}^r], r = \overline{1, k}$ , то кількість значень експресії гену, що належать відповідному інтервалу  $r$ , може бути визначена таким чином:  $m_r = \sum_{j=1}^m [e_{is}^r < e_{ij} \leq e_{ip}^r]$ .

У випадку класичного застосування критерію узгодженості Пірсона при використанні категоризованих даних, на першому етапі розраховується кількість об'єктів (значень вектора даних, що досліджуються), які потрапляють у відповідний інтервал. На другому етапі оцінюється очікувана кількість об'єктів у відповідних інтервалах через ймовірність знаходження відповідного об'єкта у тому чи іншому інтервалі:  $m'_r = p_r \cdot m$ . Критерій узгодженості  $\chi^2$  у цьому випадку розраховується таким чином:

$$\chi^2 = \sum_{r=1}^k \frac{(m_r - m'_r)^2}{m'_r}. \quad (2.27)$$

Висновок щодо прийняття або відхилення гіпотези про відповідність розподілу даних визначеному розподілу приймається на підставі порівняння значення критерію (2.27) з граничним значенням при заданих значеннях кількості ступенів волі і вірогідності отримання результату. Якщо розрахункове значення критерію узгодженості більше за граничне, то нульова гіпотеза щодо відсутності відповідності розподілу даних заданому розподілу відхиляється.

У випадку застосування даних експресії генів, під експресією гену розуміється величина, значення якої пропорційне кількості цього типу гена для відповідного об'єкта. Порівнюючи два профілі експресії генів  $e_s$  і  $e_p$ , значення експресії генів у першому профілі приймаються за очікувані, а у другому профілі – за розрахункові. Тоді формула критерію узгодженості двох профілів експресії генів  $e_s$  і  $e_p$  набуває вигляду:

$$\chi^2 = \sum_{j=1}^m \frac{(e_{sj} - e_{pj})^2}{e_{sj}}. \quad (2.28)$$

Менші значення критерію (2.28) відповідають більшому ступеню близькості відповідних профілів експресії генів.

### 2.1.3. Метод оцінки ступеня близькості профілів експресії генів на основі кореляційної відстані

При застосуванні кореляційної відстані між профілями експресії генів оцінюється ступінь узгодженості значень експресії генів, що відповідають різним об'єктам, що досліджуються. Як було зазначено раніше, основною метою етапу фільтрації даних є вибір профілів експресії генів, які дозволяють з максимальною точністю ідентифікувати (класифікувати) об'єкти, що досліджуються. У такому випадку можна передбачити, що профілі експресії генів мають корелювати між собою, і застосування кореляційної відстані може дозволити виділити необхідну кількість найбільш корельованих профілів експресії генів. Оскільки вектор профілю експресії генів представлений чисельними значеннями експресії, доцільним є розрахунок кореляційної відстані на основі кореляції Пірсона:

$$d_{cor}(e_s, e_p) = 1 - \frac{\sum_{j=1}^m (e_{sj} - \bar{e}_s)(e_{pj} - \bar{e}_p)}{\sqrt{\sum_{j=1}^m (e_{sj} - \bar{e}_s)^2} \cdot \sqrt{\sum_{j=1}^m (e_{pj} - \bar{e}_p)^2}}, \quad (2.29)$$

де  $\bar{e}_s$  і  $\bar{e}_p$  визначають середні значення експресії генів у профілях  $e_s$  і  $e_p$ , відповідно.

Зрозуміло, що у випадку високої кореляції профілів експресії генів значення критерію (2.29) є мінімальним.

Як було показано у роботі [172], метрики оцінки ступеня близькості високовимірних профілів експресії генів на основі формул відстаней Евкліда та Манхеттена суттєво менш ефективні, порівняно з кореляційною метрикою, тому в контексті наведених досліджень дані метрики не розглядаються. Також не розглядаються метрики, що засновані на порівнянні дисперсії розподілу значень експресії генів у кожному профілі та у генеральній сукупності, до якій ці профілі належать (критерій

Фішера). Даний вид критерію дозволяє порівняти характер розподілу значень експресії генів у відповідних профілях порівняно з розподілом значень експресії повного набору даних. Однак слід зазначити, що він не враховує співвідношення експресії пар генів, що належать одному об'єкту, що зменшує об'єктивність вибору відповідних профілів експресії генів для подальших досліджень. Тому, подальшим етапом досліджень є порівняльний аналіз представлених вище метрик з метою підвищення об'єктивності формування підмножини інформативних за групою критеріїв профілів експресії генів для подальшої обробки з метою створення системи діагностики відповідного захворювання або реконструкції генної регуляторної мережі і подальшого її моделювання.

#### **2.1.4. Порівняльний аналіз ефективності метрик оцінки ступеня близькості профілів експресії генів**

Як і у випадку досліджень щодо оцінки ефективності критерію максимізації взаємної інформації, при оцінці ефективності критерію узгодженості Пірсона або кореляційної метрики як головні критерії застосовано точність класифікації об'єктів, що досліджуються (2.18), F-міра (2.19) та коефіцієнт кореляції Меттьюза (2.20). Реалізація цієї процедури виконувалася відповідно до алгоритму, структурна блок-схема якого зображена на рис. 2.2. Відмінність полягає у тому, що формування матриці відстаней проводилася застосуванням критерію узгодженості Пірсона на першому етапі і кореляційної метрики – на другому. Покроковий процес реалізації цієї процедури передбачає такі етапи:

**Етап I.** Формування матриці експресії генів на функцій для розрахунку відповідних метрик.

1.1. Аналіз та обробка експериментальних даних з подальшим формуванням масиву профілів експресії генів.

1.2. Формування функцій розрахунку кількісних значень критерія узгодженості Пірсона та кореляційної відстані.

**Етап II.** Формування матриці відстаней між профілями експресії генів.

2.1. Вибір першої метрики для розрахунку відстані між відповідними профілями експресії генів на першому етапі і другої метрики – на другому.

2.2. Розрахунок відстані між усіма парами профілів експресії генів, що складають матрицю експериментальних даних застосуванням відповідної метрики.

2.3. Вибір  $k_{max}$  кількості найближчих профілів експресії генів у кожному випадку для реалізації подальшої процедури класифікації даних.

**Етап III.** Класифікація об'єктів, що досліджуються та розрахунок критеріїв якості класифікації даних.

3.1. Вибір двох найближчих за відповідною метрикою профілів експресії генів.

3.2. Класифікація об'єктів, що містять як атрибути виділені профілі експресії генів. Розрахунок критеріїв якості класифікації даних.

3.3. Збільшення кількості найближчих профілів експресії генів на одиницю та перехід на крок 3.2 даної процедури. Якщо кількість профілів експресії генів, що

використовуються як атрибути в процесі класифікації об'єктів сягає максимального значення, формування матриці критеріїв якості класифікації об'єктів, що досліджуються.

#### Етап IV. Аналіз отриманих результатів.

4.1. Побудова діаграм залежності значень критеріїв якості класифікації даних від кількості найближчих профілів експресії генів, що використовувалися у процесі моделювання.

4.2. Аналіз результатів моделювання. Формування висновків щодо ефективності відповідних метрик.

Для об'єктивного порівняння даних метрик з критерієм максимізації взаємної інформації на основі ентропії Шеннона процедура моделювання проводилася також із застосуванням даних профілів експресії генів GSE19188 пацієнтів, що досліджувалися на рак легенів. На рис. 2.11 зображені діаграми розподілу значень критерія узгодженості Пірсона та кореляційної відстані із застосуванням коробкової діаграми (box-and-whiskers diagram) та діаграми ядрової оцінки густини (kernel-density plot).

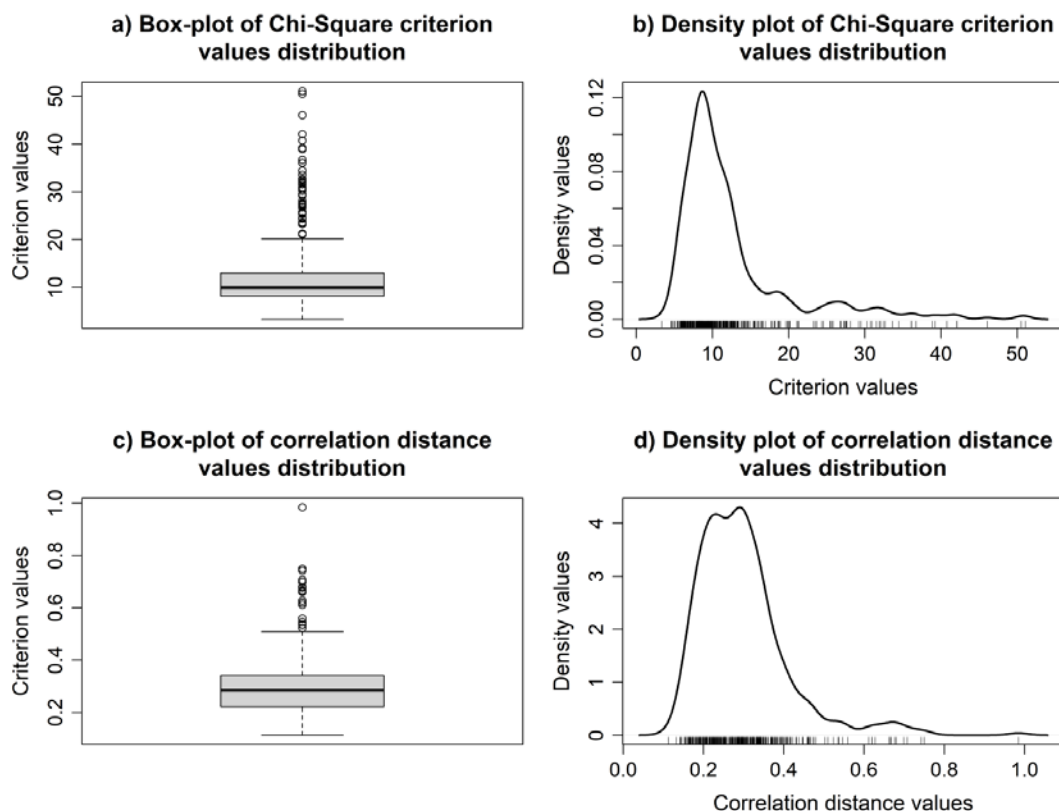


Рис. 2.11. Діаграми розподілу значень критерія узгодженості Пірсона (a,b) та кореляційної метрики (c,d)

На основі аналізу отриманих результатів можна зробити висновок, що у обох випадках дані мають викиди, які відповідають суттєво більш високим значенням відстані між відповідними профілями експресії генів. Зрозуміло, що дані профілі експресії генів не доцільно використовувати на подальшому етапі класифікації даних.

На рис. 2.12–2.14 зображені результати моделювання щодо класифікації об'єктів, атрибутами яких є найближчі за відповідною метрикою профілі експресії генів, при цьому кількість профілів експресії генів збільшувалася від 2 до 100.

Аналіз отриманих діаграм дозволяє зробити висновок щодо більш низької ефективності кореляційної метрики порівняно з критерієм узгодженості Пірсона як за абсолютним значенням, так і за чутливістю. При застосуванні кореляційної метрики при формуванні підмножини найбільш близьких профілів експресії генів результати класифікації об'єктів, що складають підмножину даних для тестування, є значно гіршими, порівняно з результатами, отриманими у випадках використання критеріїв максимізації взаємної інформації (рис. 2.10) та узгодженості Пірсона (рис. 2.12–2.14). Аналіз отриманих результатів також створює умови для підвищення об'єктивності вибору найбільш інформативних профілів експресії генів за рахунок комплексного застосування ансамблю метрик оцінки ступеня близькості відповідних профілів, при цьому необхідно врахувати характер вибору відповідних профілів експресії генів при формуванні підмножини даних за різними метриками.

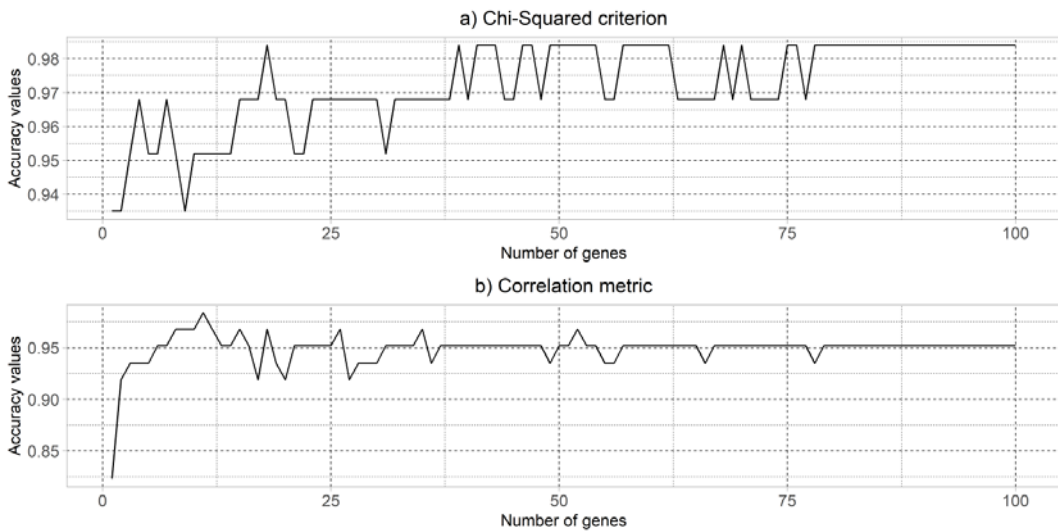


Рис. 2.12. Діаграми розподілу значень критерія точності класифікації даних (Accuracy)

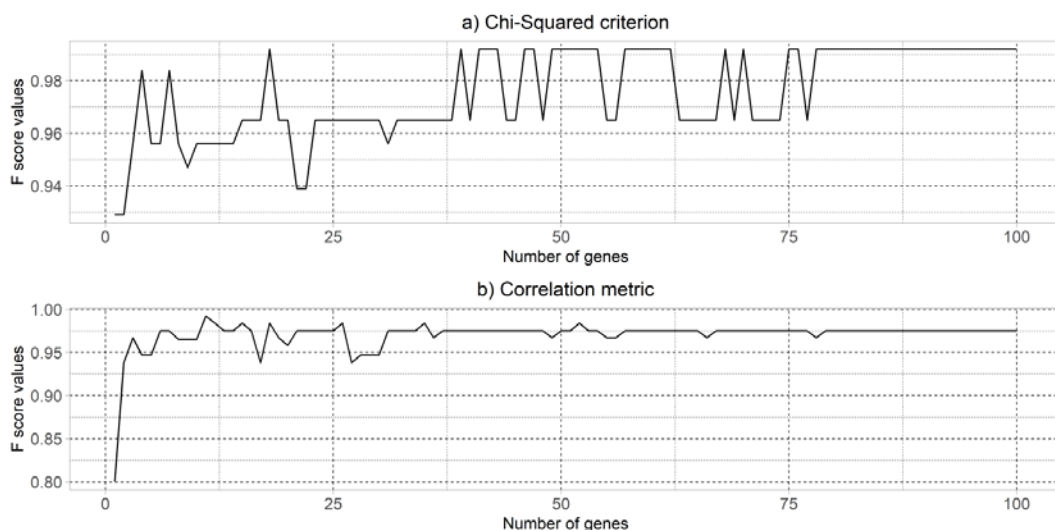


Рис. 2.13. Діаграми розподілу значень критерія якості класифікації даних F-міри (F-score)

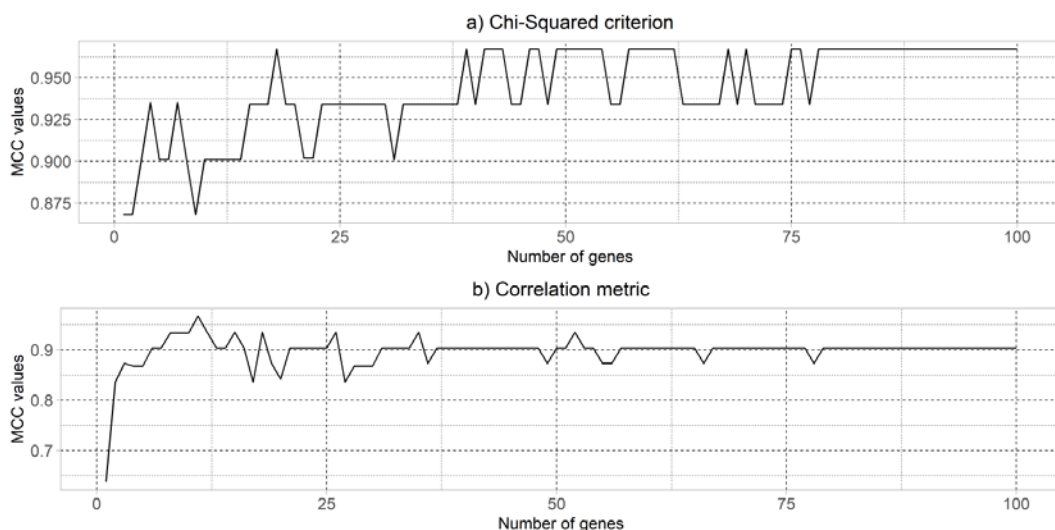


Рис. 2.14. Діаграми розподілу значень коефіцієнта кореляції Меттьюза (Matthews correlation coefficient (MCC))

Розв'язання цієї задачі здійснювалося шляхом розрахунку кількості однакових профілів експресії генів у підмножинах, що формувалися застосуванням критеріїв максимізації взаємної інформації та узгодженості Пірсона. Кількість профілів експресії генів у процесі моделювання покроково збільшувалася від 2 до 100. Результати моделювання показали, що 96 профілів експресії генів були ідентичні у сформованих підмножинах даних. Тільки 4 гена розрізнялися, що свідчить про узгодженість роботи даних критеріїв.

### 2.1.5. Метод розрахунку комплексного критерія для формування підмножини найбільш інформативних профілів експресії генів

Наведені вище результати моделювання щодо оцінки ефективності різних метрик оцінки ступеня близькості профілів експресії генів створюють передумови для розрахунку комплексного критерія, який містить як компоненти різні метрики, що виявилася найбільш ефективними за результатами попередньої оцінки. Остаточне формування значення комплексного критерію здійснюється із застосуванням функції бажаності Харрінгтона, покрокова процедура реалізації якої наведена у підрозділі 2.1.1.2. На рис. 2.15 зображено структурну блок-схему алгоритму розрахунку комплексного критерія. Реалізація цього алгоритму передбачає наступні етапи:

**Етап I.** Формування матриці профілів експресії генів та вектора, що містить назву та відповідні функції методів для розрахунку відстані між профілями експресії генів, що досліджуються.

1.1. Формування матриці експресії генів, де рядки являють собою об'єкти, що досліджуються або умови проведення експерименту щодо визначення значень експресії відповідних генів, а стовпці – профілі експресії відповідних генів.

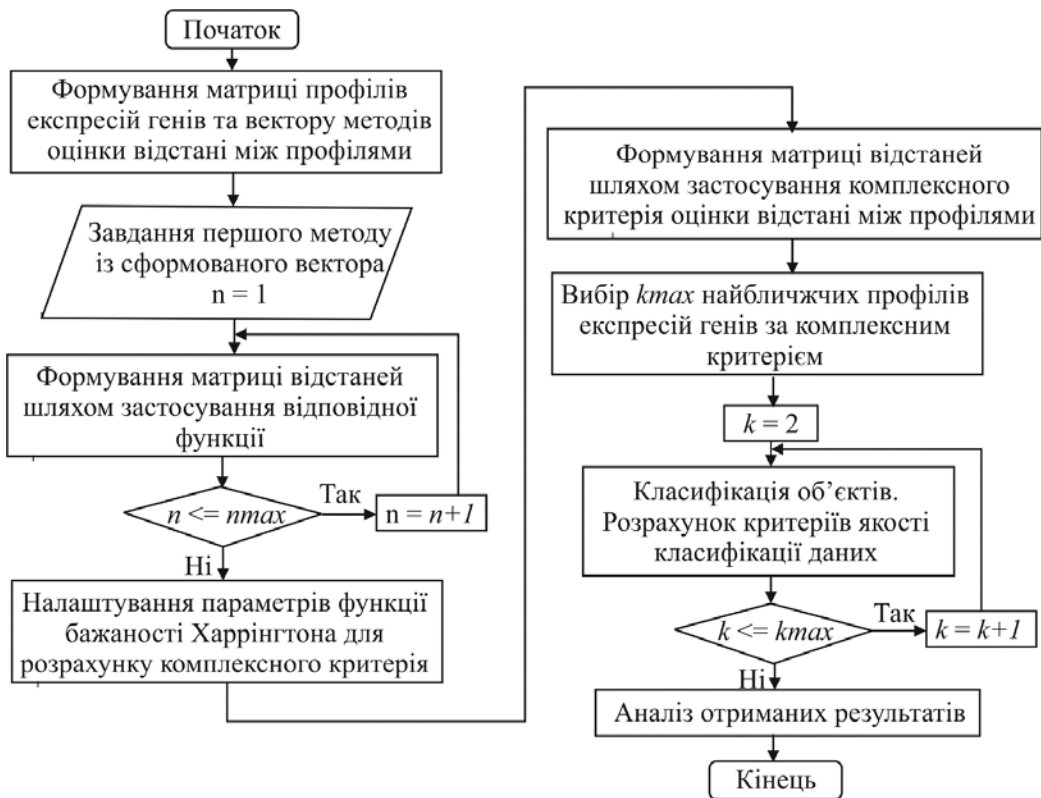


Рис. 2.15. Блок-схема алгоритму розрахунку комплексного критерія формування підмножини найбільш інформативних профілів експресії генів та оцінки його ефективності

1.2. Формування вектора метрик оцінки ступеня близькості профілів експресії генів з функціями розрахунку відповідного критерія. У рамках цього алгоритму застосовані модифікований критерій максимізації взаємної інформації (підрозділ 2.1.1.2, рис. 2.9) та критерій узгодженості Пірсона (2.28).

**Етап II.** Формування матриць відстаней між усіма парами профілів експресії генів при застосуванні відповідних критеріїв.

2.1. Формування матриць відстаней між усіма парами профілів експресії генів застосуванням відповідної функції.

2.2. Налаштування методу бажаності Харрінгтона. Розрахунок коефіцієнтів  $a$  і  $b$  для формування шкали показника  $Y$  відповідно до системи рівнянь (2.23).

2.3. Розрахунок показника  $Y$  для усіх значень отриманих відстаней за формулою (2.24).

2.4. Розрахунок приватних бажаностей для кожного критерію за формулою (2.25).

2.5. Розрахунок значень узагальненого критерію оцінки ступеня інформативності профілів експресії генів за формулою (2.26).

**Етап III.** Оцінка ефективності комплексного критерія формування найбільш інформативних профілів експресії генів.

3.1. Покрокова класифікація об'єктів, атрибутами яких є виділені профілі експресії генів. При цьому кількість профілів експресії генів збільшується від  $k = 2$  до  $k_{max}$ .

3.2. Розрахунок критеріїв якості класифікації даних за формулами (2.18–2.20) на кожному кроці реалізації цієї процедури.

3.3. Формування векторів значень відповідних критеріїв.

**Етап IV.** Аналіз отриманих результатів.

4.1. Створення діаграм залежності значень відповідних критеріїв класифікації даних від кількості найближчих профілів експресії генів.

4.2. Аналіз отриманих діаграм.

На рис. 2.16 зображено результати моделювання отже, щодо практичної реалізації вищенаведеної процедури. Як видно, запропонований метод формування підмножини близьких за гібридною метрикою профілів експресії генів дозволяє сформувати підмножини інформативних профілів експресії генів, що відповідають високим значенням критеріїв якості класифікації об'єктів, які містять як атрибути виділені профілі. Кількість генів при цьому залежить від мети поставленої задачі. Однак при цьому слід зазначити, що класифікація об'єктів може бути реалізована тільки з наявності відомих класів, на які можна розділити об'єкти, що досліджуються. Ця інформація не завжди є доступною. Тому у цьому випадку доцільним є проведення досліджень щодо оцінки ефективності критеріїв кластеризації профілів експресії генів, що розраховуються із застосуванням запропонованої метрики.

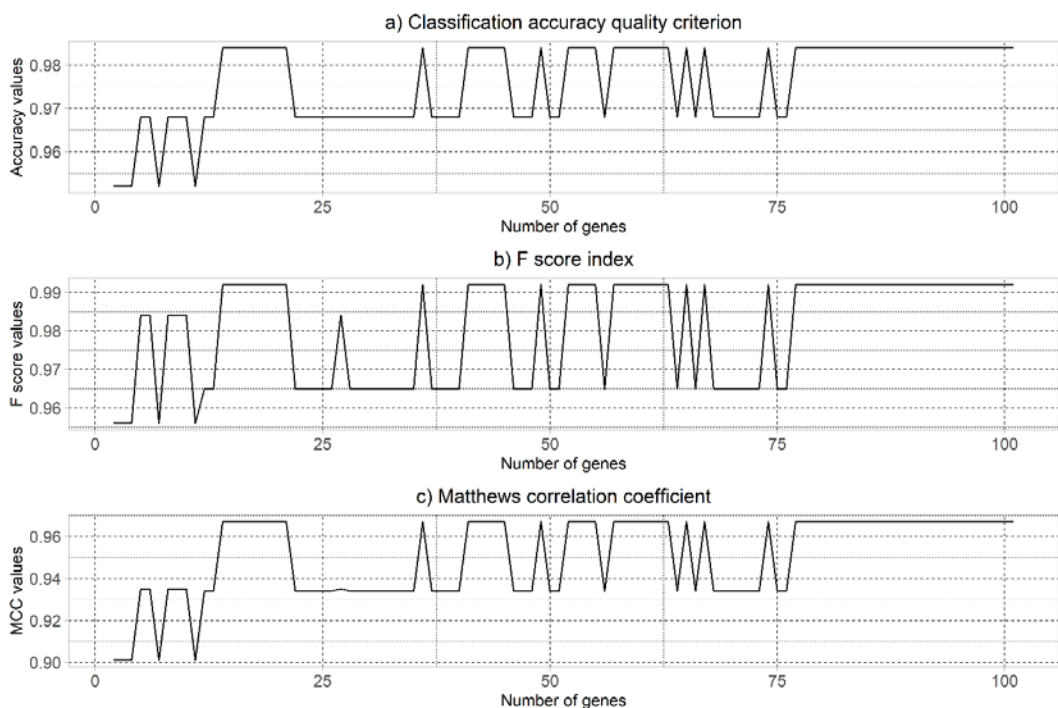


Рис. 2.16. Діаграми розподілу значень критеріїв якості класифікації даних при різних кількостях профілів експресії генів, виділених за комплексним критерієм оцінки ступеня близькості профілів

## 2.2. Теоретичні дослідження щодо оцінки ефективності застосування гібридної метрики близькості профілів експресії генів у моделях кластеризації даних

На етапі передобробки і фільтрації профілів експресії генів кластерний аналіз є невід’ємною частиною реалізації цього етапу. Якісно проведена кластеризація профілів дозволяє як виділити гени, що ідентифікуються як шум (відстань між даними профілями є суттєво більшою, порівняно з відстанню між профілями, що складають відповідні кластери), та й згрупувати близькі за відповідною метрикою профілі експресії генів у кластери для подальшого створення систем діагностики складних захворювань або реконструкції моделей генних регуляторних мереж і подальшого їх моделювання. Однак при цьому виникає необхідність у проведенні порівняльного аналізу різних метрик з оцінкою їх ефективності. У цьому підрозділі представлені дослідження щодо реалізації цього процесу, при цьому як метрики відстані досліджувалися модифікований критерій максимізації взаємної інформації, кореляційна метрика, критерій узгодженості Пірсона та комплексна метрика оцінки ступеня близькості профілів експресії генів. Як цільовий критерій, за яким оцінювалася ефективність відповідної метрики, була застосована точність класифікації об’єктів на відповідному етапі дослідження.

### 2.2.1. Аналіз ефективності метрик оцінки ступеня близькості профілів експресії генів при застосуванні однокластерної структури даних

При однокластерній структурі якість групування даних можна оцінити шляхом розрахунку середньої відстані між усіма парами профілів експресії генів:

$$QCW1 = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N d(e_i, e_j), \quad (2.30)$$

або як середнє значення відстаней профілів експресії генів до профіля, що розраховується як медіана усіх профілів експресії гена, що складають відповідний кластер:

$$QCW2 = \frac{1}{N} \sum_{i=1}^N d(e_i, C), \quad (2.31)$$

де  $N$  – кількість профілів експресії генів у кластері;  $e_i, e_j$  –  $i$ -й та  $j$ -й профілі експресії генів, відповідно;  $d(\cdot)$  – метрика оцінки відстані між відповідними профілями;  $C$  – профіль експресії гена, що визначає центр кластера. У ідеальному випадку мінімальні значення критеріїв (2.30) і (2.31) мають відповідати максимальним значенням точності класифікації об'єктів, атрибутами яких є профілі експресії генів, що складають відповідний кластер.

Розрахунок значень критеріїв (2.30) і (2.31) у рамках досліджень відповідно з такою покроковою процедурою:

1. Формування функцій розрахунку метрик відстаней, що застосовуються у процесі моделювання (модифікований критерій максимізації взаємної інформації, кореляційна метрика, критерій узгодженості Пірсона, комплексний критерій оцінки ступеня близькості профілів експресії генів).

2. Формування функцій для розрахунку внутрішніх критеріїв якості кластеризації за формулами (2.30) і (2.31).

3. Формування матриць відстаней між усіма парами профілів експресії генів за відповідними метриками, що використовуються у рамках цієї процедури.

4. Вибір двох найближчих профілів експресії генів і розрахунок критеріїв кластеризації за формулами (2.30) і (2.31).

5. Класифікація даних на цьому кроці із розрахунком точності класифікації об'єктів.

6. Збільшені кількості найближчих профілів експресії генів у кластері від 2 до  $k_{max}$  із подальшим розрахунком внутрішніх критеріїв кластеризації і класифікації для кожної метрики близькості, що використовується у рамках досліджень.

7. Побудова діаграм залежності внутрішніх критеріїв якості кластеризації та точності класифікації від кількості найближчих профілів експресії генів у кластері для кожної із метрик, що використовуються.

8. Аналіз отриманих результатів.

На рис. 2.17 зображені діаграми розподілу значень точності класифікації об'єктів, що досліджуються, відповідних внутрішніх критеріїв якості кластеризації профілів експресії генів при застосуванні модифікованого критерію максимізації взаємної інформації як метрики оцінки ступеня близькості відповідних генів при різних значеннях кількості найближчих профілів експресії генів.

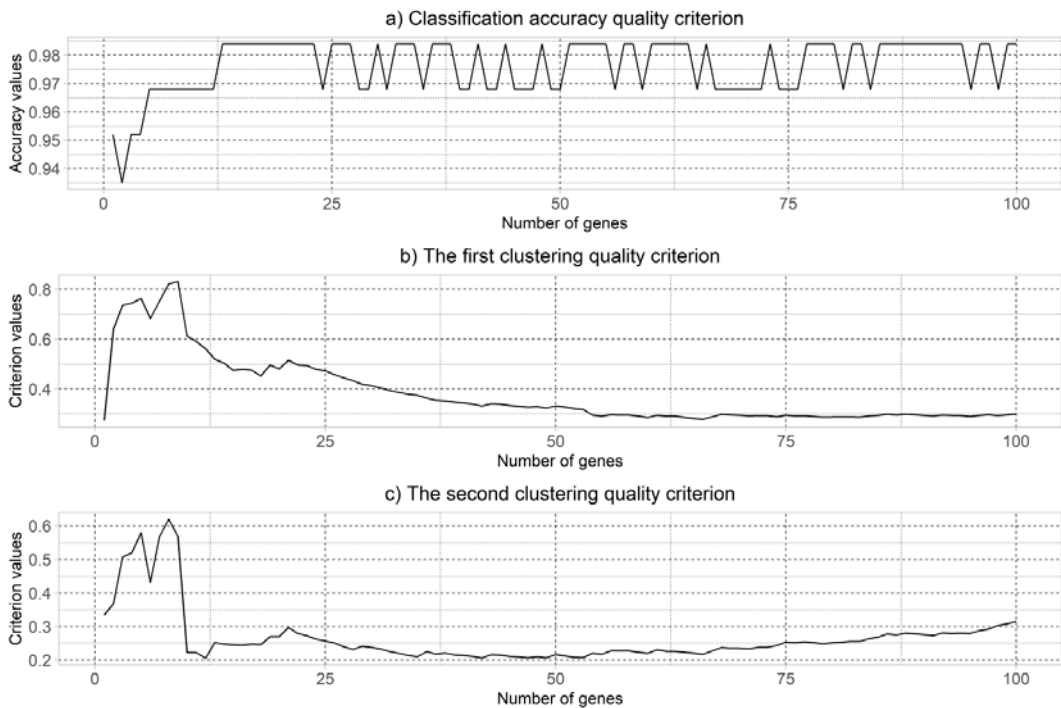


Рис. 2.17. Діаграми розподілу значень точності класифікації об'єктів, що досліджуються та внутрішніх критеріїв якості кластеризації профілів експресії генів при застосуванні модифікованого критерію максимізації взаємної інформації

На основі аналіз отриманих діаграм можна зробити висновок, що використання малої кількості профілів експресії генів ( $<15$ ) не є ефективним за усіма критеріями. При кількості профілів експресії генів більшій ніж 15, точність класифікації даних змінюється хаотично у інтервалі приблизно від 97% до 99%, значення внутрішніх критеріїв також варіюються у досить вузькому інтервалі навколо мінімуму, що свідчить про досить високу концентрацію близьких профілів експресії генів. Однак слід зазначити, що другий критерій якості кластеризації, що розрахований за формулою (2.31) за розрахунковими ресурсами є більш привабливий. Результати досліджень щодо розподілу значень точності класифікації даних та внутрішніх критеріїв якості групування профілів експресії генів у кластерах при застосуванні критерію узгодженості Пірсона, кореляційної метрики та комплексного критерію, що розраховувався на основі модифікованого критерію максимізації взаємної інформації та критерію узгодженості Пірсона, зображені на рис. 2.18–2.20.

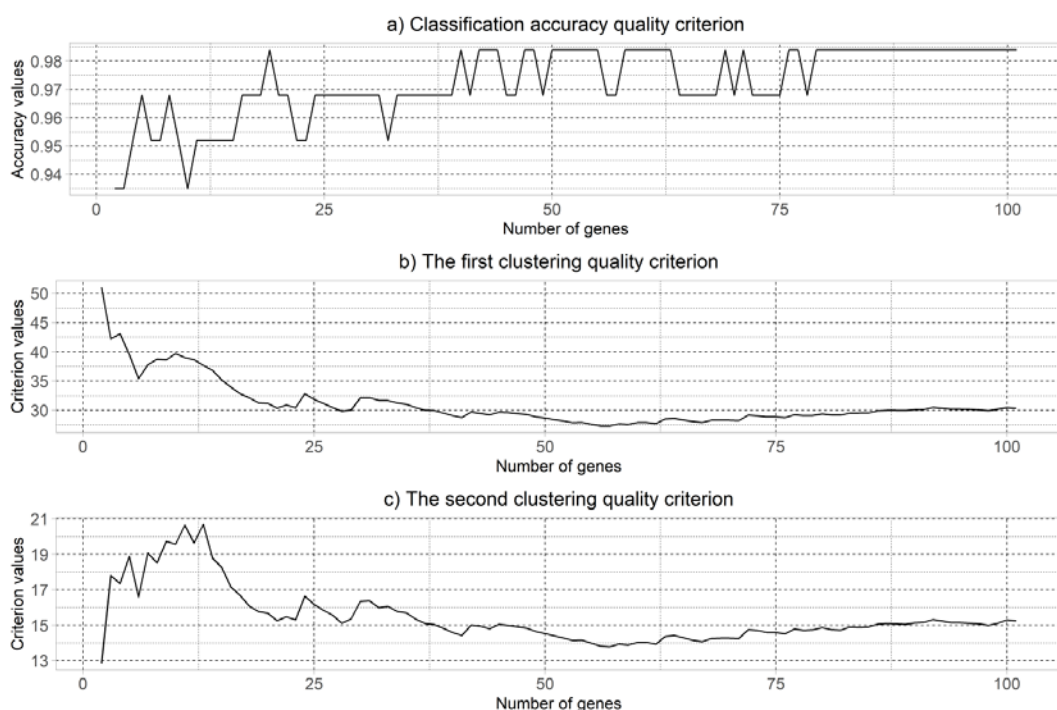


Рис. 2.18. Діаграми розподілу значень точності класифікації об'єктів, що досліджуються та внутрішніх критеріїв якості кластеризації профілів експресії генів при застосуванні критерію узгодженості Пірсона

Аналіз отриманих діаграм підтверджує раніше зроблений висновок щодо низької ефективності кореляційної метрики для виділення інформативних профілів експресії генів за внутрішніми критеріями якості групування об'єктів у одноклас-терній структурі. Значення обох внутрішніх критеріїв якості групування профілів експресії генів монотонно зменшуються, що не відповідає характеру зміни критерія точності класифікації об'єктів, що містять у якості атрибутів виділені профілі експресії генів.

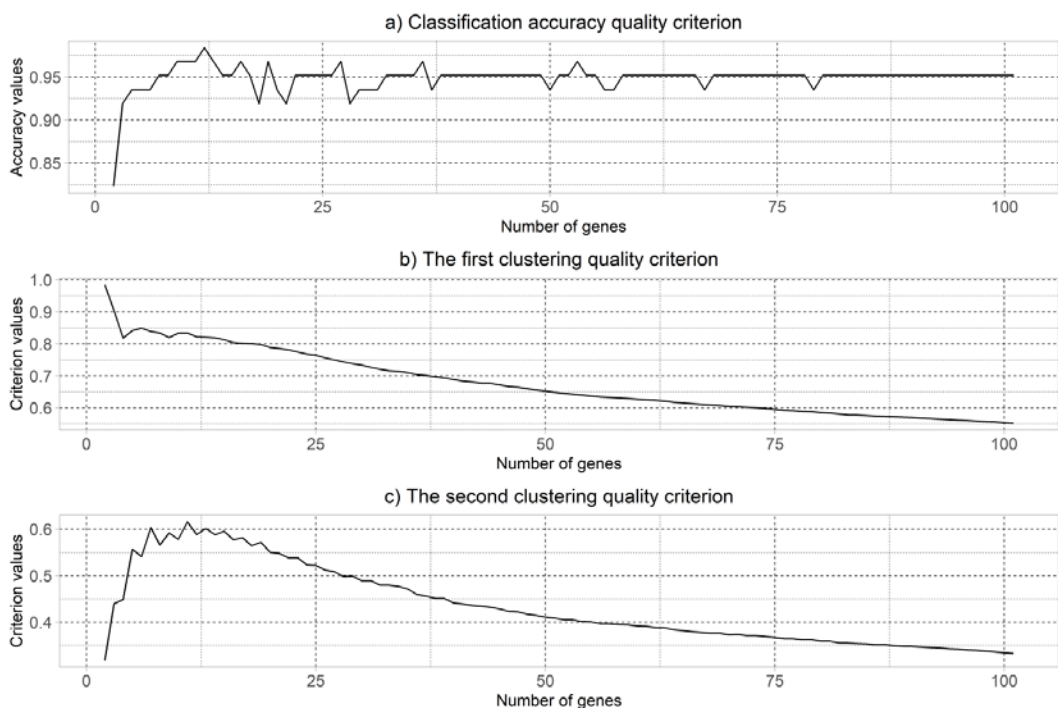


Рис. 2.19. Діаграми розподілу значень точності класифікації об'єктів, що досліджуються та внутрішніх критеріїв якості кластеризації профілів експресії генів при застосуванні кореляційної метрики

Аналіз результатів досліджень також дозволяє зробити висновок щодо більш високої ефективності другого критерію якості групування виділених профілів експресії генів, що розрахований за формулою (2.31), при застосуванні комплексного критерію оцінки ступеня близькості профілів експресії генів у однокластерній структурі. Характер зміни цього критерію найбільше відповідає характеру зміни критерія точності класифікації об'єктів, що досліджуються. Тому ця метрика буде у подальшому використовуватися для оцінки ступеня близькості профілів експресії генів.

Однак, слід зазначити, що оцінка характеру групування профілів експресії генів на основі застосування однокластерної структури не є ефективною, оскільки у більшості випадків профілі експресії генів розподіляються у різні кластери, і необхідно враховувати не тільки характер групування профілів у окремих кластерах, а – характер розподілу кластерів у просторі ознак.

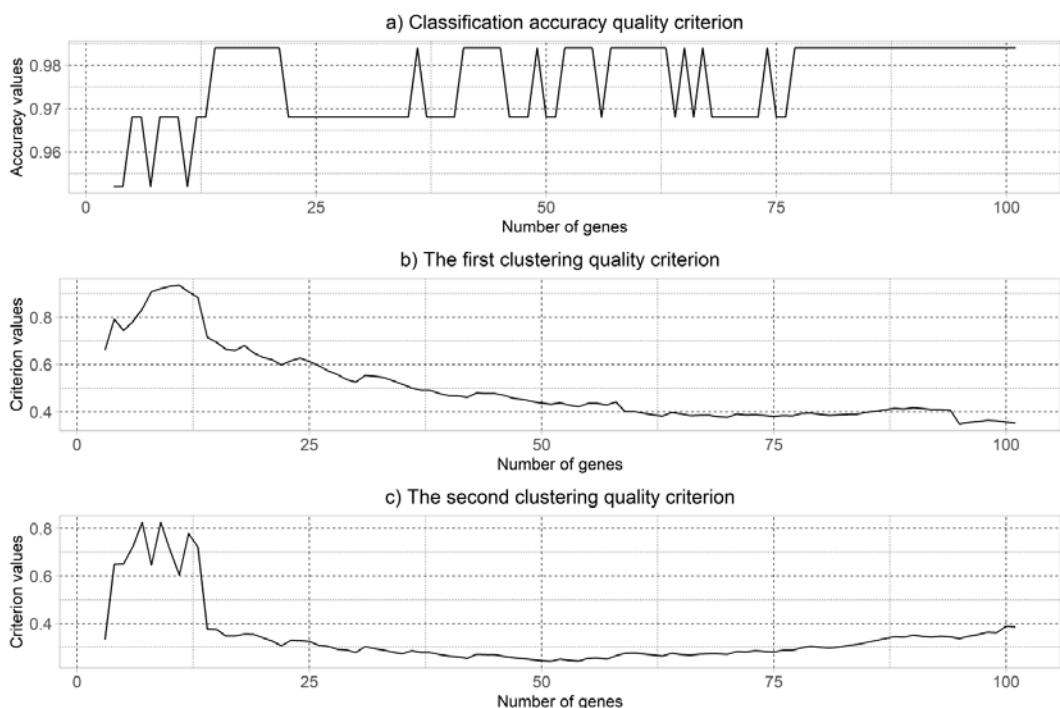


Рис. 2.20. Діаграми розподілу значень точності класифікації об'єктів, що досліджуються та внутрішніх критеріїв якості кластеризації профілів експресії генів при застосуванні комплексного критерія оцінки ступеня близькості профілів експресії генів

## 2.2.2. Розрахунок внутрішнього та зовнішнього критеріїв якості групування профілів експресії генів на основі застосування мультикластерної структури

Як було зазначено у підрозділі 2.2.1 у загальному випадку внутрішній критерій якості кластеризації профілів експресії генів повинен містити дві компоненти. Перша компонента може бути розрахована як середнє значення критерію (2.31), що розраховано для усіх виділених кластерів. Якщо вважати, що  $K$  – це кількість кластерів, то формула для розрахунку першої компоненти цього критерію може бути представлена таким чином:

$$QCW = \frac{1}{K} \sum_{k=1}^K \frac{1}{N_k} \sum_{i=1}^{N_k} d(e_i, C_k), \quad (2.32)$$

де  $N_k$  – кількість профілів експресії генів у  $k$ -му кластері;  $C_k$  – медіана  $k$ -го кластера;  $d(e_i, C_k)$  – відстань між  $i$ -м профілем та медіаною цього кластера, що розрахована із застосуванням комплексної метрики оцінки ступеня близькості відповідних профілів експресії генів.

Друга компонента внутрішнього критерія якості кластеризації може бути розрахована як середня відстань між медіанами усіх кластерів:

$$QCB = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^K d(C_i, C_j). \quad (2.33)$$

Очевидно, що у загальному випадку оптимальній кластеризації відповідає мінімальне значення критерію (2.32) і максимальне значення критерію (2.33). У праці [172] автори провели моделювання щодо оцінки ефективності різних типів внутрішніх критеріїв, що містять (2.32) і (2.33) як компоненти. Як результат дослідження, запропонований комплексний внутрішній критерій на основі мультиплікативної комбінації критеріїв Calinski-Harabasz [139] і WB індексу [173]:

$$QC^{int} = \frac{K(K-1)QCW^2}{(N-K)QCB^2}. \quad (2.34)$$

Оцінка ефективності комплексної метрики відстані між профілями експресії генів та критеріїв якості групування профілів у кластери, що розраховувалися із використанням цієї метрики проводилася на основі застосування щільнісного алгоритму кластеризації OPTICS[240], який є логічним розвитком щільнісного алгоритму DBSCAN [241, 242] і дозволяє сформувати мультикластерну структуру профілів експресії генів на основі застосування відповідної метрики оцінки відстані між профілями експресії генів. Доцільність застосування алгоритму OPTICS на цьому етапі досліджень визначається тим, що його застосування дозволяє не тільки сформувати мультикластерну структуру, яка містить кластери близьких профілів експресії генів за щільністю їх розподілу у просторі ознак, а й виділити профілі, що ідентифікуються як шум, оскільки щільність їх розподілу у просторі ознак значно менша у порівнянні зі щільністю розподілу основних груп генів.

#### 2.2.2.1. Щільнісний алгоритм формування кластерної структури OPTICS

Основна ідея функціонування щільнісного алгоритму кластеризації OPTICS (Ordering Points To Identify the Clustering Structure) [240] ілюструється рисунками 2.21 та 2.22 і полягає у наступному: для кожного об'єкта кластера (представлений точкою) встановлюється  $\varepsilon$ -околиця ( $Eps$ ), тобто визначається коло, всередині якого повинно знаходитися кількість об'єктів (точок), що перевищує мінімальний поріг ( $MinPts$ ) – граничну кількість точок всередині відповідного кола.

Функціонування щільнісних алгоритмів засновано на таких означеннях [240, 241]:

*Означення 1.* Об'єкт  $p$  є безпосередньо щільнісно досяжний з об'єкта  $q$ , якщо при заданих радіусі  $\varepsilon$ -околиці  $Eps$  та мінімальній кількості об'єктів всередині  $\varepsilon$ -околиці  $MinPts$  виконується умови:

1.  $p \in N_\varepsilon(q)$ , де  $N_\varepsilon(q)$  є підмножина об'єктів, що знаходяться всередині  $\varepsilon$ -околиці об'єкту  $q$ .

2.  $Card(N_\varepsilon(q)) \geq MinPts$ , де  $Card(N_\varepsilon(q))$  означає потужність або кількість об'єктів, що знаходяться всередині  $\varepsilon$ -околиці об'єкту  $q$ .

Слід зазначити, що друга умова є умовою кореневого об'єкта. Якщо дана умова виконується, то об'єкт  $q$  вважається кореневим.

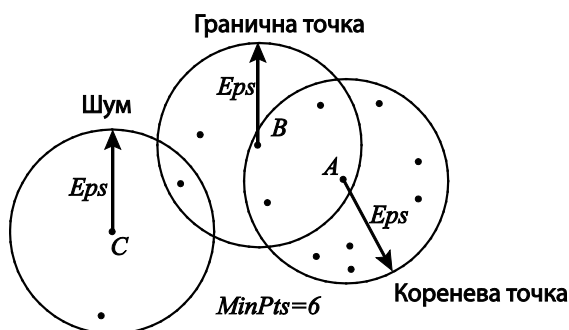


Рис. 2.21. Ілюстрація розподілу об'єктів і ключових понять при застосуванні щільнісного алгоритму



Рис. 2.22. Ілюстрація щільнісної досяжності та з'єднаності точок

**Означення 2.** Об'єкт  $p$  є щільнісно-досяжний з об'єкта  $q$ , якщо при заданих радіусі  $\varepsilon$ -околиці  $Eps$  та мінімальній кількості об'єктів всередині  $\varepsilon$ -околиці  $MinPts$  існує такий ланцюг об'єктів  $q, p_1, p_2, \dots, p_i, \dots, p$ , що об'єкт  $p_{i+1}$  є безпосередньо щільнісно-досяжний із об'єкту  $p_i$ .

**Означення 3.** Об'єкт  $p$  є щільнісно-з'єднаний з об'єктом  $q$ , якщо при заданих радіусі  $\varepsilon$ -околиці  $Eps$  та мінімальній кількості об'єктів всередині  $\varepsilon$ -околиці  $MinPts$  існує такий об'єкт  $o$ , що об'єкти  $p$  і  $q$  є безпосередньо щільнісно досяжні з об'єкта  $o$ .

**Означення 4.** Нехай  $D$  є множина об'єктів, що досліджуються, а  $C$  є кластер, що містить близькі за умовами щільнісної досяжності або з'єднаності об'єкти. У цьому випадку умови формування кластеру об'єктів або шуму є такі:

1. Якщо  $\forall p, q \in D, p \in C$  і  $q$  є щільнісно-досяжний з об'єкта  $p$  при заданих радіусі  $\varepsilon$ -околиці  $Eps$  та мінімальній кількості об'єктів всередині  $\varepsilon$ -околиці  $MinPts$ , то  $q$  також  $\in C$ .

2. Якщо  $\forall p, q \in C$ , то  $q$  є щільнісно-з'єднаний з об'єктом  $p$  при заданих радіусі  $\varepsilon$ -околиці  $Eps$  та мінімальній кількості об'єктів всередині  $\varepsilon$ -околиці  $MinPts$ .

3. Якщо об'єкт  $p$  не є щільнісно з'єднаний з будь яким об'єктом  $q \in C$ , то даний об'єкт ідентифікується як шум.

Таким чином, при застосуванні щільнісних алгоритмів результат кластеризації визначається комбінацією двох параметрів: радіусом  $\varepsilon$ -околиці ( $Eps$ ) та мінімальною кількістю об'єктів всередині  $\varepsilon$ -околиці ( $MinPts$ ). Як показує аналіз робіт [240,

241], ключовою відмінністю щільнісних алгоритмів DBSCAN і OPTICS є характер формування кластерів і ідентифікації шумових об'єктів на основі щільності з'єднання або досяжності відповідних об'єктів, що досліджуються. Для визначення оптимальних значень  $Eps$  та  $MinPts$  при застосуванні алгоритму DBSCAN, авторами у роботі [240] запропонований метод, заснований на  $k$ -distance діаграмі, яка являє собою розподіл значення радіусу  $\varepsilon$ -околиці при різних кількостях об'єктів, що мають  $k$  ( $MinPts$ ) найближчих щільнісно-з'єднаних сусідів. При цьому кількість об'єктів впорядковується за наростанням. Автори показали, що оптимальне значення  $Eps$  відповідає коліну  $k$ -distance графу. Основним недоліком цього методу є висока чутливість алгоритму до його параметрів. Як показано у роботах [243, 244], мінімальна зміна значення  $Eps$  при фіксованому значенні  $MinPts$  суттєво змінює як розподіл об'єктів у кластерах, так і розподіл кластерів у просторі ознак. Окрім того,  $k$ -distance необхідно будувати для усіх  $MinPts$  у встановленому діапазоні зміни цих значень, що ускладнює процес ефективного застосування цього алгоритму.

Як було зазначено раніше, щільнісний алгоритм кластеризації OPTICS є логічним продовженням алгоритму кластеризації DBSCAN. Для визначення оптимальних значень  $Eps$  та  $MinPts$ , авторами алгоритму OPTICS у роботі [241] запропоновано метод, заснований на діаграмі відстані досяжності (reachability-distance plot), який являє собою діаграму залежності значення  $Eps$  від упорядкованого значення відстані досяжності при формуванні кластерної структури. Як показали результати моделювання, що представлені у роботі [244], застосування щільнісного алгоритму кластеризації OPTICS до модельних синтетичних даних дозволяє отримати адекватні результати кластеризації при суттєво меншій чутливості до значення параметру  $Eps$ , тобто цей алгоритм є більш стабільним, порівняно з аналогічним алгоритмом DBSCAN.

#### 2.2.2.2. Формування кластерної структури на основі принципів індуктивної технології об'єктивної кластеризації

Слід зазначити, що внутрішній критерій, що розрахований за формулами (2.32)–(2.34) не завжди дозволяє об'єктивно вибрати адекватну кластеризацію з причини наявності похибки відтворюваності, яка властива більшості наявних алгоритмів кластеризації. Іншими словами, задовільні результати кластеризації, отримані при застосуванні однієї множини даних не завжди повторюються при використанні іншої аналогічної множини об'єктів, що досліджуються.

У працях [160–165] авторами запропоновано ідею щодо зменшення похибки відтворюваності при кластеризації даних шляхом використання «свіжих даних» (не використовувалися при створенні моделі) у процесі верифікації отриманої моделі розподілу об'єктів у кластери та прийняття остаточного рішення щодо вибору кластерної структури з послідовним розрахунком внутрішніх, зовнішніх критеріїв та критерія балансу, що враховує можливі розбіжності між внутрішніми та зовнішніми критеріями. Подальший розвиток ця ідея отримала у роботах [166–169] при розробці індуктивної технології об'єктивної кластеризації, основу якої складають три принципи:

1. *Принцип зовнішнього доповнення.* Реалізація цього принципу передбачає розбиття множини початкових даних на дві еквівалентні підмножини, що містять однакову кількість попарно близьких об'єктів за метрикою, що використовується для оцінки відстані між відповідними об'єктами. Друга підмножина даних відіграє роль «свіжих даних». Кластеризація даних здійснюється паралельно на двох еквівалентних підмножинах з подальшим розрахунком внутрішніх, зовнішніх критеріїв та критерію балансу.

2. *Принцип послідовного перебору.* Основна ідея цього принципу полягає у послідовному переборі можливих кластеризацій на двох еквівалентних підмножинах даних з вибором найкращих, які відповідають екстремумам як внутрішніх, так і зовнішніх критеріїв якості кластеризації даних.

3. *Принцип неостаточності рішень.* Реалізація цього принципу передбачає фіксацію кластеризацій, що відповідають локальним екстремумам комплексного критерію балансу, який містить як складові внутрішні та зовнішні критерії якості кластеризації. Остаточне рішення щодо формування кластерної структури приймається, враховуючи мету задачі, що розв'язується на цьому етапі обробки даних.

У роботі [166] авторами запропонований зовнішній критерій якості кластеризації, що розраховується як нормалізована різниця відповідних внутрішніх критеріїв, які розраховані на кластеризаціях двох еквівалентних підмножин на відповідному ієрархічному рівні формування кластерної структури:

$$QC^{ext} = \frac{|QC_1^{int} - QC_2^{int}|}{QC_1^{int} + QC_2^{int}}. \quad (2.36)$$

Основна ідея полягала у тому, що мінімальне значення похибки відтворюваності відповідає максимальній ступені схожості результатів кластеризації, отриманих на двох еквівалентних підмножинах даних. Оскільки внутрішні критерії враховують як характер розподілу об'єктів у кластерах, так і характер розподілу кластерів у просторі ознак, то об'єктивна кластеризація (мінімальне значення похибки відтворюваності) у цьому випадку відповідає мінімуму різниці між відповідними значеннями внутрішніх критеріїв. Нормалізуюча поправка у формулі (2.35) трансформує діапазон зміни значень зовнішніх критеріїв у процесі перебору кластеризацій у інтервал від 0 (нульова похибка відтворюваності) до 1 (максимальна похибка).

Комплексний критерій балансу розраховується на основі застосування методу бажаності Харрінгтона відповідно до алгоритму, представленою у підрозділі 2.1.1.2, при цьому у формулах (2.22)–(2.24) знак “+” змінюється на “–”, оскільки максимальне значення бажаності Харрінгтона відповідає мінімальним значенням як внутрішніх критеріїв, що розраховуються за формулою (2.34), так і зовнішніх критеріїв.

На рис. 2.23 зображено структурну блок-схему алгоритму OPTICS, реалізованого у рамках індуктивної технології об'єктивної кластеризації із застосуванням комплексної метрики оцінки відстані між профілями експресії генів, що досліджуються. Практична реалізація даного алгоритму передбачає наступні етапи.

**Етап I.** Формування даних та функцій розрахунку відповідних критеріїв.

1.1. Формування матриці даних експресії генів, де рядки є об'єктами, що досліджуються, а стовпці – генами, експресія яких визначає відносну кількість цього типу гена для відповідних об'єктів, що досліджуються.

1.2. Формування функції оцінки відстані між профілями експресії генів на основі комплексного застосування модифікованого критерія максимізації взаємної інформації та критерія узгодженості Пірсона.

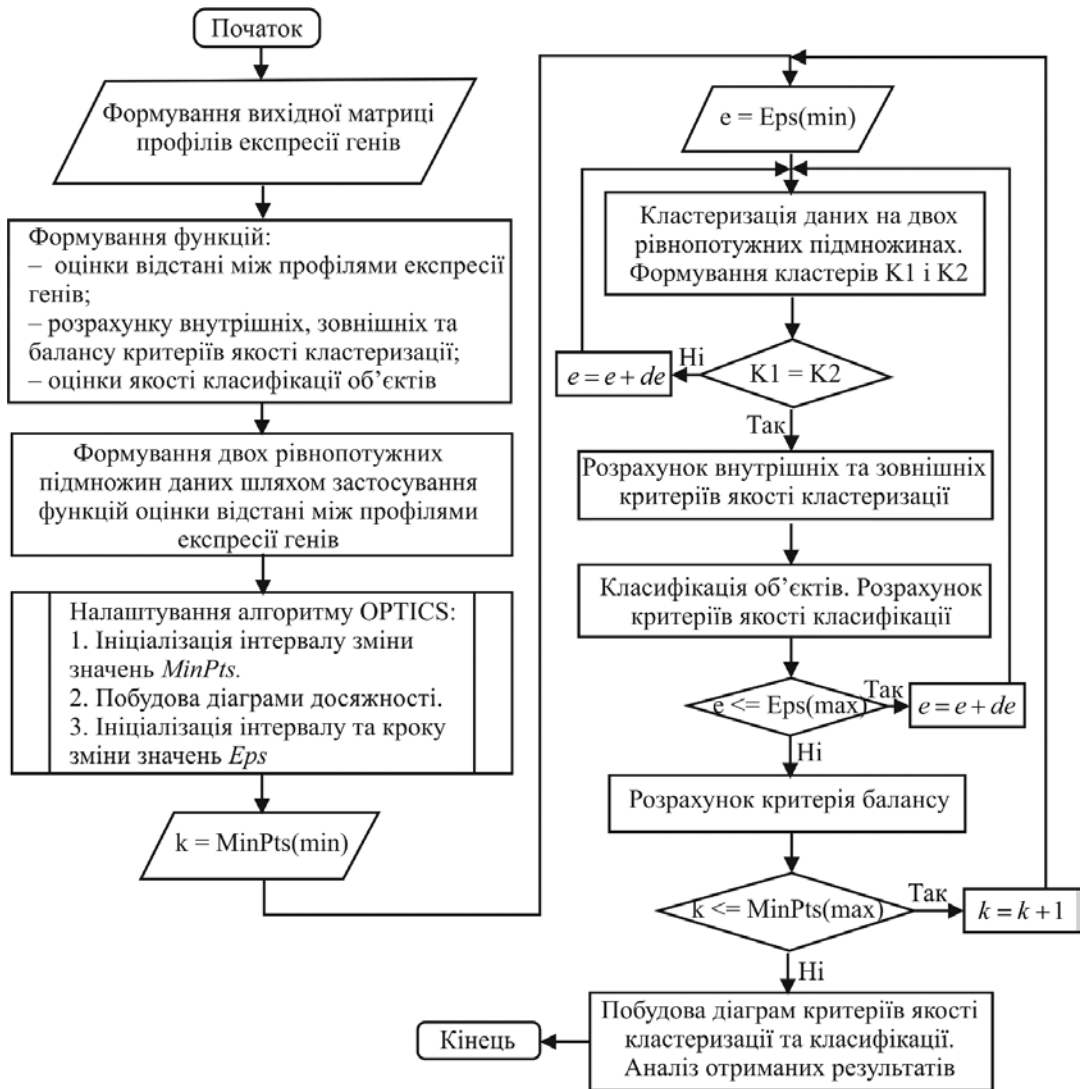


Рис. 2.23. Блок-схема алгоритму формування мультикластерної структури на основі щільнісного алгоритму кластеризації OPTICS, реалізованого у рамках індуктивної технології об'єктивної кластеризації

1.3. Формування функцій розрахунку внутрішніх критеріїв якості кластеризації профілів експресії генів за формулами (2.32)–(2.34), зовнішнього критерію відносно до формули (2.35) та критерію балансу на основі функції бажаності

Харрінгтона відповідно до алгоритму представленому у підрозділі 2.1.1.2, при цьому у формулах (2.22) – (2.24) знак “+” змінюється на “-”.

1.4. Формування функції розрахунку точності класифікації об’єктів, що досліджуються за формулою (2.18).

1.5. Формування двох еквівалентних підмножин профілів експресії генів ітеративним розподілом двох найближчих профілів експресії генів за комплексною метрикою оцінки відстані між профілями у відповідні еквівалентні підмножини даних. Якщо кількість профілів експресії генів, що складають базу даних є непарною, останній профіль розподіляється у обидві підмножини даних.

## **Етап II.** Налаштування щільнісного алгоритму кластеризації OPTICS.

2.1. Завдання інтервалу зміни мінімальної кількості об’єктів усередині  $\varepsilon$ -околиці  $MinPts_{min}$ ,  $MinPts_{max}$ .

2.2. Створення діаграми досяжності. Завдання інтервалу та кроку зміни радіусу  $\varepsilon$ -околиці  $Eps_{min}$ ,  $Eps_{max}$ ,  $dEps$ .

2.3. Розрахунок відстаней між усіма парами профілів експресії генів у еквівалентних підмножинах даних та формування матриць відстаней між відповідними профілями. Отримані матриці відстаней будуть використані як вхідні дані при реалізації процесу кластеризації застосуванням алгоритму OPTICS.

**Етап III.** Покрокова кластеризація профілів експресії генів у інтервалах зазначених діапазонів зміни відповідних параметрів алгоритму.

3.1. Ініціалізація значення  $MinPts$ :  $k = MinPts_{min}$ .

3.2. Ініціалізація вихідного значення  $Eps$ :  $e = Eps_{min}$ .

3.3. Кластеризація профілів експресії генів, що містяться у еквівалентних підмножинах даних, формування кластеризацій з кількістю кластерів  $K1$  і  $K2$ .

3.4. Якщо  $K1=K2 > 2$ , розрахунок внутрішніх і зовнішніх критеріїв за формулами (2.32)–(2.35), у протилежному випадку – збільшення значення  $Eps$  на один крок зміни цього параметру ( $e = e + de$ ).

3.5. Класифікація об’єктів, що містять профілі експресії генів у кожному з отриманих кластерів. Розрахунок критерію якості класифікації даних (Точність).

3.6. Якщо  $e \leq Eps_{max}$ , перехід на крок 3.3 даної процедури. У протилежному випадку, розрахунок критерія балансу на основі функції бажаності Харрінгтона відповідно до алгоритму представленому у підрозділі 2.1.1.2 і збільшення значення  $MinPts$  на одиницю:  $k = k + 1$ .

3.7. Якщо  $k \leq MinPts_{max}$ , перехід на крок 3.2 цієї процедури у протилежному випадку – побудова діаграм критеріїв якості кластеризації та класифікації залежно від значення  $Eps$  для кожного значення  $MinPts$ .

## **Етап IV.** Аналіз отриманих результатів.

4.1. Аналіз отриманих діаграм. Формування висновків щодо ефективності застосування комплексної метрики оцінки відстані між профілями експресії генів в процесі формування підмножин інформативних генів для подальшого їх застосування у системах діагностики складних захворювань або при реконструкції генних регуляторних мереж.

Практична реалізація запропонованої моделі здійснювалася із застосуванням даних експресії генів GSE19188 пацієнтів, що досліджувалися на ранню стадію раку легенів, при цьому як і у попередніх експериментах, було виділено 401 найбільш інформативних за точністю класифікації профілів експресії генів. Значення *MinPts* змінювалося від 3 до 5. Цей діапазон був встановлений емпіричним шляхом. Результати моделювання показали, що збільшення кількості точок всередині *Eps* околиці погіршує результати моделювання як з точки зору кількості кластерів у еквівалентних підмножинах даних, так і за критеріями кластеризації та класифікації об'єктів. Значення *Eps* змінювалося від мінімального, що розраховувалася як мінімальна відстань між профілями експресії генів у еквівалентних підмножинах даних до 1.5 мінімальної відстані. Цей інтервал також встановлений емпіричним шляхом. При більших значеннях *Eps* профілі експресії генів розділялися на 2 кластера, при цьому результати повторювалися. Отриманий діапазон зміни значень *Eps* був розділений на 20 рівних інтервалів. Ширина інтервалу дорівнювала кроку зміни значення *Eps*. Відповідно до алгоритму, структурну блок-схему якого зображено на рис. 2.23, критерії якості кластеризації профілів експресії генів та класифікації відповідних об'єктів розраховувалися тільки для випадків, якщо кількість кластерів, виділених на еквівалентних підмножинах даних є рівною. Ця умова мінімізує похибку відтворюваності.

Результати моделювання наведені на рис. 2.24 та 2.25 та у таблицях 2.3–2.4.

Таблиця 2.3

Результат розділення профілів експресії генів на кластери при *MinPts* = 3

| EPS,*10 <sup>-3</sup> | Кластери |     |   |    |   |   |
|-----------------------|----------|-----|---|----|---|---|
|                       | 1        | 2   | 3 | 4  | 5 | 6 |
| 0.66435               | 24       | 311 | 6 | 14 | 6 | 7 |
| 0.69525               | 24       | 322 | 6 | 14 | 6 | – |
| 0.71070               | 24       | 323 | 6 | 14 | 6 | – |
| 0.72615               | 24       | 329 | 6 | 14 | 6 | – |
| 0.74160               | 24       | 332 | 6 | 14 | 6 | – |
| 0.91155               | 24       | 359 | 6 | –  | – | – |
| 0.92700               | 24       | 359 | 6 | –  | – | – |

Аналіз отриманих результатів дозволяє зробити висновок щодо доцільності застосування запропонованої метрики для виділення взаємно корельованих профілів експресії генів у випадку використання мульти-кластерної структури шляхом застосування алгоритму кластеризації OPTICS. Запропонований метод дозволяє визначити оптимальні параметри алгоритму, що відповідають оптимальному характеру групування профілів експресії генів у кластери, з одного боку, так і мінімуму похибки відтворюваності, з іншого боку. З даних таблиць 2.3 і 2.4 зрозуміло, що при значенні параметра алгоритму OPTICS *MinPts* = 3, виділено сім кластеризацій.

Таблиця 2.4

Результат розділення профілів експресії генів на кластери при  $MinPts = 4,5$ 

| <i>EPS</i> ,*10 <sup>-3</sup> | <i>MinPts</i> = 4 |     |     |    | <i>EPS</i> ,*10 <sup>-3</sup> | <i>MinPts</i> = 5 |     |     |    |
|-------------------------------|-------------------|-----|-----|----|-------------------------------|-------------------|-----|-----|----|
|                               | Кластери          |     |     |    |                               | Кластери          |     |     |    |
|                               | 1                 | 2   | 3   | 4  |                               | 1                 | 2   | 3   | 4  |
| 0.64890                       | 24                | 158 | 115 | 11 | 0.64890                       | 23                | 155 | 115 | 11 |
| 0.69525                       | 24                | 322 | 14  | –  | 0.66435                       | 24                | 308 | 14  | –  |
| 0.71070                       | 24                | 323 | 14  | –  | 0.67980                       | 24                | 314 | 14  | –  |
| 0.72615                       | 24                | 326 | 14  | –  | 0.69525                       | 24                | 321 | 14  | –  |
| 0.74160                       | 24                | 331 | 14  | –  | 0.71070                       | 24                | 322 | 14  | –  |
| –                             | –                 | –   | –   | –  | 0.72615                       | 24                | 326 | 14  | –  |
| –                             | –                 | –   | –   | –  | 0.74160                       | 24                | 331 | 14  | –  |

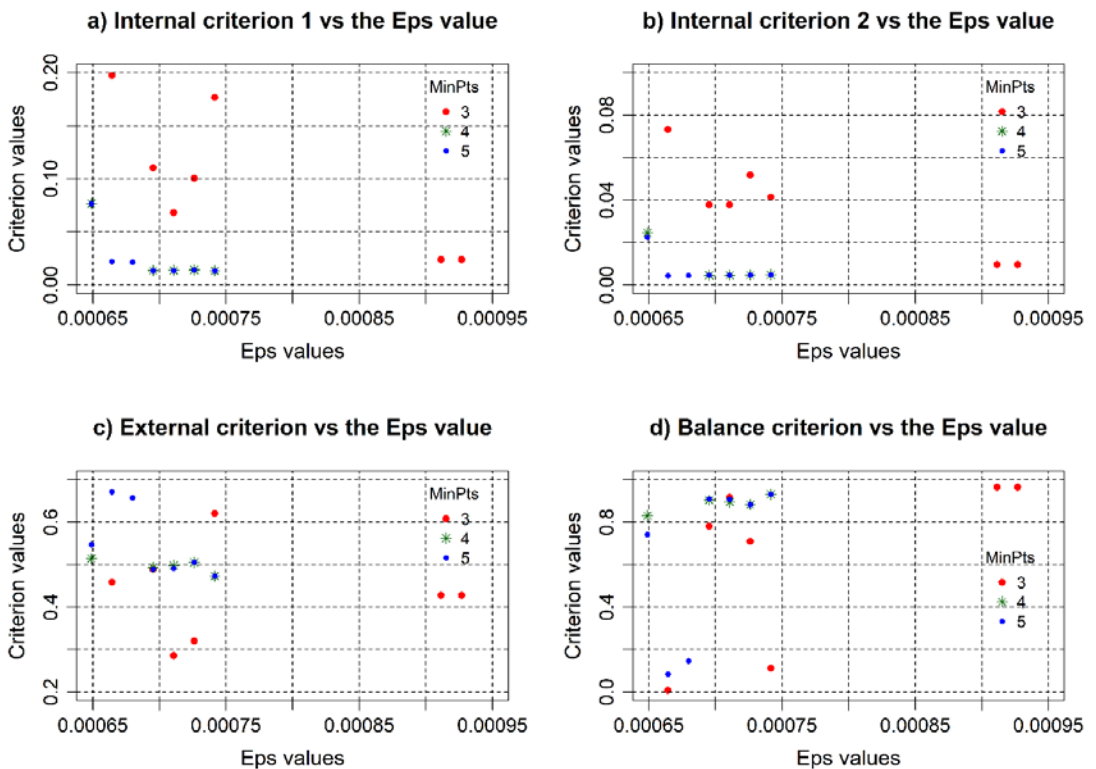


Рис. 2.24. Результати моделювання щодо критеріального аналізу кластерної структури при використанні алгоритму OPTICS, реалізованого із застосуванням індуктивної технології об'єктивної кластеризації: розподіл внутрішнього критерію якості кластеризації, розрахованого на першій (а) та другій (б) еквівалентних підмножинах профілів експресії генів, зовнішнього критерію (с) та комплексного критерію балансу (д) при різних значеннях  $Eps$  та  $MinPts$

Перша кластеризація містить шість кластерів, чотири кластеризації містять по п'ять кластерів і останні дві кластеризації містять три кластера. У випадку значень параметра *MinPts* 4 або 5, перша кластеризація містила чотири кластера, у інших випадках отримано по три кластера у кожній кластеризації. При цьому слід зазначити, що вихідні дані містили приблизно 400 профілів експресії генів, що були ретельно відібрані багатокроковим застосуванням алгоритму кластеризації SOTA. При цьому точність класифікації об'єктів, які містили як атрибути повний набір профілів експресії генів, складала приблизно 93%. Аналіз отриманих результатів показує, що в усіх випадках деяка частина профілів експресії генів ідентифікована як шум. Ці гени не містяться у жодному кластері.

Наявність «шумових» генів пояснюється тим, за метрикою, що використовується їхня щільність розташування у просторі ознак менша за умовне граничне значення, яке визначається параметрами алгоритму кластеризації OPTICS.

Аналіз діаграм, наведених на рис. 2.24, дозволяє зробити висновок, що як внутрішні, так і зовнішні критерії якості кластеризації не дозволяють визначити оптимальні параметри алгоритму кластеризації OPTICS. Мінімальні значення цих критеріїв не відповідають максимальним значенням точності класифікації об'єктів у відповідних кластерах. Максимальне значення комплексного критерія балансу, який містить як компоненти внутрішні та зовнішні критерії досягається у випадку трикластерної структури при параметрах алгоритму OPTICS:  $MinPts = 3$ ,  $Eps = 0.00091155$  або  $Eps = 0.00092700$  (досягаються однакові результати). Результати класифікації об'єктів, що містяться у відповідних кластерах і які представлені на рис. 2.25, підтверджують отримані результати.

Як можна бачити, при цих параметрах алгоритму точність класифікації є максимальною для перших двох кластерів, при цьому другий кластер містить найбільшу кількість генів, тобто він є основним з точки зору кількості профілів експресії генів. Третій кластер містить тільки шість генів. Результати класифікації у четвертому, п'ятому та шостому кластерах не представлені, оскільки вони однакові в усіх випадках і трохи гірші за результати класифікації у перших трьох кластерах. Слід відзначити, що максимальні значення комплексного критерію балансу, що визначає якість кластеризації профілів експресії генів відповідає максимальним значенням точності класифікації об'єктів, які містять як атрибути виділені профілі експресії генів, що свідчить про високу ефективність запропонованої метрики та моделі кластеризації профілів експресії генів.

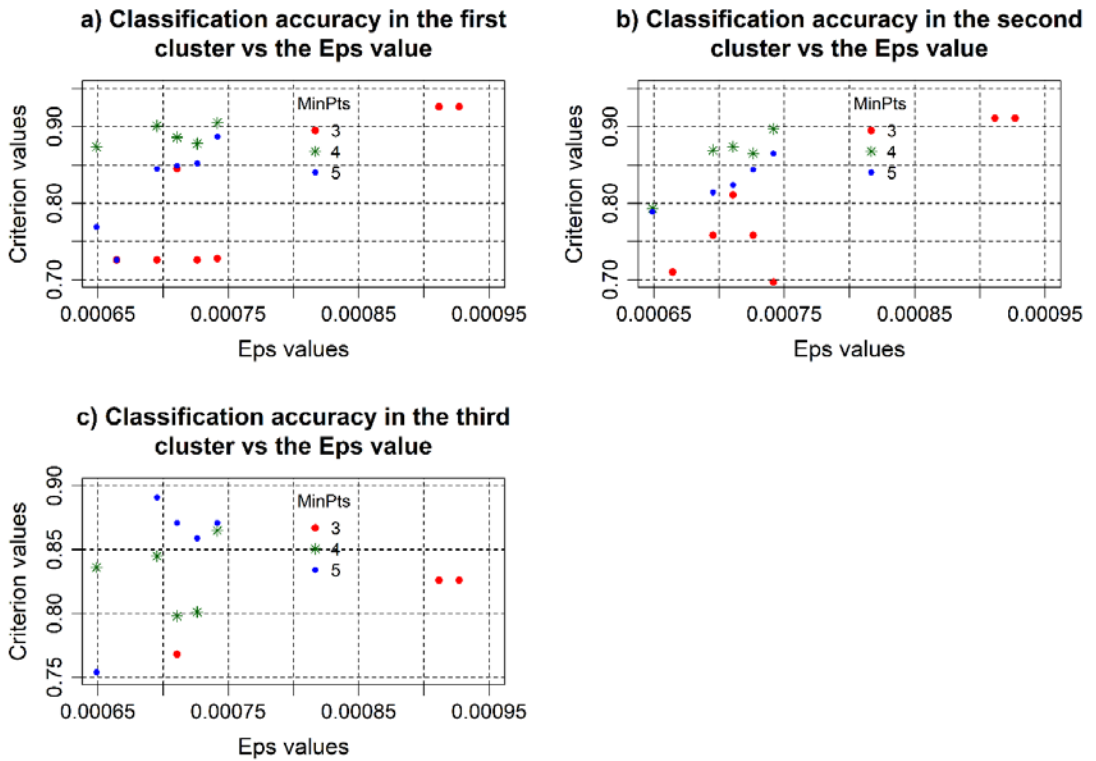


Рис. 2.25. Результати моделювання щодо оцінки точності класифікації об'єктів, атрибутами яких є значення експресії генів, профілі яких виділені у кластери при застосуванні алгоритму OPTICS: а) перший кластер; б) другий кластер; в) третій кластер

## **Розділ 3.**

### **Теоретичні та практичні засади індуктивної гібридної моделі кластеризації профілів експресії генів**

У цьому розділі представлені теоретичні та практичні дослідження щодо застосування індуктивних моделей кластеризації профілів експресії генів для формування підмножин взаємно корельованих профілів, що дозволяють з високою точністю ідентифікувати об'єкти, які містять як атрибути значення експресії виділених генів [6, 9–13, 15, 19–21, 30, 31]. Складність розв'язання поставленої задачі визначаються високою розмірністю профілів експресії генів. Кількість значень експресії генів у відповідних профілях дорівнює кількості досліджуваних об'єктів. Більшість наявних алгоритмів кластеризації орієнтовані на низьковимірні дані, що містять невелику кількість атрибутів. Крім того, більшості наявним алгоритмам кластеризації даних властива похибка відтворюваності, тобто задовільні результати кластеризації, отримані на одній множині даних не завжди повторюються при застосуванні іншої аналогічної підмножини даних. Ця проблема може бути вирішена ретельною валідацією отриманої моделі із застосуванням «свіжих» даних (не використовуються при створенні моделі), яка містить об'єкти, що близькі за своїми властивостями об'єктам початкової підмножини даних. У цьому розділі ця проблема вирішується шляхом реалізації існуючих сучасних алгоритмів кластеризації даних у рамках індуктивної технології об'єктивної кластеризації, ідея якої представлена у працях [160 – 164] і отримала подальший розвиток у працях [166, 169]. Крім того, велика кількість профілів експресії генів у початкових даних ( $\approx 10000$ ) викликає необхідність розробки ієрархічних гібридних моделей покрокової кластеризації профілів експресії генів з оцінкою якості групування профілів експресії генів на кожному ієрархічному рівні шляхом оцінки якості класифікації об'єктів, які містять виділені профілі експресії генів у якості атрибутів. Розв'язок вищенаведених задач наведено у цьому розділі.

#### **3.1. Концепція та особливості застосування індуктивних методів аналізу складних процесів при створенні індуктивних моделей кластеризації даних**

В основу індуктивної технології об'єктивної кластеризації покладені три фундаментальних принципи, що запозичені з методології індуктивного моделювання складних систем [160–164] і які дозволяють створити цілісну теорію формування кластерної структури із застосуванням внутрішніх та зовнішніх критеріїв, аналіз яких створює умови для мінімізації похибки відтворюваності при отриманні бажаної кластерної структури:

**1. Принцип послідовного перебору.** Реалізація цього принципу передбачає послідовний перебір кластеризацій на двох еквівалентних підмножинах даних (містять однакову кількість попарно близьких об'єктів) з розрахунком як внутрішніх критеріїв якості групування об'єктів у кожній кластеризації (за умовою рівної кількості класте-

рів), так і зовнішніх критеріїв, які визначають різницю між результатами кластеризацій об'єктів, що містяться у еквівалентних підмножинах даних.

**2. Принцип зовнішнього доповнення.** У рамках методології індуктивного моделювання складних систем реалізація цього принципу передбачає застосування «свіжої інформації» для верифікації отриманої моделі, тобто підмножини об'єктів, які за своїми властивостями є близькими до об'єктів, що містяться у основній множині даних. У випадку коректної моделі, результати кластеризації на основній множині даних і на додатковій підмножині повинні збігатися у межах допустимої помилки. При реалізації індуктивної технології об'єктивної кластеризації роль «свіжої інформації» відіграє еквівалентна підмножина даних. Тобто на першому кроці вихідні дані, кластеризація яких передбачається, розділяються на дві еквівалентні підмножини, що містять однакову кількість попарно близьких об'єктів відповідно до наступного алгоритму [246]:

- розрахунок  $\frac{n \cdot (n-1)}{2}$  попарно близьких об'єктів відповідно до метрики близькості, яка використовується для даних, що досліджуються. Результатом реалізації цього кроку є трикутна матриця відстаней між усіма парами об'єктів, що містить вихідна множина даних:

$$D = \begin{pmatrix} 0 & d(x_1, x_2) & d(x_1, x_3) & \dots & d(x_1, x_m) \\ & 0 & d(x_2, x_3) & \dots & d(x_2, x_m) \\ \dots & \dots & \dots & \dots & \dots \\ & & & 0 & d(x_{n-1}, x_n) \\ & & & & 0, \end{pmatrix}$$

де  $x_i$  –  $i$ -й об'єкт у вихідних даних;

- формування пустих підмножин даних  $A$  і  $B$ , які на наступному кроці будуть поповнюватися попарно близькими об'єктами;
- виділення пари об'єктів  $x_s$  і  $x_p$ , відстань між якими за метрикою, що використовується, є мінімальною:

$$d(x_s, x_p) = \min_{i,j, i \neq j} d(x_i, x_j),$$

- розподіл об'єкта  $x_s$  у підмножину  $A$ , а об'єкта  $x_p$  у підмножину  $B$ ;
- повторювання кроків 3 та 4 для усіх пар об'єктів, що залишилися. Якщо кількість об'єктів у вихідній множині даних є непарним, останній об'єкт розподіляється у обидві підмножини.

Як результат, отримаємо дві еквівалентні підмножини даних, що містять однакову кількість попарно близьких об'єктів.

**3. Принцип неостаточності рішень.** Реалізація цього принципу у рамках індуктивної технології об'єктивної кластеризації передбачає фіксацію проміжних кластеризацій, які відповідають екстремумам внутрішніх та зовнішніх критеріїв якості кластеризації. Однак слід зазначити, що внутрішні і зовнішні критерії можуть суперечити між собою. Тому остаточної вибір оптимальної кластеризації здійснюється на основі комплексного критерію балансу, який повинен включати як компо-

ненти як внутрішні, так і зовнішні критерії якості кластеризації. Для розрахунку комплексного критерію балансу у роботах [166, 169] авторами запропоновано використовувати функцію бажаності Харрінгтона робота [239], рівняння якої представлено формулою (3.1), а графік зображений на рис. 3.1.

$$d = \exp(-\exp(-Y)), \quad (3.1)$$

де  $d$  є значення бажаності, що варіюється від 0 (найгірша кластеризація) до 1 (найкраща кластеризація), а значення безрозмірного показника  $Y$  змінюється за лінійним законом від  $-2(d = 0)$  до  $5 (d = 1)$ .

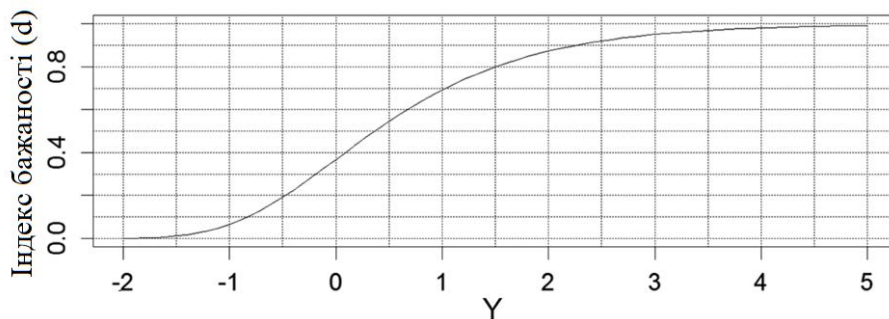


Рис. 3.1. Функція бажаності Харрінгтона

Особливості даної функції полягають у тому, що вона має дві точки перегину:  $0.37 = 1/e$  та  $0.63 = 1 - 1/e$ . Дані точки визначають межу допустимих значень за шкалою бажаності. У таблиці 3.1 наведено діапазони допустимих значень за шкалою бажаності та відповідні їм терми.

Таблиця 3.1

*Діапазони допустимих значень за шкалою бажаності та відповідні їм терми*

| № | Діапазон варіювання індексу бажаності | Терми           |
|---|---------------------------------------|-----------------|
| 1 | 0.8 – 1                               | Відмінно        |
| 2 | 0.63 – 0.8                            | Добре           |
| 3 | 0.37 – 0.63                           | Задовільно      |
| 4 | 0.2 – 0.37                            | Погано          |
| 5 | 0 – 0.2                               | Не сприйнятливо |

Метод бажаності Харрінгтона являє собою один із методів розв’язання багатокритеріальних задач, при цьому нормалізація значень різних критеріїв, що застосовуються у рамках моделі, відбувається на першому кроці шляхом трансформації шкал відповідних критеріїв у шкалу показника  $Y$  за лінійною (у більшості випадків) або нелінійною (у окремих випадках враховуючи характер розподілу даних) шкалою.

Кількість внутрішніх критеріїв якості кластеризації даних визначаються типом даних, об’єкти яких необхідно розділити на кластери. Застосування індуктивної

технології об'єктивної кластеризації передбачає формування векторів внутрішніх критеріїв на еквівалентних підмножинах  $A$  та  $B$  в процесі послідовного перебору кластеризацій:

$$QC_{int}^A = \{QC_{int_i}^A\}; QC_{int}^B = \{QC_{int_i}^B\}, i = \overline{1, m}, \quad (3.2)$$

де  $m$  є кількість внутрішніх критеріїв, що розраховуються для оцінки якості групування об'єктів у окремих кластеризаціях.

Як було зазначено раніше, найкращі кластеризації за внутрішніми критеріями якості можуть не бути об'єктивними за похибкою відтворюваності, тобто екстремуми внутрішніх критеріїв (3.2), отриманих для еквівалентних підмножин даних, можуть суперечити між собою. Для мінімізації похибки відтворюваності розраховується зовнішній критерій якості кластеризації, який враховує розбіжності між результатами групування об'єктів у еквівалентних підмножинах даних. Застосування стандартних зовнішніх критеріїв якості розподілу об'єктів у різних кластеризаціях, а саме індекси Жаккара, Кульчинського тощо не є ефективними у цьому випадку, оскільки їх використання передбачає кластеризації на однакових об'єктах. Ця умова у випадку індуктивної технології об'єктивної кластеризації не виконується, оскільки еквівалентні підмножина даних містять однакову кількість близьких за своїми властивостями об'єктів. Однак слід зазначити, що мінімум похибки відтворюваності відповідає найбільшій збіжності характеру розподілу об'єктів у кластери у еквівалентних підмножинах даних. Оскільки внутрішні критерії відображають характер розподілу об'єктів у відповідних кластерах, то можна припустити, що мінімальне значення похибки відтворюваності може бути досягнуто при мінімальній різниці між значеннями внутрішніх критеріїв, розрахованих на еквівалентних підмножинах об'єктів  $A$  та  $B$ . У роботах [166, 169] авторами запропоновано метод розрахунку зовнішнього критерію якості кластеризації як нормалізованої різниці відповідних внутрішніх критеріїв, отриманих на еквівалентних підмножинах даних:

$$QC_{ext} = \left\{ \frac{|QC_{int_i}^A - QC_{int_i}^B|}{QC_{int_i}^A + QC_{int_i}^B} \right\}, i = \overline{1, m}. \quad (3.3)$$

Алгоритм розрахунку комплексного критерію балансу на основі методу бажаності Харрінгтона передбачає наступні кроки:

1. Трансформування шкал отриманих заздалегідь внутрішніх та зовнішніх критеріїв якості кластеризації у шкалу показника  $Y$  відповідно із системою лінійних рівнянь:

$$Y_i = \begin{cases} a \pm b \cdot QC_{int_i} \\ a - b \cdot QC_{ext_i} \end{cases}, i = \overline{1, m}, \quad (3.4)$$

де знак у верхньому рівнянні визначається характером зміни значень відповідного внутрішнього критерію у процесі формування кластерної структури. Знак «+» обирається у випадку, якщо оптимальній заданим критерієм кластеризації відповідає максимальне значення цього критерія (позитивна кореляція між векторами  $Y$  і  $QC$ ), у протилежному випадку ставиться знак «-». Значення параметрів  $a$  і  $b$

визначаються емпіричним шляхом, враховуючи граничні значення відповідних критеріїв, характер їх зміни та граничні значення шкали показника  $Y$ :

$$\begin{cases} Y_{max} = a + b \cdot QC_{max} \\ Y_{min} = a + b \cdot QC_{min} \end{cases} \text{ або } \begin{cases} Y_{max} = a - b \cdot QC_{min} \\ Y_{min} = a - b \cdot QC_{max} \end{cases}. \quad (3.5)$$

2. Формування вектора значень показника  $Y = \{Y_i\}$  відповідно до лінійного рівняння:

$$Y_i = a + b \cdot QC_i \text{ або } Y_i = a - b \cdot QC_i. \quad (3.6)$$

3. Розрахунок приватних бажаностей для кожного значення показника  $Y$ , отриманого на попередньому кроці:

$$d_i = \exp(-\exp(-Y_i)). \quad (3.7)$$

4. Розрахунок комплексного критерію балансу як середнє геометричне усіх приватних бажаностей, розрахованих за формулою (3.7):

$$QC_{bal} = \sqrt[n]{\prod_{i=1}^n d_i}. \quad (3.8)$$

Відповідно до даних таблиці 3.1, максимальне значення критерію балансу відповідає оптимальній кластеризації, як з точки зору розподілу об'єктів у кластерах, так і з точки зору мінімального значення похибки відтворюваності. Слід зазначити, що у процесі моделювання можлива наявність декількох локальних максимумів критерію балансу, які мало розрізняються між собою. У цьому випадку остаточне рішення приймається, виходячи із мети поставленої задачі, враховуючи бажану кількість кластерів, на які необхідно розділити експериментальні дані.

Наведені вище теоретичні дослідження та висновки створюють умови для розробки покрокової процедури формування кластерної структури на основі принципів індуктивної технології об'єктивної кластеризації.

### **3.1.1. Архітектура та концептуальний опис покрокової процедури формування кластерної структури на основі принципів індуктивної технології об'єктивної кластеризації**

Структурна блок-схема покрокової процедури формування кластерної структури на основі принципів індуктивної технології індуктивної кластеризації зображена на рис. 3.2. Реалізація цієї процедури передбачає такі етапи:

**Етап I. Постановка проблеми.** Аналіз експериментальних даних. Формування мети. Ініціалізація моделі.

1.1. Постановка проблеми. Аналіз експериментальних даних. Формування мети.

1.2. Ініціалізація моделі. Реалізація цього кроку передбачає вибір метрики оцінки ступеня близькості даних, що досліджуються, формування вектору внутрішніх критеріїв кластеризації залежно від типу даних, вибір алгоритму кластеризації та ініціалізація діапазону та кроку зміни параметрів алгоритму ( $p_{min}, p_{max}, dp$ ).

1.3. Формування двох еквівалентних підмножин даних  $A$  і  $B$  за метрикою, визначеною на кроці 1.2 даної процедури із застосуванням алгоритму, наведеному у підрозділі 3.1.

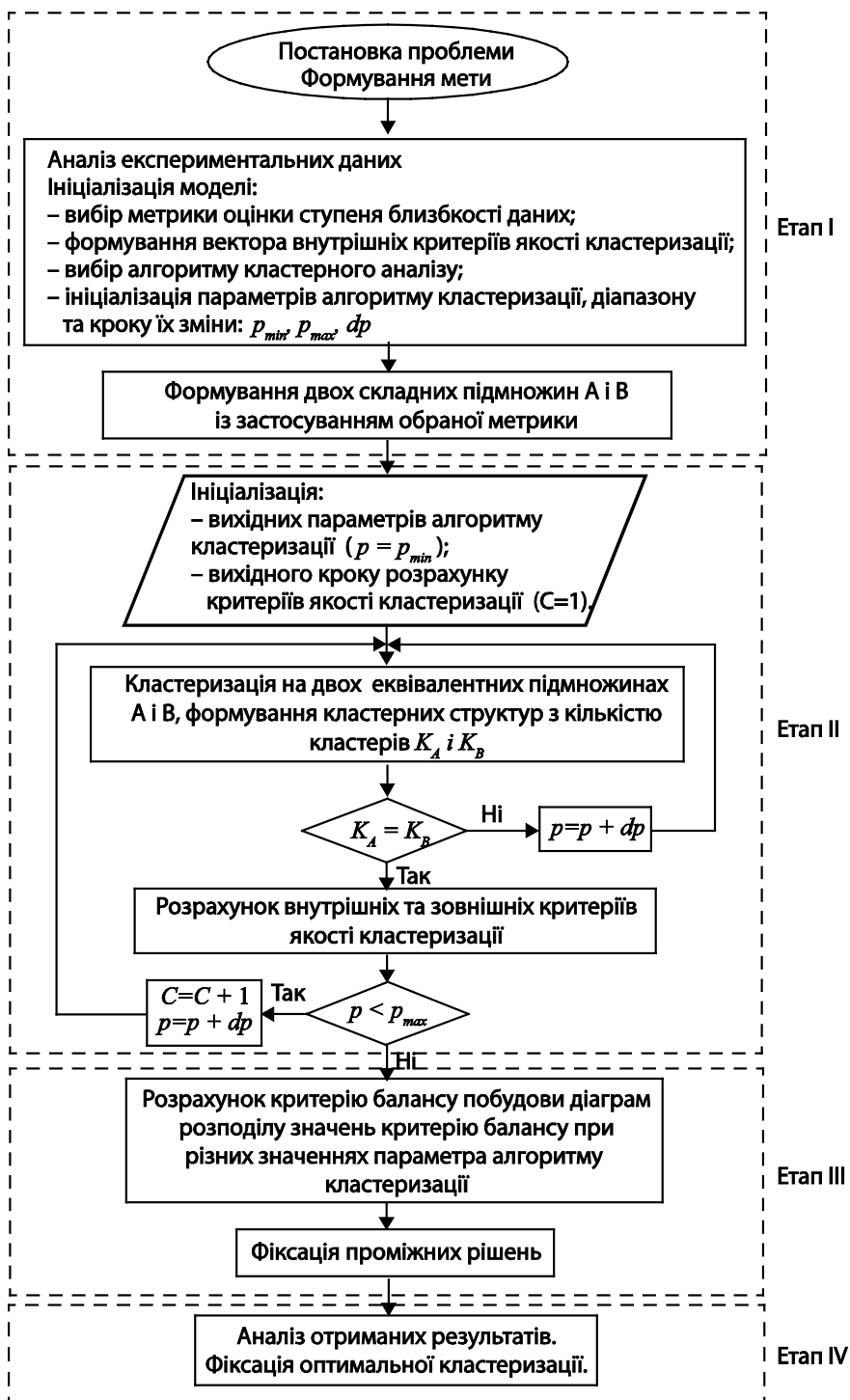


Рис. 3.2. Блок-схема покрокової процедури формування кластерної структури на основі принципів індуктивної технології об'єктивної кластеризації

**Етап II. Кластеризація даних.** Формування векторів значень внутрішніх та зовнішніх критеріїв якості кластеризації.

2.1. Ініціалізація вихідних параметрів алгоритму кластеризації ( $p = p_{\min}$ ) та вихідного кроку розрахунку внутрішніх критеріїв якості кластеризації ( $C = 1$ );

2.2. Кластеризація даних на двох еквівалентних підмножинах  $A$  і  $B$ . Формування кластерної структури з кількістю кластерів  $K_A$  і  $K_B$ .

2.3. Якщо кількість кластерів у кластеризаціях на підмножинах  $A$  і  $B$  є однаковою ( $K_A = K_B$ ), то розрахунок внутрішніх критеріїв  $QC_{int}^A$  і  $QC_{int}^B$  та зовнішніх критеріїв  $QC_{ext}$  для кожної пари внутрішніх критеріїв визначається за формулою (3.3).

**Етап III. Розрахунок критерію балансу.** Формування проміжних рішень щодо оптимальної кластерної структури.

3.1. Розрахунок критерію балансу за формулами (3.4)–(3.8).

3.2. Побудова діаграм розподілу значень критерію балансу при різних значеннях параметра алгоритму кластеризації.

3.3. Фіксація проміжних рішень, тобто параметрів алгоритму, які відповідають максимальним значенням критерію балансу.

**Етап IV. Аналіз отриманих результатів.** Фіксація оптимальної кластеризації.

4.1. Аналіз отриманих результатів. Реалізація цього кроку передбачає вибір критеріїв, за якими обирається оптимальна кластеризація. У випадку застосування цієї технології для кластеризації профілів експресії генів кластеризуються атрибути відповідних об'єктів (профілі експресії генів). Здебільшого заздалегідь відомі класи, на які можуть бути розділені дані об'єкти (анотація даних). Таким чином, оцінка кластерної структури може бути виконана шляхом реалізації процедури класифікації об'єктів, що містять як атрибути профілі експресії генів, котрі сконцентровані у відповідних кластерах із розрахунком критеріїв якості класифікації з урахуванням похибок першого та другого роду. У рамках запропонованої моделі індуктивної технології об'єктивної кластеризації якість групування профілів експресії генів оцінювалася шляхом розрахунку точності класифікації об'єктів, яка загалом визначає ймовірність коректного розподілу об'єктів у відповідні класи:

$$ACC = \frac{\sum_{i=1}^{N_{CL}} TP_i}{N_{obj}}, \quad (3.9)$$

де  $TP_i$  є кількість коректно розпізнаних об'єктів у  $i$ -му класі;  $N_{CL}$  – кількість класів;  $N_{obj}$  – загальна кількість об'єктів, що містять тестові дані.

Очевидно, що оптимальна кластеризація у цьому випадку відповідає максимальному значенню точності класифікації об'єктів, що містять виділені профілі експресії генів з урахуванням мети поставленої задачі.

### 3.2. Індуктивна гібридна модель кластеризації даних на основі щільнісних алгоритмів

Щільнісні алгоритми є одними із сучасних алгоритмів кластерного аналізу, перевага яких полягає в тому, що вони дають можливість виокреслити кластери складної форми, з одного боку, та об'єкти, які за щільністю їх розподілу у просторі можуть бути ідентифіковані як шум, оскільки вони за щільністю не попадають у жоден кластер, з іншого боку. До таких типів алгоритмів належать DBSCAN (Density-Based Spatial Clustering and Application with Noise) [241] та його логічне продовження алгоритм OPTICS (Ordering Points To Identify the Clustering Structure) [240]. Основу щільнісних алгоритмів складають твердження, наведені в працях [240, 241]:

1.  $\varepsilon$ -околицею точки  $p$  ( $Eps(p)$ ) є сфера радіусом  $\varepsilon$ , кількість точок всередині якої не менша за граничну кількість точок  $MinPts$ , що встановлюється заздалегідь:

$$N_{Eps}(p) = \{q \in D | d(q, p) \leq Eps(p)\},$$

де  $D$  – множина точок у просторі ознак;  $d(q, p)$  – відстань між точками  $q$  і  $p$  за обраною метрикою.

2. Точка  $p$  є безпосередньо щільнісно досяжна з точки  $q$ , якщо точка  $p$  знаходиться всередині  $\varepsilon$ -околиці точки  $q$  ( $p \in Eps(q)$ ) і кількість точок всередині  $Eps(q)$  не менша за граничну кількість точок  $MinPts$  ( $N_{Eps}(p) \geq MinPts$ ).

3. Точка  $p$  є щільнісно досяжна з точки  $q$ , якщо існує такий ланцюг точок  $p_1, p_2, \dots, p_n$ , де  $p_1 = q, ap_n = p$ , що для будь-якої пари сусідніх точок  $p_{i-1}, p_i$  виконується умова безпосередньої щільнісної досяжності, тобто точка  $p_i$  є безпосередньо щільнісно досяжна з точки  $p_{i-1}$ .

4. Точки  $p$  і  $q$  є щільнісно з'єднані між собою, якщо існує така точка  $s$ , що точки  $p$  і  $q$  є щільнісно досяжні з точки  $s$ .

5.  $\forall p, q_i, i = \overline{1, n}$ , якщо точки  $q_i$  є щільнісно досяжні від точки  $p$  або щільнісно з'єднані між собою та з точкою  $p$ , то точки  $p$  і  $q_i$  утворюють кластер.

6. Якщо точка  $s$  не є щільнісно досяжною або щільнісно з'єднаною з іншими точками даних, то така точка ідентифікується як шум.

Як бачимо, результат кластеризації у разі застосування щільнісних алгоритмів визначається двома параметрами:  $Eps$  та  $MinPts$ , комбінація яких визначає кластерну структуру.

У праці [241] автори запропонували метод оцінки значення  $Eps$  при заздалегідь заданому значенні  $MinPts$  на основі  $k$ -dist графу ( $k$  у такому випадку дорівнює  $MinPts$ ), що являє собою розподіл значень  $Eps$  при різній кількості точок усередині  $Eps$  околиці. Приклад такого графу зображений на рис. 3.3.

Значення  $MinPts$  у цьому випадку дорівнювало 4 (4-dist). За твердженням авторів, оптимальне значення  $Eps$  при цьому значенні  $MinPts$  знаходиться на коліні цього графу, тобто в діапазоні від 1 до 1,5. Однак слід зазначити, що точне встановлення значення  $Eps$  аналіз  $k$ -dist графу не дозволяє, можна тільки встановити інтервал, у якому знаходиться оптимальне значення  $Eps$ . Цей факт ускладнює коректне застосування алгоритму кластеризації DBSCAN, оскільки результат

роботи алгоритму має високу чутливість до значення параметра  $Eps$ , і воно зазвичай підлаштовується емпіричним шляхом у процесі моделювання.

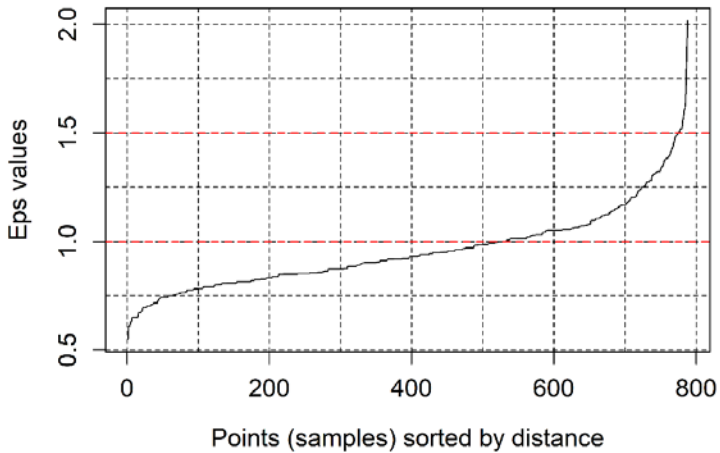


Рис. 3.3. Приклад  $4\text{-}dist$  графу для визначення оптимального значення  $Eps$

Як зазначалося вище, щільнісний алгоритм кластеризації OPTICS є логічним продовженням алгоритму DBSCAN. Автори цього алгоритму запропонували метод визначення оптимальних значень  $Eps$  і  $MinPts$  на основі діаграми досяжності, яка візуалізує розподіл значень  $Eps$  при різних значеннях щільності розподілу об'єктів у просторі ознак. Приклад діаграми досяжності для значення  $MinPts = 4$  зображений на рис. 3.4. До переваг цього методу, порівняно з методом  $k\text{-}dist$  графу, слід віднести суттєво меншу чутливість результату роботи алгоритму до значення  $Eps$ . Діапазон зміни значень  $Eps$  встановлюється шляхом аналізу діаграми досяжності так, щоб він включав необхідну кількість впадін, розділених вертикальними лініями. Кожна впадина відповідає кластеру, однак слід зазначити, що цьому алгоритму також властивий великий відсоток суб'єктивізму, оскільки оптимальне значення  $Eps$  також встановлюється емпіричним шляхом у процесі моделювання.

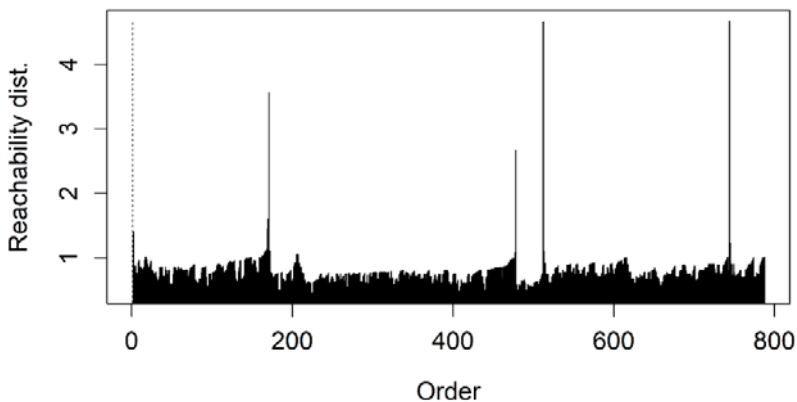


Рис. 3.4. Приклад діаграми досяжності для визначення оптимального значення  $Eps$

У праці [246] запропонована модель реалізації алгоритму кластеризації OPTICS із застосуванням принципів індуктивної технології об'єктивної кластеризації. Апробація запропонованої моделі проводилася із використанням синтетичних даних малої розмірності (набір точок у двовимірному просторі). У монографії вирішення проблеми об'єктивного визначення оптимальної комбінації значень  $Eps$  і  $MinPts$  на основі індуктивної технології об'єктивної кластеризації із застосуванням як внутрішніх, так і зовнішніх критеріїв якості кластеризації отримало подальший розвиток завдяки особливо ретельному вибору внутрішніх критеріїв кластеризації, глибокому аналізу результатів роботи моделі із застосуванням кількісних критеріїв якості та розширення типів даних, що використовуються для апробації моделі, включаючи дані експресії генів. Структурну блок-схему алгоритму реалізації індуктивної моделі об'єктивної кластеризації на основі щільнісних алгоритмів зображено на рис. 3.5. Практична реалізація цього алгоритму передбачає такі етапи:

#### **Етап I. Ініціалізація моделі.**

##### **1.1. Формування матриці експериментальних даних:**

$$G = \{g_{ij}\}, i = \overline{1, n}, j = \overline{1, m},$$

де  $g_{ij}$  – значення признаку, що відповідає  $i$ -му рядку та  $j$ -му стовпчику;  $n$  та  $m$  – кількість рядків та стовпців, відповідно.

1.2. Вибір метрики оцінки ступеня близькості об'єктів, кластерів, об'єктів та кластерів залежно від типу даних, що досліджуються.

1.3. Формування функцій розрахунку критеріїв якості кластеризації: внутрішніх, зовнішніх та балансу:  $QC_{int}$ ,  $QC_{ext}$ ,  $QC_{bal}$ .

1.4. Формування двох еквівалентних підмножин даних  $A$  і  $B$  відповідно до алгоритму, наведеному у розділі 3.1.

#### **Етап II. Кластеризація даних. Розрахунок критеріїв якості кластеризації.**

2.1. Формування діапазону зміни параметра  $MinPts$ :  $MinPts_{min} = 3, MinPts_{max}$ . Створення  $k$ -dist графу або діаграми досяжності для граничних значень цього параметра у випадку застосування алгоритму DBSCAN або OPTICS, відповідно.

Оцінка діапазону та кроку зміни параметра  $Eps$ :  $Eps_{min}$ ,  $Eps_{max}$ , шляхом аналізу  $k$ -dist графу або діаграми досяжності.

2.2. Ініціалізація вихідних значень параметрів  $MinPts$  та  $Eps$ :  $e = Eps_{min}$ ;  $k = MinPts_{min} = 3$ .

2.3. Кластеризація даних на двох еквівалентних підмножинах  $A$  і  $B$ . Формування відповідних кластерних структур.

2.4. Якщо кількість кластерів більша або дорівнює двом та кількість кластерів у двох кластеризаціях однакова ( $K_A = K_B \geq 2$ ), то виконується розрахунок внутрішніх та зовнішніх критеріїв якості кластеризації. У протилежному випадку збільшується значення параметра  $Eps$  на один крок ( $e = e + de$ ).

2.5. Якщо  $e \leq Eps_{max}$ , перехід на крок 2.3 цієї процедури. У протилежному випадку виконується розрахунок критерію балансу та збільшення значення  $k$  на одиницю ( $k = k + 1$ ).

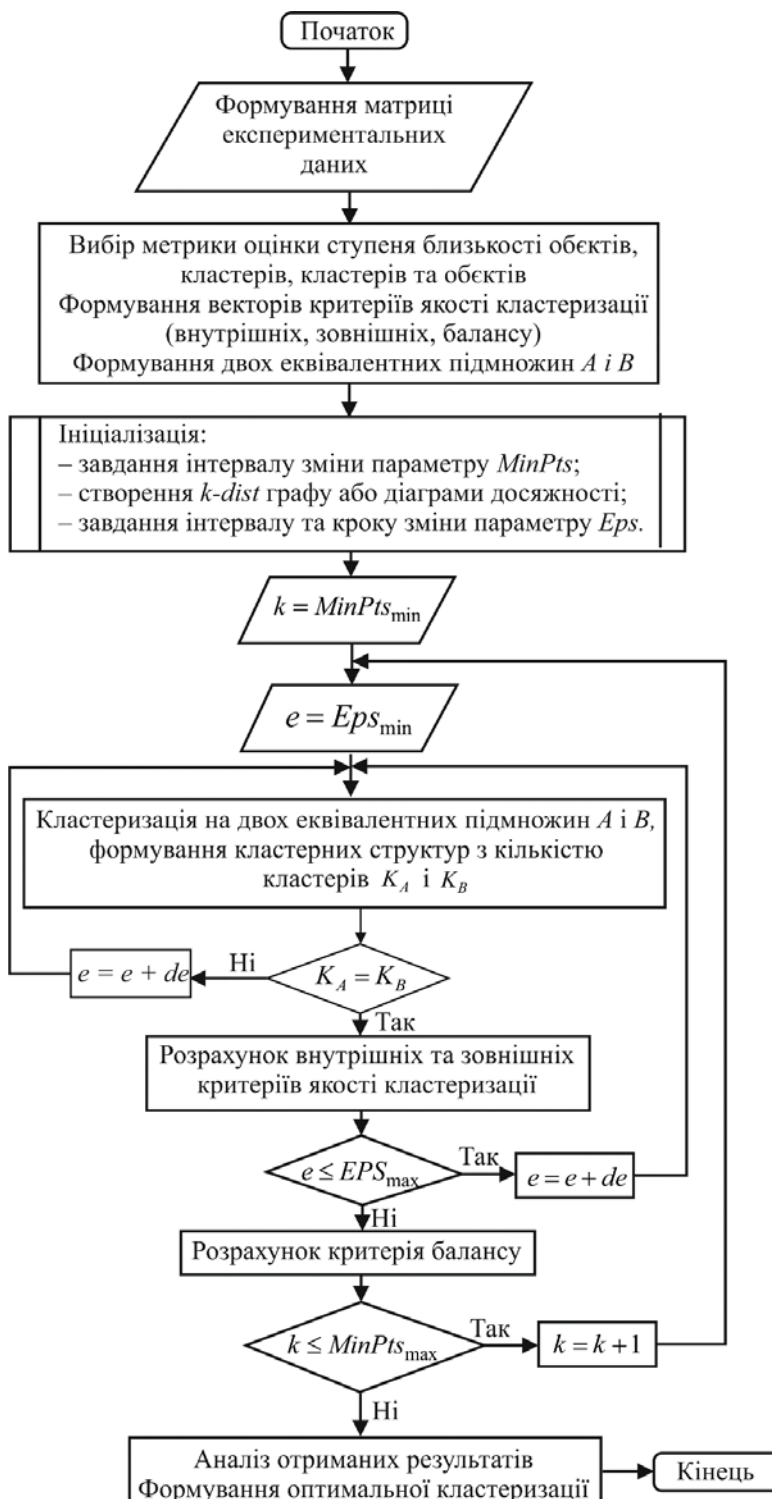


Рис. 3.5. Блок-схема алгоритму реалізації індуктивної моделі об'єктивної кластеризації на основі щільнісних алгоритмів

2.6. Якщо  $k \leq k_{max}$ , перехід на крок 2.3 цієї процедури. У протилежному випадку відбувається остаточне формування матриці критеріїв якості кластеризації.

**Етап III. Аналіз отриманих результатів.** Фіксація оптимальної кластеризації.

3.1. Створення діаграм залежності критерія балансу від значення  $Eps$  для кожного значення  $MinPts$ .

3.2. Фіксація найкращих кластеризацій за критерієм балансу у кожному випадку (відповідають максимумам критерію балансу).

3.3. Розрахунок критерію точності класифікації відповідних об'єктів за формулою (3.9) у випадку наявності необхідної інформації у анотації даних.

3.4. Фіксація оптимальних параметрів відповідного алгоритму кластерного аналізу, що відповідають максимуму критерію точності класифікації. Формування оптимальної кластеризації.

### 3.2.1. Формування векторів внутрішніх та зовнішніх критеріїв якості кластеризації залежно від типу експериментальних даних

Як було зазначено вище, одним із важливих кроків ініціалізації моделі кластеризації даних є формування векторів внутрішніх та зовнішніх критеріїв для даних, що досліджуються. У цьому підрозділі наведено дослідження щодо реалізації цього кроку, враховуючи тип експериментальних даних. Як експериментальні дані розглянуто такі:

1. Синтетичні дані «Compound» [246] та «Jain» [247], які містять кластери різної форми об'єктів у двовимірному просторі (рис. 3.6). Як бачимо, частина об'єктів за густиною їх розподілу у просторі може бути ідентифікована як шум. До того ж, відповідно до анотації даних, дані «Compound» містять шість кластерів, а дані «Jain» – два. Однак з детального аналізу видно, що перший кластер у даних «Compound» може бути ідентифікований за густиною як шум, тобто п'ять кластерів у цьому випадку є коректною кластеризацією. У випадку кластеризації даних «Jain» другий кластер за густиною розподілу об'єктів може бути розділений на два, тобто наявність трьох кластерів у цьому випадку також може бути сприйнятливою.

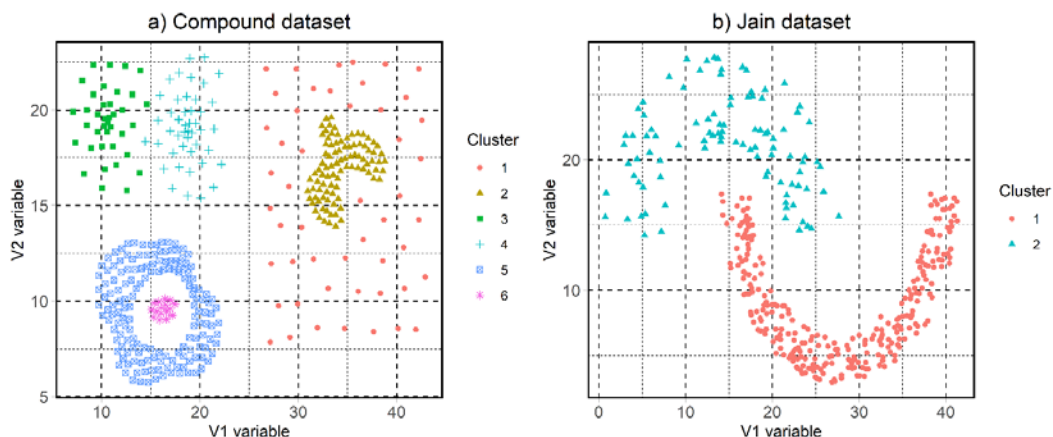


Рис. 3.6. Синтетичні дані «Compound» та «Jain», що містять кластери різних форм

2. Дані «Seeds» [248] містять інформацію щодо насіння трьох різних сортів пшениці: Ками, Рози та Канадської, по 70 елементів у кожній групі, випадково вибрані для експерименту. Інформація подана сімома атрибутами, які характеризують відповідний сорт насіння. На рис. 3.7 зображено графік у паралельних координатах об'єктів, що досліджується у цьому випадку. Як бачимо, дані можуть бути розділені на три кластери коректно.

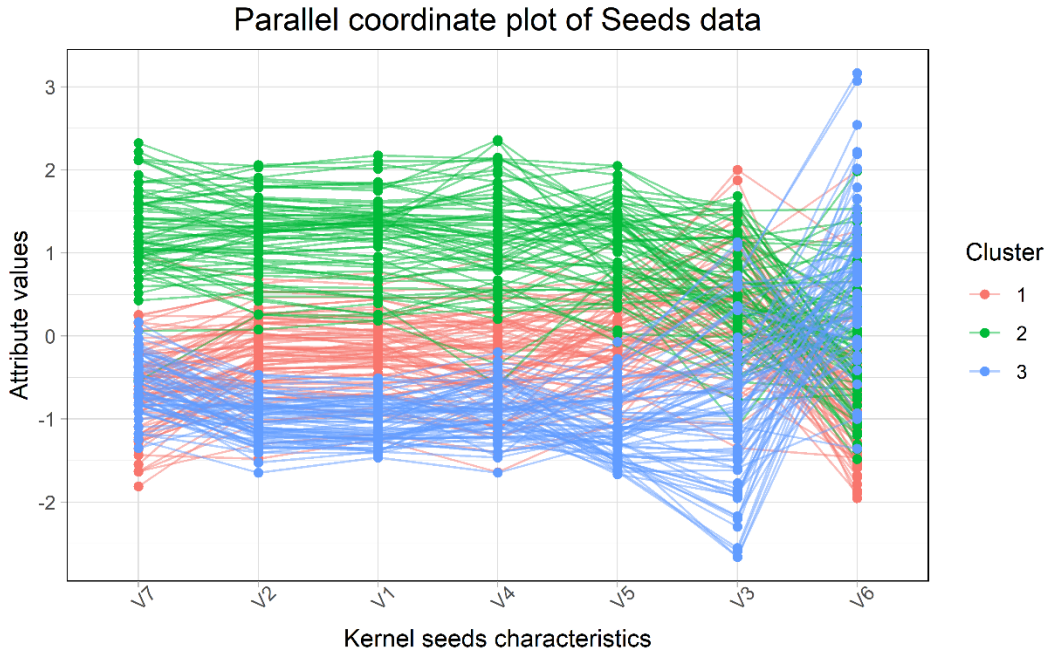


Рис. 3.7. Графік у паралельних координатах даних «Seeds»

3. Дані експресії генів *GSE 19188* пацієнтів, обстежених на ранній стадії раку легенів [249] отримані в результаті проведення експериментів з мікрочіпами ДНК, при цьому було застосовано 156 мікрочіпів. Аналіз анотації даних показав, що досліджувані зразки можна розділити на дві групи: 65 пацієнтів були здоровими, а у 91 з них була ідентифікована пухлина раку легенів. RMA (Robust Multi-Array Average) метод попередньої обробки даних (корекція фону, нормалізація, корекція РМ та сумарізація) був використаний для формування масиву даних експресії генів. Експериментальні дані у початковому стані містили 54675 генів (максимальна кількість генів на кожному з мікрочіпів). Після виділення генів з нульовим та низьким значенням експресії шляхом покрокового застосування алгоритму кластеризації SOTA з кореляційною метрикою було виділено 1126 особливо інформативних профілів експресії генів за критерієм точності класифікації об'єктів, що містять ці профілі експресії генів як атрибути [207]. Таким чином, матриця експресії генів, що використовувалася у процесі моделювання, мала розмірність 156×1126.

У таблиці 3.2 наведено внутрішні критерії якості кластеризації, що застосовувалися в процесі моделювання.

Таблиця 3.2

*Внутрішні критерії оцінки якості кластерної структури, що використовувалися в процесі моделювання*

| № | Критерій              | Правило | №  | Критерій                | Правило |
|---|-----------------------|---------|----|-------------------------|---------|
| 1 | WB index [139]        | min     | 9  | Calinski Harabasz [139] | max     |
| 2 | Banfeld-Raftery [137] | min     | 10 | PBM [149]               | max     |
| 3 | C index [138]         | min     | 11 | GDI [152]               | max     |
| 4 | Ray Turi [150]        | min     | 12 | Ratkowsky Lance [151]   | max     |
| 5 | Davies Bouldin [140]  | min     | 13 | Dunn [142]              | max     |
| 6 | McClain-Rao [148]     | min     | 14 | Gamma [144]             | max     |
| 7 | Scott-Symons [141]    | min     | 15 | Silhouette [155]        | max     |
| 8 | Xie-Beni [178]        | min     | 16 | Wemmert-Gancarski       | max     |

Процес моделювання щодо визначення оптимальної комбінації внутрішніх критеріїв, наведених у таблиці 3.1, та відповідних їм зовнішніх критеріїв передбачав штучне розділення множини даних на дві еквівалентні підмножини із застосуванням алгоритму, наведеному у підрозділі 3.1 на першому кроці. На другому кроці здійснювалося покрокове збільшення кількості штучних кластерів від 2 до 8 – для даних «Compound», від 2 до 5 – для даних «Jain» і «Seeds» та від 2 до 6 – для даних експресії генів пацієнтів, що досліджувалися на рак легенів. Розрахунок внутрішніх критеріїв здійснювався шляхом застосування функцій пакету «cluster Crit» [251] програмного середовища R.

На рис. 3.8 та 3.9 зображені діаграми залежності значень внутрішніх критеріїв від кількості кластерів у випадку застосування двох еквівалентних підмножин *A* і *B* даних «Compound». За анотацією критеріїв, перша група діаграм (рис. 3.8) відповідає правилу *min*, тобто мінімальне значення критерію відповідає оптимальній кластеризації. Діаграми, зображені на рис. 3.9, відповідають правилу *max*.

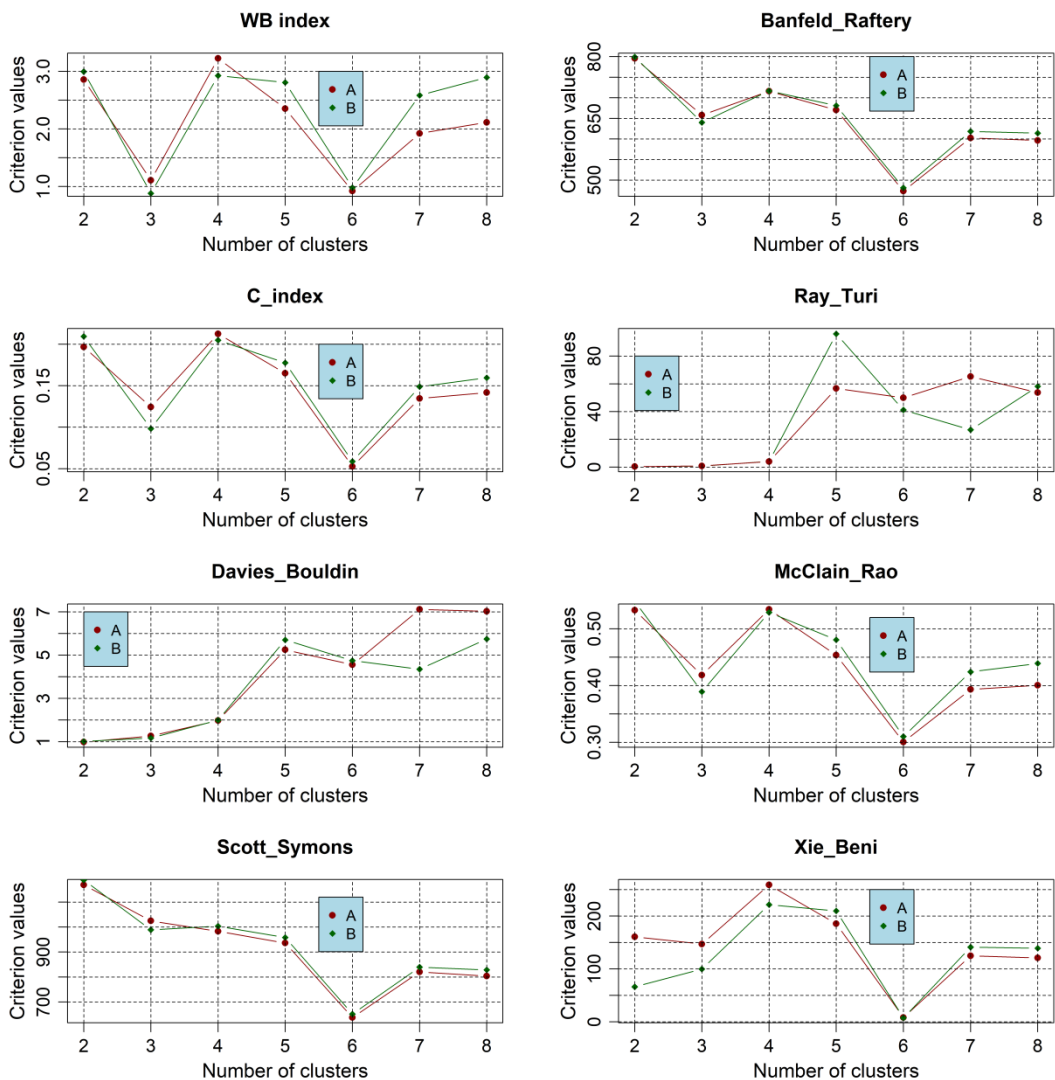


Рис. 3.8. Діаграми розподілу внутрішніх критеріїв якості кластеризації, що відповідають правилу *min*, розрахованих для двох еквівалентних підмножин *A* і *B* даних «Compound» для кластеризацій з кількістю кластерів від 2 до 8

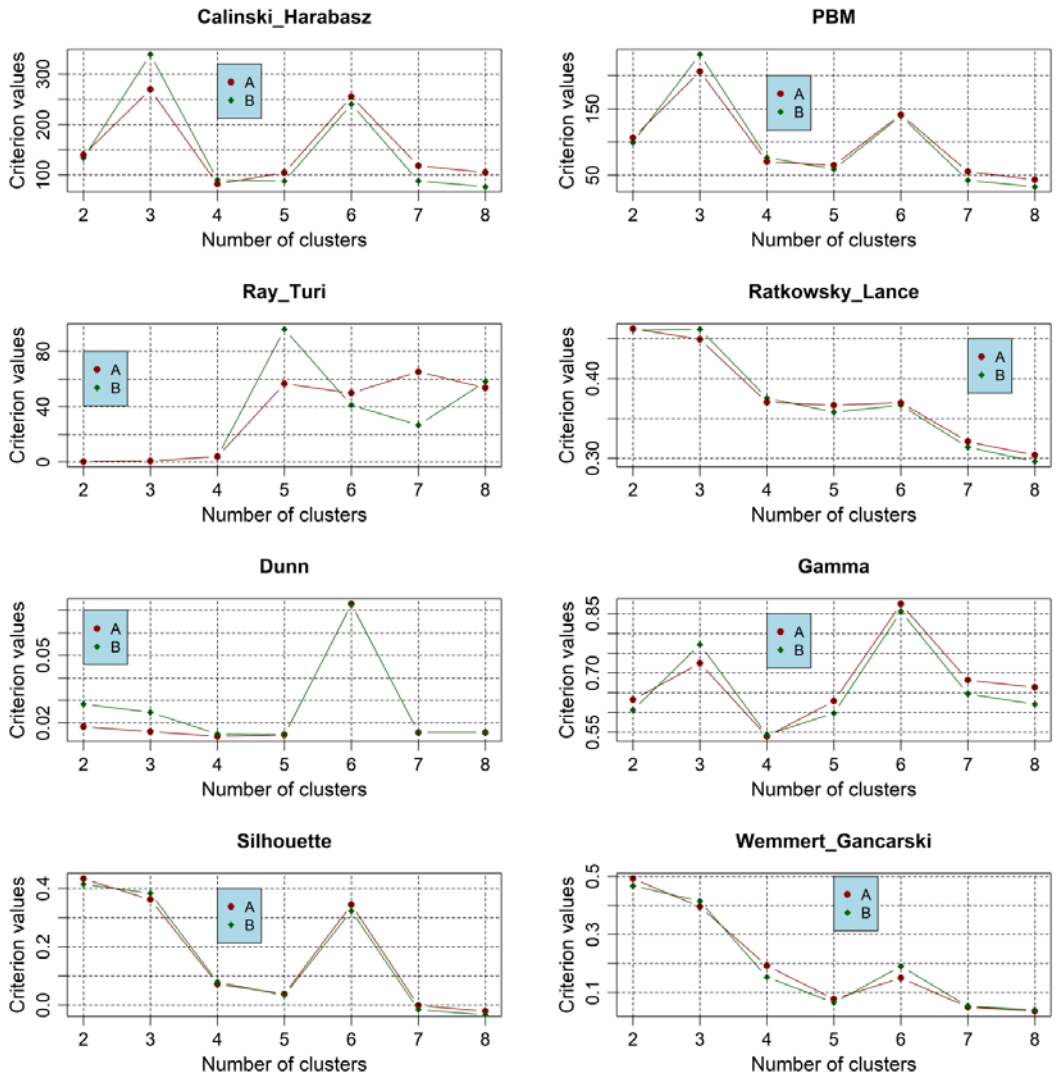


Рис. 3.9. Діаграми розподілу внутрішніх критеріїв якості кластеризації, що відповідають правилу *max*, розрахованих для двох еквівалентних підмножин *A* і *B* даних «Comround» для кластеризацій з кількістю кластерів від 2 до 8

На рис. 3.10 зображено діаграми залежності значень зовнішніх критеріїв, розрахованих як нормалізована різниця відповідних внутрішніх критеріїв за формулою (3.3).

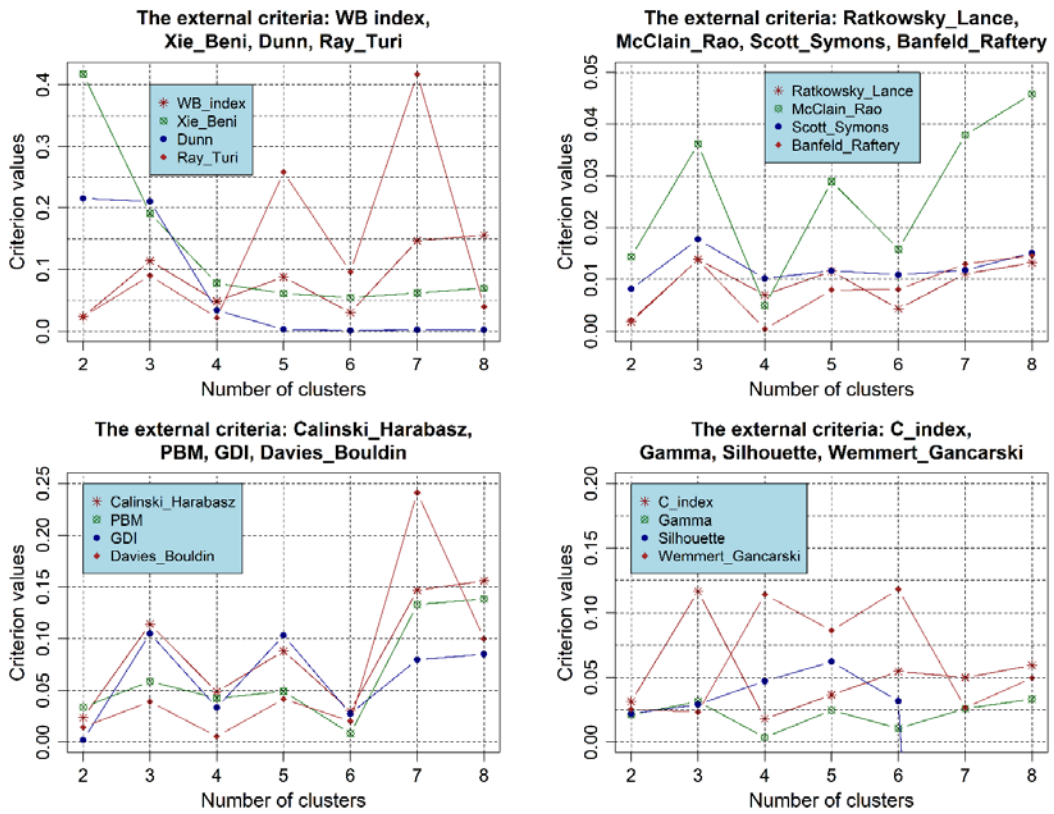


Рис. 3.10. Діаграми розподілу зовнішніх критеріїв якості кластеризації для даних «Compound»

Проаналізувавши отримані результати, можемо зробити висновок, що для даних «Compound» наступні внутрішні критерії дозволяють виділити оптимальну кластеризацію (шість кластерів) при застосуванні двох еквівалентних підмножин даних: за правилом *min*: Banfield\_Raftery, C-index, McClain\_Rao, Scott\_Symons та Xie\_Beni; за правилом *max*: Dunn і Gamma. Аналіз діаграм, зображених на рис. 3.10, свідчить, що у більшості випадків внутрішні та зовнішні критерії суперечать один одному. Цей факт підтверджує висновок щодо наявності похибки відтворюваності і як результат необхідності у компромісному рішенні шляхом застосування критерію балансу. Детальний аналіз діаграм залежності зовнішніх критеріїв від кількості кластерів показує, що із вищезазначеного переліку внутрішніх критеріїв якості кластеризації мінімуму похибки відтворюваності для кластерної структури, що містить шість кластерів (мінімальне значення зовнішнього критерія), відповідають такі критерії: Dunn, Gamma та Xie\_Beni, які були застосовані для даних «Compound». Аналогічні результати для даних «Jain» зображені на рис. 3.11–3.13.

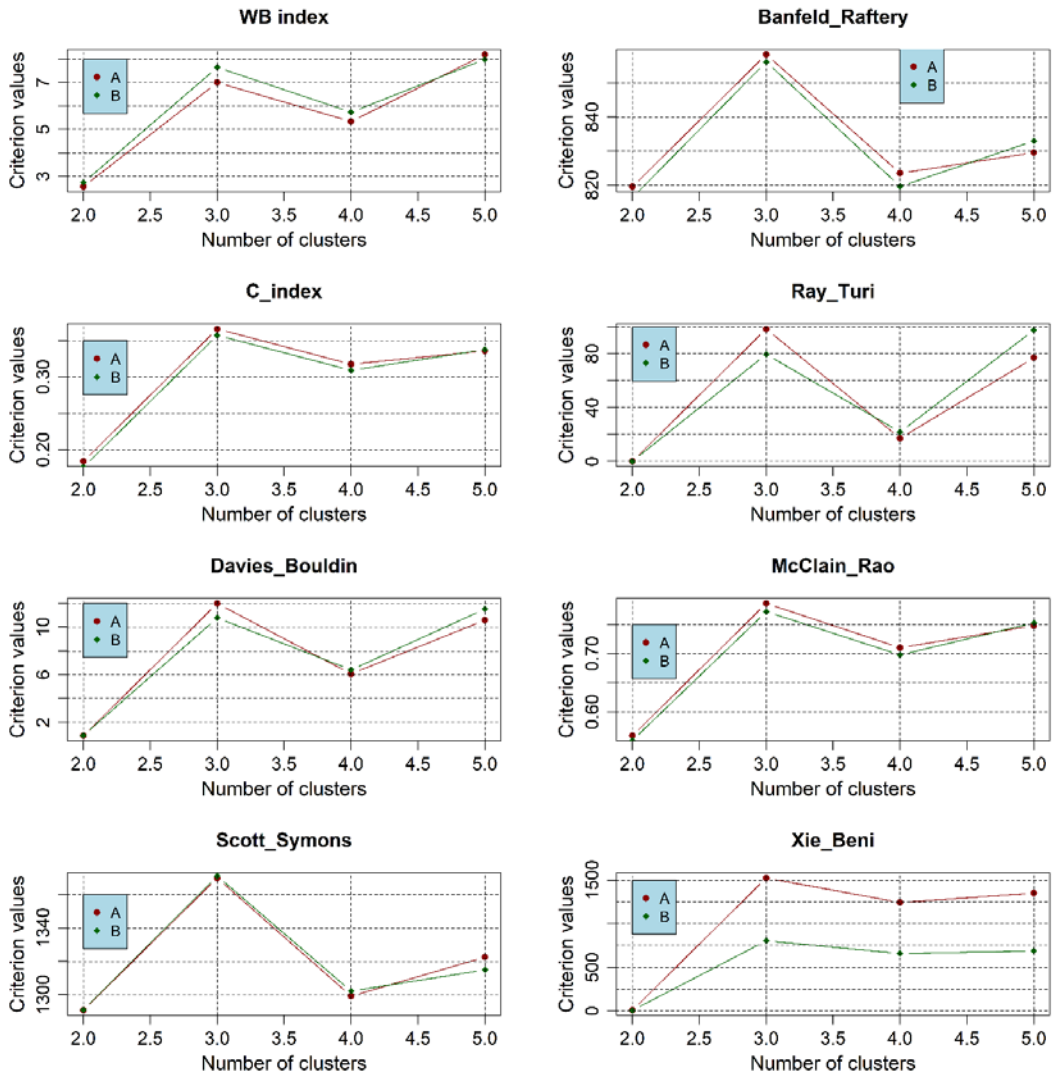


Рис. 3.11. Діаграми розподілу внутрішніх критеріїв якості кластеризації, що відповідають правилу *min*, розрахованих для двох еквівалентних підмножин *A* і *B* даних «Jain» для кластеризацій з кількістю кластерів від 2 до 5

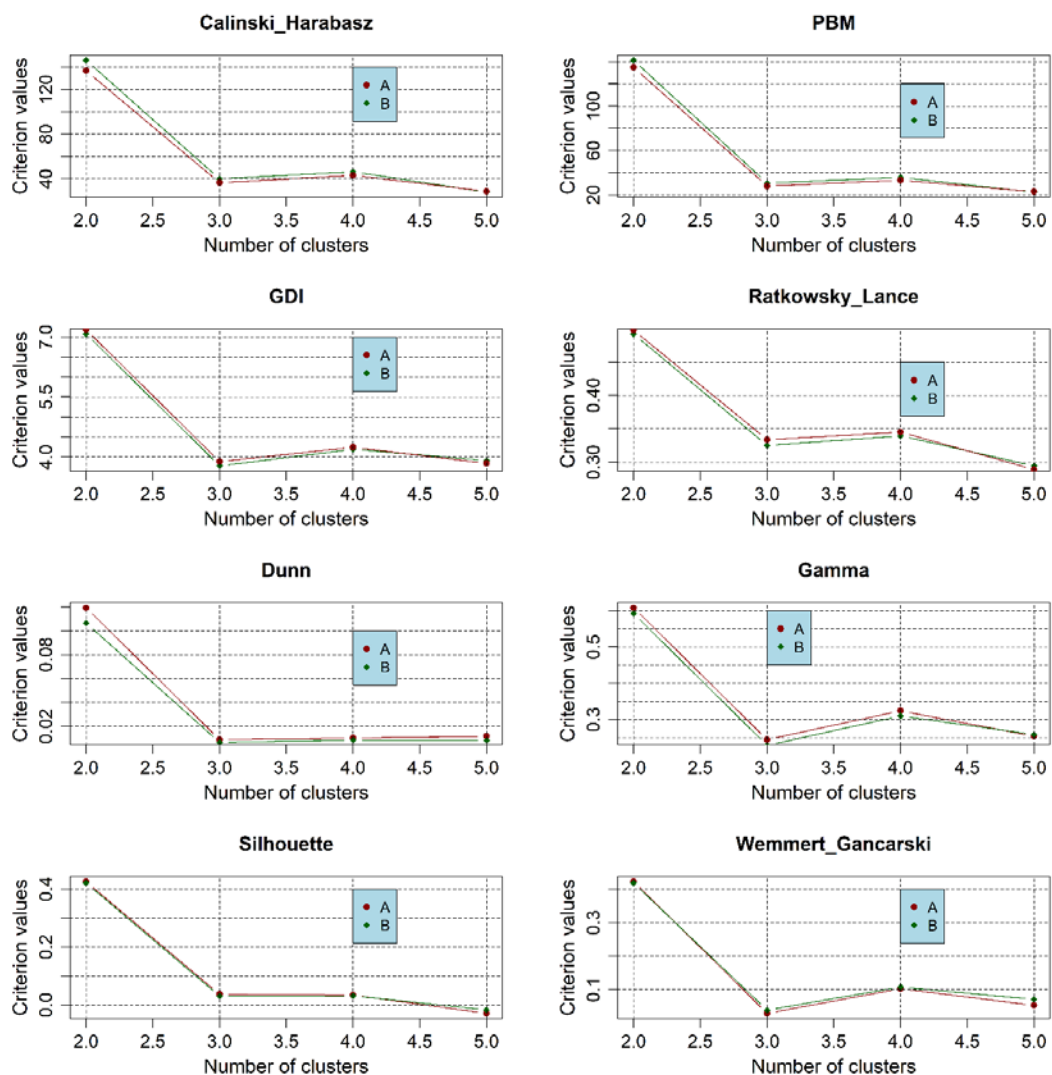


Рис. 3.12. Діаграми розподілу внутрішніх критеріїв якості кластеризації, що відповідають правилу *max*, розрахованих для двох еквівалентних підмножин *A* і *B* даних «Jain» для кластеризацій з кількістю кластерів від 2 до 5

Як можна бачити, характер розподілу об'єктів даних «Jain», зображений на рис. 3.6b не такий складний, як у випадку даних «Compound» (рис. 3.6a). Дані «Jain» містять тільки два кластери, які не перетинаються між собою. Кластери, що вкладені один в другий відсутні. Тому на рис. 3.11 і 3.12 бачимо, що майже всі внутрішні критерії дозволяють ідентифікувати оптимальну двокластерну структуру як за правилом *min*, так і за правилом *max*.

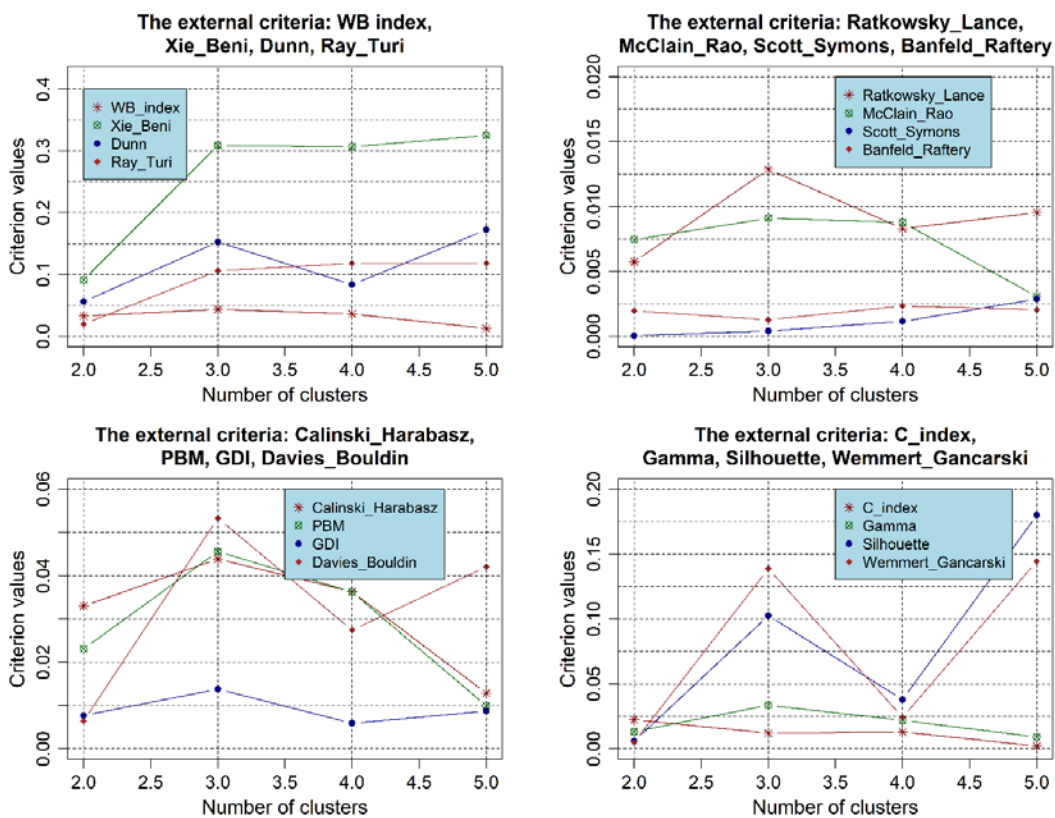


Рис. 3.13. Діаграми розподілу зовнішніх критеріїв якості кластеризації для даних «Jain»

Аналіз діаграм розподілу зовнішніх критеріїв якості кластеризації при різних кількостях кластерів, що зображені на рис. 3.13, показує, що за похибкою відтворюваності (мінімальне значення зовнішніх критеріїв, розрахованих для відповідних внутрішніх критеріїв) результати не такі однозначні. Найкращі результати у цьому випадку (мінімальне значення для двокластерної структури) отримуються у випадку застосування критеріїв: Xie\_Beni та Ray\_Turi. Ця група критеріїв була застосована для даних «Jain».

На рис. 3.14–3.16 зображено результати критеріального аналізу для даних «Seeds».

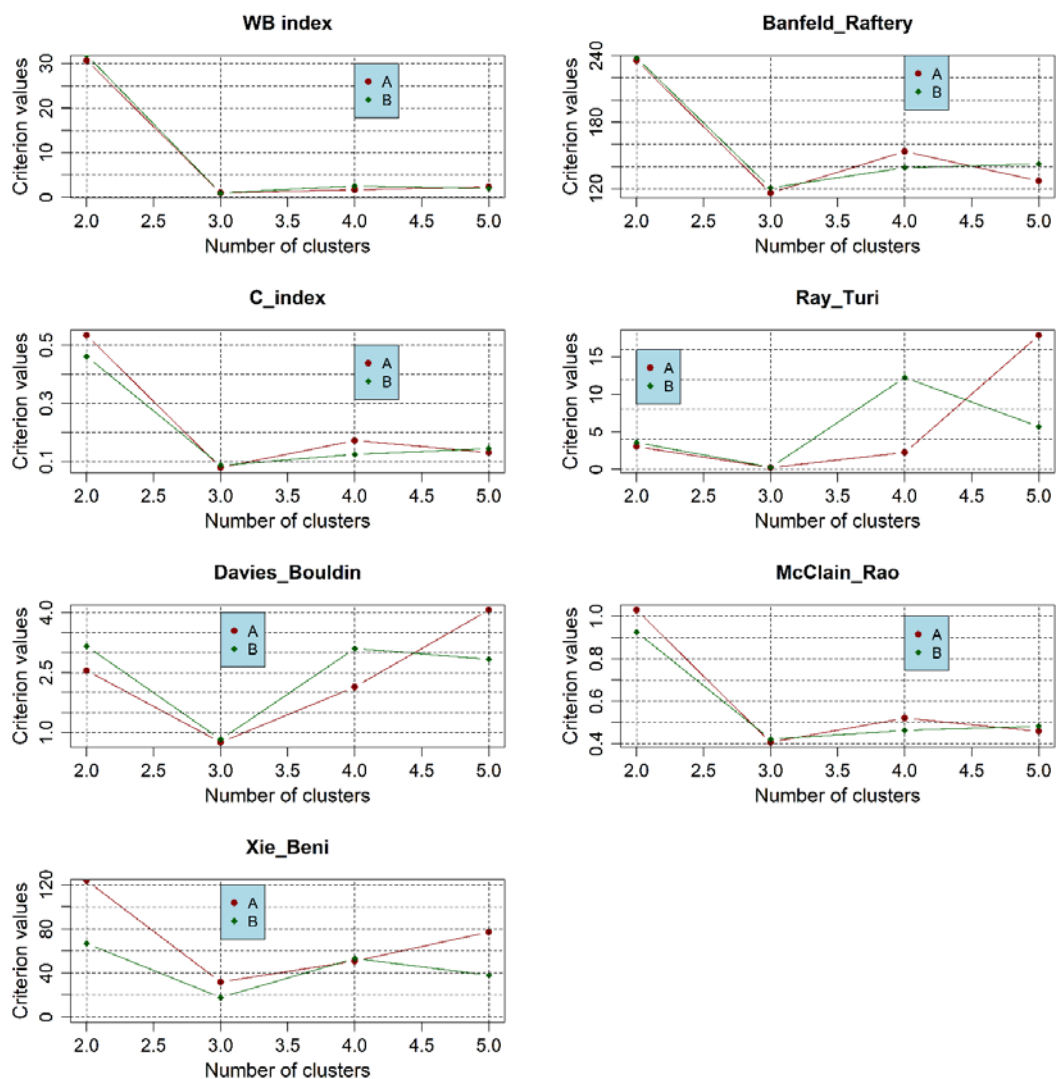


Рис. 3.14. Діаграми розподілу внутрішніх критеріїв якості кластеризації, що відповідають правилу *min*, розрахованих для двох еквівалентних підмножин *A* і *B* даних «Seeds» для кластеризацій з кількістю кластерів від 2 до 5

Для цих даних оптимальна кластерна структура відповідає трьом кластерам. У процесі моделювання кількість кластерів змінювалася від 2 до 5. Результати моделювання показали, що критерій Scott\_Symens для цього типу даних не є ефективним, тому він був заздалегідь виключений із списку критеріїв, що досліджувалися.

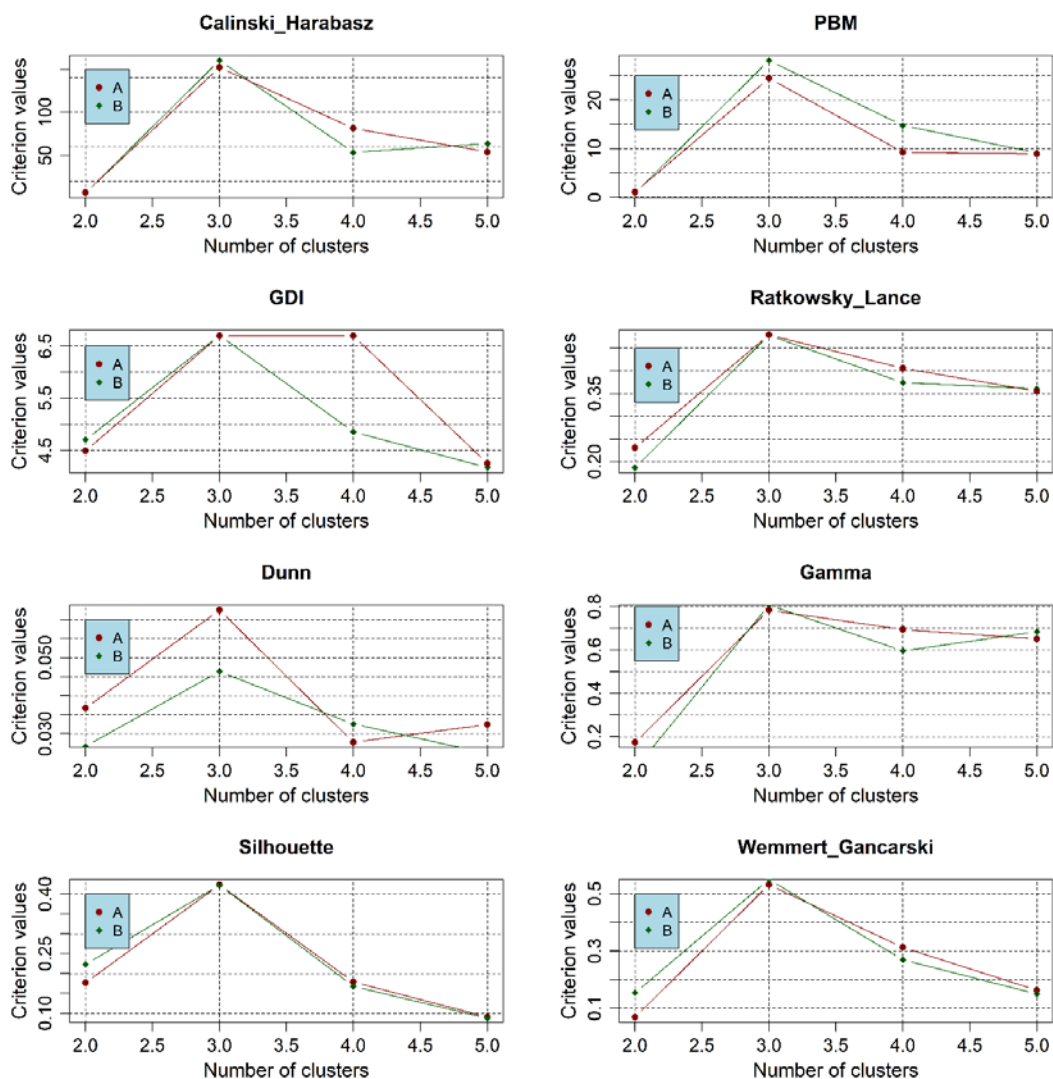


Рис. 3.15. Діаграми розподілу внутрішніх критеріїв якості кластеризації, що відповідають правилу *max*, розрахованих для двох еквівалентних підмножин *A* і *B* даних «Seeds» для кластеризацій з кількістю кластерів від 2 до 5

Результати моделювання показали, що однозначно ідентифікувати триклас-терну структуру дозволяють такі внутрішні критерії: за правилом *min* (рис. 3.14): Davies\_Bouldin; за правилом *max*: Calinski\_Harabasz, PBM, Ratkovsky\_Lance, Dunn, Silhouette та Wemmert\_Garanski.

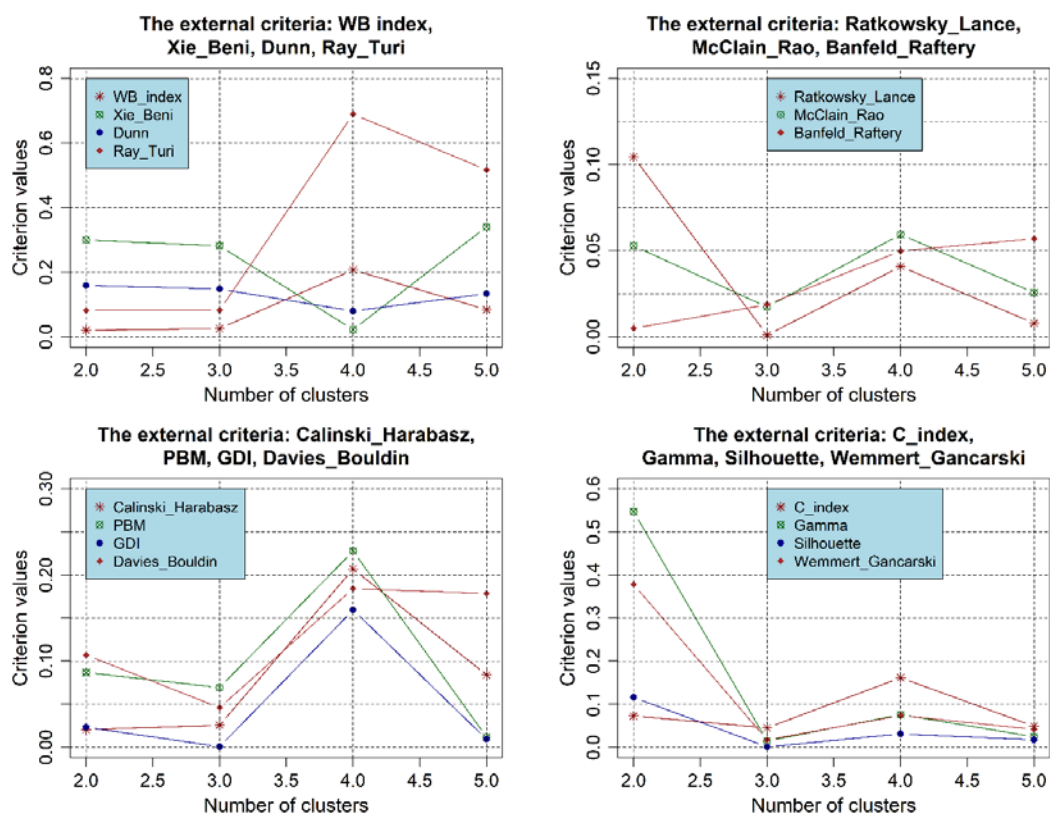


Рис. 3.16. Діаграми розподілу зовнішніх критеріїв якості кластеризації для даних «Seeds»

Аналіз діаграм розподілу зовнішніх критеріїв якості від кількості кластерів (рис. 3.16) підтвердив висновок щодо ефективності критерію Davies\_Bouldin для виділення оптимальної кластеризації даних «Seeds». Однак з другої групи виділених внутрішніх критеріїв оптимальній кластерній структурі відповідають однозначні мінімуми критерію Ratkovsky\_Lance. Таким чином, ці два критерії були застосовані в процесі подальшого моделювання для даних «Seeds».

У разі застосування даних експресії генів кластеризуються профілі експресії генів, тобто вектор значень експресії відповідного гена, що визначені для різних зразків. Внутрішні та зовнішні критерії якості кластеризації даних у цьому випадку формуються шляхом аналізу характеру розподілу профілів експресії генів у виділених кластерах. Однак, з іншого боку, оцінка адекватності відповідної кластеризації може бути здійснена тільки шляхом аналізу результатів класифікації відповідних зразків, які містять виділені профілі експресії генів у відповідних кластерах як атрибути. Таким чином, реалізація цієї процедури може бути виконана на основі комплексного застосування методів кластеризації та класифікації із розрахунком критеріїв якості кластеризації та класифікації на кожному ієрархічному рівні реалізації цієї процедури.

На рис. 3.17 зображено блок-схему гібридної моделі критеріального аналізу даних експресії генів на основі комплексного застосування кластерного аналізу та класифікації даних. Реалізація цієї процедури передбачає такі етапи:

**Етап I.** Формування двох еквівалентних підмножин профілів експресії генів  $A$  і  $B$  із розрахунком критерію точності класифікації досліджуваних зразків.

1. Розділення профілів експресії генів на дві еквівалентні підмножини шляхом застосування алгоритму, наведеному у підрозділі 3.1, із застосуванням гібридної метрики оцінки ступеня близькості профілів експресії генів, що описано у розділі 2.

2. Класифікація досліджуваних зразків, атрибутами яких є профілі експресії генів у еквівалентних підмножинах  $A$  і  $B$ . Розрахунок критеріїв точності класифікації даних підмножин  $A$  і  $B$  ( $ACC_0^A$ ,  $ACC_0^B$ ) здійснювався за формулою (2.18). Як класифікатор у рамках поточних досліджень застосовувався алгоритм «випадкового лісу» (Random Forest (RF) classifier), ефективність якого для кластеризації профілів експресії генів доведена у роботі [207].

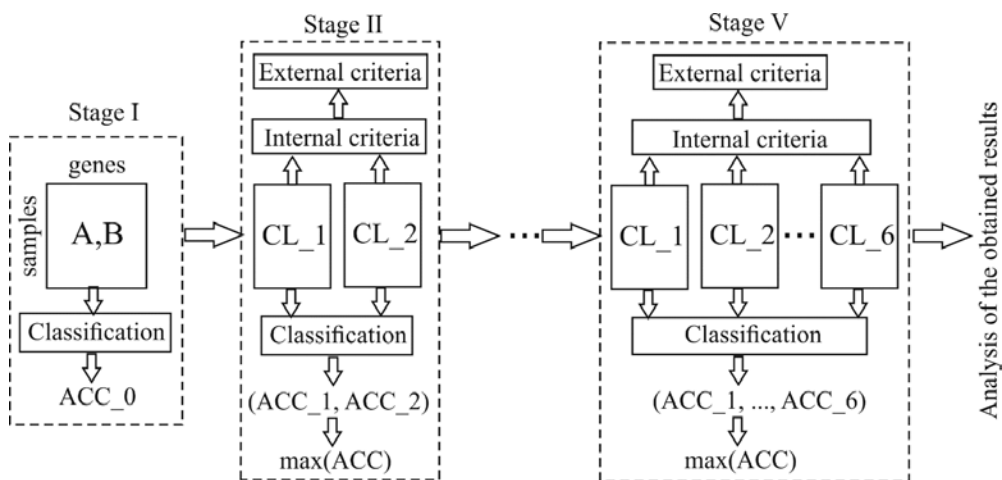


Рис. 3.17. Гібридна модель критеріального аналізу даних експресії генів

**Етапи II–V.** Ієрархічна кластеризація профілів експресії генів з розрахунком критеріїв якості кластеризації профілів експресії генів та зразків, що досліджуються на кожному ієрархічному рівні.

3. Ієрархічна кластеризація профілів експресії генів з покроковим збільшенням кількості кластерів від 2 – на першому кроці до 6 – на п'ятому кроці шляхом застосування самоорганізуючого алгоритму кластеризації SOTA (Self-Organizing Tree Algorithm) [120]. Цей алгоритм є різновидом самоорганізуючих нейронних мереж і базується на комплексному застосуванні карт Кохонена та алгоритму формування клітин у просторі Fritzke і орієнтований на ієрархічну кластеризацію високимірних профілів експресії генів [121, 252]. Результатом однієї ітерації роботи цього алгоритму є бінарне топологічне дерево, до того ж, відповідно до алгоритму вирощування клітинної структури Fritzke, вихідні вузли мережі самоорганізуються так, що кількість вузлів зростає в ділянці з більш високою щільністю

розподілу об'єктів і, відповідно, зменшується в ділянці з меншою щільністю. Таким чином регулюється неоднорідність розподілу об'єктів у просторі.

На першому кроці застосування цього алгоритму кожна множина профілів експресії генів розділяється на дві приблизно однакових підмножини, при цьому формується двокластерна структура. На подальших кроках формується ієрархічна дендрограма із збільшенням кількості кластерів від 2 до 6 з подальшим розрахунком критеріїв якості кластеризації профілів експресії генів на кожному кроці реалізації цієї процедури та точності класифікації зразків, що містять профілі експресії генів у кожному кластері. Результатом реалізації цього кроку є формування вектора критеріїв точності класифікації зразків, довжина якого дорівнює кількості кластерів.

4. Остаточне формування значення критерію точності класифікації зразків на кожному ієрархічному рівні шляхом виділення максимального значення з кожного із сформованих заздалегідь векторів точності класифікацій. Виділення максимального значення точності класифікації обумовлене тим, що основною метою передобробки профілів експресії генів є виділення групи профілів, які дозволяють з максимальною точністю ідентифікувати об'єкти, що досліджуються. При цьому необхідно сформувати групу внутрішніх критеріїв, які дозволяють утворити оптимальну кластерну структуру за максимальним значенням точності класифікації зразків з урахуванням мінімуму похибки відтворюваності.

**Етап VI.** Аналіз отриманих результатів.

5. Побудова діаграм розподілу внутрішніх та зовнішніх критеріїв якості кластеризації на різних ієрархічних рівнях формування кластерної структури, та точності класифікації зразків залежно від кількості кластерів.

6. Аналіз отриманих діаграм, формування критеріїв якості кластеризації, екстремуми яких відповідають максимуму точності класифікації зразків на відповідному ієрархічному рівні.

Результати моделювання щодо реалізації вищенаведеної процедури з розрахунком внутрішніх та зовнішніх критеріїв якості кластеризації наведені на рис. 3.18–3.20. На рис. 3.21 зображені діаграми розподілу критерію точності класифікації зразків, які містять як атрибути профілі експресії генів, що виділені у відповідні кластери.

Однокластерній структурі у цьому випадку відповідають профілі експресії генів сформованих еквівалентних підмножин даних. Отже, точність класифікації вихідних даних досягає 96% і майже однакова у випадку застосування підмножин *A* та *B*, що свідчить про те, що початкові дані профілів експресії генів розділені на еквівалентні підмножини *A* та *B* адекватно, і виділені профілі експресії генів дозволяють коректно ідентифікувати зразки, які досліджуються.

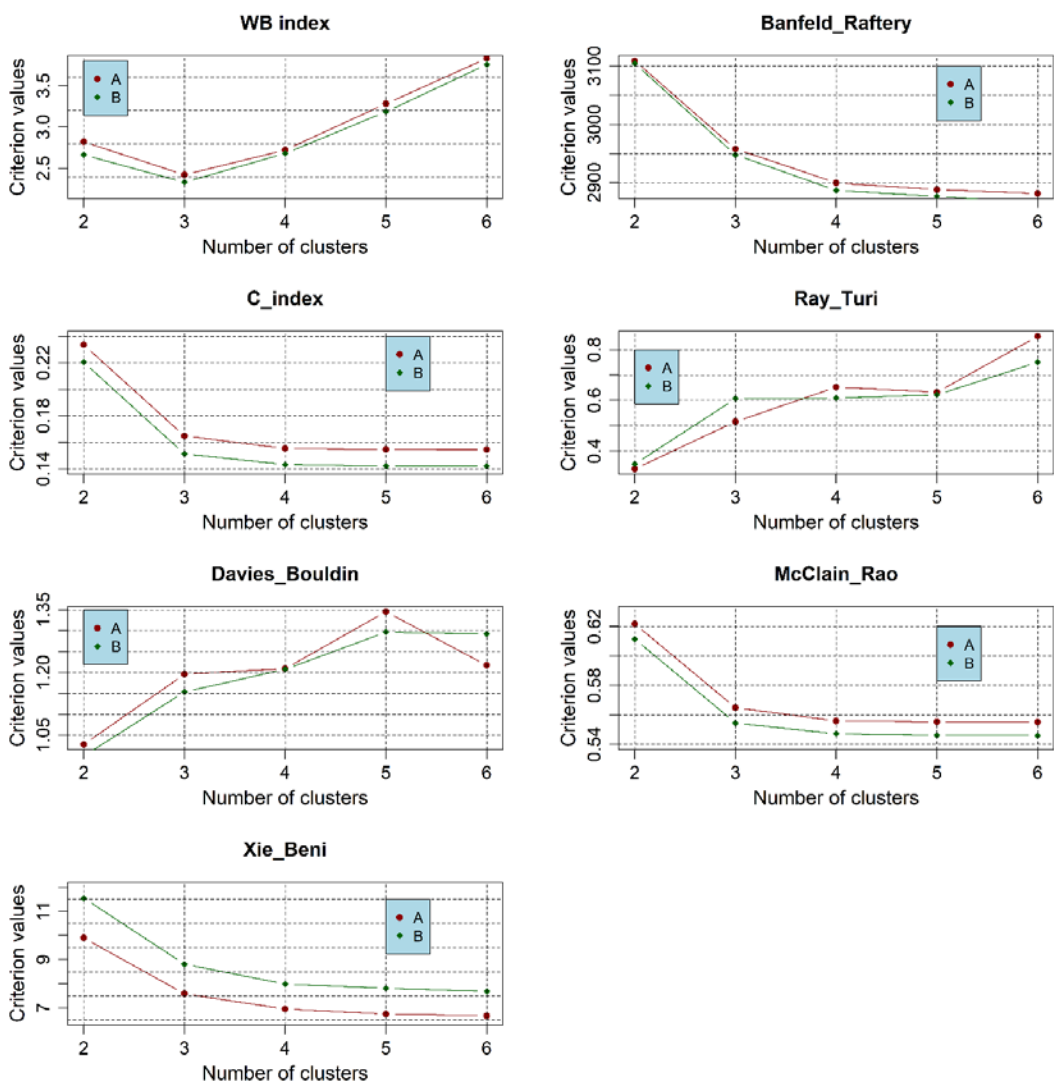


Рис. 3.18. Діаграми розподілу внутрішніх критеріїв якості кластеризації, що відповідають правилу *min*, розрахованих для двох еквівалентних підмножин *A* і *B* профілів експресії генів для кластеризацій з кількістю кластерів від 2 до 6

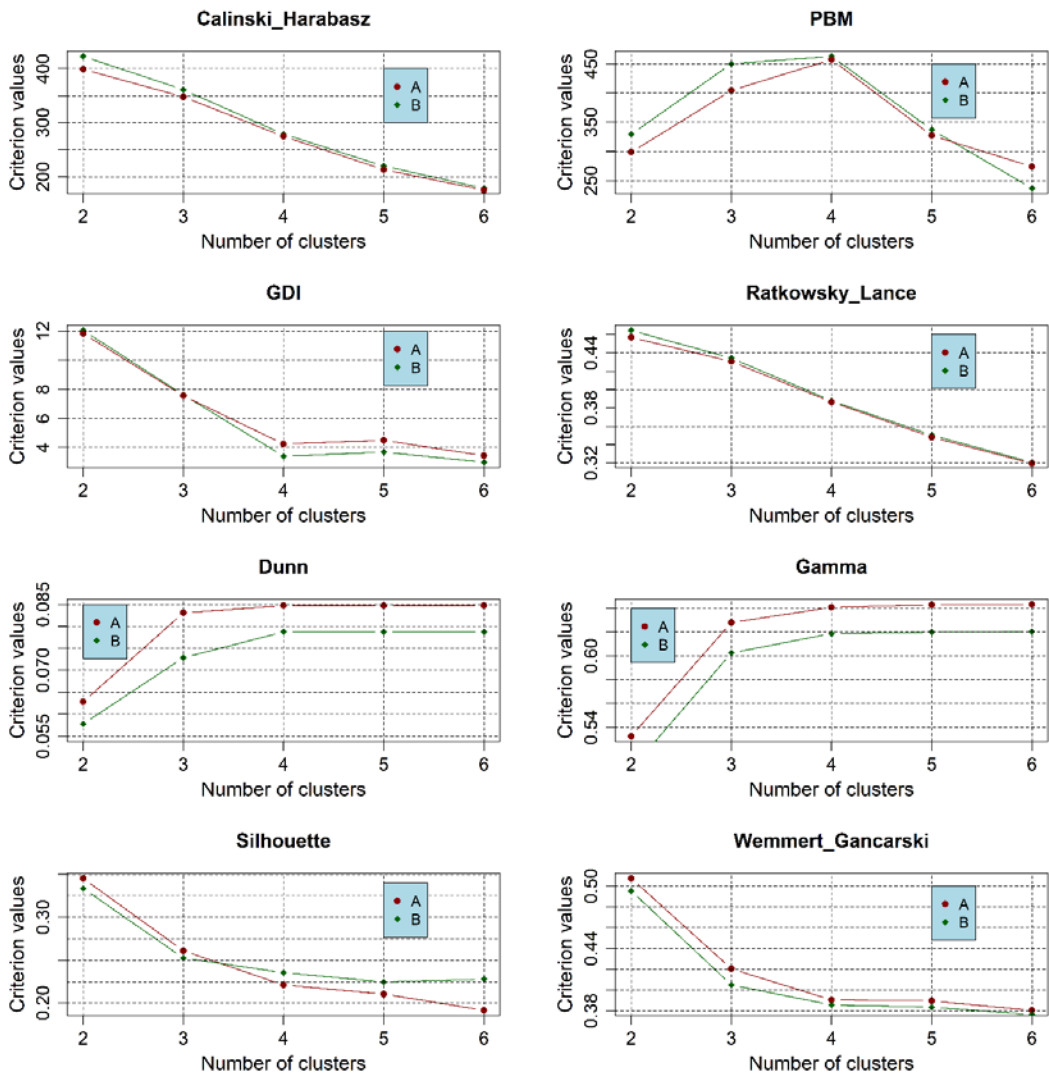


Рис. 3.19. Діаграми розподілу внутрішніх критеріїв якості кластеризації, що відповідають правилу *max*, розрахованих для двох еквівалентних підмножин *A* і *B* профілів експресії генів для кластеризацій з кількістю кластерів від 2 до 6

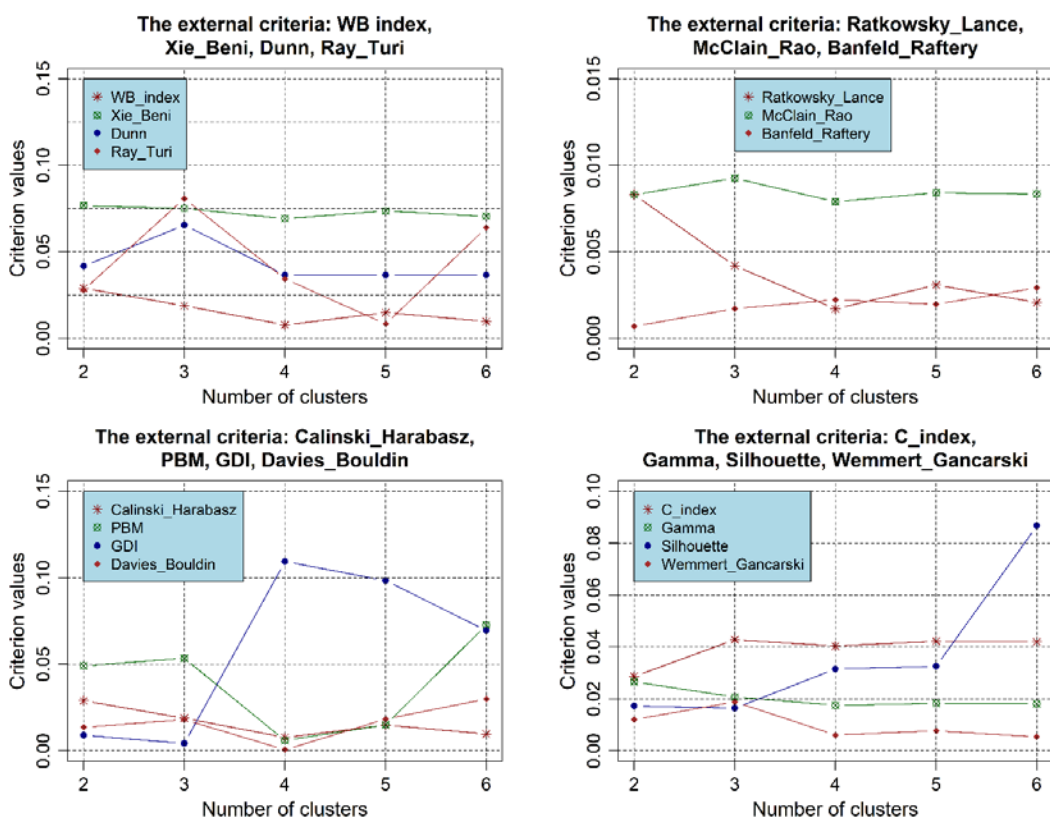


Рис. 3.20. Діаграми розподілу зовнішніх критеріїв якості кластеризації для профілів експресії генів пацієнтів, що досліджувалися на рак легенів

Аналіз діаграм розподілу точності класифікації даних експресії генів дозволяє зробити висновок, що оптимальною кластерною структурою за точністю класифікації та похибкою відтворюваності є три- або чотирікластерні структури. Якщо є три кластери то у двох випадках (для підмножин  $A$  і  $B$ ) можна виділити один кластер профілів експресії генів, який дозволяє ідентифікувати досліджувані зразки з точністю понад 90%. Профілі експресії генів у інших двох кластерах можна видалити за причиною низької точності класифікації зразків (менше ніж 70 %).

У разі наявності чотирьох кластерів у двох випадках можна виділити два кластера профілів експресії генів. Перший кластер також дозволяє ідентифікувати досліджувані зразки з точністю понад за 90%, другий кластер дозволяє отримати точність класифікації вищу за 80%, що також сприйнятливо. Якщо кластерів є більше, то точність класифікації зразків підмножини  $A$  погіршується і суттєво не збігається з аналогічними результатами, отриманими із застосуванням даних експресії генів підмножини  $B$ .

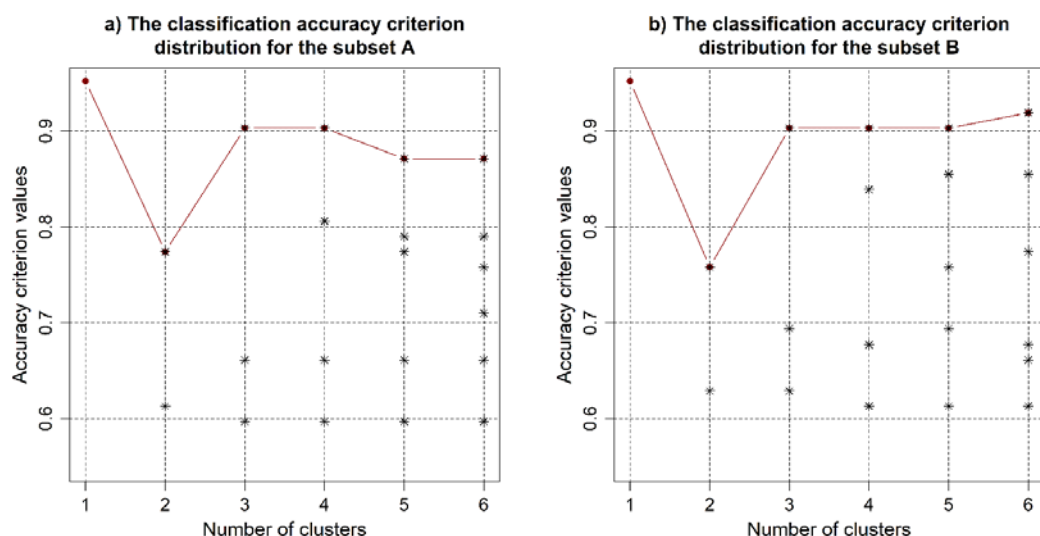


Рис. 3.21. Діаграми розподілу точності класифікації зразків, що містять як атрибути виділені у кластери профілі експресії генів

Аналіз діаграм розподілу внутрішніх критеріїв якості кластеризації (рис. 3.18 і 3.19) показує, що тільки два критерії дозволяють адекватно ідентифікувати три- або чотирьокластерні структури: WB індекс за правилом *min* і PBM критерій за правилом *max*. Екстремуми цих критеріїв відповідають бажаним структурам. Аналіз діаграм розподілу зовнішніх критеріїв (рис. 3.20) підтверджує ці висновки. Таким чином, можна зробити висновок, що для виділення оптимальних кластерних структур профілів експресії генів варто застосовувати внутрішні критерії якості кластеризації WB і PBM.

У таблиці 3.3 наведено результати моделювання щодо вибору внутрішніх критеріїв якості кластеризації для експериментальних даних, що досліджуються. Дані комбінації критеріїв застосовуються для розрахунку критерію балансу відповідно до алгоритму, наведеному у підрозділі 3.1 (формули 3.4–3.8).

Таблиця 3.3

*Оптимальні внутрішні критерії якості кластеризації для експериментальних даних, що застосовуються у процесі моделювання*

| No | Compound                | Jain                    | Seeds                          | Профілі експресії генів |
|----|-------------------------|-------------------------|--------------------------------|-------------------------|
| 1  | Dunn ( <i>max</i> )     | Xia_Beni ( <i>min</i> ) | Davies_Bouldin ( <i>min</i> )  | WBindex ( <i>min</i> )  |
| 2  | Gamma ( <i>max</i> )    | Ray_Turi ( <i>max</i> ) | Ratkowsky_Lance ( <i>max</i> ) | PBM ( <i>max</i> )      |
| 3  | Xie_Beni ( <i>min</i> ) | —                       | —                              | —                       |

### 3.2.2. Формування функцій розрахунку критерію балансу на основі виділених внутрішніх критеріїв якості кластеризації, враховуючи тип експериментальних даних

Результати моделювання, наведені у підрозділі 3.2.1, створюють умови для формування комплексного критерію балансу, враховуючи тип експериментальних даних, який включає як компоненти внутрішні критерії якості кластеризації, що наведені у таблиці 3.3, та відповідні їм зовнішні критерії. Алгоритм для розрахунку критерію балансу на основі функції бажаності Харрінгтона подано у підрозділі 3.1.

#### 3.2.2.1. Алгоритм формування функції розрахунку критерію балансу для даних «Compound»

Розрахунок критерію балансу для даних «Compound» здійснюється відповідно до такого алгоритму:

1. Розрахунок коефіцієнтів лінійного рівняння перетворення значень відповідних внутрішніх та зовнішніх критеріїв у значення показника  $Y$  для даних еквівалентних підмножин  $A$  і  $B$ , враховуючи граничні значення відповідних величин на відносний характер їх зміни:

Внутрішні критерії:

$$\begin{cases} Y_{max} = a_{XB}^{A,B} - b_{XB}^{A,B} \cdot QC_{XB}^{\min(A,B)} \\ Y_{min} = a_{XB}^{A,B} - b_{XB}^{A,B} \cdot QC_{XB}^{\max(A,B)} \\ Y_{max} = a_{Dn}^{A,B} + b_{Dn}^{A,B} \cdot QC_{Dn}^{\max(A,B)} \\ Y_{min} = a_{Dn}^{A,B} + b_{Dn}^{A,B} \cdot QC_{Dn}^{\min(A,B)} \\ Y_{max} = a_{Gm}^{A,B} + b_{Gm}^{A,B} \cdot QC_{Gm}^{\max(A,B)} \\ Y_{min} = a_{Gm}^{A,B} + b_{Gm}^{A,B} \cdot QC_{Gm}^{\min(A,B)} \end{cases} \quad (3.10a)$$

Зовнішні критерії:

$$\begin{cases} Y_{max} = a - b \cdot QC_{ext}^{\min} \\ Y_{min} = a - b \cdot QC_{ext}^{\max} \end{cases} \quad (3.10b)$$

де  $QC_{XB}^{\min(A,B)}$ ,  $QC_{XB}^{\max(A,B)}$ ,  $QC_{Dn}^{\min(A,B)}$ ,  $QC_{Dn}^{\max(A,B)}$ ,  $QC_{Gm}^{\min(A,B)}$ ,  $QC_{Gm}^{\max(A,B)}$  – мінімальне та максимальне значення критеріїв Xie\_Beni, і Gamma, що розраховані для еквівалентних підмножин  $A$  і  $B$ , відповідно;  $QC_{ext}^{\min}$ ,  $QC_{ext}^{\max}$  – мінімальне та максимальне значення відповідного зовнішнього критерію;  $Y_{max} = 5$ ,  $Y_{min} = -2$  – максимальне та мінімальне значення безрозмірного показника  $Y$  за методом бажаності Харрінгтона,  $a$  і  $b$  – коефіцієнти відповідних лінійних рівнянь.

2. Трансформування значень відповідних критеріїв якості кластеризації у значення показника  $Y$  відповідно до системи рівнянь:

$$\begin{cases} Y_{XB}^{A,B} = a_{XB}^{A,B} - b_{XB}^{A,B} \cdot QC_{XB}^{A,B} \\ Y_{Dn}^{A,B} = a_{Dn}^{A,B} + b_{Dn}^{A,B} \cdot QC_{Dn}^{A,B} \\ Y_{Gm}^{A,B} = a_{Gm}^{A,B} + b_{Gm}^{A,B} \cdot QC_{Gm}^{A,B} \\ Y_{ext}^{A,B} = a^{A,B} - b^{A,B} \cdot QC_{ext}^{A,B} \end{cases} \quad (3.11)$$

3. Розрахунок приватних бажаностей для кожного критерію:

$$d_i = \exp(-\exp(-Y_i)). \quad (3.12)$$

4. Розрахунок критерію балансу як середнє геометричне усіх приватних бажаностей (2 + 2 + 2 (внутрішні критерії для підмножин  $A$  і  $B$ ) + 3 (зовнішні критерії для кожної пари внутрішніх критеріїв)):

$$QC_{bal} = \sqrt[9]{\prod_{i=1}^9 d_i}. \quad (3.13)$$

### 3.2.2.2. Алгоритм формування функції розрахунку критерію балансу для даних «Jain»

Алгоритм розрахунку критерія балансу для даних «Jain» такий:

1. Розрахунок коефіцієнтів лінійних рівнянь перетворення значень відповідних внутрішніх та зовнішніх критеріїв у значення показника  $Y$ :

$$\begin{cases} Y_{max} = a_{XB}^{A,B} - b_{XB}^{A,B} \cdot QC_{XB}^{\min(A,B)} \\ Y_{min} = a_{XB}^{A,B} - b_{XB}^{A,B} \cdot QC_{XB}^{\max(A,B)} \\ Y_{max} = a_{RT}^{A,B} + b_{RT}^{A,B} \cdot QC_{RT}^{\max(A,B)} \\ Y_{min} = a_{RT}^{A,B} + b_{RT}^{A,B} \cdot QC_{RT}^{\min(A,B)} \end{cases} \quad (3.14)$$

Розрахунок коефіцієнтів лінійних рівнянь для зовнішніх критеріїв здійснюється відповідно до системи рівнянь (3.106).

де  $QC_{XB}^{\min(A,B)}$ ,  $QC_{XB}^{\max(A,B)}$ ,  $QC_{RT}^{\min(A,B)}$ ,  $QC_{RT}^{\max(A,B)}$  – мінімальне та максимальне значення критеріїв  $Xie\_Beni$  і  $Ray\_Turi$ , що розраховані для еквівалентних підмножин  $A$  і  $B$ .

2. Трансформування значень відповідних критеріїв якості кластеризації у значення показника  $Y$  відповідно до системи рівнянь:

$$\begin{cases} Y_{XB}^{A,B} = a_{XB}^{A,B} - b_{XB}^{A,B} \cdot QC_{XB}^{A,B} \\ Y_{RT}^{A,B} = a_{RT}^{A,B} + b_{RT}^{A,B} \cdot QC_{RT}^{A,B} \\ Y_{ext}^{A,B} = a_{ext}^{A,B} - b_{ext}^{A,B} \cdot QC_{ext}^{A,B} \end{cases} \quad (3.15)$$

3. Розрахунок приватних бажаностей для кожного критерію за формулою (3.12).

4. Розрахунок критерію балансу як середнє геометричне усіх приватних бажаностей (2 + 2 (внутрішні критерії для підмножин  $A$  і  $B$ ) + 2 (зовнішні критерії для кожної пари внутрішніх критеріїв)):

$$QC_{bal} = \sqrt[6]{\prod_{i=1}^6 d_i}. \quad (3.16)$$

### 3.2.2.3. Алгоритм формування функції розрахунку критерію балансу для даних «Seeds»

Розрахунок критерію балансу для даних «Seeds» здійснюється відповідно до алгоритму:

1. Розрахунок коефіцієнтів лінійних рівнянь перетворення значень відповідних внутрішніх та зовнішніх критеріїв у значення показника  $Y$ :

$$\begin{cases} Y_{max} = a_{DB}^{A,B} - b_{DB}^{A,B} \cdot QC_{DB}^{\min(A,B)} \\ Y_{min} = a_{DB}^{A,B} - b_{DB}^{A,B} \cdot QC_{DB}^{\max(A,B)} \\ Y_{max} = a_{RL}^{A,B} + b_{RL}^{A,B} \cdot QC_{RL}^{\max(A,B)} \\ Y_{min} = a_{RL}^{A,B} + b_{RL}^{A,B} \cdot QC_{RL}^{\min(A,B)} \end{cases} \quad (3.17)$$

Розрахунок коефіцієнтів лінійних рівнянь для зовнішніх критеріїв здійснюється відповідно до системи рівнянь (3.10б).

де  $QC_{DB}^{\min(A,B)}$ ,  $QC_{DB}^{\max(A,B)}$ ,  $QC_{RL}^{\min(A,B)}$ ,  $QC_{RL}^{\max(A,B)}$  – мінімальне та максимальне значення критеріїв Davies\_Bouldin і Ratkovsky\_Lance, що розраховані для еквівалентних підмножин  $A$  і  $B$ .

2. Трансформування значень відповідних критеріїв якості кластеризації у значення показника  $Y$  відповідно до системи рівнянь:

$$\begin{cases} Y_{DB}^{A,B} = a_{DB}^{A,B} - b_{DB}^{A,B} \cdot QC_{DB}^{A,B} \\ Y_{RL}^{A,B} = a_{RL}^{A,B} + b_{RL}^{A,B} \cdot QC_{RL}^{A,B} \\ Y_{ext}^{A,B} = a_{ext}^{A,B} - b_{ext}^{A,B} \cdot QC_{ext}^{A,B} \end{cases} \quad (3.18)$$

3. Розрахунок приватних бажаностей для кожного критерію за формулою. (3.12).

4. Розрахунок критерію балансу за формулою (3.16).

#### 3.2.2.4. Алгоритм формування функції розрахунку критерію балансу для даних профілів експресії генів пацієнтів, що досліджувалися на рак легенів

Алгоритм розрахунку критерію балансу для даних профілів експресії генів пацієнтів, що досліджувалися на рак легенів містить такі кроки:

1. Розрахунок коефіцієнтів лінійних рівнянь перетворення значень відповідних внутрішніх та зовнішніх критеріїв у значення показника  $Y$ :

$$\begin{cases} Y_{max} = a_{WB}^{A,B} - b_{WB}^{A,B} \cdot QC_{WB}^{\min(A,B)} \\ Y_{min} = a_{WB}^{A,B} - b_{WB}^{A,B} \cdot QC_{WB}^{\max(A,B)} \\ Y_{max} = a_{PBM}^{A,B} + b_{PBM}^{A,B} \cdot QC_{PBM}^{\max(A,B)} \\ Y_{min} = a_{PBM}^{A,B} + b_{PBM}^{A,B} \cdot QC_{PBM}^{\min(A,B)} \end{cases} \quad (3.19)$$

Розрахунок коефіцієнтів лінійних рівнянь для зовнішніх критеріїв здійснюється відповідно до системи рівнянь (3.10б).

де  $QC_{WB}^{\min(A,B)}$ ,  $QC_{WB}^{\max(A,B)}$ ,  $QC_{PBM}^{\min(A,B)}$ ,  $QC_{PBM}^{\max(A,B)}$  – мінімальне та максимальне значення WB індексу і PBM критерія, що розраховані для еквівалентних підмножин  $A$  і  $B$ .

2. Трансформування значень відповідних критеріїв якості кластеризації у значення показника  $Y$  відповідно до системи рівнянь:

$$\begin{cases} Y_{WB}^{A,B} = a_{WB}^{A,B} - b_{WB}^{A,B} \cdot QC_{WB}^{A,B} \\ Y_{PBM}^{A,B} = a_{PBM}^{A,B} + b_{PBM}^{A,B} \cdot QC_{PBM}^{A,B} \\ Y_{ext}^{A,B} = a_{ext}^{A,B} - b_{ext}^{A,B} \cdot QC_{ext}^{A,B} \end{cases} \quad (3.20)$$

3. Розрахунок приватних бажаностей для кожного критерія за формулою. (3.12).

4. Розрахунок критерію балансу за формулою (3.16).

### 3.2.3. Результати моделювання щодо практичної реалізації гібридної індуктивної моделі щільнісного алгоритму кластеризації DBSCAN

У цьому підрозділі подано результати моделювання щодо практичної реалізації алгоритму, зображеному на рис. 3.5, на основі щільнісного алгоритму DBSCAN. Застосування такої моделі дозволяє оптимізувати параметри алгоритму за обраними критеріями якості кластеризації, враховуючи мінімізацію похибки відтворюваності.

#### 3.2.3.1. Практична реалізація гібридної індуктивної моделі щільнісного алгоритму кластеризації DBSCAN на даних «Compound»

На рис. 3.22 зображені k-dist діаграми, побудовані для значень  $minPts$  від 3 до 8, проаналізувавши які можна ініціювати інтервал зміни значення  $Eps$  для відповідного значення  $minPts$ , які мають включати коліно відповідної діаграми. Ці інтервали виділені червоною горизонтальною лінією.

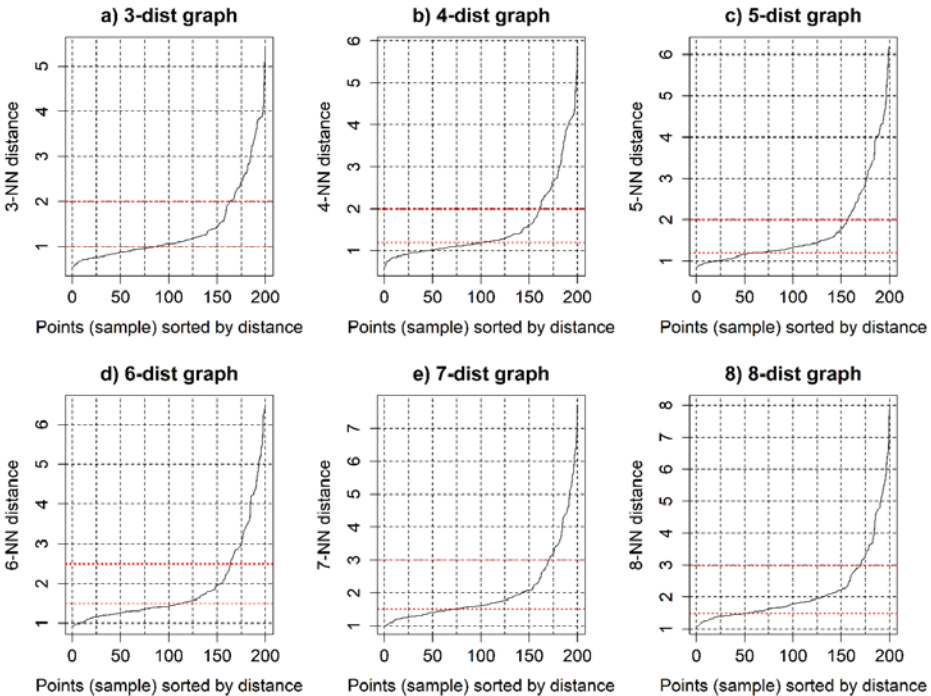


Рис. 3.22. k-dist діаграми для даних «Compound» при значеннях  $minPts$  від 3 до 8

У процесі роботи алгоритму значення  $Eps$  змінювалося від мінімального до максимального з кроком 0,004 (встановлений емпіричним шляхом) для кожного значення  $minPts$ . Результати моделювання щодо розподілу значень критерія балансу для значень  $Eps$ , при якому кількість кластерів у кластерних структурах, отриманих на еквівалентних підмножинах даних, є однаковими, зображені на рис. 3.23.

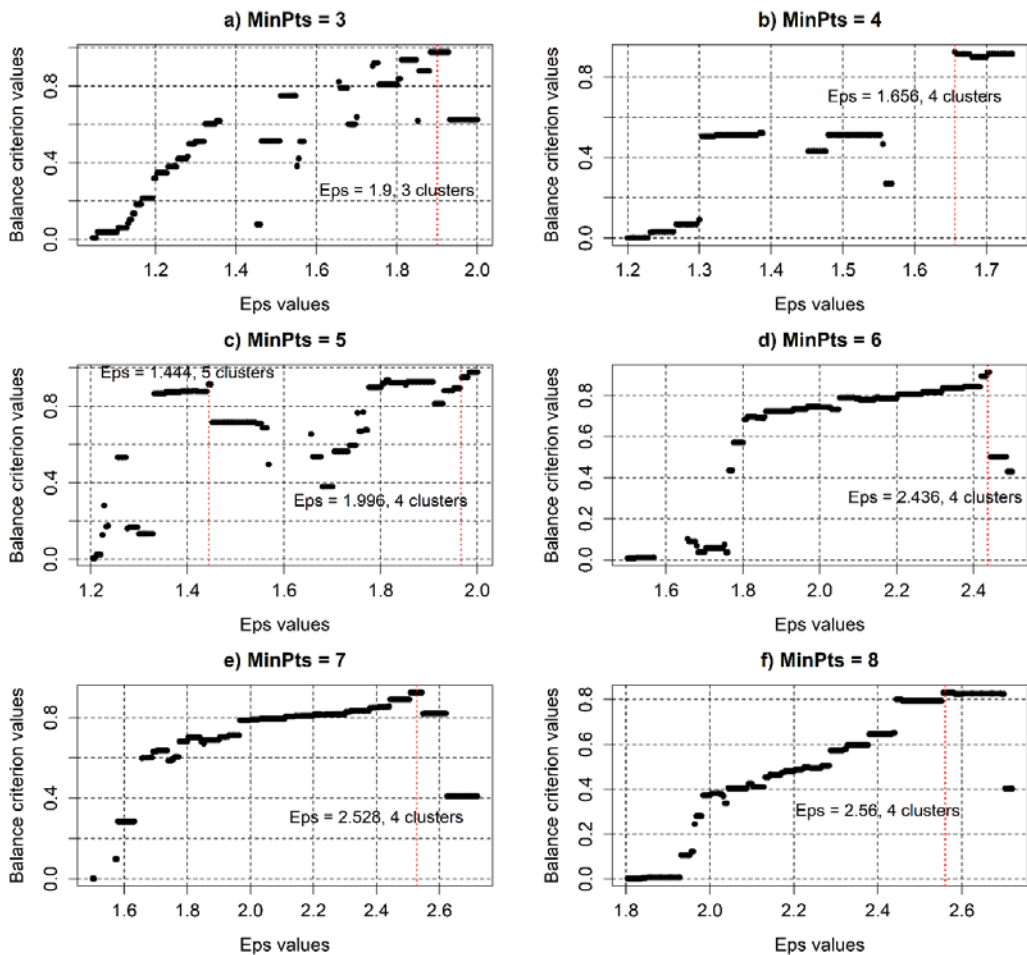


Рис. 3.23. Результати моделювання щодо визначення оптимальних параметрів алгоритму DBSCAN за критерієм балансу для даних «Compound»

Аналіз отриманих результатів дозволяє зробити висновок, що у більшості випадків оптимальна кластеризація за максимальним значенням критерію балансу відповідає чотирикластерній структурі. За характером розподілу об'єктів у кластерах усі кластеризації є коректними, але один кластер, що знаходиться у середині більшого кластера не ідентифікується. Кластеризації у таких випадках розрізняються тільки кількістю об'єктів, що ідентифікуються як шум. При значенні  $minPts = 3$  об'єкти у просторі розділені на три кластера, що також може бути сприй-

нятливим, оскільки кластери не перетинаються між собою. Отже оптимальною кластеризацією у цьому випадку є кластеризація, що відповідає параметрам алгоритму  $minPts = 5$ ,  $Eps = 1.444$ . У цьому випадку отримуються 5 кластерів. Характер розподілу об'єктів у просторі при оптимальних параметрах алгоритму зображений на рис. 3.24.

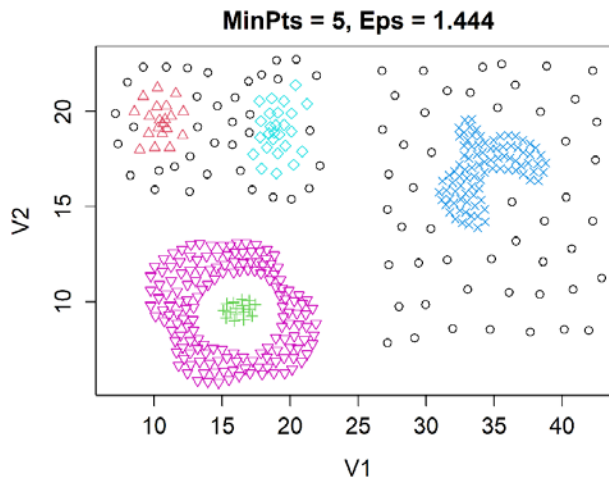


Рис. 3.24. Результати моделювання щодо кластеризації даних «Compound» при застосуванні гібридної індуктивної моделі алгоритму DBSCAN

Отже, характер розподілу об'єктів у кластерах є адекватним. До того ж, об'єкти, щільнісного розподілу яких менша за щільність розподілу об'єктів у кластерах, ідентифікуються як шум. Цей факт може бути застосований у системах фільтрації даних за щільнісними критеріями.

### 3.2.3.2. Практична реалізація гібридної індуктивної моделі щільнісного алгоритму кластеризації DBSCAN на даних «Jain»

На рис. 3.25 зображено результати розділення даних «Jain» на дві еквівалентних підмножини. Отже, за щільнісним алгоритмом один кластер виділяється однозначно, але щільність розподілу об'єктів у другому кластері варіюється у досить великому діапазоні, що ускладнює їх групування у один кластер. На рис. 3.26 зображено k-dist діаграми для значень  $minPts$  від 3 до 8, аналіз яких дозволив сформувати інтервал ті крок зміни значень  $Eps$  для кожного значення  $minPts$  (червоні лінії). Результати моделювання щодо розподілу значень критерію балансу при різних значеннях  $minPts$  та  $Eps$  зображено на рис. 3.27. Проаналізувавши отримані діаграми, можна зробити висновок, що при значеннях  $minPts$  3 і 4 результати кластеризації незадовільні. Значення критерію балансу дуже низьке.

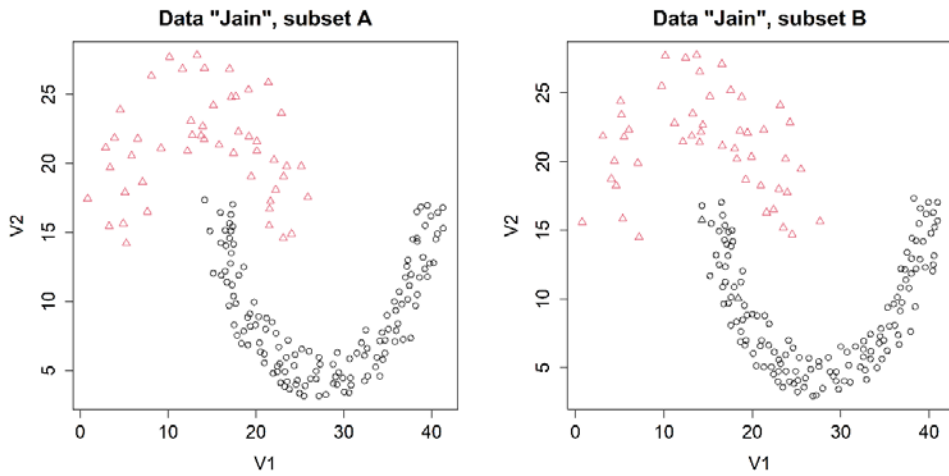


Рис. 3.25. Результати розділення даних «Jain» на дві еквівалентних підмножини

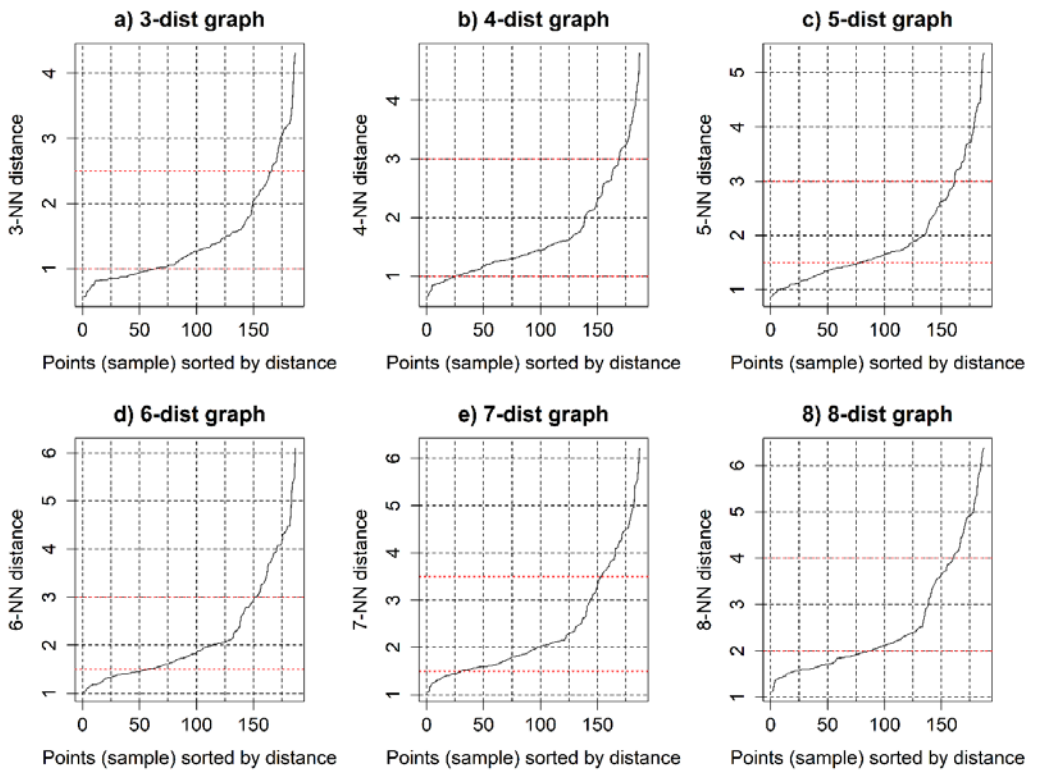


Рис. 3.26. k-dist діаграми для даних «Jain» при значеннях  $minPts$  від 3 до 8

У цих випадках алгоритм розбивав дані на велику кількість кластерів (6–8), при цьому спостерігалася велика розбіжність між кластеризаціями, отриманими на еквівалентних підмножинах А і В, що означає велику похибку відтворюваності.

Результати кластеризацій у випадках значень *minPts* 5, 6, 7 та 8, що відповідають максимумам критерію балансу, зображені на рис. 3.28. Отже, в усіх випадках один кластер виділяється однозначно, у другому кластері великий відсоток об'єктів ідентифікується як шум. Для цього типу даних це природньо, оскільки, як зазначалося раніше, густина розподілу даних об'єктів суттєво вища за густину розподілу основної групи об'єктів. Найкраща кластеризація у цьому випадку як за критерієм балансу, так і за візуальним сприйняттям досягається при параметрах алгоритму: *minPts* = 7, *Eps* = 2.625 (рис. 3.27a) і *minPts* = 8, *Eps* = 2.71 (рис. 3.27b). Результати розподілу об'єктів у даних випадках майже однакові. Отримано три кластери. У другому кластері виділено два підкластера, що, з точки зору густини розподілу об'єктів у просторі, є природнім.

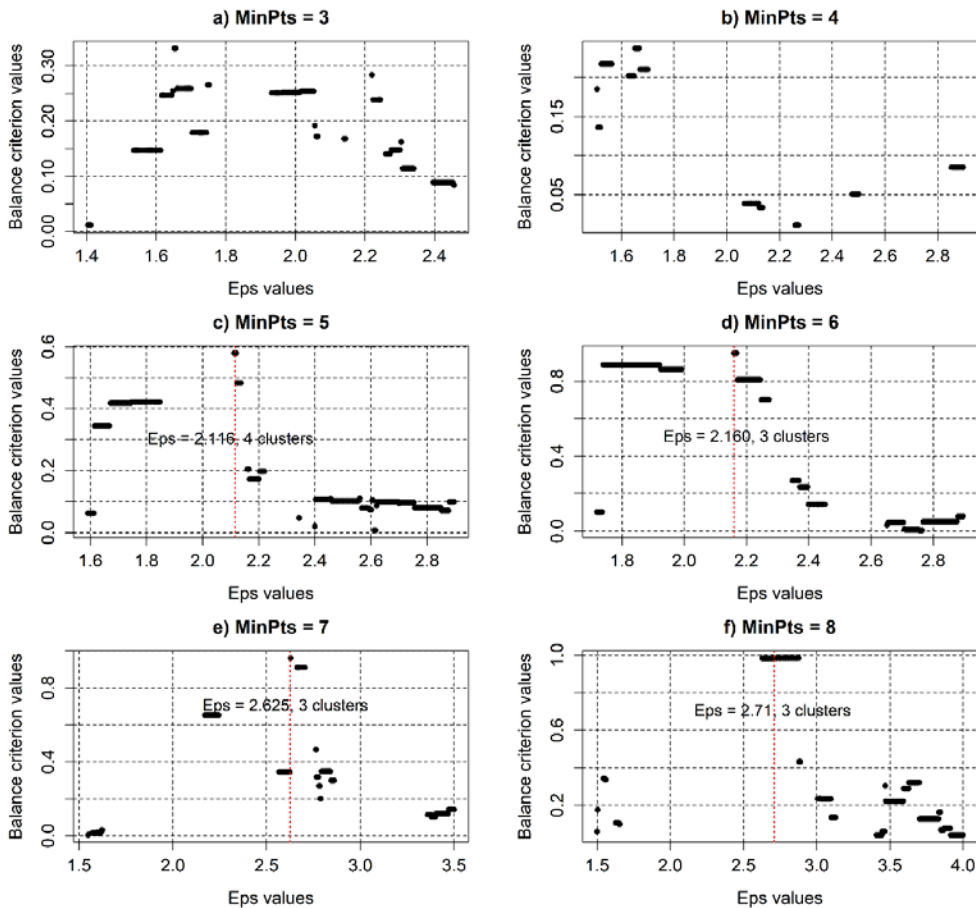


Рис. 3.27. Результати моделювання щодо визначення оптимальних параметрів алгоритму DBSCAN за критерієм балансу для даних «Jain»

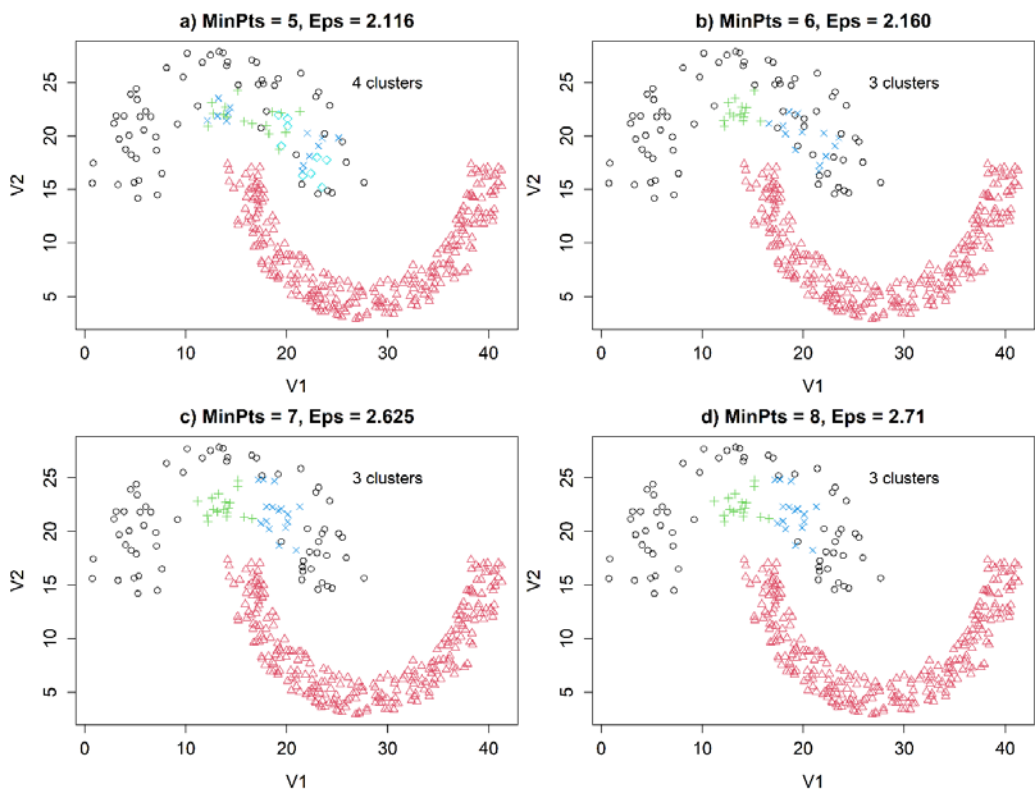


Рис. 3.28. Результати моделювання щодо кластеризації даних «Jain» при застосуванні гібридної індуктивної моделі алгоритму DBSCAN

### 3.2.3.3. Практична реалізація гібридної індуктивної моделі щільнісного алгоритму кластеризації DBSCAN на даних «Seeds»

Результати моделювання щодо визначення інтервалу та кроку зміни значення параметра *Eps* при різних *minPts* для даних *seeds* на основі *k*-dist графу наведено на рис. 3.29. На відміну від попередніх двовимірних даних, кожний об'єкт даних *seeds* характеризується сімома атрибутами, тому діапазон варіювання значень *Eps* при різних *minPts* є суттєво менший, порівняно з даними *compound* і *jain*. Тому, на основі аналізу *k*-dist діаграм, було обрано однаковий інтервал зміни значення *Eps* для усіх *minPts* від 0,75 до 2,5, при цьому крок зміни цього параметра був 0,002.

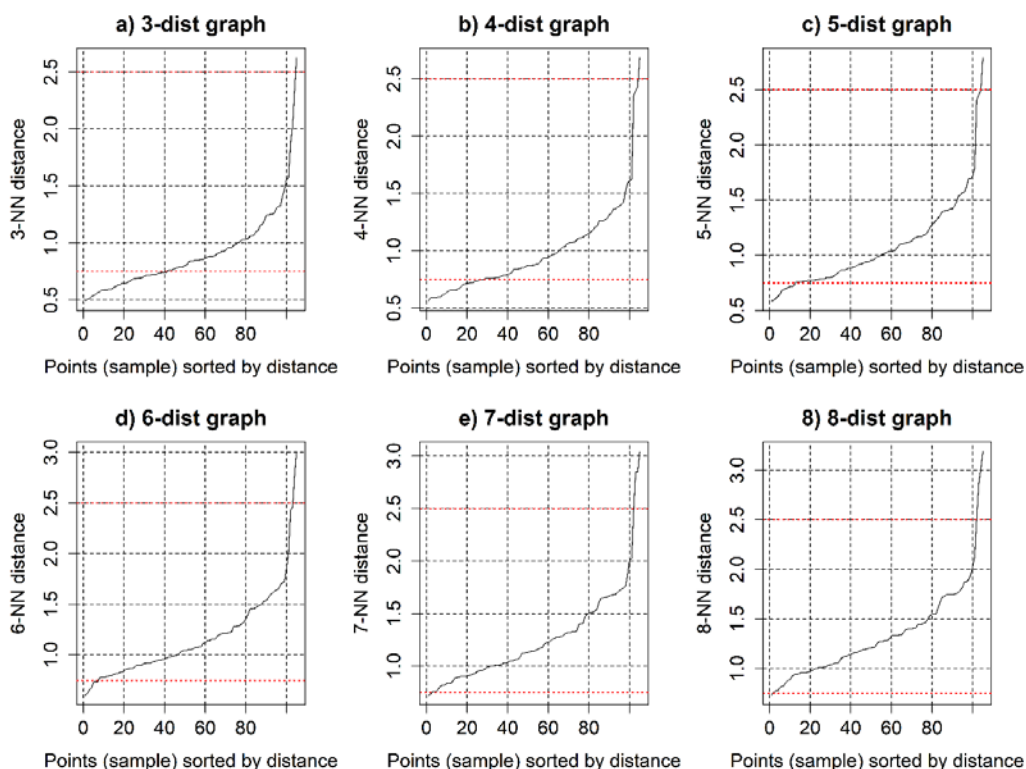


Рис. 3.29. k-dist діаграми для даних «Seeds» при значеннях  $minPts$  від 3 до 8

Результати моделювання стосовно оцінки розподілу значень критерію балансу при різних значеннях  $Eps$  та  $minPts$  зображені на рис. 3.30. На основі аналізу отриманих діаграм можна зробити висновок, що при  $minPts = 3, 5$  і  $8$  результати є незадовільними з точки зору значення критерію балансу (рис. 3.30a) і кількості кластерів (рис. 3.30c,f). Як показує попередній аналіз характеру розподілу об'єктів даних *seeds* (рис. 3.7) та їх анотація, дані містять три кластери, при чому дані кластери можуть перетинатися один із одним. Цей факт ускладнює розподіл об'єктів у кластери за щільністю. З отриманих результатів бачимо, що у випадку значень параметрів алгоритму  $Eps = 0.94$ ,  $minPts = 4$  і  $Eps = 0.91$ ,  $minPts = 6$  кількість кластерів відповідає цільовій кількості, але більшість об'єктів зосереджено у двох кластерах, тобто фактично маємо два кластера.

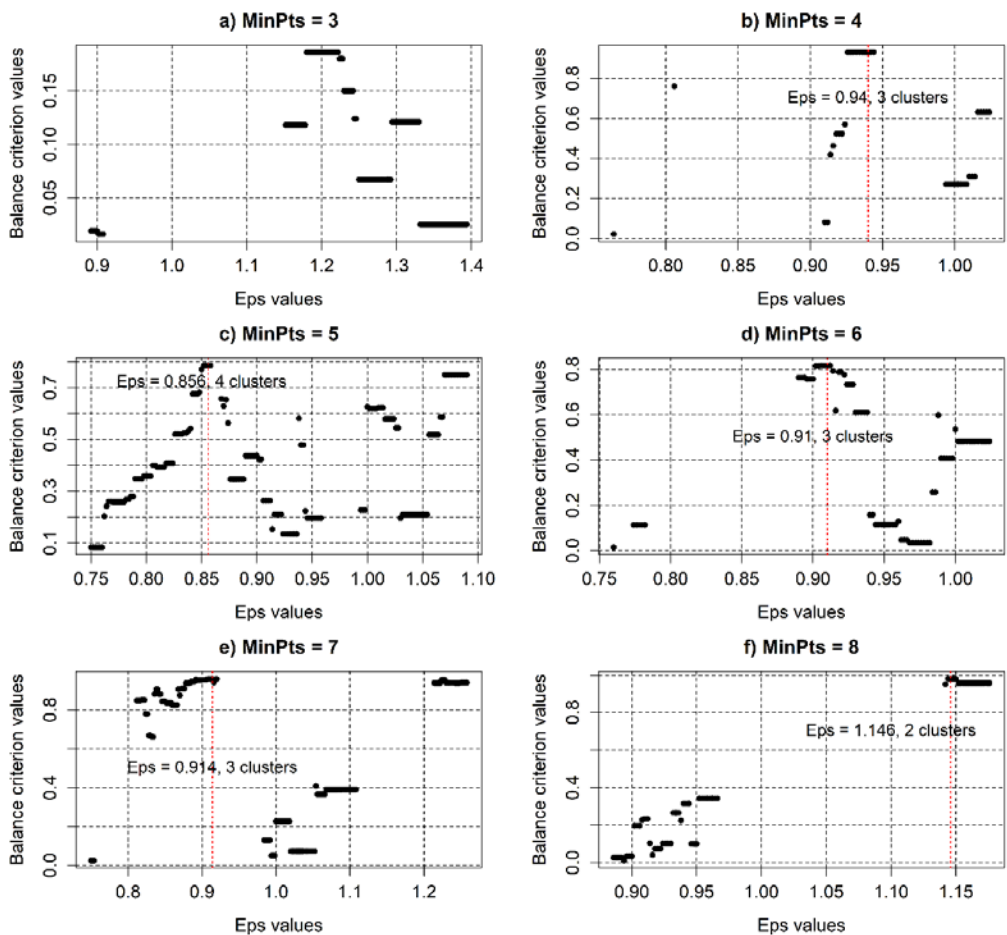


Рис. 3.30. Результати моделювання щодо визначення оптимальних параметрів алгоритму DBSCAN за критерієм балансу для даних «Seeds»

Оптимальною комбінацією значень параметрів у цьому випадку є така:  $Eps = 0.914$ ,  $minPts = 7$ . У цьому випадку отримуємо три кластери об'єктів, які не перетинаються між собою. При цьому слід зазначити, що 70 об'єктів із 210 ідентифікуються як шум, оскільки густина їх розподілу у просторі значно більша за густину розподілу основної частини об'єктів і варіюється у широкому інтервалі. Результати кластеризації вихідних даних і отриманих кластерів на основі двох головних компонент зображені на рис. 3.31. Слід зазначити, що отримані результати у цьому випадку відповідають меті дослідження, яка полягає у виділенні групи взаємно близьких за обраною метрикою профілів експресії генів, які дозволяють з високою точністю ідентифікувати об'єкти, що досліджуються.

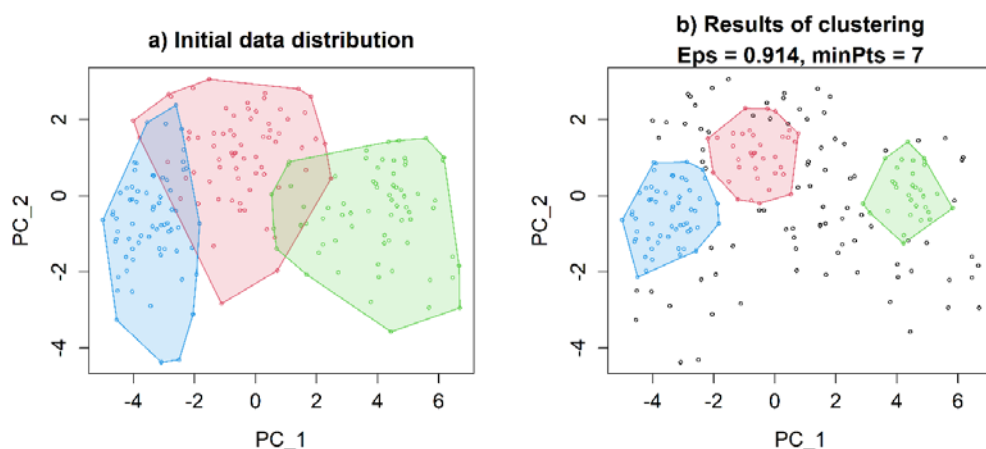


Рис. 3.31. Результати моделювання щодо кластеризації даних «Seeds» при застосуванні гібридної індуктивної моделі алгоритму DBSCAN

Таким чином, отримана кластеризація вихідних даних за їх анотацією (рис. 3.31a) підтверджує факт перетину кластерів, тобто деякі об'єкти можуть належати двом кластерам одночасно, що ускладнює їх однозначну ідентифікацію. Кластери об'єктів, які отримані шляхом застосування алгоритму DBSCAN (рис. 3.31b), не перетинаються між собою. Цей факт підтверджує доцільність застосування запропонованої гібридної індуктивної моделі для фільтрації та кластеризації високовимірних даних.

### 3.2.3.4. Практична реалізація гібридної індуктивної моделі щільнісного алгоритму кластеризації DBSCAN на даних профілів експресії генів

У процесі кластеризації профілів експресії генів на першому кроці формувалася матриця відстаней між усіма парами профілів експресії генів із застосуванням гібридної метрики відстаней, наведеної у розділі 2. Цей крок обумовлений тим, що Евклідова метрика для високовимірних профілів не є ефективною. Отримана матриця відстаней застосовувалася як вхідні дані при застосуванні алгоритму кластеризації DBSCAN. На рис. 3.32 зображені k-dist діаграми для граничних значень  $minPts$ , аналіз яких дозволив сформуванати інтервал і крок зміни значення параметра  $Eps$ :  $Eps_{min} = 10$ ;  $Eps_{max} = 60$ ;  $dEps = 0.1$ .

Оцінка якості відповідної кластерної структури здійснювалася шляхом застосування процедури класифікації об'єктів, що містять виділені у кластери профілі експресії генів як атрибути. Ця процедура виконувалася із застосуванням алгоритму класифікації «випадковий ліс» (random forest), при цьому у кожному випадку розраховувалася точність класифікації об'єктів. Результати моделювання щодо визначення комбінацій оптимальних параметрів алгоритму кластеризації за максимумом критерію балансу зображено на рис. 3.33.

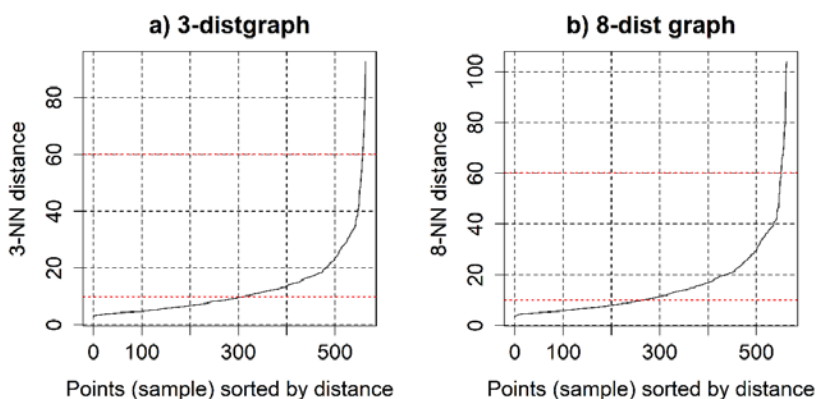


Рис. 3.32. k-dist діаграми для даних експресії генів при граничних значеннях  $minPts$  3 і 8

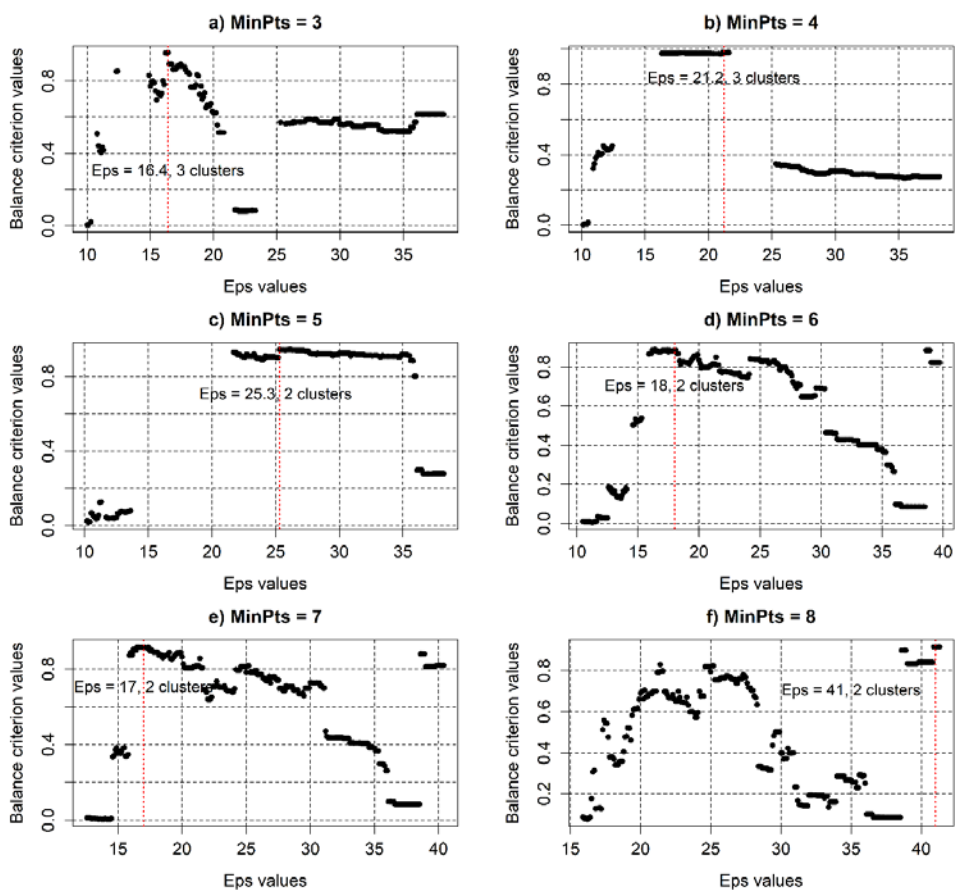


Рис. 3.33. Результати моделювання щодо визначення оптимальних параметрів алгоритму DBSCAN за критерієм балансу для даних експресії генів

У таблиці 3.4 наведено остаточні результати моделювання, а саме: в усіх випадках виділено нульовий кластер, що містить профілі експресії генів, які ідентифіковані як «шум». Точність класифікації досліджуваних об'єктів, що містять дані експресії генів є більш низькою, порівняно із основною групою профілів експресії генів.

Таблиця 3.4

*Результати моделювання щодо отримання оптимальних кластерних структур та оцінки їх якості шляхом застосування процедури класифікації*

| minPts = 3, Eps = 16.4 |                 |                       | minPts = 4, Eps = 21.2 |                 |                       |
|------------------------|-----------------|-----------------------|------------------------|-----------------|-----------------------|
| Кластер                | Кількість генів | Точність класифікації | Кластер                | Кількість генів | Точність класифікації |
| 0                      | 186             | 0.677                 | 0                      | 122             | 0.694                 |
| 1                      | 904             | 0.829                 | 1                      | 966             | 0.829                 |
| 2                      | 8               | 0.7                   | 2                      | 8               | 0.702                 |
| 3                      | 28              | 0.781                 | 3                      | 30              | 0.781                 |
| minPts = 5, Eps = 25.3 |                 |                       | minPts = 6, Eps = 18   |                 |                       |
| Кластер                | Кількість генів | Точність класифікації | Кластер                | Кількість генів | Точність класифікації |
| 0                      | 88              | 0.690                 | 0                      | 176             | 0.694                 |
| 1                      | 1006            | 0.814                 | 1                      | 922             | 0.814                 |
| 2                      | 32              | 0.780                 | 2                      | 28              | 0.781                 |
| minPts = 7, Eps = 17   |                 |                       | minPts = 8, Eps = 41   |                 |                       |
| Кластер                | Кількість генів | Точність класифікації | Кластер                | Кількість генів | Точність класифікації |
| 0                      | 198             | 0.694                 | 0                      | 24              | 0.642                 |
| 1                      | 900             | 0.812                 | 1                      | 1070            | 0.814                 |
| 2                      | 28              | 0.781                 | 2                      | 32              | 0.780                 |

До того ж аналіз отриманих результатів також показує, що застосування щільнісного алгоритму DBSCAN для кластеризації високовимірних профілів експресії генів не є ефективним, оскільки точність класифікації, порівняно з точністю класифікації вихідних даних (97%) при однакових умовах класифікації є нижчою для усіх кластерів. Однак запропонована гібридна індуктивна модель дозволяє сформувати оптимальну кластерну структуру як за характером розподілу об'єктів у кластерах, так і за мінімумом похибки відтворюваності.

#### 3.2.4. Результати моделювання щодо практичної реалізації гібридної індуктивної моделі щільнісного алгоритму кластеризації OPTICS

У цьому підрозділі наведено результати моделювання щодо практичної реалізації алгоритму, зображеному на рис. 3.5 на основі щільнісного алгоритму OPTICS.

### 3.2.4.1. Практична реалізація гібридної індуктивної моделі щільнісного алгоритму кластеризації OPTICS на даних «Compound»

На рис. 3.34 зображено діаграми досяжності для граничних значень параметра  $minPts$  (3 та 8), аналіз яких дозволяє встановити інтервал та крок зміни значень параметра  $Eps$  у процесі подальшого моделювання:

$$Eps_{min} = 0.5; Eps_{max} = 1.5; dEps = 0.005.$$

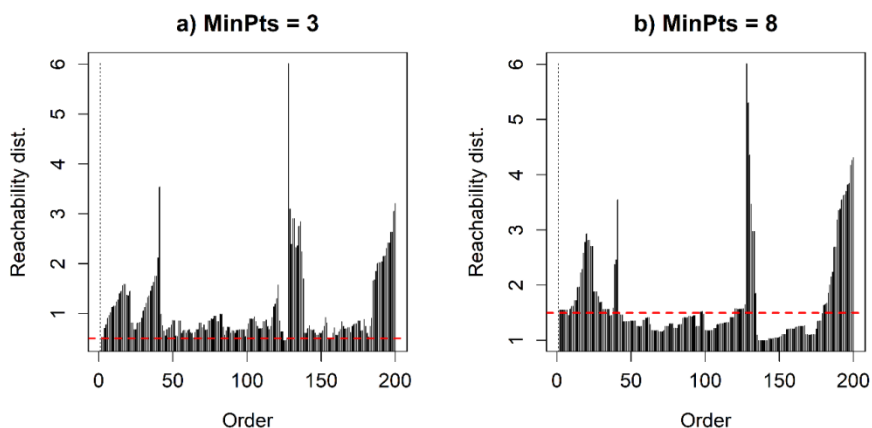


Рис. 3.34. Діаграми досяжності для граничних значень  $minPts$  (3 та 8) даних «Compound»

Результати моделювання щодо розрахунку критерію балансу при різних значеннях  $Eps$  і  $minPts$ , відповідно до запропонованої індуктивної моделі об'єктивної кластеризації, представлені на рис. 3.35.

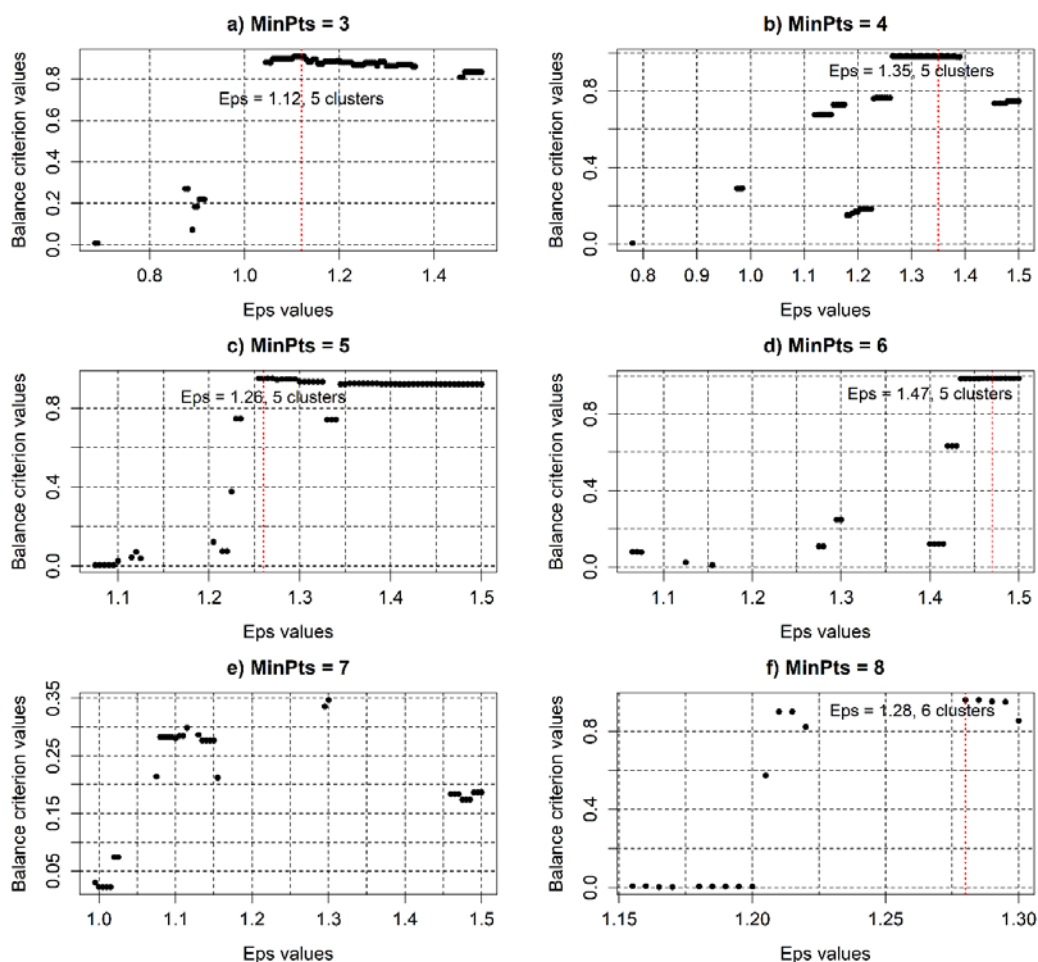


Рис. 3.35. Результати моделювання щодо визначення оптимальних параметрів алгоритму OPTICS за критерієм балансу для даних «Compound»

Проаналізувавши отримані діаграми, можемо визначити оптимальне значення *Eps* для кожного *minPts*. Результати моделювання для значень *minPts* від 3 до 6 зображені на рис. 3.36. При *minPts* = 7 значення критерію балансу показує незадовільні результати як за характером розподілу об'єктів у кластери, так і за похибкою відтворюваності. При *minPts* = 8, як показали результати моделювання, результати кластеризації також не відповідали характеру розподілу об'єктів у кластери.

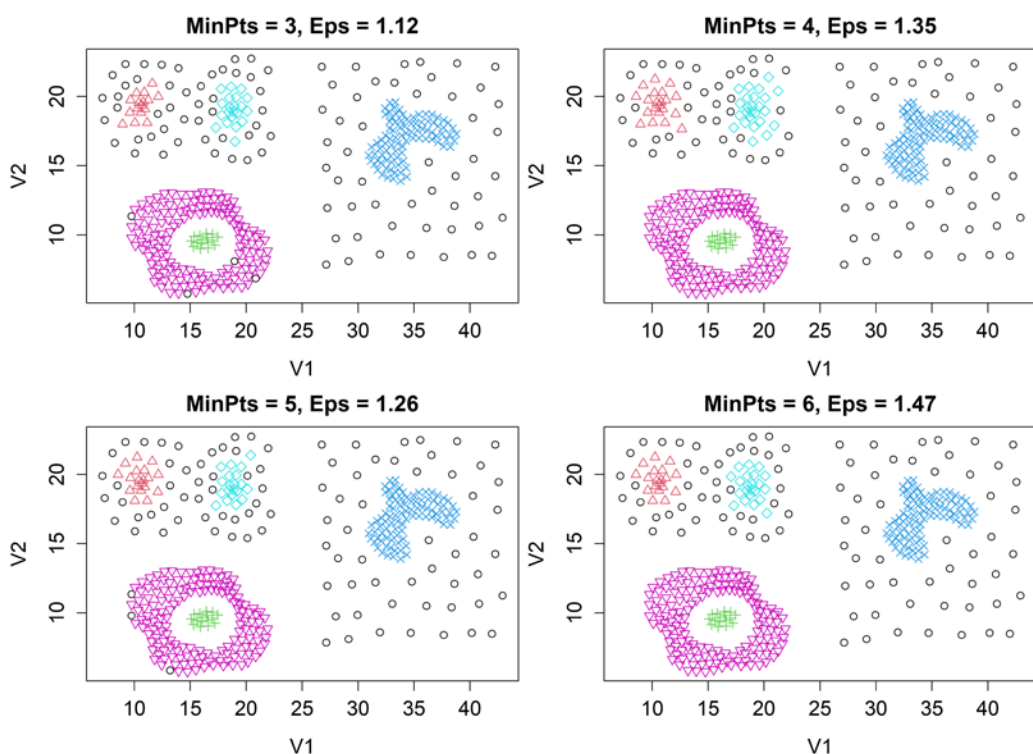


Рис. 3.36. Результати моделювання щодо кластеризації об'єктів даних «Compound» при застосуванні алгоритму кластеризації OPTICS

На основі аналізу отриманих результатів можна зробити висновок, що у усіх чотирьох випадках характер розподілу об'єктів у кластери є адекватним. Є невелика різниця у кількості об'єктів, що ідентифікована як "шум". Однак, порівняно з роботою алгоритму DBSCAN, алгоритм OPTICS стійкіший до зміни параметрів, що спрощує його адекватне застосування.

#### 3.2.4.2. Практична реалізація гібридної індуктивної моделі щільнісного алгоритму кластеризації OPTICS на даних «Jain»

На рис. 3.37 зображено діаграми досяжності для граничних значень параметру  $minPts$  (3 та 8) при застосуванні даних «Jain». Інтервал та крок зміни значень параметра  $Eps$  у цьому випадку був обраний такий:  $Eps_{min} = 0.4$ ;  $Eps_{max} = 2.8$ ;  $dEps = 0.005$ . Результати моделювання щодо визначення оптимальних комбінацій параметрів  $Eps$  і  $minPts$  за максимумом критерію балансу зображено на рис. 3.38, а на рис. 3.39 зображено результати кластеризації об'єктів даних «Jain» при визначених комбінаціях параметрів алгоритму OPTICS.

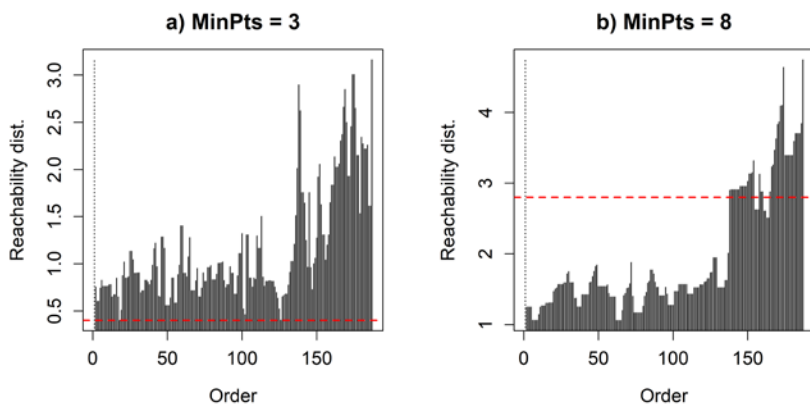


Рис. 3.37. Діаграми досяжності для граничних значень  $minPts$  (3 та 8) даних «Jain»

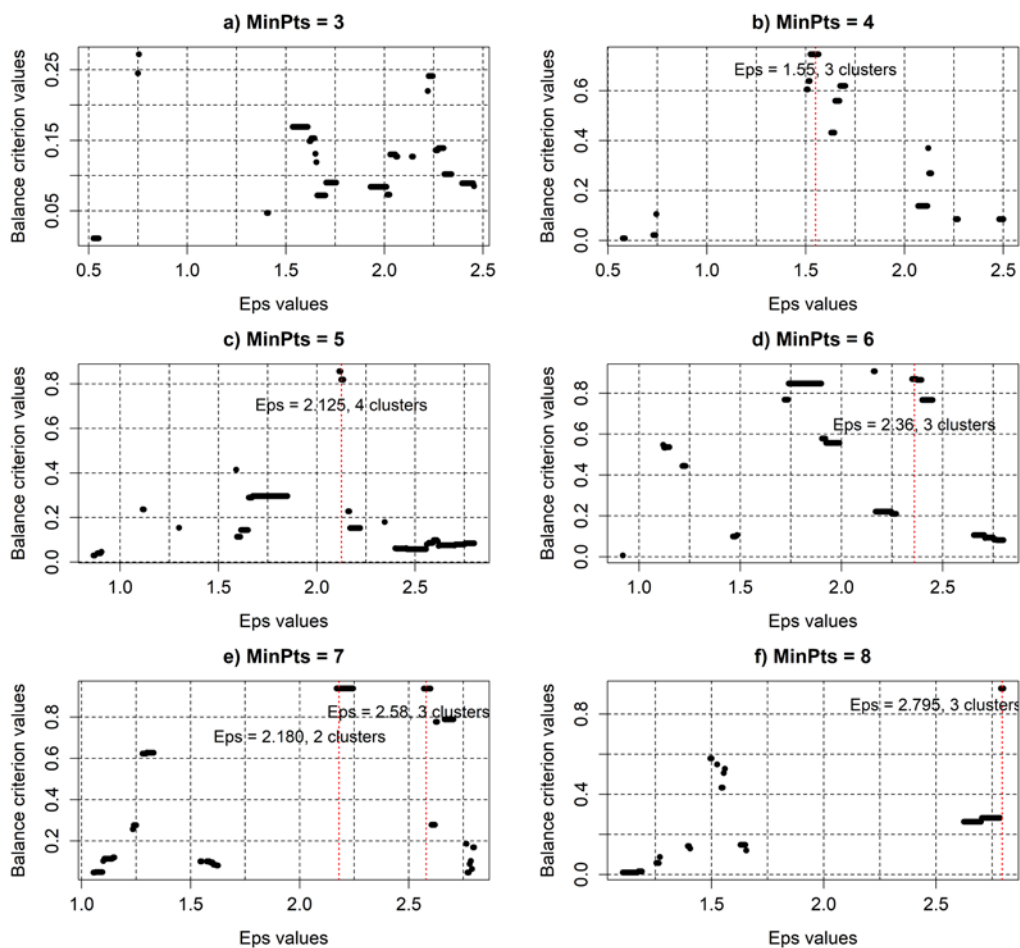


Рис. 3.38. Результати моделювання щодо визначення оптимальних параметрів алгоритму OPTICS за критерієм балансу для даних «Jain»

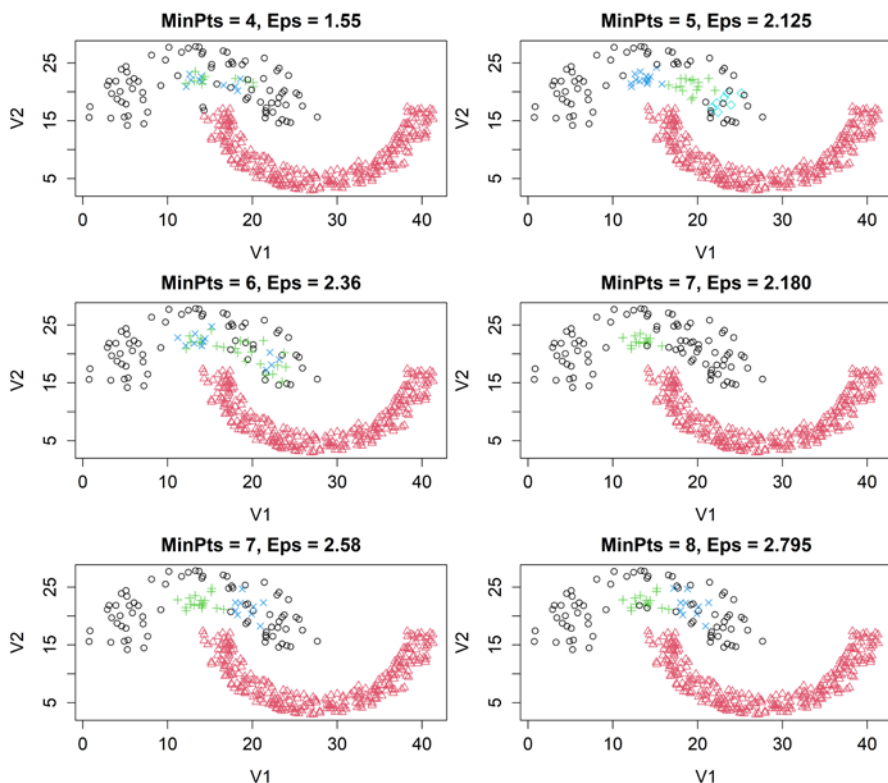


Рис. 3.39. Результати моделювання щодо кластеризації об'єктів даних «Jain» при застосуванні алгоритму кластеризації OPTICS

Отже, в усіх випадках однозначно виділений один кластер, щільність розподілу об'єктів у якому значно вища за щільність розподілу об'єктів у другому кластері. Аналогічно результатам, отриманим при застосуванні алгоритму DBSCAN, другий кластер також виділений частково. Велика кількість об'єктів у другому кластері ідентифікується як шум, що досить природно, враховуючи щільність розподілу даних об'єктів у просторі ознак. Проте слід також зазначити особливо високу стабільність алгоритму OPTICS, порівняно з алгоритмом DBSCAN, при формуванні кластерної структури. Цей факт є безумовною перевагою алгоритму кластеризації OPTICS, порівняно з алгоритмом DBSCAN.

### 3.2.4.3. Практична реалізація гібридної індуктивної моделі щільнісного алгоритму кластеризації OPTICS на даних «seeds»

На рис. 3.40 зображено діаграми досяжності для граничних значень параметра *minPts* (3 та 8) при застосуванні даних «seeds». Інтервал та крок зміни значень параметра *Eps* у цьому випадку був обраний такий:  $Eps_{min} = 0.35$ ;  $Eps_{max} = 1.2$ ;  $dEps = 0.002$ . На рис. 3.41 зображені результати моделювання щодо визначення оптимальних комбінацій параметрів *Eps* і *minPts* за максимумом критерію балансу.

Результати кластеризації об'єктів даних «seeds» при визначених комбінаціях параметрів алгоритму OPTICS для трикластерної структури (відповідно до анотації даних) зображені на рис. 3.42, на основі чого можна зробити висновок, що в усіх випадках велика кількість об'єктів за густиною їх розподілу у просторі ідентифікована як шум.

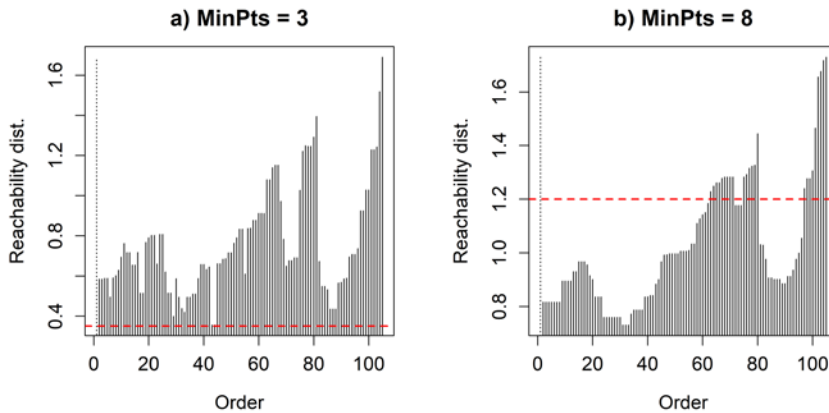


Рис. 3.40. Діаграми досяжності для даних «seeds»

Найкраща кластерна структура у цьому випадку зображена на рис. 3.41с ( $minPts = 7$ ;  $Eps = 0.934$ ) відповідає найвищому значенню критерія балансу:  $QC_{bal} = 0.975$ . Цей факт доводить адекватність запропонованої індуктивної технології щодо визначення оптимальної комбінації відповідного алгоритму кластеризації на основі аналізу як внутрішніх, так і зовнішніх критеріїв оцінки якості групування об'єктів у кластерах з урахування похибки відтворюваності.

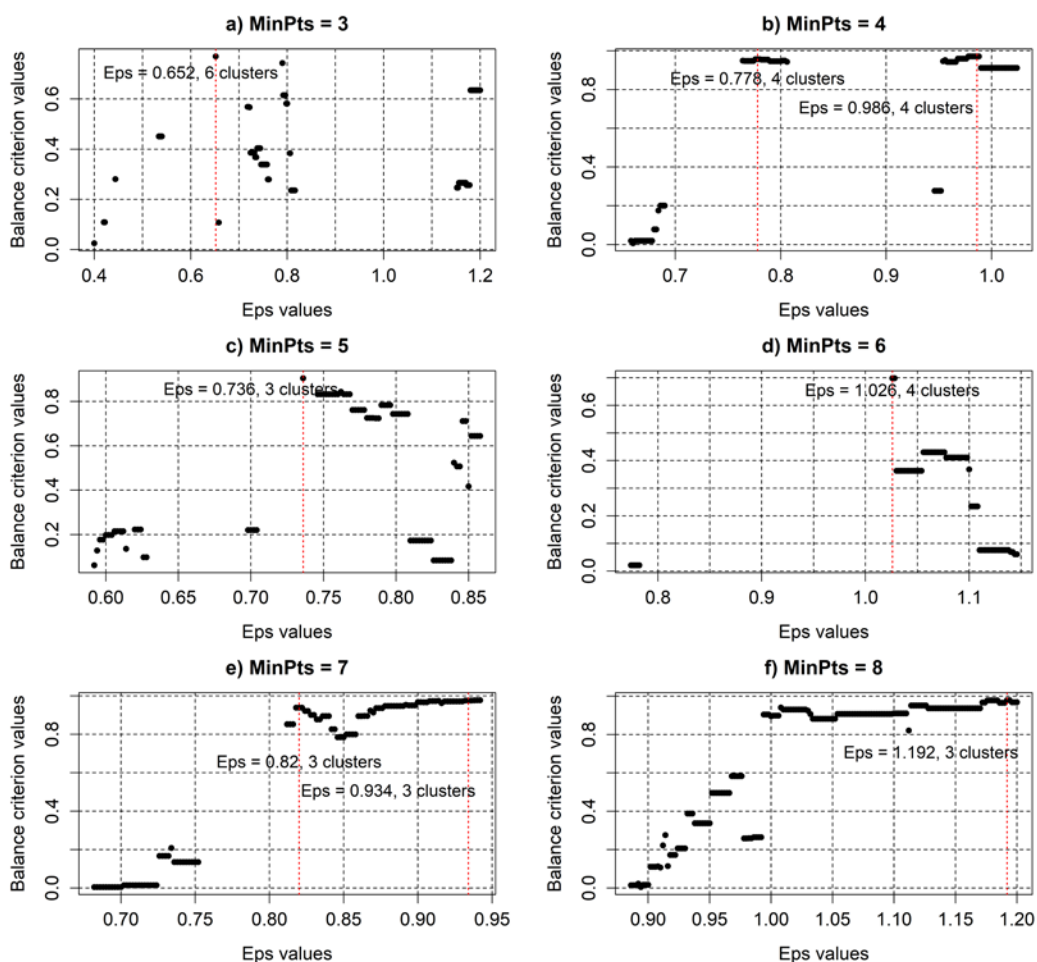


Рис. 3.41. Результати моделювання щодо визначення оптимальних параметрів алгоритму OPTICS за критерієм балансу для даних «seeds»

### 3.2.4.4. Практична реалізація гібридної індуктивної моделі щільнісного алгоритму кластеризації OPTICS на даних експресії генів

На рис. 3.43 зображено діаграми досяжності, що побудовані із застосуванням матриці відстаней даних експресії генів для граничних значень *minPts*. Проаналізувавши отримані діаграми можна зробити висновок, що профілі експресії генів переважно становлять один великий кластер. Можливе виділення декількох малих кластерів, які, не є визначальними для ідентифікації об'єктів, що досліджуються. Інтервал зміни значень *Eps* у цьому випадку був обраний такий:  $Eps_{min} = 2$ ;  $Eps_{max} = 24$ ;  $dEps = 0.05$ .

На рис. 3.44 зображено результати моделювання щодо визначення оптимальних параметрів гібридного індуктивного алгоритму кластеризації OPTICS, що відповідають максимальним значенням критерію балансу. Отже, здебільшого профілі експресії генів

розділяються на два кластери. Тільки у перших двох випадках, при *minPts* = 3 і 4, маємо три кластери.

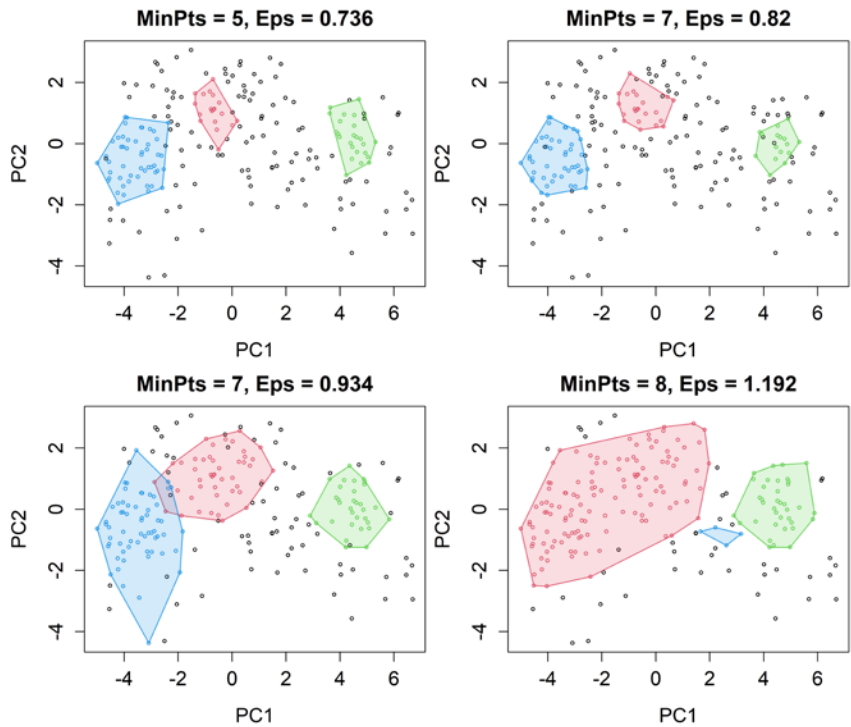


Рис. 3.42. Результати моделювання щодо кластеризації об’єктів даних «seeds» при застосуванні алгоритму кластеризації OPTICS

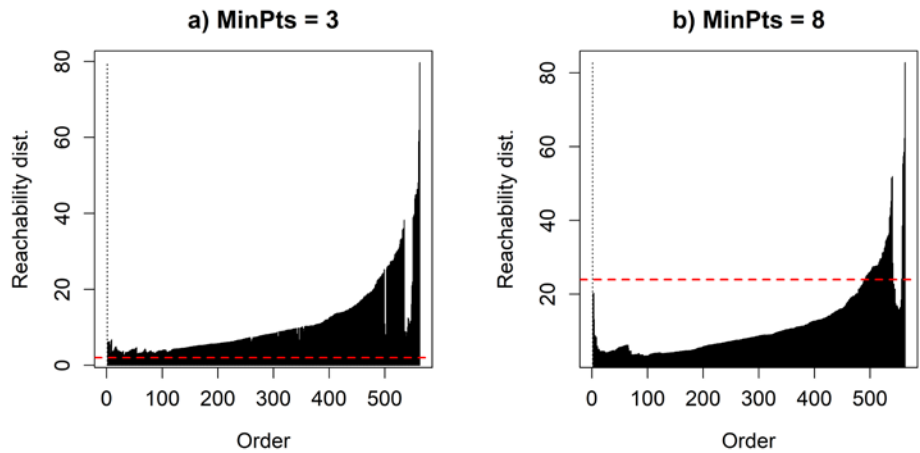


Рис. 3.43. Діаграми досяжності для даних профілів експресії генів

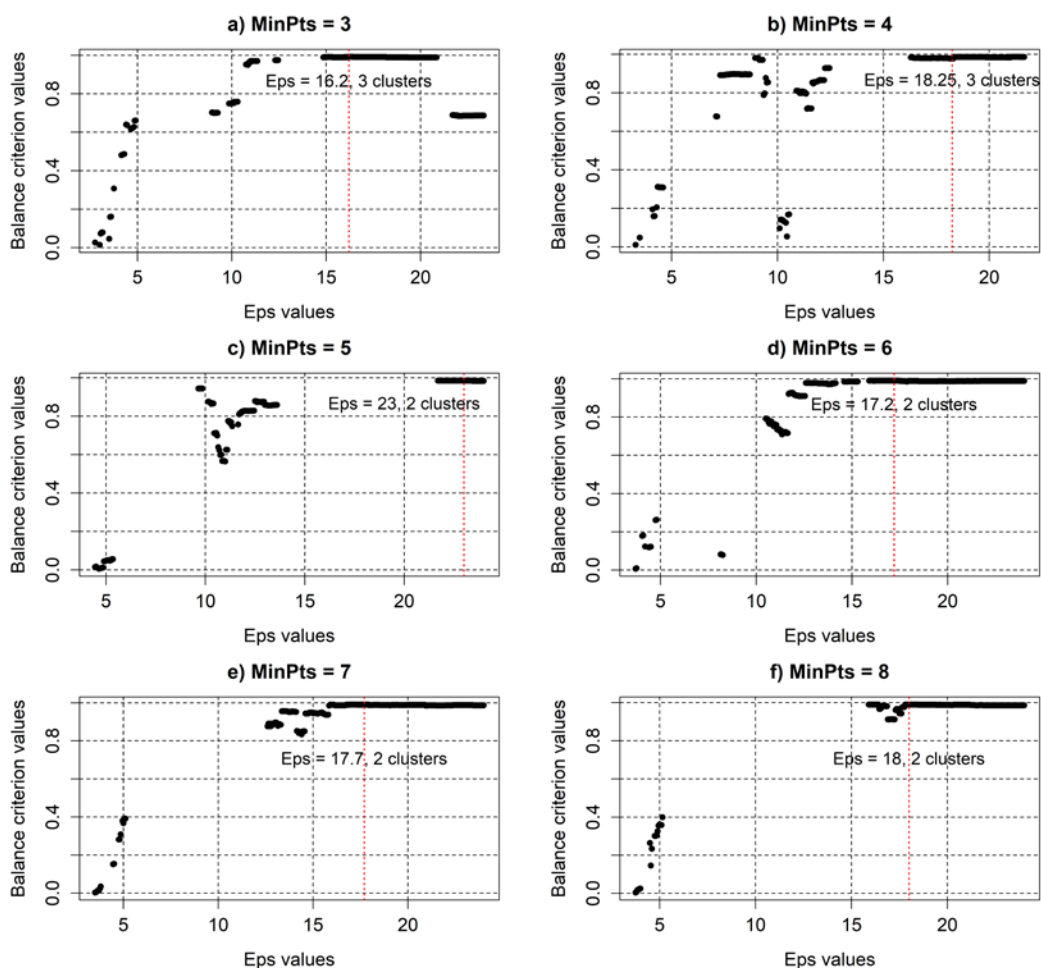


Рис. 3.44. Результати моделювання щодо визначення оптимальних параметрів алгоритму OPTICS за критерієм балансу для даних експресії генів

На основі аналізу отриманих результатів підтверджується зроблений раніше висновок, що основна група генів виділяється у один кластер. Точність класифікації об'єктів цьому випадку вища за точність класифікації об'єктів, які містять як атрибути профілі експресії генів інших кластерів. Як і у випадку застосування алгоритму DBSCAN значна кількість профілів експресії генів у кожному випадку за густиною їх розподілу у просторі ідентифікується як шум. Очевидно, що точність класифікації об'єктів, що містять дані профілі експресії генів, є нижчою порівняно з основними кластерами.

У таблиці 3.5 наведено остаточні результати оцінки кількості профілів експресії генів у виділених кластерах та точність класифікації об'єктів, що містять виділені профілі експресії генів як атрибути.

Таблиця 3.5

*Результати моделювання щодо отримання оптимальних кластерних структур та оцінки їх якості шляхом застосування процедури класифікації*

| minPts = 3, Eps = 16.2 |                 |                       | minPts = 4, Eps = 18.25 |                 |                       |
|------------------------|-----------------|-----------------------|-------------------------|-----------------|-----------------------|
| Кластер                | Кількість генів | Точність класифікації | Кластер                 | Кількість генів | Точність класифікації |
| 0                      | 190             | 0.677                 | 0                       | 158             | 0.694                 |
| 1                      | 900             | 0.813                 | 1                       | 932             | 0.813                 |
| 2                      | 8               | 0.716                 | 2                       | 8               | 0.716                 |
| 3                      | 28              | 0.797                 | 3                       | 30              | 0.797                 |
| minPts = 5, Eps = 23   |                 |                       | minPts = 6, Eps = 17.2  |                 |                       |
| Кластер                | Кількість генів | Точність класифікації | Кластер                 | Кількість генів | Точність класифікації |
| 0                      | 114             | 0.710                 | 0                       | 188             | 0.694                 |
| 1                      | 980             | 0.797                 | 1                       | 910             | 0.813                 |
| 2                      | 32              | 0.781                 | 2                       | 28              | 0.797                 |
| minPts = 7, Eps = 17.7 |                 |                       | minPts = 8, Eps = 18    |                 |                       |
| Кластер                | Кількість генів | Точність класифікації | Кластер                 | Кількість генів | Точність класифікації |
| 0                      | 186             | 0.677                 | 0                       | 174             | 0.677                 |
| 1                      | 912             | 0.829                 | 1                       | 924             | 0.813                 |
| 2                      | 28              | 0.797                 | 2                       | 28              | 0.797                 |

Праналізувавши дані таблиці 3.5 можна також зробити висновок щодо доцільності застосування індуктивної гібридної моделі для кластеризації даних, оскільки її використання сприяє підвищенню об'єктивності прийняття рішення при визначенні параметрів алгоритму за рахунок ретельного вибору критеріїв кластеризації, з одного боку, та мінімізації похибки відтворюваності за рахунок мінімізації різниці у результатах кластеризації, отриманих на еквівалентних підмножин даних, з іншого боку. Однак слід зазначити, що застосування щільнісних алгоритмів як основних для кластеризації високовимірних профілів експресії генів не є доцільним, оскільки точність класифікації об'єктів, що містять кластери профілів експресії генів як атрибути, не є високою, що потребує подальших досліджень стосовно ретельного вибору ансамблю алгоритмів кластеризації, орієнтованих на обробку профілів експресії генів та їх реалізації у рамках індуктивної моделі об'єктивної кластеризації.

## Розділ 4.

### **Гібридні моделі формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів**

Цей розділ присвячений розробці гібридних моделей формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів на основі комплексного застосування методів інтелектуального аналізу даних та машинного навчання [6, 7, 9–14, 16–25, 27, 28]. Під диференційно експресованими генами розуміються гени, рівень експресії яких максимально відрізняється для різних типів об'єктів. Іншими словами, значення експресії однакових генів для хворих та здорових пацієнтів або для хворих пацієнтів з різним станом хвороби мають максимально відрізнятися між собою. З іншого боку, профілі експресії генів мають бути максимально взаємно корельовані між собою. При наявності такої підмножини диференційно експресованих та взаємно корельованих профілів експресії генів створюються умови для розробки високоефективних систем діагностики стану здоров'я пацієнтів на основі даних експресії генів, що можуть дозволити ідентифікувати на ранньому етапі такі хвороби, як різні типи раку, хворобу Паркінсона, Альцгеймера, нейроblastoma тощо. Формування таких підмножин профілів експресії генів створює умови для створення ефективних систем ранньої діагностики складних хвороб на основі даних експресії генів або реконструкції високоінформативних генних регуляторних мереж, аналіз результатів моделювання яких дозволить краще зрозуміти характер взаємодії між цільовими генами та транскрипційними факторами з метою формування конкретних дій для зміни експресії відповідних генів та/або транскрипційних факторів для нейтралізації відповідної хвороби шляхом застосування сучасних ефективних ліків та/або методики лікування.

На основі теоретичних досліджень, що наведені у попередніх розділах, можна зробити висновок, що традиційні алгоритми інтелектуального аналізу та машинного навчання не є ефективними для профілів експресії генів, особливою яких є велика кількість атрибутів (висока розмірність). Так, щільнісні алгоритми кластеризації, які дозволяють з високою точністю виділити складні кластери об'єктів, кількість атрибутів у яких є невеликою, для профілів експресії генів не є ефективними. Тому виникає проблема вибору або розробки алгоритму кластеризації або ансамблю алгоритмів, що орієнтовані на високимірні профілі експресії генів і враховують принципи індуктивної технології об'єктивної кластеризації. З іншого боку, об'єктивна оцінка адекватності роботи алгоритму кластеризації щодо формування підмножин профілів експресії генів можлива за умови аналізу результатів класифікації об'єктів, що містять значення експресії виділених генів як атрибути. У попередньому розділі як класифікатор застосовувався алгоритм випадкового лісу (Random Forest), який показав високу ефективність при класифікації профілів експресії генів при наявності двох класів об'єктів. При більшій кількості класів застосування цього алгоритму є проблематичним. Тому виникає необхідність у

порівняльному аналізі різних типів класифікаторів, включаючи нейронні мережі з метою створення класифікатора, орієнтованого на високовимірні дані експресії генів, що дозволяє з високою точністю ідентифікувати об'єкти у мульти-класових задачах. Як результат, створюються умови для розробки гібридної моделі формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів, у якій для групування профілів застосовується кластерний аналіз, а оцінка адекватності кластеризації здійснюється шляхом застосування класифікатора, орієнтованого на цей тип даних. Цей розділ містить розв'язки вищенаведених задач.

#### 4.1. Аналіз ефективності згорткових нейронних мереж для класифікації об'єктів на основі даних експресії генів

##### 4.1.1. Архітектура, будова та модель згорткової нейронної мережі

Згорткові нейронні мережі (ЗНМ – англ. Convolutional Neural Network (CNN)) являють собою тип глибоких нейронних мереж (ГНМ – англ. Deep Neural Network (DNN)), які сьогодні успішно застосовуються для аналізу та класифікації великих даних, таких як різні типи зображень, акустичні сигнали, спектрограми, ехограми, тощо [253–256]. Загальна архітектура ЗНМ зображена на рис. 4.1.

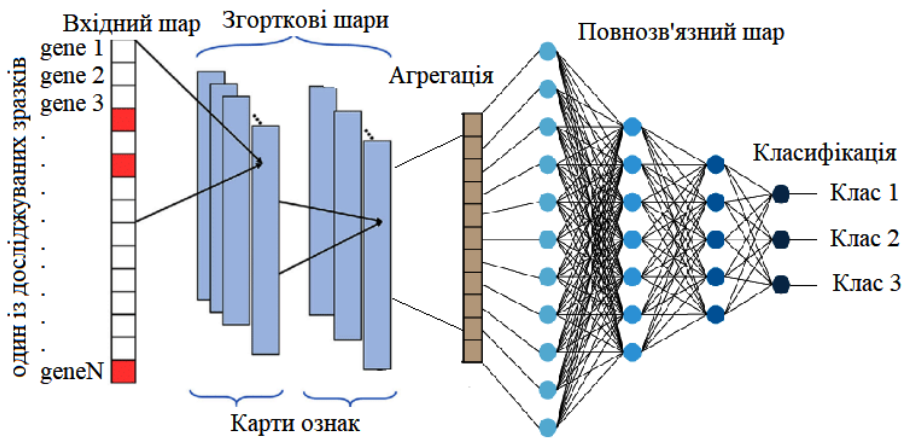


Рис. 4.1. Загальна архітектура згорткової нейронної мережі

Як вхідний шар, у цьому випадку застосовується вектор значень експресії генів одного із зразків, що досліджується. Кількість генів може досягати декілька тисяч, тому застосування класичних нейронних мереж ускладнюється за рахунок великої кількості нейронів, що ускладнює вибір оптимальної архітектури мережі. Як вхідні дані при використанні ЗНМ можуть застосовуватися зображення, речові сигнали, часові ряди, електрокардіограми тощо. При цьому застосування ЗНМ дозволяє створити класифікатор з високою точністю ідентифікації відповідних об'єктів. Як бачимо з рис. 4.1, першим етапом обробки даних у ЗНМ є операція згортки. Ця операція є однією із найголовніших і передбачає декомпозицію вхідних даних на

карти ознак шляхом застосування відповідного ядра згортки (kernel), при цьому кожна карта ознак зображена у вигляді тензора параметрів, що відповідають обмеженій частині сигналу, що досліджується. Таким чином, перший етап деякою мірою являє собою фільтр, що застосовується для декомпозиції сигналу на компоненти, кожна з яких містить інформацію щодо відповідної компоненти сигналу. З цієї причини вибір кількості фільтрів (карт ознак) є одним із найважливіших кроків на етапі ініціалізації параметрів мережі, і кількість фільтрів є одним із гіперпараметрів мережі. У цьому випадку, збільшення кількості фільтрів призводить до збільшення рівня деталізації ознак сигналу, але при цьому занадто велика кількість ознак може призвести до перенавчання мережі, тобто можуть бути виділені групи ознак, які не є характеристикою об'єкта, що досліджується. Цей факт може призвести до великої розбіжності точності класифікації об'єктів у процесі навчання мережі та подальшого її тестування і валідації отриманої моделі. З іншого боку, мала кількість фільтрів може призвести до зменшення розділяючої здатності мережі, тобто окремі суттєві компоненти сигналу, що мають значну вагу при класифікації об'єкта, не будуть виділені, що також призведе до погіршення результатів класифікації як на рівні навчання, так і на рівні тестування й валідації моделі.

Математична модель операції згортки вхідного вектора значень експресії генів з формуванням відповідної карти ознак може бути подана таким чином:

$$s(t) = (e * w)(t) = \int_{-\infty}^{+\infty} e(\tau)w(t - \tau)d\tau, \quad (4.1)$$

де  $s(t)$  – карта  $t$ -ї групи ознак;  $e$  – значення експресії відповідного гену;  $w$  – ядерна функція (kernel) або детектор ознак для формування групи взаємно корельованих ознак вектора експресії генів.

Реалізація операції згортки також передбачає, що значення ядерної функції за межами відповідної групи ознак дорівнюватиме нулю, тому нескінченна сума, що застосовується у формулі (4.1) може бути замінена на скінченні суми із значеннями у виділеному діапазоні для відповідної карти ознак. До основних гіперпараметрів функції реалізації операції згортки слід також віднести крок обходу сигналу  $s$ . Очевидно, що збільшення значення цього параметра призводить до зменшення кількості карт ознак, тобто модель у цьому випадку є більш простішою, але точність її функціонування може зменшитися через накладання ознак сигналу для різних полів. Наступним гіперпараметром моделі є коефіцієнт доповнення нулями  $p$  (padding). Застосування цього параметра дозволяє зберегти розмірність вхідних даних при формуванні карт ознак, при цьому значення, що не відповідають виділеним ознакам, доповнюються нулями. Розмірність фільтрів ознак визначається з урахуванням того, що збільшення розмірності призведе до виділення більшої кількості ознак, при цьому додатково виділені ознаки зазвичай зустрічаються не часто, але вони можуть мати більшу вагу для ідентифікації об'єкта, що досліджується [257]. Загальний процес трансформації частини вектора експресії генів у відповідний згортковий шар можна подати таким рівнянням:

$$y = f(w * e + b), \quad (4.2)$$

де  $u$  – карта ознак, що сформована на відповідному згортковому шарі;  $w$  – ядерна функція або детектор ознак;  $e$  – вектор значень експресії генів (вектор вхідних даних);  $b$  – параметр, що визначає крок зсуву фільтра в процесі реалізації операції згортки.

Наступним етапом реалізації моделі ЗНМ є агрегація виділених згорткових шарів з метою формування повнозв'язаного шару на наступному кроці цієї процедури. Як зазначалося вище, однією з переваг шару згортки є те, що згортка є локальною операцією. Просторова протяжність ядра зазвичай невелика. Один елемент карти ознак обчислюється з використанням лише невеликої кількості елементів на його вході. Необхідність формування повнозв'язаного шару визначається тим, що для обчислення будь-якого елемента на виході мережі (на підставі значень якого здійснюється ідентифікація об'єкту при розв'язанні задачі класифікації) потрібні всі елементи на вході класифікатора. Таким чином, процедура об'єднання шарів з метою формування повнозв'язаного шару дозволяє застосовувати повний вектор або матрицю вхідних даних для навчання глибинної моделі ЗНМ, при цьому на етапі формування карт ознак здійснювався етап фільтрації для виділення найвагоміших ознак для подальшого застосування у повнозв'язному шарі.

Як видно з рис. 4.1, на ранньому етапі формування повнозв'язного шару формуються шари агрегації (pooling layers), які можуть бути глобальними і містити повний вектор атрибутів вхідного сигналу (як показано на рис. 4.1) та локальними, тобто містити тільки вектор підмножини атрибутів вхідного сигналу, які будуть подаватися на відповідні нейрони поточного шару. Головною метою етапу агрегації є формування вектора атрибутів даних меншої розмірності, порівняно з розмірністю вхідних даних, при цьому зберігаються найінформативніші характеристики вхідного сигналу шляхом формування зв'язків між відповідними нейронами з попереднього шару і одним нейроном поточного шару. З цієї причини операція агрегування зазвичай застосовується покроково між сформованими згортковими шарами [258]. Агрегація може здійснюватися шляхом застосування операцій усереднення, максимізації або мінімізації, при цьому для відповідної групи нейронів розраховується середнє значення у першому випадку, виділяється максимальне або мінімальне у другому та третьому випадках, відповідно. Формально вище перелічені операції різних типів агрегування можуть бути подані таким чином:

— усереднене агрегування:

$$e(i, j) = \frac{\sum_{s,p} x(i-s, j-p)}{s \cdot p} ; \quad (4.3)$$

— максимізація:

$$e(i, j) = \max_{s,p} x(i-s, j-p) ; \quad (4.4)$$

— мінімізація:

$$e(i, j) = \min_{s,p} x(i-s, j-p), \quad (4.5)$$

де  $e(i, j)$  – значення експресії елемента поточного рівня, положення якого визначається координатами  $i, j$ ;  $x$  – значення експресії елементів на попередніх рівнях з координатами  $(i-s, j-p)$ ;  $s, p$  – розмірність поля відповідних ознак.

Слід зазначити, що результат формування шарів агрегації визначається аналогічними гіперпараметрами, які визначають результат операції згортки. Зазвичай, як гіперпараметри застосовуються крок обходу сигналу  $b$  та розмірність відповідного рецептивного поля.

Архітектура повнозв'язного шару у ЗНМ зазвичай відповідає архітектурі звичайних нейронних мереж, у яких всі нейрони поточного шару зв'язані з нейронами попереднього шару. Як видно на рис. 4.1, повнозв'язний шар у ЗНМ застосовується на передостанньому кроці після етапів згортки і агрегування, що дає можливість суттєво спростити архітектуру мережі повнозв'язного шару за рахунок зменшення розмірності вхідного вектора даних.

Розрахунок значень ваг нейронів відповідного повнозв'язного рівня здійснюється за аналогічною формулою, за якою формуються шари згортки за відмінністю, що  $w$  є масивом (або вектором) параметрів відповідного повнозв'язного шару:

$$y = f(w^T * e + b). \quad (4.6)$$

Як і у класичних нейронних мережах, нейрони повнозв'язного шару ЗНМ передбачають застосування функції активації для розрахунку значення на виході нейрону на основі вектора вхідних сигналів. Як функції активації нейронів вихідного шару ЗНМ зазвичай використовуються такі функції:

— Сігмоїдальна функція:

$$y = \frac{1}{1 + \exp(-e)}, \quad (4.7)$$

де  $e$  визначає значення експресії, що подається на вхід нейрона. Значення  $y$  на виході нейрона у цьому випадку варіюється у діапазоні  $[0,1]$ , що робить зручним застосування сігмоїдальної функції на останньому рівні повнозв'язного шару у класифікаторах.

— Softmax функція:

$$y = \max_{i=1,N} \frac{\exp(e_i)}{\sum_{j=1}^N \exp(e_j)}, \quad (4.8)$$

де  $e_i$  – значення експресії  $i$ -го елемента попереднього шару;  $N$  – кількість елементів на попередньому шарі.

Застосування цієї функції також призводить до нормалізації вхідних даних у діапазон  $[0,1]$ , але якщо застосування сігмоїдальної функції доцільне переважно для розв'язання задач бінарної класифікації, то функція Softmax зазвичай дозволяє отримати високу точність класифікації мультикласових даних, при цьому належність об'єкта до того або іншого класу визначається максимально ймовірністю на виході нейрона вихідного шару. Крім того, ця функція більш стійка до шуму, порівняно із сігмоїдальною функцією. Збільшення значень виходів нейронів на попередньому рівні на сталі значення майже не змінює значення на виході нейрона поточного шару.

Як було зазначено вище, на останньому рівні повнозв'язного шару зазвичай використовують сігмоїдальну або Softmax функції активації. Інші функції активації, такі як лінійна, тангенціальна, гіперболічний тангенс тощо, для цього шару є менш

ефективними. Для нейронів вхідного та середніх шарів особливо поширеною є лінійна функція активації, але вибір відповідної функції активації зазвичай визначається емпіричним шляхом на етапі моделювання, враховуючи тип даних, що досліджується. Слід зазначити, що вибір функцій активації для нейронів відповідних шарів суттєво впливає на результат роботи мережі, оскільки при некоректному виборі комбінації функцій активації коректне налаштування параметрів мережі у процесі її навчання може бути проблематичним.

Контроль процесу навчання мережі для налаштування параметрів нейронів її шарів здійснюється шляхом розрахунку рівню втрат на виході мережі. При цьому порівнюється значення на виході мережі з еталонним значенням, що відповідає поточному сигналу, поданому на вхід мережі. Функція втрат (loss function) у цьому випадку є одним із основних параметрів при налаштуванні параметрів мережі, оскільки вона значною мірою визначає ефективність процесу навчання мережі. До основних типів функцій втрат слід віднести функцію перехресної ентропії (cross-entropy function) та функція розрахунку середньоквадратичної похибки (mean-squared error function) [259, 260]. При застосуванні ЗНМ на виході мережі розраховується ймовірність того, що поточний вектор даних (у нашому випадку експресії генів) належить тому або іншому класу. Шляхом зіставлення поточних ймовірностей належності вектора даних до того або іншого класу з цільовими ймовірностями (0 або 1) відповідно до принципу правдоподібності параметри мережі підлаштовується так, щоб мінімізувати розбіжності на виході мережі. При цьому навчання мережі зазвичай здійснюється шляхом застосування стандартного алгоритму зворотного поширення помилки. Функція перехресної ентропії розраховується так: [261]

$$L(e, p) = - \sum_{i=1}^N t_i \cdot \log(p_i), \quad (4.9)$$

де  $e$  – поточний вектор значень експресії генів, які визначають стан об'єкта, що досліджується;  $N$  – кількість класів, до яких можуть належати об'єкти досліджуваних даних;  $t_i$  – цільовий  $i$ -й клас, до якого належить поточний вектор значень експресії генів;  $p_i$  – ймовірність належності поточного вектору значень експресії генів до  $i$ -го класу, що розраховується ЗНМ.

Слід зазначити, що ймовірність на виході мережі може бути розрахована шляхом застосування будь-якої функції активації, але зазвичай для широкого класу задач застосовується Softmax-функція у випадку як бінарної, так і мультикласової класифікацій. Для бінарної класифікації може бути також застосована сігмоїдальна функція активації. Як зазначалося раніше, вибір функції активації визначається структурою даних, що досліджуються і у кожному випадку здійснюється окремо шляхом оцінки результатів моделювання.

## 4.1.2. Моделювання щодо застосування згорткової нейронної мережі у задачах класифікації даних експресії генів

### 4.1.2.1. Експериментальні дані профілів експресії генів

Моделювання згорткової нейронної мережі та оцінка ефективності її роботи для класифікації профілів експресії генів здійснювалася із застосуванням даних експресії генів GSE19188 [262] пацієнтів, що досліджувалися на рак легенів. Дані взяті з бази GSE (Gene Expression Omnibus) [263] і містять результати аналізу ДНК 156 пацієнтів шляхом застосування технології ДНК-мікрочіпів, при цьому 65 пацієнтів за результатами аналізу ідентифіковані як здорові, а у 91 пацієнта виявлена ракова пухлина. Кожний мікрочіп містив 54675 генів. Матриці експресії генів формувалися шляхом застосування функції пакету Bioconductor [264] мови програмування R [265]. На першому етапі до профілів експресії генів (векторів значень експресії гену, що визначені для усіх зразків (довжина вектора дорівнює 156)) застосовувалася операція редукції з метою видалення неінформативних генів за статистичними та ентропійними критеріями. Як статистичні критерії середнє абсолютне значення експресії генів та дисперсія значень експресії у відповідних профілях. Як ентропійний критерій використовувалася ентропія Шеннона, розрахована за методом Джеймса та Стейна [64]. Передбачалося, що якщо абсолютне значення експресії гену для усіх зразків, що досліджувалися, та значення дисперсії профіля експресії цього гену менші за граничні значення, а значення ентропії Шеннона цього профілю більше за граничне значення, то цей профіль видалявся з бази даних як неінформативний. Ця умова визначалася тим, що функціонування біологічного організму визначають приблизно 25000 генів, тобто із 54675 генів приблизно половина є неекспресовані і їх можна видалити за абсолютним значенням без втрати корисної інформації щодо стану здоров'я відповідного біологічного організму. Із половини, що залишилася, частина генів мають низьке значення експресії для усіх зразків, тобто ці гени не є визначальними для ідентифікації відповідної хвороби і їх також можна видалити без суттєвої втрати інформації. До того ж, якщо дисперсія відповідних профілів експресії генів є малою, дані гени не є інформативними з точки зору розділяючої здатності об'єктів, що досліджуються. Високе значення ентропії Шеннона також відповідає хаотичній зміні експресії гену (шум), тобто цей критерій також може бути використаний для ідентифікації неінформативних генів. Таким чином, умова формування підмножини інформативних генів шляхом видалення неінформативних (редукція) у цьому випадку має вигляд:

$$\{e_{ij}\} = \left\{ \begin{array}{l} \max_{i=\overline{1,n}} e_{ij} \geq e_{lim}, \text{ and } var(e_j) \geq var_{lim}, \\ \text{and } entr(e_j) \leq entr_{lim} \end{array} \right\}, j = \overline{1,m}, \quad (4.10)$$

де  $n$  – кількість зразків або об'єктів, що досліджуються;  $m$  – кількість генів.

Граничні значення відповідних критеріїв визначалися емпіричним шляхом у процесі моделювання. На рис. 4.2 зображені коробкові діаграми розкиду значень критеріїв, що використовувалися для формування підмножини профілів експресії генів відповідно до формули (4.10).

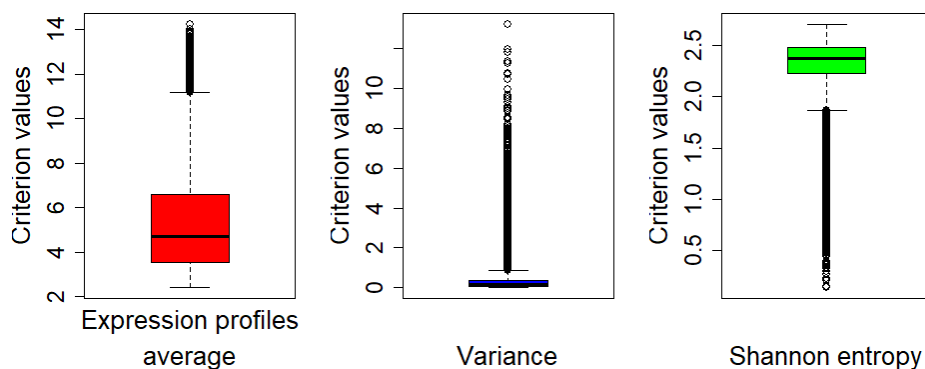


Рис. 4.2. Коробкові діаграми розкиду середніх абсолютних значень експресії генів, дисперсії та ентропії Шеннона профілів експресії генів

Як можна бачити, більшість генів мають відносно велике значення експресії, а дисперсія та ентропія Шеннона профілів експресії генів мають переважно високе та низьке значення, відповідно. Цей факт підтверджує доцільність формування підмножини інформативних профілів експресії генів за формулою (4.10). У рамках досліджень були обрані такі граничні значення критеріїв:

- абсолютне значення експресії генів – на рівні 25-го квантилю;
- дисперсія – на рівні 25-го квантилю;
- ентропія Шеннона – на рівні 75-го квантилю;

При цьому було виділено приблизно 10312 генів. Для зручності подальшого моделювання значення критеріїв незначно підлаштовувалося для виділення 10000 профілів експресії генів. Цей масив даних  $156 \times 10000$  був застосований як експериментальні дані в процесі подальшого моделювання згорткової нейронної мережі.

#### 4.1.2.2. Модель одновимірної згорткової нейронної мережі

Модель одновимірної згорткової мережі, що застосовувалася для класифікації об'єктів на основі сформованої підмножини даних експресії генів зображено на рис. 4.3.

Відповідно до розмірності експериментальних даних, довжина вектора експресії генів, що подавався на вхід мережі, становила 10000. На першому кроці до кожного вектора застосовувався фільтр довжиною 25, 50 та 100 генів. Кількість ділянок експресії генів, які покроково подавалися на вхід мережі, складали 400, 200 і 100, відповідно. При цьому розмір ядра складав 8 та 16 у першому випадку, 8, 16 і 32 – у другому і 8, 16, 32 і 64 – у третьому. До нейронів згорткового шару при цьому застосовувалася лінійна функція активації. Розмір ядра ущільненого шару складав 64, 128, 256 і 512. До нейронів ущільненого шару застосовувалася порогова функція активації ReLU (Rectified Linear Unit). До нейронів вихідного шару застосовувалася функція активації Softmax, при цьому рівень втрат визначався шляхом застосування функції перехресної ентропії (cross-entropy function).

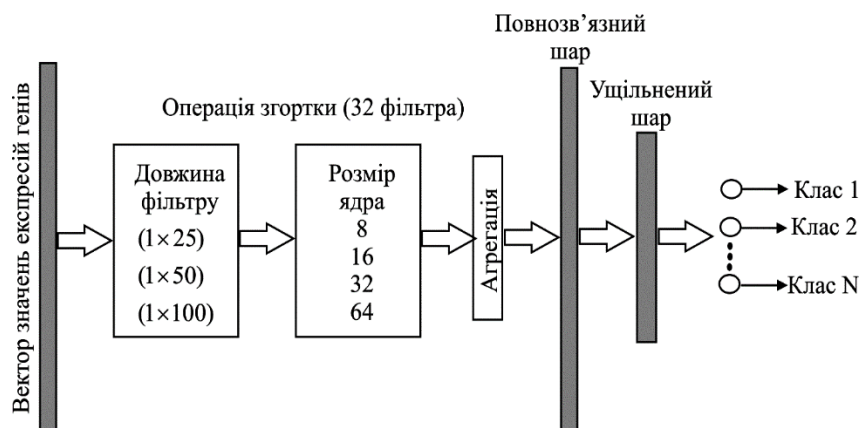


Рис. 4.3. Модель одновимірної згорткової нейронної мережі

Як випливає з анотації експериментальних даних, зразки, що досліджуються, можна розділити на дві групи (класи). Першу групу складають здорові пацієнти, а у пацієнтів другої групи ідентифікована ракова пухлина. На першому етапі 156 векторів експресії генів розділялися на навчальну підмножину (75% – 117 зразків) та тестову (25% – 39 зразків). Для контролю характеру протікання навчання мережі навчальна мережа розбивалася ще на дві підмножини: 93 вектори значень експресії генів застосовувалися безпосередньо для навчання, а 24, що залишилися, для валідації моделі. Кількість епох проходження даних експресії генів у процесі навчання було 50. Результати моделювання, що отримані на тестовій підмножині профілів експресії генів при різних комбінаціях гіперпараметрів, представлені у таблиці 4.1.

Таблиця 4.1

*Результати моделювання щодо застосування одновимірної ЗНМ для класифікації об'єктів на основі даних експресії генів (підмножина для тестування)*

| Гіперпараметри |                |             | Результати класифікації |           |             |        |
|----------------|----------------|-------------|-------------------------|-----------|-------------|--------|
| Ущільнений шар | Розмір фільтра | Розмір ядра | F-індекс                |           | Точність, % | Втрати |
|                |                |             | Кластер 1               | Кластер 2 |             |        |
| 1              | 2              | 3           | 4                       | 5         | 6           | 7      |
| 64             | (1,25)         | 8           | 0.96                    | 0.98      | 97          | 0.207  |
| 64             | (1,25)         | 16          | 0.9                     | 0.94      | 92          | 0.8    |
| 128            | (1,25)         | 8           | 0.9                     | 0.94      | 92          | 0.306  |
| 128            | (1,25)         | 16          | 0.96                    | 0.98      | 97          | 0.314  |
| 256            | (1,25)         | 8           | 0.88                    | 0.94      | 92          | 0.211  |
| 256            | (1,25)         | 16          | 0.87                    | 0.92      | 90          | 0.777  |
| 512            | (1,25)         | 8           | 0.9                     | 0.94      | 92          | 0.337  |
| 512            | (1,25)         | 16          | 0.9                     | 0.94      | 92          | 0.912  |
| 64             | (1,50)         | 8           | 0.96                    | 0.98      | 97          | 0.240  |
| 64             | (1,50)         | 16          | 0.93                    | 0.96      | 95          | 0.415  |

| 1          | 2              | 3        | 4           | 5           | 6         | 7            |
|------------|----------------|----------|-------------|-------------|-----------|--------------|
| 64         | (1,50)         | 32       | 0.96        | 0.98        | 97        | 0.458        |
| 128        | (1,50)         | 8        | 0.87        | 0.92        | 90        | 0.585        |
| 128        | (1,50)         | 16       | 0.87        | 0.92        | 90        | 0.743        |
| 128        | (1,50)         | 32       | 0.90        | 0.94        | 92        | 0.575        |
| 256        | (1,50)         | 8        | 0.96        | 0.98        | 97        | 0.145        |
| 256        | (1,50)         | 16       | 0.93        | 0.96        | 95        | 0.310        |
| 256        | (1,50)         | 32       | 0.96        | 0.98        | 97        | 0.406        |
| 512        | (1,50)         | 8        | 0.92        | 0.96        | 95        | 0.146        |
| 512        | (1,50)         | 16       | 0.96        | 0.98        | 97        | 0.148        |
| 512        | (1,50)         | 32       | 0.96        | 0.98        | 97        | 0.346        |
| 64         | (1,100)        | 8        | 0.96        | 0.98        | 97        | 0.169        |
| 64         | (1,100)        | 16       | 0.93        | 0.96        | 95        | 0.146        |
| 64         | (1,100)        | 32       | 0.89        | 0.93        | 91        | 0.226        |
| 64         | (1,100)        | 64       | 0.90        | 0.94        | 92        | 0.232        |
| 128        | (1,100)        | 8        | 0.92        | 0.96        | 95        | 0.135        |
| 128        | (1,100)        | 16       | 0.96        | 0.98        | 97        | 0.113        |
| 128        | (1,100)        | 32       | 0.96        | 0.98        | 97        | 0.154        |
| 128        | (1,100)        | 64       | 0.93        | 0.96        | 95        | 0.574        |
| <b>256</b> | <b>(1,100)</b> | <b>8</b> | <b>0.96</b> | <b>0.98</b> | <b>97</b> | <b>0.112</b> |
| 256        | (1,100)        | 16       | 0.96        | 0.98        | 97        | 0.141        |
| 256        | (1,100)        | 32       | 0.96        | 0.98        | 97        | 0.143        |
| 256        | (1,100)        | 64       | 0.92        | 0.96        | 95        | 0.242        |
| 512        | (1,100)        | 8        | 0.96        | 0.98        | 97        | 0.139        |
| 512        | (1,100)        | 16       | 0.93        | 0.96        | 95        | 0.205        |
| 512        | (1,100)        | 32       | 0.90        | 0.94        | 92        | 0.217        |
| 512        | (1,100)        | 64       | 0.93        | 0.96        | 95        | 0.463        |

На рис. 4.4 зображені коробкові діаграми розкиду відповідних критеріїв якості класифікації об'єктів, що застосовувалися у процесі моделювання.

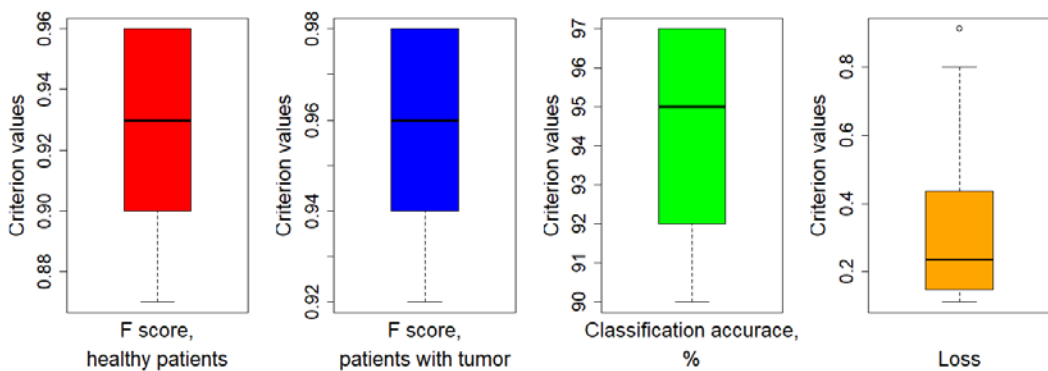


Рис. 4.4. Коробкові діаграми розкиду значень критеріїв якості класифікації об'єктів на основі сформованої підмножини тестових даних експресії генів та значень функції втрат у процесі тестування ЗНМ

Проаналізувавши отримані результати можна зробити висновок щодо високої ефективності застосування одновимірної ЗНМ для класифікації об'єктів на основі даних експресії генів. Значення F-індексу для обох кластерів, точності класифікації об'єктів варіюються у досить високому інтервалі, а значення функції втрат у відносно низькому, враховуючи абсолютні значення експресії генів. При цьому найкраща якість за цією групою критеріїв досягається при застосуванні таких параметрів: розмір фільтру (довжина вектора експресії генів, що подається на вхід мережі) – 100. При цьому повний вектор експресії генів розбивається на 100 рівних частин; густина ущільненого шару – 256; розмір ядра – 8. При застосуванні даних параметрів значення критеріїв якості роботи моделі такі: F-індекс для першого кластера – 0.96; F-індекс для другого кластера – 0.98; точність класифікації об'єктів на тестовій підмножині – 97%, при цьому із 39 об'єктів тільки один був ідентифікований помилково; значення функції втрат – 0.112.

На рис. 4.5 зображені діаграми зміни точності класифікації та значення функції втрат у процесі навчання нейронної мережі на підмножинах даних, що використовувалися для навчання та для валідації моделі на кожному кроці її навчання.

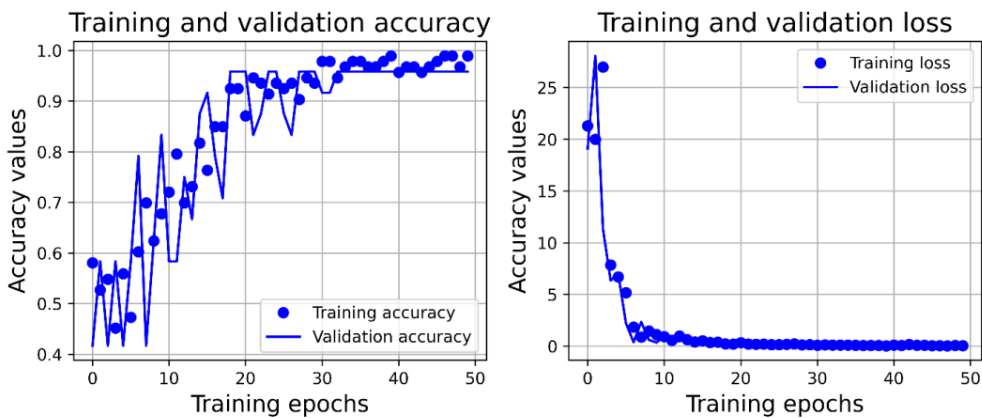


Рис. 4.5. Діаграми зміни точності класифікації та значення функції втрат у процесі навчання одновимірної нейронної мережі

Аналіз отриманих результатів свідчить про коректність навчання мережі. Перенавчання при цьому не спостерігається, оскільки характер зміни критеріїв точності класифікації та функції втрат, отриманих у процесі навчання та валідації моделі, збігається у межах допустимої помилки.

#### 4.1.2.3. Модель одновимірної двошарової згорткової нейронної мережі

На рис. 4.6 зображено структурну блок-схему моделі одновимірної двошарової згорткової нейронної мережі.

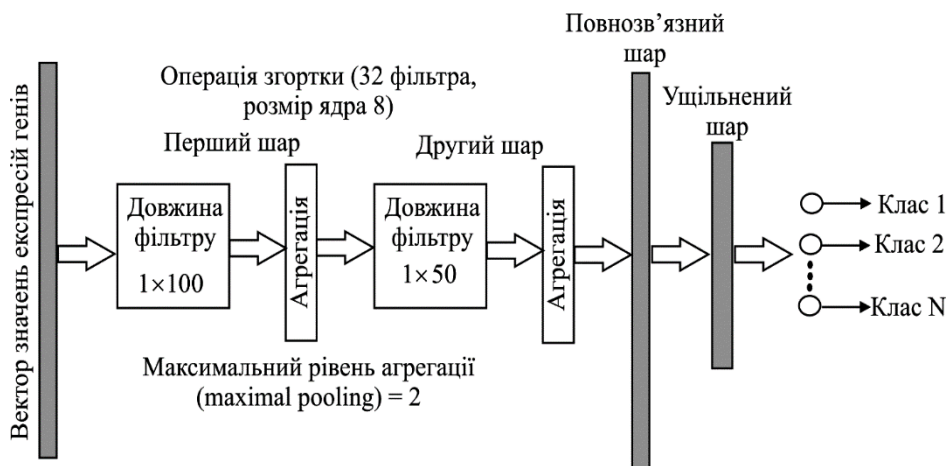


Рис. 4.6. Модель одновимірної двошарової згорткової нейронної мережі

Гіперпараметри мережі встановлено емпіричним шляхом. Результати моделювання показали, що збільшення розміру ядра більше восьми не є доцільним, оскільки значення критеріїв класифікації об'єктів та функції втрат погіршуються зі збільшенням розміру ядра. Довжина фільтра була  $1 \times 100$  для першого шару та  $1 \times 50$  для другого. Густина ущільненого шару змінювалася в інтервалі (64, 128, 256, 512). Результати моделювання показали, що при застосуванні тестової підмножини даних експресії генів в усіх випадках із 39 об'єктів коректно ідентифікуються 38, значення F-індексу для першого і другого кластерів та точності класифікації об'єктів в усіх випадках було однаковим і дорівнювало відповідно 0.96, 0.98 і 0.97. На рис. 4.7 зображено характер розподілу значень функції втрат при різних значеннях густини ущільненого шару.

На основі аналізу отриманих результатів можна зробити висновок, що двошарова мережа більш стабільна, порівняно з одношаровою. До того ж за критерієм функції втрат вона також ефективніша. Оптимальне значення густини щільнісного шару, при якому досягається мінімальне значення функції втрат 0.092, дорівнює 256. На рис. 4.8 зображені графіки зміни точності класифікації та значення функції втрат у процесі навчання нейронної мережі на підмножинах даних, що використовувалися для навчання та для валідації моделі на кожному кроці її навчання. Аналіз отриманих діаграм також свідчить про адекватність навчання нейронної мережі, оскільки характер зміни точності класифікації і функції втрат на підмножинах даних експресії генів, що були застосовані для навчання та валідації моделі збігається у межах допустимого діапазону. Аналіз отриманих результатів також показав, що збільшення кількості шарів не є доцільним, оскільки точність класифікації об'єктів та значення функції втрат при цьому не покращуються, але час навчання мережі підвищується.

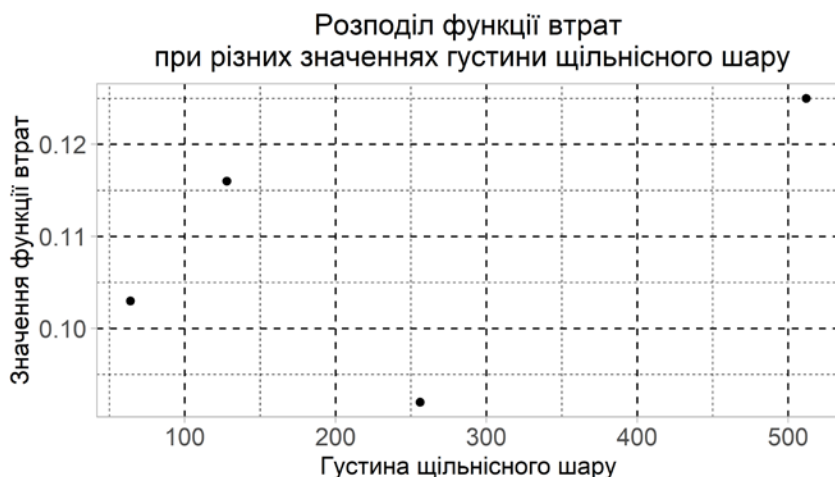


Рис. 4.7. Діаграма розподілу функції втрат при різних значеннях густини щільнісного шару

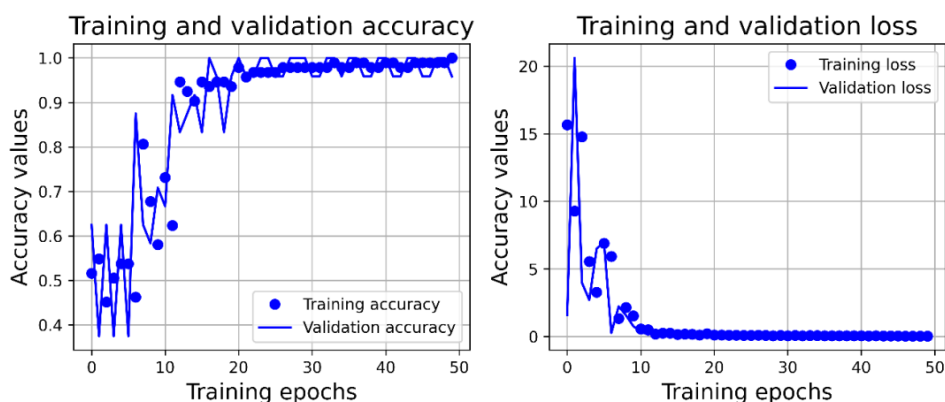


Рис. 4.8. Діаграми зміни точності класифікації та значення функції втрат у процесі навчання одновимірної двошарової нейронної мережі

#### 4.1.2.4. Модель двовимірної згорткової нейронної мережі

Модель двовимірної згорткової мережі, що застосовувалася для класифікації об'єктів на основі сформованої підмножини даних експресії генів зображено на рис. 4.9.

У цьому випадку вхідні дані формувалися як матриця експресії генів розміром  $(100 \times 100)$ . До даних застосовувалися фільтри  $(10 \times 10)$  і  $(20 \times 20)$ , максимальне об'єднання даних у згортковому шарі (Max pooling) при цьому було  $(2 \times 2)$ . Результати моделювання показали, що зменшення даних значень до  $(1 \times 1)$  не є доцільним за критеріями якості, що використовувалися в процесі моделювання. Густина ущільненого шару збільшувалася в інтервалі (64, 128, 256, 512). Результати моделювання наведено у табл. 4.2.



Рис. 4.9. Модель двовимірної згорткової нейронної мережі

Таблиця 4.2

Результати моделювання щодо застосування двовимірної ЗНМ для класифікації об'єктів на основі даних експресії генів (підмножина для тестування)

| Гіперпараметри |                    |                | Результати класифікації |             |             |              |
|----------------|--------------------|----------------|-------------------------|-------------|-------------|--------------|
| Ущільнений шар | Кількість фільтрів | Розмір ядра    | F-індекс                |             | Точність, % | Втрати       |
|                |                    |                | Кл. 1                   | Кл. 2       |             |              |
| 64             | 32                 | (10,10)        | 0.96                    | 0.98        | 97          | 1.24         |
| 64             | 64                 | (10,10)        | 0.96                    | 0.98        | 97          | 1.734        |
| <b>64</b>      | <b>32</b>          | <b>(20,20)</b> | <b>0.96</b>             | <b>0.98</b> | <b>97</b>   | <b>0.688</b> |
| 64             | 64                 | (20,20)        | 0.96                    | 0.98        | 97          | 1.08         |
| 128            | 32                 | (10,10)        | 0.96                    | 0.98        | 97          | 1.94         |
| 128            | 64                 | (10,10)        | 0.96                    | 0.98        | 97          | 2.363        |
| 128            | 32                 | (20,20)        | 0.96                    | 0.98        | 97          | 1.161        |
| 128            | 64                 | (20,20)        | 0.96                    | 0.98        | 97          | 1.412        |
| 256            | 32                 | (10,10)        | 0.96                    | 0.98        | 97          | 3.032        |
| 256            | 64                 | (10,10)        | 0.93                    | 0.96        | 95          | 4.079        |
| 256            | 32                 | (20,20)        | 0.96                    | 0.98        | 97          | 1.318        |
| 256            | 64                 | (20,20)        | 0.96                    | 0.98        | 97          | 1.470        |
| 512            | 32                 | (10,10)        | 0.96                    | 0.98        | 97          | 3.335        |
| 512            | 64                 | (10,10)        | 0.96                    | 0.98        | 97          | 4.898        |
| 512            | 32                 | (20,20)        | 0.96                    | 0.98        | 97          | 1.225        |
| 512            | 64                 | (20,20)        | 0.96                    | 0.98        | 97          | 1.604        |

На основі аналізу отриманих результатів можна зробити висновок, що для класифікації даних експресії генів двовимірна ЗНМ має більш високий рівень стабільності, порівняно з одновимірною мережею. Майже в усіх випадках (окрім одного) критерії якості класифікації є максимальними, тільки один об'єкт ідентифікований некоректно, що можна пояснити похибкою при формуванні даних. До того ж аналіз даних таблиці 4.2 дозволяє також зробити висновок, що зменшення

розміру фільтра не є доцільним за значенням функції втрат. В усіх випадках значення даної функції при застосуванні фільтра ( $10 \times 10$ ) є суттєво більшим, порівняно з фільтром ( $20 \times 20$ ). За значенням функції втрат оптимальними у даному випадку є такі параметри мережі: густина ущільненого шару – 64; кількість фільтрів – 32; розмір ядра –  $20 \times 20$ . На рис. 4.10 наведено результати моделювання у процесі навчання нейронної мережі, аналіз яких також дозволяє зробити висновок щодо коректності моделі.

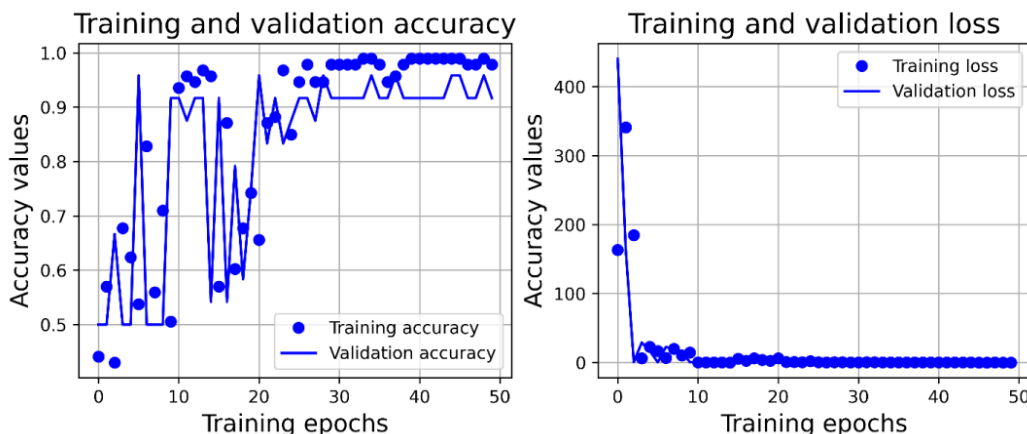


Рис. 4.10. Діаграми зміни точності класифікації та значення функції втрат у процесі навчання двовимірної нейронної мережі

Узгодженість зміни значень точності класифікації об'єктів та функції втрат, розрахованих за результатами навчання та валідації моделі на кожному кроці, свідчить про відсутність процесу перенавчання нейронів мережі.

#### 4.1.2.5. Модель двовимірної тришарової згорткової нейронної мережі

Модель двовимірної тришарової згорткової нейронної мережі, що досліджувалася у процесі моделювання, зображена на рис. 4.11.

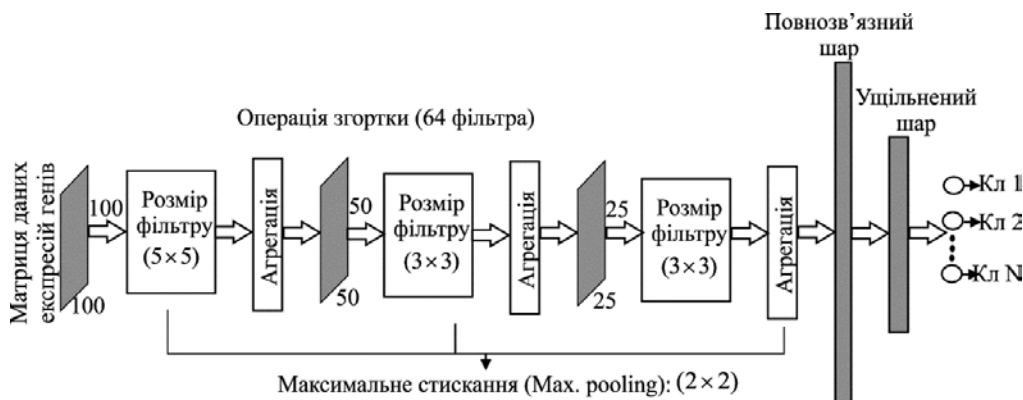


Рис. 4.11. Модель двовимірної тришарової згорткової нейронної мережі

Досліджено такі значення густини ущільненого шару: 64, 128, 256, 512. На основі результатів моделювання зроблено висновок, що значення точності класифікації на тестовій підмножині даних та F-індексу в усіх випадках однакове: ACC = 97%, F-індекс для першого класу 0.96, а для другого – 0.98. При цьому як і в інших випадках із 39 об'єктів 38 ідентифіковано коректно. На рис. 4.12 наведено характер зміни значення точності класифікації об'єктів та функції втрат для навчальної підмножини даних та підмножині, що застосовувалася для валідації моделі, аналіз яких також дозволяє зробити висновок щодо відсутності перенавчання мережі.

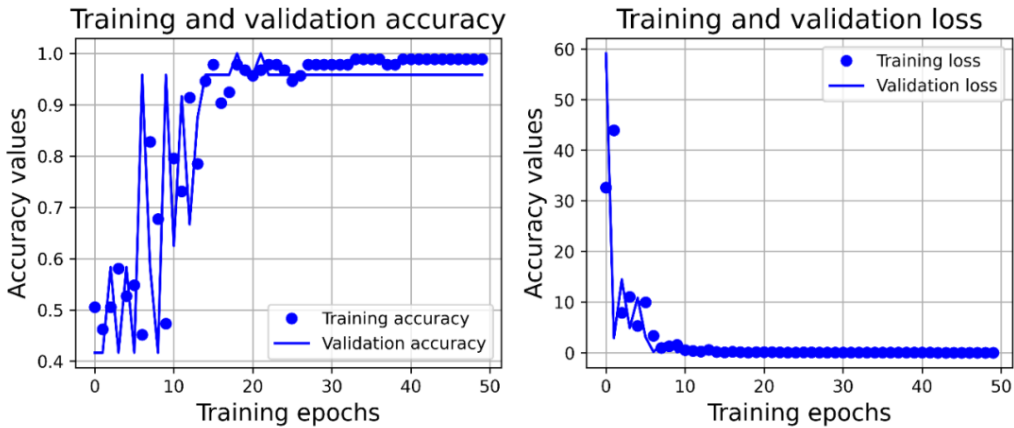


Рис. 4.12. Діаграми зміни точності класифікації та значення функції втрат у процесі навчання двовимірної тришарової нейронної мережі

#### 4.1.3. Оцінка робастності згорткової нейронної мережі при різних рівнях шумової компоненти

Для оцінки робастності згорткової нейронної мережі на дані експресії генів накладався гаусівський «білий шум» з нульовим середнім значенням, стандартне відхилення якого змінювалося у діапазоні: 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3, 3.25, 3.5, 3.75, 4. Цей діапазон був визначений емпіричним шляхом враховуючи значення середнього стандартного відхилення профілів експресії генів (рис. 4.13), що досліджувалися у процесі моделювання. Рівень шумової компоненти змінювався від мінімального (за стандартним відхиленням) до максимального, при якому шумова компонента перекривала корисну компоненту сигналу. При накладанні шумової компоненти на матрицю даних експресії генів від'ємні значення експресії генів (не коректні) замінювалися нулями (цей ген для відповідного об'єкта вважався неекспресованим). У процесі моделювання досліджувалися двошарова одновимірні та тришарова двовимірні згорткові нейронні мережі з оптимальними значеннями гіперпараметрів за результатами досліджень, що наведені у підрозділах 4.1.2.3 і 4.1.2.5. Робастність відповідної моделі оцінювалася за точністю класифікації об'єктів тестової підмножини даних.

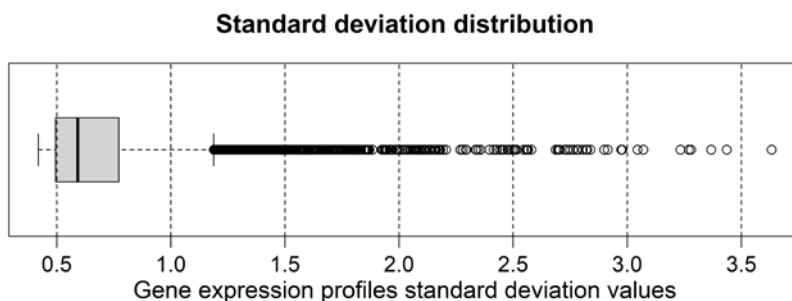


Рис. 4.13. Коробкова діаграма розкиду значень стандартного відхилення експресії генів у профілях, що досліджувалися (10000 профілів)

На рис. 4.14 зображено діаграми залежності точності класифікації об'єктів від рівня зашумленості даних експресії генів (стандартного відхилення шумової компоненти).

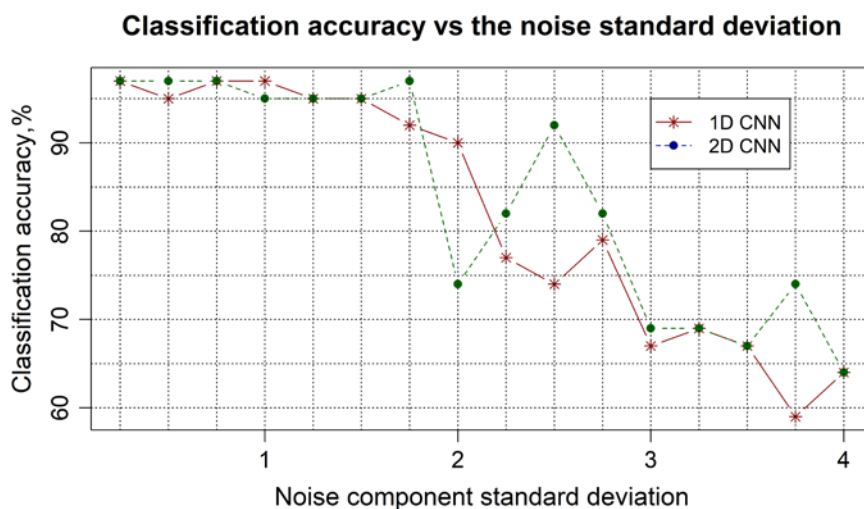


Рис. 4.14. Діаграми залежності точності класифікації об'єктів від рівня шумової складової при застосуванні одновимірної та двовимірної згорткових нейронних мереж

На основі аналізу отриманих результатів можна зробити висновок, що обидва типи нейронних мереж мають високий рівень робастності при невисоких рівнях шумової складової (для даних експресії генів, що досліджувалися, пороговий рівень стійкості мереж відповідав значенню стандартного відхилення шумової складової 1.75). При цьому точність класифікації об'єктів тестової підмножини даних варіювалася у діапазоні від 97% до 95%. У першому випадку із 39 один об'єкт був ідентифікований некоректно, у другому випадку – 2. Цей факт свідчить про те, що моделі класифікації на основі згорткових нейронних мереж не є чутливими до

довжини вектора експресії генів, що характеризують об'єкти, що досліджуються. До того ж результати моделювання дозволяють зробити висновок щодо вищою є ефективність одновимірної згорткової нейронної мережі, порівняно з двовимірною, оскільки при майже однаковій робастності час навчання одновимірної мережі є суттєво меншим.

Отримані результати створюють умови для розробки гібридних моделей формування підмножин взаємно корельованих та диференційно експресованих даних експресії генів на основі комплексного застосування алгоритмів кластеризації профілів експресії генів і згорткової нейронної мережі. Оцінка якості сформованих підмножин даних експресії генів при цьому здійснюється шляхом розрахунку критеріїв якості класифікації відповідних об'єктів, а формування вектора значень експресії генів потрібної довжини здійснюється шляхом доповнення даних профілями з нульовою експресією. Точність отриманих результатів при цьому майже не зміниться.

#### **4.2. Концептуальний опис гібридної моделі формування диференційно експресованих та взаємно корельованих профілів експресії генів на основі комплексного застосування алгоритмів кластеризації та ЗНМ**

Як зазначалося вище, основною метою цього розділу є підвищення ефективності процесу формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів, що створює умови для підвищення точності та об'єктивності класифікації об'єктів, атрибутами яких є значення експресії виділених генів. Застосування етапу кластеризації дозволяє розділити множину профілів експресії генів на підмножини взаємно близьких за відповідними метриками, при цьому дуже важливим є вибір метрики, за якими формуються кластери. У розділі 2 досліджено різні метрики, які орієнтовані на оцінку ступеня близькості високовимірних профілів експресії генів, при цьому, як результат досліджень, для підвищення об'єктивності запропоновано гібридну метрику на основі комплексного застосування метрик на основі оцінки взаємної інформації та критерія узгодженості Пірсона. Ця метрика буде застосована для формування кластерної структури в рамках подальших досліджень. У розділі 3 запропоновано модель формування кластерної структури на основі застосування індуктивних методів аналізу складних систем із використанням внутрішніх, зовнішніх критеріїв якості кластеризації та критерію балансу, аналіз якого дозволяє прийняти остаточне рішення щодо формування кластерної структури. Запропонована модель була апробована із застосуванням щільнісних алгоритмів кластеризації, при цьому вона показала високу ефективність для низьковимірних даних, але для високовимірних профілів експресії генів ця модель показала нижчу ефективність, що можна пояснити тим, що щільнісні алгоритми не орієнтовані на високовимірні дані. До того ж висока об'єктивність прийняття рішень щодо формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів може бути досягнута при застосуванні, окрім розрахунку критеріїв кластеризації профілів експресії генів, етапу класифікації об'єктів, що містять значення експресії генів у відповідних

кластерах як атрибути. Остаточне формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів при цьому відповідає екстремумам як критеріїв кластеризації, так і критеріїв класифікації даних експресії генів. У рамках цієї концепції досліджуються два сучасних алгоритми кластеризації, що орієнтовані на високовимірні дані: алгоритм кластеризації SOTA (Self-Organizing Tree Algorithm) [275] та алгоритм спектральної кластеризації [276]. Етап класифікації даних експресії генів здійснюється шляхом застосування одновимірної двошарової згорткової нейронної мережі, ефективність якої для класифікації цих даних доведена у попередньому підрозділі. На рис. 4.15 зображено структурну блок-схему покрокової процедури реалізації концепції формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів на основі комплексного застосування методів кластеризації та класифікації.



Рис. 4.15. Структурна блок-схема покрокової процедури реалізації концепції формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів

Практична реалізація вищенаведеної концепції передбачає такі етапи:

**Етап I.** Редукція неекспресованих та низькоекспресованих генів для всіх об'єктів, що складають вихідні дані експресії генів. Реалізація цього етапу передбачає видалення генів з нульовою експресією для всіх зразків та неінформативних генів за абсолютним значенням, дисперсією та ентропією Шеннона.

Зазвичай, вихідні дані містять приблизно 45000 генів, половина з яких має нульову експресію для всіх зразків, що досліджуються. Ці гени можуть бути видалені на першому кроці формування даних. На наступному кроці формуються порогові значення критеріїв, за якими гени розділяються на інформативні (для подальшої обробки) та неінформативні (для видалення). Як критерії застосовано абсолютне значення експресії, дисперсія та ентропія Шеннона. Передбачається, що якщо абсолютне значення експресії для усіх зразків та дисперсія відповідного профілю експресії генів менші, а ентропія Шеннона більша ніж відповідні пороги, то цей ген видаляється як неінформативний. Визначення порогових значень визначається емпіричним шляхом, аналізуючи характер розкиду значень критеріїв для усіх генів, враховуючи, що сформована на цьому етапі множина даних експресії генів має містити приблизно 10000 генів. Результати попередніх досліджень, що наведені у підрозділі 4.1, показують, що цієї кількості генів достатньо для отримання високої точності класифікації об'єктів, при цьому підвищення точності і об'єктивності класифікації можливе за рахунок якісної кластеризації профілів експресії генів на попередньому етапі перед класифікацією об'єктів, що досліджуються.

**Етап II.** Кластеризація профілів експресії генів та формування кластерної структури. Реалізація цього етапу передбачає розділення сформованої множини профілів експресії генів на підмножини взаємно близьких профілів за модифікованою гібридною метрикою максимізації взаємної інформації, яка детально описана у розділі 2, із застосуванням принципів індуктивної технології об'єктивної кластеризації, при цьому на кожному етапі розраховуються внутрішні та зовнішні критерії якості кластеризації. Реалізація цього кроку передбачає розділення виділеної підмножини профілів експресії генів на дві еквівалентні підмножини з подальшою паралельною кластеризацією на цих підмножинах даних у діапазоні зміни параметрів алгоритму кластеризації. Остаточне рішення щодо оптимальних параметрів відповідного алгоритму кластеризації приймається на основі аналізу значень критерію балансу, який містить як компоненти внутрішні та зовнішні критерії якості кластеризації. Оптимальні кластерні структури відповідають максимумам критерію балансу. Відповідно до принципу неостаточності рішень (розділ 3), на цьому кроці формуються декілька рішень, що відповідають екстремумам критерію балансу, при цьому формуються декілька відповідних кластерних структур. Остаточне рішення приймається на основі комплексного аналізу критеріїв якості кластеризації профілів експресії генів та класифікації об'єктів, що містять експресії виділених генів як атрибути.

**Етап III.** Класифікація об'єктів, що містять дані експресії генів у виділених кластерах шляхом застосування одновимірної двошарової згорткової нейронної

мережі. На цьому етапі, до даних експресії генів кожного кластера застосовується згорткова нейронна мережа з метою класифікації об'єктів, що досліджуються. Реалізація цього етапу потребує передобробки векторів значень експресії генів з метою коректного налаштування фільтрів. Вектори даних експресії генів доповнюються генами з нульовою експресією так, щоб довжина вектора експресії генів ділилася без залишку на довжину фільтра. Результатом реалізації етапу класифікації є розрахунок критеріїв якості класифікації об'єктів на даних експресії генів у кожному кластері. У рамках запропонованої концепції розраховується точність класифікації на тестовій підмножині даних.

**Етап IV.** Аналіз отриманих результатів. Формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів. Реалізація цього етапу передбачає комплексний аналіз як критеріїв кластеризації профілів експресії генів, так і критеріїв класифікації об'єктів, що містять дані експресії генів у відповідних кластерах, при цьому виділяються підмножини даних експресії генів, які відповідають максимальним значенням як критеріїв кластеризації, так і критеріїв класифікації. Слід зазначити, що реалізація цього етапу передбачає формування декількох підмножин диференційно експресованих та взаємно корельованих профілів експресії генів за максимальними значеннями відповідних критеріїв. Цей факт створює умови для підвищення об'єктивності класифікації відповідних об'єктів за рахунок узгодженого прийняття рішення щодо ідентифікації об'єкта, враховуючи результат класифікації цього об'єкта на основі даних експресії генів усіх виділених кластерів.

#### **4.2.1. Гібридна індуктивна модель кластеризації профілів експресії генів на основі алгоритму кластеризації SOTA**

Самоорганізуючий алгоритм кластеризації SOTA (Self-Organizing Tree Algorithm) [266] є логічним продовженням самоорганізуючих нейронних мереж на основі карт Кохонена [267–269]. Результатом застосування алгоритму SOTA є бінарне топологічне дерево, що формується відповідно до алгоритму вирощування просторової клітинної структури Fritzke [270]. Застосування алгоритму Fritzke призводить до збільшення кількості вузлів дерева в ділянці більшої щільності об'єктів (профілів експресії генів), при цьому концентрація об'єктів у області з меншою щільністю не змінюється. Процес формування клітинної структури дерева, що лежить в основі алгоритму кластеризації SOTA, проілюстровано на рис. 4.16. як вхідні дані, розглядається множина профілів експресії генів.

Отже, у початковому стані структура системи має вигляд бінарного дерева, що складається з трьох вузлів: дві клітини об'єднані через зовнішній материнський вузол (рис. 4.16а). У початковому стані кожна клітина та вузол характеризується вектором, довжина якого дорівнює довжині профілю експресії гена (вектор даних, що подається на вхід мережі), а значення лежать у інтервалі варіювання відповідних значень експресії генів вхідних даних.

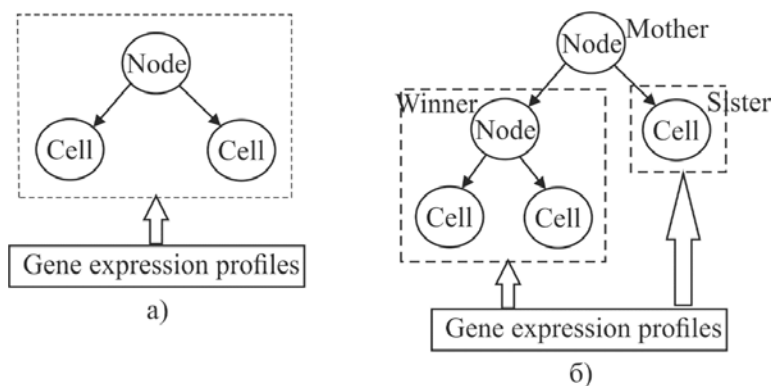


Рис. 4.16. Ілюстрація процесу формування дерева (формування клітинної структури) в процесі роботи алгоритму кластеризації SOTA: а) початковий стан системи; б) стан системи після одного циклу

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 експресії генів, що подаються на вхід системи, а значення лежать у інтервалі варіювання експресії відповідних генів. На даному етапі задаються також параметри для корекції ваг клітин і вузлів у процесі подавання на вхід мережі профілів експресії генів, відповідно до умови:  $\alpha_w > \alpha_p > \alpha_s$ , де  $\alpha_w$ ,  $\alpha_p$  та  $\alpha_s$  – параметри для корекції ваг клітини переможця (winner), кореневого (parent) вузла та сусідньої клітини (sister), відповідно. Авторами алгоритму SOTA [275] запропоновано 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, було прийнято авторський варіант співвідношень цих параметрів як найоптимальніший. 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.

**Етап II. Адаптація.** На цьому етапі на вхід усіх зовнішніх клітин поступово подаються профілі експресії генів, що складають вхідні дані. Розраховується ступінь близькості між профілями експресії генів і вектором ваг відповідної клітини. На цьому кроці виділяється клітина переможець (winner) за мінімальною відстанню між профілем експресії гену та вектором ваг і, відповідно до принципу «переможець забирає усе», профіль експресії гену поміщається у клітину-переможець. На наступному кроці підлаштовуються ваги клітини переможця, сусідньої клітини та кореневого вузла відповідно до формули:

$$W_i(\tau + 1) = W_i(\tau) + \eta \cdot (P_j - W_i(\tau)), \quad (4.11)$$

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

$$\eta_\tau = \alpha_i \cdot \frac{1-\tau}{N} \cdot (1-bt), \quad (4.12)$$

де  $\eta_\tau$  – значення  $\eta$ -параметра на ітерації  $\tau$ ;  $\alpha_i$  – параметр для корекції ваг  $i$ -ї клітини;  $N$  – максимальна кількість профілів експресії генів, що подаються на вхід мережі;  $t$  – максимальна кількість операцій в одному циклі;  $b$  – коефіцієнт, що визначає швидкість зміни параметра  $\eta$ .

Підлаштування ваг клітин та кореневого вузла здійснюється відповідно до 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 III.** Формування структури дерева (мережі). На першому кроці цього етапу розраховується коефіцієнт варіації профілів експресії генів у кожній клітині:

$$R_i = \sum_{j=1}^K d(P_j, W_i) / K, \quad (4.13)$$

де  $W_i$  – вектор ваг у  $i$ -й клітині;  $P_j$  –  $j$ -й профіль експресії генів;  $K$  – кількість профілів експресії генів у  $i$ -й клітині.

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

$$\varepsilon = \sum_{i=1}^p R_i. \quad (4.14)$$

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

$$E = \frac{|\varepsilon_\tau - \varepsilon_{\tau-1}|}{\varepsilon_{\tau-1}} \leq E_{zp}, \quad (4.15)$$

де  $E_{zp}$  – граничне значення відносної зміни коефіцієнта варіації.

Якщо виконується умова (4.15), алгоритм зупиняється з фіксацією структури дерева і відповідних кластерів, що містять профілі експресії генів у зовнішніх клітинах. У протилежному випадку (умова (4.15) не виконується) здійснюється зростання дерева. Клітина, що має найбільше значення коефіцієнта варіації, ділиться на дві частини і стає вузлом (рис. 4.16б), при цьому усім клітинам і вузлу присвоюються ідентичні значення ваг. Далі повторюються етапи II і III. Умовою зупинки алгоритму є або виконання умови (4.15), або досягнення максимальної кількості ітерацій, що задається на етапі ініціалізації початкових параметрів алгоритму.

На основі аналізу вищенаведеного процесу можна зробити висновок, що результат зробити 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 відповідно до умови:  $\alpha_w = 2\alpha_p$ ,  $\alpha_p = 5\alpha_s$ ). В рамках дисертаційного дослідження, граничне значення відносної зміни коефіцієнта варіації було прийнято за 0. За цією умовою зупинка алгоритму визначалась або повторюванням характеру розподілу профілів експресії генів у кластери на двох послідовних ітераціях, або досягненням граничної кількості ітерацій. Визначення оптимального значення параметра для корекції ваг сусідньої клітини  $\alpha_s$  здійснювалося із застосуванням індуктивних методів аналізу складних систем (індуктивної технології об'єктивної кластеризації). Блок-схема гібридної індуктивної моделі формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів на основі алгоритму кластеризації SOTA зображена на рис. 4.17.

Реалізація алгоритму передбачає такі етапи:

**Етап I. Формування даних та ініціалізація алгоритму кластеризації SOTA.**

1.1. Формування матриці профілів експресії генів, де рядки є об'єктами, що досліджуються, а стовпці – генами, значення експресії яких визначають стан відповідного об'єкта.

1.2. Видалення неінформативних генів за критеріями: абсолютне значення експресії гена, дисперсія та ентропія Шеннона профілів експресії генів відповідно до методики, описаної у підрозділі 4.1.2.

1.3. Формування метрики оцінки ступеня близькості профілів експресії генів і відповідних кластерів. Відповідно до результатів досліджень, наведених у розділі 2, 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 експресії генів.

1.5. Формування критеріїв якості кластеризації для оцінки якості кластерної структури. Враховуючи результати досліджень, що наведені у розділі 3, як внутрішні критерії для оцінки характеру групування профілів експресії генів у кластерах, були використані WB-індекс та PBM – критерій, зовнішні критерії розраховувався як нормалізована різниця відповідних внутрішніх критеріїв, а критерій балансу розраховувався відповідно до алгоритму, наведену у підрозділі 3.2.2.3.

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

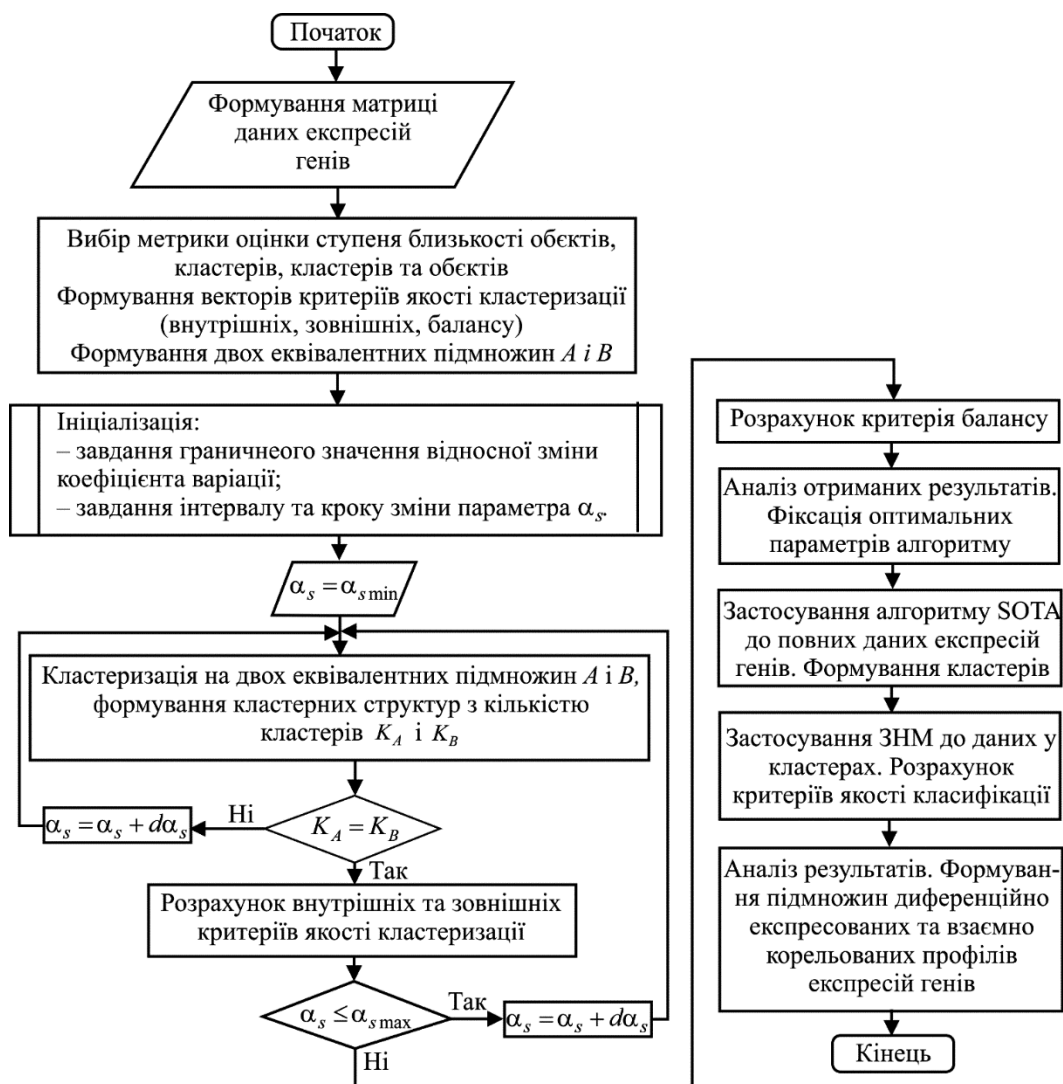


Рис. 4.17. Блок-схема гібридної індуктивної моделі формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів на основі алгоритму кластеризації SOTA

**Етап II.** Кластеризація профілів експресії генів та визначення оптимальних параметрів алгоритму 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 експресії генів, що містяться у еквівалентних підмножинах A і B. Формування кластерів.

2.3. Якщо кількість кластерів, що сформовані на еквівалентних підмножинах профілів експресії генів, 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 відбувається без розрахунку критеріїв.

2.4. Якщо  $\alpha_s \leq \alpha_{s_{\max}}$ , перехід на крок 2.2 цієї процедури, в іншому випадку – розрахунок критерію балансу.

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

2.6. Застосування алгоритму кластеризації SOTA з визначеними оптимальними значеннями параметра для корекції ваг сусідньої клітини до повної множини профілів експресії генів, що сформована на кроці 1.2 цієї процедури. Формування кластерної структури.

**Етап III.** Застосування згорткової нейронної мережі до даних експресії генів у сформованих кластерах.

3.1. Передобробка даних експресії 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 експресії генів у виділених кластерах. Розрахунок критеріїв якості класифікації даних.

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

#### **4.2.2. Практична реалізація гібридної індуктивної моделі кластеризації профілів експресії генів на основі алгоритму кластеризації SOTA**

Результати моделювання щодо визначення оптимального значення параметра для корекції ваг сусідньої клітини  $\alpha_s$  зображено на рис. 4.18.

Значення параметра  $\alpha_s$  при цьому змінювалося в інтервалі від 0.0001 до 0.01 з кроком 0.0002. Моделювання здійснювалося у програмному середовищі R із застосуванням функцій пакету *clValid* [271]. Функція *sota()* передбачає можливість реалізації алгоритму SOTA із застосуванням евклідової та кореляційної метрик. Враховуючи високу розмірність профілів експресії 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.

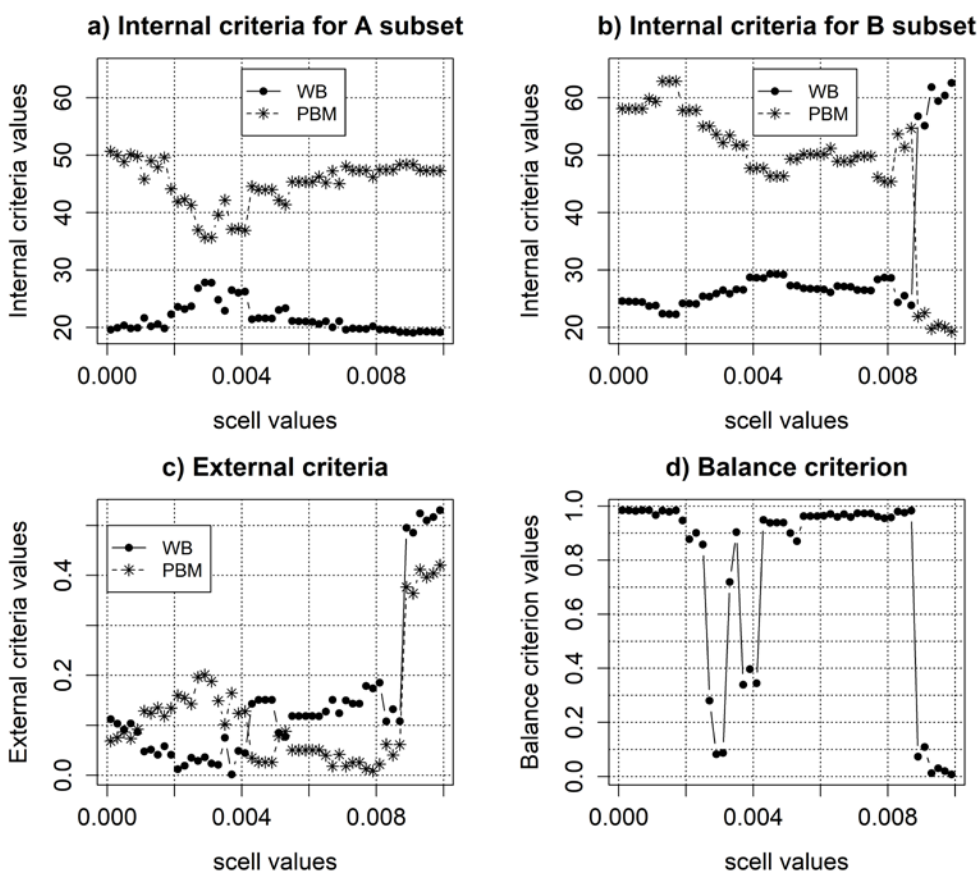


Рис. 4.18. Результати моделювання щодо визначення оптимальних параметрів алгоритму кластеризації SOTA на основі застосування індуктивних методів аналізу складних систем

Відповідно до першої частини алгоритму, структурну блок-схему якого зображено на рис. 4.17, на першому етапі виділена підмножина профілів експресії генів (1000 профілів) розділялася на дві еквівалентні підмножини A і B, що містили однакову кількість попарно близьких об'єктів (за модифікованою гібридною метрикою). Далі, на кожному кроці реалізації алгоритму SOTA на еквівалентних підмножинах профілів експресії генів розраховувалися внутрішні критерії якості кластеризації. На рис. 4.18a,b зображені діаграми залежності значень цих критеріїв від параметра алгоритму SOTA  $\alpha_s$ , до того ж за цими критеріями, оптимальна кластеризація відповідає мінімальному значенню WB-індексу і максимальному значенню PBM – критерію. За отриманими діаграмами, значення цих критеріїв, розрахованих на підмножинах A і B, у деякій мірі суперечать один одному. На рис. 4.18c зображені діаграми зовнішніх критеріїв, розрахованих із застосуванням відповідних внутрішніх критеріїв. Аналіз цих діаграм також не дозволяє однозначно

визначити оптимальне значення параметра  $\alpha_s$  (за мінімальним значенням). На рис. 4.18d зображено діаграму залежності критерію балансу від значення параметра алгоритму, аналіз якої дозволяє виділити інтервали варіювання значень параметра  $\alpha_s$ , що відповідають максимумам критерія, з одного боку, і стійкій кластеризації, з іншого боку. Детальніший 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 експресії генів у кластери є досить висока. Слід 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. Таким чином, значення  $\alpha_s = 0.0007$  було обрано як оптимальне для подальших досліджень.

На другому етапі реалізації алгоритму, зображеному на рис. 4.17, алгоритм кластеризації SOTA з визначеними параметрами застосовувався до повного набору профілів експресії генів (10000), у результаті чого отримано два кластери: 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 ( $50 \times 121$ ) на першому згортковому шарі і ( $25 \times 242$ ) – на другому. Другий кластер доповнювався до загальної кількості 4000 профілів, при цьому фільтри мали розміри ( $50 \times 80$ ) на першому шарі і ( $25 \times 160$ ) – на другому. Результати моделювання щодо застосування згорткових нейронних мереж до даних експресії генів, що містяться у сформованих кластерах, зображені на рис. 4.19 і 4.20. Точність класифікації об'єктів із застосуванням відповідних тестових підмножин даних становила при цьому 95% для першого кластера і 97% – для об'єктів другого кластера, при цьому із 39 об'єктів було коректно ідентифіковано 37 і 38 у першому та у другому випадках, відповідно.

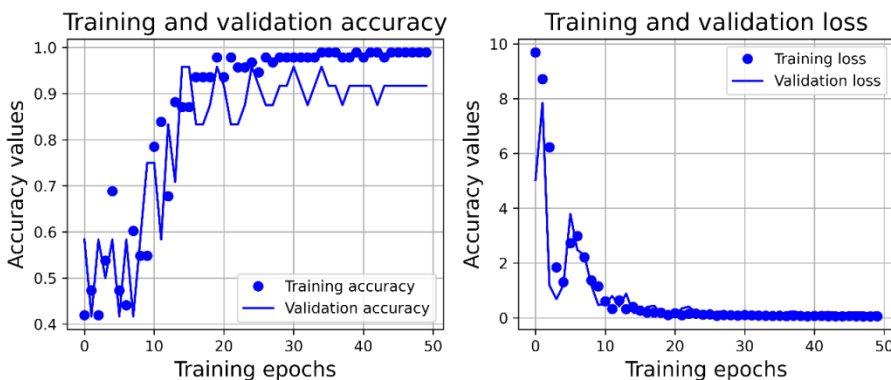


Рис. 4.19. Діаграми зміни точності класифікації та значення функції втрат у процесі навчання одновимірної двошарової нейронної мережі на даних у першому кластері

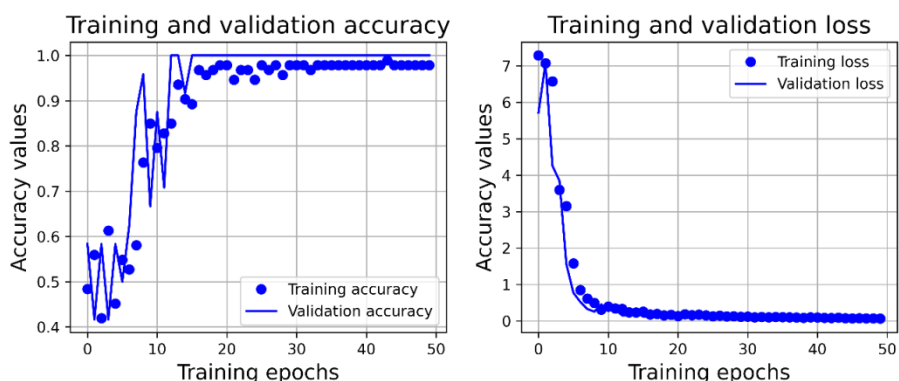


Рис. 4.20. Діаграми зміни точності класифікації та значення функції втрат у процесі навчання одновимірної двошарової нейронної мережі на даних у другому кластері

На наступному етапі до профілів експресії 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 наведено у табл. 4.3.

Таблиця 4.3

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

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

На рис. 4.21 зображено дендрограму розподілу початкової кількості профілів експресії генів на кластери з відповідною точністю класифікації об'єктів. На основі аналізу отриманих результатів можна зробити висновок, що комплексне застосування алгоритму кластеризації SOTA і згорткової нейронної мережі дозволяє виділити кластери профілів експресії генів, які дозволяють з високою точністю ідентифікувати об'єкти, що містять як атрибути значення експресії генів, які локалізовані у відповідних кластерах.

Таким чином, на першому етапі 10000 профілів експресії генів були розділені на два кластери, 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 слід зазначити, що подальше розділення профілів експресії генів, що містилися у меншому кластері на підмножини, що складали 1934 та 2046 профілів експресії генів, погіршило точність класифікації, тобто розділення меншого кластера на підкластери у цьому випадку не доцільно.

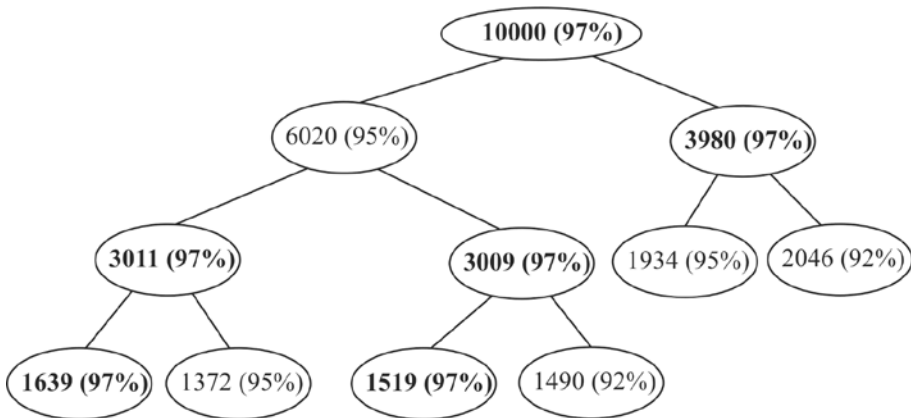


Рис. 4.21. Дендрограма розподілу профілів експресії генів у кластери з точністю класифікації об'єктів, що містять як атрибути значення експресії генів, які містяться у відповідних кластерах

До іншого висновку можна прийти у випадку розділення більшого кластера (6020 профілів експресії генів) на підкластери. Розділення цього кластера на дві підмножини (3011 і 3009 профілів експресії генів) підвищує точність класифікації об'єктів, що містять як атрибути значення експресії генів цих кластерів. Подальше їхнє розділення на більш дрібні підмножини дозволяє на кожній гілці виділити підмножину більш інформативних за критеріями класифікації профілів експресії генів.

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

#### 4.2.3. Практична реалізація гібридної індуктивної моделі кластеризації профілів експресії генів на основі алгоритму спектральної кластеризації

Алгоритм спектральної кластеризації є одним із сучасних алгоритмів, який дозволяє ідентифікувати кластери довільної форми на основі застосування матриць подібності об'єктів, що досліджуються [272–275]. Порівнянні з традиційними алгоритмами, такими як  $k$ -середніх, ієрархічних агломеративних та дивізійних тощо, алгоритм спектральної кластеризації має багато фундаментальних переваг. Результати, що отримуються шляхом застосування алгоритму спектральної кластеризації, часто перевершують за якістю результати, що отримуються шляхом застосування традиційних підходів, при цьому реалізація алгоритму спектральної кластеризації дуже проста та може бути ефективно реалізована на основі стандартних методів лінійної алгебри. Основною відмінністю алгоритму спектральної кластеризації від більшості традиційних алгоритмів є те, що формування кластерів відбувається не на основі абсолютного розташування об'єктів у просторі, а на основі аналізу рівня спорідненості цих об'єктів, що особливо корисне при формуванні кластерів складної форми. У загальному випадку реалізація алгоритму спектральної кластеризації передбачає такі етапи:

1. *Створення графу подібності.* Граф подібності  $G=(V, E)$  являє собою неорієнтований граф, що складається з множини вузлів  $V = \{v_1, v_n\}$  (об'єкти, що досліджуються) та множини дуг (ребра)  $E = s_{ij}$ , що з'єднують вузли  $i$  та  $j$ , відповідно, і визначають міру близькості відповідних вузлів. Вважається, що дві вершини з'єднані між собою, якщо значення подібності  $s_{ij}$  між відповідними об'єктами (вузлами графа) більше за певний поріг (трешолдінг), а ребро при цьому має відповідну вагу  $w_{ij}$ . Проблема кластеризації у цьому випадку може бути формалізована так: необхідно сформувати структуру графа так, щоб ребра між різними групами (кластерами) мали дуже малу вагу (що означає, що об'єкти різних кластерах максимально відрізняються один від одного), а ребра між вузлами всередині групи матимуть великі значення ваг (це означає, що об'єкти в межах одного кластера матимуть якомога більший рівень подібності між собою). Формування графу подібності передбачає розрахунок матриці подібності  $W = (w_{ij})_{i,j=1,n}$  із застосування відповідної метрики близькості враховуючи характер об'єктів, що досліджуються. У випадку кластеризації профілів експресії генів як міра близькості профілів буде використана гібридна модифікована метрика на основі критеріїв максимізації взаємної інформації та узгодженості Пірсона. Таким чином, можна зробити висновок, що граф подібності є неорієнтованим зваженим графом, сила зв'язку між відповідними вузлами якого визначається вагою дуги, що їх з'єднує. Ступень вузла  $v_i \in V$  визначається як сума ваг дуг, що з'єднують цей вузол із сусідніми:

$$d_i = \sum_{j=1}^m w_{ij}, \quad (4.16)$$

де  $m$  – визначає кількість вузлів, що напряму з'єднані з  $i$ -м вузлом.

Слід зазначити, умова  $w_{ij} = 0$  означає, що вузли  $v_i$  і  $v_j$  напряму не зв'язані між собою. На основі значень ступенів вузлів формується матриця ступенів  $D$ , яка являє собою діагональну матрицю зі ступенями вузлів  $d_1, \dots, d_m$ , значення яких розташовані на головній діагоналі. Побудова графу подібності створює умови для формування кластерів шляхом ініціалізації значення коефіцієнта трешолдінгу, що обмежує кількість зв'язків із ненульовою вагою. При цьому вважається, що у підмножині об'єктів  $A$  всі компоненти зв'язані, якщо ваги прямих або непрямих зв'язків між усіма вузлами, що належать підмножині  $A$ , більше ніж нуль і ваги між вузлами підмножини  $A$  і вузлами, що містяться у інших підмножинах даних, дорівнюють нулю.

Залежно від способу трансформації множини об'єктів з відповідною матрицею подібності між усіма парами об'єктів у граф подібності можна виділити такі типи графів:

- Граф на основі  $\epsilon$ -околиці ( $\epsilon$ -neighborhood). Цей тип графу передбачає з'єднання всіх точок (ідентифікатори об'єктів), попарні відстані між якими менші за апріорі визначену  $\epsilon$ -околицю. Оскільки відстані між усіма парами об'єктів визначаються у однаковому масштабі (не більше  $\epsilon$ -околиці), формування дуг графу здійснюється на основі значень матриці подібності, тобто цей тип графу зазвичай ідентифікується як незважений, та не потребує застосування додаткової інформації щодо сили (ваги) відповідних зв'язків.
- Граф на основі  $k$ -найближчих сусідів ( $k$ -nearest neighbors). У цьому випадку передбачається, що вузол  $v_i$  з'єднується з вузлом  $v_j$ , якщо  $v_j$  знаходиться серед  $k$ -найближчих сусідів вузла  $v_i$ . Ініціалізація ваги дуг здійснюється відповідно до значень матриці подібності, тобто граф на основі  $k$ -найближчих сусідів є зваженим.
- Повнозв'язаний граф. Цей тип графу передбачає з'єднання всіх вузлів з позитивним значенням ступеня зв'язку на основі аналізу матриці подібності. Локальні  $\epsilon$ -околиці при цьому формуються шляхом застосування відповідних функцій подібності. Наприклад, у випадку застосування гаусової функції подібності:

$$s(x_i, x_j) = \exp \left( -\frac{\|x_i - x_j\|^2}{2\sigma^2} \right),$$

де параметр  $\sigma$  визначає ширину  $\epsilon$ -околиці, на основі якої формуються зв'язки між відповідними вузлами.

2. *Формування матриці та графу Лапласа і розрахунок власних векторів матриці.* Граф Лапласа, що створений на основі матриці Лапласа, є основним інструментом при реалізації алгоритму спектральної кластеризації. Вважатимемо, що граф  $G$  є неорієнтованим, зваженим з ваговою матрицею  $W$ , при цьому  $w_{ij} = w_{ji} \geq 0$ , і власні вектори матриці подібності можуть бути як нормовані, так і ненормовані. Власні значення матриці  $W$  розташовуються за зростанням, при цьому

перші  $k$  власних векторів відповідають першим  $k$  найменшим власним числам. Матриця Лапласа може розраховуватися із застосуванням нормалізованих та ненормалізованих значень.

Ненормалізована матриця Лапласа, на основі якої будується граф Лапласа, визначається так:

$$L = D - W. \quad (4.17)$$

При застосуванні перших  $k$  – компонентів зв'язку (перших  $k$  власних векторів або власних чисел) без втрати загальності будемо вважати, що вершини графу впорядковані відповідно до компонентів зв'язку, яким вони відповідають. У цьому випадку матриця подібності  $W$  має блочну діагональну форму. Такий самий висновок можна зробити про матрицю Лапласа:

$$L = \begin{pmatrix} L_1 & & & \\ & L_2 & & \\ & & \ddots & \\ & & & L_k \end{pmatrix}. \quad (4.18)$$

Слід зазначити, що кожен з блоків  $L_i$  матриці (4.19) є власним лапласіаном графу, що відповідає підграфу  $i$ -ї зв'язкової компоненти. Повний спектр матриці  $L$  при цьому формується шляхом об'єднання спектрів під графів  $L_i$ , а власні вектори матриці  $L$  являють собою комбінацію власних векторів матриць  $L_i$ , значення яких заповнюються нулями у позиціях, що не відповідають  $i$ -му блоку. Оскільки кожна компонента матриці  $L$  є власним лапласіаном зв'язаного графу, то кожен  $L_i$  має власне значення 0 з кратністю 1, а відповідний цьому власному числу власний вектор є сталим одиничним вектором на  $i$ -ї й зв'язуючій компоненті. Таким чином, у матриці  $L$  кількість власних нульових значень дорівнює кількості зв'язуючих компонент, а відповідні власні вектори є індикаторами наявності зв'язаних компонент.

Нормалізований граф Лапласа може бути сформованим із застосування двох типів матриць, які тісно пов'язані одна з одною і визначаються таким чином:

$$\begin{aligned} L_{sym} &= D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2} \\ L_{rw} &= D^{-1} L = I - D^{-1} W, \end{aligned} \quad (4.19)$$

де перший тип матриці  $L_{sym}$  є симетричною матрицею Лапласа, а друга  $L_{rw}$  (від словосполучення random walk) називається матрицею випадкового блукання.

3. Формування кластерної структури шляхом застосування методу  $k$  найближчих сусідів.

#### 4.2.3.1. Покрокова процедура реалізації алгоритму спектральної кластеризації

Припустимо, що експериментальні дані складаються з  $n$  об'єктів (точок у  $m$  - вимірному просторі), при цьому відстань між усіма парами точок визначається

матрицею подібності  $W = (w_{ij})_{i,j=1,n}$ . Залежно від типу формування графу подібності та методу розрахунку матриці Лапласа можна виділити декілька типів покрокової процедури, що лежить в основі алгоритму спектральної кластеризації.

1. Алгоритм спектральної кластеризації на основі ненормалізованої матриці Лапласа.

*Вхідні дані:* Матриця подібності  $W \in R^{n \times n}$ , кількість кластерів  $k$ .

- Формування графу подібності одним із описаних вище способів із застосуванням значень матриці  $W$  для ініціалізації ваги відповідних зв'язків.
- Розрахунок ненормалізованої матриці Лапласа  $L$ .
- Розрахунок перших  $k$  власних векторів матриці  $L$ :  $u_1, \dots, u_k$ . Формування матриці  $U \in R^{n \times k}$ , що містить, як стовпці, власні вектори  $u_1, \dots, u_k$ .
- Для  $i = 1, \dots, n$ , виділення вектора  $y_i \in R^k$ , що відповідає  $i$ -му рядку матриці  $U$ .
- Кластеризація точок, що відповідають значенням вектора  $y_i \in R^k$ , шляхом застосування алгоритму  $k$ -means. Формування кластерів  $C_1, \dots, C_k$ .

*Вихідні дані:* Кластери  $A_1, \dots, A_k$ , де  $A_i = \{j \mid y_j \in C_i\}$ .

2. Алгоритм спектральної кластеризації на основі нормалізованої матриці Лапласа із застосуванням методу «Shi and Malik».

*Вхідні дані:* Матриця подібності  $W \in R^{n \times n}$ , кількість кластерів  $k$ .

- Формування графу подібності одним із описаних вище способів із застосуванням значень матриці  $W$  для ініціалізації ваги відповідних зв'язків.
- Розрахунок нормалізованої матриці Лапласа  $L$ .
- Розрахунок перших  $k$  власних векторів матриці  $L$ :  $u_1, \dots, u_k$  відповідно до рівності:  $Lu = \lambda Du$ , де  $\lambda$  – власне число, що відповідає власному вектору  $u$ . Формування матриці  $U \in R^{n \times k}$ , що містить, як стовпці, власні вектори  $u_1, \dots, u_k$ .
- Для  $i = 1, \dots, n$ , виділення вектора  $y_i \in R^k$ , що відповідає  $i$ -му рядку матриці  $U$ .
- Кластеризація точок, що відповідають значенням вектору  $y_i \in R^k$  шляхом застосування алгоритму  $k$ -means. Формування кластерів  $C_1, \dots, C_k$ .

*Вихідні дані:* Кластери  $A_1, \dots, A_k$ , де  $A_i = \{j \mid y_j \in C_i\}$ .

3. Алгоритм спектральної кластеризації на основі нормалізованої матриці Лапласа із застосуванням методу «Ng, Jordan and Weiss».

*Вхідні дані:* Матриця подібності  $W \in R^{n \times n}$ , кількість кластерів  $k$ .

- Формування графу подібності одним із описаних вище способів із застосуванням значень матриці  $W$  для ініціалізації ваги відповідних зв'язків.
- Розрахунок нормалізованої матриці Лапласа  $L_{sym}$ .
- Розрахунок перших  $k$  власних векторів матриці  $L_{sym}$ :  $u_1, \dots, u_k$ . Формування матриці  $U \in R^{n \times k}$ , що містить, як стовпці, власні вектори  $u_1, \dots, u_k$ .

— Формування матриці  $T \in R^{n \times k}$  шляхом нормалізації рядків матриці  $U$  відповідно до рівняння:

$$t_{ij} = u_{ij} / \left( \sum_{j=1}^k u_{ij}^2 \right)^{1/2}.$$

— Для  $i = 1, \dots, n$ , виділення вектора  $y_i \in R^k$ , що відповідає  $i$ -му рядку матриці  $T$ .

— Кластеризація точок, що відповідають значенням вектору  $y_i \in R^k$  шляхом застосування алгоритму  $k$ -means. Формування кластерів  $C_1, \dots, C_k$ .

*Вихідні дані:* Кластери  $A_1, \dots, A_k$ , де  $A_i = \{j \mid y_j \in C_i\}$ .

Слід зазначити, що в усіх випадках результат роботи алгоритму визначається методом, що використовується для формування матриці подібності (методом оцінки ступеня близькості об'єктів) і кількістю бажаних кластерів. Однак у багатьох випадках кількість кластерів заздалегідь визначити неможливо, тому виникає необхідність перебору різних кластеризацій із розрахунком кількісних критеріїв якості кластеризації. Вибір метрики близькості об'єктів також залежить від типу даних. Для профілів експресії генів, що досліджуються у цій дисертаційній роботі, застосовується модифікована гібридна метрика на основі комплексного застосування модифікованого критерія максимізації взаємної інформації та критерія узгодженості Пірсона. Кількість кластерів визначається із застосуванням методів індуктивної технології об'єктивної кластеризації.

#### 4.2.3.2. Гібридна індуктивна модель кластеризації профілів експресії генів на основі алгоритму спектральної кластеризації

Блок-схему покрокової процедури реалізації гібридної індуктивної моделі алгоритму спектральної кластеризації зображено на рис. 4.22. Практична реалізація алгоритму передбачає наявність таких етапів:

**Етап I.** Формування даних та ініціалізація параметрів моделі.

1.1. Формування матриці експресії генів  $\{e_{ij}\}, i = \overline{1, n}, j = \overline{1, m}$ , де  $n$  – кількість об'єктів, що досліджуються, а  $m$  – кількість генів.

1.2. Формування метрики оцінки подібності профілів експресії генів.

1.3. Формування векторів та функцій розрахунку критеріїв якості кластеризації профілів експресії генів: внутрішніх, зовнішніх та балансу.

1.4. Розділення множини профілів експресії генів на дві еквівалентні підмножини  $A$  і  $B$ .

1.5. Розрахунок матриць подібності для профілів експресії генів, що містяться у еквівалентних підмножинах.

1.6. Ініціалізація інтервалу зміни кількості кластерів  $k_{min}, k_{max}$ .

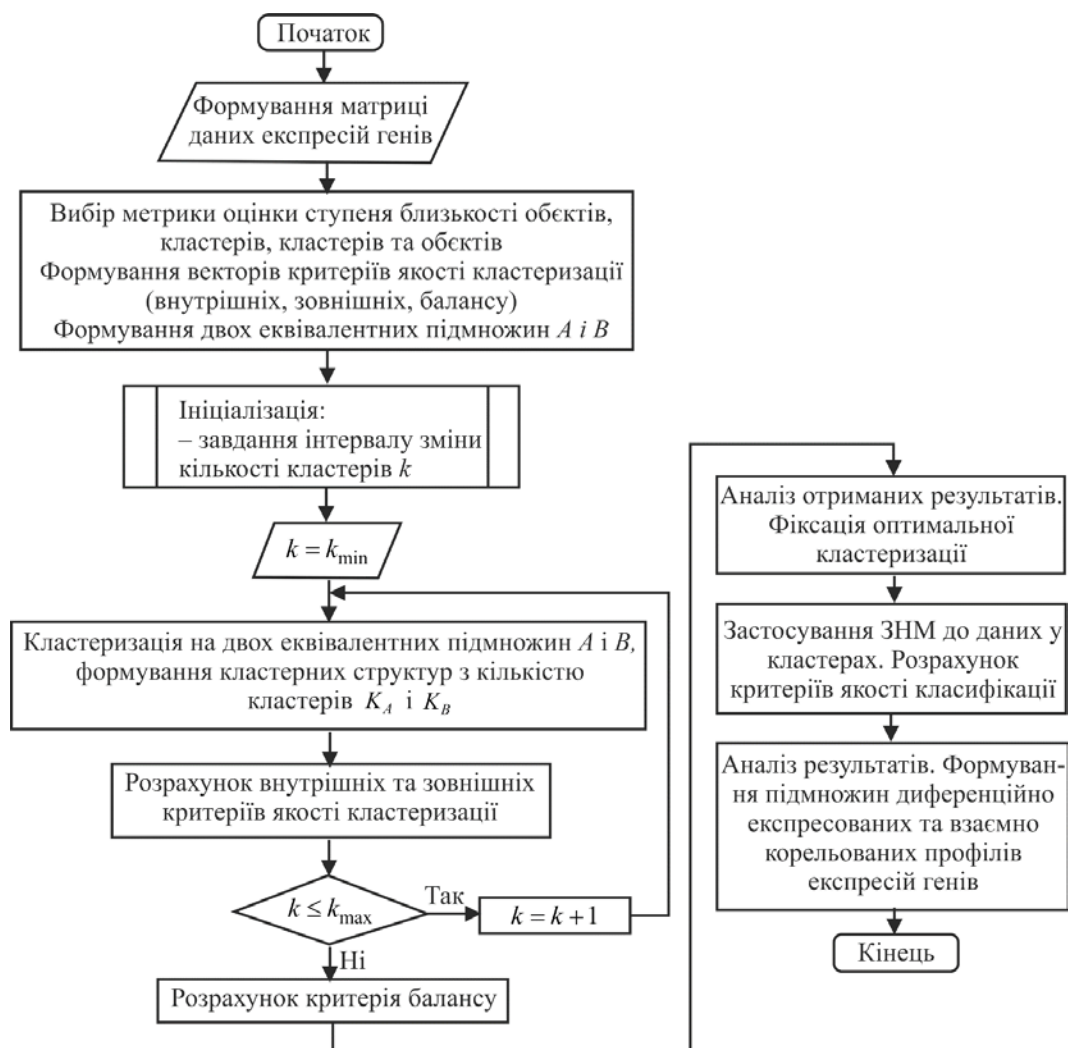


Рис. 4.22. Блок-схема алгоритму гібридної індуктивної моделі формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів на основі алгоритму спектральної кластеризації

**Етап II.** Кластеризація профілів експресії генів. Розрахунок внутрішніх та зовнішніх критеріїв якості кластеризації.

2.1. Ініціалізація кількості кластерів  $k = k_{\min}$

2.2. Кластеризація профілів експресії генів, що містяться у еквівалентних підмножинах даних. Формування кластерних структур.

2.3. Розрахунок внутрішніх та відповідних їм зовнішніх критеріїв якості кластеризації.

2.4. Якщо  $k < k_{\max}$ , збільшення кількості кластерів на 1 та перехід на крок 2.2 цієї процедури, в протилежному випадку – розрахунок критерію балансу на основі отриманих значень внутрішніх та зовнішніх критеріїв.

2.5. Аналіз отриманих результатів, фіксація оптимальних кластеризацій, що відповідають максимуму критерію балансу.

### **Етап III. Класифікація даних експресії генів.**

3.1. Формування підмножин даних експресії генів виділених кластерів для подальшого застосування як вхідні дані у згортковій нейронній мережі.

3.2. Застосування ЗНМ до даних експресії генів, що містяться у кожному кластері на виділеному ієрархічному рівні формування кластерної структури. Розрахунок критеріїв якості класифікації.

3.3. Аналіз отриманих результатів. Формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів.

#### **4.2.3.3. Результати моделювання щодо практичної реалізації гібридної індуктивної моделі кластеризації профілів експресії генів на основі алгоритму спектральної кластеризації**

На рис. 4.23 зображені результати моделювання щодо застосування запропонованої гібридної індуктивної моделі кластеризації профілів експресії генів на основі ієрархічного алгоритму спектральної кластеризації. Кількість кластерів у процесі моделювання змінювалося з 2 до 10. Таким чином, внутрішні та зовнішні критерії якості кластеризації у деяких випадках суперечать один одному, що підтверджує доцільність розрахунку критерію балансу, який містить як компоненти внутрішні та відповідні їм зовнішні критерії. Максимального значення цей критерій досягає при розділенні множини профілів експресії генів на три кластери (рис. 4.23d). При цьому за внутрішнім критерієм WB-index оптимальна кластеризація відповідає трьом кластерам для підмножини *A* і двом кластерам для підмножини *B* (рис. 4.23a). При застосування PBM внутрішнього критерію оптимальна кластеризація (відповідає максимальному значенню) у випадку застосування обох підмножин відповідають трикластерній структурі (рис. 4.23b). За зовнішніми критеріями оптимальна кластеризація відповідає трикластерній структурі при застосуванні WB-індексу і чотирикластерній структурі – при застосуванні PBM – критерію (рис. 4.23c).

Наступним етапом реалізації алгоритму, структурну блок-схему якого наведено на рис. 4.22, є застосування згорткової нейронної мережі до даних експресії генів у виділених кластерах. Для підтвердження вищенаведених висновків щодо ефективності критеріїв якості кластеризації для формування кластерної структури були досліджені структури, що містять 2, 3 та 4 кластери. Експериментальні дані містили 10000 профілів експресії генів для 156 пацієнтів, що досліджувалися на рак легенів. Результати моделювання наведено у табл. 4.4.

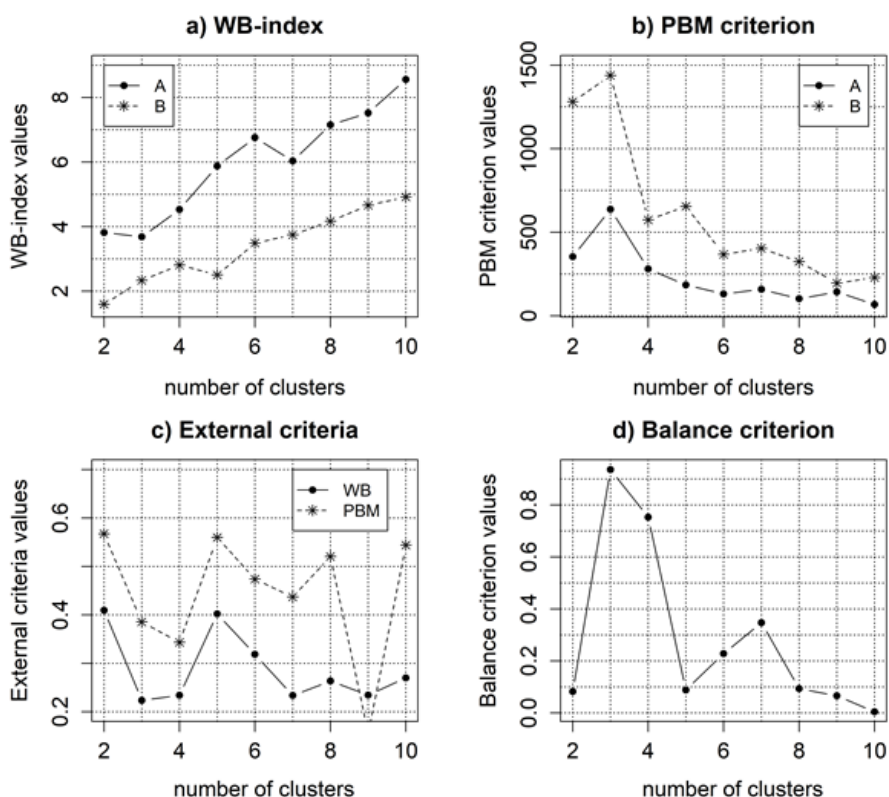


Рис. 4.23. Результати моделювання щодо практичної реалізації ієрархічного алгоритму спектральної кластеризації на основі індуктивних методів аналізу складних систем

Таблиця 4.4

Результати моделювання щодо покрокового застосування гібридної моделі обробки даних експресії генів на основі алгоритму спектральної кластеризації

| Етап і параметри моделі    |                 | Кластер 1    | Кластер 2    | Кластер 3    | Кластер 4 |
|----------------------------|-----------------|--------------|--------------|--------------|-----------|
| Двокластерна структура     | Кількість генів | 4074         | 4926         | —            | —         |
|                            | Розмір фільтра  | 60×68        | 50×119       | —            | —         |
|                            |                 | 30×136       | 25×238       | —            | —         |
|                            | Точність,%      | 95           | 97           | —            | —         |
| Трикластерна структура     | Втрати          | 0.254        | 0.067        | —            | —         |
|                            | Кількість генів | 2487         | 2549         | 4964         | —         |
|                            | Розмір фільтра  | 50×50        | 50×51        | 60×83        | —         |
|                            |                 | 25×100       | 25×102       | 30×166       | —         |
| Чотириклас-терна структура | Точність,%      | <b>97</b>    | <b>97</b>    | <b>100</b>   | —         |
|                            | Втрати          | <b>0.141</b> | <b>0.123</b> | <b>0.058</b> | —         |
|                            | Кількість генів | 1615         | 2779         | 4715         | 891       |
|                            | Розмір фільтра  | 30×54        | 50×56        | 50×95        | 50×18     |
|                            |                 | 15×108       | 25×112       | 25×190       | 25×36     |
|                            | Точність,%      | 97           | 97           | 97           | 95        |
|                            | Втрати          | 0.169        | 0.142        | 0.189        | 0.295     |

Аналіз отриманих результатів підтверджує факт, що трикластерна структура є оптимальною з точки зору як точності класифікації, так і значення функції втрат у процесі навчання нейронної мережі. Однак слід зазначити, що точність класифікації є досить високою в усіх випадках, що пояснюється високою ефективністю згорткової нейронної мережі для цього типу даних і стійкістю до наявності шумової компоненти. Точність класифікації об'єктів, що досліджувалися, оцінювалася на тестовій підмножині даних, яка не застосовувалась при навчанні нейронної мережі. При цьому у випадку трикластерної структури при застосування третього кластеру, що містить 4964 генів, досягається стовідсоткова точність класифікації при мінімальному значенні функції втрат. При застосуванні інших кластерів цієї кластерної структури, із 39 об'єктів, що складають тестову підмножину даних експресії генів, 38 ідентифіковані коректно. Представлені дослідження створюють умови для підвищення об'єктивності діагностики складних захворювань шляхом прийняття компромісного рішення за результатами класифікації об'єктів, що містять дані експресії генів у різних кластерах із застосуванням методу альтернативного голосування.

#### **4.3. Концептуальна схема інформаційної технології діагностики складних об'єктів на основі комплексного застосування алгоритму спектральної кластеризації, ЗНМ та методу альтернативного голосування**

На рис. 4.24 зображено концептуальну схему інформаційної технології обробки даних експресії генів для діагностики складних об'єктів, застосування якої передбачає покрокове використання алгоритму спектральної кластеризації, згорткової нейронної мережі та системи альтернативного голосування для прийняття остаточного рішення стосовно наявності і стану об'єкта.

Вхідними даними у запропонованій моделі є матриця експресії генів, рядки якої являють собою об'єкти, що досліджуються, а стовпці – гени, експресія яких характеризує стан відповідного об'єкта. Реалізація інформаційної технології передбачає такі етапи:

**Етап I.** Налаштування та реалізація алгоритму спектральної кластеризації із застосуванням принципів індуктивної технології об'єктивної кластеризації.

1.1. Формування метрики близькості профілів експресії генів, функцій для розрахунку внутрішніх, зовнішніх критеріїв якості кластеризації та критерія балансу.

1.2. Розділення множини профілів експресії генів на дві еквівалентні підмножини.

1.3. Ініціалізація інтервалу зміни кількості кластерів у процесі реалізації алгоритму спектральної кластеризації.

1.4. Кластеризація профілів експресії генів на еквівалентних підмножинах даних в рамках заданого на попередньому кроці інтервалу. Розрахунок внутрішніх та зовнішніх критеріїв якості кластеризації на кожному кроці реалізації цієї процедури.

1.5. Розрахунок критерію балансу. Формування кластерної структури, що відповідає максимальному значенню критерію балансу.



Рис. 4.24. Блок-схема інформаційної технології діагностики складних об'єктів на основі даних експресії генів із застосуванням алгоритму спектральної кластеризації, ЗНМ та методу альтернативного голосування

**Етап II.** Застосування згорткової нейронної мережі для класифікації об'єктів на основі даних експресії генів, що містяться у виділених кластерах.

2.1. Корекція даних експресії генів у виділених кластерах шляхом додавання необхідної кількості профілів з нульовою експресією для коректної роботи одновимірної двошарової згорткової нейронної мережі.

2.2. Формування гіперпараметрів мережі, враховуючи кількість профілів експресії генів у відповідних кластерах.

2.3. Формування підмножин даних експресії генів для навчання мережі, її валідації та тестування.

2.3. Застосування ЗНМ до даних експресії генів у виділених кластерах. Розрахунок точності класифікації об'єктів із застосуванням тестових підмножин даних

експресії генів. Формування векторів, що визначають належність об'єктів до відповідного класу для даних експресії генів, що містяться у кожному із виділених кластерів.

**Етап III.** Застосування методу альтернативного голосування до отриманих на попередньому етапі результатів для прийняття остаточного рішення щодо стану об'єкта, що досліджується.

3.1. Налаштування моделі, враховуючи кількість альтернатив, отриманих на етапі II цієї процедури. Формування правил для прийняття остаточного рішення на основі альтернатив, отриманих на попередньому етапі реалізації цієї процедури.

3.2. Застосування методу альтернативного голосування до векторів даних, що визначають результат класифікації об'єктів, що досліджуються, на основі даних експресії генів відповідних кластерів.

3.3. Аналіз отриманих результатів. Формування остаточного рішення щодо наявності або відсутності відповідної хвороби.

## **Розділ 5.**

### **Застосування розроблених моделей, методів та алгоритмів для створення системи діагностики ракових захворювань**

Цей розділ присвячений практичній реалізації розроблених моделей, методів та алгоритмів шляхом створення гібридних моделей діагностики різних захворювань на основі даних експресії генів із застосуванням методів інтелектуального аналізу даних та машинного навчання [9, 10, 12, 17, 25, 29]. За результатами моделювання, наведеними у попередніх розділах цієї монографії, підвищення об'єктивності формування інформативних підмножин профілів експресії генів і діагностики відповідних захворювань із застосуванням виділених даних експресії генів можливе шляхом застосування ансамблю методів реалізації відповідного етапу, з одного боку, та застосування ансамблю відповідних критеріїв якості для оцінки ефективності відповідного етапу обробки інформації з подальшим формування інтегрального (комплексного) критерія, з іншого боку. В дисертаційному дослідженні були застосовані дані експресії генів пацієнтів, що досліджувалися на різні типи раку з Атласу Геному Раку (The Cancer Genome Atlas (TCGA)) [276]. Формування експериментальних даних здійснювалося із застосуванням функцій пакету TCGAAbiolinks [277] модуля Bioconductor [278] програмного середовища R [279]. Дані отримані із застосуванням як ДНК-мікрочіпових експериментів, так і шляхом секвенування молекул mRNA, і включали значення експресії генів для більш ніж 10000 проб, що відповідають 33-м типам раку. Відповідно до результатів дослідження, що наведені у попередніх розділах, обробка даних та формування остаточних рішень проводилася шляхом реалізації таких етапів:

1. Попереднє формування матриці експресії генів, видалення генів, що мають нульову експресію для всіх зразків, що досліджуються.
2. Видалення неінформативних генів за статистичними та ентропійними критеріями за методикою, що представлена у розділі 2.
3. Кластеризація інформативних профілів експресії генів шляхом застосування індуктивної технології об'єктивної кластеризації відповідно до методики описаної у розділі 3. Формування оптимальної кластерної структури.
4. Класифікація зразків, що містять, як атрибути, дані експресії генів у сформованих кластерах шляхом застосування методів та алгоритмів, представлених у розділі 4. Формування проміжних рішень щодо стану відповідного об'єкта.
5. Формування остаточного рішення щодо стану об'єкта шляхом застосування моделі, концепція якої представлена у розділі 4 (рис. 4.24).

У наступних підрозділах представлено детальну реалізацію вищезазначеної процедури.

## 5.1. Формування та передобробка експериментальних даних

### 5.1.1. Формування та аналіз даних експресії генів із застосуванням функцій та модулів пакетів TCGAAbiolinks і Bioconductor

Вибір типу експериментальних даних для реалізації дисертаційних досліджень визначається тим фактом, що ракові захворювання на цей час є однією з основних причин смерті в усьому світі, і лікування раку варіюється від клінічних процедур, наприклад, хірургічне втручання зі складними комбінаціями ліків, до відповідних методів хіміопроменевої терапії. Атлас геному раку (The Cancer Genome Atlas (TCGA)), створення якого було розпочато в 2006 році з метою збору та аналізу як клінічних, так і молекулярних даних про понад 33 різні типи ракових пухлин шляхом відбору 500 випадків кожного типу пухлини, на сьогодні є одним із найповніших баз даних раку людини, що містить як молекулярні, так і клінічні дані. Пухлини, що профільовані із застосуванням методів TCGA, варіюються від суцільних до гематологічних, від легко до сильно агресивних з точки зору виживання та від доброякісних до метастатичних. Для кожного випадку раку було виділено ДНК, РНК і білок, а потім було виконано геномне, транскриптомне, епігеномне та протеомне профілювання з використанням різноманітного набору платформ, від спеціальних мікрочіпів до великомасштабного геномного секвенування. Таким чином, експериментальні дані експресії генів, що містяться у TCGA, є дуже потужною платформою для апробації та оцінки ефективності результатів дисертаційних досліджень з обробки даних експресії генів для створення гібридної моделі діагностики складних захворювань. До переваг пакету TCGAAbiolinks слід також віднести те, що його застосування дозволяє завантажувати та формувати дані з різних платформ, які застосовуються сьогодні у різних науково-дослідницьких установах для формування експериментальних даних на основі аналізу геному. До того ж функції цього пакету дозволяють реалізувати різні типи передобробки для формування та аналізу експериментальних даних із застосуванням різних методів візуалізації характеру розподілу даних у просторі відповідних ознак.

У рамках дисертаційного дослідження були використані дані експресії генів, отримані шляхом застосування методу геномного секвенування молекул РНК на платформі Illumina [280]. Кожний зразок початкових даних містив, як атрибути, 19947 типів генів, при цьому експресія відповідного гену визначалася рівнем його активності, який є пропорційним кількості відповідного типу гена. База експериментальних даних містила дев'ять типів даних, вісім з яких відповідали восьми типам ракових захворювань, а дев'яті дані відповідали зразкам, для яких за результатами клінічних випробувань ракова пухлина не ідентифікована. Загальна кількість зразків, що досліджувалися, дорівнювала 3269. Таким чином, на початковому етапі моделювання матриця експресії генів мала такий вигляд:  $E=(3269 \times 19947)$ . Класифікація експериментальних даних, що використовувалися в процесі моделювання, наведена в таблиці 5.1 та на діаграмі 5.1.

Таблиця 5.1

*Класифікація експериментальних даних експресії генів, що використовувалися в процесі моделювання*

| № | Тип раку                                                                | Кількість зразків |
|---|-------------------------------------------------------------------------|-------------------|
| 1 | Адренокортикальна карцинома (Adrenocortical carcinoma - ACC)            | 79                |
| 2 | Мультиформна гліобластома (Glioblastoma multiforme - GBM)               | 169               |
| 3 | Саркома (Sarcoma - SARC)                                                | 263               |
| 4 | Плоскоклітинний рак легень (Lung squamous cell carcinoma - LUSC)        | 502               |
| 5 | Аденокарцинома легені (Lung adenocarcinoma - LUAD)                      | 541               |
| 6 | Аденокарцинома шлунка (Stomach adenocarcinoma - STAD)                   | 415               |
| 7 | Світлоклітинний рак нирки (Kidney renal clear cell carcinoma - KIRC)    | 542               |
| 8 | Гліома головного мозку нижчого ступеня (Brain Lower Grade Glioma - LGG) | 534               |
| 9 | Ракова пухлина не виявлена (Normal)                                     | 224               |

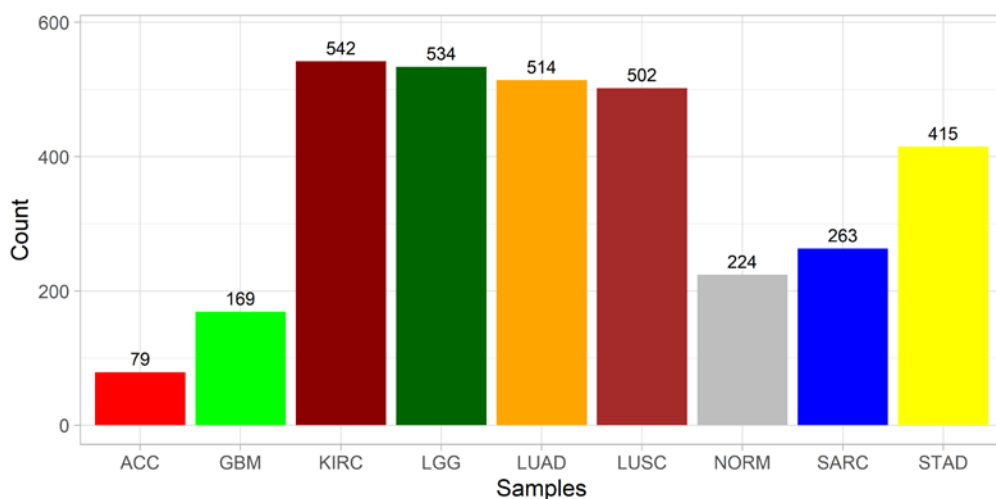


Рис. 5.1. Характер розподілу зразків у експериментальних даних експресії генів з урахуванням типу ракових захворювань

У таблиці 5.2 наведено загальну статистику розподілу максимальних значень кількостей генів, що містили експериментальні дані для усіх зразків, що досліджувалися.

Таблиця 5.2

*Загальна статистика розподілу максимальних значень кількостей генів для усіх зразків, що досліджувалися*

| Min | 1st quartile | Median | Mean  | 3rd quartile | Max     |
|-----|--------------|--------|-------|--------------|---------|
| 0   | 3775         | 12030  | 44324 | 30945        | 6303246 |

Аналіз даних таблиці 5.2 свідчить про те, що є певна кількість неекспресованих генів для всіх зразків (нульова кількість) і певна кількість низькоекспресованих генів (кількість активних генів для всіх зразків суттєво менша за кількість інших генів). Очевидно, що ці гени можуть бути видалені з даних без суттєвої втрати інформації. До того ж аналіз даних таблиці 5.2 свідчить про необхідність трансформування абсолютних значень кількості експресованих генів у більш зручні значення з меншим діапазоном їх варіації. У пакеті Bioconductor [276] запропоновано метод трансформації абсолютних значень кількості генів у відносне значення (кількість на мільйон (count-per-million-cpm)), що розраховується так:

$$count'_{ij} = \left( count_{ij} / \sum_{j=1}^m count_{ij} \right) \cdot 10^6, \quad (5.1)$$

де  $count_{ij}$  – кількість  $j$ -го гена для  $i$ -го зразка;  $m$  – кількість генів, що досліджується.

Загальна статистика розподілу максимальних трансформованих за формулою (5.1) значень кількості генів для усіх зразків, що досліджуються, подано у таблиці 5.3. Таким чином, інтервал варіювання трансформованих значень максимальної кількості генів для усіх зразків є суттєво меншим, порівняно з початковими даними, які представлені у таблиці 5.2, що спрощує процедуру подальшої передобробки експериментальних даних.

Таблиця 5.3

*Загальна статистика розподілу трансформованих максимальних значень кількостей генів для усіх зразків, що досліджувалися*

| Min | 1st quartile | Median | Mean   | 3rd quartile | Max       |
|-----|--------------|--------|--------|--------------|-----------|
| 0   | 76.73        | 237.87 | 886.58 | 598.56       | 172693.66 |

Наступним етапом є видалення з експериментальних даних неекспресованих профілів експресії генів за відповідним трешолдінговим коефіцієнтом, значення якого відповідає умові  $\log_2(count'_{ij}) = 0$  для усіх зразків, що досліджуються. Відповідно до цієї умови, гени, для профілів яких виконується умова:  $\max(count') \leq 1$ , вважаються неекспресованими і видаляються з даних. На цьому етапі кількість генів була зменшена на 682 і матриця експресії генів набула такої форми:  $E = (3269 \times 19265)$ . На наступному етапі значення експресії генів були трансформовані у більш зручніший діапазон шляхом застосування до усіх значень матриці  $E$  функції  $\log_2$ , при цьому, як показали результати моделювання, деякі значення, що менші за 1, були трансформовані у відповідні негативні значення, що є несприйнятливим, враховуючи той факт, що мінімальне значення експресії генів є 0 (ген є неактивний). Тому на цьому етапі негативні значення були замінені на нуль, що відповідає неекспресованим генам. На рис. 5.2 зображено характер розподілу нормалізованих значень експресії дев'яти генів для одного зразка адренокортикальної карциноми (ACC). На рис. 5.3 зображено характер розподілу одного типу гена для усіх зразків, що досліджувалися в процесі моделювання, при цьому досліджувалися гени A1BG.1 (рис. 5.3a), A2M.2 (рис. 5.3b),

NAT1.9 (рис. 5.3с) та NAT2.10 (рис. 5.3d). Аналіз отриманих результатів дозволяє зробити висновок, що характер розподілу значень експресії генів суттєво розрізняється для різних зразків як за абсолютним значенням, так і за дисперсією, що створює умови для ідентифікації зразків за значеннями експресії шляхом застосування відповідного класифікатора. Наступним кроком етапу передобробки даних експресії генів є видалення неінформативних генів за статистичними і ентропійними критеріями із застосуванням функції бажаності Харрінгтона, алгоритм застосування для розрахунку комплексного критерія наведено у розділі 2 дисертаційної роботи.

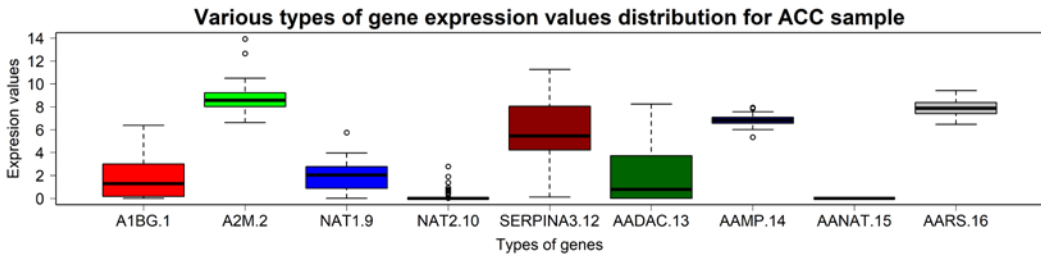


Рис. 5.2. Характер розподілу значень експресії різних типів генів для одного зразка адренкортикальної карциноми (АСС)

### 5.1.2. Видалення неінформативних генів за статистичними критеріями та ентропією Шеннона

На рис. 5.4 зображені діаграми розподілу максимальних абсолютних значень профілів експресії генів, що досліджувалися, дисперсії та ентропії Шеннона. Аналіз отриманих діаграм свідчить про зростання рівня інформативності профілів експресії генів при зростанні максимальних абсолютних значень експресії генів і дисперсії відповідних профілів і зменшенні значення ентропії Шеннона. Відповідно до вищезазначеного, алгоритм для розрахунку комплексного критерія на основі методу Харрінгтона, який визначає рівень інформативності генів, що досліджуються, передбачає такі кроки:

1. Визначення значень коефіцієнтів у лінійних рівняннях, що дозволяють трансформувати шкали статистичних критеріїв та ентропії Шеннона, що розрахована за методом Джеймса та Стейна [281], у шкалу безрозмірного показника  $Y$ , враховуючи граничні значення відповідних критеріїв ( $Y_{\min} = -2$ ;  $Y_{\max} = 5$ ):

$$\begin{cases} Y_{\min}^{(abs, var)} = a^{(abs, var)} + b^{(abs, var)} \cdot \min(abs, var) \\ Y_{\max}^{(abs, var)} = a^{(abs, var)} + b^{(abs, var)} \cdot \max(abs, var) \end{cases} \quad (5.2)$$

$$\begin{cases} Y_{\min}^{(entr)} = a^{(entr)} - b^{(entr)} \cdot \max(entr) \\ Y_{\max}^{(entr)} = a^{(entr)} - b^{(entr)} \cdot \min(entr) \end{cases} \quad (5.3)$$

2. Розрахунок показника  $Y$  для кожного значення статистичних критеріїв та ентропії Шеннона:

$$\begin{cases} Y^{(abs)} = a^{(abs)} + b^{(abs)} \cdot \max\_abs \\ Y^{(var)} = a^{(var)} + b^{(var)} \cdot var \\ Y^{(entr)} = a^{(entr)} - b^{(entr)} \cdot entr \end{cases} \quad (5.4)$$

де  $\max\_abs$  – максимум абсолютних значень експресії профіля відповідного гена;  
 $var$  і  $entr$  – дисперсія та ентропія Шеннона цього профіля експресії гена відповідно.

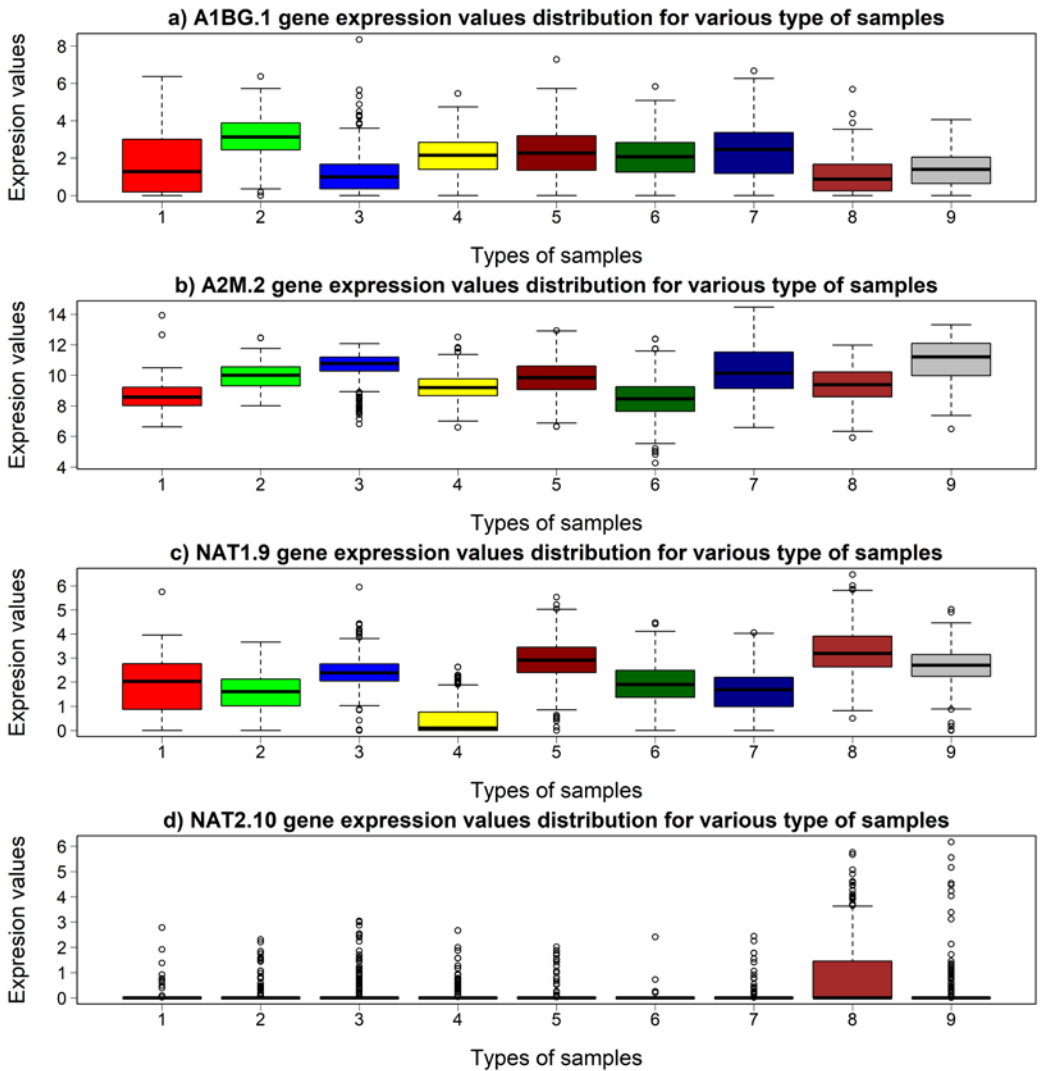


Рис. 5.3. Характер розподілу значень експресії одного типу гена для різних зразків, що досліджуються. Досліджувалися гени: а) A1BG.1; б) A2M.2; в) NAT1.9; д) NAT2.10

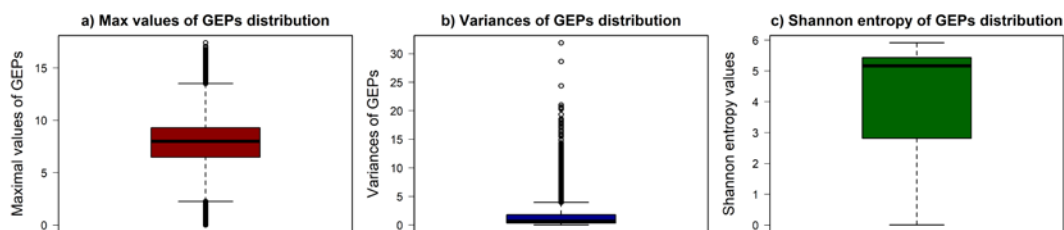


Рис. 5.4. Характер розподілу статистичних значень та ентропії Шеннона профілів експресії генів, що досліджуються: а) максимальне значення експресії; б) дисперсія; в) ентропія Шеннона

3. Розрахунок приватних бажаностей для кожного значення параметра  $Y$ :

$$\begin{cases} d^{(abs)} = \exp(-\exp(-Y^{(abs)})); \\ d^{(var)} = \exp(-\exp(-Y^{(var)})); \\ d^{(entr)} = \exp(-\exp(-Y^{(entr)})). \end{cases} \quad (5.5)$$

4. Розрахунок узагальненого показника якості профіля експресії гену як геометричне середнє усіх приватних бажаностей:

$$D^{(general)} = \sqrt[3]{d^{(abs)} \cdot d^{(var)} \cdot d^{(entr)}}. \quad (5.6)$$

Більше значення показника (5.6) відповідає більшому ступеню інформативності профіля експресії гену за статистичними критеріями та ентропією Шеннона. У таблиці 5.4 представлено загальну статистику характеру розподілу значень узагальненого індексу.

Таблиця 5.4

*Загальна статистика розподілу значень узагальненого індексу бажаності Харрінгтона*

| Min      | 1st quartile | Median   | Mean     | 3rd quartile | Max      |
|----------|--------------|----------|----------|--------------|----------|
| 0.007239 | 0.029438     | 0.039786 | 0.074478 | 0.075635     | 0.947162 |

Фільтрація профілів експресії генів на цьому етапі передбачала видалення профілів, узагальнений індекс бажаності для яких менший за значення індексу, що відповідає першому квантилю, тобто 0,029438. Кількість профілів експресії генів у цьому випадку була зменшена на 4814, тобто фільтрована матриця значень експресії генів набула форми:  $E = (3269 \times 14451)$ .

Оцінка адекватності моделі фільтрації профілів експресії генів здійснювалася шляхом застосування бінарного класифікатора до зразків, що містять як атрибути повний набір експресії генів (19265) і набір експресії генів після видалення неінформативних за статистичними критеріями та ентропією Шеннона (14451), при цьому на першому кроці повна множина зразків розділялася на вісім підмножин, кожна з яких містила експресії генів зразків, для яких пухлина не виявлена (224 зразка) і зразки з раковою пухлиною з урахуванням типу захворювання. Відповідно до методики, що представлена у попередніх розділах дисертаційної роботи, експериментальні дані

розділялися на дві підмножини, 65% зразків застосовувалися для навчання класифікатору, а 35% – для тестування отриманої моделі з розрахунком критеріїв якості класифікації даних. На цьому етапі як класифікатор застосовувався алгоритм «Випадкового лісу (Random Forest)», ефективність якого для бінарної класифікації об'єктів на основі високовимірних даних експресії генів доведена у попередніх розділах цієї монографії. У таблиці 5.5 представлені результати бінарної класифікації об'єктів при застосуванні повної множини профілів експресії генів (19265). Слід зазначити, що при застосуванні фільтрованих даних (14451) результати майже не відрізнялися.

Таблиця 5.5

*Результати бінарної класифікації об'єктів при застосуванні повної множини профілів експресії генів*

| Тип даних | Критерії класифікації |            |               |        |       |
|-----------|-----------------------|------------|---------------|--------|-------|
|           | Точність              | Чутливість | Специфічність | F-міра | MCC   |
| acc       | 1                     | 1          | 1             | 1      | 1     |
| gbm       | 0.993                 | 0.983      | 1             | 0.988  | 0.985 |
| kirc      | 1                     | 1          | 1             | 1      | 1     |
| lgg       | 0.992                 | 0.995      | 0.987         | 0.994  | 0.982 |
| luad      | 0.989                 | 0.984      | 1             | 0.987  | 0.973 |
| lusc      | 0.992                 | 0.989      | 1             | 0.990  | 0.982 |
| sarc      | 1                     | 1          | 1             | 1      | 1     |
| stad      | 0.987                 | 0.963      | 1             | 0.975  | 0.971 |

Аналіз даних таблиці 5.5 свідчить про високу точність бінарної класифікації із застосуванням повних даних експресії генів, у трьох випадках модель зі стовідсотковою точністю ідентифікувала наявність або відсутність ракової пухлини у пацієнта. У інших п'яти випадках точність ідентифікації стану пацієнта за критеріями класифікації, що використовувалися не є стовідсотковою, але також є високою, порівняно з даними, що представлені у відповідних наукових публікаціях. У таблицях 5.6 і 5.7 наведено результати застосування згорткової нейронної мережі для розв'язання мультикласової задачі на основі повних та фільтрованих даних експресії генів, відповідно. Враховуючи результати моделювання, що представлені у розділі 4 цієї монографії, було застосовано одновимірну двошарову ЗНМ, структурну топологію якої зображено на рис. 4.6.

Таблиця 5.6

*Результати моделювання щодо застосування ЗНМ для класифікації різних типів ракових захворювань при застосуванні повних даних (19265 генів)*

| Клас   | Кількість об'єктів | Чутливість | Специфічність | F-міра | Точність |
|--------|--------------------|------------|---------------|--------|----------|
| ACC    | 24                 | 0.85       | 0.96          | 0.9    | 97%      |
| GBM    | 73                 | 0.97       | 0.93          | 0.95   |          |
| KIRC   | 169                | 0.99       | 0.98          | 0.99   |          |
| LGG    | 191                | 0.97       | 0.98          | 0.98   |          |
| LUAD   | 200                | 0.96       | 0.96          | 0.96   |          |
| LUSC   | 169                | 0.96       | 0.96          | 0.96   |          |
| SARC   | 73                 | 0.96       | 0.96          | 0.96   |          |
| STAD   | 96                 | 0.94       | 0.96          | 0.95   |          |
| NORMAL | 149                | 0.99       | 0.98          | 0.99   |          |

Таблиця 5.7

*Результати моделювання щодо застосування ЗНМ для класифікації різних типів ракових захворювань при застосуванні фільтрованих даних (14451 генів)*

| Клас   | Кількість об'єктів | Чутливість | Специфічність | F-міра | Точність |
|--------|--------------------|------------|---------------|--------|----------|
| ACC    | 24                 | 0.74       | 0.96          | 0.84   | 96%      |
| GBM    | 73                 | 0.96       | 0.92          | 0.94   |          |
| KIRC   | 169                | 0.99       | 0.99          | 0.99   |          |
| LGG    | 191                | 0.97       | 0.98          | 0.97   |          |
| LUAD   | 200                | 0.94       | 0.96          | 0.95   |          |
| LUSC   | 169                | 0.96       | 0.93          | 0.95   |          |
| SARC   | 73                 | 0.97       | 0.96          | 0.97   |          |
| STAD   | 96                 | 0.95       | 0.93          | 0.94   |          |
| NORMAL | 149                | 0.99       | 0.99          | 0.99   |          |

Для коректної ініціалізації фільтрів першого та другого шару експериментальні дані були доповнені профілями з нульовою експресією до 19300 та 14500 у випадку застосування повних та фільтрованих даних експресії генів. Розмір ядра був визначений на рівні 8, а густина ущільненого шару – 256. Результати оцінювалися на тестовій підмножині даних, кількість об'єктів якої становила 35% від повної кількості об'єктів, що досліджувалися (1144 з 3269). Тестова підмножина даних при цьому не використовувалася для навчання мережі.

Аналіз отриманих результатів свідчить про високу ефективність застосування ЗНМ для розв'язання мультикласових задач. Так, при застосуванні повної множини профілів експресії генів із 1144 об'єктів, що використовувалися для тестування моделі, тільки 36 і 44 об'єктів були ідентифіковані некоректно при застосуванні повних і фільтрованих даних профілів експресії генів, відповідно, при цьому точність класифікації склала 97% при використанні повних даних і 96% при використанні фільтрованих даних. До того ж аналіз значень критеріїв класифікації за класами свідчить про високу точність ідентифікації зразків, для яких ракова пухлина не виявлена, що підтверджує результати бінарної класифікації, які отримані із застосуванням алгоритму класифікації Random Forest (RF). Цей факт свідчить про те, що ЗНМ можуть ефективно використовуватися у системах ідентифікації наявності або відсутності ракової пухлини на основі даних експресії генів.

На рисунках 5.5 і 5.6 зображені діаграми зміни точності класифікації об'єктів та функцій втрат, що отримані на множині об'єктів, які використовувалися безпосередньо для навчання мережі і для валідації моделі в процесі її навчання. В процесі моделювання множина даних, що виділялася для навчання мережі, розділялася на дві підмножини у співвідношенні 0,7/0,3. 70% зразків при цьому використовувалися безпосередньо для навчання, а 30% – для валідації моделі в процесі навчання. Збіжність як точності класифікації, так і функцій втрат, що отримані на двох підмножинах даних, свідчить про відсутність перенавчання мережі. У цьому випадку результати, що отримані на тестовій підмножині даних, можна вважати адекватними.

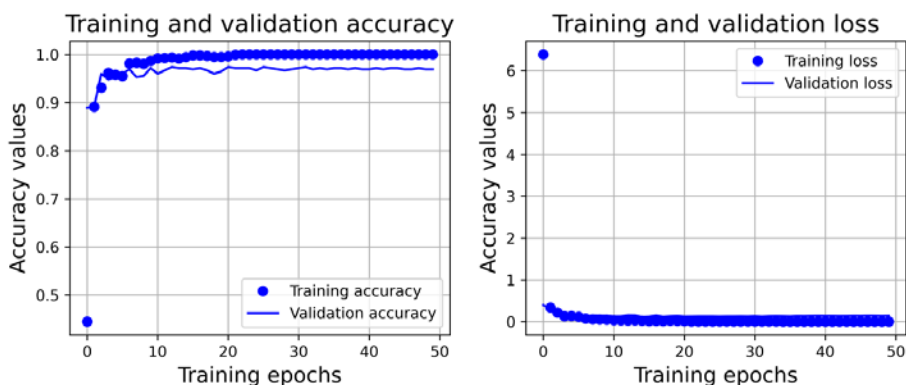


Рис. 5.5. Діаграми варіювання точності класифікації об'єктів та функцій втрат, що отримані на множині об'єктів, які використовувалися безпосередньо для навчання мережі і для валідації моделі при застосуванні повних профілів експресії генів

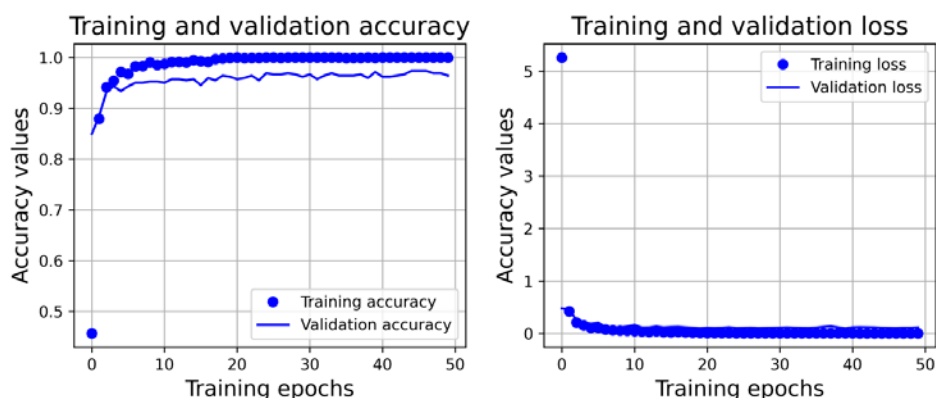


Рис. 5.6. Діаграми варіювання точності класифікації об'єктів та функцій втрат, що отримані на множині об'єктів, які використовувалися для навчання мережі і для валідації моделі при застосуванні фільтрованих профілів експресії генів

Аналіз отриманих результатів свідчить також про малу розбіжність результатів класифікації зразків, що отримані на повному і фільтрованому наборі профілів експресії генів. Цей факт підтверджує доцільність видалення неінформативних генів на етапі передобробки даних за статистичними і непропійними критеріями, оскільки зменшення кількості профілів експресії генів у цьому випадку майже не впливає на результати класифікації зразків, а час на обробку даних і обсяг комп'ютерних ресурсів, які необхідні для обробки інформації суттєво зменшуються.

## 5.2. Застосування індуктивної моделі об'єктивної кластеризації на основі алгоритму спектральної кластеризації для формування підмножин диференційно експресованих та взаємно корельованих підмножин профілів експресії генів

Як зазначалося у підрозділі 4.3 цієї монографії, підвищення об'єктивності ідентифікації стану об'єкту можливе за рахунок розпаралелювання процесу обробки даних експресії генів шляхом формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів з подальшою класифікацією об'єктів, що містять як атрибути дані експресії генів у виділених кластерах, при цьому остаточне рішення приймається шляхом аналізу результатів класифікації даних усіх кластерів із застосуванням системи нечіткого логічного виводу. Структурна блок-схема реалізації цієї процедури зображена на рис. 5.7.

Першим етапом реалізації процедури, що зображена на рис. 5.7, є формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів із застосуванням індуктивного алгоритму спектральної кластеризації, переваги і ефективність якого для кластеризації високовимірних профілів експресії генів доведена у розділі 4 цієї монографії. Ступень близькості профілів експресії генів, кластерів та профілів експресії генів і кластерів у рамках поточного моделювання оцінювалася шляхом використання гібридної метрики близькості профілів експресії генів, що містить, як компоненти, метрики на основі оцінки взаємної інформації із застосування різних методів розрахунку ентропії (модифікована метрика близькості на основі максимізації взаємної інформації). При цьому максимальне значення взаємної інформації за цим критерієм відповідало мінімальній відстані між відповідними профілями експресії генів.

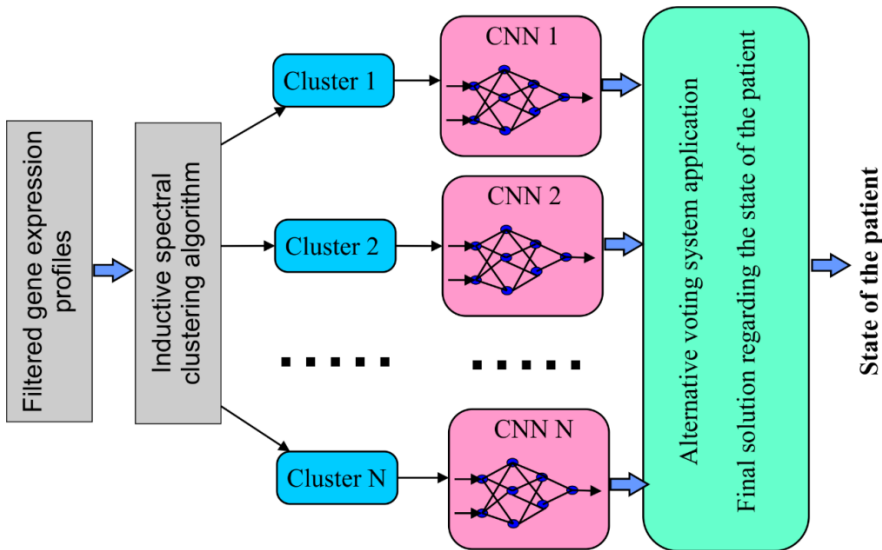


Рис. 5.7. Структурна блок-схема покрокової процедури обробки даних експресії генів у системі діагностики стану об'єкта, що досліджується

Внутрішній критерій якості кластеризації профілів експресії генів у цьому випадку враховував як характер розподілу у окремих кластерах щодо медіани відповідного кластера та характер розподілу окремих кластерів (медіан) у просторі ознак. Формула для розрахунку внутрішнього критерію якості кластеризації у цьому випадку має такий вигляд:

$$QC_{int} = \frac{K(K-1)ADW^2}{(N-K)ADB^2}, \quad (5.7)$$

де  $K$  – кількість кластерів, що формують кластерну структуру;  $ADW$  (Average Distance Within) – середня відстань між профілями експресії генів до медіан кластерів, де ці профілі знаходяться:

$$ADW = \frac{1}{m} \sqrt{\sum_{k=1}^K \sum_{i=1}^{m_k} MI(e_i, M_k)^2}, \quad (5.8)$$

де  $m$  – загальна кількість профілів експресії генів;  $m_k$  – кількість профілів експресії генів у  $k$ -му кластері;  $M_k$  – медіана  $k$ -го кластера;  $e_i$  – вектор значень експресії  $i$ -го гена;  $MI(\cdot)$  – значення взаємної інформації (більше значення взаємної інформації відповідає меншій відстані між профілями експресії генів).

Друга компонента критерію (5.7) розраховується як середнє значення відстаней між усіма парами медіан кластерів, що складають кластерну структуру:

$$ADB = \frac{2}{K(K-1)} \sqrt{\sum_{i=1}^{K-1} \sum_{j=i+1}^K MI(M_i, M_j)^2}. \quad (5.9)$$

При цьому слід зазначити, що менше значення критерію (5.7) відповідає кращій кластеризації за цим типом критерію якості. Зовнішній та балансу критерії розраховувалися відповідно до методики, що представлена у розділі 3 цієї монографії.

Відповідно до методики, що представлена у розділі 3 цієї монографії, реалізація алгоритму кластеризації в рамках індуктивної технології об'єктивної кластеризації передбачає формування двох еквівалентних підмножин профілів експресії генів із застосуванням відповідної метрики близькості профілів експресії генів. Як зазначалося вище, в рамках поточних досліджень застосовувалася модифікована метрика близькості на основі комплексного застосування різних методів максимізації взаємної інформації (передбачає застосування різних методів розрахунку ентропії Шеннона). Формування оптимальної кластерної структури шляхом одночасного застосування алгоритму спектральної кластеризації до еквівалентних підмножин профілів експресії генів з формуванням кластерних структур на наступному етапі реалізації цієї процедури, при цьому кількість кластерів у відповідних кластерних структурах змінювалося від 2 до 7. На кожному кроці розраховувалися внутрішні критерії якості кластеризації за формулами (5.7–5.9). Зовнішній критерій розраховувався як нормалізована різниця внутрішніх критеріїв на кожному кроці реалізації цієї процедури:

$$QC_{ext} = \frac{|QC_{int}^A - QC_{int}^B|}{QC_{int}^A + QC_{int}^B}. \quad (5.10)$$

Менше значення критерію (5.10) відповідає меншій похибці відтворюваності, що виникає в процесі формування кластерної структури. Критерій балансу, що

містить як компоненти внутрішні та зовнішній критерії, розраховувався із застосуванням функції бажаності Харрінгтона за методикою, детальний опис якої представлено у розділі 3 цієї монографії. Результати моделювання щодо визначення оптимальної кластерної структури наведено на рис. 5.8.

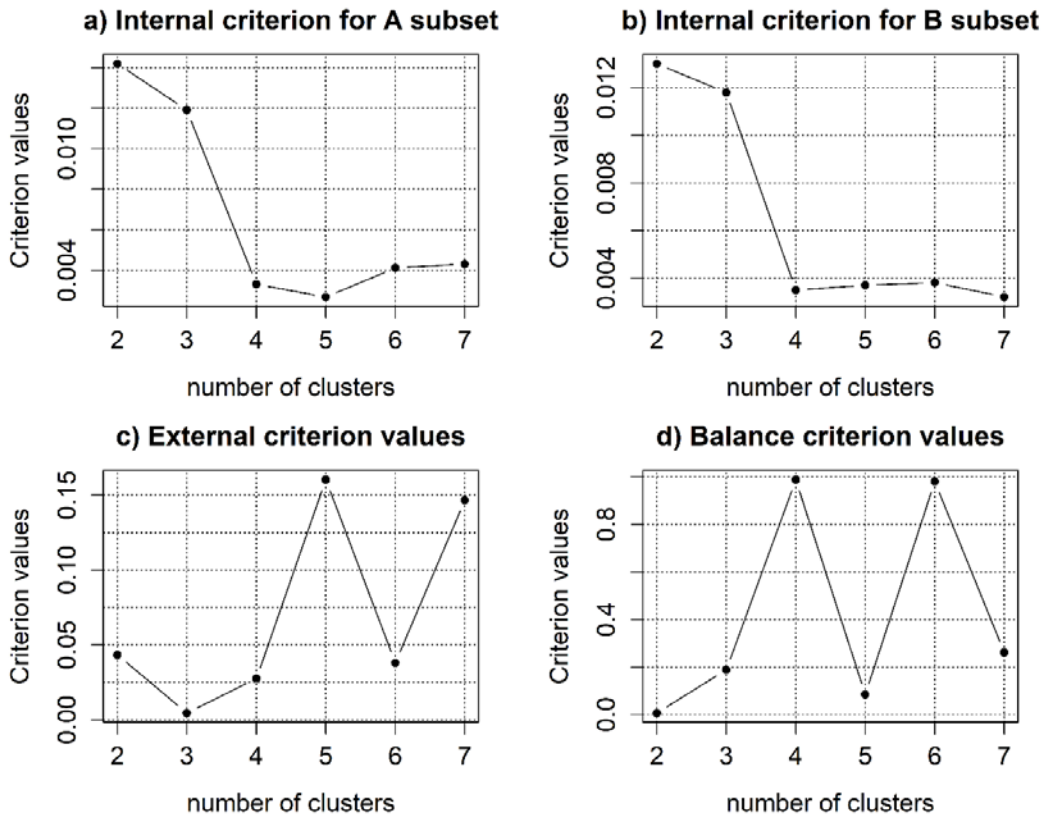
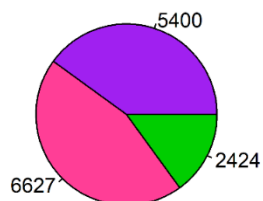
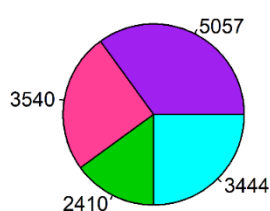


Рис. 5.8. Результати моделювання щодо визначення оптимальної кластерної структури при застосуванні індуктивного алгоритму спектральної кластеризації: а, б) значення внутрішніх критеріїв якості кластеризації залежно від кількості кластерів, що розраховані на еквівалентних підмножинах А і В, відповідно; с) зовнішній критерій якості кластеризації; д) критерій балансу

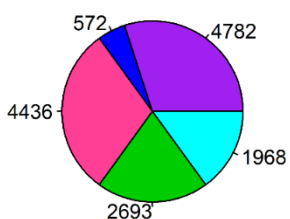
**a) Tree-cluster structure**



**b) Four-cluster structure**



**c) Five-cluster structure**



**d) Six-cluster structure**

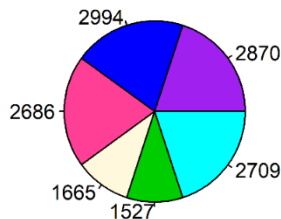
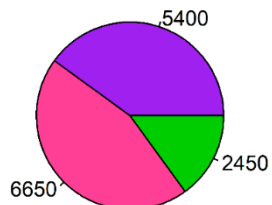
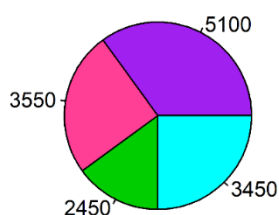


Рис. 5.9. Характер розподілу профілів експресії генів у кластерах при застосуванні індуктивного алгоритму спектральної кластеризації

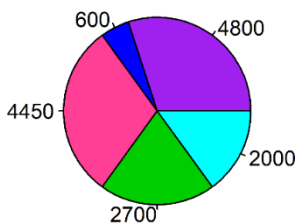
**a) Tree-cluster structure**



**b) Four-cluster structure**



**c) Five-cluster structure**



**d) Six-cluster structure**

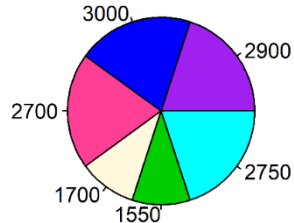


Рис. 5.10. Характер розподілу профілів експресії генів у кластерах після додавання профілів з нульовою експресією

Аналіз отриманих результатів підтверджує висновки, що отримані у попередніх розділах цієї монографії щодо можливого неузгодження внутрішніх та зовнішніх критеріїв кластеризації даних в процесі формування оптимальної кластерної структури. Так, як видно на рис 5.10a,b, мінімальне значення внутрішнього критерію відповідає п'ятикластерній структурі на підмножині А профілів експресії генів і

семикластерній структурі на підмножині В. Однак слід зазначити, що різниця між значеннями внутрішніх критеріїв в інтервалі зміни кількості кластерів від 4 до 7 є незначна. Цей факт свідчить про проблематичність вибору оптимальної кластерної структури на основі аналізу значень внутрішніх критеріїв якості кластеризації. За похибкою відтворюваності найкращою є трикластерна структура, оскільки значення зовнішнього критерія при цьому є мінімальним, але чотирьох- та шестикластерні структури також сприйнятливі за зовнішнім критерієм. Аналіз характеру розподілу значень критерію балансу, який містить як компоненти внутрішні та зовнішній критерії якості кластеризації дозволяє зробити висновок, що оптимальними є чотири- та шестикластерні структури, які відповідають максимальному значенню критерію балансу.

Підтвердження отриманих результатів проведено шляхом реалізації процедури класифікації об'єктів, атрибутами яких є значення експресії генів у сформованих кластерах, при цьому досліджувалися три-, чотири-, п'яти- та шестикластерні структури. На рис. 5.9 зображено характер розподілу профілів експресії генів у сформованих кластерах. Як видно з рисунку, у випадку чотирьох- та шестикластерних структур характер розподілу профілів експресії генів у сформованих кластерах приблизно рівномірний. Цей факт може визначати більш високу якість розподілу профілів експресії генів у кластери за критеріями, що використовуються в рамках запропонованої моделі.

Для коректного застосування згорткової нейронної мережі до даних експресії генів у сформованих кластерах матриці експресії генів доповнювалися профілями експресії генів з нульовими значеннями. Кількість додаткових профілів визначалася так, щоб сумарна кількість профілів була кратна п'ятдесяти (один з параметрів фільтра у ЗНМ). На рис. 5.10 зображено остаточний характер розподілу профілів експресії генів у кластерах після додавання профілів з нульовою експресією генів. Наступним кроком є застосування ЗНМ до даних експресії генів у сформованих кластерах.

### **5.3. Застосування згорткової нейронної мережі до даних експресії генів у сформованих кластерах**

Класифікація об'єктів, що містять як атрибути дані експресії генів у сформованих кластерах, здійснювалася із застосуванням одновимірної двошарової згорткової нейронної мережі. Розмір ядра був також визначений на рівні 8, а густина ущільненого шару – 256. Параметри фільтрів, що застосовувалися при використанні даних експресії генів різних кластерних структур, представлені у таблиці 5.8. Як і у попередній моделі класифікації об'єктів на основі даних експресії генів із застосуванням ЗНМ, на попередньому етапі множина об'єктів, що досліджувалися (3269), розділялися на три підмножини: 65% від загальної кількості об'єктів (2125) використовувалися для навчання та валідації моделі, а 35% (1144) – для тестування. При цьому з першої підмножини 70% об'єктів (1487) використовувалися безпосередньо для навчання мережі, а 30% (638) для валідації моделі.

Таблиця 5.8

*Параметри фільтрів, що використовувалися у одновимірній двошаровій ЗНМ*

| Шар | Трикластерна структура   |          |          | Чотирикластерна структура |          |          |          |
|-----|--------------------------|----------|----------|---------------------------|----------|----------|----------|
|     | Кл. 1                    | Кл. 2    | Кл. 3    | Кл. 1                     | Кл. 2    | Кл. 3    | Кл. 4    |
| 1   | 108 × 50                 | 133 × 50 | 49 × 50  | 102 × 50                  | 71 × 50  | 49 × 50  | 69 × 50  |
| 2   | 216 × 25                 | 266 × 25 | 98 × 25  | 204 × 25                  | 142 × 25 | 98 × 25  | 138 × 25 |
| Шар | П'ятикластерна структура |          |          |                           |          |          |          |
|     | Кл. 1                    | Кл. 2    | Кл. 3    | Кл. 4                     | Кл. 5    |          |          |
| 1   | 96 × 50                  | 12 × 50  | 89 × 50  | 54 × 50                   | 40 × 50  |          |          |
| 2   | 192 × 25                 | 24 × 25  | 198 × 25 | 108 × 25                  | 80 × 25  |          |          |
| Шар | Шестикластерна структура |          |          |                           |          |          |          |
|     | Кл. 1                    | Кл. 2    | Кл. 3    | Кл. 4                     | Кл. 5    | Кл. 6    |          |
| 1   | 58 × 50                  | 60 × 50  | 54 × 50  | 34 × 50                   | 31 × 50  | 55 × 50  |          |
| 2   | 116 × 25                 | 120 × 25 | 108 × 25 | 68 × 25                   | 62 × 25  | 110 × 25 |          |

Оцінка якості моделі здійснювалася шляхом оцінки характеру зміни точності класифікації та функції втрат у процесі навчання мережі, що розраховувалися на підмножинах даних експресії генів, які використовувалися безпосередньо для навчання та для валідації моделі. Адекватність моделі оцінювалася шляхом застосування тестових даних з розрахунком критеріїв якості класифікації даних.

У таблицях 5.9–5.11 наведено результати моделювання щодо використання згорткової нейронної мережі до зразків, атрибутами яких є значення експресії генів у кластерах при застосуванні трикластерної структури. У таблицях 5.12–5.26 представлені аналогічні результати при застосуванні чотири-, п'яти- та шестикластерних структур, відповідно. На рис. 5.11–5.13 зображено характер розподілу відповідних критеріїв якості класифікації зразків, що розраховувалися в процесі моделювання при ідентифікації наявності або відсутності відповідної хвороби при застосуванні даних експресії генів, що знаходилися у виділених кластерах.

Таблиця 5.9

*Результати моделювання щодо використання ЗНМ для класифікації різних типів ракових захворювань при застосуванні першого кластера у трикластерній структурі (45 зразків з 1144 ідентифіковані некоректно)*

| Клас   | Кількість об'єктів | Чутливість | Специфічність | F-міра | Точність |
|--------|--------------------|------------|---------------|--------|----------|
| ACC    | 24                 | 0.85       | 0.96          | 0.90   | 0.961    |
| GBM    | 73                 | 0.96       | 0.92          | 0.94   |          |
| KIRC   | 169                | 0.99       | 0.99          | 0.99   |          |
| LGG    | 191                | 0.97       | 0.98          | 0.98   |          |
| LUAD   | 200                | 0.95       | 0.95          | 0.95   |          |
| LUSC   | 169                | 0.94       | 0.94          | 0.94   |          |
| SARC   | 73                 | 0.97       | 0.95          | 0.96   |          |
| STAD   | 96                 | 0.94       | 0.97          | 0.95   |          |
| NORMAL | 149                | 0.99       | 0.97          | 0.98   |          |

Аналіз отриманих результатів дозволяє зробити висновок щодо високої збіжності критеріїв якості кластеризації, розподіл яких для отриманих кластерних структур подано на рис. 5.8.

Таблиця 5.10

*Результати моделювання щодо використання ЗНМ для класифікації різних типів ракових захворювань при застосуванні другого кластера у трикластерній структурі (42 зразка з 1144 ідентифіковані некоректно)*

| Клас   | Кількість об'єктів | Чутливість | Специфічність | F-міра | Точність |
|--------|--------------------|------------|---------------|--------|----------|
| ACC    | 24                 | 0.79       | 0.96          | 0.87   | 0.963    |
| GBM    | 73                 | 1          | 0.89          | 0.94   |          |
| KIRC   | 169                | 0.98       | 0.99          | 0.98   |          |
| LGG    | 191                | 0.96       | 0.99          | 0.97   |          |
| LUAD   | 200                | 0.95       | 0.95          | 0.95   |          |
| LUSC   | 169                | 0.96       | 0.95          | 0.95   |          |
| SARC   | 73                 | 0.99       | 0.93          | 0.96   |          |
| STAD   | 96                 | 0.94       | 0.98          | 0.96   |          |
| NORMAL | 149                | 1          | 0.98          | 0.96   |          |

При цьому слід зазначити, що майже в усіх випадках точність класифікації зразків є досить високою, при цьому у кожній кластерній структурі можна виділити один кластер профілів експресії генів, при застосуванні яких точність класифікації більш низька, порівняно з іншими випадками.

Таблиця 5.11

*Результати моделювання щодо використання ЗНМ для класифікації різних типів ракових захворювань при застосуванні третього кластера у трикластерній структурі (74 зразка з 1144 ідентифіковані некоректно)*

| Клас   | Кількість об'єктів | Чутливість | Специфічність | F-міра | Точність |
|--------|--------------------|------------|---------------|--------|----------|
| ACC    | 24                 | 0.74       | 0.96          | 0.84   | 0,935    |
| GBM    | 73                 | 0.94       | 0.85          | 0.89   |          |
| KIRC   | 169                | 0.98       | 0.98          | 0.98   |          |
| LGG    | 191                | 0.95       | 0.97          | 0.96   |          |
| LUAD   | 200                | 0.90       | 0.90          | 0.90   |          |
| LUSC   | 169                | 0.91       | 0.91          | 0.91   |          |
| SARC   | 73                 | 0.94       | 0.93          | 0.93   |          |
| STAD   | 96                 | 0.93       | 0.95          | 0.95   |          |
| NORMAL | 149                | 0.99       | 0.96          | 0.96   |          |

Таблиця 5.12

*Результати моделювання щодо використання ЗНМ для класифікації різних типів ракових захворювань при застосуванні першого кластеру у чотирикластерній структурі (41 зразок з 1144 ідентифіковані некоректно)*

| Клас   | Кількість об'єктів | Чутливість | Специфічність | F-міра | Точність |
|--------|--------------------|------------|---------------|--------|----------|
| ACC    | 24                 | 0.96       | 0.92          | 0.94   | 0.964    |
| GBM    | 73                 | 0.94       | 0.93          | 0.94   |          |
| KIRC   | 169                | 0.99       | 0.99          | 0.99   |          |
| LGG    | 191                | 0.97       | 0.97          | 0.97   |          |
| LUAD   | 200                | 0.95       | 0.96          | 0.96   |          |
| LUSC   | 169                | 0.95       | 0.95          | 0.95   |          |
| SARC   | 73                 | 0.96       | 0.94          | 0.94   |          |
| STAD   | 96                 | 0.93       | 0.95          | 0.95   |          |
| NORMAL | 149                | 0.99       | 0.99          | 0.99   |          |

За критеріями кластеризації, що наведені на рис. 5.8, оптимальними є чотири- та шестикластерні структури. За критеріями якості класифікації чотирикластерна структура є найбільш стабільною для всіх типів зразків, що досліджувалися.

Таблиця 5.13

*Результати моделювання щодо використання ЗНМ для класифікації різних типів ракових захворювань при застосуванні другого кластеру у чотирикластерній структурі (41 зразок з 1144 ідентифіковані некоректно)*

| Клас   | Кількість об'єктів | Чутливість | Специфічність | F-міра | Точність |
|--------|--------------------|------------|---------------|--------|----------|
| ACC    | 24                 | 0.88       | 0.96          | 0.92   | 0.965    |
| GBM    | 73                 | 0.97       | 0.92          | 0.94   |          |
| KIRC   | 169                | 0.99       | 0.98          | 0.98   |          |
| LGG    | 191                | 0.96       | 0.98          | 0.97   |          |
| LUAD   | 200                | 0.94       | 0.96          | 0.95   |          |
| LUSC   | 169                | 0.96       | 0.94          | 0.95   |          |
| SARC   | 73                 | 0.96       | 0.93          | 0.94   |          |
| STAD   | 96                 | 0.94       | 0.98          | 0.96   |          |
| NORMAL | 149                | 1.00       | 0.98          | 0.99   |          |

Таблиця 5.14

*Результати моделювання щодо використання ЗНМ для класифікації різних типів ракових захворювань при застосуванні третього кластеру у чотирикластерній структурі (74 зразків з 1144 ідентифіковані некоректно)*

| Клас   | Кількість об'єктів | Чутливість | Специфічність | F-міра | Точність |
|--------|--------------------|------------|---------------|--------|----------|
| ACC    | 24                 | 0.68       | 0.96          | 0.79   | 0.937    |
| GBM    | 73                 | 0.91       | 0.86          | 0.89   |          |
| KIRC   | 169                | 0.98       | 0.98          | 0.98   |          |
| LGG    | 191                | 0.94       | 0.96          | 0.95   |          |
| LUAD   | 200                | 0.94       | 0.93          | 0.93   |          |
| LUSC   | 169                | 0.92       | 0.93          | 0.92   |          |
| SARC   | 73                 | 0.93       | 0.92          | 0.92   |          |
| STAD   | 96                 | 0.94       | 0.94          | 0.94   |          |
| NORMAL | 149                | 0.97       | 0.92          | 0.94   |          |

При застосуванні шестикластерної структури п'ять кластерів дозволяють отримати високу точність класифікації за критеріями, що використовувалися, а п'ятий кластер дає суттєво нижчі результати класифікації зразків.

Таблиця 5.15

*Результати моделювання щодо використання ЗНМ для класифікації різних типів ракових захворювань при застосуванні четвертого кластеру у чотирикластерній структурі (39 зразків з 1144 ідентифіковані некоректно)*

| Клас   | Кількість об'єктів | Чутливість | Специфічність | F-міра | Точність |
|--------|--------------------|------------|---------------|--------|----------|
| ACC    | 24                 | 0.85       | 0.96          | 0.90   | 0.967    |
| GBM    | 73                 | 0.97       | 0.93          | 0.95   |          |
| KIRC   | 169                | 0.99       | 0.99          | 0.99   |          |
| LGG    | 191                | 0.97       | 0.99          | 0.98   |          |
| LUAD   | 200                | 0.94       | 0.95          | 0.95   |          |
| LUSC   | 169                | 0.96       | 0.93          | 0.94   |          |
| SARC   | 73                 | 0.97       | 0.95          | 0.96   |          |
| STAD   | 96                 | 0.95       | 0.97          | 0.96   |          |
| NORMAL | 149                | 1.00       | 0.99          | 1.00   |          |

Таблиця 5.16

*Результати моделювання щодо використання ЗНМ для класифікації різних типів ракових захворювань при застосуванні першого кластеру у п'ятикластерній структурі (44 зразка з 1144 ідентифіковані некоректно)*

| Клас   | Кількість об'єктів | Чутливість | Специфічність | F-міра | Точність |
|--------|--------------------|------------|---------------|--------|----------|
| ACC    | 24                 | 0.88       | 0.96          | 0.92   | 0.962    |
| GBM    | 73                 | 0.96       | 0.93          | 0.94   |          |
| KIRC   | 169                | 0.99       | 0.98          | 0.99   |          |
| LGG    | 191                | 0.98       | 0.98          | 0.98   |          |
| LUAD   | 200                | 0.94       | 0.94          | 0.94   |          |
| LUSC   | 169                | 0.93       | 0.93          | 0.93   |          |
| SARC   | 73                 | 0.95       | 0.96          | 0.95   |          |
| STAD   | 96                 | 0.95       | 1.00          | 0.97   |          |
| NORMAL | 149                | 1.00       | 0.97          | 0.98   |          |

Цей факт свідчить про адекватність розподілу профілів експресії генів у кластери в процесі формування кластерної структури.

Результати класифікації зразків при застосуванні чотирикластерної структури є найбільш стабільними для усіх типів даних. Однак слід зазначити, що найгірші результати класифікації зразків отримані при застосуванні зразків з ACC раковою хворобою. Це можна пояснити найменшою кількістю зразків, порівняно з іншими типами даних, що, безумовно, може вплинути на чутливість моделі.

Таблиця 5.17

*Результати моделювання щодо використання ЗНМ для класифікації різних типів ракових захворювань при застосуванні другого кластеру у п'ятикластерній структурі (49 зразків з 1144 ідентифіковані некоректно)*

| Клас   | Кількість об'єктів | Чутливість | Специфічність | F-міра | Точність |
|--------|--------------------|------------|---------------|--------|----------|
| ACC    | 24                 | 0.79       | 0.92          | 0.85   | 0.957    |
| GBM    | 73                 | 0.99       | 0.92          | 0.95   |          |
| KIRC   | 169                | 0.98       | 0.97          | 0.98   |          |
| LGG    | 191                | 0.97       | 0.98          | 0.98   |          |
| LUAD   | 200                | 0.93       | 0.96          | 0.95   |          |
| LUSC   | 169                | 0.97       | 0.93          | 0.95   |          |
| SARC   | 73                 | 0.92       | 0.89          | 0.90   |          |
| STAD   | 96                 | 0.93       | 0.96          | 0.94   |          |
| NORMAL | 149                | 0.99       | 0.99          | 0.99   |          |

Найкращі результати отримані при класифікації зразків, для яких ракова пухлина не виявлена (NORMAL). Це підтверджує високі результати бінарної класифікації, що отримані при застосуванні алгоритму випадкового лісу для повного набору профілів експресії генів.

Таблиця 5.18

Результати моделювання щодо використання ЗНМ для класифікації різних типів ракових захворювань при застосуванні третього кластеру у п'ятикластерній структурі (49 зразків з 1144 ідентифіковані некоректно)

| Клас   | Кількість об'єктів | Чутливість | Специфічність | F-міра | Точність |
|--------|--------------------|------------|---------------|--------|----------|
| ACC    | 24                 | 0.88       | 0.92          | 0.90   | 0.958    |
| GBM    | 73                 | 0.96       | 0.90          | 0.93   |          |
| KIRC   | 169                | 0.99       | 0.99          | 0.99   |          |
| LGG    | 191                | 0.97       | 0.98          | 0.98   |          |
| LUAD   | 200                | 0.94       | 0.94          | 0.94   |          |
| LUSC   | 169                | 0.93       | 0.93          | 0.93   |          |
| SARC   | 73                 | 0.96       | 0.93          | 0.94   |          |
| STAD   | 96                 | 0.93       | 0.97          | 0.95   |          |
| NORMAL | 149                | 0.98       | 0.98          | 0.98   |          |

Таблиця 5.19

Результати моделювання щодо використання ЗНМ для класифікації різних типів ракових захворювань при застосуванні четвертого кластеру у п'ятикластерній структурі (41 зразок з 1144 ідентифіковані некоректно)

| Клас   | Кількість об'єктів | Чутливість | Специфічність | F-міра | Точність |
|--------|--------------------|------------|---------------|--------|----------|
| ACC    | 24                 | 0.74       | 0.96          | 0.84   | 0.968    |
| GBM    | 73                 | 1.00       | 0.92          | 0.96   |          |
| KIRC   | 169                | 0.98       | 0.98          | 0.98   |          |
| LGG    | 191                | 0.97       | 0.99          | 0.98   |          |
| LUAD   | 200                | 0.95       | 0.94          | 0.94   |          |
| LUSC   | 169                | 0.95       | 0.94          | 0.95   |          |
| SARC   | 73                 | 0.99       | 0.95          | 0.97   |          |
| STAD   | 96                 | 0.96       | 1.00          | 0.98   |          |
| NORMAL | 149                | 1.00       | 0.99          | 0.99   |          |

Таблиця 5.20

Результати моделювання щодо використання ЗНМ для класифікації різних типів ракових захворювань при застосуванні п'ятого кластеру у п'ятикластерній структурі (116 зразків з 1144 ідентифіковані некоректно)

| Клас   | Кількість об'єктів | Чутливість | Специфічність | F-міра | Точність |
|--------|--------------------|------------|---------------|--------|----------|
| ACC    | 24                 | 0.57       | 0.88          | 0.69   | 0.904    |
| GBM    | 73                 | 0.85       | 0.78          | 0.81   |          |
| KIRC   | 169                | 0.95       | 0.95          | 0.95   |          |
| LGG    | 191                | 0.92       | 0.96          | 0.94   |          |
| LUAD   | 200                | 0.88       | 0.89          | 0.88   |          |
| LUSC   | 169                | 0.90       | 0.90          | 0.90   |          |
| SARC   | 73                 | 0.87       | 0.89          | 0.88   |          |
| STAD   | 96                 | 0.92       | 0.80          | 0.86   |          |
| NORMAL | 149                | 0.94       | 0.91          | 0.92   |          |

Таблиця 5.21

*Результати моделювання щодо використання ЗНМ для класифікації різних типів ракових захворювань при застосуванні першого кластеру у шестикластерній структурі (44 зразка з 1144 ідентифіковані некоректно)*

| Клас   | Кількість об'єктів | Чутливість | Специфічність | F-міра | Точність |
|--------|--------------------|------------|---------------|--------|----------|
| ACC    | 24                 | 0.92       | 0.96          | 0.94   | 0.962    |
| GBM    | 73                 | 0.96       | 0.90          | 0.93   |          |
| KIRC   | 169                | 0.99       | 0.99          | 0.99   |          |
| LGG    | 191                | 0.97       | 0.98          | 0.97   |          |
| LUAD   | 200                | 0.94       | 0.94          | 0.94   |          |
| LUSC   | 169                | 0.94       | 0.93          | 0.94   |          |
| SARC   | 73                 | 0.97       | 0.97          | 0.97   |          |
| STAD   | 96                 | 0.93       | 0.99          | 0.96   |          |
| NORMAL | 149                | 0.99       | 0.97          | 0.98   |          |

Таблиця 5.22

*Результати моделювання щодо використання ЗНМ для класифікації різних типів ракових захворювань при застосуванні другого кластеру у шестикластерній структурі (42 зразка з 1144 ідентифіковані некоректно)*

| Клас   | Кількість об'єктів | Чутливість | Специфічність | F-міра | Точність |
|--------|--------------------|------------|---------------|--------|----------|
| ACC    | 24                 | 0.82       | 0.96          | 0.88   | 0.965    |
| GBM    | 73                 | 0.99       | 0.92          | 0.95   |          |
| KIRC   | 169                | 0.99       | 0.96          | 0.98   |          |
| LGG    | 191                | 0.97       | 0.98          | 0.98   |          |
| LUAD   | 200                | 0.95       | 0.94          | 0.95   |          |
| LUSC   | 169                | 0.95       | 0.95          | 0.95   |          |
| SARC   | 73                 | 0.92       | 0.96          | 0.94   |          |
| STAD   | 96                 | 0.95       | 0.98          | 0.96   |          |
| NORMAL | 149                | 1.00       | 0.99          | 0.99   |          |

Отримані результати свідчать також про доцільність застосування системи альтернативно голосування на останньому кроці при прийнятті остаточного рішення щодо наявності або відсутності відповідної хвороби, оскільки рішення на основі застосування однієї множини даних експресії генів не є об'єктивним.

Таблиця 5.23

Результати моделювання щодо використання ЗНМ для класифікації різних типів ракових захворювань при застосуванні третього кластеру у шестикластерній структурі (54 зразка з 1144 ідентифіковані некоректно)

| Клас   | Кількість об'єктів | Чутливість | Специфічність | F-міра | Точність |
|--------|--------------------|------------|---------------|--------|----------|
| ACC    | 24                 | 0.92       | 0.96          | 0.94   | 0.953    |
| GBM    | 73                 | 0.94       | 0.85          | 0.89   |          |
| KIRC   | 169                | 0.98       | 0.98          | 0.98   |          |
| LGG    | 191                | 0.95       | 0.97          | 0.96   |          |
| LUAD   | 200                | 0.95       | 0.94          | 0.95   |          |
| LUSC   | 169                | 0.96       | 0.94          | 0.95   |          |
| SARC   | 73                 | 0.91       | 0.93          | 0.92   |          |
| STAD   | 96                 | 0.92       | 1.00          | 0.96   |          |
| NORMAL | 149                | 0.99       | 0.95          | 0.97   |          |

Таблиця 5.24

Результати моделювання щодо використання ЗНМ для класифікації різних типів ракових захворювань при застосуванні четвертого кластеру у шестикластерній структурі (49 зразків з 1144 ідентифіковані некоректно)

| Клас   | Кількість об'єктів | Чутливість | Специфічність | F-міра | Точність |
|--------|--------------------|------------|---------------|--------|----------|
| ACC    | 24                 | 0.88       | 0.96          | 0.92   | 0.957    |
| GBM    | 73                 | 0.93       | 0.90          | 0.92   |          |
| KIRC   | 169                | 0.98       | 0.99          | 0.99   |          |
| LGG    | 191                | 0.97       | 0.97          | 0.97   |          |
| LUAD   | 200                | 0.94       | 0.95          | 0.95   |          |
| LUSC   | 169                | 0.95       | 0.93          | 0.94   |          |
| SARC   | 73                 | 0.97       | 0.93          | 0.95   |          |
| STAD   | 96                 | 0.92       | 0.95          | 0.93   |          |
| NORMAL | 149                | 0.99       | 0.99          | 0.99   |          |

Таблиця 5.25

Результати моделювання щодо використання ЗНМ для класифікації різних типів ракових захворювань при застосуванні п'ятого кластеру у шестикластерній структурі (195 зразків з 1144 ідентифіковані некоректно)

| Клас   | Кількість об'єктів | Чутливість | Специфічність | F-міра | Точність |
|--------|--------------------|------------|---------------|--------|----------|
| ACC    | 24                 | 0.45       | 0.62          | 0.53   | 0.834    |
| GBM    | 73                 | 0.85       | 0.68          | 0.76   |          |
| KIRC   | 169                | 0.90       | 0.89          | 0.89   |          |
| LGG    | 191                | 0.90       | 0.92          | 0.91   |          |
| LUAD   | 200                | 0.77       | 0.82          | 0.79   |          |
| LUSC   | 169                | 0.79       | 0.86          | 0.82   |          |
| SARC   | 73                 | 0.82       | 0.81          | 0.81   |          |
| STAD   | 96                 | 0.80       | 0.80          | 0.80   |          |
| NORMAL | 149                | 0.90       | 0.77          | 0.84   |          |

Таблиця 5.26

Результати моделювання щодо використання ЗНМ для класифікації різних типів ракових захворювань при застосуванні шостого кластеру у шестикластерній структурі (47 зразків з 1144 ідентифіковані некоректно)

| Клас   | Кількість об'єктів | Чутливість | Специфічність | F-міра | Точність |
|--------|--------------------|------------|---------------|--------|----------|
| ACC    | 24                 | 0.96       | 0.92          | 0.94   | 0.959    |
| GBM    | 73                 | 0.94       | 0.92          | 0.93   |          |
| KIRC   | 169                | 0.98       | 0.98          | 0.98   |          |
| LGG    | 191                | 0.96       | 0.97          | 0.97   |          |
| LUAD   | 200                | 0.95       | 0.95          | 0.95   |          |
| LUSC   | 169                | 0.95       | 0.94          | 0.94   |          |
| SARC   | 73                 | 0.97       | 0.92          | 0.94   |          |
| STAD   | 96                 | 0.90       | 0.99          | 0.95   |          |
| NORMAL | 149                | 1.00       | 0.98          | 0.99   |          |

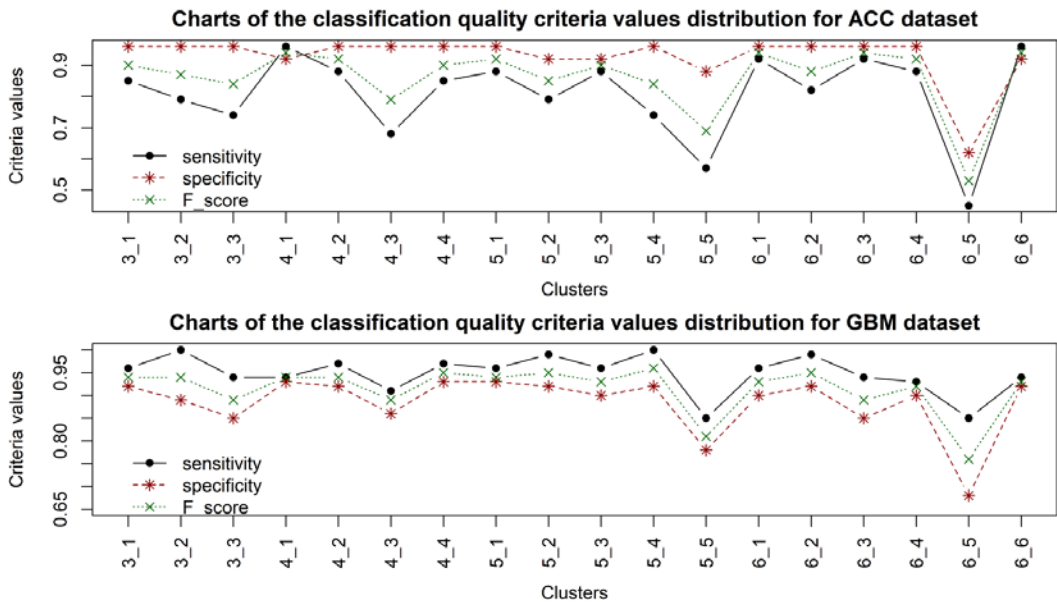


Рис. 5.11. Діаграми розподілу значень критеріїв якості класифікації зразків пацієнтів, що мають ACC та GBM ракові хвороби при застосуванні даних експресії генів різних кластерів

#### 5.4. Застосування системи альтернативного голосування для прийняття остаточного рішення щодо стану пацієнтів, що досліджуються

Доцільність застосування системи альтернативного голосування на останньому кроці прийняття остаточного рішення щодо стану здоров'я пацієнта визначається тим фактом, що результати класифікації, які отримані із застосування різних

підмножин профілів експресії генів, що сформовані в процесі реалізації процедури кластеризації, можуть суперечити між собою. Реалізація процедури кластеризації передбачає формування підмножин взаємоподібних за відповідними метриками профілів експресії генів, при цьому можливе формування кластерів профілів експресії генів, що визначають процеси функціонування біологічного організму, які не мають прямого відношення до відповідної хвороби.

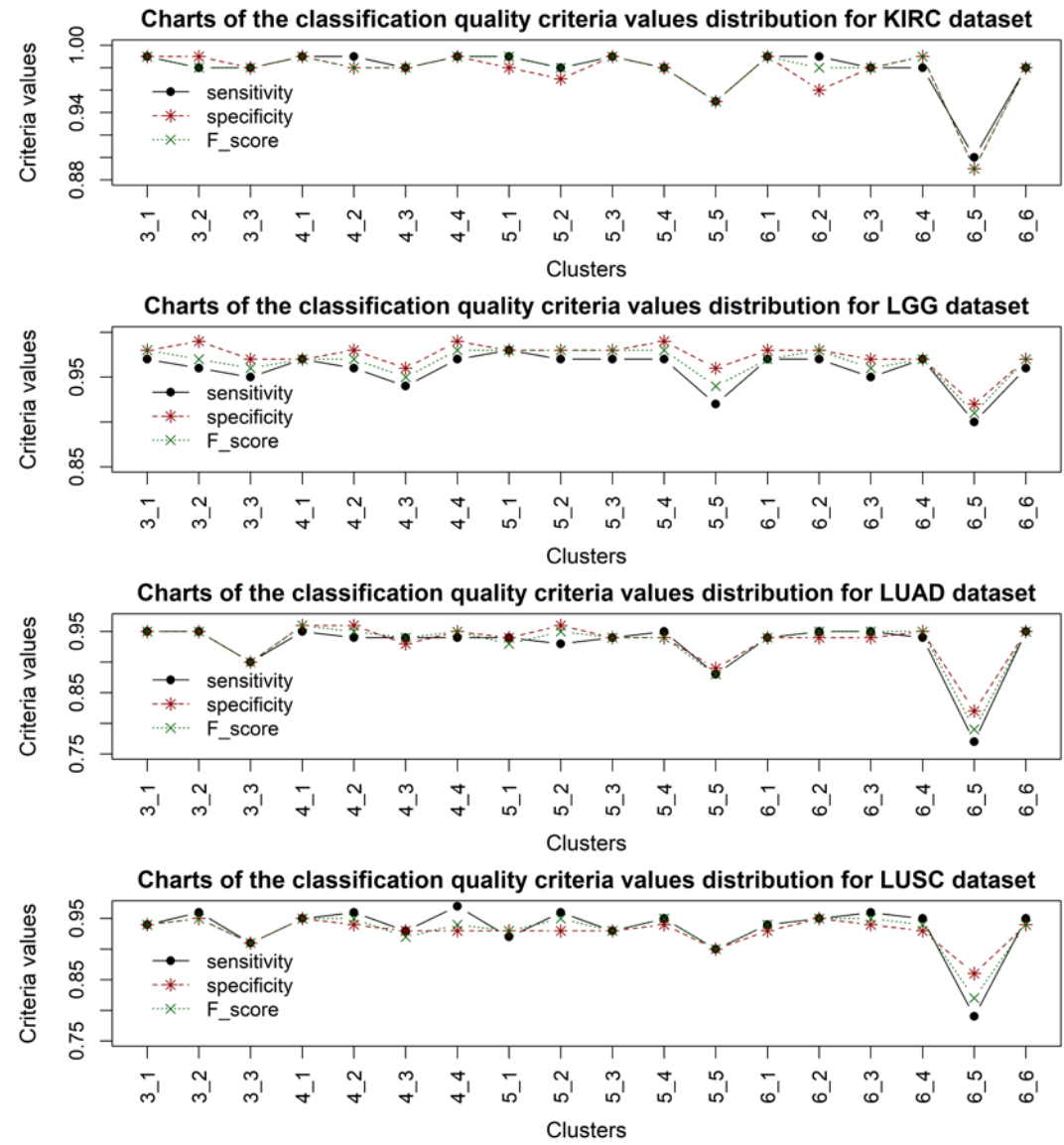


Рис. 5.12. Діаграма розподілу значень критеріїв якості класифікації зразків пацієнтів, що мають KIRC, LGG, LUAD та LUSC ракові хвороби при застосуванні даних експресії генів різних кластерів

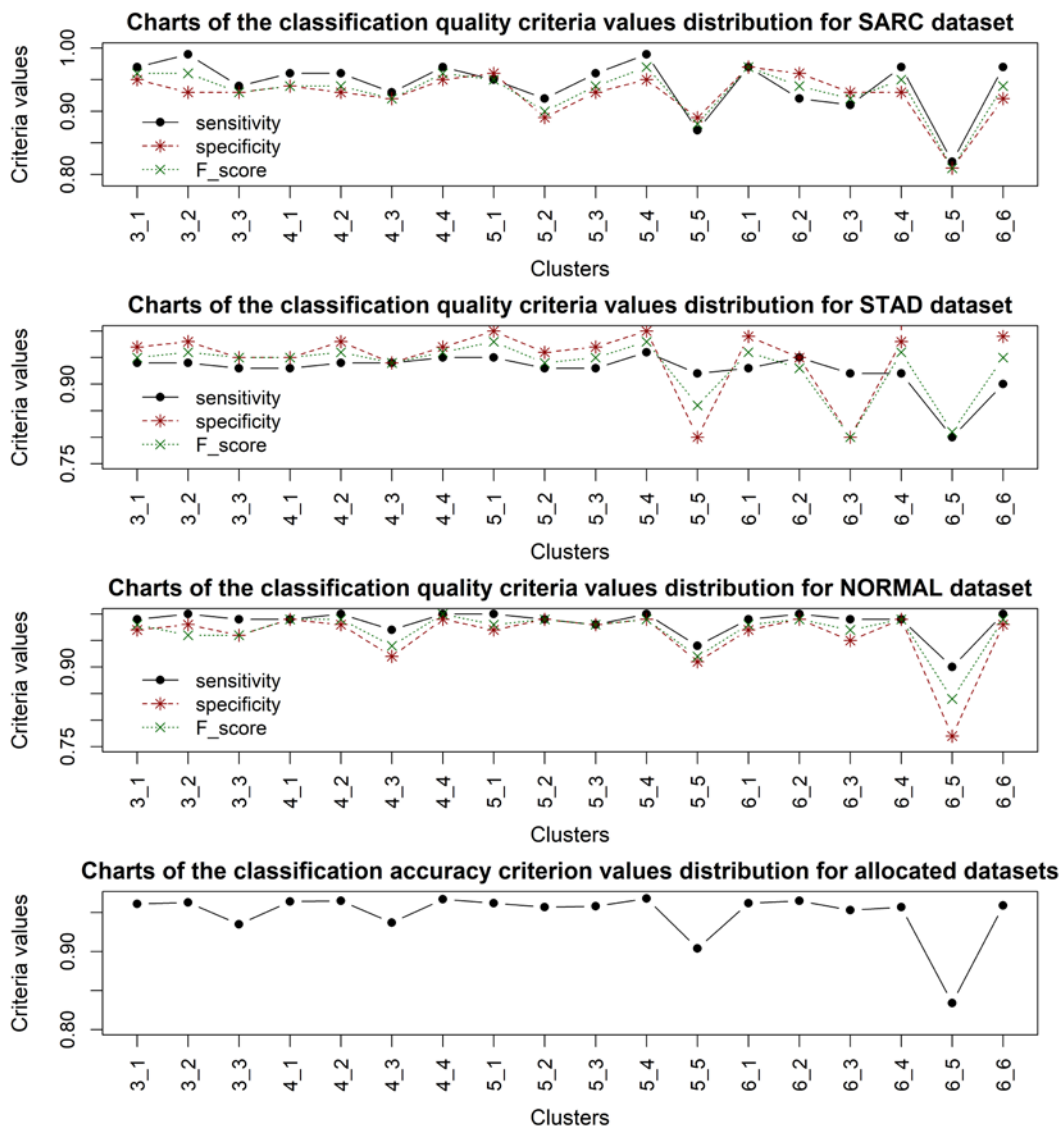


Рис. 5.13. Діаграма розподілу значень критеріїв якості класифікації зразків пацієнтів, що мають SARC, STAD ракові хвороби, здорових пацієнтів (NORMAL), та точності класифікації зразків тестової підмножини даних при застосуванні даних експресії генів різних кластерів

За результатами моделювання, що наведені у попередньому підрозділі, при застосуванні кожної кластерної структури у більшості випадків результати класифікації зразків, що містять як атрибути сизначення експресії генів одного з кластерів, є більш низькі, порівняно з результатами клафікації із застосуванням даних інших кластерів. Це третій кластер у три- та чотирикластерних структурах та п'ятий кластер у п'яти- та шестикластерних структурах.

Очевидно, що цей кластер можна не розглядати в процесі прийняття остаточного рішення щодо стану пацієнта, але у цьому випадку виникає необхідність проведення глибокого кластерного аналізу кожному випадку при аналізі нових даних.

До того ж при застосуванні три- та чотирікластерних структур результати класифікації, що отримані на даних експресії генів третього кластера є також досить високими, і вони можуть вплинути на прийняття остаточного рішення щодо наявності або відсутності хвороби.

Застосування системи альтернативного голосування передбачає застосування результатів класифікації, отриманих на всіх кластерах, при цьому зникає необхідність вибору найбільш інформативних кластерів за розділяючою здатністю відповідної хвороби. Рішення приймається за узгодженістю результатів класифікації на основі даних більшості кластерів. При цьому можливі два випадки:

1. Зразок належить до відповідного класу (однозначно ідентифіковано ракову хворобу відповідного типу або пацієнта визнано здоровим).

2. Пацієнт потребує додаткового обстеження щодо визначення його стану здоров'я.

Так, при застосуванні експериментальних даних експресії генів пацієнтів, що використовуються в рамках поточного моделювання, при використанні даних трикластерної структури, об'єкт відносять до відповідного класу, якщо результати класифікації збігаються при застосуванні даних двох будь-яких кластерів з трьох. При застосуванні чотири- та п'ятикластерних структур результати мають збігатися на основі даних експресії генів трьох будь-яких кластерів. При застосуванні шестикластерної структури кількість таких кластерів повинно бути не менше ніж чотири. При невиконанні вищезазначених умов (класифікатор ідентифікує різні стани пацієнта (різні типи раку або відсутність хвороби (NORMAL)) на основі даних експресії генів різних кластерів) стан пацієнта ідентифікується як невизначений, що потребує подальших клінічних досліджень.

Слід зазначити, що збільшення кількості кластерів сприяє підвищенню об'єктивності прийняття адекватного рішення щодо стану здоров'я пацієнтів за рахунок розпаралелювання процесу обробки даних експресії генів, з одного боку, і підвищення кількості можливих альтернатив, з іншого боку.

#### **5.4.1. Практична реалізація моделі бінарної класифікації стану здоров'я пацієнтів на основі мультикластерної структури даних експресії генів із застосуванням методу альтернативного голосування**

Результати бінарної класифікації зразків, що містять як атрибути дані експресії генів при застосуванні різних кластерних структур наведені у таблицях 5.27–5.30 та на рис. 5.14. На цьому етапі застосовувався алгоритм класифікації Random Forest (RF). Доцільність застосування цього типу алгоритму визначається тим фактом, що при бінарній класифікації початкові дані розділялися на підгрупи, при цьому кількість зразків у кожній групі (з урахуванням типу хвороби) є суттєво меншою загальної кількості зразків. Для таких обмежених даних застосування згорткової

нейронної мережі не є доцільним. Окрім того, доцільність застосування алгоритму RF для бінарної класифікації зразків на основі даних експресії генів доведена у праці [207].

Процедура моделювання здійснювалася шляхом застосування таких кроків:

1. Формування підмножин даних, зразки яких відповідають одному типу для кожної кластерної структури. На цьому етапі було сформовано дев'ять підмножин, вісім з яких відповідали різним типам ракових захворювань, а дев'ята підмножина містила дані експресії генів пацієнтів, для яких хвороба не виявлена.

2. Об'єднання підмножин даних експресії генів, що відповідають різним типам захворювань, з даними експресії генів здорових пацієнтів. На цьому етапі формуються вісім підмножин даних експресії генів для кожної кластерної структури.

3. Розділення кожної групи даних на дві підмножини у відношенні 0.65/0.35, перша з яких використовується для навчання моделі, а друга – для її тестування.

4. Навчання моделі шляхом застосування алгоритму RF до першої підмножини даних.

5. Застосування моделі до тестових даних. Розрахунок критеріїв якості бінарної класифікації.

6. Аналіз отриманих результатів.

*Таблиця 5.27*

*Результати бінарної класифікації зразків на основі даних експресії генів при застосуванні трикластерної структури даних*

| Тип зразків | Критерії якості класифікації |            |               |        |
|-------------|------------------------------|------------|---------------|--------|
|             | Точність                     | Чутливість | Специфічність | F-міра |
| ACC         | 0.990                        | 1.000      | 0.963         | 0.981  |
| NORM_ACC    | 0.990                        | 0.987      | 1.000         | 0.994  |
| GBM         | 0.993                        | 0.983      | 1.000         | 0.992  |
| NORM_GBM    | 0.993                        | 1.000      | 0.987         | 0.994  |
| SARC        | 1.000                        | 1.000      | 1.000         | 1.000  |
| NORM_SARC   | 1.000                        | 1.000      | 1.000         | 1.000  |
| LUSC        | 0.996                        | 1.000      | 0.994         | 0.997  |
| NORM_LUSC   | 0.996                        | 0.987      | 1.000         | 0.994  |
| LUAD        | 0.996                        | 0.995      | 1.000         | 0.997  |
| NORM_LUAD   | 0.996                        | 1.000      | 0.987         | 0.994  |
| STAD        | 0.987                        | 0.975      | 0.987         | 0.981  |
| NORM_STAD   | 0.987                        | 0.993      | 0.986         | 0.990  |
| KIRC        | 0.996                        | 0.995      | 1.000         | 0.997  |
| NORM_KIRC   | 0.996                        | 1.000      | 0.987         | 0.994  |
| LGG         | 0.992                        | 0.995      | 0.995         | 0.995  |
| NORM_LGG    | 0.992                        | 0.987      | 0.987         | 0.987  |

Таблиця 5.28

*Результати бінарної класифікації зразків на основі даних експресії генів при застосуванні чотирикластерної структури даних*

| Тип зразків | Критерії якості класифікації |            |               |        |
|-------------|------------------------------|------------|---------------|--------|
|             | Точність                     | Чутливість | Специфічність | F-міра |
| ACC         | 0.990                        | 1.000      | 0.963         | 0.981  |
| NORM_ACC    | 0.990                        | 0.987      | 1.000         | 0.994  |
| GBM         | 1.000                        | 1.000      | 1.000         | 1.000  |
| NORM_GBM    | 1.000                        | 1.000      | 1.000         | 1.000  |
| SARC        | 1.000                        | 1.000      | 1.000         | 1.000  |
| NORMP_SARC  | 1.000                        | 1.000      | 1.000         | 1.000  |
| LUSC        | 1.000                        | 1.000      | 1.000         | 1.000  |
| NORM_LUSC   | 1.000                        | 1.000      | 1.000         | 1.000  |
| LUAD        | 0.992                        | 0.990      | 1.000         | 0.995  |
| NORM_LUAD   | 0.992                        | 1.000      | 0.974         | 0.987  |
| STAD        | 0.991                        | 0.975      | 1.000         | 0.987  |
| NORM_STAD   | 0.991                        | 1.000      | 0.986         | 0.993  |
| KIRC        | 1.000                        | 1.000      | 1.000         | 1.000  |
| NORM_KIRC   | 1.000                        | 1.000      | 1.000         | 1.000  |
| LGG         | 0.992                        | 0.995      | 0.995         | 0.995  |
| NORM_LGG    | 0.992                        | 0.987      | 0.987         | 0.987  |

Таблиця 5.29

*Результати бінарної класифікації зразків на основі даних експресії генів при застосуванні п'ятикластерної структури даних*

| Тип зразків | Критерії якості класифікації |            |               |        |
|-------------|------------------------------|------------|---------------|--------|
|             | Точність                     | Чутливість | Специфічність | F-міра |
| ACC         | 0.990                        | 1.000      | 0.963         | 0.981  |
| NORM_ACC    | 0.990                        | 0.987      | 1.000         | 0.994  |
| GBM         | 1.000                        | 1.000      | 1.000         | 1.000  |
| NORM_GBM    | 1.000                        | 1.000      | 1.000         | 1.000  |
| SARC        | 1.000                        | 1.000      | 1.000         | 1.000  |
| NORM_SARC   | 1.000                        | 1.000      | 1.000         | 1.000  |
| LUSC        | 1.000                        | 1.000      | 1.000         | 1.000  |
| NORM_LUSC   | 1.000                        | 1.000      | 1.000         | 1.000  |
| LUAD        | 0.993                        | 0.990      | 1.000         | 0.995  |
| NORM_LUAD   | 0.993                        | 1.000      | 0.974         | 0.987  |
| STAD        | 0.995                        | 0.987      | 1.000         | 0.994  |
| NORM_STAD   | 0.995                        | 1.000      | 0.993         | 0.997  |
| KIRC        | 0.996                        | 0.995      | 1.000         | 0.997  |
| NORM_KIRC   | 0.996                        | 1.000      | 0.987         | 0.994  |
| LGG         | 0.992                        | 0.995      | 0.995         | 0.995  |
| NORM_LGG    | 0.992                        | 0.987      | 0.987         | 0.987  |

*Результати бінарної класифікації зразків на основі даних експресії генів при застосуванні шестикластерної структури даних*

| Тип зразків | Критерії якості класифікації |            |               |        |
|-------------|------------------------------|------------|---------------|--------|
|             | Точність                     | Чутливість | Специфічність | F-міра |
| ACC         | 1.000                        | 1.000      | 1.000         | 1.000  |
| NORM_ACC    | 1.000                        | 1.000      | 1.000         | 1.000  |
| GBM         | 1.000                        | 1.000      | 1.000         | 1.000  |
| NORM_GBM    | 1.000                        | 1.000      | 1.000         | 1.000  |
| SARC        | 1.000                        | 1.000      | 1.000         | 1.000  |
| NORM_SARC   | 1.000                        | 1.000      | 1.000         | 1.000  |
| LUSC        | 0.996                        | 1.000      | 0.994         | 0.997  |
| NORM_LUSC   | 0.996                        | 0.987      | 1.000         | 0.994  |
| LUAD        | 0.996                        | 0.995      | 1.000         | 0.997  |
| NORM_LUAD   | 0.996                        | 1.000      | 0.987         | 0.993  |
| STAD        | 0.991                        | 0.975      | 1.000         | 0.987  |
| NORM_STAD   | 0.991                        | 1.000      | 0.986         | 0.993  |
| KIRC        | 1.000                        | 1.000      | 1.000         | 1.000  |
| NORM_KIRC   | 1.000                        | 1.000      | 1.000         | 1.000  |
| LGG         | 1.000                        | 1.000      | 1.000         | 1.000  |
| NORM_LGG    | 1.000                        | 1.000      | 1.000         | 1.000  |

На основі аналізу отриманих результатів можна зробити висновок, що в усіх випадках точність бінарної класифікації зразків є досить високою (варіюється від 98 до 100%). Детальніший аналіз показує більш високу ефективність моделі при застосуванні чотири- та шестикластерних структур, що узгоджується з результатами оцінки якості кластерної структури на основі застосування внутрішніх та зовнішніх критеріїв якості кластеризації, які наведені у розділі 4 цієї монографії. Аналіз результатів моделювання дозволяє також зробити висновок, що найкращі результати отримані при застосуванні шестикластерної структури, що підтверджує доцільність застосування моделі розпаралелювання потоку обробки даних експресії генів з подальшим прийняттям остаточного рішення щодо стану об'єкта шляхом застосування методу альтернативного голосування.

При цьому слід зазначити важливість попереднього етапу видалення неінформативних за статистичними та ентропійними критеріями профілів експресії генів і формування кластерів взаємно корельованих та диференційно експресованих профілів експресії генів, що підтверджує значущість досліджень, представлених у попередніх розділах монографії.

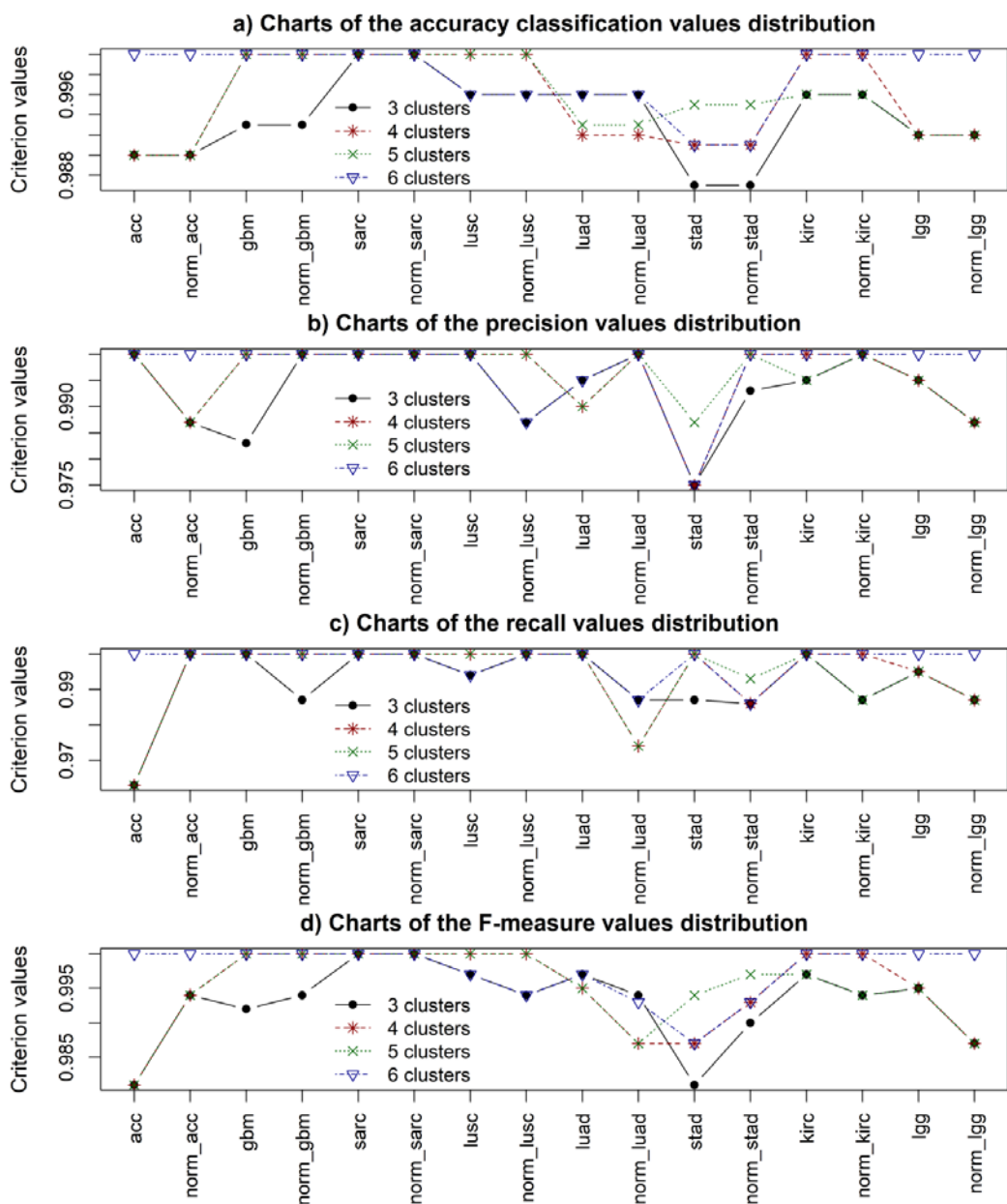


Рис. 5.14. Результати моделювання щодо застосування алгоритму Random Forest (RF) для бінарної класифікації зразків, що містять як атрибути дані експресії генів різних кластерних структур

#### **5.4.2. Практична реалізація мультикласової моделі ідентифікації стану здоров'я пацієнтів на основі мультикластерної структури даних експресії генів із застосуванням методу альтернативного голосування**

Практична реалізація мультикласової моделі ідентифікації стану здоров'я пацієнтів (ідентифікація наявності або відсутності відповідного типу раку) на основі даних, що представлені у таблиці 5.1 здійснювалася шляхом застосування двох типів класифікаторів: класифікатора на основі алгоритму RF та згорткової нейронної мережі.

Процедура реалізації цього процесу передбачала наявність таких кроків:

1. Формування експериментальних даних для кожної кластерної структури у вигляді матриці, де рядки є зразками, що досліджуються, а стовпці – значення експресії генів профілів відповідних кластерів. При використанні даних таблиці 5.1 загальна кількість зразків дорівнює 3269.

2. Розділення даних на дві підмножини також у відношенні 0.65/0.35.

3. Застосування класифікатора до даних, що використовуються для навчання моделі (65% від загальної кількості зразків).

4. Тестування моделі шляхом застосування тестових даних. Формування узагальнених результатів класифікації по кожній кластерній структурі у вигляді матриці, де рядки є зразки, що складають тестову підмножину даних, а стовпці – результат класифікації по кожному кластеру поточної кластерної структури.

5. Формування остаточного результату класифікації шляхом застосування методу альтернативного голосування до сформованої матриці результатів. Відповідно до процедури, що описана у підрозділі 5.4, ідентифікація зразка здійснюється на основі узгодженості результатів класифікації із застосуванням даних експресії генів більшості кластерів. При наявності розбіжності результатів класифікації зразок ідентифікується як невизначений (undefined), що вимагає подальших клінічних досліджень щодо його об'єктивної ідентифікації.

6. Розрахунок критеріїв якості класифікації на основі застосування однозначно ідентифікованих зразків.

7. Аналіз отриманих результатів.

##### **5.4.2.1. Результати діагностики стану пацієнтів на основі застосування алгоритму Random Forest**

У таблицях 5.31–5.34 та на рис. 5.15 представлені результати моделювання щодо застосування алгоритму RF до даних експресії генів різних кластерних структур. Як було зазначено вище, оцінка результатів класифікації здійснювалася на тестових наборах даних. Аналіз отриманих результатів дозволяє зробити висновок, що за критеріями якості класифікації чотири- та шестикластерні структури є більш привабливими, порівняно з три- та п'ятикластерними структурами, що узгоджується з результатами, які отримані при застосуванні алгоритму спектральної кластеризації у розділі 5.2 цієї монографії (рис. 5.8). Однак слід зазначити що із 1140 зразків, що використовувалися для тестування моделі, стан 31 та 21 зразків у випадку застосування чотири- та шестикластерних структур, відповідно, був ідентифікований як невизначений. Це означає, що у цих пацієнтів тип раку не ідентифікований однозначно, що потребує подальших клінічних досліджень для однозначної ідентифікації типу захворювання.

Таблиця 5.31

Результати застосування алгоритму RF до тестової підмножини зразків при використанні трикластерної структури даних експресії генів

| Тип зразка | Загальна кількість зразків | Кількість ідентифікованих зразків | Критерії класифікації |            |               |        |
|------------|----------------------------|-----------------------------------|-----------------------|------------|---------------|--------|
|            |                            |                                   | Точність              | Чутливість | Специфічність | F-міра |
| ACC        | 1140                       | 1136                              | 0.977                 | 1.000      | 0.962         | 0.980  |
| GBM        |                            |                                   |                       | 0.914      | 0.914         | 0.914  |
| KIRC       |                            |                                   |                       | 0.989      | 0.995         | 0.992  |
| LGG        |                            |                                   |                       | 0.957      | 0.973         | 0.965  |
| LUAD       |                            |                                   |                       | 0.968      | 0.968         | 0.968  |
| LUSC       |                            |                                   |                       | 0.971      | 0.966         | 0.968  |
| NORM       |                            |                                   |                       | 0.973      | 0.910         | 0.940  |
| SARC       |                            |                                   |                       | 0.989      | 1.000         | 0.995  |
| STAD       |                            |                                   |                       | 0.986      | 1.000         | 0.993  |

Таблиця 5.32

Результати застосування алгоритму RF до тестової підмножини зразків при використанні чотирікластерної структури даних експресії генів

| Тип зразка | Загальна кількість зразків | Кількість ідентифікованих зразків | Критерії класифікації |            |               |        |
|------------|----------------------------|-----------------------------------|-----------------------|------------|---------------|--------|
|            |                            |                                   | Точність              | Чутливість | Специфічність | F-міра |
| ACC        | 1140                       | 1109                              | 0.99                  | 1.000      | 0.962         | 0.980  |
| GBM        |                            |                                   |                       | 1.000      | 1.000         | 1.000  |
| KIRC       |                            |                                   |                       | 0.995      | 1.000         | 0.997  |
| LGG        |                            |                                   |                       | 0.984      | 1.000         | 0.992  |
| LUAD       |                            |                                   |                       | 0.978      | 0.989         | 0.984  |
| LUSC       |                            |                                   |                       | 0.988      | 0.977         | 0.983  |
| NORM       |                            |                                   |                       | 1.000      | 0.946         | 0.972  |
| SARC       |                            |                                   |                       | 0.989      | 1.000         | 0.995  |
| STAD       |                            |                                   |                       | 1.000      | 1.000         | 1.000  |

Таблиця 5.33

Результати застосування алгоритму RF до тестової підмножини зразків при використанні n'ятикластерної структури даних експресії генів

| Тип зразка | Загальна кількість зразків | Кількість ідентифікованих зразків | Критерії класифікації |            |               |        |
|------------|----------------------------|-----------------------------------|-----------------------|------------|---------------|--------|
|            |                            |                                   | Точність              | Чутливість | Специфічність | F-міра |
| ACC        | 1140                       | 1130                              | 0.981                 | 1.000      | 0.926         | 0.962  |
| GBM        |                            |                                   |                       | 0.947      | 0.982         | 0.964  |
| KIRC       |                            |                                   |                       | 0.989      | 0.995         | 0.992  |
| LGG        |                            |                                   |                       | 0.978      | 0.978         | 0.978  |
| LUAD       |                            |                                   |                       | 0.984      | 0.973         | 0.978  |
| LUSC       |                            |                                   |                       | 0.977      | 0.988         | 0.983  |
| NORM       |                            |                                   |                       | 0.973      | 0.910         | 0.940  |
| SARC       |                            |                                   |                       | 0.968      | 1.000         | 0.984  |
| STAD       |                            |                                   |                       | 0.993      | 1.000         | 0.997  |

Таблиця 5.34

Результати застосування алгоритму RF до тестової підмножини зразків при використанні шестикластерної структури даних експресії генів

| Тип зразка | Загальна кількість зразків | Кількість ідентифікованих зразків | Критерії класифікації |            |               |        |
|------------|----------------------------|-----------------------------------|-----------------------|------------|---------------|--------|
|            |                            |                                   | Точність              | Чутливість | Специфічність | F-міра |
| ACC        | 1140                       | 1119                              | 0.985                 | 1.000      | 0.926         | 0.962  |
| GBM        |                            |                                   |                       | 1.000      | 1.000         | 1.000  |
| KIRC       |                            |                                   |                       | 0.989      | 1.000         | 0.995  |
| LGG        |                            |                                   |                       | 0.984      | 0.995         | 0.989  |
| LUAD       |                            |                                   |                       | 0.984      | 0.973         | 0.978  |
| LUSC       |                            |                                   |                       | 0.983      | 0.983         | 0.983  |
| NORM       |                            |                                   |                       | 0.971      | 0.918         | 0.944  |
| SARC       |                            |                                   |                       | 0.968      | 1.000         | 0.984  |
| STAD       |                            |                                   |                       | 0.993      | 1.000         | 0.997  |

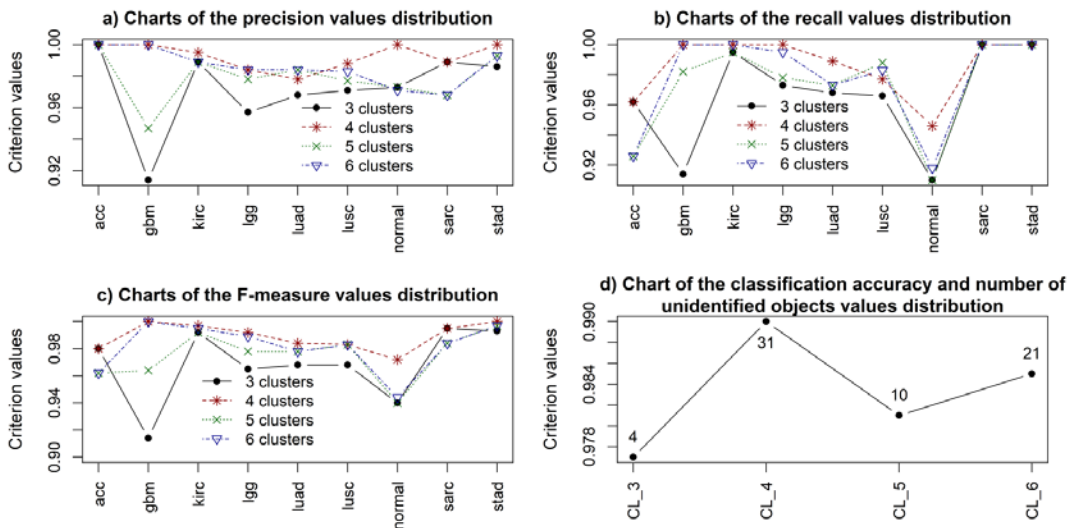


Рис. 5.15. Результати моделювання щодо застосування алгоритму Random Forest для розв'язання мультикласової задачі ідентифікації стану пацієнтів, що досліджуються на різні типи раку

Проаналізувавши результати моделювання, бачимо найвищу ефективність застосування чотирикластерної структури для ідентифікації типу ракового захворювання. Аналіз значень критеріїв якості класифікації зразків, що відповідають різним типам раку та здоровому стану пацієнтів (рис. 5.15a,b,c) свідчить про більш високе значення та більшу стабільність цих критеріїв при застосуванні чотирикластерної структури даних, порівняно із застосуванням інших кластерних структур. Однак слід зазначити, що при застосуванні чотирикластерної структури стан відносно великої кількості об'єктів ідентифікований як невизначений (31 із 1140), що, без сумнівів, є недоліком моделі на основі алгоритму RF. До другого суттєвого недоліку слід віднести велику

кількість часу для навчання моделі. На поточних даних навчання кожної моделі (при застосуванні кожної кластерної структури) здійснювалося протягом приблизно 80 годин.

5.4.2.2. Результати діагностики стану пацієнтів на основі застосування згорткової нейронної мережі

При поточному моделюванні щодо ідентифікації типу раку або стану пацієнта застосовувалася одновимірна двошарова згорткова нейронна мережа (ЗНМ), ефективність якої для класифікації зразків на основі даних експресії генів доведена у четвертому розділі цієї монографії. На підставі попередніх результатів моделювання, розмір ядра був визначений на рівні 8, а густина ущільненого шару – 256. Параметри фільтрів, що застосовувалися при використанні даних експресії генів різних кластерних структур, представлені у таблиці 5.8. Розподіл даних для навчання та тестування моделі був таким самим, як і у попередніх моделях: 65% даних використовувалися для навчання мережі, а 35% (1140 зразків) – для її тестування. Результати моделювання щодо застосування ЗНМ для класифікації зразків, що складають тестову підмножину даних, представлені у таблицях 5.35–5.38 та на рис. 5.16.

Таблиця 5.35

Результати застосування ЗНМ до тестової підмножини зразків при використанні трикластерної структури даних експресії генів

| Тип зразка | Загальна кількість зразків | Кількість ідентифікованих зразків | Критерії класифікації |            |               |        |
|------------|----------------------------|-----------------------------------|-----------------------|------------|---------------|--------|
|            |                            |                                   | Точність              | Чутливість | Специфічність | F-міра |
| ACC        | 1140                       | 1136                              | 0.975                 | 1.000      | 0.926         | 0.962  |
| GBM        |                            |                                   |                       | 0.917      | 0.932         | 0.924  |
| KIRC       |                            |                                   |                       | 0.989      | 1.000         | 0.995  |
| LGG        |                            |                                   |                       | 0.968      | 0.968         | 0.968  |
| LUAD       |                            |                                   |                       | 0.978      | 0.963         | 0.971  |
| LUSC       |                            |                                   |                       | 0.977      | 0.983         | 0.980  |
| NORM       |                            |                                   |                       | 0.986      | 0.936         | 0.961  |
| SARC       |                            |                                   |                       | 0.948      | 1.000         | 0.974  |
| STAD       |                            |                                   |                       | 0.993      | 0.993         | 0.993  |

Таблиця 5.36

*Результати застосування ЗНМ до тестової підмножини зразків при використанні чотирикластерної структури даних експресії генів*

| Тип зразка | Загальна кількість зразків | Кількість ідентифікованих зразків | Критерії класифікації |            |               |        |
|------------|----------------------------|-----------------------------------|-----------------------|------------|---------------|--------|
|            |                            |                                   | Точність              | Чутливість | Специфічність | F-міра |
| ACC        | 1140                       | 1119                              | 0.981                 | 1.000      | 0.926         | 0.962  |
| GBM        |                            |                                   |                       | 0.911      | 0.962         | 0.936  |
| KIRC       |                            |                                   |                       | 0.995      | 1.000         | 0.997  |
| LGG        |                            |                                   |                       | 0.978      | 0.968         | 0.973  |
| LUAD       |                            |                                   |                       | 0.989      | 0.962         | 0.975  |
| LUSC       |                            |                                   |                       | 0.961      | 0.988         | 0.974  |
| NORM       |                            |                                   |                       | 1.000      | 0.949         | 0.974  |
| SARC       |                            |                                   |                       | 0.957      | 1.000         | 0.978  |
| STAD       |                            |                                   |                       | 1.000      | 1.000         | 1.000  |

Таблиця 5.37

*Результати застосування ЗНМ до тестової підмножини зразків при використанні п'ятикластерної структури даних експресії генів*

| Тип зразка | Загальна кількість зразків | Кількість ідентифікованих зразків | Критерії класифікації |            |               |        |
|------------|----------------------------|-----------------------------------|-----------------------|------------|---------------|--------|
|            |                            |                                   | Точність              | Чутливість | Специфічність | F-міра |
| ACC        | 1140                       | 1131                              | 0.976                 | 1.000      | 0.926         | 0.962  |
| GBM        |                            |                                   |                       | 0.931      | 0.915         | 0.923  |
| KIRC       |                            |                                   |                       | 0.995      | 1.000         | 0.997  |
| LGG        |                            |                                   |                       | 0.963      | 0.978         | 0.971  |
| LUAD       |                            |                                   |                       | 0.989      | 0.962         | 0.975  |
| LUSC       |                            |                                   |                       | 0.966      | 0.988         | 0.977  |
| NORM       |                            |                                   |                       | 1.000      | 0.948         | 0.973  |
| SARC       |                            |                                   |                       | 0.958      | 1.000         | 0.979  |
| STAD       |                            |                                   |                       | 1.000      | 1.000         | 1.000  |

На основі аналізу результатів моделювання можна зробити висновок, що при застосуванні згорткової нейронної мережі використання чотири- та шестикластерних структур є також більш привабливим для ідентифікації типу ракового захворювання, порівняно з три- та п'ятикластерними структурами. Однак різниця між результатами ідентифікації зразків у випадку застосування ЗНМ є суттєво меншою, порівняно із застосування класифікатора на основі алгоритму Random Forest. При цьому слід зазначити, що час навчання моделі при застосуванні ЗНМ є значно менший, порівняно із застосуванням, як класифікатора алгоритму Random Forest.

Таблиця 5.38

Результати застосування ЗНМ до тестової підмножини зразків при використанні шестикластерної структури даних експресії генів

| Тип зразка | Загальна кількість зразків | Кількість ідентифікованих зразків | Критерії класифікації |            |               |        |
|------------|----------------------------|-----------------------------------|-----------------------|------------|---------------|--------|
|            |                            |                                   | Точність              | Чутливість | Специфічність | F-міра |
| ACC        | 1140                       | 1109                              | 0.985                 | 1.000      | 0.962         | 0.980  |
| GBM        |                            |                                   |                       | 0.930      | 0.946         | 0.938  |
| KIRC       |                            |                                   |                       | 0.995      | 1.000         | 0.997  |
| LGG        |                            |                                   |                       | 0.973      | 0.978         | 0.975  |
| LUAD       |                            |                                   |                       | 0.983      | 0.978         | 0.980  |
| LUSC       |                            |                                   |                       | 0.982      | 0.982         | 0.982  |
| NORM       |                            |                                   |                       | 1.000      | 0.960         | 0.980  |
| SARC       |                            |                                   |                       | 0.979      | 1.000         | 0.989  |
| STAD       |                            |                                   |                       | 1.000      | 1.000         | 1.000  |

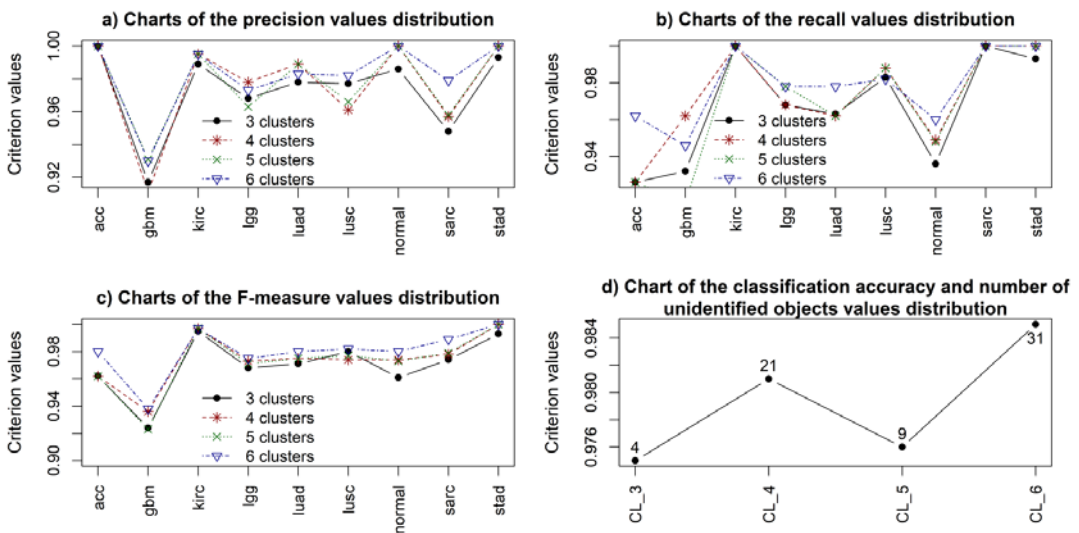


Рис. 5.16. Результати моделювання щодо застосування згорткової нейронної мережі для розв'язання мультикласової задачі ідентифікації стану пацієнтів, що досліджуються на різні типи раку

На основі аналізу результатів моделювання можна зробити висновок, що за похибками першого та другого роду та за точністю класифікації зразків при застосуванні ЗНМ використання шестикластерної структури даних дозволяє отримати кращі та більш стабільні результати класифікації усіх зразків, що відповідають усім типам ракових захворювань, включаючи зразки пацієнтів, для яких ракова хвороба не виявлена. Однак слід відзначити, що при застосуванні шестикластерної структури стан 31 зразка із 1140, що складають тестові дані, ідентифікований як невизна-

чений, що означає, що результати застосування даних експресії генів різних кластерів можуть суперечити між собою щодо визначення типу ракового захворювання. У цьому випадку виникає необхідність у додатковому клінічному дослідженні для однозначного визначення стану пацієнта або типу ракового захворювання.

Отримані у цьому розділі результати моделювання щодо застосування ЗНМ для ідентифікації стану об'єкта, що досліджується, на основі даних експресії генів свідчать про високу ефективність системи діагностики, що використовує як класифікатор ЗНМ. Застосування повного набору даних експресії генів також дозволяє отримати високу точність класифікації даних, але при цьому можуть виникнути питання щодо об'єктивності отриманих результатів. Формування кластерів диференційно експресованих та взаємно корельованих профілів експресії генів підвищує об'єктивність прийняття адекватного рішення щодо стану об'єкта за рахунок розпаралелювання потоку інформації, формування проміжних рішень на основі даних експресії генів різних кластерів, і прийняття остаточного рішення щодо стану відповідного об'єкта із застосуванням методу альтернативного голосування.

Порівняльний аналіз результатів моделювання, отриманих при застосуванні до даних експресії генів алгоритму Random Forestа згорткової нейронної мережі, дозволяє зробити висновок щодо високої ефективності обох моделей, але, як було зазначено вище, за часом навчання ЗНМ є суттєво кращою. Якщо на даних, що використовувалися, модель на основі алгоритму Random Forest навчалася приблизно 70 годин, то час навчання моделі на основі ЗНМ складав приблизно 10 хвилин. Однак за точністю класифікації та значеннями критеріїв, що визначають похибки першого та другого роду, чотирикластерна структура при застосуванні алгоритму Random Forest дозволяє отримати кращі результати ідентифікації об'єктів, порівняно з шестикластерною структурою, при застосуванні ЗНМ. У обох випадках при цьому стан 31 об'єкта ідентифікований як невизначений.

## Висновки

На основі виконаних теоретичних та експериментальних досліджень в монографії вирішено важливу науково-прикладну проблему в галузі інформаційних технологій: підвищення ефективності обробки даних експресії генів для ранньої діагностики складних захворювань, що створює умови для підвищення об'єктивності прийняття рішення щодо стану об'єкта, що досліджується, за рахунок розпаралелювання потоку обробки даних експресії генів та використання гібридних критеріїв оцінки якості реалізації відповідного етапу й застосування гібридних глибинних моделей інтелектуального аналізу даних та машинного навчання для ідентифікації стану відповідного об'єкта.

При цьому отримані такі нові результати:

1. Виконано аналіз сучасного стану досліджень в області обробки даних експресії генів для виділення диференційно експресованих та взаємно корельованих профілів експресії генів та подальшого їх застосування у системах діагностики складних об'єктів. Проаналізовано такі етапи обробки інформації, як: формування матриці експресії генів та редукції неінформативних генів за статистичними й ентропійними критеріями, кластеризація та бікластеризація даних експресії генів з метою формування підмножин взаємно корельованих й диференційно експресованих профілів експресії генів, методи машинного навчання для обробки даних експресії генів для прийняття об'єктивного рішення щодо стану об'єкта, що досліджується. Це дало змогу розробити структурну покрокову процедуру процесу обробки даних експресії генів для створення системи діагностики складних захворювань, сформулювати головну науково-практичну проблему, сформулювати мету дослідження, виділити завдання, розв'язання яких сприяє досягненню поставленої мети і вирішенню поставленої проблеми.

2. Розроблено інформаційну технологію обробки даних експресії генів для створення системи діагностики складних об'єктів, яка заснована на комплексному використанні методів інтелектуального аналізу даних, методів індуктивного моделювання складних систем, методів кластеризації високовимірних даних, методів глибинного навчання та методу альтернативного голосування для прийняття остаточного рішення щодо стану об'єктів, що досліджуються. Застосування цієї технології дозволяє підвищити об'єктивність прийняття рішення щодо стану досліджуваного об'єкта за рахунок розпаралелювання процесу обробки інформації та коректного застосування критеріїв якості обробки даних на відповідних етапах.

3. Удосконалено метрику оцінки близькості профілів експресії генів шляхом застосування модифікованого методу максимізації взаємної інформації, що заснована на комплексному використанні різних методів розрахунку ентропії Шеннона, при цьому остаточне рішення щодо ступеня близькості даних, що досліджуються, приймається на основі застосування функції бажаності Харрінгтона. Застосування цієї метрики дозволяє підвищити об'єктивність при оцінці ступеня близькості профілів експресії генів за рахунок всебічної оцінки близькості даних при застосуванні методу максимізації взаємної інформації, що заснований на різних методах оцінки ентропії Шеннона.

4. Уперше запропоновано гібридну метрику оцінки близькості профілів експресії генів, яка містить як компоненти метрики, що засновані на модифікованому критерії максимізації взаємної інформації та критерії узгодженості Пірсона. Висока ефективність цієї метрики доведена шляхом оцінки результатів класифікації об'єктів, що містять виділені за запропонованою метрикою профілі експресії генів та розрахунку відповідних критеріїв якості класифікації даних експресії генів.

5. Уперше запропоновано гібридну модель формування кластерів профілів експресії генів для створення системи ранньої діагностики складних захворювань на основі застосування щільнісного алгоритму кластеризації OPTICS у рамках індуктивної технології об'єктивної кластеризації, при цьому відстань між профілями експресії генів розраховувалася шляхом застосуванням гібридної метрики близькості даних експресії генів. Розроблено структурну блок-схему алгоритму визначення оптимальних параметрів алгоритму кластеризації OPTICS за комплексним критерієм балансу при оцінці кластерної структури, який містить як компоненти внутрішні та зовнішні критерії оцінки якості кластеризації профілів експресії генів. Висока ефективність запропонованої моделі підтверджується збіжністю критеріїв якості кластеризації профілів експресії генів та критеріїв якості класифікації об'єктів, які містять виділені профілі експресії генів як атрибути.

6. Набула подальшого розвитку концепція індуктивної технології об'єктивної кластеризації складних даних за рахунок більш ретельного вибору внутрішніх критеріїв якості кластеризації даних з урахуванням типу даних, формування критерію балансу на основі ансамблю внутрішніх критеріїв якості кластеризації та зовнішнього критерія, який враховує розбіжності між кластеризаціями, отриманими на еквівалентних підмножинах даних. Показано, що застосування цієї концепції дозволяє підвищити об'єктивність прийняття рішення щодо формування кластерної структури з одного боку, та зменшити похибку відтворюваності за рахунок застосування еквівалентних підмножин даних, з іншого боку.

7. Уперше запропоновано гібридну модель критеріального аналізу даних експресії генів, що заснована на комплексному використанні методів кластеризації та класифікації, застосування якої передбачає паралельну оцінку кластерної структури, отриманої на еквівалентних підмножинах даних із розрахунком внутрішніх та зовнішніх критеріїв якості кластеризації, при цьому остаточне рішення щодо формування ансамблю внутрішніх критеріїв та кластерної структури здійснюється на основі оцінки точності класифікації об'єктів, які містять виділені у кластери профілі експресії генів як атрибути. Застосування цієї моделі дозволяє підвищити об'єктивність при формуванні підмножин взаємно корельованих та диференційно експресованих профілів експресії генів за рахунок коректного застосування ансамблю внутрішніх критеріїв якості кластеризації з урахуванням типу даних.

8. Отримали подальший розвиток методи багатокритеріальної оптимізації в системах прийняття рішень і методи оптимізації визначення параметрів алгоритмів кластеризації профілів експресії генів та класифікації об'єктів на основі даних експресії генів за рахунок комплексного використання внутрішніх і зовнішніх критеріїв якості обробки інформації та узагальненого індексу бажаності Харрінгтона.

9. Уперше запропоновано гібридну модель формування диференційно експресованих та взаємно корельованих профілів експресії генів на основі комплексного застосування алгоритмів кластеризації та згорткової нейронної мережі, що представлено у вигляді структурної блок-схеми покрокової процедури реалізації запропонованої концепції. Показано, що застосування запропонованої моделі створює умови для підвищення об'єктивності класифікації відповідних об'єктів за рахунок узгодженого прийняття рішення щодо ідентифікації об'єкта, враховуючи результат класифікації цього об'єкта на основі даних експресії генів усіх виділених кластерів.

10. Уперше запропоновані та практично реалізовані гібридні індуктивні моделі формування підмножин диференційно експресованих та взаємно корельованих профілів експресії генів на основі комплексного застосування алгоритму кластеризації SOTA і згорткової нейронної мережі у першій моделі та алгоритму спектральної кластеризації і згорткової нейронної мережі у другій моделі. Показано, що застосування запропонованої технології дозволяє виділити підмножини (кластери) диференційно експресованих та взаємно корельованих профілів експресії генів за найбільшою точністю класифікації об'єктів, атрибутами яких є значення експресії генів відповідних кластерів. Показано перевагу алгоритму спектральної кластеризації порівняно з алгоритмом кластеризації SOTA, яка полягає у більш високій точності класифікації відповідних об'єктів з одного боку, і більш простою реалізацією, з іншого боку.

11. Уперше запропоновано концептуальну схему інформаційної технології діагностики складних об'єктів на основі комплексного застосування алгоритму спектральної кластеризації, згорткової нейронної мережі та методу альтернативного голосування, яка представлена у вигляді блок-схеми покрокової процедури реалізації процесу обробки даних експресії генів з метою підвищення об'єктивності прийняття рішення щодо стану об'єкта, що досліджується.

12. Проведено практичну реалізацію запропонованих методів, моделей та інформаційної технології обробки даних експресії генів для створення системи діагностики стану складних об'єктів. Як експериментальні дані використані дані експресії генів пацієнтів, що досліджувалися на різні типи ракових захворювань. Проведено моделювання процесу редукції профілів експресії генів шляхом видалення неінформативних генів за статистичними критеріями та інформаційної ентропії Шеннона, що дозволило зменшити кількість профілів експресії генів. Проведено практичну реалізацію гібридної індуктивної моделі формування диференційно експресованих та взаємно корельованих профілів експресії генів на основі алгоритму спектральної кластеризації. Здійснено оцінку адекватності моделі формування кластерів диференційно експресованих та взаємно корельованих профілів експресії генів шляхом застосування класифікатора до зразків, що містять дані експресії генів у сформованих кластерах як атрибуту.

13. Проведено практичну реалізацію гібридної моделі діагностики різних типів ракових захворювань на основі комплексного застосування індуктивного алгоритму спектральної кластеризації, класифікатора на основі алгоритму Random Forest та згорткової нейронної мережі і методу альтернативного голосування, що дозволило

підвищити точність та об'єктивність ідентифікації стану об'єкта за рахунок розпаралелювання процесу обробки інформації. Показано, що висока точність ідентифікації об'єктів досягається при застосуванні обох класифікаторів, але за швидкістю обробки даних згорткова нейронна мережа має суттєво більшу ефективність порівняно із алгоритмом Random Forest.

## Conclusions

Based on the theoretical and experimental studies performed, this monograph addresses an important scientific and applied problem in the field of information technology: improving the efficiency of gene expression data processing for early diagnosis of complex diseases. This improvement enhances the objectivity of decision-making regarding the state of the studied object through the parallelization of data processing and the application of hybrid criteria for evaluating the quality of each processing stage, as well as the use of hybrid deep models for intelligent data analysis and machine learning to identify the object's state.

The following novel results were obtained:

1. An analysis of the current state of research in gene expression data processing was performed, focusing on the selection of differentially expressed and mutually correlated gene profiles and their application in diagnostic systems for complex objects. The following processing stages were examined: formation of a gene expression matrix; reduction of uninformative genes using statistical and entropy-based criteria; clustering and biclustering of gene expression data to form subsets of mutually correlated and differentially expressed profiles; and machine learning techniques for objective decision-making about the state of the studied object. This analysis enabled the development of a structural step-by-step procedure for gene expression data processing aimed at constructing diagnostic systems for complex diseases. It also allowed the formulation of the main scientific and practical problem, the study objective, and the identification of tasks whose solution contributes to achieving this goal.

2. An information technology for processing gene expression data was developed to build diagnostic systems for complex objects. This technology is based on the integrated use of intelligent data analysis methods, inductive modeling of complex systems, high-dimensional data clustering, deep learning methods, and the method of alternative voting to make final decisions about the state of the studied objects. The application of this technology increases the objectivity of decision-making by parallelizing the processing and correctly applying data quality criteria at appropriate stages.

3. The metric for assessing the similarity of gene expression profiles was improved through a modified mutual information maximization method. This method integrates various approaches for calculating Shannon entropy, while the final decision on data proximity is based on the application of the Harrington desirability function. This improvement allows for more objective assessment of similarity between gene expression profiles due to a comprehensive evaluation approach.

4. A hybrid metric for evaluating the similarity of gene expression profiles is proposed. It combines metrics based on the modified mutual information maximization criterion and the Pearson correlation coefficient. The high effectiveness of this metric is demonstrated through classification results of gene expression profiles selected using the proposed metric, along with the corresponding classification quality indicators.

5. A hybrid model for clustering gene expression profiles has been proposed to support the early diagnosis of complex diseases. This model is based on the application of the OPTICS density-based clustering algorithm within the framework of inductive

objective clustering technology, with distances between gene expression profiles computed using a hybrid proximity metric. A structural block diagram was developed to determine optimal OPTICS parameters using a complex balance criterion that incorporates both internal and external clustering quality measures. The effectiveness of the proposed model is confirmed by the convergence of clustering quality criteria and classification quality metrics of objects that include gene expression profiles as features.

6. The concept of inductive technology for objective clustering of complex data has been further refined through more rigorous selection of internal clustering quality criteria, adapted to the data type. A composite balance criterion was developed, combining an ensemble of internal clustering measures with an external criterion that reflects differences in clusterings obtained from equivalent data subsets. The application of this approach improves the objectivity of cluster structure formation and reduces reproducibility error by leveraging equivalent data subsets.

7. A novel hybrid model for criterion-based analysis of gene expression data is proposed. It integrates clustering and classification techniques, including parallel evaluation of cluster structures obtained from equivalent data subsets. Both internal and external clustering quality criteria are calculated, while the final decision regarding the ensemble of internal criteria and cluster structure is based on the classification accuracy of objects with cluster-derived gene expression attributes. This model enhances the objectivity of forming subsets of mutually correlated and differentially expressed gene expression profiles by properly applying internal quality criteria tailored to the data type.

8. Methods for multi-criteria optimization in decision-making systems have been further developed. This includes optimizing parameter selection for clustering and classification algorithms based on gene expression data by integrating internal and external information quality criteria, together with the generalized Harrington desirability function.

9. A novel hybrid model for selecting differentially expressed and mutually correlated gene expression profiles has been proposed. It combines clustering algorithms with convolutional neural networks and is formalized through a structural block diagram outlining a step-by-step implementation procedure. The use of this model supports more objective classification of relevant objects by enabling coordinated decision-making that incorporates classification results based on gene expression data from all selected clusters.

10. A novel hybrid inductive models for selecting subsets of differentially expressed and mutually correlated gene expression profiles have been proposed and practically implemented. The first model combines the SOTA clustering algorithm with a convolutional neural network, while the second integrates spectral clustering with a convolutional neural network. The proposed technology enables the selection of gene expression clusters based on achieving the highest classification accuracy of objects whose attributes are the expression values of these clusters. The spectral clustering approach demonstrates an advantage over the SOTA algorithm, offering both higher classification accuracy and simpler implementation.

11. A conceptual framework for an information technology system for diagnosing complex objects has been proposed for the first time. It is based on the integrated use of

the spectral clustering algorithm, convolutional neural network, and the alternative voting method. The technology is presented as a flowchart outlining a step-by-step procedure for processing gene expression data, with the goal of increasing the objectivity of decision-making regarding the state of the studied object.

12. The proposed methods, models, and information technology for processing gene expression data were practically implemented to develop a diagnostic system for complex objects. Gene expression data from patients with various types of cancer were used as experimental input. Uninformative genes were removed using statistical measures and Shannon entropy, which reduced the dimensionality of the data. The hybrid inductive model based on spectral clustering for identifying differentially expressed and mutually correlated gene expression profiles was successfully implemented. The adequacy of the model was validated by applying classifiers to samples containing gene expression data from the identified clusters as features.

13. A hybrid model for diagnosing various types of cancer was practically implemented. It combines inductive spectral clustering, a Random Forest-based classifier, a convolutional neural network, and an alternative voting method. This integrated approach improved both the accuracy and objectivity of object state identification through parallel information processing. High classification accuracy was achieved using both classifiers; however, the convolutional neural network demonstrated significantly higher processing speed compared to the Random Forest algorithm.

## Список використаних джерел

1. Babichev S., Yasinskyi M., Yasinska-Damri L., Ratushniak Y., Lytvynenko V. Current state of the problem of gene expression data processing and extraction to solve the reverse engineering tasks in the field of bioinformatics. *CEUR Workshop Proceedings*. 2021. Vol. 2853. Pp. 62–71.
2. Yasinska-Damri L., Liakh I., Babichev S., Durnyak B. Evaluation of the gene expression profiles complex proximity metric effectiveness based on a hybrid technique of gene expression data extraction. *CEUR Workshop Proceedings*. 2021. Vol. 3038. Pp. 150–160.
3. Babichev S., Yasinska-Damri L., Liakh I., Durnyak B. Comparison analysis of gene expression profiles proximity metrics. *Symmetry (Switzerland)*. 2021. Vol. 13(10), art. no. 1812. doi: [10.3390/sym13101812](https://doi.org/10.3390/sym13101812).
4. Yasinska-Damri L., Liakh I., Babichev S., Durnyak B. Current State of Methods, Models, and Information Technologies of Genes Expression Profiling Extraction: A Review. *Lecture Notes on Data Engineering and Communications Technologies*. 2022. Vol. 77. Pp. 69–81. URL: [https://link.springer.com/chapter/10.1007/978-3-030-82014-5\\_5](https://link.springer.com/chapter/10.1007/978-3-030-82014-5_5).
5. Yasinska-Damri L., Babichev S., Liakh I. Comparison Analysis of the Pearson's Phi-Square Test and Correlation Metric Effectiveness to Form the Subset of Differently Expressed and Mutually Correlated Genes. *CEUR Workshop Proceedings*. 2022. Vol. 3156. Pp. 93–102.
6. 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. no. 11795. doi: [10.3390/app122211795](https://doi.org/10.3390/app122211795).
7. 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. doi: [10.1007/978-3-031-16203-9\\_1](https://doi.org/10.1007/978-3-031-16203-9_1).
8. Ясінська-Дамрі Л.М. Метод оцінки ступеня близькості складних об'єктів на основі модифікованого індексу максимізації взаємної інформації. *Наукові записки [Української академії друкарства]*. 2021. № 1(62). С. 42–51.
9. 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. *Applied Sciences*. 2023. Vol. 13(10), art. no. 6022. doi: [10.3390/app13106022](https://doi.org/10.3390/app13106022).
10. Yasinska-Damri L., Babichev S., Spivakovsky A., Lemeshchuk O. Formation and Analysis of Gene Expression Data Based on the Joint Use of Data Mining and Machine Learning Techniques. *CEUR Workshop Proceedings*. 2023. Vol. 3373. Pp. 87–98.
11. Ясінська-Дамрі Л.М. Індуктивна гібридна модель кластеризації даних експресії генів на основі щільнісних алгоритмів. *Наукові записки [Української академії друкарства]*. 2021. № 2 (63). С. 64–76.

12. Ясінська-Дамрі Л. М., Лях І. М., Дурняк Б. В., Бабічев С. А. Гібридна індуктивна модель кластеризації профілів експресії генів на основі алгоритму SOTA. *Наукові записки [Української академії друкарства]*. 2022. № 1 (64). С. 48–62.
13. Огірко І. В., Ясінський М. Ф., Ясінська-Дамрі Л. М., Огірко О. І. Інформаційні технології кваліметрії. *Комп'ютерні технології друкарства*. 2019. № 2 (42). С. 89–99.
14. Дурняк Б., Бабічев С., Ясінська-Дамрі Л. Застосування згорткових нейронних мереж у системах класифікації великих даних. *Комп'ютерні технології друкарства*. 2022. № 1 (47). С. 8–20.
15. Ясінська-Дамрі Л. М., Дурняк Б. В. Модель оцінки якості нормалізації даних на основі застосування критеріїв якості класифікації об'єктів. *Поліграфія і видавнича справа*. 2021. № 1 (81). С. 35–44.
16. Ясінська-Дамрі Л. М., Дурняк Б. В., Бабічев С. А. Методологічні засади розробки гібридних індуктивних моделей кластеризації великих даних. *Поліграфія і видавнича справа*. 2021. № 2 (82). С. 141–150.
17. Бабічев С. А., Ясінська-Дамрі Л. М. Моделювання процесу реконструкції генної регуляторної мережі на основі комплексного застосування топологічних параметрів. *Прикладні питання математичного моделювання*. 2021. № 1 (4). С. 32–41.
18. Огірко І. В., Ясінський М. Ф., Ясінська-Дамрі Л. М. Жорсткі і м'які математичні моделі та їх застосування. *Наукові записки [Української академії друкарства]*. 2015. № 1 (15). С. 102–117.
19. Yuzevych L., Yankovska L., Sopilnyk L., Yuzevych V., Skrynkovsky R., Koman B., Yasinska-Damri L., Neorhiadi N., Dzhala R., Yasynskyi M. Improvement of the toolset for diagnosing underground pipelines of oil and gas enterprises considering changes in internal working pressure. *Eastern-European Journal of Enterprise Technologies*. 2019. Vol. 6 (5–102). Pp. 23–29.
20. Огірко І. В., Ясінський М. Ф., Ясінська-Дамрі Л. М., Огірко О. І. Інформаційна технологія безпеки впровадження нанотехнологій у поліграфічне виробництво. *Комп'ютерні технології друкарства*. 2020. № 2 (44). С. 119–133.
21. Огірко І. В., Ясінський М. Ф., Ясінська-Дамрі Л. М., Пілат О. Інформаційна система оцінки якості електронних видань. *Комп'ютерні технології друкарства*. 2014. № 31. С. 135–143.
22. Огірко І. В., Ясінський М. Ф., Ясінська-Дамрі Л. М. Інформаційна технологія і математична модель створення графічних засобів захисту цінних паперів і документів з використанням 3D-голограми. *Комп'ютерні технології друкарства*. 2015. № 1 (33). С. 90–101.
23. Огірко І. В., Ясінський М. Ф., Огірко О. І., Ясінська-Дамрі Л. М. Інформаційна технологія і модель корозіометрії для поверхневих шарів металів. *Наукові записки [Української академії друкарства]*. 2016. № 1 (52). С. 69–77.
24. Ясінський М. Ф., Ясінська-Дамрі Л. М., Огірко О. І., Огірко І. В. Методи комп'ютерного моделювання фізико-механічних полів у поліграфії. *Квалілогія книги*. 2020. № 2 (38). С. 98–112.

25. Babichev S., Yasinska-Damri L. Technique of gene regulatory network topology optimization based on the use of ensemble of topological parameters : матеріали XXII Міжнародної конференції з математичного моделювання МКММ2021. Херсон, 2021. С. 14–15.

26. Yasinska-Damri L., Liakh I., Babichev S., Durnyak B. Comparison analysis of the Pearson  $X^2$  coefficient and correlation metric to evaluate the gene expression profiles proximity. *Інформаційні управляючі системи і технології ІУСТ-Одеса-2021*: матеріали X Міжнародної науково-практичної конференції (2021, Одеса). С. 103–104.

27. Огірко І. В., Ясінський М. Ф., Ясінська-Дамрі Л. М., Огірко О. І. Інформаційні технології моделювання та алгоритмізація. *Інформаційні технології друкарства* : збірник наукових праць VII Міжнародної науково-технічної конференції (Львів, 15–16 листопада). 2018. С. 64–70.

28. Огірко І. В., Ясінський М. Ф., Огірко О. І., Ясінська-Дамрі Л. М. Діагностика тріщин конструкцій за допомогою нечітких баз знань. *Вібрації в техніці та технологіях* : тези доповідей XIV Міжнародно-науково-технічної конференції (Дніпро, 2015). С. 29–30.

29. Ясінський М. Ф., Огірко І. В., Кашевська С. І., Огірко О. І., Ясінська-Дамрі Л. М. Інформаційні технології у медицині. *Сучасна медицина: тенденції та перспективи розвитку* : збірник матеріалів Міжнародної науково-практичної медичної конференції (Жешув, Польща, 2018). С. 137–143.

30. Ясінська-Дамрі Л. М., Дурняк Б. В., Бабічев С. А. Методологічні засади та покрокова процедура реалізації індуктивних моделей кластеризації даних. *Тези доповідей науково-технічної конференції проф.-викл. складу наук. працівників і аспірантів*. Львів : УАД, 2022. С. 116.

31. Ясінська-Дамрі Л. М. Застосування щільнісних алгоритмів OPTICS та DBSCAN у індуктивних моделях кластеризації даних. *Тези доповідей науково-технічної конференції проф.-викл. складу наук. працівників і аспірантів*. Львів : УАД, 2022. С. 120.

32. Array Express. URL: <https://www.ebi.ac.uk/arrayexpress/>.

33. Crick F. On Protein Synthesis. The Symposia of the Society for Experimental Biology, 1958. Pp. 139–163.

34. Crick F. Central Dogma of Molecular Biology. *Nature*. 1970. Vol. 227. Pp. 561–563.

35. Сиволюб А. В. Генетика: навч. посіб. 2008. URL: <https://lifelib.info/genetics/genetics/6.html>.

36. Bear D. G. Gene expression and regulation. Lecture notes for sale and distribution. Jones & Bartlett Learning, Chapter 4. Pp. 103–168.

37. Kitano H. Foundations of Systems Biology. MIT Press, Cambridge, 2001.

38. Davidson E., Levin M. Gene regulatory networks for development. *Proceedings of the National Academy of Sciences of the USA*. 2005. № 102(14). Pp. 4936–4942.

39. Emmert-Streib F., Dehmer M., Haibe-Kains B. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Frontiers in Cell and Developmental Biology*. 2014. Vol. 2. № 38.

40. Taub E. F., Deleo J. M., Thompson E. B. Sequential Comparative Hybridizations Analyzed by Computerized Image Processing Can Identify and Quantitate Regulated RNAs. *DNA*. 1983. Vol. 2(4). Pp. 309–327.
41. Chu Y., Corey D. R. RNA sequencing: Platform selection, experimental design, and data interpretation. *Nucleic Acid Therapeutics*. 2012. Vol. 22 (4). Pp. 271–274.
42. Wang Z., Gerstein M., Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*. 2009. Vol. 10 (1). Pp. 57–63.
43. Kurella M., Hsiao L.-L., Yoshida T. et al. DNA microarray analysis of complex biologic processes. *Journal of the American Society of Nephrology*. 2001. Vol. 12 (5). Pp. 1072–1078.
44. Myles L., Mace J., Montagu J. et al. Microarray biochip technology. EatonPublishing, 2000. Pp. 39–64.
45. Baldi P., Hatfield G.W. DNA Microarrays and gene expression: From experiments to data analysis modeling. Cambridge University Press, 2002. Pp. 22–23.
46. RNA Sequencing. Science Direct. URL:  
<https://www.sciencedirect.com/topics/medicine-and-dentistry/rna-sequencing>.
47. Agapito G. Computer Tools to Analyze Microarray Data. *Methods in Molecular Biology*. 2019. Vol. 1986. Pp. 267–282.
48. Gautier L., Cope L., Bolstad B. M., Irizarry R. A. Affy - Analysis of Affymetrix Gene Chip data at the probe level. *Bioinformatics*. 2004. Vol. 20 (3). Pp. 307–315.
49. Irizarry R. A., Bolstad B. M., Collin F. et al. Summaries of Affymetrix GeneChip probe level data. *Nucleic acids research*. 2003. Vol. 31(4), art. no. e15.
50. Bolstad B. M., Irizarry R. A., Astrand M., Speed T. P. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. *Bioinformatics*. 2003. Vol. 19. Pp. 185–193.
51. Affymetrix. Statistical Algorithms Description Document. Affymetrix, Inc., Santa Clara, CA, 2002. Pp. 1–27.
52. Irizarry R. A., Hobbs B., Collin F. et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003. Vol. 4(2). Pp. 249–264.
53. Chen Z., McGee M., Liu Q. Distribution-Free Convolution Model for background correction of oligonucleotide microarray data. *BMC genomics*. 2009. Vol. 10. Pp. 1–19.
54. Gentleman R., Carey V., Huber W., Irizarry R., Dudoit S. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, 2005. 473 p.
55. Smyth G.K., Speed T.P. Normalization of cDNA microarray data. *PubMed*. 2003. Vol. 31. Pp. 265–273.
56. Astrand M. Contrast Normalization of Oligonucleotide Arrays. *Journal of Computational Biology*. 2003. Vol. 10 (1). Pp. 95–102.
57. Li C., Tseng G. C., Wong W. H. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Science USA*. 2001. Vol. 98. Pp. 31–36.
58. BeadStudio Normalization Algorithms for Gene Expression Data. Illumina, 2011. Pp. 1–6.

59. Barbará D., Wu X. An Approximate Median Polish Algorithm for Large Multidimensional Data Sets. Springer-Verlag London Ltd. *Knowledge and Information Systems*. 2003. Vol. 5. Pp. 416–438.

60. Lazaridis E. N., Sinibaldi D., Bloom G., Mane S., Jone R. A simple method to improve probe set estimates from oligonucleotide arrays. *Mathematical Biosciences*. 2002. Vol. 176(1). Pp. 53–58.

61. Andrews D. F., Bickel P.J., Hampel F. R., Huber P. J., Rogers W. H., Tukey J. W. Robust Estimation of Location, Survey and Advances. *Princeton University Press, New Jersey*, 1972. 371 p.

62. Babichev S., Kornelyuk A., Lytvynenko V., Osypenko V. Computational analysis of microarray gene expression profiles of lung cancer. *Biopolymers and Cell*. Kyiv: NAS of Ukraine, 2016. Vol. 32(1). Pp. 70–79.

63. Babichev S., Durnyak B., Zhydetsky V., Pikh I., Senkivsky V. Techniques of DNA microarray data pre-processing based on the complex use of bioconductor tools and shannon entropy. *CEUR Workshop Proceedings*. 2019. Vol. 2353. Pp. 365–377.

64. Hausser J., Strimmer K. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*. 2009. Vol. 10. Pp. 1469–1484.

65. Buermans H. P. J., Ariyurek Y., van Ommen G. et al. New methods for next generation sequencing based microRNA expression profiling. *BMC Genomics*. 2010. Vol. 11(1), art. no. 716.

66. Hackenberg M., Sturm M., Langenberger D., Falcón-Pérez J. M., Aransay A. M. miRanalyzer: A microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Research*. 2009. Vol. 37 (SUPPL. 2). Pp. 68–76.

67. Farazi T. A., Brown M., Morozov P. et al. Bioinformatic analysis of barcoded cDNA libraries for small RNA profiling by next-generation sequencing. *Methods*. 2012. Vol. 58 (2). Pp. 171–187.

68. Hackenberg M., Rodríguez-Ezpeleta N., Aransay A. M. MiRanalyzer: An update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Research*. 2011. Vol. 39 (SUPPL. 2). Pp. 132–138.

69. Huang P.-J., Liu Y.-C., Lee C.-C. et al. DSAP: Deep-sequencing small RNA analysis pipeline. *Nucleic Acids Research*. 2010. Vol. 38(SUPPL. 2), art. no. gkq392. Pp. 385–391.

70. Morin R.D., O'Connor M. D., Griffith M. et al. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Research*. 2008. Vol. 18 (4). Pp. 610–621.

71. Pantano L., Estivill X., Martí E. SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. *Nucleic Acids Research*. 2009. Vol. 38(5), art. no. gkp1127. Pp. e34.1–e34.13.

72. Li Y., Zhang Z., Liu F., Vongsangnak W., Jing Q., Shen B. Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis. *Nucleic Acids Research*. 2012. Vol. 40 (10). Pp. 4298–4305.

73. Ronen R., Gan I., Modai S. et al. miRNAkey: A software for microRNA deep sequencing analysis. *Bioinformatics*. 2010. Vol. 26(20), art. no. btq493. Pp. 2615–2616.
74. Babichev S., Durnyak B., Senkivskyy V. et al. Exploratory analysis of neuroblastoma data genes expressions based on bioconductor package tools. *CEUR Workshop Proceedings*. 2019. Vol. 2488. Pp. 268–279.
75. Bioconductor. Open-source software for bioinformatics. URL: <https://www.bioconductor.org/developers/package-guidelines/>.
76. R and Bioconductor Manual. Institute for Integrative Genome Biology. URL: [http://manuals.bioinformatics.ucr.edu/home/R\\_BioCondManual](http://manuals.bioinformatics.ucr.edu/home/R_BioCondManual).
77. Dillies M. A., Rau A., Aubert J. at. al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*. 2013. Vol. 14(6), art. no. bbs046. Pp. 671–683.
78. Garmire L. X., Subramaniam S. Evaluation of normalization methods in mammalian microRNA-Seq data. *RNA*. 2012. Vol. 18 (6). Pp. 1279–1288.
79. Robinson M. D., Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*. 2010. Vol. 11(3), art. no. r 25.
80. Calinski T., Harabasz J. A dendrite method for cluster analysis. *Communications in Statistics*. 1974. Vol. 3. Pp. 1–27.
81. Zhao Q., Xu M., Franti P. Sum-of-squares based cluster validity index andsignificance analysis. *Lecture notes in computer science (including subseries lecturenotes in artificial intelligence and lecture notes in bioinformatics)*. 2009. Vol. 5495. Pp. 313–322.
82. Babichev S., Durnyak B., Pikh I., Senkivskyy V. An Evaluation of the Objective Clustering Inductive Technology Effectiveness Implemented Using Density-Based and Agglomerative Hierarchical Clustering Algorithms. *Advances in Intelligent Systems and Computing*. 2020. Vol. 1020. Pp. 532–553.
83. Fang L., Zhou X., Cui L. Biclustering high-frequency MeSH terms based on the co-occurrence of distinct semantic types in a MeSH tree. *Scientometrics*. 2020. Vol. 124 (2). Pp. 1179–1190.
84. Michalak M., Jaksik R., Słezak D. Heuristic Search of Exact Biclusters in Binary Data. *International Journal of Applied Mathematics and Computer Science*. 2020. Vol. 30 (1). Pp. 161–171.
85. Helgeson E. S., Liu Q., Chen G., Kosorok M. R., Bair E. Biclustering via sparse clustering. *Biometrics*. 2020. Vol. 76(1). Pp. 348–358.
86. Huang Q., Yang J., Feng X., Liew A. W.-C., Li X. Automated trading point forecasting based on bicluster mining and fuzzy inference. *IEEE Transactions on Fuzzy Systems*. 2020. Vol. 28(2), art. no. 8667357. Pp. 259–272.
87. Mounir M., Hamdy M., Khalifa M. E. Identifying Different Types of Biclustering Patterns Using a Correlation-Based Dilated Biclusters Algorithm. *Advances in Intelligent Systems and Computing*. 2020. Vol. 921. Pp. 261–271.
88. Kavitha Sri N., Porkodi R. A study and analysis of biclustering algorithms to identify ppi network and significant enrichments of genes in biclusters based on gene

ontology and kegg pathway database. *International Journal of Scientific and Technology Research*. 2020. Vol. 9(1). Pp. 2525–2530.

89. Kakati P., Bhattacharyya D. K., Kalita J. K. BicBioEC: biclustering in biomarker identification for ESCC. *Network Modeling Analysis in Health Informatics and Bioinformatics*. 2019. Vol. 8(1), art. no. 19.

90. Karim M. B., Kanaya S., Altaf-Ul-Amin M. Implementation of BiClusO and its comparison with other biclustering algorithms. *Applied Network Science*. 2019. Vol. 4(1), art. no. 79.

91. Mirkin B. Clustering for data mining a data recovery approach. CRC Press, 2012.

92. Pontes B., Giráldez R., Aguilar-Ruiz J. S. Biclustering on expression data: A review. *Journal of Biomedical Informatics*. 2015. Vol. 57. Pp. 163–180.

93. Kaiser S. Biclustering: Methods, Software and Application. Thesis of Doctor of Philosophy. München, 2011. 163 p.

94. Eren K., Deveci M., Kucuktunc O., Catalyurek U. V. A comparative analysis of biclustering algorithms for gene expression data. *Briefings in Bioinformatics*. 2012. Vol. 14(3). Pp. 279–292.

95. Kluger Y., Basry R., Chang J. T., Gerstein M. Spectral biclustering of microarray data: co-clustering genes and conditions. *Genome Resources*. 2003. Vol. 13(4). Pp. 703–716.

96. Mukhopadhyay A., Maulik U., Bandyopadhyay S. On biclustering of gene expression data. *Current Bioinformatics*. 2010. Vol. 5. Pp. 204–216.

97. Califano A., Stolovitzky G., Tu Y. Analysis of gene expression microarrays for phenotype classification. In *Proceedings of the International Conference on Computational Molecular Biology*, 2000. Pp. 75–85.

98. Getz G., Levine E., Domany E. Coupled two-way clustering analysis of gene microarray data. In *Proceedings of the National Academy of Sciences USA*. 2000. Vol. 27(22). Pp. 12079–12084.

99. Sheng Q., Moreau Y., De Moor. B. Biclustering microarray data by Gibbs sampling. *Bioinformatics*. 2003. Vol. 19(2). Pp. 196–205.

100. Cheng Y., Church G.M. Biclustering of expression data. *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB'00)*, 2000. Pp. 93–103.

101. Babichev S., Osypenko V., Lytvynenko V., Voronenko M., Korobchynskyi M. Comparison Analysis of Biclustering Algorithms with the use of Artificial Data and Gene Expression Profiles. 2018 IEEE 38th International Conference on Electronics and Nanotechnology, ELNANO 2018 - Proceedings, 2018. art. no. 8477439, Pp. 298–304.

102. Roy S., Bhattacharyya D. K., Kalita J. K. CoBi: Pattern Based Co-Regulated Biclustering of Gene Expression Data. *Pattern Recognition Letters*. 2013. Vol. 34 (14). Pp. 1669–1678.

103. Prelic A., Bleuler S., Zimmermann P. et al. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*. 2006. Vol. 22(9). Pp. 1122–1129.

104. Kaiser S., Leisch F. A toolbox for bicluster analysis in R. Proc. in Comp. Statistics. Physics Verlag, Heidelberg, Germany, 2008. Pp. 201–208.
105. Shirkhorshidi A.S., Aghabozorgi S., Wah T.Y., Herawan T. Big Data Clustering: A Review. International Conference on Computational Science and Its Applications ICCSA 2014: Computational Science and Its Applications (ICCSA), 2014. Pp. 707–720.
106. Hu Z., Bodyanskiy Y. V., Tyshchenko O. K., Tkachov V. M. Fuzzyclustering data arrays with omitted observations. *International Journal of Intelligent Systems and Applications*. 2017. Vol. 9(6). Pp. 24–32.
107. Литвиненко В. И. Кластерный анализ данных на основе модифицированной иммунной сети. *Управляющие системы и машины*. 2009. №1. С. 54–61.
108. 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.
109. Whitaker V., Oldham M., Boyd J. et al. Clustering of health-related behaviours within children aged 11–16: a systematic review. *BMC Public Health*. 2021. Vol. 21(1), art. no. 137.
110. Priami C. Computational approaches to understanding nutrient metabolism and metabolic disorders. *Current Opinion in Biotechnology*. 2021. Vol. 70. Pp. 7–14.
111. Kang H. S., Kim E. S. Recent advances in heterologous expression of natural product biosynthetic gene clusters in *Streptomyces* hosts. *Current Opinion in Biotechnology*. 2021. Vol. 69. Pp. 118–127.
112. Askari S. Fuzzy C-Means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development. *Expert Systems with Applications*. 2021. Vol. 165, art. no. 113856.
113. Yang S., Hu C., Yu J., Jiang H. Finite-time cluster synchronization in complex-variable networks with fractional-order and nonlinear coupling. *Neural Networks*. 2021. Vol. 135. Pp. 212–224.
114. Karim M. R., Beyan O., Zappa A. et al. Deep learning-based clustering approaches for bioinformatics. *Briefings in bioinformatics*. 2021. Vol. 22 (1). Pp. 393–415.
115. Pathak S., Jain S., Borah S. Clustering Algorithms for MANETs: A Review on Design and Development. *Advances in Intelligent Systems and Computing*. 2021. Vol. 1248. Pp. 563–578.
116. Rui Xu, Donald C. Wunsch H. Survey of Clustering Algorithms. *IEEE Transactions on neural Networks*. 2005. Vol. 16. Pp. 645–678.
117. Bodyanskiy Y. V., Deineko A. O., Kutsenko Y. V., Zayika O. O. Data streams fast EM-fuzzy clustering based on Kohonen's self-learning. Proceedings of the 2016 IEEE 1st International Conference on Data Stream Mining and Processing, DSMP 2016. Pp. 309–313.
118. Kohonen T. Self-Organizing Maps (Third Extended Edition), New York, 2001. 501 p.

119. Bodyanskiy Y., Gorshkov Y., Kolodyazhniy V., Stephan A. Combined learning algorithm for a self-organizing map with fuzzy inference. *Advances in Soft Computing*. 2006. Vol. 33. Pp. 641–650.
120. 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.
121. Fritzke B. Growing Cell Structures. A Self-Organizing Network for Unsupervised and Supervised Learning. *Neural Networks*. 1994. Vol. 7 (9). Pp. 1441–1420.
122. Hämmäläinen J., Kärkkäinen T., Rossi T. Improving scalable k-means. *Algorithms*. 2021. Vol. 14 (1), art. no. 6. Pp. 1–20.
123. Davids D. M., Christopher C. S. An efficient video summarization for surveillance system using normalized k-means and quick sort method. *Microprocessors and Microsystems*. 2021. Vol. 83, art. no. 103960.
124. Saxena S., Kumari N., Pattnaik S. Brain tumour segmentation in flair mri using sliding window texture feature extraction followed by fuzzy c-means clustering. *International Journal of Healthcare Information Systems and Informatics*. 2021. Vol. 16 (3). Pp. 1–20.
125. Gan C., Cao W. H., Liu K. Z., Wu M. A new spatial modeling method for 3D formation drillability field using fuzzy c-means clustering and random forest. *Journal of Petroleum Science and Engineering*. 2021. Vol. 200, art. no. 108371.
126. Ester M., Kriegel H. Sander J. A density-based algorithm for discovering clusters in large spatial databases. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996. Pp. 226–231.
127. Ester M., Kriegel H., Sander J., Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD-1996. Proceeding, 1996. Pp. 226–231.
128. Kriegel H.P., Kröger P., Sander J., Zimek A. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. 2011. Vol. 1(3). Pp. 231–240.
129. Whitley D. Andrew M. Sutton. Genetic Algorithms – A Survey of Models and Methods. Handbook of Natural Computing, Springer-Verlag Berlin Heidelberg, 2012. Pp. 638–671.
130. Литвиненко В. І. Методологія застосування алгоритму клонального добору для прогнозування спектрів поглинання лазерних барвників складних молекулярних структур. *Наукові вісті Національного технічного університету України “Київський політехнічний інститут”*. 2006. № 1. С. 31–40.
131. Grosan C., Abraham A. Hybrid Evolutionary Algorithms: Methodologies, Architectures, and Reviews. *Studies in Computational Intelligence (SCI)*. 2007. Vol. 75. Pp. 1–17.
132. Berkhin P. A Survey of Clustering Data Mining Techniques. Grouping Multi-dimensional Data. Recent Advances in Clustering. Springer-Verlag Berlin Heidelberg, 2006. Pp. 25–72.

133. Karypis G., Han E.H., Kumar V. Chameleon: Hierarchical clustering using dynamic modelling. *Computer*. 1999. Vol. 32 (8). Pp. 68–75.
134. Li J., Wang K., Xu L. Chameleon based on clustering feature tree and its application in customer segmentation. *Annals of Operations Research*. 2009. Vol. 168 (1). Pp. 225–245.
135. Kamel N., Boucheta R. A New Clustering Algorithm Based on Chameleon Army Strategy. *Advances in Intelligent Systems and Computing*. 2014. Vol. 265. Pp. 23–32.
136. Ball G. H., Hall D. J. Isodata: A novel method of data analysis and pattern classification. Menlo Park: Stanford Research Institute, 1965. NTIS No. AD 699616).
137. Banfeld J. D., Raftery A. E. Model-based gaussian and non-gaussian clustering. *Biometrics*. 1993. Vol. 49. Pp. 803–821.
138. Hubert L., Schultz J. Quadratic assignment as a general data-analysis strategy. *British Journal of Mathematical and Statistical Psychologie*. 1976. Vol. 29. Pp. 190–241.
139. Calinski T., Harabasz J. A dendrite method for cluster analysis. *Communications in Statistics*. 1974. Vol. 3 (1). Pp. 1–27.
140. Davies D. L., Bouldin D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1*. 1979. Vol. 2. Pp. 224–227.
141. Scott A. J., Symons M. J. Clustering methods based on likelihood ratio criteria. *Biometrics*. 1971. Vol. 27. Pp. 387–397.
142. Dunn J. Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*. 1974. Vol. 4. Pp. 95–104.
143. Bezdek J. C., Pal N.R. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics NPART B: CYBERNETICS*. 1998. Vol. 28 (3). Pp. 301–315.
144. Baker F. B., Hubert L. J. Measuring the power of hierarchical cluster analysis. *Journal of the American Statistical Association*. 1975. Vol. 70. Pp. 31–38.
145. Rohlf F. J. Methods of comparing classifications. *Annual Review of Ecology and Systematics*. 1974. Vol. 5. Pp. 101–113.
146. Marriot F. H. B. Practical problems in a method of cluster analysis. *Biometrics*. 1975. Vol. 27. Pp. 456–460.
147. Hartigan J. A. Clustering algorithms. New York: Wiley, 1975.
148. McClain J. O., Rao V. R. Clustisz: A program to test for the quality of clustering of a set of objects. *Journal of Marketing Research*. 1975. Vol. 12. Pp. 456–460.
149. Bandyopadhyay S., Pakhira M. K., Maulik U. Validity index for crisp and fuzzy clusters. *Pattern Recognition*. 2004. Vol. 37. Pp. 487–501.
150. Milligan G. W. A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*. 1981. Vol. 46 (2). Pp. 187–199.
151. Ratkowsky D. A., Lance G. N. A criterion for determining the number of groups in a classification. *Australian Computer Journal*. 1978. Vol. 10. Pp. 115–117.
152. Ray S. Turi Rose H. Determination of number of clusters in k-means clustering and application in colour image segmentation. In Proceedings of the 4th International Conference on Advances in Pattern Recognition and Digital Techniques, 1999. Pp. 137–143.

153. Halkidi M., Batistakis Y., Vazirgiannis M. On clustering validation techniques. *Journal of Intelligent Information Systems*. 2001. Vol. 17(2–3). Pp. 107–145.
154. Halkidi M., Vazirgiannis M. Clustering validity assessment: Finding the optimal partitioning of a data set. *Proceedings IEEE International Conference on Data Mining*, 2001. Pp. 187–194.
155. Rousseeuw P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. 1987. Vol. 20. Pp. 53–65.
156. Edwards A. W. F., Cavalli-Sforza L. A method for cluster analysis. *Biometrika*. 1965. Vol. 56. Pp. 362–375.
157. Friedman H. P., Rubin J. On some invariant criteria for grouping data. *Journal of the American Statistical Association*. 1967. Vol. 62. Pp. 1159–1178.
158. Xie X.L., Beni G. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1991. Vol. 13 (4). Pp. 841–846.
159. Desgraupes B. Clustering Indices. Univ. Paris Ouest, Lab Modal'X; Nanterre, France, 2013.
160. Ивахненко А. Г. Индуктивный метод самоорганизации моделей сложных систем. Киев: Наукова думка, 1982. 296 с.
161. Ивахненко А. Г. Объективная кластеризация на основе теории самоорганизации моделей. *Автоматика*. Киев, 1987. №5. С. 6–15.
162. Ивахненко А. Г., Коппа Ю. В., Петухова С. А., Ивахненко М. А. Применение самоорганизации для разбиения множества данных на заранее заданное число кластеров. *Автоматика*. Киев, 1985. №5. С. 9–16.
163. Madala H. R., Ivakhnenko A. G. Inductive Learning Algorithms for Complex Systems Modeling. *CRC Press*, 1994. 365 p.
164. Степашко В. С. Елементи теорії індуктивного моделювання. Стан та перспективи розвитку інформатики в Україні: монографія. Київ: Наукова думка, 2010. С. 471–486.
165. Yefimenko S., Stepashko V. Intelligent recurrent-And-parallel computing for solving inductive modeling problems. *Proceedings - 2015 16th International Conference on Computational Problems of Electrical Engineering, CPEE 2015*, 2015, art.no. 7333385. Pp. 236–238.
166. 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.
167. 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.
168. Babichev S. Objective clustering inductive technology of gene expression sequences. *Системные технологии*. 2016. №6(107). С. 3–20.

169. Бабічев С. А. Концептуальні основи і методологія створення індуктивної технології об'єктивної кластеризації. *Індуктивне моделювання складних систем*. Київ: Міжнародний науково-навчальний центр інформаційних технологій та систем НАН України та МОН України, 2016. Вип. 8. С. 12–32.
170. Babichev S., Taif M. A., Lytvynenko V. Estimation of the inductive model of objects clustering stability based on the k-means algorithm for different levels of data noise. *Радиоелектроника, інформатика, управління*. Запоріжжя: ЗНТУ МОН України, 2016. №4. С. 54–60.
171. Babichev S., Durnyak B., Pikh I., Senkivskyy V. An Evaluation of the Objective Clustering Inductive Technology Effectiveness Implemented Using Density-Based and Agglomerative Hierarchical Clustering Algorithms. *Advances in Intelligent Systems and Computing*. 2020. Vol. 1020. Pp. 532–553.
172. Babichev S., Taif M. A., Lytvynenko V., Osypenko V. Criterial analysis of gene expression sequences to create the objective clustering inductive technology. Materials of 2017 IEEE 37th International Conference on Electronics and Nanotechnology (ELNANO). Kiev, 2017. Pp. 244–248.
173. Zhao Q., Xu M., Fränti P. Sum-of-Squares Based Cluster Validity Index and Significance Analysis. Proceeding of International Conference on Adaptive and Natural Computing Algorithms, 2009. Pp. 313–322.
174. Babichev S., Lytvynenko V., Skvor J., Korobchynskiy M., Voronenko M. Information Technology of Gene Expression Profiles Processing for Purpose of Gene Regulatory Networks Reconstruction. *Proceedings of the 2018 IEEE 2nd International Conference on Data Stream Mining and Processing, DSMP 2018*, 2018. Art. no. 8478452. Pp. 336–341.
175. Wang L., Song F., Yin H. et al. Comparative microRNAs expression profiles analysis during embryonic development of common carp, *Cyprinus carpio*. *Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics*. 2021. Vol. 37, art. no. 100754.
176. Marchetti M. A., Coit D. G., Dusza S. W. et al. Performance of Gene Expression Profile Tests for Prognosis in Patients with Localized Cutaneous Melanoma: A Systematic Review and Meta-Analysis. *JAMA Dermatology*. 2020. Vol. 156(9). Pp. 953–962.
177. Almugren N., Alshamlan H. A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. *IEEE Access*. 2019. Vol. 7, art. no. 8736725, Pp. 78533–78548.
178. Lu H., Chen J., Yan K. et al. A hybrid feature selection algorithm for gene expression data classification. *Neurocomputing*. 2017. Vol. 256. Pp. 56–62.
179. Lee C.P., Leu Y. A novel hybrid feature selection method for microarray data analysis. *Appl. Soft Comput.* 2011. Vol. 11 (1). Pp. 208–213.
180. Chuang L.-Y., Yang C.-H., Wu K.-C., Yang C.-H. A hybrid feature selection method for DNA microarray data. *Computers in Biology and Medicine*. 2011. Vol. 41 (4). Pp. 228–237.

181. Dashtban M., Balafar M. Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. *Genomics*. 2017. Vol. 109 (2). Pp. 91–107.
182. Salem H., Attiya G., El-Fishawy N. Classification of human cancer diseases by gene expression profiles. *Appl. Soft Comput.* 2017. Vol. 50. Pp. 124–134.
183. Sharbat F. V., Mosafer S., Moattar M. H. A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization. *Genomics*, 2016. Vol. 107 (6). Pp. 231–238.
184. Vijay S. A. A., Kumar P. G. Fuzzy expert system based on a novel hybrid stem cell (HSC) algorithm for classification of micro array data. *J. Med. Syst.* 2018. Vol. 42, art.no. 61.
185. Dashtban M., Balafar M., Suravajhala P. Gene selection for tumor classification using a novel bio-inspired multi-objective approach. *Genomics*. 2018. Vol. 110 (1). Pp. 10–17.
186. Aziz R., Verma C. K., Srivastava N. A novel approach for dimension reduction of microarray. *Comput. Biol. Chem.* 2017. Vol. 71. Pp. 161–169.
187. Alshamlan H., Badr G., Alohal Y. mRMR-ABC: A hybrid gene selection algorithm for cancer classification using microarray gene expression profiling. *Biomed. Res. Int.* 2018. Vol. 2015, art. no. 604910.
188. Moradi P., Gholampour M. A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. *Applied Soft Computing*. 2016. Vol. 43. Pp. 117–130.
189. Jain I., Jain V.K., Jain R. Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification. *Applied Soft Computing*. 2018. Vol. 62. Pp. 203–215.
190. Pashaei E., Ozen M., Aydin N. Gene selection and classification approach for microarray data based on random forest ranking and BBHA. In Proc. IEEE-EMBS Int. Conf. Biomed. Health Inform. (BHI), 2016. Pp. 308–311.
191. Li X., Yin M. Multiobjective binary biogeography based optimization for feature selection using gene expression data. *IEEE Trans. Nanobiosci.* 2013. Vol. 12 (4). Pp. 343–353.
192. Shreem S. S., Abdullah S., Nazri M. Z. A Hybrid feature selection algorithm using symmetrical uncertainty and a harmony search algorithm. *International Journal of Systems Science*. 2016. Vol. 47 (6). Pp. 1312–1329.
193. Tumuluru P., Ravi B. GOA-based DBN: Grasshopper optimization algorithm-based deep belief neural networks for cancer classification. *Int. J. Appl. Eng. Res.* 2017. Vol. 12 (24). Pp. 14218–14231.
194. Djellali H., Guessoum S., Ghoualmi-Zine N., Layachi S. Fast correlation based filter combined with genetic algorithm and particle swarm on feature selection. In Proc. 5th Int. Conf. Elect. Eng.-Boumerdes (ICEE-B), 2017. Pp. 1–6.

195. Alshamlan H.M., Badr G.H., Alohal Y.A. Genetic bee colony (GBC) algorithm: A new gene selection method for microarray cancer classification. *Comput. Biol. Chem.* 2015. Vol. 56. Pp. 49–60.
196. Alon U., Barkai N., Notterman D. A. et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Nat. Acad. Sci. USA.* 1999. Vol. 96 (12). Pp. 6745–6750.
197. UCI Machine Learning Repository: Data Sets. URL: <https://archive.ics.uci.edu/ml/datasets.html>.
198. Khan J., Wei J. S., Ringnér M. et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Med.* 2001. Vol. 7 (6). Pp. 673–679.
199. Golub T.R., Slonim D. K., Tamayo P. et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science.* 1999. Vol. 286 (5439). Pp. 531–537.
200. GEMS: Gene Expression Model Selector. URL: <http://www.gems-system.org/>.
201. Gene Expression Correlates of Clinical Prostate Cancer Behavior. *Cancer Cell.* URL: [http://www.cell.com/cancer-cell/fulltext/S1535-6108\(02\)00030-2](http://www.cell.com/cancer-cell/fulltext/S1535-6108(02)00030-2).
202. Microarray Datasets. URL: <http://csse.szu.edu.cn/staff/zhuzx/Datasets.html>.
203. Armstrong S. A., Staunton J. E., Silverman L. B. et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genet.* 2002. Vol. 30 (1). Pp. 41–47.
204. Alizadeh A. A. et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature.* 2000. Vol. 403 (6769). Pp. 503–511.
205. Armstrong S. A., Staunton J. E., Silverman L. B. et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics.* 2002. Vol. 30 (1). Pp. 41–47.
206. Babichev S., Khamula O., Durnyak B., Škvor J. Technique of gene expression profiles selection based on SOTA clustering algorithm using statistical criteria and Shannon entropy. *Advances in Intelligent Systems and Computing.* 2021. Vol. 1246. Pp. 23–38.
207. Babichev S., Škvor J. Technique of Gene Expression Profiles Extraction Based on the Complex Use of Clustering and Classification Methods. *Diagnostics.* 2020. Vol. 10 (8), art. no. 584.
208. Babichev S., Barilla J., Fišer J., Škvor J. A hybrid model of gene expression profiles reducing based on the complex use of fuzzy inference system and clustering quality criteria. *Proceedings of the 11th Conference of the European Society for Fuzzy Logic and Technology, EUSFLAT 2019*, 2020. Pp. 128–133.
209. Hou J., Aerts J., den Hamer B., van Ijcken W. et al. Gene expression-based classification of non-small cell 561 lung carcinomas and survival prediction. *PLoS ONE.* 2010. 5(4), art. no. e10312.
210. XinY., Kong L., Liu Z. et al. Machine learning and deep learning methods for cybersecurity. *IEEE Access.* 2018. Vol. 6. Pp. 35365–35381.

211. Hassan C. A., Khan M. S., Shah M. A. Comparison of Machine Learning Algorithms in Data classification. *2018 24th International Conference on Automation and Computing (ICAC)*, 2018. Pp. 1–6.
212. Kumar C. A., Ramakrishnan S. Binary classification of cancer microarray gene expression data using extreme learning machines. *2014 IEEE International Conference on Computational Intelligence and Computing Research*, Coimbatore, India, 2014. Pp. 1–4. doi: 10.1109/ICCIC.2014.7238297.
213. Vural S., Wang X., Guda C. Classification of breast cancer patients using somatic mutation profiles and machine learning approaches. *BMC Syst. Biol.* 2016. Vol. 10 (S. 3). Pp. 264–276.
214. Pineda A. L., Ogoe H. A., Balasubramanian J. B. et al. On predicting lung cancer subtypes using “omic” data from tumor and tumoradjacent histologically-normal tissue. *BMC Cancer*. 2016. Vol. 16. Pp. 1–11.
215. Podolsky M. D., Barchuk A. A., Kuznetsov V. I. et al. Evaluation of Machine Learning Algorithm Utilization for Lung Cancer Classification Based on Gene Expression Levels. *Asian Pacific journal of cancer prevention: APJCP*. 2016. Vol. 17(2). Pp. 835–838.
216. Yuan F., Lu L., Zou Q. Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. *Biochimica et biophysica acta. Molecular basis of disease*. 2020. Vol. 1866(8), art. no.165822.
217. Cano A. Kent Ridge Biomedical Data Set Repository. ELVIRA Biomedical Data Set Repository, 2019. URL: <http://leo.ugr.es/elvira/DBCRepository/>.
218. Pati J. Gene expression analysis for early lung cancer prediction using machine learning techniques: An eco-genomics approach. *IEEE Access*. 2019. Vol. 7. Pp. 4232–4238.
219. Xu J., Wu P., Chen Y., Zhang L. Comparison of different classification methods for breast cancer subtypes prediction. In *Proc. Int. Conf. Secur., Pattern Anal., Cybern. (SPAC)*, 2018. Pp. 91–96.
220. Zhang W. J., Yang G., Lin Y., Ji C., Gupta M. M. On definition of deep learning. In *Proc. World Automation Congr. (WAC)*, USA, 2018. Pp. 1–5.
221. Le Cun Y., Bengio Y., Hinton G. Deep learning. *Nature*. 2015. Vol. 521. Pp. 436–444.
222. Zhu W., Xie L., Han J., Guo X. The Application of Deep Learning in Cancer Prognosis Prediction. *Cancers (Basel)*. 2020. Vol. 12(3), art no. 603.
223. Divite M., Tyagi A., Richard D. J. et al. Deep Learning-Based Pan-Cancer Classification Model Reveals Tissue-of-Origin Specific Gene Expression Signatures. *Cancers*. 2022. Vol. 14(5), art no. 1185.
224. Xiao Y., Wu J., Lin Z., Zhao X. A deep learning-based multi-model ensemble method for cancer prediction. *Computer methods and programs in biomedicine*. 2018. Vol. 153. Pp. 1–9.
225. Mostavi M., Chiu Y. C., Huang Y. et al. Convolutional neural network models for cancer type prediction based on gene expression. *BMC Med Genomics*. 2020. Vol. 13(5), art no. 44.

226. Thomas M. C., Joy A. T. Elements of Information Theory. Wiley, 2nd Edition, 2006. 792 p.
227. Shannon C. E. A mathematical theory of communication. *Bell System Technical Journal*. 1948. Vol. 27. Pp. 379–423, 623–656.
228. Mitic Vojislav V., Lazovic Goran M., Manojlovic Jelena Z. Entropy and Fractal Nature. *Thermal Science*. 2020. Vol. 24. Pp. 2203–2212.
229. Hausser J., Strimmer K. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*. 2009. Vol. 10. Pp. 1469–1484.
230. Miller G. Note on the Bias of Information Estimates. In *Information Theory in Psychology II-B*, Free Press : Glencole, IL, USA, 1955. Pp. 95–100.
231. Horvitz D. G., Thompson D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 1952. Vol. 47. Pp. 663–685.
232. Orlitsky A., Santhanam N.P., Zhang J. Always Good Turing: Asymptotically optimal probability estimation. *Science*. 2003. Vol. 302 (5644). Pp. 427–431.
233. Archer E., Park I. M., Pillow J. W. Bayesian Entropy Estimation for Countable Discrete Distributions. *Journal of Machine Learning Research*. 2014. Vol. 15. Pp. 2833–2868.
234. Ihaka R. Gentleman R. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*. 1996. Vol. 5 (3). Pp. 299–314.
235. Hou J., Aerts J., den Hamer B. et al. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS ONE*. 2010. Vol. 5, art. no. e10312.
236. Breiman, L. Random forests. *Mach. Learn.* 2001. Vol. 45. Pp. 5–32.
237. Sasaki Y. The truth of the f-measure. In *Research Fellow*. School of Computer Science, The University of Manchester, UK, 2007. Pp. 1–5.
238. Matthews B. W. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *BBA–Protein Struct.* 1975. Vol. 405. Pp. 442–451.
239. Harrington J. The desirability function. *Industrial Quality Control*. 1965. Vol. 21(10). Pp. 494–498.
240. Ankerst M., Breunig M. M., Kriegel H.-P., Sander J. OPTICS: Ordering Points to Identify the Clustering Structure. *SIGMOD Record (ACM Special Interest Group on Management of Data)*. 1999. 28 (2). Pp. 49–60. doi: 10.1145/304181.304187.
241. Ester M., Kriegel H. P., Sander J., Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Datasets with Noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, 1996. Pp. 226–231.
242. Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms. URL: <https://cran.r-project.org/web/packages/dbscan/dbscan.pdf>.
243. Babichev S., Lytvynenko V., Osypenko V. Implementation of the objective clustering inductive technology based on DBSCAN clustering algorithm. *Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2017*. Vol. 1. Pp. 479–484. art. no. 8098832.

244. Babichev S., Durnyak B., Pikh I., Senkivskyy V. An Evaluation of the Objective Clustering Inductive Technology Effectiveness Implemented Using Density-Based and Agglomerative Hierarchical Clustering Algorithms. *Advances in Intelligent Systems and Computing*. 2020. Vol. 1020. Pp. 532–553.
245. Сарычева Л. В. Объективный кластерный анализ данных на основе метода группового учета аргументов. *Проблемы управления и информатики*. 2008. №2. С. 86–104.
246. Zahn C. T. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Transactions on Computers*. 1971. Vol. 100 (1). Pp. 68–86.
247. Jain A., Law M. Data clustering: A user's dilemma. *Lecture Notes in Computer Science*. 2005. Vol. 3776. Pp. 1–10.
248. Charytanowicz M., Niewczas J., Kulczycki P. et al. A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images. *Information Technologies in Biomedicine*. 2010. Pp. 15–24.
249. Hou J., Aerts J., den Hamer B. et al. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS ONE*. 2010. Vol. 5, art. no. e10312.
250. Zhao Q., Xu M., Fränti P. Sum-of-squares based cluster validity index and significance analysis. In *Proceedings of the International Conference on Adaptive and Natural Computing Algorithms*, Kuopio, Finland, 2019. Vol. 5495. Pp. 313–322.
251. The R Project for Statistical Computing. Package Cluster Crit. URL: <https://cran.r-project.org/web/packages/clusterCrit>
252. Babichev S., Lytvynenko V., Skvor J., Fiser J. Model of the objective clustering inductive technology of gene expression profiles based on SOTA and DBSCAN clustering algorithms. *Advances in Intelligent Systems and Computing*. 2018. Vol. 689. Pp. 21–39.
253. 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.
254. 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.
255. 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.
256. 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.
257. Khan S., Rahmani H., Ali Shakh S. A., Bennamoun M. A Guide to Convolutional Neural Networks for Computer Vision. Morgan & Claypool Publishers, 2018. 207 p.
258. Springenberg J.T., Dosovitskiy A., Brox T., Riedmiller M. Striving for Simplicity: The All Convolutional Net. Cornell University, arXiv:1412.6806, 2014.

259. Brownlee J. Loss and Loss Functions for Training Deep Learning Neural Networks. *Deep Learning Performance*. 2019. URL: <https://machinelearningmastery.com/loss-and-loss-functions-for-training-deep-learning-neural-networks/>.
260. Brownlee J. How to Choose Loss Functions When Training Deep Learning Neural Networks. *Deep Learning Performance*. 2019. URL: <https://machinelearningmastery.com/how-to-choose-loss-functions-when-training-deep-learning-neural-networks/>.
261. Khan S., Rahmani H., Ali Shakh S. A., Bennamoun M. A Guide to Convolutional Neural Networks for Computer Vision. Morgan & Claypool Publishers, 2018. 207 p.
262. Hou J., Aerts J., den Hamer B. et al. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS ONE*. 2010. Vol. 5, art. no. e10312.
263. Gene Expression Omnibus. URL: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi>.
264. Bioconductor. URL: <https://www.bioconductor.org/>.
265. Morandat F., Hill B., Osvald L., Vitek J. Evaluating the Design of the R Language. In: Noble J. (eds). ECOOP 2012 – Object-Oriented Programming. ECOOP 2012. Lecture Notes in Computer Science, 2012. Vol. 7313. Springer, Berlin, Heidelberg. URL: [https://doi.org/10.1007/978-3-642-31057-7\\_6](https://doi.org/10.1007/978-3-642-31057-7_6).
266. 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.
267. Kohonen T. Self-Organizing Maps (Third Extended Edition), New York, 2001.
268. Bodyanskiy Y., Gorshkov Y., Kolodyazhnyi V., Stephan A. Combined learning algorithm for a self-organizing map with fuzzy inference. *Advances in Soft Computing*. 2006. Vol. 33. Pp. 641–650.
269. Bodyanskiy Y. V., Deineko A. O., Kutsenko Y. V., Zayika O. O. Data streams fast EM-fuzzy clustering based on Kohonen’s self-learning. Proceedings of the 2016 IEEE 1st International Conference on Data Stream Mining and Processing, DSMP 2016. Pp. 309–313.
270. Fritzke B. Growing Cell Structures. A Self-Organizing Network for Unsupervised and Supervised Learning. *Neural Networks*. 1994. Vol. 7(9). Pp. 1441–1420.
271. 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.
272. Von Luxburg U. A tutorial on spectral clustering. *Statistics and Computing*. 2007. Vol. 17 (4). Pp. 395–416.
273. Romero M., Ramírez Ó., Finke J., Rocha C. Supervised Gene Function Prediction Using Spectral Clustering on Gene Co-expression Networks. *Studies in Computational Intelligence*. 2022. Vol. 1016. Pp. 652–663.
274. Yu K., Xie W., Wang L., Zhang S., Li W. Determination of biomarkers from microarray data using graph neural network and spectral clustering. *Scientific Reports*. 2021. Vol. 11(1), art. no. 23828.
275. Liu J., Ge S., Cheng Y., Wang X. Multi-View Spectral Clustering Based on Multi-Smooth Representation Fusion for Cancer Subtype Prediction. *Frontiers in Genetics*. 2021. Vol. 12, art. no. 718915.

276. The Cancer Genome Atlas (TCGA). URL: <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>
277. Colaprico A., Silva T. C., Olsen C. et al. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research* 2016. Vol. 44(8), art. no. e71.
278. Bioconductor. URL: <https://www.bioconductor.org/>
279. Ihaka R., Gentleman R. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*. 1996. Vol. 5(3). Pp. 299–314.
280. Illumina. URL: <https://www.illumina.com/>.
281. Hausser J., Strimmer K. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*. 2009. Vol. 10. Pp. 1469–1484.

Наукове видання

Ясінська-Дамрі Людмила Михайлівна  
Сергій Анатолійович Бабічев  
Богдан Васильович Дурняк  
Михайло Федорович Ясінський

## **МОДЕЛІ ТА МЕТОДИ ОБРОБКИ ДАНИХ ЕКСПРЕСІЇ ГЕНІВ У СИСТЕМАХ ДІАГНОСТИКИ СКЛАДНИХ ОБ'ЄКТІВ**

Колективна монографія

Художнє оформлення та верстання *В. І. Сабат*  
Обкладинка *В. І. Сабат*  
Редактор *М. М. Кулинич*

Літературна агенція «ПІРАМІДА»  
м. Львів, вул. Плугова, 6.  
e-mail: [pyramidabook@ukr.net](mailto:pyramidabook@ukr.net)  
[www.pyramidabooks.net](http://www.pyramidabooks.net)

Підписано до друку \_\_. \_\_. \_\_. Формат 70×100/16  
Папір офсетний. Гарнітура «Times».  
Умов. друк. арк. \_\_. Обл.-вид. арк. \_\_.  
Друк офсетний. Тираж 300 примірників.  
Зам. № \_\_.

Віддруковано в ДП «Всеукраїнське спеціалізоване видавництво «Світ»  
79008, Львів, вул. Галицька, 21