УДК 004.048

НЕЧІТКА МОДЕЛЬ ВИДАЛЕННЯ НЕІНФОРМАТИВНИХ ПРОФІЛІВ ЕКСПРЕСІЇ ГЕНІВ ЗА СТАТИСТИЧНИМИ ТА ЕНТРОПІЙНИМИ КРИТЕРІЯМИ

І. М. Лях

Ужгородський національний університет,
пл. Народна, 3, Ужгород, 88000, Україна

Подано результати досліджень щодо формування підмножин взаємно-експресованих профілів експресії генів для подальшої реконструкції генних регуляторних мереж. Запропоновано технологію видалення неінформативних генів за статистичними критеріями та ентропією Шеннона з урахуванням ступеня пріоритетності відповідного критерію. Розроблено нечітку модель формування підмножини інформативних профілів експресії генів, валідація якої здійснювалася шляхом застосування класифікатора до об'єктів, що містять значення експресії виділених у підмножини генів як атрибути. Результати класифікації об'єктів, що містять як атрибути дані експресії генів у виділених підмножинах, показали високу ефективність запропонованої моделі, оскільки значення критеріїв класифікації об'єктів відповідали рівню інформативності відповідної групи профілів експресії генів.

Ключові слова: *експресія генів, статистичні критерії, ентропія Шеннона, нечітка логіка, критерії класифікації, ROC-аналіз.*

Постановка проблеми. Дані експресії генів, що використовуються для реконструкції ГРМ, зазвичай подають у вигляді матриці $(e_{ij}), i = \overline{1,n}, j = \overline{1,m}$, де n і m — це кількість генів та об'єктів, що досліджуються. Зазвичай після видалення генів з нульовою експресією для всіх об'єктів (неекспресовані гени) залишається приблизно 25000 генів, що визначають геном відповідного біологічного організму. Потрібно зазначити, що велика кількість генів є слабоекспресованими для усіх об'єктів. Вони визначають певні процеси, що протікають у біологічному організмі, але не є визначальними з погляду стану здоров'я об'єкта (хвороби, що ідентифікується та досліджується). Отже, на першому кроці формування експериментальних даних доцільним є видалення генів з малою експресією для усіх досліджуваних об'єктів. На другому кроці доцільним є видалення генів, значення експресії яких слабо змінюється при аналізі різних типів об'єктів (за дисперсією або стандартним відхиленням) або хаотично змінюється (високе значення ентропії Шеннона), що відповідає шуму. Ці профілі експресії генів не дають змоги за рівнем експресії однозначно ідентифікувати відповідні об'єкти за станом здоров'я біологічного організму, і вони можуть бути також видалені із бази даних. Під профілем експресії гена у цьому випадку розуміється вектор значень експресії

відповідного гена, що визначені для усіх об'єктів, що досліджуються. У контексті цього означення під взаємно-експресованими генами розуміються гени, значення експресії яких змінюється узгоджено при переході від одного об'єкта до іншого. При цьому профілі взаємно-експресованих генів дають змогу з високою точністю ідентифікувати об'єкти, враховуючи стан здоров'я біологічного організму.

Процедура формування взаємно-експресованих профілів експресії генів передбачає два етапи. Перший етап полягає у скороченні кількості генів за малим абсолютним значенням на першому кроці і за малою дисперсією та високою ентропією Шеннона на другому. При цьому виникає проблема визначення коефіцієнта трешолдінгу для кожного з критеріїв. Зазвичай коефіцієнти трешолдінгу визначаються емпіричним шляхом у процесі моделювання, враховуючи приблизну кількість генів, що мають залишитися після реалізації цього етапу. Але, враховуючи той факт, що взаємно-експресовані профілі експресії генів мають дозволяти ідентифікувати об'єкти з максимально-можливою точністю, значення коефіцієнтів трешолдінгу для кожного із критеріїв можна визначити за максимальним значенням точності класифікації об'єктів, що досліджуються.

На другому етапі до сформованої підмножини профілів експресії генів застосовується кластерний аналіз для розділення виділеної підмножини профілів експресії генів на менші кластери взаємно-експресованих генів, при цьому кількість кластерів визначається критеріями якості кластеризації з одного боку, і високою точністю класифікації об'єктів, що містять як атрибути виділені у кластери профілі експресії генів, з іншого боку. Отже, реалізація концепції формування взаємно-експресованих профілів експресії генів передбачає застосування гібридної моделі, що містить як методи інтелектуального аналізу даних, так і методи машинного навчання.

Аналіз останніх досліджень та публікації. На сьогодні вирішенню проблеми фільтрації даних експресії генів присвячено велику кількість наукових праць. У публікації [1] автори представили модуль «limma» (Linear Models for Microarrayand RNA-SeqData), який містить різні функції для формування, фільтрації та інтерпретації даних експресії генів, отриманих як шляхом проведення ДНК-мікрочіпових експериментів, так і з використанням методу секвенування молекул мРНК. Цей модуль реалізований у програмному середовищі інтелектуального аналізу даних і машинного навчання R [2] і деякою мірою альтернативою модулю «bioconductor». Він заснований на використанні лінійних моделей для виділення різно-експресованих генів при проведенні багатофакторного експерименту і містить функції для аналізу онтології генів, що є дуже суттєвим для реконструкції ГРМ, оскільки інтерпретація генів та їх взаємодії на основі аналізу концептуальних зв'язків дає змогу виділити цільові гени, встановити характер зв'язків між цільовими та іншими генами з погляду хвороби, на дослідження якої орієнтована відповідна мережа. У працях [3, 4, 5] автори навели порівняльний аналіз різних методів фільтрації даних, доступних у модулі «bioconductor» із застосування кількісних критеріїв якості для даних експресії генів, отриманих шляхом реалізації ДНК-мікрочіпових експериментів [3, 4] та методу секвенування молекул мРНК [5]. Як результат моделювання автори запропонували покроковий алгоритм виділення високо- та взаємно-експресованих профілів експресії

генів для подальшого їх групування у кластери. В огляді [6] автори подали аналіз сучасного програмного забезпечення для обробки даних експресії генів з метою виділення найбільш інформативних ознак (генів). Аналіз досліджень авторів дає змогу зробити висновок щодо доцільності застосування програмного середовища R для передобробки та обробки профілів експресії генів з метою формування кластерів високо- та взаємно-експресованих генів, оскільки він містить усі необхідні модулі та функції, які дають змогу обробити дані експресії генів відповідно до поставленої задачі.

Кластеризація або/та бікластеризація даних експресії генів є наступним кроком реалізації процедури передобробки профілів експресії генів з метою подальшої реконструкції ГРМ. Ключова відмінність цих двох технологій полягає у такому. Якщо при реалізації процедури кластеризації формуються кластери взаємно-близьких за обраною метрикою генів або об'єктів (умови проведення експерименту), то у випадку застосування алгоритму бікластеризації виділяються бікластери, що містять взаємно корельовані гени та об'єкти [7]. При цьому можливі випадки перетину бікластерів, коли один і той самий ген або об'єкт може належати різним бікластерам. На сьогодні питанням бікластерного аналізу даних експресії генів присвячено велику кількість наукових праць. Так, у публікаціях [8–10] здійснено порівняльний аналіз різних алгоритмів бікластеризації з виділенням переваг та недоліків відповідних алгоритмів, наведено класифікацію наявних алгоритмів бікластеризації. У праці [11] автори запропонували алгоритм спектральної бікластеризації для групування даних експресії генів, на прикладі застосування модельних даних представлено діаграми розподілу генів та об'єктів у бікластери, розглянуто особливості формування різних типів бікластерів при застосуванні запропонованого алгоритму. Класифікація математичних моделей різних типів бікластерів наведено у публікації [12].

Однак потрібно зазначити, що незважаючи на певні досягнення у цій предметній галузі, задача формування інформативних взаємно-експресованих профілів експресії генів на цей час немає однозначного розв'язку.

Мета статті — розробка нечіткої моделі видалення неінформативних профілів експресії генів за статистичними критеріями та ентропією Шеннона.

Виклад основного матеріалу дослідження. Проблема видалення неінформативних профілів експресії генів за статистичними та ентропійними критеріями вирішувалася у працях [13−15]. За думкою авторів, профіль експресії гена вважався інформативним, якщо максимальне значення експресії цього профілю та дисперсія більше, а ентропія Шеннона менша за відповідні порогові значення:

$$\{e_{ij}\} = \begin{Bmatrix} \max\limits_{i=1,n} e_{ij} \geq e_{lim}\,, and\ var(e_j) \geq var_{lim}\,, \\ and\ entr(e_j) \leq entr_{lim} \end{Bmatrix}, j = \overline{1,m}, \qquad (1)$$

де n — кількість зразків або об'єктів, що досліджуються; m — кількість генів. Граничні значення у цьому випадку також визначалися емпіричним шляхом у процесі моделювання, враховуючи приблизну кількість генів, які мають становити підмножину експериментальних даних для подальшого дослідження.

Але потрібно зазначити, що концепція, яку запропонували автори, має суттєвий недолік. Високе значення дисперсії відповідного профілю експресії гена або низьке значення ентропії Шеннона (за цими критеріями цей профіль експресії гена вважається інформативним) при низьких абсолютних значеннях експресії генів для всіх досліджуваних об'єктів не означає, що цей профіль експресії гена є інформативним, оскільки за абсолютними значеннями експресії він не сприяє високій точності ідентифікації об'єктів, що досліджуються. Отже, виникає необхідність у встановлені пріоритетів виконання відповідних операцій або шляхом введення послідовності їх застосування, або шляхом ініціалізації ваг тієї або іншої операції, але при цьому необхідно обґрунтувати вибір значення відповідної ваги.

У межах поточного дослідження ця задача розв'язується на основі застосування системи нечіткої логіки, при цьому пріоритетність тієї або іншої операції враховується при створенні бази нечітких правил, що становлять основу нечіткої моделі. Формування бази нечітких правил при цьому передбачає наявність таких кроків:

– визначити множину вхідних змінних: $\underline{X} = \{x_1, \underline{x_2}, ..., x_n\}$ з відповідними термами для кожної змінної: $T_{input} = \left\{t_i^p\right\}, i = \overline{1,n}, p = \overline{1,q}$, де q є кількість термів, що відповідають i-й вхідній змінній;
– визначити множину термів для вихідної змінної y: $T_{out} = \left\{t^r\right\}, r = \overline{1,q}$, q у цьому випадку — кількість термів, що відповідають вихідній змінній;
– сформувати кінцеву множину нечітких правил, узгоджених з вхідними та вихідними змінними, що використовуються в рамках моделі:

$$\bigcup_{k=1}^{m}\left[\bigcap_{p=1}^{q}\left(x_i = t_p^k\right), при \quad \omega_k\right] \to \left(y = t^r\right), i = \overline{1,n},\ r = \overline{1,q}, \tag{2}$$

де $k = \overline{1,m}$ — кількість логічних правил, що становлять нечітку базу даних; ω_k — вага k-го правила (визначається у випадку наявності пріоритету правил).

У загальному випадку процедура нечіткого логічного виводу передбачає наявність таких етапів:

– фазифікація, або встановлення відповідності між конкретними значеннями вхідних змінних, що використовуються в межах моделі, і значенням відповідних функцій належності з урахуванням відповідного терму, що відповідає цій функції належності. На етапі фазифікації, функції належності, що заздалегідь визначені на вхідних змінних, застосовуються для їх вхідних значень, тобто визначаються $\mu^{t_i^k}\left(x_i\right)$ — значення функції належності вхідної змінної x_i для терму t_i^k. Результатом етапу фазифікації є матриця значень функцій належності усіх вхідних змінних, що визначені для усіх нечітких правил, що входять у базу даних нечіткої моделі;
– агрегація, або визначення ступеня істинності умов для кожного з нечітких правил шляхом знаходження рівня «відсікання» для передумов кожного правила з використанням операції *min*:

$$\alpha_k = \bigwedge_{i=1}^{n}\left[\mu^{t_i^k}\left(x_i\right)\right]. \tag{3}$$

– активація, або знаходження ступеня істинності для кожного з нечітких правил шляхом визначення відсічених функцій належності нечітких множин для кожного з нечітких правил:

$$\mu'_k(y) = \left(\alpha_k \wedge \mu_k(X)\right), \tag{4}$$

де $\mu_k(X)$ — відсічені функції належності для вектора вхідних змінних, що відповідають *k*-му правилу; $\mu'_k(y)$ — результуюча функція належності для вихідної змінної, що відповідає *k*-му правилу;

– акумуляція, або формування функції належності результуючої нечіткої множини для вихідної змінної із застосуванням операції *max:*

$$\mu_\Sigma(y) = \bigvee_{k=1}^{m} \left[\mu'_k(y)\right]. \tag{5}$$

– дефазифікація, або знаходження чіткого значення вихідної змінної шляхом застосування до отриманої функції належності результуючої нечіткої множини відповідної операції. Операція дефазифікація може бути реалізована із застосуванням різних методів: розрахунок центру тяжіння отриманої функції, центру площини, лівого або правого модальних значень. У межах запропонованої моделі застосований найрозповсюдженіший метод центру тяжіння:

$$Y = \frac{\int_{\min}^{\max} y \cdot \mu_\Sigma(y)\,dy}{\int_{\min}^{\max} \mu_\Sigma(y)\,dy}. \tag{6}$$

Практична реалізація моделі нечіткого логічного виводу у межах досліджень передбачає такі кроки:

1. Визначення діапазонів варіювання значень вхідних статистичних критеріїв, ентропії Шеннона та вихідного параметра (значущість профілю за здатністю ідентифікувати відповідний об’єкт).

2. Визначення функцій належності нечітких множин для вхідних і вихідного параметрів.

3. Формування бази нечітких правил, що формують нечіткий логічний вивід.

4. Визначення алгоритму нечіткого логічного виводу та методу формування чіткого значення вихідної змінної.

5. Визначення кількісних критеріїв оцінки адекватності моделі у процесі її тестування.

Діапазон варіювання значень вхідних параметрів у межах запропонованої моделі визначається шляхом аналізу загальної статистики, при цьому для абсолютних значень експресії генів на першому кроці визначається максимальне значення експресії для кожного профілю. Далі для отриманого вектора максимальних значень експресії генів, вектора дисперсії профілів експресії генів та ентропії Шеннона формується загальна статистика. Для створення нечіткої моделі застосовувався міжквартильний інтервал зміни максимальних абсолютних значень, дисперсії та ентропії Шеннона профілів експресії генів. При цьому сформовані діапазони були розділені на три інтервали з відповідними термами. Для дисперсії та

максимальних абсолютних значень експресії генів: $0\% \le x < 25\%$ — «Низьке» (Н); $25\% \le x < 75\%$ — «Середнє» (С); $x \ge 75\%$ — «Високе» (В). Для ентропії Шеннона: $x \ge 75\%$ — «Високе» (В); $25\% \le x < 75\%$ «Середнє» (С); $x < 25\%$ — «Низьке» (Н). Діапазон варіювання вихідного параметра (значущість профілю) у запропоновані моделі змінювався від 0 до 100 і розділявся на п'ять рівних інтервалів: $0 \le y < 20$ — «Дуже низьке» (ДН); $20 \le y < 40$ — «Низьке» (Н); $40 \le y < 60$ — «Середнє» (С); $60 \le y < 80$ — «Високе» (В); $80 \le y \le 100$ — «Дуже високе» (ДВ). Щодо функцій належності нечітких множин для вхідних параметрів до значень з термами «Низьке» та «Високе» застосовувалася трапецеїдальна функція належності, а до середнього діапазону значень (С) — трикутникова функція належності. До всіх підмножин вихідного параметра застосовувалися трикутникові функції належності нечітких множин. Потрібно зазначити, що параметри функцій належності нечітких множин вхідних критеріїв передбачають підлаштування у процесі моделювання, враховуючи характер розподіл значень експресії генів в експериментальних даних, що досліджуються.

У табл. 1 наведені терми нечіткої бази правил, що використовувалися у процесі налаштування нечіткої моделі. Аналіз даних таблиці дає змогу зробити висновок, що пріоритетним показником для ідентифікації ступеня значущості профілю експресії гена є максимум абсолютних значень експресії цього гена, що визначені для всіх досліджуваних об'єктів. Як зазначалося вище, гени, значення експресії яких порівняно низьке для усіх об'єктів, що досліджуються, не є визначальними для ідентифікації об'єктів і можуть бути видаленими, незважаючи на високе значення дисперсії та/або низьке значення ентропії Шеннона. Комбінації значень дисперсії та ентропії Шеннона у цьому випадку є коригуючими.

На останньому кроці визначається чітке значення вихідної змінної, як центр тяжіння отриманої фігури відповідно до формули (6). Апробація запропонованого методу формування груп взаємно-експресованих профілів експресії генів за статистичними та ентропійними критеріями здійснювалася із застосуванням даних експресії генів пацієнтів, що досліджувалися на ранню стадію раку легенів. Дані GSE19188 [16] були взяті із вільно доступної бази даних Gene Expression Omnibus [17] і містили дані експресії генів 156 пацієнтів, з яких 65 були ідентифіковані за результатами клінічних досліджень як здорові, а у 91 була ідентифікована ракова пухлина у початковій стадії. Передобробка даних здійснювалася шляхом застосування функцій пакету Bioconductor мови програмування R. У початковому стані дані містили 54675 генів. На першому етапі для кожного профілю експресії генів розраховувалося максимальне значення експресії гена, дисперсія та ентропія Шеннона за методом Джеймса та Стейна.

На рис. 1 зображені діаграми розмаху отриманих векторів значень статистичних критеріїв та ентропії Шеннона, аналіз яких дає змогу зробити висновок щодо доцільності застосування вищезазначених діапазонів відповідних значень у процесі налаштування системи нечіткого логічного виводу.

Таблиця 1

Терми бази знань нечіткої моделі формування підмножин взаємно-експресованих профілів експресії генів

№	Максимальне абсолютне значення	Дисперсія	Ентропія Шеннона	Значущість профілю
1	В	В	Н	ДВ
2	В	С	Н	В
3	В	В	С	В
4	В	С	С	В
5	В	Н	С	В
6	В	В	В	В
7	В	Н	Н	В
8	В	Н	В	С
9	В	С	В	С
10	С	В	Н	В
11	С	Н	Н	С
12	С	С	Н	С
13	С	В	С	С
14	С	С	С	С
15	С	Н	С	С
16	С	В	В	С
17	С	Н	В	Н
18	Н	В	Н	Н
19	Н	С	Н	Н
20	Н	В	С	Н
21	Н	Н	Н	Н
22	Н	С	С	Н
23	Н	Н	В	ДН

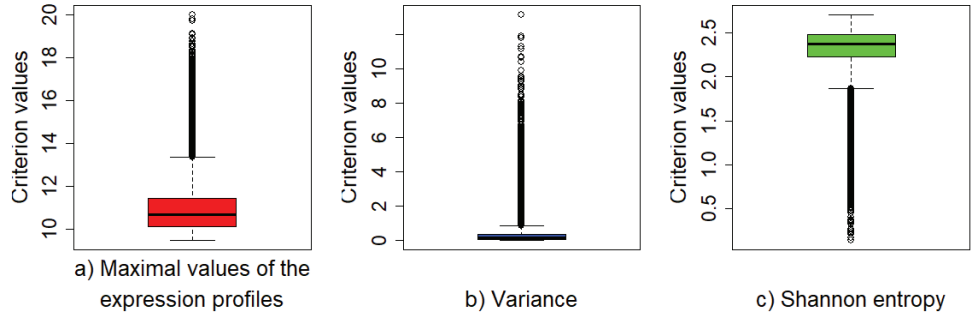

Рис. 1. Характер розподілу статистичних критеріїв та ентропії Шеннона профілів експресії генів пацієнтів, що досліджувалися на ранню стадію раку легенів

Дійсно, найбільш інформативні профілі експресії генів мають високі значення експресії та дисперсії і низьке значення ентропії Шеннона. На рис. 2 зображені функції належності нечітких множин для вхідних критеріїв та вихідного параметра, що використовувалися в межах запропонованої моделі.

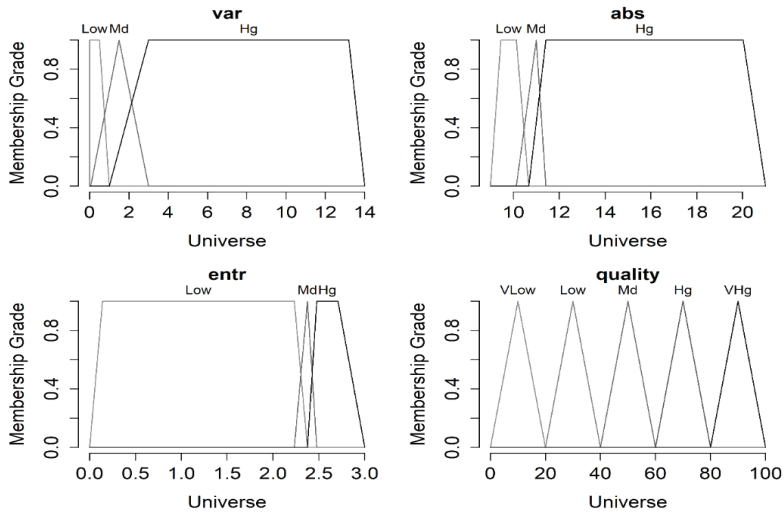

Рис. 2. Функції належності нечітких множин вхідних змінних та вихідного параметра, що застосовувалися у нечіткій моделі формування взаємно-експресованих профілів експресії генів

Покрокову процедуру, що була описана вище та реалізована у процесі моделювання і яка дає змогу оцінити ефективність запропонованої нечіткої моделі шляхом аналізу результатів класифікації об’єктів, які містять як атрибути сформовані підмножини даних експресії генів, зображено на рис. 3.

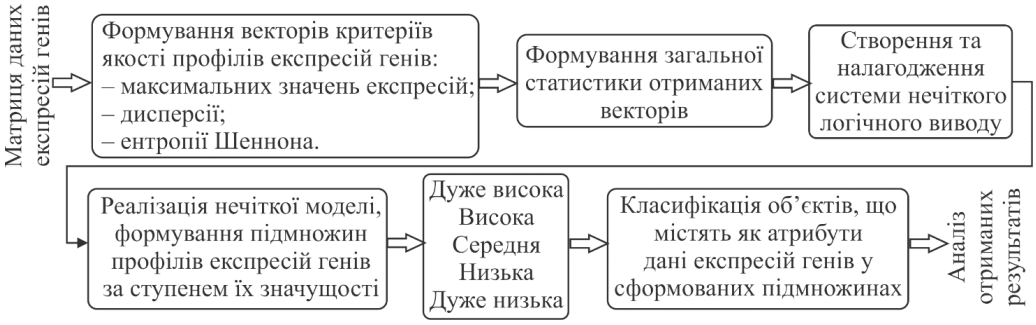

Рис. 3. Структурна блок-схема покрокової процедури формування підмножин профілів експресії генів за статистичними та ентропійними критеріями на основі застосування системи нечіткого логічного виводу

Як можна побачити, реалізація нечіткої моделі передбачає такі етапи:

Етап 1. Формування векторів критеріїв якості профілів експресії генів та загальної статистики для значень отриманих векторів.

1.1. Розрахунок для кожного профілю експресії генів максимального значення експресії, дисперсії та ентропії Шеннона. Довжина отриманих векторів при цьому дорівнює кількості генів, що становлять експериментальні дані.

1.2. Розрахунок загальної статистики отриманих векторів. Фіксація інтервалу зміни відповідних значень та квантілів, що відповідають 25, 50 та 75 відсоткам інтервалу варіювання значень відповідних критеріїв.

Етап II. Створення, налагодження та реалізація системи нечіткого логічного виводу.

2.1. Формування структури системи нечіткого логічного виводу, формалізація моделі, визначення функцій належності нечітких множин для вхідних та вихідних параметрів, формування бази правил моделі.

2.2. Застосування моделі нечіткого логічного виводу до профілів експресії генів, формування підмножин профілів експресії генів за рівнем їх значущості, враховуючи відповідні значення статистичних критеріїв та ентропії Шеннона.

Етап III. Оцінка адекватності моделі шляхом застосування класифікатора до об'єктів, що містять як атрибути сформовані підмножини даних експресії генів.

3.1. Вибір класифікатора, враховуючи тип даних, формування критеріїв якості класифікації.

3.2. Реалізація процедури класифікації об'єктів, що містять дані експресії генів у виділених підмножинах як атрибути.

3.3. Розрахунок критеріїв якості класифікації об'єктів.

Етап IV. Аналіз отриманих результатів.

4.1. Формування відповідних рішень щодо адекватності нечіткої моделі, враховуючи кореляцію між результатами класифікації об'єктів на основі даних експресії генів у сформованих підмножинах та ступенем значущості відповідних профілів експресії генів.

На рис. 4 наведено результати моделювання щодо застосування моделі нечіткого логічного виводу для формування підмножин профілів експресії генів різного ступеня значущості.

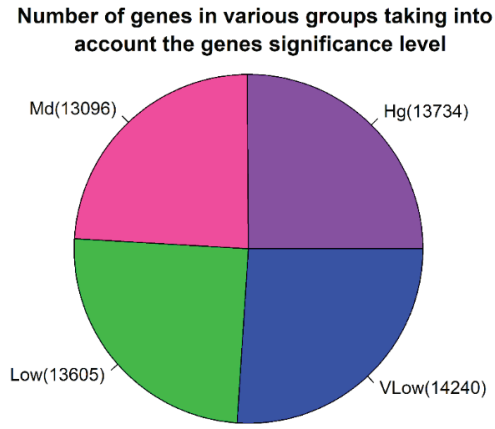

Рис. 4. Результати моделювання щодо застосування моделі нечіткого логічного виводу для формування підмножин профілів експресії генів різного ступеня значущості за статистичними критеріями та ентропією Шеннона

Враховуючи, що із 54675 генів тільки 29 були ідентифіковані як «Дуже високий» ступінь значущості, групи зі ступенем значущості «Дуже високий» і «Високий» були об'єднані для подальшого моделювання. Аналіз отриманих результатів дає змогу зробити висновок щодо адекватності роботи нечіткої моделі із розділення множини генів на відповідні підмножини за кількістю генів. Як добре відомо, геном людини становлять приблизно 25000 активних генів. З цього погляду виділення 13734 генів дуже високого та високого ступеня значущості та 13096 генів середнього ступеня значущості для подальшої обробки є доцільним. Гени з низьким та дуже низьким ступенем значущості можна видалити з даних як неінформативні.

Наступним кроком, який може підтвердити або спростувати висновок щодо адекватності отриманих результатів із видалення системою нечіткого логічного виводу неінформативних профілів експресії генів за вищезазначеною групою критеріїв, є застосування класифікатора для ідентифікації об'єктів, що містять як атрибути дані експресії генів у відповідних підмножинах.

Оцінка якості класифікації об'єктів в межах досліджень здійснювалася із застосуванням похибок першого та другого роду, при цьому розраховувалися такі критерії:

– точність класифікації об'єктів (Accuracy):

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}, \tag{7}$$

де TP і TN — коректно ідентифіковані позитивні та негативні випадки, відповідно (наприклад, наявність або відсутність хвороби); FP і FN — помилково ідентифіковані позитивні та негативні випадки (похибки першого та другого роду, відповідно);

– F-індекс визначається як гармонічне середнє точності (Precision (PR)) і чутливості (Recall (RC)):

$$F = \frac{2 \cdot \mathrm{PR} \cdot \mathrm{RC}}{PR + RC}, \tag{8}$$

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

– Matthews Correlation Coefficient (MCC):

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}}. \tag{9}$$

У табл. 2 наведено процедуру формування матриці збіжності, що лежить в основі розрахунку критеріїв якості класифікації об'єктів.

Таблиця 2

Таблиця збіжності для ідентифікації похибок першого та другого роду

Стан об'єкта за результатами клінічних випробувань	Результат класифікації об'єктів	
	Хворий (True — 1)	Здоровий (False — 0)
Хворий (True — 1)	TP (True Positives)	FN (False Negatives)
Здоровий (False — 0)	FP (False Positives)	TN (True Negatives)

У цьому випадку точність класифікації максимальна (100 %), якщо всі об'єкти ідентифіковані коректно і помилки першого (FP) та другого (FN) роду відсутні. Значення F-індексу та MCC критерію при цьому також є максимальними і дорівнюють 1.

Другий тип критерію, що використовувався в межах досліджень для оцінки якості класифікації об'єктів, заснований на ROC-аналізі (Received Operating Characteristic), при цьому розраховувалася площа під ROC-кривою AUC (Area Under Curve). Більша площа відповідає більш високій якості класифікації об'єктів.

Вибір класифікатора визначається особливістю експериментальних даних експресії генів, ключовою відмінною рисою яких є велика кількість атрибутів (кількість генів, що визначають стан організму, що досліджується). Як зазначалося вище, у процесі моделювання застосовано як дані експресії генів 156 пацієнтів, з яких 65 були ідентифіковані за результатами клінічних досліджень як здорові, а у 91 була ідентифікована ракова пухлина на початковій стадії. Початкова кількість генів (54675) при цьому була розділена на чотири групи за статистичними критеріями та ентропією Шеннона (рис. 4). Кожна група містила приблизно 13000 генів, що подаються на вхід класифікатора. У цьому випадку класифікатор має бути орієнтований на великі дані. У працях [18–20] автори подали результати досліджень щодо застосування як класифікатор згорткових нейронних мереж (convolutional neural networks) для ідентифікації об'єктів на основі даних експресії генів. Автори дослідили різні топологічні структури цього типу мереж і довели їх ефективність для класифікації об'єктів на основі високо-вимірних даних експресії генів. Але потрібно зазначити, що коректне застосування згорткових

нейронних мереж передбачає при формуванні згорткових шарів доповнення даних профілями з нульовою експресією для отримання необхідної кількості генів. У межах поточних досліджень цей крок може вплинути на результати, що є небажаним. З цієї причини застосування згорткових нейронних мереж на цьому етапі моделювання не є доцільним. У публікації [13] автори навели результати досліджень щодо порівняння бінарних класифікаторів для ідентифікації об'єктів на основі високовимірних даних експресії генів. За результатами досліджень автори зробили висновок, що алгоритм бінарної класифікації «Випадковий ліс» (Random-forest) за критеріями якості класифікації та значенням площі під ROC-кривою має більш високу ефективність для ідентифікації об'єктів на основі даних експресії генів, порівняно з іншими аналогічними класифікаторами. З цієї причини цей класифікатор був використаний в межах поточного моделювання. Результати моделювання щодо ідентифікації об'єктів, що містять як атрибути дані експресії генів, наведені у табл. 3 та на рис. 5.

Таблиця 3

Результати моделювання щодо класифікації об'єктів на основі даних експресії генів різного ступеня значущості

Значущість генів	Критерії якості класифікації об'єктів				
	Точність, %	Чутливість	Специфічність	F-індекс	MCC
Висока	98,4	1	0,973	0,992	0,967
Середня	93,5	1	0,9	0,967	0,873
Низька	90,3	0,917	0,917	0,894	0,801
Дуже низька	85,5	0,870	0,846	0,862	0,701

Аналіз отриманих результатів підтверджує доцільність застосування запропонованої нечіткої моделі для формування груп профілів експресії генів різного ступеня значущості за статистичними критеріями та ентропією Шеннона. Значення критеріїв якості класифікації, що наведені у табл. 3, поступово збільшуються при переході від підмножин профілів експресії генів з дуже низьким ступенем значущості до підмножини профілів експресії генів з високим ступенем значущості, при цьому характер зміни критерію точності класифікації об'єктів, F-індексу та MCC-критерію збігається в межах допустимої похибки. Аналіз значень AUC-критерію (площа під ROC-кривою) також підтверджує доцільність застосування системи нечіткого логічного виводу для розділення множини профілів експресії генів на підмножини генів з різним ступенем значущості, але потрібно зазначити, що цей тип критерію є суттєво менш чутливим для профілів експресії генів, що становлять підмножини з високим та середнім ступенями значущості. Крім того, значення AUC-критерію для підмножини генів з дуже низьким ступенем значущості є вищим за аналогічне значення для підмножини профілів експресії генів з низьким ступенем значущості, що не є коректним і суперечить значенням критеріїв класифікації, що наведені у табл. 3. Але потрібно зазначити, що цей критерій дає змогу розділити

множину профілів експресії генів на дві підмножини: підмножина інформативних генів у цьому випадку містить гени з високим і середнім ступенями значущості; інші гени видаляються як неінформативні.

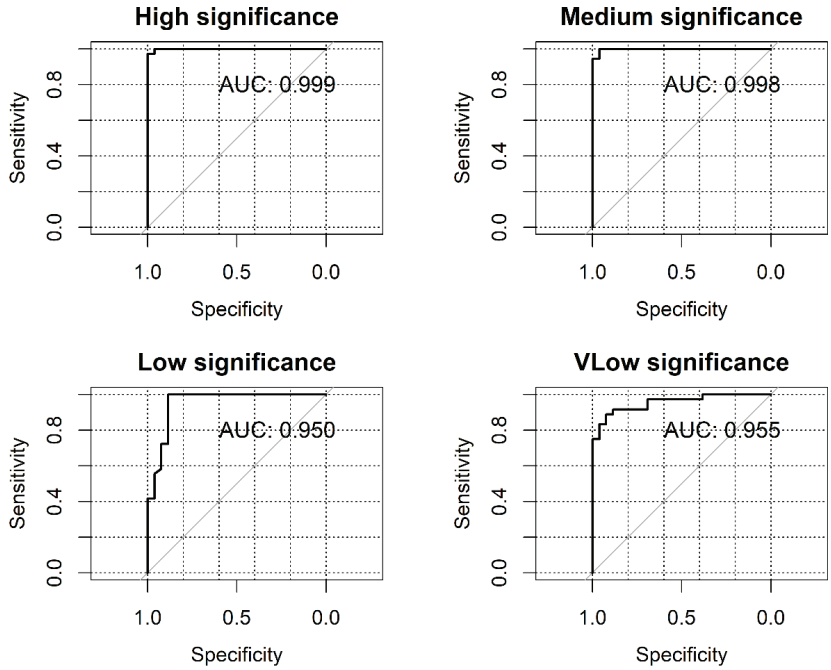

Рис. 5. Результати ROC-аналізу для оцінки ефективності нечіткої моделі формування підмножин профілів експресії генів за ступенем їх значущості

Висновки. Наведено результати досліджень щодо формування підмножин взаємно-експресованих профілів експресії генів для подальшої реконструкції генних регуляторних мереж. Запропоновано технологію видалення неінформативних генів за статистичними критеріями та ентропією Шеннона з урахуванням ступеня пріоритетності відповідного критерію із застосуванням нечіткої моделі, яка дозволяє рівень значущості відповідного профілю експресії гена за відповідними критеріями. Розроблено базу нечітких правил, що становлять основу нечіткої моделі, сформовані функції належності нечітких множин вхідних та вихідних параметрів моделі. Проведено моделювання процесу формування груп профілів експресії генів різного ступеня інформативності, доведено адекватність запропонованої моделі шляхом аналізу значень критеріїв класифікації зразків на основі даних експресії генів у сформованих кластерах. Аналіз результатів моделювання дав змогу зробити висновок щодо доцільності застосування запропонованої нечіткої моделі для формування підмножин профілів експресії генів різного ступеня значущості за статистичними та ентропійними критеріями.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Ritchie M. E., Phipson B., Wu D. et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*. 2015. Vol. 43 (7), art. no. e47.
2. 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.
3. Computational analysis of microarray gene expression profiles of lung cancer / Babichev S., Kornelyuk A., Lytvynenko V., Osypenko V. *Biopolymers and Cell*. Kyiv : NAS of Ukraine, 2016. Vol. 32 (1). Pp. 70–79.
4. Techniques of DNA microarray data pre-processing based on the complex use of bioconductor tools and Shannon entropy / Babichev S., Durnyak B., Zhydetskyy V., Pikh I., Senkivskyy V. *TCEUR Workshop Proceedings*. 2019. Vol. 2353. Pp. 365−377.
5. 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.
6. Tan C. S., Ting W. S., Mohamad M. S. et al. A Review of Feature Extraction Software for Microarray Gene Expression Data. *BioMed Research International*. 2014. Vol. 2014, art. no. 213656. doi: 10.1155/2014/213656.
7. Mirkin B. Clustering for data mining a data recovery approach. CRC Press, 2012.
8. Pontes B., Giráldez R., Aguilar-Ruiz J. S. Biclustering on expression data: A review. *Journal of Biomedical Informatics*. 2015. Vol. 57. Pp. 163−180. doi: 10.1016/j.jbi.2015.06.028.
9. Kaiser S. Biclustering: Methods, Software and Application. Thesis of Doctor of Philosophy. Minchin, 2011. 163 p.
10. A comparative analysis of biclustering algorithms for gene expression data / Eren K., Deveci M., Kucuktunc O., Catalyurek U. V. *Briefings in Bioinformatics*. 2012. Vol. 14 (3). Pp. 279–292.
11. Spectral biclustering of microarray data: co-clustering genes and conditions / Kluger Y., Basry R., Chang J. T., Gerstein M. *Genome Resources*. 2003. Vol. 13 (4). Pp. 703–716.
12. Mukhopadhyay A., Maulik U., Bandyopadhyay S. On biclustering of gene expression data. *Current Bioinformatics*. 2010. Vol. 5. Pp. 204–216.
13. 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.
14. A hybrid model of gene expression profiles reducing based on the complex use of fuzzy inference system and clustering quality criteria / Babichev S., Barilla J., Fišer J., Škvor J. *Proceedings of the 11th Conference of the European Society for Fuzzy Logic and Technology, EUSFLAT 2019*, 2020. Pp. 128−133.
15. A fuzzy model for gene expression profiles reducing based on the complex use of statistical criteria and Shannon entropy / Babichev S., Lytvynenko V., Gozhyj A., Korobchynskyi M., Voronenko M. *Advancesin Intelligent Systems and Computing*. 2019. Vol. 754. Pp. 545−554.
16. Hou J., Aerts J., denHamer B. etal. Gene expression – based classification of non-small cell lung carcinomas and survival prediction. *PLoS ONE*. 2010. Vol. 5, art. no. e10312.
17. Gene Expression Omnibus. El. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi.

18. 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.
19. 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.
20. Li J., Sun W., Feng X. etal. A denseconnection encoding–decoding convolutional neural network structure for semantic segmentation of thymoma. *Neurocomputing*. 2021. Vol. 451. Pp. 1−11.

REFERENCES

1. Ritchie, M. E., Phipson, B., & Wu, D. et al. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies: Nucleic Acids Research, 43 (7), art. no. e47 (in English).
2. Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics: Journal of Computational and Graphical Statistics, 5 (3), 299−314 (in English).
3. Babichev, S., Kornelyuk, A., Lytvynenko, V., & Osypenko, V. (2016). Computational analysis of microarray gene expression profiles of lung cancer: Biopolymers and Cell. Kyiv : NAS of Ukraine, 32 (1), 70–79 (in English).
4. Babichev, S., Durnyak, B., Zhydetskyy, V., Pikh, I., & Senkivskyy, V. (2019). Techniques of DNA microarray data pre-processing based on the complex use of bioconductor tools and Shannon entropy: TCEUR Workshop Proceedings, 2353, 365−377 (in English).
5. Babichev, S., Durnyak, B., & Senkivskyy, V. et al. (2019). Exploratory analysis of neuroblastoma data genes expressions based on bioconductor package tools: *CEUR Workshop Proceedings*, 2488, 268−279 (in English).
6. Tan, C. S., Ting, W. S., & Mohamad, M. S. et al. (20140. A Review of Feature Extraction Software for Microarray Gene Expression Data: *BioMed Research International*, 2014, art. no. 213656. doi: 10.1155/2014/213656 (in English).
7. Mirkin, B. (2012). Clustering for data mining a data recovery approach. CRC Press (in English).
8. Pontes, B., Giráldez, R., & Aguilar-Ruiz, J. S. (2015). Biclustering on expression data: A review: *Journal of Biomedical Informatics*, 57, 163−180. doi: 10.1016/j.jbi.2015.06.028 (in English).
9. Kaiser, S. (2011). Biclustering: Methods, Software and Application. Thesis of Doctor of Philosophy. Minchin (in English).
10. Eren, K., Deveci, M., Kucuktunc, O., & Catalyurek, U. V. (2012). A comparative analysis of biclustering algorithms for gene expression data: Briefings in Bioinformatics, 14 (3), 279–292 (in English).
11. Kluger, Y., Basry, R., Chang, J. T., & Gerstein, M. (2003). Spectral biclustering of microarray data: co-clustering genes and conditions: Genome Resources, 13 (4), 703–716 (in English).
12. Mukhopadhyay, A., Maulik, U., & Bandyopadhyay, S. (2010). On biclustering of gene expression data: Current Bioinformatics, 5, 204–216 (in English).
13. Babichev, S., & Škvor, J. (2020). Technique of Gene Expression Profiles Extraction Based on the Complex Use of Clustering and Classification Methods: *Diagnostics*, 10 (8), art. no. 584 (in English).

14. Babichev, S., Barilla, J., Fišer, J., & Škvor, J. (2020). 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*, 128−133 (in English).
15. Babichev, S., Lytvynenko, V., Gozhyj, A., Korobchynskyi, M., & Voronenko, M. (2019). A fuzzy model for gene expression profiles reducing based on the complex use of statistical criteria and Shannon entropy: *Advancesin Intelligent Systems and Computing*, 754, 545−554 (in English).
16. Hou, J., Aerts, J., & denHamer, B. etal. (2010). Gene expression – based classification of non-small cell lung carcinomas and survival prediction: PLoS ONE, 5, art. no. e10312 (in English).
17. Gene Expression Omnibus. El. Retrieved from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi (in English).
18. Chuang, Y.-H., Huang, S.-H., & Hung, T.-M. et al. (2021). Convolutional neural network for human cancer types prediction by integrating protein interaction networks and omics data: *Scientific Reports*, 11 (1), art. no. 20691 (in English).
19. Busaleh, M., Hussain, M., & Aboalsamh, H. A. (2021). Breast mass classification using diverse contextual information and convolutional neural network: *Biosensors*, 11(11), art. no. 419 (in English).
20. Li, J., Sun, W., & Feng, X. etal. (2021). A denseconnection encoding–decoding convolutional neural network structure for semantic segmentation of thymoma: *Neurocomputing*, 451, 1−11 (in English).

doi: 10.32403/1998-6912-2023-1-66-39-55

A FUZZY MODEL FOR THE REMOVAL OF UNINFORMATIVE GENE EXPRESSION PROFILES USING STATISTICAL AND ENTROPY CRITERIA

I. M. Liakh

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

The results of research on the formation of subsets of mutually expressed gene expression profiles for further reconstruction of gene regulatory networks are presented. A technology for removing uninformative genes based on statistical criteria and Shannon's entropy, considering the degree of priority of the corresponding criterion, is proposed. The range of variation of the values of the input parameters within the framework of the proposed model is determined by analyzing general statistics, while for the absolute values of gene expression, the maximum value of expression for each profile is determined in the first step. Next, general statistics are formed for the obtained vector of maximum values of gene expression, vector of dispersion of gene expression profiles

and Shannon entropy. To create a fuzzy model, the interquartile interval of changes in maximum absolute values, dispersion and Shannon entropy of gene expression profiles are used. At the same time, the formed ranges are divided into three intervals with corresponding terms. A fuzzy model of the formation of a subset of informative gene expression profiles is developed, the validation of which is carried out by applying a classifier to objects containing the expression values of the genes selected in the subset as attributes. The results of classification of objects containing gene expression data in selected subsets as attributes shows the high efficiency of the proposed model, since the values of the object classification criteria correspond to the level of informativeness of the corresponding group of gene expression profiles.

Further perspectives of the author's research are the practical implementation of the proposed model for the formation of subsets of informative gene expression profiles for the purpose of reconstructing gene regulatory networks.

Keywords: *gene expression, statistical criteria, Shannon entropy, fuzzy logic, classification criteria, ROC analysis.*

Стаття надійшла до редакції 24.04.2023.
Received 24.04.2023.