УДК 004.8

Н. Р. Кондратенко, к. т. н., доц.; О. О. Манаєва

НЕЧІТКА КЛАСТЕРИЗАЦІЯ АБОНЕНТІВ ІНТЕРНЕТ-ПРОВАЙДЕРА

Запропоновано генетичний алгоритм нечіткої кластеризації на основі неоднорідних хромосом із початковою ініціалізацією координат центрів кластерів. Його було досліджено на збіжність і функціонування, проілюстровано комп'ютерним експериментом.

Ключові слова: *нечітка кластеризація, провайдер, генетичний алгоритм, неоднорідна хромосома, тестові функції, середньозважене відхилення, ступені належності.*

Вступ

Інтернет – глобальна інформаційна мережа, яка об'єднує велику кількість регіональних мереж і водночас мільйони комп'ютерів в усіх кінцях планети з метою обміну даними та доступу до інформаційних і технологічних ресурсів [1]. Наданням послуг доступу до мережі Інтернет та інших послуг, пов'язаних із нею, займаються інтернет-провайдери. Такого роду організації володіють величезними обсягами інформації про своїх користувачів; цю інформацію необхідно певним чином систематизувати, структурувати, узагальнювати тощо. Ці завдання тісно пов'язані із задачею кластерного аналізу [2].

Існує велика кількість методів кластеризації, які можна класифікувати на чіткі і нечіткі. Чіткі методи кластеризації розбивають вихідну множину об'єктів $\boldsymbol{X}$ на декілька підмножин, що не перетинаються. При цьому будь-який об'єкт із $\boldsymbol{X}$ належить лише одному кластера. Нечіткі методи кластеризації дозволяють одному й тому самому об'єкту належати одночасно до декількох (або навіть до всіх) кластерів, але з різним ступенем належності. Єдиною відмінністю є те, що при нечіткому розбитті ступінь належності об'єкта до кластера набуває значення з інтервалу [0, 1], а при чіткому – з двоелементної множини {0, 1}. Нечітка кластеризація в багатьох ситуаціях є "природнішою" за чітку, наприклад, для об'єктів, розташованих на межі кластерів [3, 4].

Для розв'язання поставленої задачі запропонуємо підхід, який ґрунтується саме на нечіткому розбитті простору об'єктів на кластери. У задачі розподілу абонентів провайдера інтернет-послуг на групи саме такий підхід має практичне значення. Це пов'язано з основною вимогою до тарифних планів організації – їхньою гнучкістю. Для виконання цієї вимоги при побудові розбиття необхідно допускати деяку невизначеність щодо належності абонента до певної групи.

Нехай кожен із абонентів виступає як об'єкт, що характеризується певними значеннями заданих показників (швидкість доступу, обсяг спожитого вхідного та вихідного трафіку тощо). Відповідно вони можуть бути представлені як точки в багатовимірному просторі. Практичний сенс такого розуміння подібності означає, що абоненти вважаються тим більш схожими, чим менше розходження між однойменними показниками, за допомогою яких вони описуються [4].

За таких умов доцільно розв'язати подібну задачу в масштабах окремого провайдера інтернет-послуг. У нашому випадку як об'єкти розглядаються абоненти згаданого провайдера, представлені набором параметрів. Задача полягає в розбитті сукупності абонентів, заданих таким чином, на однорідні нечіткі множини.

Метою представленого **дослідження** є математичне моделювання поведінки абонентів щодо провайдера телекомунікаційних послуг і розподіл їх на однорідні групи з можливістю подальшого аналізу отриманих результатів.

Постановка задачі

Поставимо задачу розбиття множини абонентів інтернет-провайдера на нечіткі однорідні підмножини відповідно до заданого набору показників. При цьому кожен абонент може міститися в певному кластері з деяким ступенем належності в межах від 0 до 1. Необхідно визначити всі ступені належності μ_{ij} абонента j до кластера і, а також місця розташування центрів кластерів $c_i, i = \overline{1, m}$.

Для розв'язання поставленої задачі запропонуємо генетичний алгоритм, що здійснює нечітку кластеризацію абонентів інтернет-провайдера відповідно до зазначених показників, та дослідимо його на збіжність на ряді тестових функцій.

Математична модель

Нехай маємо набір абонентів $I=\{I_1, I_2, \ldots, I_n\}$ деякого провайдера інтернет-послуг. Кожен із n абонентів характеризується множиною ознак (вимірів) $X_i=\{x_1, x_2, \ldots, x_p\}$, серед яких швидкість передачі даних, а також обсяги вхідного та вихідного трафіку за заданий період часу. Задача нечіткої кластеризації полягає в тому, щоб на основі даних, що містяться у множині I, розбити множину абонентів I на $1<m<n$ кластерів, тобто визначити ступені належності μ_{ij} кожного абонента до кожного з m кластерів, що задаються центрами $c_i, i = \overline{1, m}$.

На величини μ_{ij} накладаються такі обмеження:

1. $0 \le \mu_{ij} \le 1$;
2. $\sum_{i=1}^{m} \mu_{ij} = 1$ для всіх j. (1)

Для оцінки якості нечіткого розбиття використовується середньозважене відхиле точок-абонентів від центрів кластерів:

$$E = \sum_{i=1}^{m} \sum_{j=1}^{n} \mu_{ij}^{m} \left\| x_j - c_i \right\|^{2},$$

де $m \ge 1$ – експоненційна вага, що визначає нечіткість, розсіяність кластерів.

Необхідно знайти таке розміщення центрів кластерів c_i та величини $\{\mu_{ij}\}$, за яких велич цього критерію була б мінімальною при одночасному дотриманні умов обмеження (1) [illegible]. Для розв'язання цієї задачі запропонуємо генетичний алгоритм оптимізації. Класичний генетичний алгоритм являє собою ітераційний процес, на кожній ітерації якого популяція послідовно піддається операціям відбору, схрещування та мутації. Зупинивши ітераційний процес у певний момент та вибравши кращу особину з популяції, можна отримати цілком прийнятний розв'язок задачі.

Запропонуємо неоднорідну хромосому як формалізоване представлення розв'язку. Хромосомний набір складається з двох якісно відмінних частин. Перша з них являє собою прямокутну матрицю, що містить ступені належності точок-абонентів до відповідних кластерів. Її елементи визначають ступінь зв'язку відповідного абонента з певним кластером. Друга визначає координати центрів кластерів у просторі ознак. Для кожної такої хромосоми обчислюється деяке значення цільової функції.

На підготовчому етапі відбувається початкова ініціалізація координат центрів кластерів. Вона спирається на геометричне представлення абонентів у вигляді точок у просторі. Для цього кожна з осей розбивається на $N = 2 + E(3{,}322 lgn)$ інтервалів. Таким чином, увесь простір ознак розділяється на N^d рівних за об'ємом кубів, де d – кількість ознак (вимірів). За початкові центри кластерів приймаються геометричні центри кубів, всередину яких потрапляє найбільша кількість точок.

Надалі над початковою популяцією виконуються операції схрещування та мутації в такій послідовності:

– одноточкове схрещування: дві хромосоми розрізаються у випадково вибраній точці та обмінюються отриманими частинами. Операція проводиться за однаковою схемою над обома компонентами хромосоми;

– двоточкове схрещування: хромосоми розцінюються як цикли, які формуються з'єднанням кінців лінійної хромосоми. Для заміни сегмента одного циклу сегментом іншого циклу вибираються дві точки розрізу;

– рівномірне схрещування: кожен ген нащадка створюється копіюванням відповідного гена від одного або іншого з батьків згідно з випадково згенерованою маскою. Якщо у відповідній позиції маски стоїть 1, то ген копіюється з першої батьківської хромосоми, якщо 0, то з другої. Процес повторюється з батьками, яких обміняли, для створення другого нащадка. Для кожної пари батьків випадково генерується нова маска;

– мутація: генерування нових ступенів належності для однієї, випадково обраної точки, а також випадкова зміна положення центра кожного кластера за одним виміром у просторі ознак.

Особини, повністю ідентичні за хромосомним набором, із популяції вилучаються та замінюються на мутантів, утворених за наведеною вище схемою.

У результаті виконання таких дій отримуємо $7n$ генетично унікальних нащадків, для кожного з яких підраховуємо цільову функцію, після чого в межах популяції реалізується механізм природного відбору на основі стратегії елітизму. При цьому розв'язки з нижчим значенням цільової функції гарантовано переходять у популяцію наступного покоління, що сприяє швидшій збіжності генетичного алгоритму. Переважно за рахунок схрещування відбувається опрацювання найбільш перспективних варіантів розв'язку, тоді як мутації реалізують механізм виходу оптимізаційного процесу з локальних мінімумів. Як результат, алгоритм із високою ймовірністю сходиться до розв'язку, максимально близького до оптимального.

Комп'ютерний експеримент

Розв'яжемо задачу кластеризації абонентів провайдера телекомунікаційних послуг за трьома ознаками: швидкістю передачі даних та обсягами вхідного і вихідного трафіку за фіксований період часу. Обсяг досліджуваної вибірки становить 100 користувачів.

Результат кластеризації абонентів за даними показниками наведено в таблицях 1 та 2.

Таблиця 1

Результати кластеризації: ступені належності

Код абонента	Кластер 1	Кластер 2	Кластер 3	Кластер 4
0	0,004049983	0,058683567	0,268992839	0,668273611
1	0,233707476	0,039341806	0,080751559	0,646199159
2	0,083366999	0,115781932	0,751763956	0,049087113
3	0,189849775	0,73390408	0,010152989	0,066093155
4	0,997684019	0,000579759	0,001224275	0,000511947
5	0,010728999	0,784860715	0,19198237	0,012427916
6	0,003554482	0,098927767	0,250335671	0,64718208
7	0,01029047	0,864558804	0,010456979	0,114693747
8	0,989020468	0,002236017	0,007642451	0,001101064
9	0,006364244	0,85193206	0,015814621	0,125889074
…	…			
99	0,059285226	0,903040624	0,002809315	0,034864835

Таблиця 2

Результати кластеризації: положення центрів кластерів

Вимір простору ознак	Кластер 1	Кластер 2	Кластер 3	Кластер 4
Швидкість передачі даних	3208,12	430,24	484,15	819,71
Обсяг вхідного трафіку	4133,44	1613,21	887,96	701,37
Обсяг вихідного трафіку	514,09	2503,17	105,46	88,01

Із розташування центрів кластерів бачимо, що в першому кластері знаходяться абоненти, що мають найвищу швидкість доступу, другий – невеликий сегмент користувачів, у яких обсяг вихідного трафіку за добу, яка розглядається, наближається до вхідного або перевищує його. До третього кластера було віднесено абонентів, швидкість передачі даних для яких переважно невисока, а вхідний трафік значно переважає вихідний. Четвертий кластер відрізняється від третього вищою швидкістю доступу, співвідношення ж вхідного та вихідного трафіку приблизно таке саме, як у третьому.

Дослідження генетичного алгоритму на збіжність

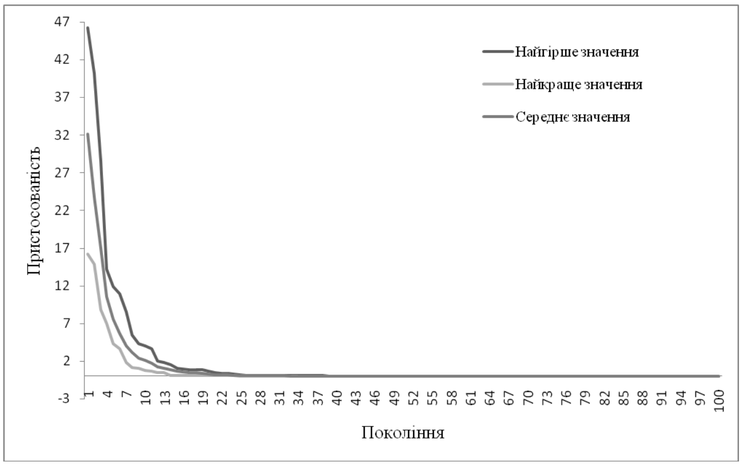
Оцінимо ефективність запропонованого генетичного алгоритму за допомогою тестових функцій. Дослідимо його оптимізаційні можливості для числа змінних n=10 та n=100 при граничному числі поколінь 100 та 1000 відповідно. Для цього виконаємо для кожної з наведених нижче тестових функцій серію з десяти експериментів.

1. Сферична функція (перша функція де Джонга) – неперервна випукла унімодальна тестова функція, вважається найпростішою для оптимізації:

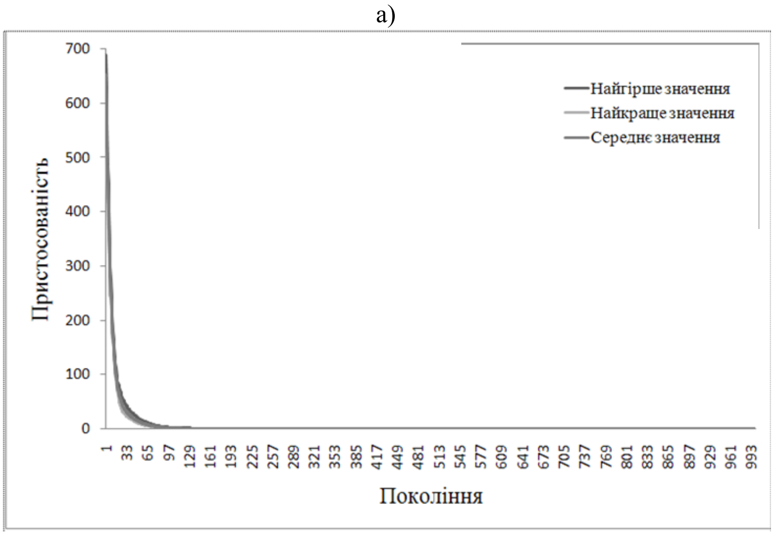
$$f_1(\mathbf{x}) = \sum_{i=1}^{n} x_i^2, \quad (3)$$

де $-5{,}12 \le x_i \le 5{,}12$, i=1…n. Має один глобальний мінімум, який дорівнює 0 в точці, де x_i=0, i=1…n.

Зміни найкращої, найгіршої та середньої пристосованості найкращої особини в популяції для цієї функції в серії з десяти експериментів показано на рис. 1.



а)



б)

Рис. 1. Зміна найкращої, найгіршої та середньої пристосованості найкращої особини в популяції для сферичної функції при n=10 (а) та n=100 (б)

Аналогічним чином запропонований алгоритм було випробувано на інших поширених тестових функціях. У таблиці 3 наведено найкращі, найгірші та середні значення пристосованості найкращої особини в популяції в останньому поколінні, отримані для різних тестових функцій.

Таблиця 3

Результати дослідження алгоритму на збіжність

Тестова функція	Значення пристосованості найкращої особини в популяції					
	Найкраще		Найгірше		Середнє	
	n=10	n=100	n=10	n=100	n=10	n=100
Сферична	0,0005884	0,000238	0,002541	0,000478	0,001472	0,000324
Крокова	0	0	0	0	0	0
Растрігіна	0,088847	0,031292	0,775811	0,053048	0,335579	0,044519
Швефеля	0,417962	0,20015783	3,743808	0,468031	0,914933	0,338687
Гріванка	0,3411336	0,0493624	1,001547	0,141237	0,65774	0,092017

З табл. 3 видно, що середня помилка знаходження глобального екстремуму для жодної з тестових функцій не перевищує порядку першого знака після коми. Це підтверджує високу ефективність роботи запропонованого генетичного алгоритму, у тому числі й у разі великого числа змінних.

Висновки

У статті запропоновано підхід для розв'язання задачі нечіткої кластеризації абонентів провайдера інтернет-послуг та розроблено генетичний алгоритм із використанням неоднорідних хромосом.

Дослідження запропонованого генетичного алгоритму на збіжність засвідчило, що він є потужним оптимізаційним алгоритмом, а тому може використовуватись у задачі кластеризації, якій властива наявність великої кількості параметрів та, як правило, значного числа локальних екстремумів.

За допомогою запропонованого алгоритму було проведено кластеризацію абонентів інтернет-провайдера за показниками, що характеризують особливості використання ними послуг цієї організації. У результаті множину користувачів було розбито на компактні групи, між якими існують суттєві відмінності за певними показниками. Проведений кластерний аналіз дозволив дійти висновку, що застосування нечіткості, зокрема в задачах кластерного аналізу, дозволяє працювати й отримувати результати в умовах значної зашумленості даних. Тому подальші дослідження в цьому напрямку є перспективними.

СПИСОК ЛІТЕРАТУРИ

1. Олифер В. Г. Компьютерные сети: принципы, технологии, протоколы / В. Г. Олифер, Н. А. Олифер. – СПб.: Питер, 2006. – 958 с.

2. Муссель К. Предоставление и биллинг услуг связи. Системная интеграция / К. Муссель. – М.: Эко-Трендз, 2003. – 319 с.

3. Дюран Б. Кластерный анализ / Б. Дюран, П. Оделл; Пер. с англ. Е.З. Демиденко. – М.: Статистика, 1977. – 128 с.

4. Мандель И. Д. Кластерный анализ / И. Д. Мандель. – М.: Статистика, 1988. – 176 с.

5. Зайченко Ю. П. Нечеткие модели и методы в интеллектуальных системах / Ю. П. Зайченко. – К.: Издательский дом «Слово», 2008. – 344 с.

Кондратенко Наталія Романівна – к. т. н., доцент, професор кафедри захисту інформації.

Манаєва Ольга Олексіївна – магістрант.

Вінницький національний технічний університет.