УДК 275:004.738.5:339 **DOI: 10.31673/2412-9070.2022.063842**

І. М. АВЕРІЧЕВ, канд. екон. наук;
Д. А. ДІБРІЙ, магістр,
Державний університет телекомунікацій, Київ

РОЗРОБЛЕННЯ МЕТОДУ ПОШУКУ СХОЖИХ АКАУНТІВ У СОЦІАЛЬНИХ МЕРЕЖАХ НА ОСНОВІ КЛАСТЕРНОГО АНАЛІЗУ

В останні десятиліття істотно зросла кількість користувачів мережі «Інтернет». Водночас збільшується кількість людей, які використовують соціальні мережі для зв'язку та отримання інформації. Активний розвиток соціальних мереж також надзвичайно вплинув на ринок реклами, адже завдяки інформації про користувачів, яку збирають ці мережі, рекламодавці мають змогу таргетувати рекламу з величезною точністю, що значно впливає на зріст продажів товарів та послуг. Адже за допомогою визначення схожих акаунтів можна легко встановити, що товар, який сподобався одному користувачу, швидше за все привабить й іншого.

Щоб відшукати найбільш доцільний за вибраною тематикою метод кластерного аналізу, було проведено системний аналіз методів кластерного аналізу. Проаналізовано, що одним із підходів до пошуку схожих акаунтів у соціальних мережах є аналіз акаунтів за певними ознаками, що широко використовується сучасними компаніями, які мають соціальні мережі для таргетованої реклами. Аналіз методів кластеризації показав, що не існує певного універсального методу для аналізу кластеризації даних, а отже, було проаналізовано різні методи кластеризації та зроблено висновок про те, що поліпшений алгоритм k-means – k-means mini batch найкраще підходить для пошуку схожих акаунтів.

Ключові слова: метод кластерного аналізу; інформаційні технології; кластер; методи кластеризації; спрощений алгоритм кластеризації; метод статистичного аналізу; системний аналіз; великі дані.

Вступ

Постановка проблеми. З появою нових технологій у сучасному світі почали по-іншому взаємодіяти соціальні мережі з користувачами. Ця парадигма працює з великою кількістю даних, її застосовує компанія Amazon. Сервіс використовує збір персональних рекомендацій і динамічного ціноутворення. Але існує значна проблема — кількість інформації про користувачів настільки велика, що стає неможливим обробляти її звичайними методами. Через це постала потреба в аналізі великих даних.

Аналіз останніх досліджень і публікацій. Вітчизняні та зарубіжні вчені, такі як Дж. Гурвіц, А. Ньюджент, Ф. Халпер, М. Кауфма, К. О. Кірей у своїх дослідженнях все частіше звертаються до теоретичних і практичних проблем розвитку та трансформації Big Data.

Великі дані набули особливого поширення в багатьох сферах нашого життя, хоча ми цього і не помічаємо. Їх використовують не тільки для соціальних мереж, а й у медицині, телекомунікаціях та фінансових компаніях, а також у державному управлінні. За допомогою технологій Big Data підприємства мають змогу обробляти великі масиви даних і виявляти корисні закономірності, що дають їм конкурентні переваги.

Для пошуку таких закономірностей широко використовується кластерний аналіз, адже він надає можливість розподіляти користувачів на групи за спільними параметрами, що дає змогу оптимізувати оброблення даних для подальшого аналізу та використання. Сьогодні найбільші компанії світу, зокрема Amazon, Google, Meta, активно вкладають ресурси в дослідження способів аналізу користувачів за допомогою кластеризації.

Мета і задачі дослідження. Зі збільшенням інтернет-ринку значно зросла конкуренція за гроші юзера, тепер уже недостатньо просто аналізувати користувача за первинною рисою, зростає потреба в аналізі різних даних, які передусім можуть бути не зв'язаними між собою.

Але такі величезні масиви даних неможливо аналізувати людям, особливо в таких чималих компаніях, як Amazon, Meta, Google, які мають мільярди користувачів. Також слід зауважити, що дані про кожного окремого користувача зазвичай не мають великої значущості з фінансового погляду, тому рекламодавці зацікавлені в групах таких користувачів, котрих об'єднують спільні риси.

Через це виникає потреба у використанні певних алгоритмів для аналізу та сортування схожих акаунтів за певними рисами та подальший розподіл цих акаунтів на групи.

Для розв'язання цієї задачі широко використовуються алгоритми кластеризації, оскільки вони можуть аналізувати велику кількість даних за певними рисами та розподіляти схожі дані в групи.

Основна частина

Кластерний аналіз застосовується для виявлення розподілу об'єктів. Традиційний кластерний аналіз зазвичай проводиться для групування спостережень або змінних окремо, але є й одночасна кластеризація (або бікластеризація) рядків і стовпців матриці даних.

У статті використовуватимемо спрощений алгоритм кластеризації, наведений далі.

1. Завантаження початкових даних та випадковий вибір центрів.

2. На першій ітерації кожен об'єкт віднесено до певної групи.

3. Перерахунок метричності кожного об'єкта до відповідного центру кластерів та перевизначення його центрів.

4. Завершення кластеризації.

Слід зазначити, що етапи 1-3 можуть неодноразово повторюватись, оскільки кластеризація проходить ітеративно (рис. 1 – рис. 4).

Рис. 1. Початок кластеризації, вибираються точки

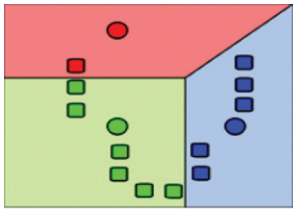

Рис. 2. Перша ітерація. Точки відносять до центрів

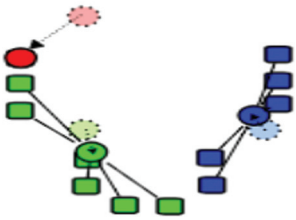

Рис. 3. Перерахунок центрів кластерів

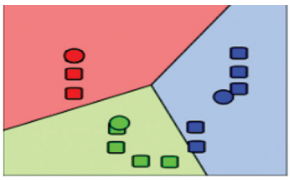

Рис. 4. Кінець кластеризації, точки віднесені до своїх кластерів

Алгоритми кластеризації можна поділити на категорії з огляду на їх кластерну модель.

Наведемо способи порівняння таких моделей (рис. 5 – рис. 8).

Рис. 5. Ієрархічна кластеризація — кластеризація за допомогою дендрограм

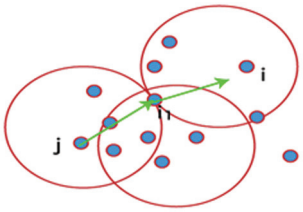

Рис. 6. Кластеризація на основі щільності — враховує щільність розташування об'єктів

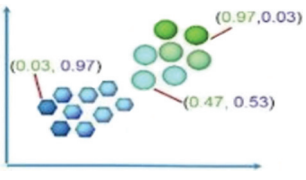

Рис. 7. *c*-means — кластеризація відносить об'єкти до кластерів із певною вірогідністю

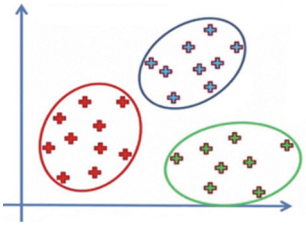

Рис. 8. *k*-means — кластеризація випадково вибирає початкові центри та визначає найближчі об'єкти

Алгоритми з'єднують «об'єкти» в «кластери» залежно від відстані. Кластер можна описати загалом максимальною відстанню, необхідною для з'єднання частин кластера.

На різних відстанях утворюватимуться різні кластери, які можна уявити за допомогою дендрограми, що пояснює, звідки з'явилася загальна назва «ієрархічна кластеризація».

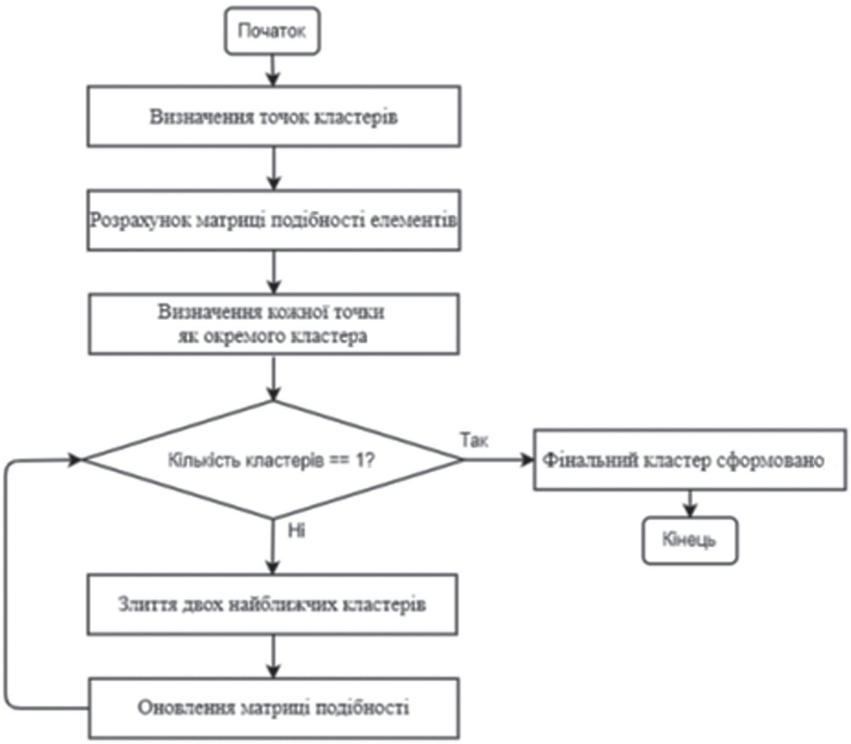

Рис. 9. Алгоритм ієрархічної кластеризації

Оскільки вважається, що початкові дані матимуть досить великий обсяг через те, що в деяких соціальних мережах кількість користувачів сягає мільярду, доречно буде використовувати модифікований для роботи з великою кількістю даних *k*-means mini batch (у звичайного *k*-means кількість обчислень збільшується зі збільшенням кількості елементів).

Але попередній аналіз алгоритмів кластеризації показав, що не існує певного методу «срібної кулі», який дає змогу аналізувати будь-який масив даних однаково ефективно, вибір необхідного методу суцільно залежить від висунутих у бізнес-вимогах задач. Оскільки неможливо описати роботу всіх методів, у цій статті буде розглянуто лише алгоритм *k*-means як найбільш універсальний та відповідний до аналізу акаунтів користувачів (рис. 10).

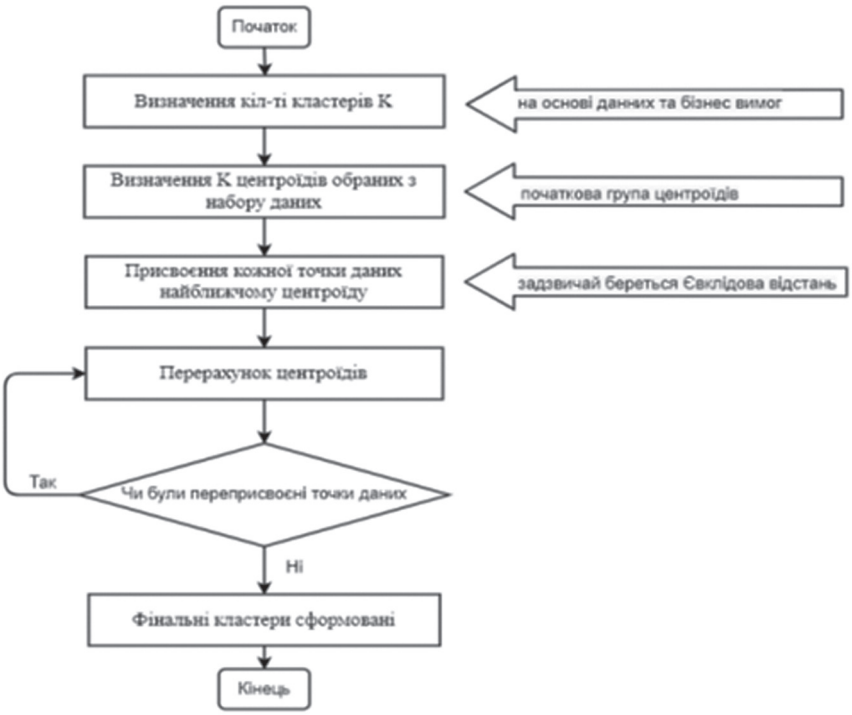

Рис. 10. Алгоритм *k*-means

Основною відмінністю *k*-means mini batch від звичайного *k*-means є те, що він ітеративно розбиває вхідні дані на рандомізовані батчі — тобто невеликі частини початкового масиву даних, які вибираються випадково. Порівняння швидкості роботи обох алгоритмів наведено на рис. 11.

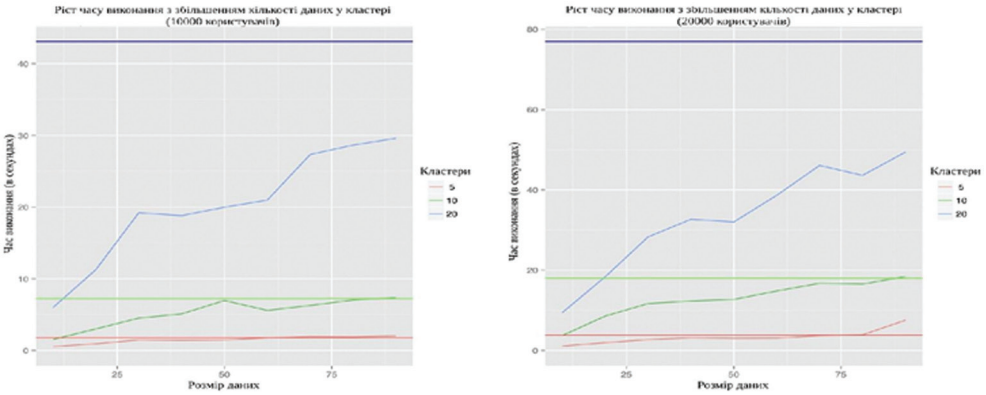

Рис. 11. Порівняння швидкості роботи *k*-means і *k*-means mini batch зі збільшенням кількості вхідних даних

Визначення схожості акаунтів потребує атомізованого типу даних, а оскільки в *k*-means використовують вектори, то дані потрібно формувати саме через них. Після того, як дані було зведено до векторної моделі, необхідно визначити індекс схожості між акаунтами. Схему методу пошуку схожих акаунтів у соціальних мережах за допомогою кластеризації зображено на рис. 12.

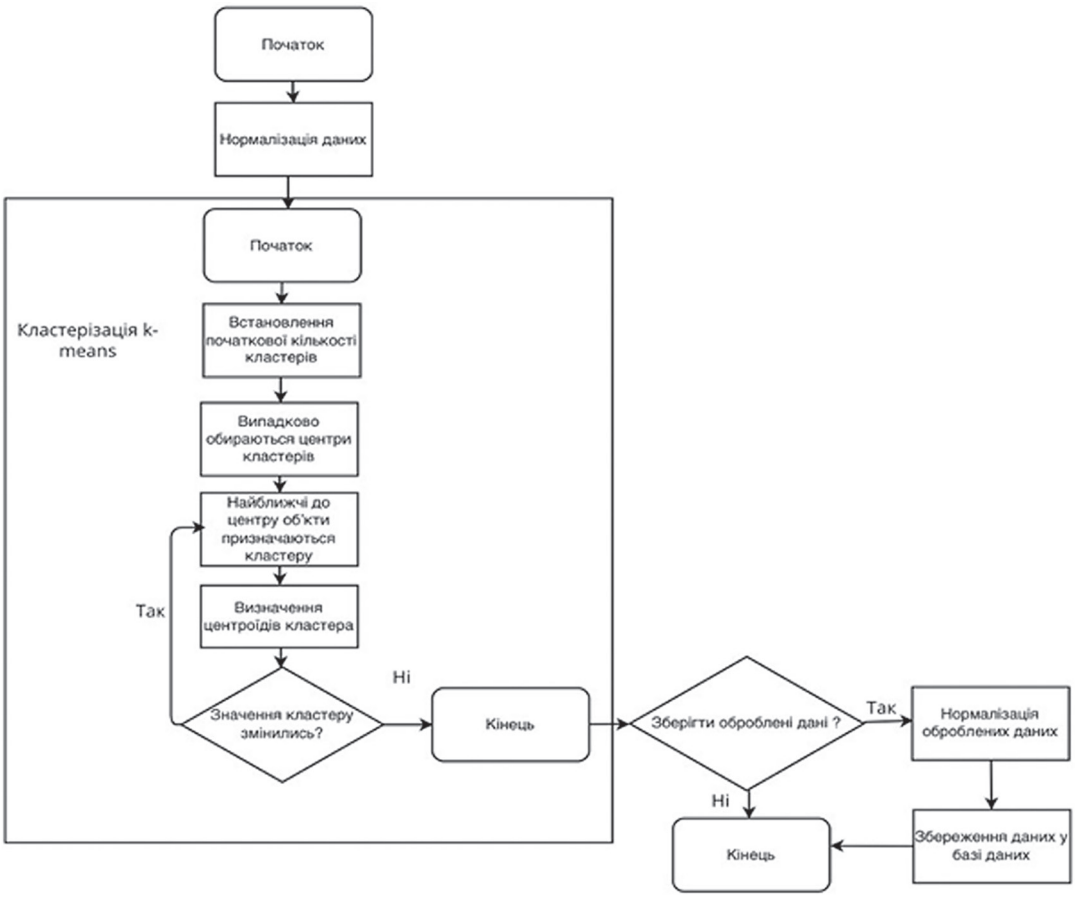

Рис. 12. Схема методу пошуку схожих акаунтів у соціальних мережах за допомогою кластеризації

Правильного алгоритму кластеризації немає, тому використовувати кластеризацію потрібно ґрунтуючись на бізнес-логіці проєкту, в якій вони будуть використовуватись, оскільки саме в цій логіці буде визначено, яким саме буде вхідний масив даних та за якими рисами потрібно групувати інформацію.

Алгоритм кластеризації для конкретної задачі часто доводиться вибирати експериментально, якщо немає математичної причини віддати перевагу одній кластерній моделі іншій. Алгоритм, розроблений для одного типу моделі, зазвичай не працює на наборі даних, що містить зовсім інший тип моделі.

Тобто ці алгоритми не забезпечують єдине розбиття набору даних, а натомість забезпечують велику ієрархію кластерів, які зливаються один з одним на певних відстанях.

Висновки

Аналіз методів кластеризації показав, що не існує універсального методу для аналізу кластеризації даних. Було проаналізовано різні методи кластеризації. Проведений порівняльний аналіз алгоритмів *k*-means та *k*-means mini batch встановив, що час виконання *k*-means mini batch значно менший для великої кількості вхідних даних.

Список використаної літератури

1. ***Internet*** *users 2017 forecast [Електронний ресурс]. URL: https://stats.areppim.com/stats/stats_internetxfcstx2017.htm*

2. ***Hadoop:*** *The Definitive Guide. T. White, 2012.*

3. ***Association*** *Rule Mining: Models and Algorithms. Chengqi Zhang, 2002.*

4. ***Big data*** *market forecast worldwide from 2011 to 2026 [Електронний ресурс]. URL: https://www.statista.com/statistics/255970/global-big-data-market-forecast-by-segment/*

5. ***Hari S., Narasimha M., Tripathy B. K.*** *Modern Technologies for Big Data Classification and Clustering. IGI Global, 2015 p.*

6. ***Grid-Based*** *Clustering - STING, WaveCluster & CLIQUE [Електронний ресурс]. URL: https://www.datamining365.com/2020/04/grid-based-clustering.html*

I. M. Averichev, D. A. Dibrii

DEVELOPMENT OF A METHOD FOR SEARCHING FOR SIMILAR ACCOUNT IN SOCIAL NETWORKS BASED ON CLUSTER ANALYSIS

In recent decades, the number of Internet users has been growing significantly. At the same time, the number of people using social networks to communicate and receive information is growing. The active development of social networks has also had a significant impact on the advertising market, as the information about users collected by these networks allows advertisers to target ads with great precision, which has a significant impact on the growth of sales of goods and services. After all, by identifying similar accounts, you can easily determine that a product that one user likes is likely to attract another.

In order to select the most appropriate method of cluster analysis for the chosen topic, a systematic analysis of cluster analysis methods was carried out. It has been analyzed that one of the approaches to finding similar accounts in social networks is to analyze accounts by certain characteristics and is widely used by modern companies that own social networks for targeted advertising. The analysis of clustering methods showed that there is no specific universal method for analyzing data clustering, which is why various clustering methods were analyzed and it was concluded that the improved k-means algorithm — k-means mini batch is best suited for finding similar accounts.

Keywords: cluster analysis method; information technology; cluster; clustering methods; simplified clustering algorithm; statistical analysis method; system analysis; big data.