СТАТИСТИКА

УДК 519.237.8

МЕТОДИЧНІ ЗАСАДИ ВИКОРИСТАННЯ КЛАСТЕРНОГО АНАЛІЗУ

METHODOLOGICAL PRINCIPLES OF CLUSTER ANALYSIS USE

Пономаренко І.В.
кандидат економічних наук, доцент,
доцент кафедри економічної кібернетики та маркетингу,
Київський національний університету технологій та дизайну

Бойко Д.Ю.
магістр,
Київський національний університету технологій та дизайну

Статтю присвячено особливостям використання методів кластерного аналізу для виокремлення груп, що містять подібні за певними показниками одиниці сукупності. Доведено важливість використання зазначеного методу в межах комплексного статистичного аналізу соціально-економічних явищ та процесів. Наведено ключові етапи виконання кластеризації, що дасть змогу сформувати групи на основі науково обґрунтованих підходів та розробити згідно з отриманими результатами ефективні управлінські рішення.
Ключові слова: *вибірка, дендрограма, кластерний аналіз, одиниці сукупності, система показників.*

Статья посвящена особенностям использования методов кластерного анализа для выделения групп, содержащих подобные по определенным показателям единицы совокупности. Доказана важность использования указанного метода в рамках комплексного статистического анализа социально-экономических явлений и процессов. Приведены ключевые этапы выполнения кластеризации, что позволит сформировать группы на основе научно обоснованных подходов и разработать согласно полученным результатам эффективные управленческие решения.
Ключевые слова: *выборка, дендрограмма, кластерный анализ, единицы совокупности, система показателей.*

The article is devoted to the peculiarities of cluster analysis methods use for the selection of groups containing similar in certain units of aggregate. The importance of using this method in the comprehensive statistical analysis framework of socio-economic phenomena and processes is proved. The key stages of clusterization implementation are described, which will allow forming groups based on scientifically based approaches and develop effective management decisions according to the obtained results.
Key words: *sampling, dendrogram, cluster analysis, aggregate units, system of indicators.*

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями. У сучасних умовах компанії накопичують великі обсяги інформації, яка знаходиться у структурованому, напівструктурованому та неструктурованому вигляді. Зазначені дані необхідно розглядати як цінний ресурс, який можливо використати після застосування комплексу економіко-матетематичних та статистичних методів для прийняття ефективних управлінських рішень. На етапі дослідження будь-яких даних необхідно використати комплекс методів, що дають змогу виявити структуру. Цілий клас методів орієнтований на ідентифікацію певних груп у досліджуваній сукупності. Класифікація дає можливість створити певну кількість груп, у кожній з яких одиниці сукупності характеризуються спільними характеристиками згідно з наявною системою показників. Одним із методів, який широкого використовується для проведення групування, є кластерний аналіз. Науковцями розроблено значну кількість методів кластеризації, а переваги їх використання пояснюються відсутністю суворих обмежень до якості даних та простотою інтерпретації отриманих результатів.

Аналіз останніх досліджень і публікацій, в яких започатковано розв'язання цієї проблеми і на які спираються автори. Дослідженню питань

використання методів кластерного аналізу присвячено праці таких іноземних учених, як: М. Алдендерфер, Д. Гарсон, Г. Джеймс, Б. Еверіт, А. Кассамбара, Р. Кінг, С. Ландау, М. Лісі, М. Мейла, Ч. Ромесбург, Д. Стахл, К. Хеннінг та ін. Проте існує потреба в постійному дослідженні еволюції методів кластеризації та виявленні особливостей їх застосування для розподілу одиниць сукупності на групи згідно з науково обґрунтованими підходами і виходячи зі специфіки явищ та процесів, що вивчаються.

Формулювання цілей статті **(постановка завдання)**. Мета статті – дослідити й узагальнити сучасні методичні підходи до використання кластерного аналізу для групування сукупності одиниць за системою статистичних показників.

Виклад основного матеріалу дослідження з повним обґрунтуванням отриманих наукових результатів. Розвиток суспільства нерозривно пов'язаний із використанням інформаційних технологій, що дають можливість накопичувати великі обсяги інформації щодня. Джерелами продукування даних виступають країни, види економічної діяльності, компанії та громадяни. Інформацію отримують у графічній формі зі супутників, фото- та відеоапаратури різноманітних компаній, які застосовують отримані візуалізації як дані. Медичні центри та спеціалізовані наукові лабораторії проводять значну кількість досліджень у сфері біомедицини, що дає змогу накопичувати великі обсяги відповідної статистичної інформації. Маркетингові компанії в рамках реалізації комплексних стратегій розвитку проводять складні дослідження ринків та населення для визначення ефективних інструментів налагодження комунікацій із цільовою аудиторією та отримання конкурентних переваг, а в процесі накопичують великі масиви даних. Важливим джерелом генерування інформації виступає мережа Інтернет, оскільки щоденно користувачі у соціальних мережах публікують графічні, аудіо– та відеоматеріали, додаючи відповідні коментарі, а компанії розміщують на власних ресурсах тематичний контент. У багатьох компаніях різноманітні процеси безперервно або періодично вимірюються за допомогою спеціалізованих приладів, що призводить до накопичення значної інформації [1; 2].

Великі дані є цінним ресурсом, проте необхідно проводити комплексне дослідження отриманої інформації за допомогою математичних та статистичних методів, а на основі здійсненого аналізу розробляти ефективні управлінські рішення виходячи з особливостей прикладної сфери. Одним із важливим напрямів аналізу виступає класифікація досліджуваної сукупності об'єктів на групи виходячи з наявних ознак та мети дослідження [3]. Ефективним методом для вирішення зазначеного завдання є кластерний аналіз, який дає змогу вирішити низку важливих питань класифікації (рис. 1).

Кластерний аналіз застосовується у багатьох сферах: антропології, археології, біології, геології, державному управлінні, маркетингу, медицині, психології, соціології, філології, хімії тощо. Розглянемо приклади застосування даного методу:

– маркетинг – сегментування ринку за окремими показниками: продуктами, соціально-еко-

Рис. 1. Основні завдання кластерного аналізу [4]

номічними характеристиками споживачів, наявними на ринку конкурентами тощо;

– банківські установи – визначення різноманітних груп клієнтів із певними типовими рисами поведінки, що дає можливість ідентифікувати сумлінних і неблагонадійних позичальників кредитів, а також розробляти спеціалізовані банківські продукти для окремих категорій користувачів;

– страховий бізнес – формування типових профілів клієнтів для подальшого віднесення майбутніх клієнтів до групи з певним рівнем ймовірності настання страхового випадку, а також створення спеціалізованих страхових продуктів на основі отриманих класифікації потенційних споживачів даних послуг;

– медицина – групування типових клінічних випадків із віднесенням їх згідно з виявленою симптоматикою до певного діагнозу та формування специфічної процедури лікування; класифікація медико-біологічних об'єктів;

– телекомунікаційний бізнес – виявлення однорідних груп споживачів телекомунікаційного контенту та формування спеціалізованих пакетів послуг;

– соціологія – групування респондентів, які надали відповіді на тематичні питання, за певними ознаками для виявлення відношення населення до важливих соціально-економічних процесів тощо [4].

Існує понад 100 алгоритмів кластеризації, на рис. 2 представлено основні методи кластерного аналізу, які використовуються в сучасних умовах для розподілу одиниць сукупності на групи. Наведені методи володіють певними недоліками та перевагами, а їх використання дає можливість отримати різноманітні результати групування одиниць сукупності. Під час вирішення реальних завдань дослідники вибирають найкращий варіант кластеризації виходячи із цілей дослідження та наявних обмежень (часових, матеріальних тощо).

Під час проведення кластерного аналізу виконуються такі етапи:

1. Формування мети дослідження та обґрунтування доцільності використання методів кластерного аналізу для групування одиниць сукупності у досліджуваних явищах.

2. Збір необхідної інформації, бажано на основі науково обґрунтованої статистичної методології. Формування бази даних передбачає використання певної системи показників та відбір репрезентативної вибіркової сукупності. Іноді аналітикам доводиться досліджувати неструктуровані дані, а також використовувати малі вибірки, що пояснюється специфікою досліджуваних явищ або неможливістю поліпшити процедуру збору інформації в певних умовах. Однією з переваг кластерного аналізу є можливість його використання до зазначених даних.

3. Стандартизація даних або використання методу експертних оцінок. Здебільшого показники, що передбачається використовувати під час кластеризації, мають різну розмірність. За використання фактичних даних показники з великою

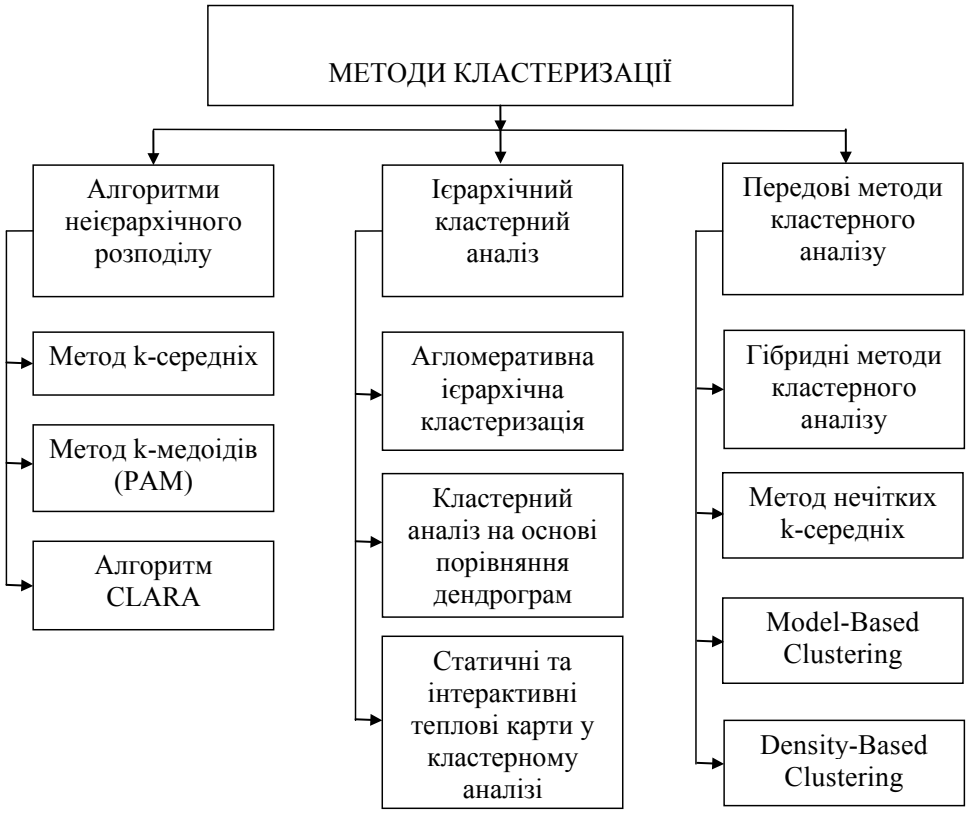

Рис. 2. Основні методи кластерного аналізу [5]

розмірністю істотно впливають на формування кластерів порівняно з показниками з малою розмірністю. Для нівелювання чинника розмірності дані доцільно стандартизувати. Метод експертних оцінок доцільно застосовувати для надання різної ваги показникам, що використовуються в процесі кластеризації. У процесі кластеризації відбір показників можливий на основі обчислення *p*-value. Поряд із цим у процесі реалізації процедури кластерного аналізу можливо перевіряти будь-які групування сукупності шляхом перебору різноманітних комбінації показників, що включаються в розрахунки.

4. Визначення кількості кластерів. Для вибору оптимального числа кластерів може бути використаний метод «локтя», графік GAP-статистики, графік ширини силуетів та ін. У процесі відбору кількості кластерів необхідно орієнтуватися не лише на результати, які було отримано після проведення алгоритму підрахунку. Іноді методи розділяють сукупність на кластери, в яких міститься лише одна одиниця сукупності, що вважається недопустимим. Під час розподілу сукупності з кількістю кластерів, що містять одиничні елементи, необхідно використати інший метод виділення кількості груп.

5. Перевірка адекватності кластеризації. Існує декілька підходів до валідації кластерів:

– зовнішня валідація, яка полягає у порівнянні підсумків кластерного аналізу із заздалегідь відомим результатом (тобто мітки кластерів відомі апріорі);

– відносна валідація, яка оцінює структуру кластерів, змінюючи різні параметри одного і того ж алгоритму (наприклад, число груп k);

– внутрішня валідація, яка використовує внутрішню інформацію процесу об'єднання в кластери (якщо зовнішня інформація відсутня);

– оцінка стабільності об'єднання в кластери (або спеціальна версія внутрішньої валідації), яка використовує методи ресемплінгу [6].

6. Використання адекватних методів кластерного аналізу та відбір найкращого згідно з потребами дослідження. Під час відбору найкращого варіанту кластеризації доцільно скористатися отриманими в ході розрахунків візуалізаціями. У разі використання ієрархічного кластерного аналізу буде створено дендрограми, які дають змогу наочно оцінити виокремлені групи одиниць сукупності. Поряд із цим застосовуються методи кластерного аналізу, які дають змогу наносити на двовимірний простір одиниці сукупності, що розподілені на групи за великою кількістю показників (більше двох). У цьому разі n-вимірний простір за допомогою спеціалізованих методів кластерного аналізу наближено переноситься на двовимірну площину. На основі отриманих результатів дослідниками вибираються кластеризації, які вважаються оптимальними у даних умовах.

7. Застосування методів статистичного аналізу для дослідження виокремлених кластерів та формування висновків і рекомендацій. Для сформованих у результаті кластеризації груп необхідно обчислити основні показники: середні значення, дисперсії, моди, медіани, коефіцієнти варіації тощо. Для проведення візуального порівняльного аналізу між кластерами здебільшого застосовується графік типу Box Plot. Також для отриманих груп можливо будувати регресійні рівняння, що дає змогу оцінити взаємозв'язки між чинниками в кожній сукупності [7]. На основі обчислених показників проводиться порівняльний аналіз виділених кластерів із зазначенням наявних між ними відмінностей.

Отже, кластерний аналіз широко застосовується у різноманітних сферах, оскільки дає змогу проводити групування одиниць у різних за обсягами сукупностях. До переваг кластеризації належать відсутність жорстких вимог до первинних даних, наявність значної кількості методів виділення груп та можливість створення різноманітних графіків з ідентифікованими кластерами.

Висновки з цього дослідження і перспективи подальших розвідок у даному напрямку. Кластерний аналіз належить до одного з важливих методів пошуку взаємозв'язків між одиницями певної сукупності за умови використання багатовимірних даних. Гнучкість зазначено підходу дає можливість змінювати набір показників та використовувати різноманітні методи кластерного аналізу, застосовуючи для аналізу найкращі з позиції аналітика результати. Отримані у результаті кластеризації групи характеризуються схожими рисами для одиниць сукупності, які входять до окремої групи, згідно з використаними показниками, що дає можливість розробляти ефективні управлінські рішення для кожного кластера.

БІБЛІОГРАФІЧНИЙ СПИСОК:

1. Kaggle Datasets. URL: https://www.kaggle.com/datasets.

2. Чубукова О.Ю., Раллє Н.В. Складові інноваційної економіки – освіта, технологічні уклади, когнітивні технології. Науковий вісник Полісся. 2016. № 3(7). С. 130–133.

3. Поиск структуры в данных. URL: https://ru.coursera.org/learn/unsupervised-learning.

4. Brian S. Everitt, Sabine Landau, Morven Leese, Daniel Stahl. Cluster Analysis, 5th Edition. John Wiley & Sons, 2011. 346 p. URL: https://pdfs.semanticscholar.org/b05a/ca28ced3751d6302d8ce7396f80c90d612fd.pdf.

5. Kassambara A. Practical Guide to Cluster Analysis in R. New York: STHDA, 2017. 187 p.

6. Шитиков В.К., Мастицкий С.Э. Классификация, регрессия, алгоритмы Data Mining с использованием R. URL: https://github.com/ranalytics/data-mining.

7. Rzepka A., Ślusarczyk B. Correlation and dependence between: Business-Globalisation-Information Society and Global Society. International Journal of Management Invention. 2016. Volume 5. Issue 1. P. 39–45. ISSN 2319-8028.