

Погорєлова Т.В.

к.е.н., доцент

Одеський національний економічний університет (Україна)

Марчук Д.О.

здобувач першого (бакалаврського) рівня ВО

Одеський національний економічний університет (Україна)

ДОСЛІДЖЕННЯ ПРОБЛЕМИ КЛАСТЕРНОГО АНАЛІЗУ МНОЖИНИ ОБ'ЄКТІВ

Актуальність теми дослідження проблеми кластерного аналізу полягає в його широкому застосуванні у різних сферах, таких як маркетинг, біоінформатика, соціологія та економіка. Постійний розвиток методів кластеризації і необхідність адаптації їх до нових типів даних та вимог роблять це дослідження надзвичайно важливим для сучасної науки і практики.

Кластеризація – це метод аналізу даних, при якому об'єкти групуються в "кластери" на основі їхньої схожості. Метою є об'єднання подібних об'єктів та розділення тих, які відрізняються. Термін уперше введений Тріоном (Tryon) у 1939 році. На відміну від класифікації, кластерний аналіз не потребує попередніх припущень про дані і може працювати з різними типами даних. Головне, щоб змінні вимірювались у порівнюваних шкалах [1].

Основні задачі кластерного аналізу досить стандартні і включають: 1) розробку типології або класифікації; 2) дослідження схем групування об'єктів; 3) формування гіпотез на основі даних; 4) перевірку гіпотез для визначення правильності групування [1].

При практичному застосуванні даного методу вирішується декілька з наведених груп задач.

Кластерний аналіз вирішує ці задачі, створюючи розбиття, що задовольняє певному критерію оптимальності, який часто виражається цільовою функцією.

Методи кластеризації не є універсальними рішеннями, оскільки: 1) різні методи дають різні результати, і немає однозначного критерію перевірки правильності, хоча існують метрики (як-от Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index) для оцінки результатів; 2) визначення кількості кластерів є складною задачею, яку можна вирішити за допомогою різних методів, таких як метрика (Діаграма ліктя), аналіз "відстаней" між об'єктами та оцінка щільності об'єктів, що дають оптимальні результати для різних алгоритмів [2].

Аналіз наукових джерел [2; 3] дає можливість виокремити типи рішень проблеми кластерного аналізу множини об'єктів, а саме:

1. Алгоритми ієрархічної кластеризації – це метод групування об'єктів у ієрархічні кластери шляхом послідовного з'єднання або розділення залежно від їх схожості. Процес починається з окремих кластерів для кожного об'єкта, які поступово об'єднуються або розділяються до утворення ієрархічної структури. Алгоритми визначають відстань між кластерами і стратегії об'єднання або розділення: перші об'єднують найближчі кластери

до одного, другі розділяють до окремих об'єктів. Цей підхід дозволяє візуалізувати структуру кластерів у вигляді дендрограми, що спрощує аналіз взаємозв'язків між групами даних.

2. Параметричні методи кластеризації – передбачають припущення щодо параметрів розподілу даних у кластерах, використовуючи статистичні методи та теорію ймовірностей для призначення об'єктів до кластерів. Ці методи вимагають вибору конкретних моделей розподілу та припущень щодо їхніх параметрів. На відміну від них, непараметричні методи, такі як ієрархічна кластеризація або методи, засновані на відстанях, менш чутливі до припущень про форму та розподіл кластерів.

Проаналізувавши існуючі програмні продукти, виокремимо ті, які вирішують проблему кластерного аналізу множини об'єктів:

1. Рішення від компанії Dataviz – Dataviz.io надає онлайн-сервіси для аналізу даних у роздрібній торгівлі та харчовій промисловості, автоматизуючи аналіз показників продажів, визначення головних товарів, побудову стратегій розвитку, планування акцій та виявлення неприбуткових товарів за допомогою алгоритму k-means. Перевагами рішень Dataviz є надання повного апаратно-програмного комплексу для кластерного аналізу, побудова стратегій покращення та визначення їх недоліків. Недоліками є вузька специфіка та обмеженість використання лише алгоритму k-means, що може бути менш точним для специфічних датасетів.

2. Рішення від компанії sciencehunter – Sciencehunter.net надає інструменти аналітики, такі як візуалізація, гістограми з різними метриками та методами поєднання, і аналіз даних. Переваги сервісу включають інтерактивні графіки для зрозумілого пояснення даних, простий набір угруповань завдяки зручному інтерфейсу, і якісні візуальні налаштування. Недоліки полягають у використанні переважно ієрархічного (агломеративного) кластерного аналізу та обмеженні візуалізації лише методом PCA [3].

Завдяки здатності виявляти природні групи об'єктів без попередніх припущень, кластерний аналіз допомагає виявити приховані структури в даних, що сприяє більш точному прийняттю рішень і підвищенню ефективності бізнес-процесів. Проте, аналіз існуючих рішень показує, що наразі недостатньо програм, які б могли використовувати всі методи кластерного аналізу, візуалізувати результати і порівнювати їх за допомогою метрик, що дозволить покращити точність та ефективність аналізу.

Список використаних джерел

1. Sharma R. What is Clustering and Different Types of Clustering Methods [Електронний ресурс] / Rohit Sharma // upGrad. 2020 URL: <https://www.upgrad.com/blog/clustering-and-types-of-clustering-methods/>.
2. Wierzchoń S., Kłopotek M. Modern algorithms of cluster analysis, 2018. URL: <https://doi.org/10.1007/978-3-319-69308-8>
3. Brian S. Everitt. Cluster Analysis: Basic Concepts and Algorithms. Беверлі Хіллз : Sage Publications, 1984. Глава 8 .