Технічні науки

УДК 519.6, 519.8

Якимець Роман Вікторович
студент
Національний технічний університет України
«Київський політехнічний інститут»
Якимец Роман Викторович
студент
Национальный технический университет Украины
«Киевский политехнический институт»
Yakymets Roman V.
student
National Technical University of Ukraine «Kyiv Polytechnic Institute»

МЕТОДИ КЛАСТЕРИЗАЦІЇ ТА ЇХ КЛАСИФІКАЦІЯ
МЕТОДЫ КЛАСТЕРИЗАЦИИ И ИХ КЛАССИФИКАЦИЯ
METHODS OF CLUSTERING AND CLASSIFICATION

Анотація: Описано суть кластеризації. Досліджені методи кластеризації та їх особливості.

Ключові слова: кластеризація, добування даних, методи кластеризації, K-середніх.

Аннотация: Описана суть кластеризации. Исследованы методы кластеризации и их особенности.

Ключевые слова: кластеризация, добыча данных, методы кластеризации, K-средних.

Summary: Describe the essence of clustering. Investigated clustering methods and their features.

Key words: clustering , data mining , clustering methods, K-Means.

Вступ

Існує безліч способів застосування кластерного аналізу. Найчастіше він виступає як інструмент, що дозволяє поглянути на дані в цілому. Також кластерний аналіз може використовуватись для попередньої обробки або як проміжний етап інших алгоритмів, таких як класифікації або прогнозування, чи для data mining. В задачах data mining за допомогою кластерного аналізу створюється комплексне зведення даних для класифікації , відбувається виявлення шаблонів, формування і перевірка гіпотез і т.і. Крім того, кластерний аналіз часто застосовується для виявлення даних, що «вибиваються» з-поміж інших, оскільки таким даним відповідають точки, розташовані на відстані від будь-якого кластера . Також кластерний аналіз використовується для стиснення та узагальнення даних.

Кластерний аналіз

Кластер-колекція об'єктів даних містить схожі об'єкти в одному кластері. Це означає, що об'єкти є аналогічними один до одного в межах однієї групи, і в той же час вони досить різні, або пов'язані з об'єктами в іншій групі або в інших кластерах. Кластерний аналіз також називають кластеризацією або сегментацією даних. Кластерний аналіз розподіляє даний набір точок даних в набір кластерів або груп . Ці точки даних якомога більше схожі в межах однієї групи та віддалені наскільки це можливо від інших груп. Кластерний аналіз відноситься до навчання без вчителя (unsupervised learning) з огляду на те, що на початку немає визначених класів . Це суттєво відрізняє його від класифікації, де потребується навчання з учителем (supervised learning) або завдання міток класу для побудови моделі класифікації.

Класифікація алгоритмів кластеризації

Існує дві основні класифікації алгоритмів кластеризації :

1. Ієрархічні і неієрархічні (плоскі) . Ієрархічні алгоритми будують систему вкладених розбиттів , тобто на виході алгоритму представляється дерево кластерів, з коренем у якості всієї вибірки і листками – у якості найменших кластерів. Неієрархічні алгоритми будують лише одне розбиття об'єктів на кластери.

2. Чіткі і нечіткі.

Чіткі алгоритми надають всім об'єктам вибірки відповідний номер кластера, що означає , що кожен об'єкт повинен відноситься лише до одного кластеру.

Нечіткі алгоритми надають кожному об'єкту у відповідність набір значень , які демонструють ступінь належності об'єкта до кластерів. Отже, кожен об'єкт відноситься до кожного кластеру з певною ймовірністю .

Плоскі методи на прикладі K-Means

Метод K-Means полягає в тому, щоб виявити угруповання в даних . Вхідна множина розділяється на К груп, при цьому мінімізується функція, що визначає відстані як суми квадратів помилок – Sum of Squared Errors (SSE):

$$\mathrm{SSE(C)} = \sum_{j=1}^{K} \sum_{i=1}^{n_j} \|x_i^j - \mathrm{c}_j\|^2$$

Після цього ітеративно оптимізується якість такого поділу. Таким чином, К-секціонування – це метод, що розділяє набір даних D з n об'єктів в набір К кластерів.

Кожний кластер представляється центром кластера. Для К кластерів метод K-Means працює наступним чином :

1. Обирає К точок центроїдами.

2. В циклі виконує наступні дії до того моменту, поки не досягає критерія збіжності:

a. Формує К кластерів шляхом присвоєння кожної точки до найближчого до неї центроїда.

b. Перевизначає центроїди.

3. Алгоритм може використовувати різні міри відстані, наприклад Манхеттенську, Евклідову відстані.

Особливості методу полягають в наступному:

- обчислювальна складність O(tKn), де n – це кількість об’єктів, K – кількість кластерів, t – кількість ітерацій. Звичайно K, t<< n, тобто метод є ефективним;
- кластеризація може завершитись на локальному оптимумі, тому для високоякісного результату необхідна початкова ініціалізація;
- необхідно заздалегідь задати К – кількість кластерів;
- чутливість до «шумних» даних та значень, що сильно відрізняються;
- можливе застосування тільки для чисельних даних;
- неможливо будувати кластери неопуклої форми.

На даний час існує багато варіацій цього методу, що частково усувають недоліки, серед них: K-Medoids, K-Medians, K-Modes, K-means++, Intelligent K-Means, Genetic K-Means .

Ієрархічні методи кластеризації

Ієрархічна кластеризація – це така кластеризація, за якої, починаючи з кластера, що складається з одного елемента, кластери ітеративно зливаються в кластери вищого рівня . Також можливо починати з єдиного великого макрокластера, який ітеративно розділяється на маленькі кластери. Таким чином формуються ієрархія кластерів. Для їх формування не потрібно задавати кількість кластерів К, такий тип кластеризації є більш детермінованим та не потребує ітеративних уточнень. Ієрархічні методи кластеризації включають в себе дві категорії

алгоритмів. Перша категорія має назву агломераційної. Вона починається з одноелементного кластера, що зливаються два кластери, щоб побудувати ієрархію кластерів «знизу вверх». Друга група – Divisive methods – методом розділення великий макрокластер, що містить всі елементи, розділяється на дві групи, кожна з них також на дві групи і так далі. Таким чином генерується ієрархія кластерів «зверху вниз». На рисунку 1 зображено обидва підходи.

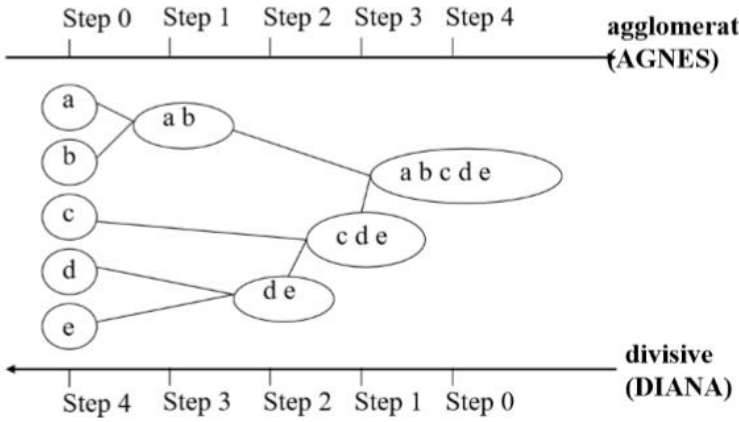

Рисунок 1 – Хід алгоритмів AGNES, DIANA

Агломеративні методи ієрархічної кластеризації

Агломеративні алгоритми – це такі, що кластеризують «знизу верх». На початку алгоритму кожна точка розглядається як кластер, потім алгоритм намагається об'єднати найближчі сусідні точки в один більший кластер і так далі, щоб зрештою об'єднати всі кластери в один великий кластер. Агломеративні алгоритми також називають AGNES (AGglomerative NESting) . Хід алгоритму виглядає наступним чином(рис.2):

- використовується метод одноканального зв'язку «найближчий сусід» та матриця відмінностей;
- вузли, що мають найменші відмінності, зливаються;
- всі вузли об'єднуються в один кластер.

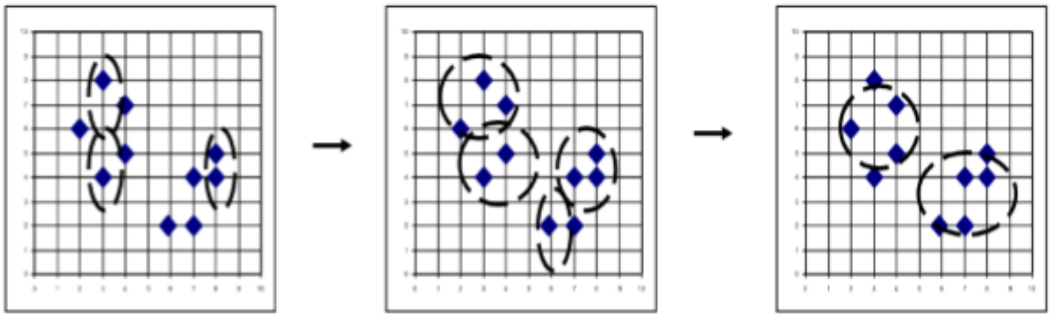

Рисунок 2 – Хід алгоритму AGNES

Агломеративна кластеризація залежить від використання мір подібності кластерів:

- одноканальний зв'язок (найближчий сусід);
- повний зв'язок (діаметр);
- середній зв'язок (середнє по групі);
- центроїдний зв'язок (подібність центроїдів).

Одноканальний зв'язок (найближчий сусід):

- подібність двох кластерів – це подібність між їх найбільш подібними членами (найближчий сусід);
- приділяється увага найближчим точкам, ігнорується структура кластера;
- можливість будувати кластери неправильної форми;
- такий вид зв’язку чутливий до даних з шумами та значень, що вибиваються з множини. Схематичне зображення наведено на рисунку 3.

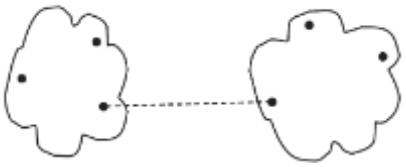

Рисунок 3 – Одноканальний зв'язок

Повний зв'язок:

- подібність двох кластерів рахується як подібність їх найменш подібних членів;
- два кластери об’єднуючись формують кластер з щонайменшим діаметром;
- на виході – кластери компактної форми;
- чутливий до значень, що суттєво відрізняються.

Схематичне зображення наведено на рисунку 4.

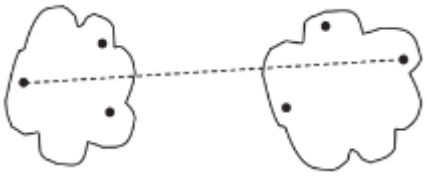

Рисунок 4 – Повний зв'язок

Середній зв'язок – середня відстань між елементами в парі кластерів (рис. 5). Особливістю є затратне обчислення.

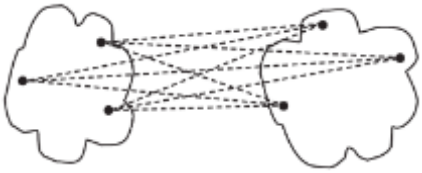

Рисунок 5 – Середній зв'язок

Центроїдний зв’язок – відстань між центроїдами двох кластерів (рис. 6). Даний алгоритм не потребує задання кількості кластерів та дозволяє об’єднувати в кластери дані без істотних змін через значення, що вибиваються, та шуми.

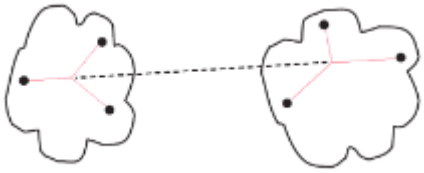

Рисунок 6 – Центроїдний зв'язок

Методи ієрархічної кластеризації розділенням

DIANA (Divisive Analysis) – зворотній порядок дій від AGNES: в результаті кожний елемент представляє собою кластер. На рис. 7 зображено хід алгоритму DIANA.

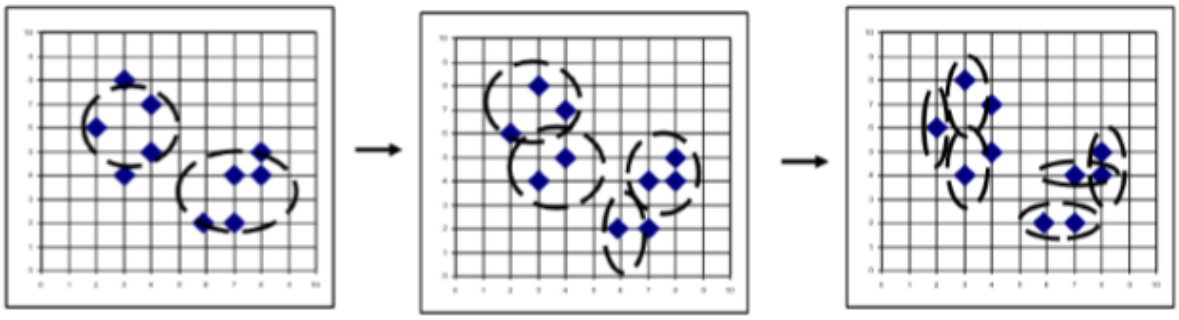

Рисунок 7 – Хід алгоритму DIANA

Ієрархічна кластеризація розділенням – це підхід «зверху вниз»:

- процес починається з кореня, розглядаючи всі точки множини як кластер;
- кластери вищого рівня рекурсивно розщеплюються для побудови діаграми;
- може розглядатись в якості глобального підходу;
- може вважатись ефективнішим, але більш чутливим до шумів за AGNES.

Висновки

В даній статті були описані та класифіковані деякі методи кластеризації. А саме , плоскі методи на прикладі K-Means та ієрархічні методи кластеризації. Також описані особливості цих методів.

Література:

1. A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Prentice Hall, 1988

2. R. Ng and J. Han. Efficient and Effective Clustering Method for Spatial Data Mining. VLDB'9

3. L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, 199

4. Котов А. Кластеризация данных./ Котов А., Красильников Н. 2006.