КЛАСТЕРНИЙ АНАЛІЗ. ІЄРАРХІЧНІ КЛАСТЕР-ПРОЦЕДУРИ

Аліна Алєксєєвець, Анатолій Плічко

Науковий керівник: д. фіз.-мат. наук, професор Плічко А.М.

У статті розглянуто історію становлення кластерного аналізу, основні поняття та методи. Детально охарактеризовано поняття ієрархічних кластерних процедур. Показано переваги та недоліки використання агломеративних та дивізимних алгоритмів кластерного аналізу. .Наведено приклад розв'язання задачі класифікації шести об'єктів довільної природи, кожен з яких характеризується двома ознаками. Для розв'язання використано метод ієрархічних кластер-процедур. Детально описано процедуру класифікації на основі вказаного методу та результати представлено у вигляді дендродрами.

Ключові слова: *кластерний аналіз, дендрограма, ієрархічні кластер-процедури, агломеративний та дивізимний метод.*

CLUSTER ANALYSIS. HIERARCHIC CLUSTER PROCEDURES

Alina Alekseevets, Anatoliy Plichko

Scientific adviser: Doctor of Phys. and Math. sciences, Professor Plichko A.M.

The history of cluster analysis, basic concepts and methods are considered in the article. The concept of hierarchical cluster procedures is described in detail. The advantages and disadvantages of using agglomerative and divisive algorithms of cluster analysis are shown. An example of solving the problem of classification of six objects of arbitrary nature, each of which is characterized by two features. The method of hierarchical cluster of procedures is used for the decision. The classification procedure based on this method is described in detail and the results are presented in the form of arboretums.

Key words*: cluster analysis, dendrogram, hierarchical cluster procedures, agglomerative and divisional methods.*

Мета статті: визначити що являє собою кластерний аналіз та ієрархічні кластер процедури, навести приклад задачі, яку можна розв'язати цим методом, а також визначити для чого необхідний саме такий метод кластерного аналізу.

Основні результати дослідження. Сам кластерний аналіз запропонував американський вчений Тріон ще у 1939 р. Дослівно термін «кластер» з англійської "cluster" означає група. У 60-х роках минулого століття відбувся бурхливий розвиток кластерного аналізу, передумовами якого стали поява

швидкісних комп’ютерів та визнання класифікацій фундаментальним методом наукових досліджень. *Кластерним аналізом* називається завдання розбиття об’єктів на підмножини (кластери) так, щоб кожен кластер складався зі схожих об’єктів, а об’єкти різних кластерів істотно відрізнялися (Рис.1).

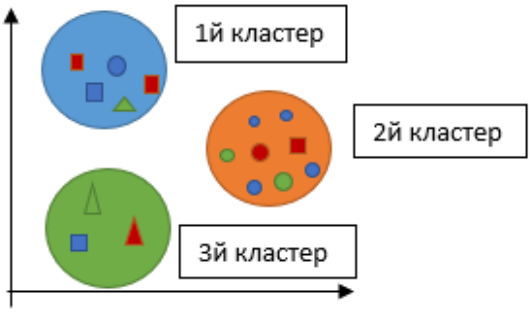

Рис.1. Класифікація на три кластери

Подані об’єкти прості і мають обмежену кількість характеристик (координати, форма, колір). Залежно від того, які характеристики використовуються для групування, кластеризація може дати різні результати.

Кластерний аналіз має два методи: ієрархічний та неієрархічний. Неієрархічні методи мають за основу вже задану кількість кластерів та використовують складні алгоритми знаходження їхньої кількості. Серед неієрархічних методів кластеризації особливої уваги заслуговують ітеративні методи. Ітеративні методи працюють безпосередньо з первинними даними. Тому за їх допомогою можна обробляти доволі великі обсяги даних.

Суть ієрархічної кластеризації полягає в послідовному об’єднанні менших кластерів у більші або поділі більших кластерів на менші. Ієрархічні алгоритми використовуються в задачах класифікації невеликого числа об’єктів (в основному до 150 об’єктів), де основний інтерес представляє аналіз структури множини об’єктів і наочна інтерпретація проведеного аналізу у вигляді дендрограми (дерево об’єднання кластерів). Дендрограма це деревоподібна діаграма, яка містить *n* рівнів, кожен з яких відповідає одному з кроків процесу послідовного укрупнення кластерів. Дендрограма являє собою групи об’єктів, які змінюються на різних рівнях ієрархії (Рис.2).

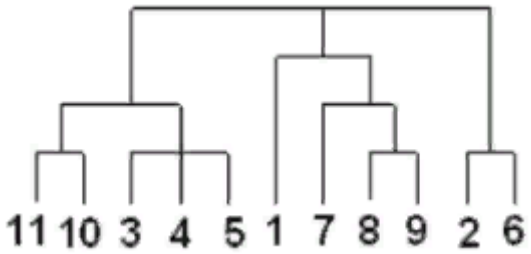

Рис.2. Приклад дендрограми.

Числа 11, 10, 3 і т.д. – це номери об’єктів або спостережень вихідної вибірки. На першому кроці кожне спостереження представляє один кластер (вертикальна лінія), на другому кроці спостерігаємо об’єднання таких спостережень: 11 і 10; 3, 4 і 5; 8 і 9; 2 і 6. На другому кроці продовжується об’єднання в кластери: спостереження 11, 10, 3, 4, 5 і 7, 8, 9. Цей процес триває доти, поки всі спостереження не об’єднаються в один кластер.

Ієрархічні (деревоподібні) процедури разом є з їхніми реалізаціями на комп’ютерах є найбільш поширеними алгоритмами кластерного аналізу. Вони бувають агломеративними і дивізимними. В агломеративних процедурах початковим є розбиття, яке складається з *n* одноелементних класів, а кінцевим з одного класу; у дивізимних навпаки. Опишемо принцип роботи ієрархічних кластер процедур:

1) для агломеративних він полягає в послідовному об’єднанні елементів спочатку найближчих, а потім все більше віддалених один від одного;

2) для дивізимних він полягає в послідовному поділі елементів спочатку далеких, а потім все більш близьких один до одного.

Якщо говорити про переваги та недоліки ієрархічних кластер-процедур, то можна сказати, що:

- недоліком є громіздкість обчислювальної реалізації ієрархічних процедур;
- перевагами є більш повний і точний аналіз структури досліджуваної множини спостережень, а також можливість наочної інтерпретації проведеного аналізу на основі дендрограми.

Приклад. Проведемо класифікацію 6-ти об’єктів (*n=6*), кожен з яких характеризується двома ознаками(Табл.1 та Рис.3):

Таблиця 1.

Постановка задачі класифікації

№ об'єкта *i*	1	2	3	4	5	6
x_{i_1}	5	6	5	10	11	10
x_{i_2}	10	12	13	9	9	7

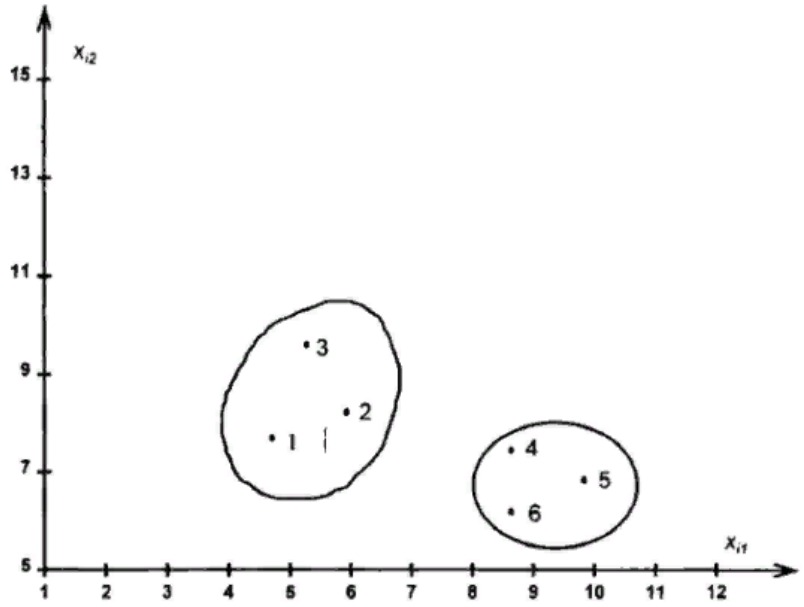

Рис.3. Постановка задачі.

Використаємо агломеративний ієрархічний алгоритм класифікації. Нехай відстань між об’єктами – звичайна евклідова відстань, тоді за формулою:

$$\rho(X_i, X_j) = \sqrt{\sum_{l-1}^{k} (x_{il} - x_{jl})^2}$$

відстань між I та II об’єктами буде дорівнювати:

$\rho_{12} = \sqrt{(5-6)^2 + (10-12)^2} = \sqrt{5} = 2{,}24,$

відстань між I та III об’єктами буде дорівнювати:

$\rho_{13} = \sqrt{(5-5)^2 + (10-13)^2} = 3.$

Очевидно, що $\rho_{11} = 0.$

Аналогічно знаходимо відстань між шістьма об’єктами і будуємо матрицю відстані:

$$R_1 = \{\rho(X_i, X_j)\} = \begin{pmatrix} 0 & 2{,}24 & 3 & 5{,}10 & 6{,}08 & 5{,}83 \\ 2{,}24 & 0 & 1{,}41 & 5 & 5{,}83 & 6{,}40 \\ 3 & 1{,}41 & 0 & 6{,}40 & 7{,}21 & 7{,}81 \\ 5{,}10 & 5 & 6{,}40 & 0 & 1 & 2 \\ 6{,}08 & 5{,}83 & 7{,}21 & 1 & 0 & 2{,}24 \\ 5{,}83 & 6{,}40 & 7{,}81 & 2 & 2{,}24 & 0 \end{pmatrix}$$

З матриці відстаней випливає, що четвертий та п’ятий об’єкти найближчі: $\rho_{4,5} = 1{,}00$ і тому об’єднуються в один кластер.

Після об’єднання об’єктів маємо п’ять кластерів:

Номер кластера	1	2	3	4	5
Склад кластера	(1)	(2)	(3)	(4,5)	(6)

Відстань між кластерами визначимо за принципом «найближчого сусіда», використавши формулу перерахунку:

$$\rho_{l,(m,q)} = \rho(S_l, S_{(m,q)}) = \alpha\rho_{lm} + \beta\rho_{lq} + \gamma\rho_{mq} + \delta|\rho_{lm} - \rho_{lq}|$$

Відстань між об’єктом S_1 і кластером $S_{(4,5)}$:

$$\rho_{1,(4,5)} = \rho(S_1, S_{(4,5)}) = \frac{1}{2}\rho_{14} + \frac{1}{2}\rho_{15} - \frac{1}{2}|\rho_{14} - \rho_{15}|$$

$$= \frac{1}{2}[5{,}10 + 6{,}08] - \frac{1}{2}[|5{,}10 - 6{,}08|] = 6{,}08$$

Отже, відстань $\rho_{1,(4,5)}$ дорівнює відстані від об’єкта 1 до найближчого до нього об’єкта, який входить в кластер $S_{(4,5)}$ тобто $\rho_{1,(4,5)} = \rho_{1,4} = 6{,}08$. Тоді матриця відстані виглядатиме так:

$$R_2 = \begin{pmatrix} 0 & 2{,}24 & 3 & 5{,}10 & 5{,}83 \\ 2{,}24 & 0 & 1{,}41 & 5 & 6{,}40 \\ 3 & 1{,}41 & 0 & 6{,}40 & 7{,}81 \\ 5{,}10 & 5 & 6{,}40 & 0 & 2 \\ 5{,}83 & 6{,}40 & 7{,}81 & 2, & 0 \end{pmatrix}$$

Після об’єднання другого та третього об’єктів, відстань між якими найменша: $\rho_{2,3} = 1{,}41$, маємо чотири класи: $S_{(1)}$, $S_{(2,3)}$, $S_{(4,5)}$, $S_{(6)}$. Знаходимо матрицю відстаней. Для того, щоб порахувати відстань до кластера $S_{(2,3)}$ використаємо

матрицю відстаней R_2. Наприклад, відстань між кластерами $S_{(4,5)}$ та $S_{(2,3)}$ дорівнює:

$$\rho_{(4,5),(2,3)} = \frac{1}{2}\rho_{(4,5),2} + \frac{1}{2}\rho_{(4,5),3} - \frac{1}{2}\left|\rho_{(4,5),2} - \rho_{(4,5),3}\right| = \frac{5}{2} + \frac{6{,}40}{2} - \frac{1{,}40}{2} = 5$$

Провівши аналогічні підрахунки, отримаємо:

$$R_3 = \begin{pmatrix} 0 & 2{,}24 & 5{,}10 & 5{,}83 \\ 2{,}24 & 0 & 5 & 6{,}40 \\ 5{,}10 & 5 & 0 & 2 \\ 5{,}83 & 6{,}40 & 2 & 0 \end{pmatrix}$$

Об'єднаємо кластери $S_{(4,5)}$ та $S_{(6)}$, відстань між якими згідно з матрицею R_3 найменша: $\rho_{(4,5),6} = 2$. В результаті отримаємо три кластери: $S_{(1)}$, $S_{(2,3)}$, $S_{(4,5,6)}$. Матриця відстаней буде мати вигляд:

$$R_4 = \begin{pmatrix} 0 & 2{,}24 & 5{,}10 \\ 2{,}24 & 0 & 5 \\ 5{,}10 & 5 & 0 \end{pmatrix}$$

Об'єднаємо тепер кластери $S_{(1)}$ і $S_{(2,3)}$, відстань між якими $\rho_{1,(2,3)} = 2{,}24$. В результаті отримаємо два кластери $S_{(1,2,3)}$ і $S_{(4,5,6)}$, відстань між якими знайдена за принципом «найближчого сусіда», $\rho_{(1,2,3),(4,5,6)} = 5$.

Результати ієрархічної класифікації об'єктів зобразимо у вигляді дендрограми:

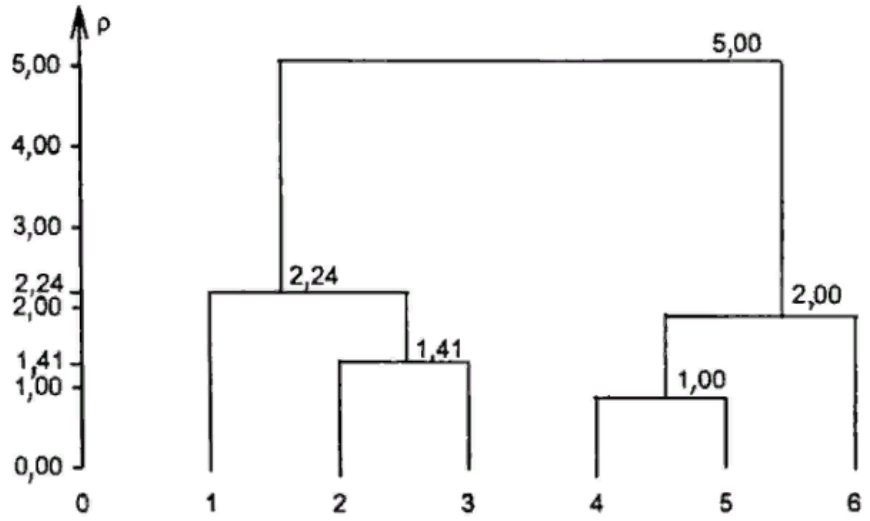

Рис.4. Дендрограма ієрархічної класифікації

На поданій дендрограмі вказано відстані між об’єднаними на даному етапі кластерами (об’єктами). В даному прикладі слід надати перевагу передостанньому етапу класифікації, коли всі об’єкти об’єднані в два кластери $S_{(1,2,3)}$ і $S_{(4,5,6)}$.

Висновки. Отже, кластерний аналіз являє собою технологію, що дозволяє розбити вхідні дані на класи – групи однотипних елементів вибірки, або (іншими словами) кластери – компактні області групування елементів вибірки у просторі ознак. Одним з методів кластерного аналізу є метод ієрархічних кластер-процедур, який використовується в задачах класифікації невеликого числа об’єктів, де основний інтерес представляє аналіз структури множини об’єктів та наочна інтерпретація проведеного аналізу у вигляді дендрограми.

Список використаних джерел

1. А.М. Дубров, В.С. Мхитарян, Л.И. Трошин, Многомерные статистические методы, Москва, “Финансы и статистика”, 2003

2. С.А.Айвазян, В.М.Бухштабер, И.С. Енюков, Л.Д. Мешалкин, Прикладная статистика. Классификация и снижение размерности, Москва, “Финансы и статистика”, 1989

3. Торопчина Г. Н., Двоерядкина Н. Н., Вохминцева Г. П. Элементы кластерного анализа. Учебное пособие. Благовещенск: Амурский гос. ун-т, 2006.