Кредитний модуль «Основи роботи з сучасними програмними комплексами 2. Статистичний аналіз і візуалізація даних»

Розділ 3. Статистичне оброблення даних у ППП STATISTICA

Конспект лекцій

Лекція 8.

Кластерний аналіз.

Реалізування кластерного аналізу в ППП STATISTICA

Прикладну статистику можна охарактеризувати як наукову дисципліну про прийоми, математичні методи й моделі обробляння статистичних даних з метою аналізування й ухвалювання рішень. Прикладна статистика охоплює

- статистику випадкових величин;
- багатомірний статистичний аналіз;
- статистику часових рядів та випадкових величин[1];
- статистику об'єктів нечислової природи.

Багатомірний статистичний аналіз вивчає оброблення багатомірних статистичних даних з метою виявлення взаємозв'язків, їх характеру та структури. До методів багатомірного статистичного аналізу можна віднести множинний кореляційний та регресійний аналіз, методи ***зниження розмірності*** багатомірного простору (наприклад, метод факторного аналізу), методи багатомірного ***класифікування*** (наприклад, методи кластерного й дискримінантного аналізу) та інші.

Метод класифікування полягає у розбиванні великої кількості об'єктів на однорідні групи. ***Класифікація*** – це певна закономірність, яка дає змогу зробити висновок щодо визначення характеристик цих груп. Розрізняють одномірне[2] і багатомірне класифікування. Для класифікування використовують різні методи:

- дерева розв'язків;
- баєсове класифікування;
- генетичні алгоритми;
- статистичні методи, зокрема, лінійну регресію;
- штучні нейронні мережі

та багато інших.

---

[1] елемент вибірки – функція
[2] за однією ознакою

Кластерний аналіз (автоматичне класифікування «без вчителя», класифікаційний аналіз, числова таксономія) об'єднує методи класифікування багатомірних об'єктів або подій у відносно однорідні групи (кластери). Об'єкти всередині кластеру повинні бути сході один на одного у більшій мірі ніж на об'єкти інших кластерів і відрізнятись від об'єктів інших кластерів більше ніж від об'єктів власного кластеру. Характеристиками кластеру є внутрішня однорідність та зовнішня ізольованість.

Центр кластеру – це середнє геометричне місце точок у просторі змінних. ***Радіус кластера*** – це максимальна відстань точок від центра кластера. Розмір кластера визначається його радіусом, середньоквадратичним відхиленням його об'єктів або кількістю об'єктів кластеру.

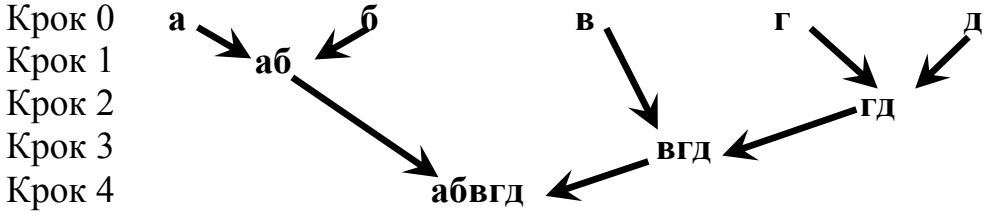
Кластери можуть перекриватись. Об'єкт, який неможливо однозначно віднести до одного з кластерів називається ***спірним***.

«Пре»пояснення

Алгоритм середнього зв'язку

На першому кроці кожний об'єкт розглядають як окремий кластер. На кожному наступному кроці об'єднують два найближчі кластери. Від між кластерами розраховують як середнє арифметичне відстаней між парами об'єктів, один з яких входить у перший кластер, а інший – у другий. Нарешті усі об'єкти об'єднують разом. Урезультаті отримуємо дерево послідовних об'єднань – дендрограму.

Дендрограма – деревовидна діаграма, яка містить n рівнів, кожний з яких відповідає одному з кроків процесу послідовного укрупнення кластерів.



Цей алгоритм відносять до агломеративних методів кластерного аналізу.

Міри подібності

Вирізняють такі міри подібності: А – коефіцієнти кореляції, Б – міри відстані (віддаленості, неподібності), В – коефіцієнти асоціативності, Г – ймовірнісні коефіцієнти подібності.

Введемо позначення: $d(C_i, C_j)$ – відстань між i та j об'єктами; C_{ik} – значення k змінної для i об'єкту; C_{jk} – значення k змінної для j об'єкту; $k = \overline{1,V}$, де V – кількість змінних, якими описують об'єкти.

1(Б) ***Евклідова відстань*** (англ. Euclidean distance) – одна з на використовуваніших типів відстаней

$$d_{Euc}(C_i, C_j) = \sqrt{\sum_k (C_{ik} - C_{jk})^2}. \quad (1)$$

2(Б) ***Квадратична евклідова відстань*** (англ. Squared euclidean distance)

$$\mathrm{d}^2_{\mathrm{Euc}}(\mathrm{C_i}, \mathrm{C_j}) = \sum_{\mathrm{k}} (\mathrm{C_{ik}} - \mathrm{C_{jk}})^2. \quad (2)$$

3(Б) ***Манхеттенвська відстань*** (англ. Manhattan distance) або відстань міських кварталів (англ. city-block distance)

$$\mathrm{d_{c-b}}(\mathrm{C_i}, \mathrm{C_j}) = \sum_{\mathrm{k}} |\mathrm{C_{ik}} - \mathrm{C_{jk}}|. \quad (3)$$

Міра впливу окремих викидів зменшується, оскільки вони не підносяться у квадрат.

4(Б) ***Відстань Махаланобіса*** (англ. Mahalanobis distance) пов'язана з кореляціями змінних

$$\mathrm{d_{Mah}}(\overrightarrow{\mathrm{C_l}}, \overrightarrow{\mathrm{C_j}}) = (\overrightarrow{\mathrm{C_l}} - \overrightarrow{\mathrm{C_j}})^{\mathrm{T}} \Sigma^{-1} (\overrightarrow{\mathrm{C_l}} - \overrightarrow{\mathrm{C_j}}), \quad (4)$$

де Σ – внутрішньогрупова дисперсійно-коваріаційна матриця.

Якщо кореляція між параметрами відсутня, відстань Махаланобіса еквівалентна квадратичній евклідовій відстані.

5(Б) ***Метрика домінування*** (супремум-норма) або відстань Чебишева (англ. Chebyshev distance),

$$\mathrm{d_{sup}}(\mathrm{C_i}, \mathrm{C_j}) = \max_{\mathrm{k}} |\mathrm{C_{ik}} - \mathrm{C_{jk}}|. \quad (5)$$

6(Б) Узагальнена ***метрика Мінковського***

$$\mathrm{d_{Min}}(\mathrm{C_i}, \mathrm{C_j}) = \left(\sum_{\mathrm{k}} |\mathrm{C_{ik}} - \mathrm{C_{jk}}|^{\mathrm{p}}\right)^{1/\mathrm{p}}. \quad (6)$$

Якщо *p*=2, то отримуємо евклідову відстань, якщо *p*=1 – Манхеттенівську відстань, якщо *p* прямує до нескінченності – відстань Чебишева.

7(Б) Степенева відстань (англ. power distance)

$$d_{pow}(C_i, C_j) = \left(\sum_k |C_{ik} - C_{jk}|^p\right)^{1/r}. \quad (7)$$

Якщо *r*=2, то отримуємо евклідову відстань.

8 (А) Лінійний коефіцієнт кореляції Пірсона

$$r_{ij} = \frac{\sum_k (C_{ik} - \overline{C_i})(C_{jk} - \overline{C_j})}{\sqrt{\sum_k (C_{ik} - \overline{C_i})^2 \sum_k (C_{jk} - \overline{C_j})^2}}. \quad (8)$$

Методи кластеризування

Існує багато методів кластеризування. Нижче наведено деякі родини кластерних методів.

1) Ієрархічні методи
   - агломеративні;
   - дивизимні.

2) Алгоритми розбивання, зокрема ітераційні методи
   - метод k-середніх;
   - метод «сходження на пагорб».

3) Факторні методи.

4) Нейромережні методи.

5) Методи ущільнення об’єктів.

6) Грід-методи.

7) Методи моделювання

та інші. Також вирізняють двовхідне об’єднання, коли кластеризування здійснюють одночасно і за об’єктами, і за змінними, які ці об’єкти характеризують.

Деякі ієрархічні агломеративні методи[3]

① *Алгоритм «близького сусіди»* (одиночний зв’язок, англ. single linkage)

За правилом об’єднання для методу одиночного зв’язку новий кандидат

[3] так звані правила об’єднання

приєднується до кластеру, якщо він має найвищу ступінь подібності до якогось з членів кластеру

$$\min\{d(C_i, C_j): C_i \in A, C_j \in B\}. \quad (9)$$

Відстань від деякого об'єкту до групи об'єктів визначають як відстань до найближчого об'єкту з цієї групи. Тому отримувані кластери мають тенденцію бути представленими довгими «ланцюгами».

② *Алгоритм «далекого сусіди»* (повний зв'язок, англ. complete linkage)

Відстань між кластерами визначається найбільшою відстанню між будь-якими двома об'єктами у різних кластерах

$$\max\{d(C_i, C_j): C_i \in A, C_j \in B\}. \quad (10)$$

Відстань від деякого об'єкту до групи об'єктів визначають як відстань до найвіддаленішого об'єкту з цієї групи. Цей алгоритм добре працює для об'єктів, що утворюють «гаї», і погано – для тих, що утворюють «ланцюги».

③ Алгоритм середнього зв'язку (див. вище, $\frac{1}{|A|\times|B|}\sum_{C_i\in A}\sum_{C_j\in B} d(C_i, C_j)$):

- незважений (англ. unweighted pair-group average, абр. UPGMA);
- зважений (англ. weighted pair-group average, абр. WPGMA), де як ваговий коефіцієнт використовують розмір кластерів тобто кількість об'єктів у них.

④ Центроїдний метод, за яким відстань між двома кластерами визначають як відстань між їх центрами ваги:

- незважений (англ. unweighted pair-group centroid, абр. UPGMC);
- зважений (англ. weighted pair-group centroid, абр. WPGMC).

⑤ Метод Варда (або Уорда, англ. Ward's method).

Метод мінімізує суму квадратів відстаней для будь-яких двох кластерів, які може бути сформовано на кожному кроці, та намагається створювати кластери малих розмірів. Метод побудовано таким чином, щоб мінімізувати дисперсію всередині кластерів. Тобто сума квадратів відхилень $C_j^2 - \frac{1}{n}\left(\sum C_j\right)^2$

на першому кроці дорівнюватиме 0. За методом Варда об’єднуються ті групи або об’єкти, для яких сума квадратів отримує мінімальний приріст.

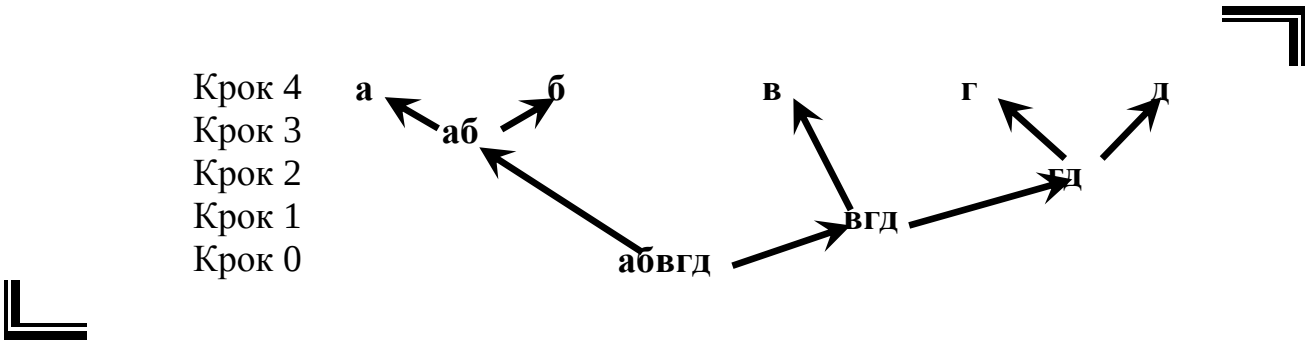
У процедурах кластеризування, що використовують послідовне об’єднання елементів, застосовується така узагальнена формула для перерахування відстані між кластером «**а**» та кластером «**бв**», який є об’єднанням кластерів «**б**» та «**в**».

$$d(C_{\text{а}}, C_{\text{бв}}) = \alpha * d_{\text{аб}} + \beta * d_{\text{ав}} + \gamma * d_{\text{бв}} + \delta * |d_{\text{аб}} - d_{\text{ав}}|. \qquad (11)$$

У випадку single linkage маємо $\alpha = \beta = -\delta = \frac{1}{2}$, $\gamma = 0$, у випадку complete linkage – $\alpha = \beta = \delta = \frac{1}{2}, \gamma = 0$.

Ієрархічні дивизимні методи

На початковому етапі уся вибірка розглядається як єдиний кластер, після чого починається процес його розділення доки кожний об’єкт не стане окремим кластером.



Розрізняють монотетичне розділення на основі однієї ознаки та політетичне розділення на основі усіх ознак об’єктів.

Приклад 1 – Single Linkage

Таблиця 1. Вихідні дані для розрахунку

Об’єкти (cases)	*Ознаки (variables)*	
	x	y
а	0	-2
б	-1	0
в	1	2
г	4	0

Відобразимо об’єкти у двомірному просторі *xy* (рис. 1).

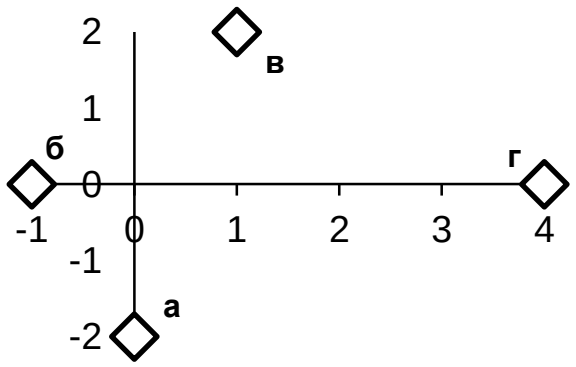


Рис. 1. Об’єкти для класифікування

Розглядаємо кожний об’єкт як окремий кластер. Розрахуємо відстані між об’єктами:

$$d^2_{евс_{аб}} = (а_x - б_x)^2 + \left(а_y - б_y\right)^2 = (0+1)^2 + (-2-0)^2 = 5;$$

$$d^2_{евс_{ав}} = (а_x - в_x)^2 + \left(а_y - в_y\right)^2 = (0-1)^2 + (-2-2)^2 = 17;$$

$$d^2_{евс_{аг}} = (а_x - г_x)^2 + \left(а_y - г_y\right)^2 = (0-4)^2 + (-2-0)^2 = 20;$$

$$d^2_{евс_{бв}} = (б_x - в_x)^2 + \left(б_y - в_y\right)^2 = (-1-1)^2 + (0-2)^2 = 8;$$

$$d^2_{евс_{бг}} = (б_x - г_x)^2 + \left(б_y - г_y\right)^2 = (-1-4)^2 + (0-0)^2 = 25;$$

$$d^2_{евс_{вг}} = (в_x - г_x)^2 + \left(в_y - г_y\right)^2 = (1-4)^2 + (2-0)^2 = 13;$$

Сформуємо матрицю відстаней (англ. distance matrix)

$$D_1^{SL} = \begin{pmatrix} 0 & \mathbf{5} & 17 & 20 \\ \mathbf{5} & 0 & 8 & 25 \\ 17 & 8 & 0 & 13 \\ 20 & 25 & 13 & 0 \end{pmatrix},$$

згідно якої об’єднуємо кластери «а» та «б» як найближчі. Отримуємо нову сукупність кластерів – «аб», «в» та «г». Визначаємо відстань між кластерами «аб» і «в» та «аб» і «г» скориставшись правилом (9):

– $d^2_{евс_{(аб)в}} = \min\left(d^2_{евс_{ав}}; d^2_{евс_{бв}}\right) = \min(17; 8) = 8$ або розрахуємо її згідно формули (11)

$$d^2_{евс_{(аб)в}} = \frac{1}{2} d^2_{евс_{ав}} + \frac{1}{2} d^2_{евс_{бв}} - \frac{1}{2}\left|d^2_{евс_{ав}} - d^2_{евс_{бв}}\right| = 8;$$

– $d^2_{евс_{(аб)г}} = \min\left(d^2_{евс_{аг}}; d^2_{евс_{бг}}\right) = \min(20; 25) = 20.$

Сформуємо матрицю відстаней

$$D_2^{SL} = \begin{pmatrix} 0 & \mathbf{8} & 20 \\ \mathbf{8} & 0 & 13 \\ 20 & 13 & 0 \end{pmatrix},$$

згідно якої об'єднуємо кластери «аб» та «в» як найближчі. Визначаємо відстань між кластерами «абв» і «г» скориставшись правилом (9):

$d^2_{\text{євис}_{(\text{абв})\text{г}}} = \min\left(d^2_{\text{євис}_{(\text{аб})\text{г}}}; d^2_{\text{євис}_{\text{вг}}}\right) = \min(20; 13) = 13.$ Сформуємо матрицю відстаней

$$D_3^{SL} = \begin{pmatrix} 0 & 13 \\ 13 & 0 \end{pmatrix}$$

та об'єднуємо два кластери. Побудуємо дендрограму ходу об'єднування кластерів (рис. 2).

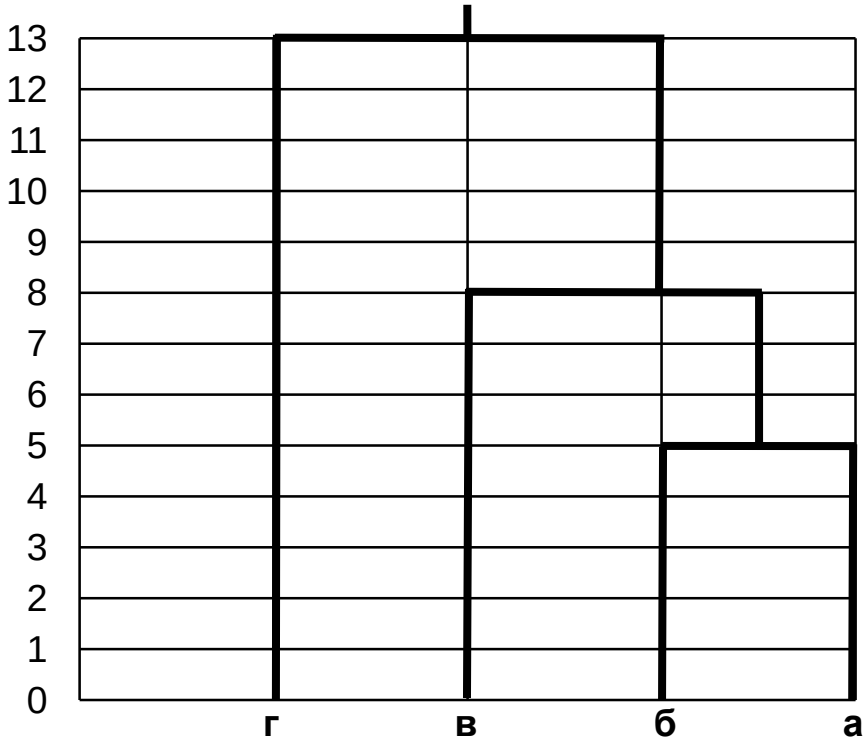


Рис. 2. Дендрограма для алгоритму «близького сусіди»

Зазвичай ознаки виміряно у різних одницях, тому необхідно їх попереднє нормування. У пакеті STATISTICA реалізовано такий спосіб нормування вихідних даних як $z = \frac{x - \bar{x}}{\sigma}$, де $\bar{x}$ – середнє значень змінної, яка характеризує об'єкти, σ - середнє квадратичне відхилення (див. меню «Edit», підменю «Standardize Block», команда «Standardize Columns» для прикладу 1).

Нормування змінює геометрію вихідного простору, що може змінити результати кластеризування.

Приклад 2 – Complete Linkage

Вихідні дані для розрахунку представлено у табл. 1 та на рис. 1. Розглядаємо кожний об’єкт як окремий кластер. Розрахуємо відстані між об’єктами (див. приклад 1) та сформуємо матрицю відстаней

$$D_1^{CL} = D_1^{SL} = \begin{pmatrix} 0 & \mathbf{5} & 17 & 20 \\ \mathbf{5} & 0 & 8 & 25 \\ 17 & 8 & 0 & 13 \\ 20 & 25 & 13 & 0 \end{pmatrix},$$

згідно якої об’єднуємо кластери «а» та «б» як найближчі. Отримуємо нову сукупність кластерів – «аб», «в» та «г». Визначаємо відстань між кластерами «аб» і «в» та «аб» і «г» скориставшись правилом (10):

- $d^2_{евис_{(аб)в}} = \max\left(d^2_{евис_{ав}}; d^2_{евис_{бв}}\right) = \max(17; 8) = 17$ або розрахуємо її згідно формули (11)

  $$d^2_{евис_{(аб)в}} = \frac{1}{2} d^2_{евис_{ав}} + \frac{1}{2} d^2_{евис_{бв}} + \frac{1}{2}\left|d^2_{евис_{ав}} - d^2_{евис_{бв}}\right| = 17;$$

- $d^2_{евис_{(аб)г}} = \max\left(d^2_{евис_{аг}}; d^2_{евис_{бг}}\right) = \max(20; 25) = 25.$

Сформуємо матрицю відстаней

$$D_2^{CL} = \begin{pmatrix} 0 & 17 & 25 \\ 17 & 0 & 13 \\ 25 & 13 & 0 \end{pmatrix},$$

згідно якої об’єднуємо кластери «в» та «г» як найближчі. Визначаємо відстань між кластерами «аб» і «вг» скориставшись правилом (10):

$d^2_{евис_{(аб)(вг)}} = \max\left(d^2_{евис_{(аб)в}}; d^2_{евис_{(аб)г}}\right) = \max(17; 25) = 25$. Сформуємо матрицю відстаней

$$D_3^{CL} = \begin{pmatrix} 0 & 25 \\ 25 & 0 \end{pmatrix}$$

та об’єднуємо два кластери. Побудуємо дендрограму ходу об’єднування кластерів (рис. 3).

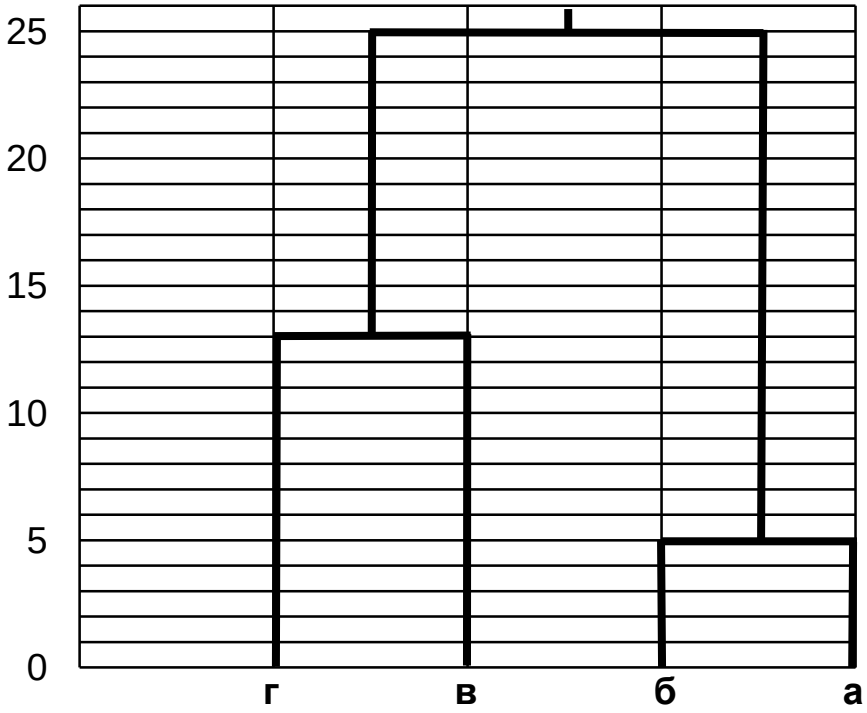


Рис. 3. Дендрограма для алгоритму «далекого сусіди»

Метод k-середніх
(швидкий кластерний аналіз)

Алгоритм k-середніх будує k кластерів, розташованих на максимальній відстані один від одного. Середні у кластері для усіх змінних максимально відрізняються.

Крок 1.
Початковий розподіл об'єктів за кластерами

Вибір числа k та початкових об'єктів-центроїдів:

а) вибір k-спостережень для максимізування початкової відстані;

б) сортування відстаней і задавання спостережень на постійних інтервалах;

в) вибір перших k-спостережень.

Результатом цього кроку є призначення кожного об'єкту певному кластеру.

Крок 2.
Ітераційний процес

Обчислення центрів[4] кластерів та перерозподілення об'єктів. Процедуру

[4] покоординатні середні

повторюють доки стабілізуються кластерні центри або досягнуто максимальну кількість ітерацій.

Крок 3.

Перевірення якості кластеризування та ступеню впливу ознак

У якості ***критерію якості кластеризування*** візьмемо відношення середньої внутрішньокластерної[5] відстані до середньої міжкластерної відстані.

$$Q = \frac{d_w/n_w}{d_b/n_b}, \tag{12}$$

де d_w та d_b – суми внутрішніх та міжкластерних відстаней, n_w та n_b – кількість внутрішніх та міжкластерних відстаней. Чим ближче значення такого критерію до нуля, тим краще проведено кластеризування.

Ступінь впливу ознак визначають як відношення функції Фішера, розрахованої емпірично на основі внутрішньо- і міжкластерних відстаней у деякому наборі об’єктів, до функції Фішера, обчисленої теоретично за відповідних ступенів свободи вибірки для внутрішньо- і міжкластерних відстаней за умови рівності їхніх дисперсій.

$$K = \frac{F}{F_{\text{кр}}},\ F_{\text{кр}} = F\big(\alpha, \max(f_b, f_w), \min(f_b, f_w)\big), \tag{13}$$

де f_b та f_w – число ступенів свободи міжкластерних та внутрішньокластерних відстаней. Чим більше значення K для певної змінної, тим вищою є ступінь впливу цієї ознаки на хід та результати кластеризування.

Кластерний аналіз у пакеті STATISTICA

Кластерний аналіз у пакеті STATISTICA проводять за допомогою модуля «Cluster Analysis», який можна викликати за допомогою підменю «Multivariate Exploratory Techniques» меню «Statistics». Якщо ознаки об’єктів представлено у різних одиницях вимірювання варто виконати попереднє нормування даних за допомогою підменю «Standardize Block» меню «Edit».

У діалоговому вікні «Clustering Method» представлено такі методи

- «Joining (tree clustering)» – агломеративні методи;
- «K-means clustering» – метод *k*-середніх;
- «Two-way joining» – двовхідне об’єднання.

[5] об’єктів до центру кластеру

У разі вибору агломеративних методів з’являється діалогове вікно «Cluster Analysis: Joining (tree clustering)», у якому можна вибрати алгоритм об’єднання кластерів («Single Linkage», «Complete Linkage», «Unweighted pair-group average», «Weighted pair-group average», «Unweighted pair-group centroid», «Weighted pair-group centroid» або «Ward’s method») та міру відстаней («Squared Euclidean distances», «Euclidean distances», «City-block (Manhattan) distances», «Chebychev distance metric», «Power[6]», «Percent disagreement[7]», «1-Person r»). Опцію «Distance matrix» у рядку «Input file» передбачено для випадку, якщо початкові дані представлено у вигляді мір подібності. «Блок MD deletion» призначено для вибору способу оброблення неповних даних

– «Casewise» – неповні дані порядково видаляють;
– «Mean substitution» – пропущені дані замінюють середніми значеннями.

У вікні «Joining Results» можна переглянути хід об’єднання у вигляді горизонтальної та вертикальної дендрограми, схеми об’єднання і графіка схеми об’єднання, отримати вихідну матрицю відстаней і описові характеристики, зокрема середніми значеннями та стандартними відхиленнями для кожного об’єкта.

У разі вибору методу k-середніх з’являється діалогове вікно «Cluster Analysis: K-Means Clustering», яке дає змогу налаштувати параметри кластеризування цим методом: показники, за якими проходитиме кластеризування, кількість кластерів, максимальну кількість ітерацій, спосіб задавання початкових центрів кластерів. За варіантом «Choose observations to maximize initial between-cluster distances» обирають перші k об’єктів, які стають центрами кластерів. Центри замінюють наступними об’єктами, якщо найменша відстань до будь-якого з них більша ніж найменша відстань між кластерами. Таким чином, початкові відстані між кластерами максимізовано. За варіантом «Sort distances and take observations at constant intervals» відстані між об’єктами відсортовують та обирають в якості початкових центрів кластерів об’єкти на

[6] з можливістю вибору значень *p* та *r* (див. формулу 7)
[7] для категоріальних даних

постійних інтервалах. За варіантом «Choose the first N (Number of cluster)» у якості початкових центрів кластерів обирають перші k спостережень.

У вікні «K-Means Clustering Results» можна переглянути результати класифікування. У верхній частині діалогового вікна представлено вихідні дані, параметри кластеризування та кількість ітерацій, за які досягнуто стабілізування центрів кластерів. У нижній частині розташовано кнопки виведення результатів процедури: результати дисперсійного аналізу, середні значення кластерів, отримані відстані, графік середніх, статистичні характеристики кожного кластеру, об’єкти кожного кластеру та відстані від кожного з членів кластеру до центру кластеру.

Приклад 3 – K-means clustering

Вихідні дані для розрахунку представлено у табл. 2 та на рис. 4.

Таблиця 2. Вихідні дані для розрахунку

Об’єкт	*Ознаки*	
	x	y
1	1	1
2	2	1
3	2	2
4	3	2
5	5	4
6	4	5
7	5	5
8	6	5
9	5	6
10	6	6

Відобразимо об’єкти у двомірному просторі *xy* (рис. 4).

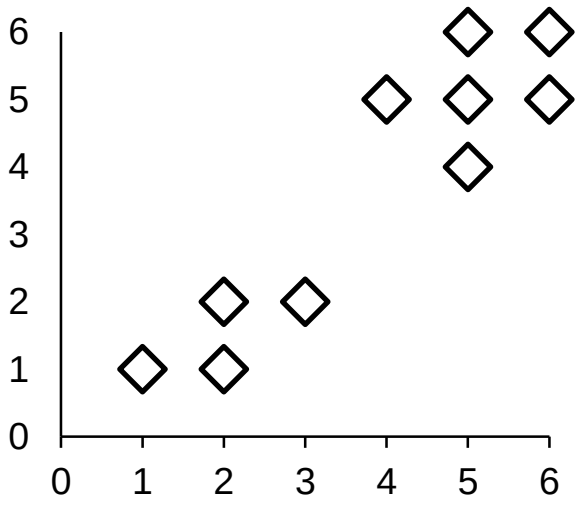


Рис. 4. Об’єкти для класифікування

Приймемо *k*=2 та скористаємось інструментами пакету STATISTICA для отримання проміжних результатів. Відкриваємо підменю «Multivariate Exploratory Techniques» меню «Statistics» і натискаємо команду «Cluster Analysis». У діалоговому вікні «Clustering Method» обираємо «K-Means clustering», а у діалоговому вікні «Cluster Analysis» вказуємо, що класифікуватимемо об’єкти («Cluster: Cases (rows)») з отриманням двох кластерів («Number of clusters: 2»). У вікні «K-Means Clustering Results» обираємо команду «Members of each cluster & distances» на вкладці «Advanced». Бачимо, що один кластер утворили об’єкти 1, 2, 3 та 4, а другий – решта об’єктів. Розрахуємо суму внутрішньокластерних відстаней

$$d_w = (0{,}79 + 0{,}35 + 0{,}35 + 0{,}79) + (0{,}83 + 0{,}83 + 0{,}17 + 0{,}60 + 0{,}60 + 0{,}83) = 6{,}1$$

для n_w = 10. У вікні «K-Means Clustering Results» обираємо команду «Summary: Cluster means & Euclidean distances» та під головною діагоналлю отриманої матриці відстаней знаходимо відстань між двома кластерами d_b = 3,4 для n_b = 1. Розраховуємо значення критерію якості кластеризування за формулою (12)

$$Q = \frac{6{,}1/10}{3{,}4/1} = 0{,}18,$$

Приймемо *k*=3 та за допомогою пакету STATISTICA отримаємо, що перший кластер утворили об'єкти 1, 2, 3 та 4, другий – об'єкти 5, 6 та 7, а третій – об'єкти 8, 9 та 10. Сума внутрішньокластерних відстаней становить

d_w = (0,79 + 0,35 + 0,35 + 0,79) + (0,53 + 0,53 + 0,33) + (0,53 + 0,53 + 0,33) = 5,1

для n_w = 10, сума відстаней між кластерами d_b = 2,9 + 1,0 + 3,9 = 7,8 для n_b = 3. Розраховуємо значення критерію якості кластеризування

$$Q = \frac{5{,}1/10}{7{,}8/3} = 0{,}19.$$

Таким чином, розбиття на два кластери є якіснішим, а частка залишкової дисперсії є меншою ніж для розбиття на три кластери.

Визначимо, яка з ознак більше вплинула на результати кластеризування. Для *k*=2 у вікні «K-Means Clustering Results» обираємо команду «Analysis of variance» та отримуємо таблицю результатів дисперсійного аналізу, за якою приймаємо $f_b = 1$, $f_w = 8$, $F_x = 39{,}8$ та $F_y = 67{,}3$. Отже, за формулою (13) для рівня значимості α = 0,01 отримуємо[8]

$$F_{кр} = F(0{,}01;\ 8;\ 1) = 5981,\ K_x = \frac{F_x}{F_{\text{кр}}} = \frac{39{,}8}{5981} = 0{,}007,\ K_y = \frac{F_y}{F_{\text{кр}}} = \frac{67{,}3}{5981} = 0{,}011.$$

Таким чином, змінна *y* має вищий ступінь впливу на результати кластеризування ніж змінна *x*.

[8] для визначення значення можна скористатись функцією MS Excel FРАСПОБР()

Список використаної та рекомендованої літератури

Айвазян, Сергей Артемьевич. Прикладная статистика в задачах и упражнениях: Учебник для студ. экономич. спец. вузов / С.А. Айвазян, В.С. Мхитарян. – М.: ЮНИТИ-ДАНА, 2001. – 270 с. *(фонд НТБ, читальна зала № 11)*

Боровиков, Владимир Павлович. Программа Statistica для студентов и инженеров / В. Боровиков. - М.: КомпьютерПресс, 2001. - 301 с. *(фонд НТБ, читальна зала № 11)*

Мандель, Игорь Давидович. Кластерный анализ / И. Д. Мандель. - М.: Финансы и статистика, 1988. - 176 с. *(фонд НТБ, науково-техн. від.: абон-т)*

Статистика : Учеб. для студ. / В.С. Мхитарян, Т.А. Дуброва, В.Г. Минашкин и др.; Под ред. В.С. Мхитаряна. – М.: Академия, 2003. – 272 с. *(фонд НТБ, науково-техн. від.: абон-т)*

Факторный, дискриминантный и кластерный анализ: [Сборник] / Дж.-О. Ким, Ч. У. Мьюллер, У. Р. Клекка и др.; пер. с англ. А. М. Хотинского, С. Б. Королева; под ред. И. С. Енюкова. – Москва: Финансы и статистика, 1989. – 215 с. *(фонд НТБ, науково-техн. від.: абон-т)*

Додаток А

Роздавальний матеріал до лекції

«Реалізування кластерного аналізу в ППП STATISTICA»

Приклад А.1 – Single Linkage та Complete Linkage

Таблиця А.1. Вихідні дані для розрахунку

Об'єкти (cases)	*Ознаки (variables)*	
	x	y
а	1	0
б	0	2
в	3	0
г	2	4
д	3	3

Відобразимо об'єкти у двомірному просторі *xy* (рис. А.1).

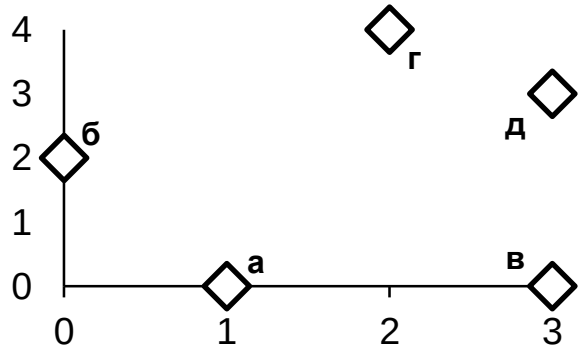


Рис. А.1. Об'єкти для класифікування

Розглядаємо кожний об'єкт як окремий кластер. Розрахуємо відстані між об'єктами:

$d^2_{евис_{аб}} = (1-0)^2 + (0-2)^2 = 5;$ $d^2_{евис_{бг}} = (0-2)^2 + (2-4)^2 = 8;$

$d^2_{евис_{ав}} = (1-3)^2 + (0-0)^2 = 4;$ $d^2_{евис_{бд}} = (0-3)^2 + (2-3)^2 = 10;$

$d^2_{евис_{аг}} = (1-2)^2 + (0-4)^2 = 17;$ $d^2_{евис_{вг}} = (3-2)^2 + (0-4)^2 = 17;$

$d^2_{евис_{ад}} = (1-3)^2 + (0-3)^2 = 13;$ $d^2_{евис_{вд}} = (3-3)^2 + (0-3)^2 = 9;$

$d^2_{евис_{бв}} = (0-3)^2 + (2-0)^2 = 13;$ $d^2_{евис_{гд}} = (2-3)^2 + (4-3)^2 = 2;$

Сформуємо матрицю відстаней (англ. distance matrix)

$$D_1^{SL} = D_1^{CL} = \begin{pmatrix} 0 & 5 & 4 & 17 & 13 \\ 5 & 0 & 13 & 8 & 10 \\ 4 & 13 & 0 & 17 & 9 \\ 17 & 8 & 17 & 0 & \mathbf{2} \\ 13 & 10 & 9 & \mathbf{2} & 0 \end{pmatrix},$$

згідно якої об'єднуємо кластери «г» та «д» як найближчі. Отримуємо нову сукупність кластерів – «а», «б», «в» та «е», який є об'єднанням кластерів «г» та «д». Визначаємо відстань між кластерами скориставшись правилами (9) та (10):

– $d^{2\,SL}_{еис_{ае}} = \min\left(d^2_{еис_{аг}}; d^2_{еис_{ад}}\right) = \min(17; 13) = 13;$

$d^{2\,CL}_{еис_{ае}} = \max\left(d^2_{еис_{ае}}; d^2_{еис_{ад}}\right) = \max(17; 13) = 17;$

– $d^{2\,SL}_{еис_{бе}} = \min\left(d^2_{еис_{бг}}; d^2_{еис_{бд}}\right) = \min(8; 10) = 8,$

$d^{2\,CL}_{еис_{бе}} = \max\left(d^2_{еис_{бе}}; d^2_{еис_{бд}}\right) = \max(8; 10) = 10;$

– $d^{2\,SL}_{еис_{ве}} = \min\left(d^2_{еис_{вг}}; d^2_{еис_{вд}}\right) = \min(17; 9) = 9,$

$d^{2\,CL}_{еис_{ве}} = \max\left(d^2_{еис_{вг}}; d^2_{еис_{вд}}\right) = \max(17; 9) = 17.$

Сформуємо матриці відстаней

$$D_2^{SL} = \begin{pmatrix} 0 & 5 & \mathbf{4} & 13 \\ 5 & 0 & 13 & 8 \\ \mathbf{4} & 13 & 0 & 9 \\ 13 & 8 & 9 & 0 \end{pmatrix}, D_2^{CL} = \begin{pmatrix} 0 & 5 & \mathbf{4} & 17 \\ 5 & 0 & 13 & 10 \\ \mathbf{4} & 13 & 0 & 17 \\ 17 & 10 & 17 & 0 \end{pmatrix},$$

згідно якої об'єднуємо кластери «а» та «в» як найближчі й для випадку Single Linkage, і для випадку Complete Linkage. Визначаємо відстань між кластерами «ж», який є об'єднанням кластерів «а» та «в», і «б» та «е» скориставшись правилами (9) та (10):

– $d^{2\,SL}_{еис_{бж}} = \min\left(d^2_{еис_{аб}}; d^2_{еис_{бв}}\right) = \min(5; 13) = 5,$

$d^{2\,CL}_{еис_{бж}} = \max\left(d^2_{еис_{аб}}; d^2_{еис_{бв}}\right) = \max(5; 13) = 13;$

– $d^{2\,SL}_{еис_{еж}} = \min\left(d^{2\,SL}_{еис_{ае}}; d^{2\,SL}_{еис_{ве}}\right) = \min(13; 9) = 9,$

$d^{2\,CL}_{еис_{еж}} = \max\left(d^{2\,CL}_{еис_{ае}}; d^{2\,CL}_{еис_{ве}}\right) = \max(17; 17) = 17.$

Сформуємо матриці відстаней

$$D_3^{SL} = \begin{pmatrix} 0 & \mathbf{5} & 9 \\ \mathbf{5} & 0 & 8 \\ 9 & 8 & 0 \end{pmatrix}, D_3^{CL} = \begin{pmatrix} 0 & 13 & 17 \\ 13 & 0 & 10 \\ 17 & 10 & 0 \end{pmatrix},$$

згідно якої об'єднуємо кластери «б» та «ж» як найближчі для випадку Single Linkage, і кластери «б» та «е» як найближчі для випадку Complete Linkage. Визначаємо відстань між кластерами «е» та «з», який є об'єднанням кластерів «б» та «ж», для випадку Single Linkage і «ж» та «і», який є об'єднанням кластерів «б» та «е», для випадку Complete Linkage скориставшись правилами (9) та (10):

– $d_{еиc_{ез}}^{2\,SL} = \min\left(d_{еиc_{еж}}^{2\,SL}; d_{еиc_{бе}}^{2\,SL}\right) = \min(9; 8) = 8;$

$d_{еиc_{жі}}^{2\,CL} = \max\left(d_{еиc_{бж}}^{2\,CL}; d_{еиc_{еж}}^{2\,CL}\right) = \max(13; 17) = 17.$

Сформуємо матриці відстаней

$$D_4^{SL} = \begin{pmatrix} 0 & 8 \\ 8 & 0 \end{pmatrix}, D_4^{CL} = \begin{pmatrix} 0 & 17 \\ 17 & 0 \end{pmatrix},$$

згідно якої об'єднуємо кластери «е» та «з» у кластер «к» для випадку Single Linkage, і кластери «ж» та «і» у кластер «л» для випадку Complete Linkage.

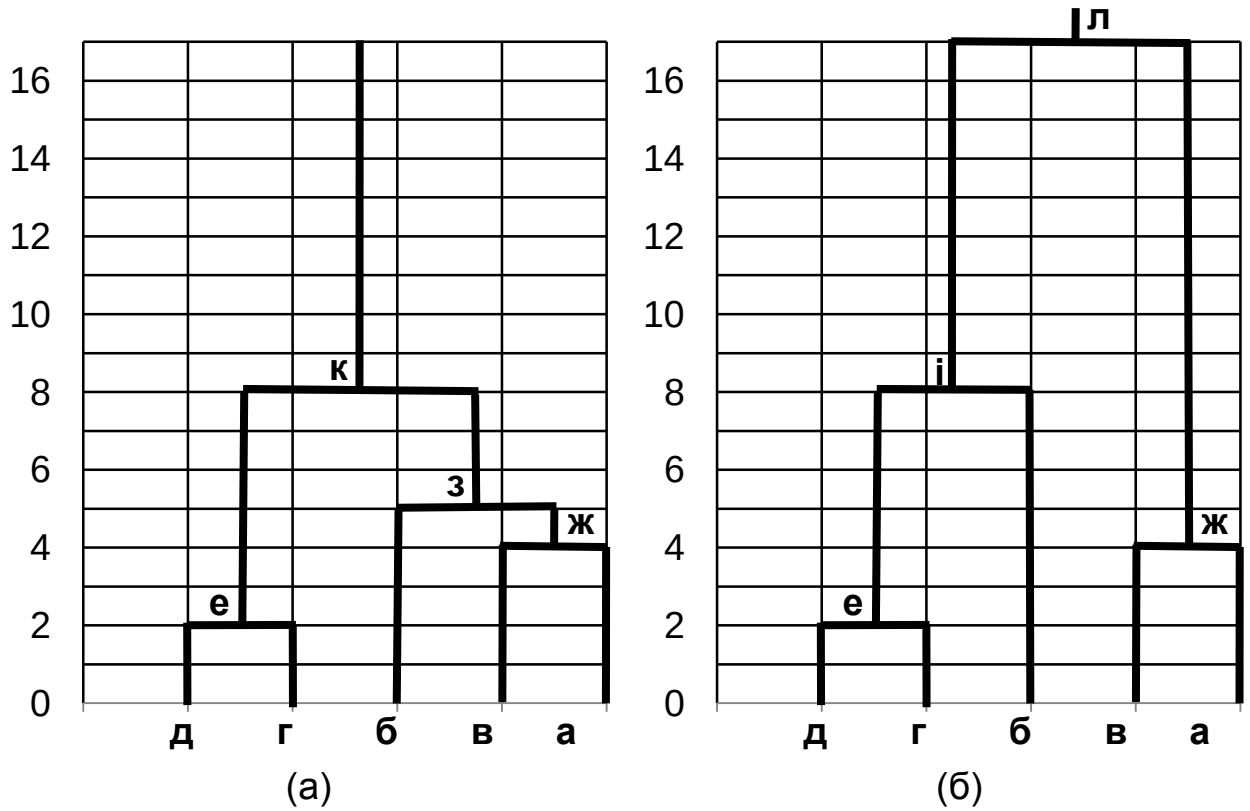


Рис. А.2. Дендрограми для алгоритмів «близького сусіди» (а) та «далекого сусіди» (б)