**НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОТЕХНОЛОГІЙ І ПРИРОДОКОРИСТУВАННЯ**

**Факультет інформаційних технологій**

Кафедра комп’ютерних наук

**Терія розпізнавання образів та класифікація в системах штучного інтелекту**

**Лабораторна робота №5**

«Дослідження методів к**ластерного аналізу в розпізнаванні образів»**

**(4 години)**

Київ - 2025

**Мета роботи:**

**1.** Дослідження методів класифікації образів на основі методів кластерного аналізу.

**2.** Закріплення теоретичного й практичного матеріалу, набуття навичок кластеризації за ієрархічними агломеративними процедурами в табличних редакторах *OpenOffice Calc або Microsoft Office Excel.*

**3.** Розробка програмного додатку на мові *C++* в середовищі розробки *MS Visual Studio*, що реалізує методи розпізнавання образів, що на основі теоретичних відомостей.

**4.** Дослідження функціонування програмного додатку при вирішенні задач класифікації образів на основі методів кластерного аналізу. Отримання практичних навичок з розробки програмних систем розпізнавання образів.

**5.** Порівняння отриманих результатів класифікації образів на основі різних методів кластерного аналізу при використанні табличних редакторів *OpenOffice Calc* або *Microsoft Office Excel* і створеного програмного додатку на

**Підготовка до роботи:** Вивчити і уяснити теретичні відомості щодо способів класифікації оброазів на основі образів на основі методів кластерного аналізу..

**Завдання:**

1. Розробити програмний додаток на мові програмування C/C++ в середовищі розробки MS Visual Studio, що реалізує розпізнавання образів на основі зазначеного методу.
2. Дослідити функціонування розробленого програмного додатку на прикладах.
3. За результатами досліджень скласти звіт з описом отриманих результатів та обґрунтованими висновками.

**Зміст звіту:**

1. Назва та мета роботи.
2. Теоретичні відомості.
3. Опис алгоритму функціонування програмного додатку.
4. Похідний програмний від коментований код на мові програмування *C++* в середовищі розробки *MS Visual Studio*.
5. Скрин-шоти роботи розробленого програмного засобу.

6. Опис ходу досліджень і отриманих результатів класифікації образів на основі різних методів кластерного аналізу при використанні табличних редакторів *OpenOffice Calc* або *Microsoft Office Excel.*

7. Обгрунтовані висновки щодо отриманих порівняння результатів проведених досліджень з використанням створеного програмного додатку і табличних редакторів *OpenOffice Calc* або *Microsoft Office Excel.*.

## Теоретичні відомості

Кластерний аналіз - це угруповання образів з використанням функцій відстані.

**Кластером** зазвичай називають групу образів $\{x_i\}$, Що задовольняють умові:

$$\|x_i - x_k\| < d \tag{1}$$

де $\|x_i - x_k\|$ - міра подібності між образами, $d$ - задане граничне обмеження по цьому заході.

Іноді кластери називають **таксонами,** а кластерний аналіз -таксономій.

Найбільш поширеною мірою подібності є відстань між точками-образами в просторі вимірів (ознак) X. У більшості випадків використовується евклідова метрика

$$\|\mathbf{a} - \mathbf{b}\| = \sqrt{\sum_{i=1}^{n} (a_j^2 - b_j^2)}\,. \tag{2}$$

Кластеризація по кореляційної міру подібності в завданні розпізнавання має ще й інше, самостійне значення. Така кластеризація застосовується для стиснення даних, зокрема, для скорочення розмірності простору ознак. Цей клас завдань називають **факторний аналіз.**

Використання поняття кластера передбачає, що образи досліджуваних об'єктів або явищ мають природну тенденцію до угрупованні навколо деяких характерних значень, які називають **центрами кластерів.** Чим сильніше виражена ця тенденція, тим більше успішно при вирішенні задачі можуть використовуватися методи кластерного аналізу.

При появі багатозональних космічних сканерів кластерний аналіз був одним з перших підходів, використаним при цифровий обробці багатозональних сканерних зображень. Це пояснюється тим, що при розрішенні цифрового зображення кілька десятків або навіть сотень метрів на піксель надійно розділяються тільки великі елементи ландшафту підстильної поверхні (водні об'єкти, лісові масиви, сільськогосподарські угіддя, відкриті грунти, забудова, і т.п.). ці об'єкти, як правило, добре розрізняються по спектральним характеристикам в тому чи іншому спектральному діапазоні і досить однорідні за яскравістю завдяки згладжуванню сцени. При таких умовах повинна існувати тенденція до утворення груп в просторі спектральних характеристик яскравості ознак **X.**

Недоліком такого способу класифікації є відсутність взаємозв'язку системи координат яскравості простору **X** з системою координат зображення. Виявлені кластери далеко не завжди збігаються з тими об'єктами, які

цікавлять обробника. Тому методи кластерного аналізу в пакетах обробки даних називають **неконтрольованою** або **Непоміченою** (Англійський термін **unsupervise)** класифікацією.

**Неконтрольована класифікація** застосовується зазвичай в таких цілях:

= для визначення кількості розділяючих по спектральних характеристиках класів об'єктів на обстежуваній території;

- для оцінки інформативності наявного набору вимірювань при вирішенні конкретної прикладної задачі;

- для вибору ділянок, які можуть бути використані при формуванні навчальних та контрольних вибірок в процесі класифікації з навчанням.

При відсутності даних наземних обстежень або нестачі довідкових (фондових) матеріалів неконтрольована класифікація може, однак, виявитися єдиним доступним способом тематичної обробки. В цьому випадку прагнуть виділити якомога більшу кількість кластерів і потім домагаються інтерпретується результату, групуючи кластери вже з урахуванням розташування класифікованих точок на зображенні.

Найбільш розповсюдження методи кластерного аналізу можна умовно розділити на дві групи:

1) Методи виявлення (вирощування) кластерів при заданому пороговому обмеженні на відстань між точками множини.

2) Методи формування кластерів при заданій кількості груп.

**У першому підході** кількість кластерів, як правило, апріорі невідомо. Вихідними даними при такій постановці завдання є порогове обмеження відстані (1) d і правила об'єднання елементів множини. В результаті кількість і форма кластерів сильно залежать від обраного методу аналізу, величини порога і початкових умов.

За методами формування кластерів в цьому підході виділяються Однозв'язуючі методи (аналіз елементів, найближчих до поточного), полізв'язуючі методи (аналіз найбільш віддалених елементів), і средньозв'язуючі методи.

**У другому підході** задається початкова кількість центрів кластерів, які в процесі аналізу переміщуються таким чином, щоб задані вимоги до кластерів виконувалися найкращим чином. Як правило, тут є критерій якості кластеризації, який в процесі формування кластерів максимізується (або мінімізується).

До цієї групи алгоритмів відносяться алгоритми внутрішньогрупових середніх, ISODATA (ітеративний алгоритм, самоорганізується метод аналізу даних) і дисперсно-коваріаційні критерії оптимізації.

Алгоритми цієї групи становлять найбільший інтерес, тому нижче ми розглянемо всі три типи таких алгоритмів.

В тому і в іншому підході можуть бути задані додаткові вимоги до кластерів: мінімальна кількість точок в групі, мінімальна відстань між групами і деякі інші.

Існує велика кількість більш-менш ефективних методів кластеризації, що грунтуються на теорії графів, проте в методології тематичної обробки даних такі алгоритми не знайшли застосування.

***Методи виявлення (вирощування) кластерів.***

**Однозв'язний метод.** Найпростішим способом виявлення кластерів є такий: вибираємо довільну точку (вектор) множини X, призначаємо її центром першого кластера $\left(x_k, k=1\right)$ і приєднуємо до цього кластеру всі точки, що задовольняють умові (1) ($\left\|x_i - x_k\right\| < d$) . Перша точка, для якої умова (1) не виконується, призначається центром наступного кластера.

Далі для кожної точки обчислюється відстань вже до двох центрів, і вона ставиться до того кластеру, відстань до якого менше. Якщо відстань до існуючих кластерів більше заданого порогу, утворюється новий кластер і т.д.

В кінцевому підсумку отримаємо деяке розбиття на класи, вид якого, як уже говорилося, дуже сильно залежить від порядку перегляду образів, особливо в тих випадках, коли тенденція до утворення груп простежується слабо. На рис. 1 показані два варіанти такого розбиття при різному порядку переглянути образів. Легко показати, що розділяючі функції являють собою геометричне місце точок, що лежать на перпендикулярах до відрізків, що з'єднують центри кластерів.

Недоліком алгоритмів такого типу часто є утворення так званих "ланцюгових" або "серпантинових" кластерів. Для утворення компактних груп образів, близьких до гіперсфери, зручніше використовувати повнозв'язний метод, в якому виділення кластерів починається з аналізу найбільш віддалених точок.

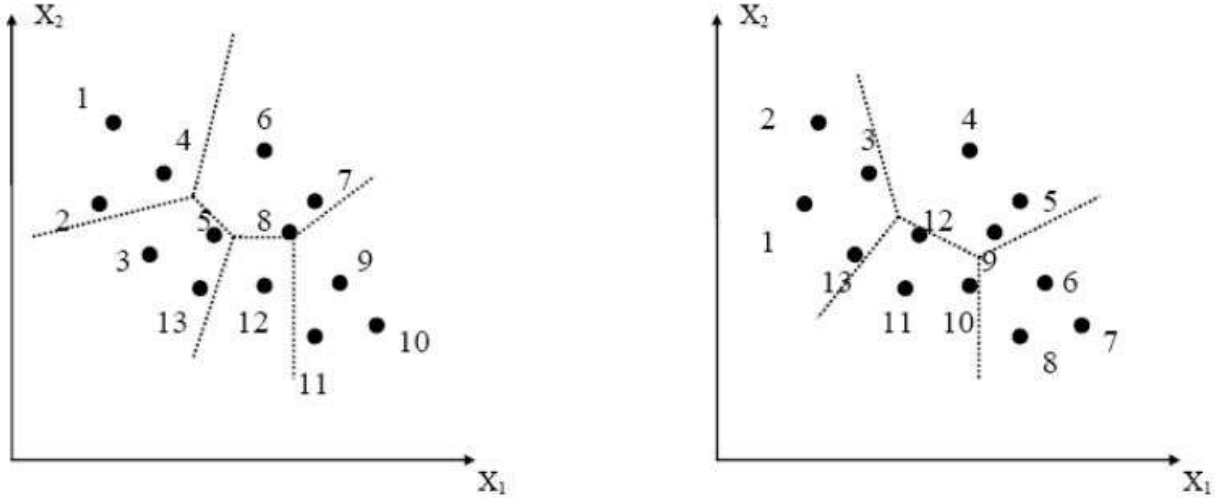



Рис.1. Результати кластеризації по порогу при різній послідовності аналізу образів

***Повнозв'зний метод. Алгоритм максимінної відстані.***

В якості вихідного образу виберемо деяку "крайню" точку, наприклад, з мінімальними координатами в просторі X. Назвемо її центром m1 кластера К1. В якості другого центру m2 кластера К2 виберемо найбільш віддалену від неї точку по всій множині образів. Визначимо граничне значення d як

$$d = || m1-m2 || / 2. \quad (3)$$

**Крок 1.** Обчислюємо відстані до центрів m1 і m2 || x-m1 ||, || x- m2 || для всіх елементів x нашої множини образів. З кожної пари відстаней вибираємо мінімальне.

**Крок 2.** Визначаємо максимальне значення

$$M = \max \{\min (|| x-m1 ||, || x- m2 ||)\} \quad (4)$$

по всій множині образів. Нехай цим значенням відповідає образ xi. Якщо M>d, призначаємо xi центром кластера К3. В якості нової порогової величини d можна взяти величину d=M/2 або половину середнього значення по всім мінімальним відстаням, розрахованим на попередньому кроці.

**Крок 3.** Для всіх x обчислюємо мінімальний з K відстаней до центрів вже утворених кластерів: min || x-mk ||, k = 1, ..., K.

К**рок 4.** Обчислюємо середнє мінімальне відстань способу x до центру кластера $p_{пор}(X, mk) = || x-mk ||/N$, де N - загальна кількість пар (x, mk).

Призначаємо новий поріг $d = p_{ср}(X, mk)$.

**Крок 5.** Шукаємо xi, відповідний значенням M (3) по всій множині образів. Якщо $M \le d$, процес закінчується. В іншому випадку призначаємо xi черговим центром кластера і переходимо до кроку 3. Процес має сенс припинити також в тому випадку, коли величина d стає менше середньоквадратичної похибки вимірювань ознак, з якими ми працюємо.

| | X1 | X2 | X3 (m1) | X4 (M2) |
|---|---|---|---|---|
| Зріст | 180 | 60 | 190 | 40 |
| вага | 90 | 20 | 100 | 10 |

Якщо то Xi $\|m_1 - X_i\| < d \in \Omega_1$
Якщо <d, то Xi $\|m_2 - X_i\| \in \Omega_2$

d = {[(190-40) 2+ (100-10) 2] ½} / 2 = 87,46

[(190-180) 2 + (100-90) 2] 1/2 = 14,14 <d→$\Omega_1$
[(190-60) 2+ (100-20) 2] 1/2 = 152,6> d
[(190-190) 2+ (100-100) 2] 1/2 = 0 <d→$\Omega_1$
[(190-40) 2+ (100-10) 2] 1/2 = 174,9> d

[(40-180) 2+ (10-90) 2] 1/2 =161,2452 > d

[(40-60) 2+ (10-20) 2] 1/2 =22,36068 <d $\rightarrow\Omega_2$

[(40-190) 2+ (10-100) 2] 1/2 =174,9286 > d

[(40-40) 2+ (10-10) 2] 1/2 = 0 <d $\rightarrow\Omega_2$

M = max {min (|| x-m1 ||, || x- m2 ||)}= max {min (14,14; 22,36068 )}= 22,36068

Якщо M> d (22,36068 <87,46), Призначаємо хі центром кластера К3 інакше закінчуємо роботу.

min || x-mk ||, k = 1,2, ... До

$\rho_{пор.}(X, mk) = || x-mk || / N$

Повторюємо процедуру для нового d.

$d = \rho_{ср}(X, mk)$.

## Методичні рекомендації до виконання завдання використанні табличних редакторів *OpenOffice Calc* або *Microsoft Office Excel*

Маємо шість об'єктів, описаних трьома показниками x1-x3. Геометрично кожному з об'єктів відповідає точка в трьовимірному просторі (де осями координат виступають шкали значень показників x1- x3). Графічно це може бути представлено графіком на рис. 1.

Кластеризація об'єктів за агломеративною процедурою припускає, що на першому етапі всі № об'єктів розглядаються як окремі кластери.

| | A | B | C |
|---|---|---|---|
| 1 | № спостереженя | X | |
| 2 | 1 | 313,10 | |
| 3 | 2 | 1387,50 | |
| 4 | 3 | 438,40 | |
| 5 | 4 | 425,50 | |
| 6 | 5 | 290,50 | |
| 7 | 6 | 124,60 | |
| 8 | 7 | 262,90 | |
| 9 | 8 | 330,20 | |
| 10 | 9 | 470,20 | |
| 11 | 10 | 14772,90 | |
| 12 | 11 | 11854,00 | |
| 13 | 12 | 10735,20 | |

Рис. 1. – Похідні дані

Розраховується відстань між ними та об’єднуються найближчі кластери. Матриця відстаней перераховується для отриманої меншої кількості кластерів (зазвичай *N*-1) і знову об’єднуються найближчі кластери. Потім знову перераховується матриця відстаней між кластерами і т. д. Процедура триває до тих пір, доки всі об’єкти не об’єднуються в один кластер.

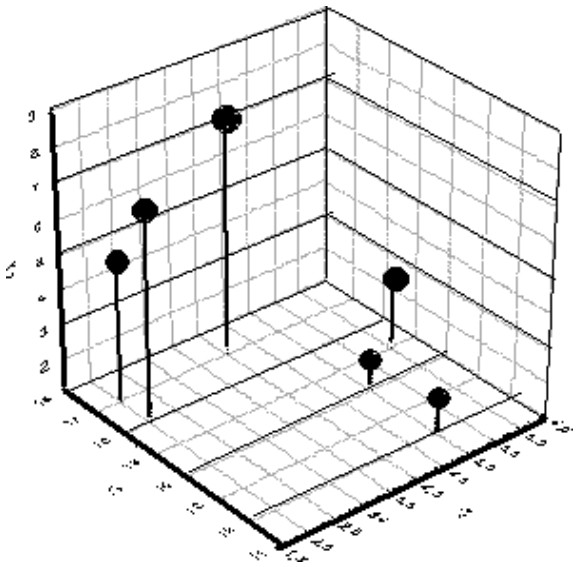


Рис. 2. - Геометричне представлення вихідних даних

**Агломеративна процедура за методом найближчого сусіда**

Для цього поперше розрахуємо матрицю відстаней між об’єктами. Геометрична відстань між точками на рис. 2 відповідає Евклідовій відстані:

$$d_{ij} = \sqrt{\sum_{k=1}^{m} \left( x_{ik} - x_{jk} \right)^2}, \qquad (5)$$

де $d_{ij}$ - відстань між *i*-м і *j*-м об’єктами;

$x_{jk}$ – *k*-та координата *j*-того об’єкта (значення *k*-того показника для *i*-го об’єкта);

*m* - кількість характеристик (показників), за якими описані об’єкти.

Щоб отримати матрицю відстаней у пакеті *Calc* або *Excel* набираємо таблицю з вихідними даними двічі: так як дані були надані в завданні та транспоновану матрицю (рис. 3). Це дозволить ввести формулу для розрахунку Евклідової відстані в першу з комірок таблиці відстаней, зафіксувати потрібні комірки знаком $, та потім розтягнути такий тип на всю таблицю (рис. 3).

I8 =КОРЕНЬ(($C4-I$3)^2+($D4-I$4)^2+($E4-I$5)^2)

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | | | | | | | | |
| 2 | | | Характеристики | | | | | | 1 | 2 | 3 | 4 | 5 | 6 | | | | |
| 3 | | № об'єкта | x1 | x2 | x3 | | | x1 | 2 | 3 | 2 | 7 | 8 | 5 | | | | |
| 4 | | 1 | 2 | 12 | 5 | | | x2 | 12 | 15 | 14 | 16 | 17 | 17 | | | | |
| 5 | | 2 | 3 | 15 | 6 | | | x3 | 5 | 6 | 5 | 2 | 4 | 2 | | | | |
| 6 | | 3 | 2 | 14 | 5 | | | | | | | | | | | | | |
| 7 | | 4 | 7 | 16 | 2 | | 1 Этап | | 1 | 2 | 3 | 4 | 5 | 6 | | Матриця евклідових відстаней | | |
| 8 | | 5 | 8 | 17 | 4 | | | 1 | 0 | 3,31662 | 2 | 7,07107 | 7,87401 | 6,55744 | | | | |
| 9 | | 6 | 5 | 17 | 2 | | | 2 | 3,31662 | 0 | **1,73205** | 5,74456 | 5,74456 | 4,89898 | | | | |
| 10 | | | | | | | | 3 | 2 | **1,73205** | 0 | 6,16441 | 6,78233 | 5,19615 | | мінімальна | | |
| 11 | | | | | | | | 4 | 7,07107 | 5,74456 | 6,16441 | 0 | 2,44949 | 2,23607 | | відстань = | **1,732** | |
| 12 | | № етапа | Об'єкти | Відстань | | | | 5 | 7,87401 | 5,74456 | 6,78233 | 2,44949 | 0 | 3,60555 | | | | |
| 13 | | 1 | 2+3 | 1,732 | | | | 6 | 6,55744 | 4,89898 | 5,19615 | 2,23607 | 3,60555 | 0 | | | | |

Рис. 3. Розрахунок матриці відстаней між об’єктами

У комірку I8 вводимо формулу:

для Calc - =SQRT(($C4-I$3)^2+($D4-I$4)^2+($E4-I$5)^2);

для Excel - =КОРЕНЬ(($C4-I$3)^2+($D4-I$4)^2+($E4-I$5)^2).

Завдяки тому, що в формулі розставлені знаки $, маємо можливість розтягнути введену формулу на весь діапазон I8-*N*13. Перевіряємо, якщо формула введена правильно – матриця відстаней буде симетричною матрицею з нулями на головній діагоналі.

Знаходимо мінімальну відстань між об’єктами. Як видно з рис. 3 – це відстань між 2 та 3 об’єктами. Об’єднуємо їх у новий кластер.

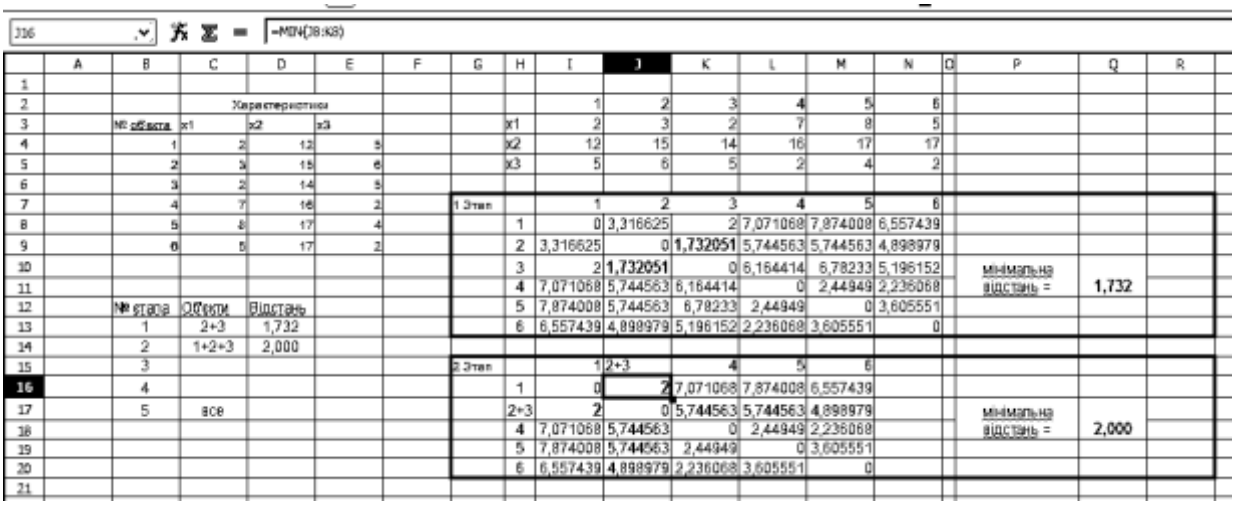
Для того, щоб формалізувати результати розрахунків сформуємо маленьку таблицю з трьох стовбців, що будуть містити № етапу, назви об’єктів, що на цьому етапі об’єднано, та відстані, на яких об’єднано об’єкти (рис. 2.2). Вносимо в цю таблицю дані щодо першого об’єднання (2 та 3 об’єкти).

Після цього будуємо нову таблицю для матриці відстаней між кластерами (рис. 4). На цьому етапі в нас кластерів вже 5, а не шість. Їх представляють 1, 2+3, 4, 5 та 6 об’єкти, відповідно.

Відстані між 1, 4, 5 та 6 кластерами вже розраховані в попередній таблиці. Перерахувати треба тільки відстані між новим кластером "2+3" та 1, "2+3" та 4, "2+3" та 5, "2+3" та 6 кластерами.

За методом найближчого сусіда за відстань між кластерами береться відстань між найближчими об’єктами двох кластерів. Тож, наприклад відстань між кластером "2+3" та 1 кластером, що містить тільки один 1 об’єкт, розраховується, як мінімальна відстань з $d_{13}$ 2 та $d_{12}$=3,316, і дорівнює, відповідно, $\min d=d_{2+3,1}=2$.

Для розрахунків використовуємо вбудовану функцію пакета Calc MIN(_) або Excel – МИН(_).



| | № об'єкта | x1 | x2 | x3 |
|---|---|---|---|---|
| 1 | 1 | 2 | 12 | 5 |
| 2 | 2 | 3 | 15 | 6 |
| 3 | 3 | 2 | 14 | 5 |
| 4 | 4 | 7 | 16 | 2 |
| 5 | 5 | 8 | 17 | 4 |
| 6 | 6 | 5 | 17 | 2 |

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| x1 | 2 | 3 | 2 | 7 | 8 | 5 |
| x2 | 12 | 15 | 14 | 16 | 17 | 17 |
| x3 | 5 | 6 | 5 | 2 | 4 | 2 |

1 Этап

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 0 | 3,316625 | 2 | 7,071068 | 7,874008 | 6,557439 |
| 2 | 3,316625 | 0 | **1,732051** | 5,744563 | 5,744563 | 4,898979 |
| 3 | 2 | **1,732051** | 0 | 6,164414 | 6,78233 | 5,196152 |
| 4 | 7,071068 | 5,744563 | 6,164414 | 0 | 2,44949 | 2,236068 |
| 5 | 7,874008 | 5,744563 | 6,78233 | 2,44949 | 0 | 3,605551 |
| 6 | 6,557439 | 4,898979 | 5,196152 | 2,236068 | 3,605551 | 0 |

мінімальна відстань = **1,732**

| № етапа | Об'єкти | Відстань |
|---|---|---|
| 1 | 2+3 | 1,732 |
| 2 | 1+2+3 | 2,000 |
| 3 | | |
| 4 | | |
| 5 | все | |

2 Этап

| | 1 | 2+3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| 1 | 0 | 2 | 7,071068 | 7,874008 | 6,557439 |
| 2+3 | 2 | 0 | 5,744563 | 5,744563 | 4,898979 |
| 4 | 7,071068 | 5,744563 | 0 | 2,44949 | 2,236068 |
| 5 | 7,874008 | 5,744563 | 2,44949 | 0 | 3,605551 |
| 6 | 6,557439 | 4,898979 | 2,236068 | 3,605551 | 0 |

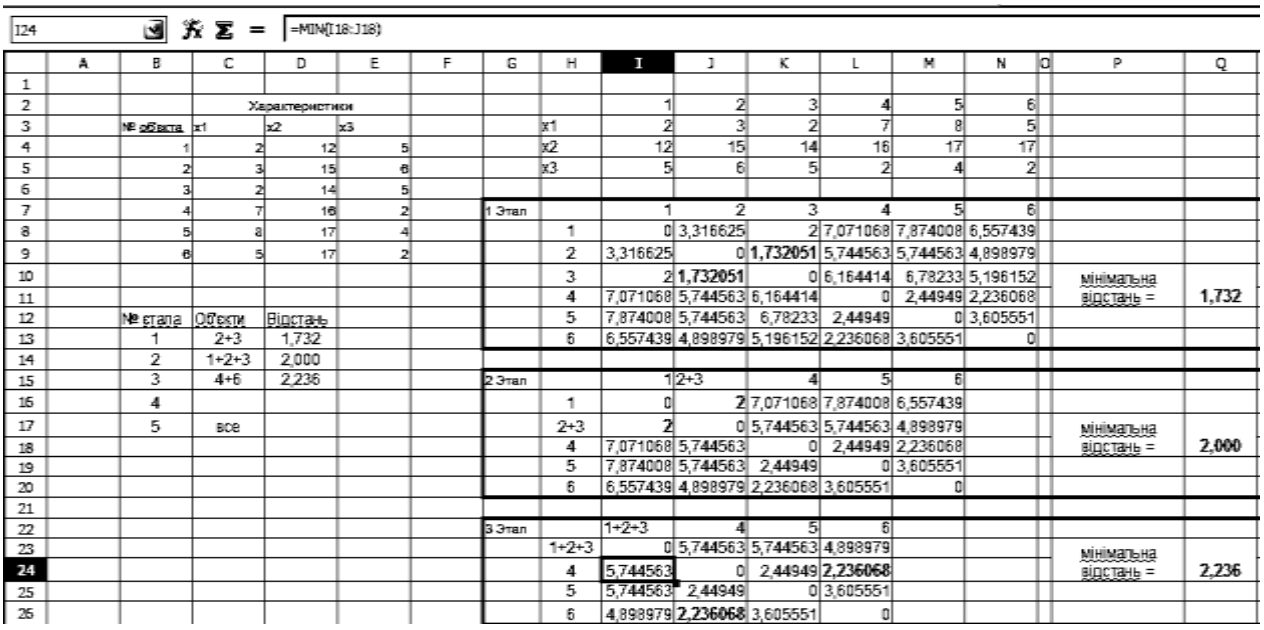
мінімальна відстань = **2,000**

Рис. 4. - Розрахунки другого етапу кластеризації за методом найближчого сусіда

Після розрахунку матриці відстаней знову знаходимо найближчі кластери. Оскільки мінімальна відстань, рівна 2 знаходиться між 1 та "2+3" кластерами на наступному етапі об'єднуємо саме їх.

На наступному етапі будуємо матрицю відстаней вже 4*4, бо маємо лише 4 кластери: "1+2+3", 4, 5, 6 (рис. 5).

Перераховуємо відстані. Обираємо мінімальну відстань та об'єднуємо кластери 4 та 6 ($\min d = d_{46} = 2{,}236$).



| | № об'єкта | x1 | x2 | x3 |
|---|---|---|---|---|
| 1 | 1 | 2 | 12 | 5 |
| 2 | 2 | 3 | 15 | 6 |
| 3 | 3 | 2 | 14 | 5 |
| 4 | 4 | 7 | 16 | 2 |
| 5 | 5 | 8 | 17 | 4 |
| 6 | 6 | 5 | 17 | 2 |

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| x1 | 2 | 3 | 2 | 7 | 8 | 5 |
| x2 | 12 | 15 | 14 | 16 | 17 | 17 |
| x3 | 5 | 6 | 5 | 2 | 4 | 2 |

1 Этап

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 0 | 3,316625 | 2 | 7,071068 | 7,874008 | 6,557439 |
| 2 | 3,316625 | 0 | **1,732051** | 5,744563 | 5,744563 | 4,898979 |
| 3 | 2 | **1,732051** | 0 | 6,164414 | 6,78233 | 5,196152 |
| 4 | 7,071068 | 5,744563 | 6,164414 | 0 | 2,44949 | 2,236068 |
| 5 | 7,874008 | 5,744563 | 6,78233 | 2,44949 | 0 | 3,605551 |
| 6 | 6,557439 | 4,898979 | 5,196152 | 2,236068 | 3,605551 | 0 |

мінімальна відстань = **1,732**

| № етапа | Об'єкти | Відстань |
|---|---|---|
| 1 | 2+3 | 1,732 |
| 2 | 1+2+3 | 2,000 |
| 3 | 4+6 | 2,236 |
| 4 | | |
| 5 | все | |

2 Этап

| | 1 | 2+3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| 1 | 0 | 2 | 7,071068 | 7,874008 | 6,557439 |
| 2+3 | 2 | 0 | 5,744563 | 5,744563 | 4,898979 |
| 4 | 7,071068 | 5,744563 | 0 | 2,44949 | 2,236068 |
| 5 | 7,874008 | 5,744563 | 2,44949 | 0 | 3,605551 |
| 6 | 6,557439 | 4,898979 | 2,236068 | 3,605551 | 0 |

мінімальна відстань = **2,000**

3 Этап

| | 1+2+3 | 4 | 5 | 6 |
|---|---|---|---|---|
| 1+2+3 | 0 | 5,744563 | 5,744563 | 4,898979 |
| 4 | 5,744563 | 0 | 2,44949 | **2,236068** |
| 5 | 5,744563 | 2,44949 | 0 | 3,605551 |
| 6 | 4,898979 | **2,236068** | 3,605551 | 0 |

мінімальна відстань = **2,236**

Рис. 5. Розрахунки третього етапу кластеризації за методом найближчого сусіда

На наступному етапі маємо вже 3 кластери: "1+2+3", "4+6", "5". Будуємо матрицю відстаней розмірності 3*3 (рис. 6).

Знаходимо мінімальну відстань $d = d_{4+6,5} = 2{,}236$ та об'єднуємо кластери "4+6" та "5".

На останньому етапі маємо 2 кластери: "1+2+3", "4+5+6". Будуємо матрицю (рис. 6), знаходимо мінімальну відстань $d = 4{,}898$ та об'єднуємо всі об'єкти в один кластер.

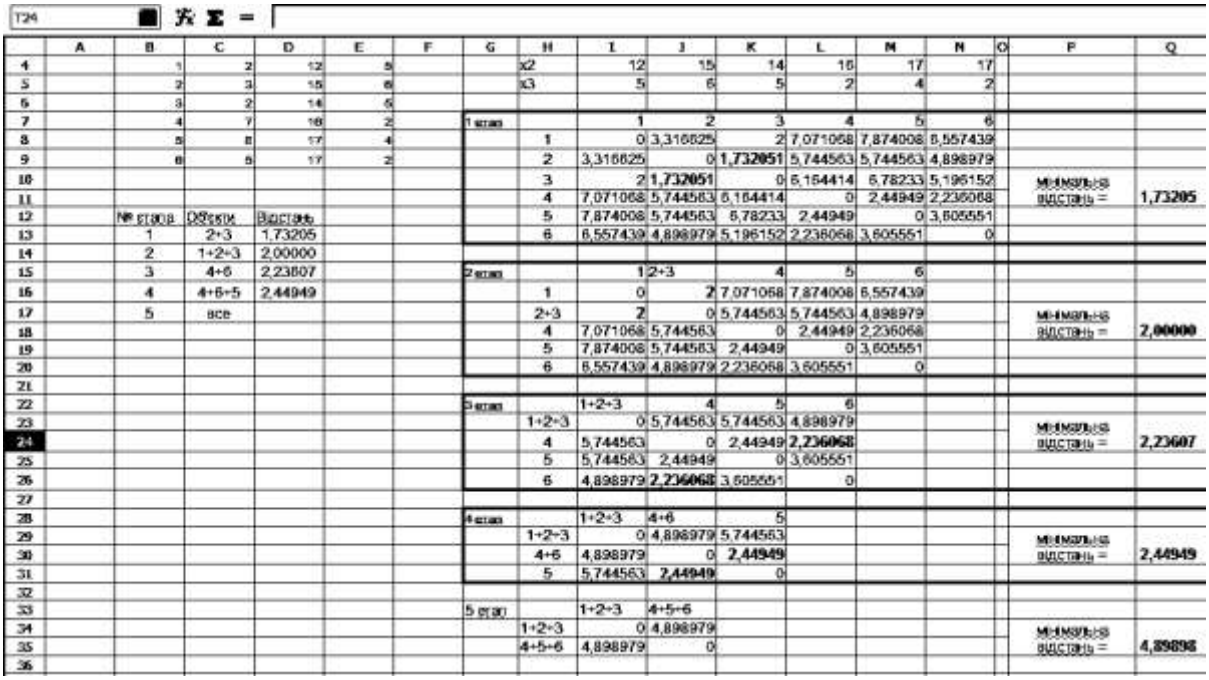


Рис. 6. Розрахунки четвертого етапу кластеризації за методом найближчого сусіда

За результати розрахунків будуємо дендрограму. Будуємо вручну на аркуші паперу, бо пакети не дозволяють це зробити автоматично. На осі ординат відмічаємо об'єкти, а по осі абсцис – відстані, на яких об'єкти об'єднуються у кластери. Отримуємо наступний малюнок – рис. 7.

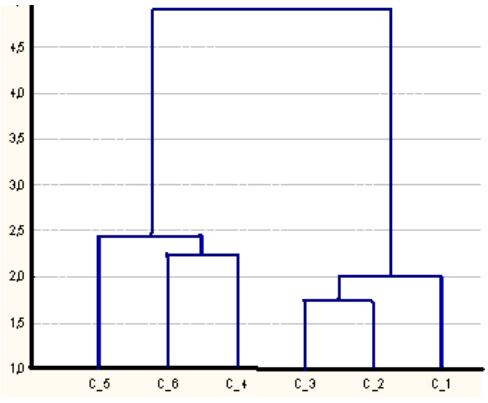



Рис. 7. Дендрограма за методом найближчого сусіда

Аналізуємо отриману дендрограму: проводимо лінію відсікання (рис.

8).

Бачимо, що існує природне розбиття сукупності об’єктів на кластери. А саме, можна отримати два досить істотно відмінні кластери. У першому будуть міститися об’єкти 4, 5 та 6. В другому – 1, 2 та 3.

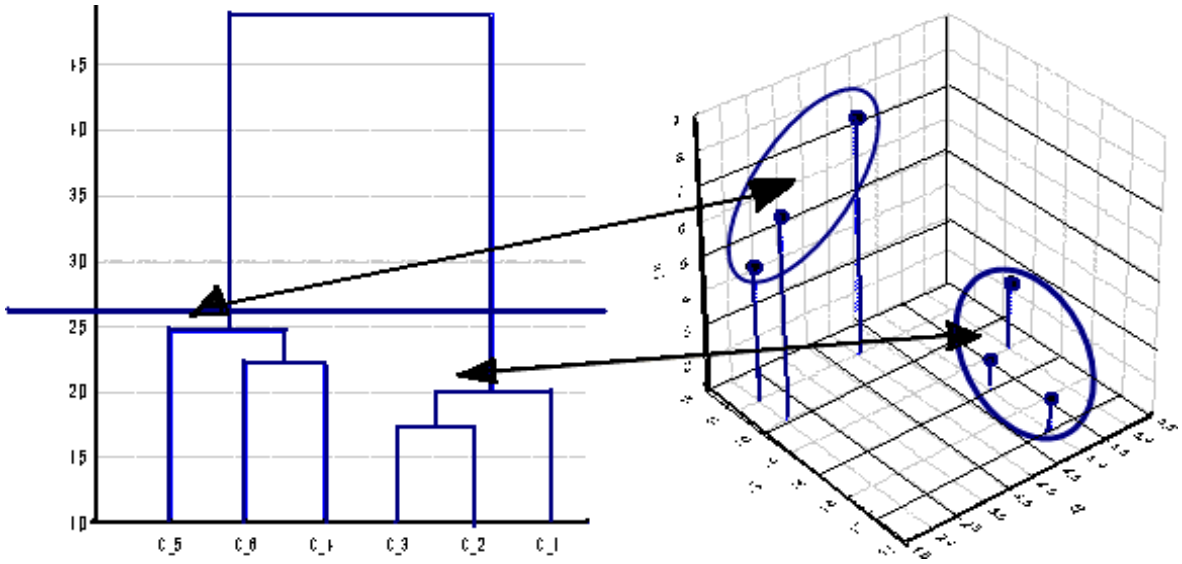



Рис. 8. Відповідність між дендрограмою та розташуванням об’єктів у просторі ознак

**Агломеративна процедура кластеризації за методом дальнього сусіда (рис.9).**

J16 =MAX(J8:K8)

| № об'єкта | x1 | x2 | x3 |
|---|---|---|---|
| 1 | 2 | 12 | 5 |
| 2 | 3 | 15 | 6 |
| 3 | 2 | 14 | 5 |
| 4 | 7 | 16 | 2 |
| 5 | 8 | 17 | 4 |
| 6 | 5 | 17 | 2 |

| № етапа | Об'єкти | Відстань |
|---|---|---|
| 1 | 2+3 | 1,732 |
| 2 | 4+6 | 2,236068 |
| 3 | 1+2+3 | 3,317 |
| 4 | 4+5+6 | 3,606 |
| 5 | все | 7,874 |

| | | 1 | 2 | 3 | 4 | 5 | 6 | |
|---|---|---|---|---|---|---|---|---|
| | x1 | 2 | 3 | 2 | 7 | 8 | 5 | |
| | x2 | 12 | 15 | 14 | 16 | 17 | 17 | |
| | x3 | 5 | 6 | 5 | 2 | 4 | 2 | |
| 1 етап | | 1 | 2 | 3 | 4 | 5 | 6 | |
| | 1 | 0 | 3,316625 | 2 | 7,071068 | 7,874008 | 6,557439 | |
| | 2 | 3,316625 | 0 | **1,732051** | 5,744563 | 5,744563 | 4,898979 | |
| | 3 | 2 | **1,732051** | 0 | 6,164414 | 6,78233 | 5,196152 | мінімальна відстань = **1,732** |
| | 4 | 7,071068 | 5,744563 | 6,164414 | 0 | 2,44949 | 2,236068 | |
| | 5 | 7,874008 | 5,744563 | 6,78233 | 2,44949 | 0 | 3,605551 | |
| | 6 | 6,557439 | 4,898979 | 5,196152 | 2,236068 | 3,605551 | 0 | |
| 2 етап | | 1 | 2+3 | 4 | 5 | 6 | | |
| | 1 | 0 | 3,316625 | 7,071068 | 7,874008 | 6,557439 | | |
| | 2+3 | 3,316625 | 0 | 6,164414 | 6,78233 | 5,196152 | | мінімальна відстань = **2,23607** |
| | 4 | 7,071068 | 6,164414 | 0 | 2,44949 | **2,236068** | | |
| | 5 | 7,874008 | 6,78233 | 2,44949 | 0 | 3,605551 | | |
| | 6 | 6,557439 | 5,196152 | **2,236068** | 3,605551 | 0 | | |
| 3 етап | | 1 | 2+3 | 4+6 | 5 | | | |
| | 1 | 0 | **3,316625** | 7,071068 | 7,874008 | | | мінімальна відстань = **3,317** |
| | 2+3 | **3,316625** | 0 | 6,164414 | 6,78233 | | | |
| | 4+6 | 7,071068 | 6,164414 | 0 | 3,605551 | | | |
| | 5 | 7,874008 | 6,78233 | 3,605551 | 0 | | | |
| 4 етап | | 1+2+3 | 4+6 | 5 | | | | |
| | 1+2+3 | 0 | 7,071068 | 7,874008 | | | | мінімальна відстань = **3,606** |
| | 4+6 | 7,071068 | 0 | **3,605551** | | | | |
| | 5 | 7,874008 | **3,605551** | 0 | | | | |
| 5 етап | | 1+2+3 | 4+6+5 | | | | | |
| | 1+2+3 | 0 | 7,874008 | | | | | мінімальна відстань = **7,874** |
| | 4+6+5 | 7,874008 | 0 | | | | | |

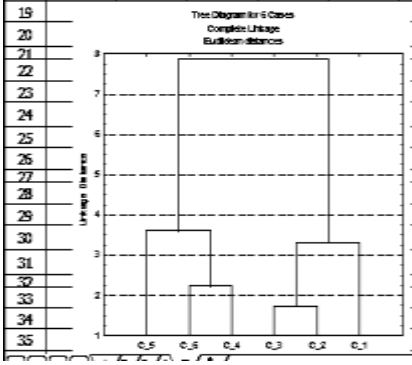



Рис. 9. Процедура кластеризації за методом дальнього сусіда

Єдина відмінність: за правилом дальнього сусіда за відстань між кластерами береться відстань між найбільш далекими один від одного об’єктами у кластерах. Наприклад, на другому етапі кластеризації (рис. 9) у комірці *J*16 потрібно розрахувати відстань між кластерами "2+3" та "1". За таку відстань береться максимальне значення з відстаней між "1" та "2" і "1" та "3"

$$d_{2+3,1} = \max\left(d_{1,2}; d_{1,3}\right) = d_{1,2} = 3{,}31$$

**Агломеративна процедура за методом середнього зв’язку.**

Відмінність полягає у розрахунку відстаней між клас- терами.

За відстань між двома кластерами береться середньоарифметичне всіх відстаней між об’єктами двох кластерів. Так на другому етапі відстань між кластерами "2+3" та "1", що міститься в комірці *J*16, треба розрахувати як

$$d_{2+3,1} = \frac{d_{1,2} + d_{1,3}}{2} = \frac{3{,}31 + 2}{2} = 2{,}658$$

Вся процедура за етапами представлена на рис. 10.

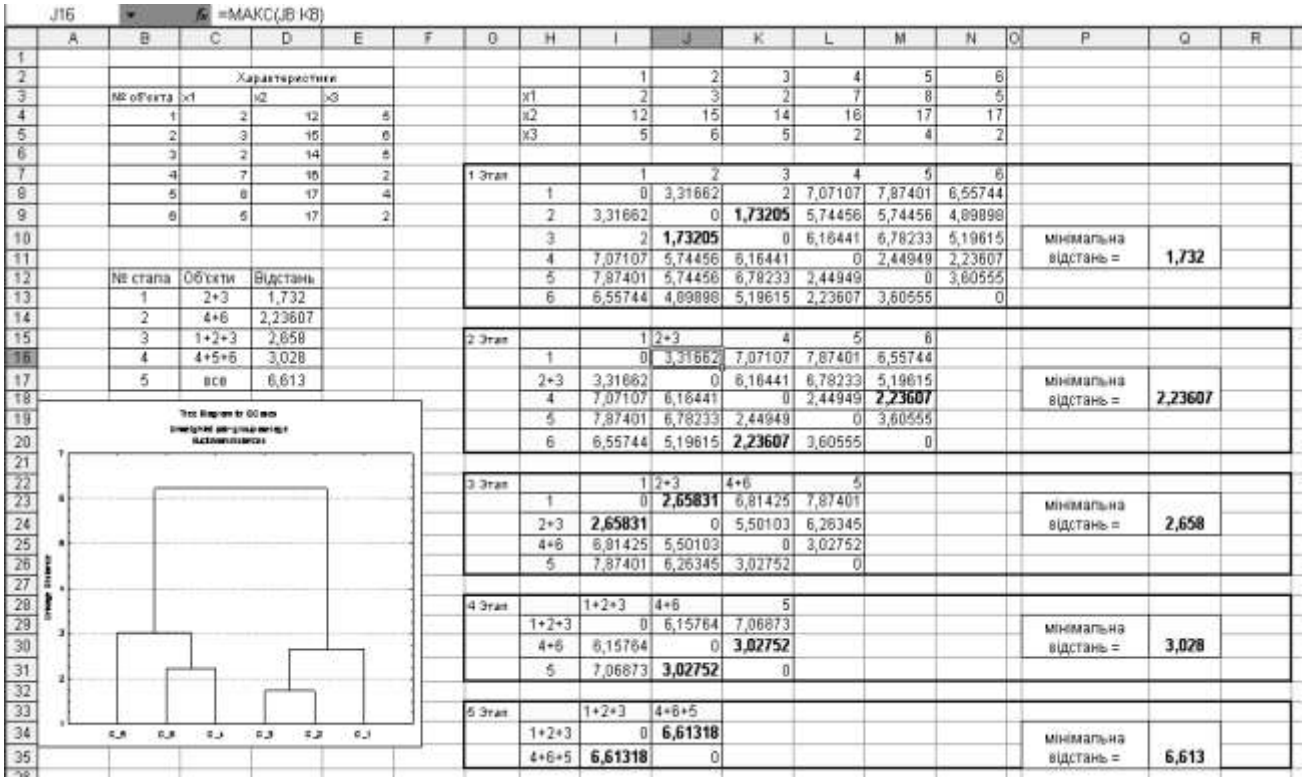



Рис. 10. Процедура кластеризації за методом середнього зв’язку

**Агломеративна процедура за методом центрів тяжіння.**

Ця процедура припускає додатковий етап: для кожного побудованого кластера розраховується координати його центру тяжіння (рис. 11).

Таким чином, перший етап – розрахунок матриці відстаней між кластерами, що містять кожний тільки по одному об’єкту, той же самий, що й в трьох попередніх методах. Але другий етап починається з розрахунку центра новоствореного об’єднаного кластеру: комірки T17-V17. Координати центру кластера розраховуються як середнє арифметичне відповідних координат елементів кластера. Наприклад, координата по ознаці x1 кластеру "2+3" визначається як T17=СРЗНАЧ(C5:C6) в *Excel* та T17=AVERAGE(C5:C6) в *Calc.*

Після розрахунку центра об’єднаного кластера відстань до інших кластерів від цього нового визначається за евклідовою відстанню між центрами кластерів. Наприклад, відстань між кластерами "2+3" та "1", що міститься в комірці J16 дорівнює:

в Excel - =КОРЕНЬ((C4-T$17)^2+(D4-U$17)^2+(E4-V$17)^2);

в Calc - =SQRT((C4-T$17)^2+(D4-U$17)^2+(E4-V$17)^2).

Вся процедура за етапами представлена на рис. 11.

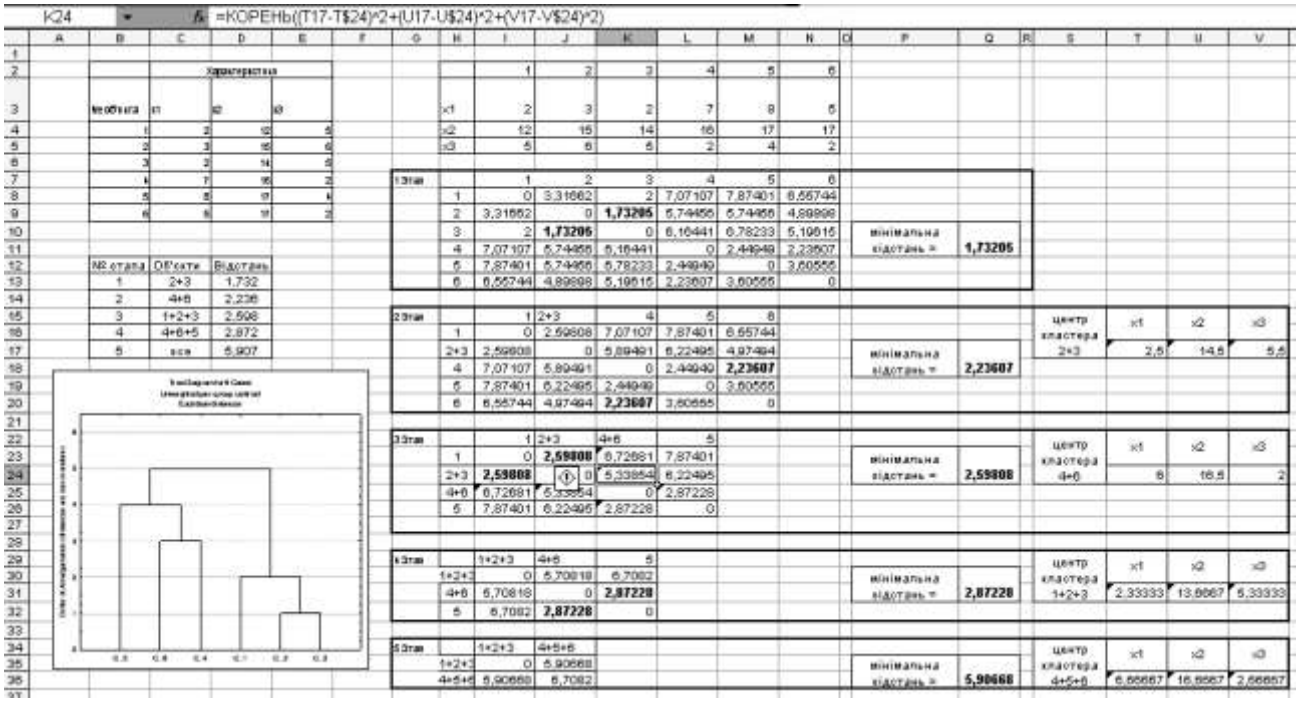


Рис. 11. Процедура кластеризації за методом центрів тяжіння

За результатами 4 процедур кластеризації будуємо дендрограми та у висновках аналізуємо отримані розбиття сукупності об’єктів на кластери, наявність природного розбиття та ін.

**Завдання**

1. За допомогою пакету *Calc* або *Excel* на підставі даних з табл. 1 провести кластеризацію об’єктів, описаних трьома показниками, за

агломеративними процедурами. Кластеризацію провести методом найближчого сусіда, дальнього сусіда, середнього зв’язку, центрів тяжіння. При обчислені відстаней використовувати просту Евклідову відстань. Результати кластеризації представити у вигляді дендрограм. Порівняти результати кластеризації, отримані за різними методами. Зробити висновки щодо наявності природного розбиття сукупності об’єктів на кластери.

Таблиця 1. Вихідні дані для кластерного аналізу

| № об'єкта | Характеристики | | |
|---|---|---|---|
| | x1 | x2 | x3 |
| 1 | 2 | 12 | 5 |
| 2 | 3 | 15 | 6 |
| 3 | 2 | 14 | 5 |
| 4 | 7 | 16 | 2 |
| 5 | 8 | 17 | 4 |
| 6 | 5 | 17 | 2 |

2. Розробити програмний додаток на мові програмування C/C++ в середовищі розробки MS Visual Studio для виконання кластеризації провести методом найближчого сусіда, дальнього сусіда, середнього зв’язку, центрів тяжіння. При обчислені відстаней використовувати просту Евклідову відстань. Результати кластеризації представити у вигляді дендрограм. Порівняти результати кластеризації, отримані за різними методами. Зробити висновки щодо наявності природного розбиття сукупності об’єктів на кластери.  що реалізує розпізнавання образів на основі зазначених в п.1 методів. Похідні дані для обчислень взяти з табл.. 1

3. За результатами досліджень скласти звіт з описом отриманих результатів та обґрунтованими висновками.