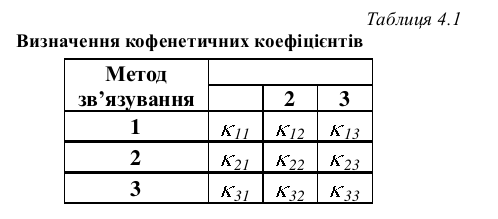
**Лабораторна робота №4**

**ІЄРАРХІЧНІ МЕТОДИ КЛАСТЕРНОГО АНАЛІЗУ**

**Мета роботи**: Практичне засвоєння методів ієрархічного кластерного аналізу даних.

### Завдання до роботи

1. Ознайомитися з конспектом лекцій та рекомендованою літературою, а також додатком Д, що містить короткі теоретичні відомості про ієрархічні методи кластерного аналізу та особливості їх застосування.
2. За допомогою команди >>help вивчити функції pdist, squareform, linkage, dendrogram, cophenet, cluster, gscatter.
3. Завантажити висхідний набір даних для обробки та аналізу згідно до вашого варіанту (табл. 4.2). Побудувати графічне зображення експериментальних даних.
4. Обчислити відстань між об'єктами. Використовувати міри для розрахунку відстаней відповідно до варіанту (табл. 4.2).
5. Використовуючи рекомендоване програмне забезпечення здійснити кластерний аналіз висхідних даних методом ієрархічної кластеризації за методами зв'язування відповідно до варіанту (табл. 4.2).
6. Виконати аналіз якості кластеризації за допомогою обчислення кофенетичного кореляційного коефіцієнту. Заповнити таблицю для кофенетичного кореляційного коефіцієнту (табл. 4.1).



1. Визначити найбільш і найменш ефективні способи ієрархічної кластеризації для аналізу висхідного набору даних (максимальні і мінімальні коефіцієнти та відповідні до них способи кластеризації). Для найбільш ефективного способу ієрархічної кластериації побудувати дендрограму результатів кластерного аналізу.
2. Визначити кількість вірогідних кластерів. Для виділення значущих кластерів використовувати порогове значення, розраховане за метрикою відстаней або методом завдання фіксованого числа кластерів.
3. Розрахувати центри та внутрішньокластерну дисперсію отриманих кластерів, геометричні відстані від елементів до центрів кластерів, відстані між центрами кластерів. Відобразити графічно знайдені кластери та ії центри (скористатися діаграмою розсіювання у кольорі).
4. Оформити звіт з роботи.
5. Відповісти на контрольні питання.

### Зміст звіту

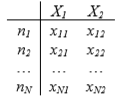
1. Тема та мета роботи.
2. Короткі теоретичні відомості.
3. Номер варіанту та набір даних для обробки (якщо він великий – навес
4. ти фрагмент). Графічне зображення експериментальних даних.
5. Опис процесу використання програмного забезпечення для обробки набору даних, що має бути ілюстрований зображеннями екранних форм.
6. Тексти програм.
7. Результати роботи програмного забезпечення (зображення знайдених кластерів, таблиця для кофенетичного кореляційного коефіцієнту, дендрограма результатів кластерного аналізу тощо).
8. Висновки, що містять відповіді на контрольні запитання, а також відображують результати виконання роботи та їх критичний аналіз

 \* Метрики відстаней: 1 – Евклідова, 2 – стандартизована Евклідова, 3 – міста, 4 – Махаланобіса, 5 – Мінковського (p = 4), 6 –Чебишева.

\*\* Методы зв'язування: a – ближчого сусіда, b – дальшого сусіда, c – середнього зв'язку, d – центроїдний, e – медіанного зв'язку.

*Таблиця 4.3*

**Вигляд висхідних даних для кластеризації**



### Контрольні запитання

1. В чому полягає задача кластерного аналізу?
2. Для яких задач обробки експериментальних даних використовуються методи ієрархічного кластерного аналізу?
3. Наведіть основні міри порівняння об'єктів між собою.
4. Що таке дендрограма?
5. Що являють собою ієрархічні агломеративні методи кластерного аналізу?
6. Що являють собою ієрархічні дивізимні методи кластерного аналізу?
7. Наведіть основні способи зв'язування об'єктів у кластери
8. Що таке кофенетичний кореляційний коефіцієнт?
9. В чому полягають основні етапи ієрархічного кластерного аналізу?
10. Яким чином визначити значущу кількість кластерів?
11. З якою метою використовують функції MATLAB: pdist, square- form, linkage, dendrogram, cophenet, cluster, gscatter?

**Додаток Д**

**ІЄРАРХІЧНІ МЕТОДИ КЛАСТЕРНОГО АНАЛІЗУ У ПАКЕТІ MATLAB**

Вирішення задачі розбиття множини елементів без навчання називають кластерним аналізом. Кластерний аналіз не потребує апріорної інформації про данні та дозволяє розділити множину досліджуваних об'єктів на групи схожих об'єктів - кластери. Це дає можливість значно скорочувати великі об'єми даних, робити їх компактними та наочними.

Задачу кластеризації можна сформулювати таким чином.

Є деяка кінцева множина об'єктів довільної природи , кожний з яких  характеризується набором з *K* – ознак (вимірюваних характеристик об'єктів). Необхідно кластеризувати ці об'єкти, тобто розбити їх множину на задану або довільну кількість груп (кластерів або класів) таким чином, щоб в кожну групу було включено об'єкти близькі між собою в тому чи іншому сенсі. Апріорна інформація про класифікаціє об'єктів при цьому відсутня. Таким чином необхідно розбити множину векторів *C* на *k* попарно неперетинаючихся класів



Для знаходження певного рішення даної задачі необхідно задати:

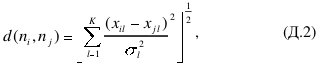
* спосіб порівняння об'єктів між собою (міру схожості);
* спосіб кластерізації;
* розбиття даних за кластерами (встановлення кількості кластерів).

Для порівняння двох об'єктів ni та nj можуть бути використані наступні міри:

1. Евклідова відстань



1. Стандартизована Евклідова відстань



де  - дисперсія за *l*-ю ознакою.

Метрика міста



1. Відстань Мінковського





1. Відстань Махалонобіса

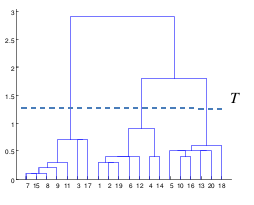


де *С-1* - коваріаційна матриця вхідних даних;

1. Метрика Чебишева



За способами кластеризації методи кластерного аналізу можна розділити на дві великі групи: ієрархічні та неієрархічні методи. Ієрархічні методи, що ми розглядаємо, використовують послідовне об'єднання об'єктів у кластери, малих кластерів у великі, що може бути візуально представлене у вигляді дерева вкладених кластерів - дендрограми (рис. Д.1), при цьому кількість кластерів приховане або умовне. Як правило, на графі дендрограми вздовж осі x розташовують номера об'єктів а вздовж осі y – значення міри схожості.



*Рисунок Д.1* Дендрограма

Ієрархічні методи поділяються на:

* *агломеративні*, що характеризуються послідовним об'єднанням висхідних об'єктів та відповідним зменшенням числа кластерів (побудова кластерів знизу вгору);
* *дивізимні* (ділимі), в яких кількість кластерів зростає починаючи з одного, в результаті чого з'являється послідовність розщеплюваних груп (побудова кластерів згори вниз).

Методи ієрархічного кластерного аналізу відрізняються за способом або методом зв'язування об'єктів у кластери. Якщо кластери *i* та j об'єднуються в кластер *r* та необхідно розрахувати відстань до кластеру *s*, то найбільш часто використовуються такі методи зв'язування об'єктів/кластерів *r* і *s*:

1. метод ближчого сусіда (відстань між найближчими сусідами - найближчими об'єктами кластерів)



1. метод дальнього сусіда (відстань між найбільш далекими сусідами)



1. метод середнього зв'язку (середня відстань між всіма об'єктами пари кластерів з урахуванням відстані всередині кластерів)



де *mi* та *mj* – кількість об'єктів в *i*-ому та *j*-ому кластерах відповідно;

1. центроїдний метод



1. метод медіанного зв'язку



Таким чином, в великий кластер об'єднуються ті малі кластери, відстань між якими мінімальна.

*Оцінка відмінності*. Використовується для визначення найбільш оптимального вибору метрічної відстані та методу зв'язування об'єктів. Два довільно вибраних об'єкти на ієрархічному дереві пов'язані між собою так званою кофенетичною відстанню, величина якої визначається відстанню між двома кластерами, в яких знаходяться данні об'єкти. Мірою лінійного зв'язку між кофенетичними та метричними відстанями є кофенетичний кореляційний коефіцієнт *К*. Побудову ієрархічного дерева вважають успішною, якщо кофенетичний кореляційний коефіцієнт близький до 1. Для найбільш успішної кластеризації будують дендрограму ієрархічного дерева.

*Виділення значущих кластерів*. Для виділення значущих кластерів можна задати деяке порогове значення *T* міри відстаней схожості (горизонтальна перпендикулярна вісь *T* на дендрограмі рис. 4.1). Кількість значущих кластерів визначається кількістю перетинів ліній порогу *T* та зв'язків ієрархічного дерева. Причому кожна з гілок дерева, що відсікається лінією порогу, буде формувати окремий кластер. На практиці часто обирають порогове значення *T* на основі візуального аналізу щільності гілок побудованої дендрограми.

Альтернативний спосіб виділення значущих кластерів – метод завдання фіксованої кількості кластерів. Порогове значення міри схожості *T* встановлюється у корені ієрархічного дерева. Потім значення порогу *T* поступово знижується доти, доки не буде встановлено кількість перетинів лінії порогу *T* і зв'язків ієрархічного дерева таке, що дорівнює заданій кількості кластерів.

**Основні етапи ієрархічного кластерного аналізу в MATLAB**

Етап 1. Обчислення матриці відстанєй між об'єктами D (функція pdist, squareform).

Етап 2. Зв'язування або групування об'єктів в бінарні ієрархічні дерева (дендрограми) (функції linkage, dendrogram).

Етап 3. Оцінка якості кластеризації (функції cophenet).

Етап 4. Виділення значущих кластерів (функція cluster).

Етап 5. Візуалізація і аналіз значущих кластерів (функція gscatter).