МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
СХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМ. В. ДАЛЯ
ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА ЕЛЕКТРОНІКИ
КАФЕДРА КОМП'ЮТЕРНИХ НАУК ТА ІНЖЕНЕРІЇ

До захисту допускається
Завідувач кафедри
_____ І.С. Скарга-Бандурова
«____»_____ 20__ р.

**ДИПЛОМНИЙ ПРОЕКТ (РОБОТА) БАКАЛАВРА**

**ПОЯСНЮВАЛЬНА ЗАПИСКА**

НА ТЕМУ:

# Кластерізація медичних діагностичних даних

Освітній рівень"бакалавр"
Спеціальність 123 "Комп'ютерна інженерії"

Науковий керівник роботи: _____ (підпис) О.І.Рязанцев (ініціали, прізвище)

Консультант з охорони праці: _____ (підпис) Я.О.Критська (ініціали, прізвище)

Здобувач вищої освіти: _____ (підпис) І.М.Гриньків (ініціали, прізвище)

Група: КІ-16з

Сєвєродонецьк 2020

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
СХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВОЛОДИМИРА ДАЛЯ

Факультет Інформаційних технологій та електроніки
Кафедра Комп'ютерних наук та інженерії
Освітній рівень Бакалавр
Спеціальність 123 "Комп'ютерна інженерія"
(шифр і назва)

**ЗАТВЕРДЖУЮ:**

Т.в.о. завідувача кафедри _____
_____ С.О. Сафонова
«_____» _____20____р.

**З А В Д А Н Н Я**
**НА ДИПЛОМНИЙ ПРОЕКТ (РОБОТУ) БАКАЛАВРА**

Гриньків Ірині Миколаївні
(прізвище, ім'я, по батькові)

1. Тема роботи Кластерізація медичних діагностичних даних

керівник проекту (роботи) Рязанцев Олександр Іванович, д.т.н., проф.
(прізвище, м.'я, по батькові, науковий ступінь, вчене звання)
затверджені наказом вищого навчального закладу від «30» 04 2020 р. № 77/15.15
2. Строк подання студентом роботи 10.06.2020
3. Вихідні дані до роботи Матеріали переддипломної практики, теоретичні відомості про кластерний аналіз та задачі кластеризації, математичні та комп'ютерні моделі методів кластерного аналізу, методи графічного представлення результатів кластеризації

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити) Аналіз видів тестування, математична постановка завдання, методи та алгоритми, програмна реалізація, аналіз результатів, охорона праці, висновки

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень)
Електронні плакати

6. Консультанти розділів проекту (роботи)

| Розділ | Прізвище, ініціали та посада консультанта | Підпис, дата | |
|---|---|---|---|
| | | завдання видав | завдання прийняв |
| Охорона праці | Критська Я.О. ст. викл. кафедри КНІ | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

7. Дата видачі завдання 30.04.2020

Керівник _____
(підпис)

Завдання прийняв до виконання _____
(підпис)

**КАЛЕНДАРНИЙ ПЛАН**

| № з/п | Назва етапів дипломного проекту (роботи) | Строк виконання етапів проекту ( роботи ) | Примітка |
|---|---|---|---|
| 1 | Розробка технічного завдання | 01.05.2020-07.05.2020 | |
| 2 | Аналіз завдання, огляд літератури | 08.05.2020-10.05.2020 | |
| 3 | Аналіз технічних засобів та розробка алгоритму | 10.05.2020-13.05.2020 | |
| 4 | Розробка частини проекту "Охорона праці" | 13.05.2020-15.05.2020 | |
| 5 | Проведення навантажувального тестування | 15.05.2020-01.06.2020 | |
| 6 | Оформлення пояснювальної записки та презентації | 2.06.2020-09.06.2020 | |
| | | | |
| | | | |

**Здобувач вищої освіти** _____ І.М.Гриньків
( підпис ) (прізвище та ініціали)

**Науковий керівник** _____ О.І.Рязанцев
( підпис ) (прізвище та ініціали)

# РЕФЕРАТ

Пояснювальна записка дипломної роботи бакалавра: 80 с., 31 рис., 6 табл., 30 бібліографічних джерел посилань.

Робота присвячена вирішенню проблеми кластеризації у прикладені до аналізу часових рядів та візуалізації її результатів. Розглянуті найпоширеніші методи кластеризації та репрезентації відкластеризованих даних. Завдання кластеризації полягає в поділі досліджуваної множини об'єктів на групи «схожих» об'єктів, називаних кластерами.

У ході роботі було розроблено програмний засіб, що дозволяє кластеризувати вхідні дані за допомогою методів , що застосовуються у аналізі часових рядів. Програмний засіб тестувався на вхідному наборі даних медицинських часових рядів, що складається з колекції сигналів серцебиття, отриманих з бази даних діагностичних електрокардіограм.

**Ключові слова:** кластерний аналіз, часові ряди, метрики, кластери, міри близькості, групування, центроїди, матриця трансформацій, візуалізація, електрокардіограми серцебиття.

# ЗМІСТ

# ПЕРЕЛІК СКОРОЧЕНЬ ТА УМОВНИХ ПОЗНАЧЕНЬ

DBA – DTW Barycenter Averaging, метод DTW з усередненням центрів ваги

DBSCAN – Density-Based Spatial Clustering of Applications with Noise, щільнісний алгоритм просторової кластеризації з присутністю шуму

DTW – Dynamic Time Warping, динамічна трансформація часової шкали

CSV – comma-separated values, значення, розділені комою

FCM – Fuzzy Classifier Means, засоби нечіткої класифікації

FOREL – Formal Element, формальний елемент

QRS – поєднання трьох графічних відхилень (кодів Q, R та S), що спостерігаються на типовій електрокардіограмі

SOM – Self-organizing map, карта самоорганізації

VQ – Vector Quantization, векторне квантування

ЕКГ – електрокардіограма

# ВСТУП

Завдання кластеризації полягає в поділі досліджуваної множини об'єктів на групи «схожих» об'єктів, називаних кластерами. Слово кластер англійського походження (cluster), переводиться як згусток, пучок, група. Іноді використовують назви клас, таксон, згущення. Розв'язання задачі розбивки множини елементів на кластери називають кластерним аналізом.

Кластерний аналіз даних передбачає поділ множини об'єктів не кластери (cluster) або класи, таксони, згущення, групи. Для задач кластеризації характерна відсутність відмінностей об'єктів по атрибутах (змінним). Термін кластерний аналіз уперше введений Тріоном (Tryon) в 1939 році [1]. При проведенні кластерного аналізу не будують апріорних припущень про заданий набір даних, не вводять обмежень на представлення об'єктів аналізу й типи даних. Кластерний аналіз також можна використовувати для скорочення розмірності й візуалізації даних. У цей час кластерний аналіз розвивається в напрямках, пов'язаних з комерційною діяльністю, технічними науками, біологією й психологією.

Кластерний аналіз виконує наступні основні завдання:

– розробка типології або класифікації;

– дослідження корисних концептуальних схем групування об'єктів;

– породження гіпотез на основі дослідження даних;

– перевірка гіпотез або дослідження для визначення, чи дійсно групи, виділені тим чи іншим способом, присутні в наявних даних.

Кластеризація може бути сформульована також як задача багатокритеріальної оптимізації.

Завдання кластеризації може відноситися до статистичної обробки вихідних даних, а також до широкого класу задач навчання без учителя [2].

Кластеризація — процес розбивки заданої вибірки об'єктів (спостережень) на підмножини (як правило, що не перетинаються), називані

кластерами, так, щоб кожний кластер складався зі схожих об'єктів, а об'єкти різних кластерів суттєво відрізнялися.

Однієї із цілей кластеризації є виявлення внутрішніх зв'язків між даними шляхом визначення кластерної структури. Розбивка спостережень на групи схожих об'єктів дозволяє спростити подальшу обробку даних і прийняття розв'язків, застосовуючи до кожного кластера свій метод аналізу.

Одним з додатків кластеризації є розв'язок задачі стискання даних.

У випадку, якщо вихідна вибірка надлишково велика, то можна скоротити її, залишивши трохи найбільш характерних представників від кожного кластера.

Іншою сферою використання кластеризація є виявлення новизни в досліджуваній множині об'єктів. Виділяються нетипові об'єкти, які не вдається приєднати до жодного із кластерів. Для розв'язку задач методами кластерного аналізу, необхідно задавати кількість кластерів заздалегідь. В одному випадку число кластерів намагаються зробити поменше. В іншому випадку важливіше забезпечити високий ступінь подібності об'єктів усередині кожного кластера, а кластерів може бути скільки завгодно. У третьому випадку найбільший інтерес представляють окремі об'єкти, що не вписуються ні в один із кластерів.

# 1 ОГЛЯД ОСНОВНИХ ЗАСОБІВ КЛАСТЕРНОГО АНАЛІЗУ

## 1.1 Постановка задачі кластеризації

Звичайною задачею кластерного аналізу є розбивка на групи всієї множини об'єктів $\mathrm{Q} = \{\mathrm{q_j}\}_{\mathrm{j=1}}^{\mathrm{n}}$, де n – їх загальне число [3].

Формальна постановка задачі кластеризації здійснюється в такий спосіб. Визначається множина об'єктів даних

$$Q = \{q_j\}_{j=1}^{n} = \{q_1, q_2, \dots, q_n\}. \quad (1.1)$$

Кожний об'єкт $\mathrm{q_j}$ може мати набір атрибутів:

$$\mathrm{x_j} = (\mathrm{x_{j1}}, \mathrm{x_{j2}}, \dots, \mathrm{x_{jm}}). \quad (1.2)$$

Прикладом такої множини об'єктів може бути, наприклад, множина аварій протягом певного часу на промислових об’єктах, або об’єктах мобільного зв’язку, або електростанціях, кожна з яких характеризується набором показників (атрибутів) тривалості, категорії наслідків, ступеня пожежної і техногенної безпеки, зв’язку з часом доби, пори року, завантаження об’єкта тощо.

Розв’язком задачі кластеризації є множина сформованих кластерів

$$C = \{c_k\}_{k=1}^{g} = \{c_1, c_2, \dots, q_g\}, \quad (1.3)$$

де $c_k$ - кластер, що містить схожі об'єкти із множини $Q$:

$$c_k = \{q_i, q_j | d_{i,j} < \sigma\}. \quad (1.4)$$

В цьому виразі $d_{i,j} = d(x_i, x_j)$ - міра близькості між об'єктами $q_i, q_j$, яка визначається з урахуванням наборів (векторів) їх атрибутів $x_i, x_j$.

Величина $\sigma$ визначає порогове значення для міри близькості між об'єктами.

Усі алгоритми (методи) кластеризації в цілому розділяють на ієрархічні й неієрархічні алгоритми.

Ієрархічні алгоритми кластерного аналізу у свою чергу розділяють на агломеративні (що збирають) й дивізимні (що розділяють).

## 1.2 Основні поняття, які використовуються в задачах кластеризації

До основних засобів, які використовуються в задачах кластеризації, відносяться відстані, метрики і міри близькості.

Ці засоби, іноді суттєво вирізняючись між собою, використовуються в різноманітних алгоритмах (методах) кластеризації, яких загалом відомо декілька сотень.

### 1.2.1 Метрики й відстані

Метрика – це функція на парах елементів множини [7]. Метрики належать до основних засобів формування альтернатив класів у задачах розпізнавання і кластеризації, де оцінки приналежності до класу обчислюються однозначним способом.

Ці функції (метрики) використовуються для обчислення відстаней між об’єктами $q_i, q_j$ і у термінах метричних просторів і є відстанями між

об’єктами цих просторів. Ці обчислені відстані є мірами близькості цих об’єктів.

Якщо набори атрибутів $x_j = (x_{j1}, x_{j2}, \dots, x_{jm})$ можна представити у вигляді m-вимірних числових векторів, які є елементами m-вимірного метричного простору дійсних чисел, то кожна змінна з набору $x_j$ може приймати значення із множини дійсних чисел, що дає можливість застосувати відповідні метрики метричного простору дійсних чисел для групування (кластеризації) об’єктів.

Інакше кажучи, множина об’єктів $Q = \{q_j\}_{j=1}^{n}$ повністю визначається множиною даних $X_Q$, яка є підмножиною m-вимірного простору дійсних чисел $R^m$ [3]:

$$X_Q = \{x_j\}_{j=1}^{n} \subseteq R^m, \tag{1.1}$$

$$x_j = (x_{j1}, x_{j2}, \dots, x_{jm}), \quad j = \overline{1,n}. \tag{1.2}$$

Для множини даних $X_Q$ можна визначити m-вимірний вектор середніх значень точок даних $\bar{x}$ і коваріаційну матрицю $S$ розмірності $m \times m$, які використовуються в розв’язаннях задач кластеризації:

$$\bar{x} = (\overline{x_k})_{k=1}^{m} = (\overline{x_1}, \overline{x_1}, \dots, \overline{x_{jm}}), \tag{1.3}$$

$$\overline{x_k} = \frac{1}{n}\sum_{j=1}^{n} x_{jk}, \tag{1.4}$$

$$S = \left\| \frac{1}{n-1}\sum_{j=1}^{n} (x_j - \bar{x})(x_j - \bar{x})^T \right\|. \tag{1.5}$$

Ненегативне значення $d_{i,j} = d(x_i, x_j)$ називається відстанню між елементами $x_i, x_j$, якщо виконуються наступні чотири умови:

$$\forall x_i, x_j: \quad d(x_i, x_j) > 0, \tag{1.6}$$

$$x_i = x_j \Leftrightarrow d(x_i, x_j) = 0, \tag{1.7}$$

$$d(x_i, x_j) = d(x_j, x_i), \tag{1.8}$$

$$d(x_i, x_j) \leq d(x_i, x_l) + d(x_l, x_j). \tag{1.9}$$

Тобто:

а) $d(x_i, x_j) > 0$, для всіх $x_i, x_j$, – відстань завжди невід'ємна (1.6);

б) $d(x_i, x_j) = 0$, тоді й тільки тоді, коли $x_i = x_j$, – об'єкти співпадають (1.7);

в) $d(x_i, x_j) = d(x_j, x_i)$ – відстань від об'єкта $q_i$ до об'єкта $q_j$ завжди дорівнює зворотній відстані від об'єкта $q_j$ до об'єкта $q_i$ (1.8);

г) $d(x_i, x_j) \leq d(x_i, x_l) + d(x_l, x_j)$ – відстань від об'єкта $q_i$ до об'єкта $q_j$ завжди менша або дорівнює сумі відстаней між ними та третім об'єктом (1.9).

Якщо відстань $d(x_i, x_j)$ менше деякого значення $\sigma$, то приймають рішення, що елементи близькі і їх містять в один кластер. А якщо ні, то приймають рішення, що елементи відмінні друг від друга і їх містять у різні кластери.

Велика кількість алгоритмів розв'язання задачі кластеризації використовують у якості формату вхідних даних матрицю відмінності (відстаней) $D$. Рядки й стовпці матриці відповідають елементам множини $X_Q$. Елементами матриці є значення $d(x_i, x_j)$ у рядку $i$ і стовпці $j$. Очевидно, що на головній діагоналі значення будуть дорівнювати нулю:

$$D = \begin{pmatrix} 0 & d_{1,2} & \dots & d_{1,n} \\ d_{2,1} & 0 & \dots & d_{2,n} \\ \dots & \dots & \dots & \dots \\ d_{n,1} & d_{n,2} & \dots & 0 \end{pmatrix}. \tag{1.10}$$

Таблиця 1.1 – Зразок матриці відстаней

| $x_j$ | Елемент | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 0,471666 | 1 | 0 | 1,662236 | 1,285075 | 0,924313 | 1,52347 |
| -1,19057 | 2 | 1,662236 | 0 | 0,377161 | 2,586548 | 0,138766 |
| -0,81341 | 3 | 1,285075 | 0,377161 | 0 | 2,209388 | 0,238394 |
| 1,395979 | 4 | 0,924313 | 2,586548 | 2,209388 | 0 | 2,447782 |
| -1,0518 | 5 | 1,52347 | 0,138766 | 0,238394 | 2,447782 | 0 |

Іноді для підготовки вхідних даних використовують більш спеціальні методи, наприклад, з використанням карт самоорганізації [17].

Для розглянутої множини об'єктів із простору $R^m$ при кластеризації найчастіше використовують наступні відстані:

Евклідова відстань (Euclidian Distance):

$$d_{E1}(x_i, x_j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}. \quad (1.11)$$

Цю відстань часто підносять до квадрата, щоб додати більше ваги більш віддаленим одне від одного об'єктам, тобто

$$d_{E2}(x_i, x_j) = \sum_{k=1}^{m} (x_{ik} - x_{jk})^2. \quad (1.12)$$

Відстань за Хемінгом:

$$d_H(x_i, x_j) = \sum_{k=1}^{m} |x_{ik} - x_{jk}|. \quad (1.13)$$

Ця відстань є просто середньою різниць по координатах. У більшості випадків дана міра відстані приводить до таких же результатів, як і для

звичайної відстані Евкліда, однак для неї вплив окремих більших різниць (викидів) зменшується, тому що вони не підносяться до квадрата).

Відстань Чебишева:

$$d_{\infty}(x_i, x_j) = \max_{k=1,m} |x_{ik} - x_{jk}|. \quad (1.14)$$

Ця відстань може виявитися корисною, коли бажають визначити два об'єкти як «різні», якщо вони різняться по якій-небудь одній координаті (яким-небудь одним виміром).

Відстань Махаланобіса (Mahalanobis Distance):

$$d_M(x_i, x_j) = \sqrt{(x_i - x_j)S^{-1}(x_i - x_j)^T}. \quad (1.15)$$

Дана міра відстані використовує коваріаційну матрицю. Але вона погано працює, якщо коваріаційна матриця вираховується на всій множині вхідних даних. У той же час, будучи зосередженою на конкретному класі (групі даних), дана міра відстані показує гарні результати.

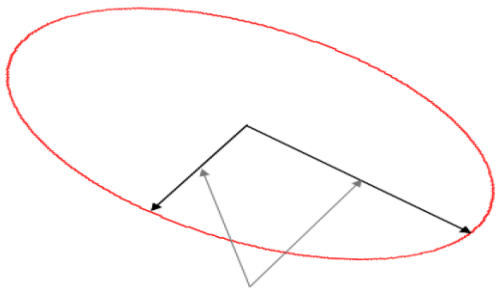


Рисунок 1.1 – Вектори коваріаційної матриці і еліпс множини рівновіддалених із центра точок в сенсі відстані Махаланобіса

Пікова відстань:

$$d_I(x_i, x_j) = \frac{1}{m}\sum_{k=1}^{m} \frac{|x_{ik} - x_{jk}|}{x_{ik} + x_{jk}}. \quad (1.16)$$

Дана міра відстані припускає незалежність між випадковими змінними, що передбачає відстань в ортогональному просторі. В практичних додатках ці змінні не є незалежними.

У багатьох випадках замість відстані як міри близькості використовується значення косинуса кута між двома векторами

$$\cos\varphi_{ij} = \frac{x_i \cdot x_j}{|x_i| \cdot |x_j|} \tag{1.17}$$

або коефіцієнт кореляції

$$\rho_{ij} = \frac{\sum_{i,j=1}^{n}(x_i - \overline{x_i})(x_j - \overline{x_j})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x_i})^2}\sqrt{\sum_{j=1}^{n}(x_j - \overline{x_j})^2}}. \tag{1.18}$$

Для визначення відстаней між кластерами часто використовується розрахунок відстаней до об’єктів або елементів, які входять в ці кластери. Відстань до всіх елементів бази можна оцінити шляхом обчислення відстані до одного з елементів. Це значно скорочує час процедур порівняння, тому що внутрішні відстані можна обчислити заздалегідь на попередньому етапі.

Найбільш застосовуваними для розрахунку відстаней між групами (кластерами) є відстані за принципом «близького сусіда», за принципом «далекого сусіда» і відстань між «центрами ваги» груп (кластерів) [14].

Відстань за принципом «близького сусіда» (метод одиночного зв’язку):

$$d_{bs}(x_i, x_j) = \min_{p,t} d(x_{ip}, x_{jt}). \tag{1.19}$$

В цьому виразі $x_{ip}, x_{jt}$ – це $p$-й і $t$-й елемент відповідної групи (кластера) В свою чергу, відстань $d(x_{ip}, x_{jt})$ між відповідними елементами

об’єктів і кластерів може визначатися різними способами, наприклад, за Евклідом, Хемінгом або Чебишевим.

Відстань за принципом «далекого сусіда» (метод повного зв’язку):

$$d_{ds}(x_i, x_j) = \max_{p,t} d(x_{ip}, x_{jt}). \tag{1.20}$$

Тут також $x_{ip}, x_{jt}$ – це $p$-й і $t$-й елемент відповідної групи (кластера) В свою чергу, відстань $d(x_{ip}, x_{jt})$ між відповідними елементами об’єктів і кластерів може визначатися різними способами, наприклад, за Евклідом, Хемінгом або Чебишевим.

Відстань методом середнього зв’язку:

$$d_{ms}(x_i, x_j) = \frac{1}{n_i n_j} \sum_{x_{ip} \in X_i} \sum_{x_{jt} \in X_j} d(x_{ip}, x_{jt}). \tag{1.21}$$

Тут $n_i, n_j$ – кількість елементів в кластерах $X_i, X_j$:

$$n_i \mathrm{card} X_i, \tag{1.22}$$

$$n_j \mathrm{card} X_j. \tag{1.23}$$

Відстань між двома кластерами вважається рівною середній відстані між елементами цих кластерів.

Відстань між «центрами ваги» груп:

$$d_{ct}(x_i, x_j) = d(\overline{x_i}, \overline{x_j}). \tag{1.24}$$

В цій метриці $\overline{x_i}, \overline{x_j}$ є середніми значеннями точок даних відповідних груп і розраховуються за формулами (1.3) і (1.4). Відстань між кластерами покладається рівним відстані між їх центроїдами (центрами маси).

Відстань Уорда.

Для оцінки відстаней між кластерами тут використовуються методи дисперсійного аналізу. Як відстань між кластерами береться приріст суми квадратів відстаней об'єктів до центру кластера $X_k = X_i \cup X_j$, одержаного в результаті їх об'єднання:

$$\Delta = \sum_{x_k \in X_k} (d(x_k, \overline{x_k})^2) - \sum_{x_i \in X_i} (d(x_i, \overline{x_i})^2) - \sum_{x_j \in X_j} \left(d\left(x_i, \overline{x_j}\right)^2\right). \quad (1.25)$$

Цей метод застосовується для задач з близько розташованими кластерами.

Застосування відстаней у порівнянні із іншими мірами близькості має перевагу в тому, що при пошуку об'єктів в об'ємних базах даних можна спиратися при обчисленні відстаней на відстані усередині бази даних.

### 1.2.2 Міри близькості

Відстань $d\left(x_i, x_j\right)$ є мірою близькості об’єктів. Але існують і інші міри близькості об’єктів.

Для визначення подібності об'єктів у теорії кластеризації й розпізнавання використовується також більш загальне поняття величини міри близькості $r\left(q_i, q_j\right)$ об'єктів.

Приклад міри близькості:

$$r(q_i, q_j) = \frac{v(q_i, q_j)}{v_0}. \quad (1.26)$$

де $v(q_i, q_j)$ – число однакових ознак (атрибутів) у порівнюваних об'єктів $q_i, q_j$, а величина $v_0$ – загальне число ознак (атрибутів) або число відповідностей ознак.

Умови, яким відповідають міри близькості, аналогічні розглянутим раніше для відстаней:

$$\forall(q_i, q_j): \quad r(q_i, q_j) > 0, \quad (1.27)$$

$$q_i = q_j \Leftrightarrow r(q_i, q_j) = 0, \quad (1.28)$$

$$r(q_i, q_j) = r(q_i, q_j), \quad (1.29)$$

$$r(q_i, q_j) \leq r(q_i, q_j) + r(q_i, q_j). \quad (1.30)$$

Для міри близькості, на відміну від відстані, не завжди потрібне обов’язкове виконання умови (1.30). Ця умова обов’язкова лише для тих мір близькості, які є відстанями.

При виконанні нерівності $r(q_i, q_j) < \sigma$ об'єкти із множини Q розглядаються як близькі й містяться в один кластер. Інакше об'єкти містяться в різні кластери.

Ми вже розглядали міри близькості, які є відстанями: Евклідова відстань, відстань за Хемінгом, відстань Чебишева, відстань Махаланобіса, тощо. Розглянемо деякі інші міри близькості.

Існує ряд асоціативних мір близькості на основі множин. Визначимо для множин атрибутів об’єктів (груп, кластерів) операції їх перетину $X \cap Y$ й різниці $X \backslash Y$, $Y \backslash X$.

$$X = v(q_i), Y = v(q_j) \big| q_i, q_j \in Q \quad (1.31)$$

Нехай потужності (число елементів) цих операцій дорівнюють:

$$a = \mu(X \backslash Y), \tag{1.32}$$

$$b = \mu(Y \backslash X), \tag{1.33}$$

$$c = \mu(X \cap Y). \tag{1.34}$$

Існує ряд асоціативних мір близькості на основі вказаних потужностей:

Міра близькості Жаккара:

$$r_J(q_i, q_j) = \frac{c}{a + b + c}. \tag{1.35}$$

Історично була першою запропонованою мірою близькості. Міра близькості Жаккара пов'язана однією монотонно зростаючою залежністю з мірою близькості Сьоренсена и мірою Сокала-Сніта.

Міра близькості Сьоренсена:

$$r_S(q_i, q_j) = \frac{2c}{a + b}. \tag{1.36}$$

Використовується досить часто, зокрема в наукових дослідженнях для визначення ступеня взаємозалежності двох ознак.

Міра близькості Сокала-Сніта:

$$r_{SS}(q_i, q_j) = \frac{c}{2(a + b) + c}. \tag{1.37}$$

Міри близькості Дейка:

$$r_{D_1}(q_i, q_j) = \frac{2c}{a + b + 2c}, \tag{1.38}$$

$$r_{D_2}(q_i, q_j) = \frac{c - \min(a, b)}{c + \min(a, b)}. \tag{1.39}$$

Міра близькості Кульчинського:

$$r_K(q_i, q_j) = \frac{a+b}{2ab}. \quad (1.40)$$

Використовується рідко.

Міра незгоди Танімото:

$$r_T(q_i, q_j) = \frac{a+b}{a+b+c}. \quad (1.41)$$

Ця міра використовується в основному для підтвердження того, що об'єкти не входять до одного кластера. Цю міру можна виразити через інші операції алгебри множин (симетричну різницю $X\Delta Y$ й об'єднання $X \cup Y$) [7)]:

$$r_T(q_i, q_j) = \frac{\mu(X \backslash Y) + \mu(Y \backslash X)}{\mu(X \backslash Y) + \mu(Y \backslash X) + \mu(X \cap Y)} = \frac{\mu(X \Delta Y)}{\mu(X \cup Y)}. \quad (1.42)$$

Міра незгоди Танімото і міра близькості Жаккара доповнюють одне одного до 1:

$$r_T(q_i, q_j) + r_J(q_i, q_j) = \frac{a+b}{a+b+c} + \frac{c}{a+b+c} = 1. \quad (1.43)$$

Міра близькості Отіаї:

$$r_O(q_i, q_j) = \frac{c}{\sqrt{ab}}. \quad (1.44)$$

Часто використовується міра близькості Отіаї, піднесена до квадрату.

Міра близькості Шимкевича-Сімпсона:

$$r_{Sh}(q_i, q_j) = \frac{c}{\min(a, b)}. \quad (1.45)$$

Ця міра зустрічається під назвою «коефіцієнт перекриття» (overlap coefficient).

Міра близькості Браун-Бланке:

$$r_B(q_i, q_j) = \frac{c}{\max(a, b)}. \quad (1.46)$$

Міра близькості Браун-Бланке спочатку була записана через неправильну інтерпретацію міри Жаккара, але потім знайшла використання в біології, екології та інших галузях [16].

Для порівняння груп об'єктів іноді використовуються багатовимірні міри близькості (multiple-site similarity measure, multidimensional coefficient, multiple-community measure), тобто $r(q_1, q_2, \dots, q_n)$. До них належать: середня подібність Альохіна, індекс біотичної дисперсії Коха, коефіцієнт розсіювання (дисперсності) Шеннікова, міра бета-різноманітності Уіттекера, міра гомотонности й двоїста їй міра гетеротонності Міркіна-Розенберга, коефіцієнт подібності серії описів Семкіна.

Найбільш відома багатовимірна міра була запропонована Л.Кохом [16]:

$$r(q_1, q_2, \dots, q_n) = \frac{\sum_{j=1}^{n} v(q_j) - v(q_1, q_2, \dots, q_n)}{(n-1) \cdot v(q_1, q_2, \dots, q_n)}, \quad (1.47)$$

де $\sum_{j=1}^{n} v(q_j)$ – сума числа ознак (атрибутів) у всіх порівнюваних об'єктів;

$v(q_1, q_2, \dots, q_n)$ –кількість однакових ознак (атрибутів) у всіх порівнюваних об'єктів, $n$ – кількість порівнюваних об'єктів,

### 1.2.3 Нормування й порівняння векторів даних

Компоненти векторів набору атрибутів $x_j = (x_{j1}, x_{j2}, \dots, x_{jm})$ часто приймаються нормованими на відрізку [0,1], щоб можна було контролювати діапазон зміни метрик і мір близькості, а також вибирати пороги для прийняття рішень.

Виконаємо нормування компонентів вектора набору атрибутів $x_j \in X_Q$, привівши їх значення до однакового діапазону, наприклад, до відрізка [0,1], наприклад:

$$x_{jk}^{\cdot} = \frac{x_{jk} - \min\limits_{k} x_{jk}}{\max\limits_{k} x_{jk} - \min\limits_{k} x_{jk}}, \tag{1.48}$$

де $x_{jk}^{\cdot}$ - нормовані значення атрибутів даного вектору;

$x_{jk}$ – ненормовані значення атрибутів всіх вимірів $k$ вектору $x_j$ розмірності $m$, $k = \overline{1,m}$.

Іншим варіантом нормування є

$$x_{jk}^{\cdot} = \frac{x_{jk}}{\|x_j\|}, \tag{1.49}$$

де $\|x_j\|$ – норма вектору $x_j$, яка теж може визначатися різними способами, наприклад,

$$\|x_j\| = \sqrt{\sum_{k=1}^{m} {x_{jk}}^2}. \tag{1.50}$$

В результаті такого нормування всіх $x_j \in X_Q$ отримаємо нормований векторний простір з евклідовими метриками. В геометричному трактуванні

такого m-вимірного простору всі вектори $x_j$ будуть вирізнятися не довжиною, а напрямками.

Якщо застосувати нормування, то це приведе до нормування на відрізку [-1,1].

$$x_{jk}^{\cdot} = \frac{x_{jk}}{\max\limits_{k}|x_{jk}|}, \tag{1.51}$$

Нормовані вектори $\dot{x}_\iota, \dot{x}_J$ можна порівнювати на основі метрики відстані $d(\dot{x}_\iota, \dot{x}_J)$, будь-якої міри подібності $r(\dot{x}_\iota, \dot{x}_J)$ або з використанням норм $\|x_i\|$ і $\|x_j\|$.

Крім класичних мір подібності, для нормованих векторів іноді доцільно застосовувати міри виду $r(x_{\iota k}^{\cdot}, x_{jk}^{\cdot})$, пов'язані з аналізом окремих компонентів, що точніше враховує їхні відхилення.

Для таких мір умови близькості (порогові значення) звичайно вибираються не для однієї міри одного компонента (виміра $k$), а декількох, наприклад:

$$\forall k = \overline{1,m}: \quad r(x_{\iota k}^{\cdot}, x_{jk}^{\cdot}) \le \sigma. \tag{1.52}$$

При цьому з пороговим значенням порівнюються по черзі усі міри близькості (відхилення) нормованих компонент вектора.

Інший приклад умов близькості

$$\sum_{k=1}^{m} r(x_{\iota k}^{\cdot}, x_{jk}^{\cdot}) \le \sigma. \tag{1.53}$$

При цьому з пороговим значенням порівнюється одразу сума усіх мір близькості (відхилень) нормованих компонент вектора.

## 1.3 Особливості кластеризації часових рядів

Актуальною є задача поділу множини об'єктів на окремі групи – кластери, для подальшого прогнозування поведінки часового ряду.

Дані часового ряду є однієї з найпоширеніших форм даних, з якими зустрічаються у великій різноманітності сценаріїв, таких як фондові ринки, дані датчика, контроль відмови, контроль стану машини, екологічні застосування або медичні дані. Проблема кластеризації знаходить численне застосування в областях часового ряду, таке як визначення груп об'єктів з подібними тенденціями. Кластеризація часового ряду має численні застосування в різноманітних проблемних областях.

Дані часового ряду перебувають у межах класу контекстних подань даних. Багато видів даних, таких як дані часового ряду, дискретні ряди й просторові дані знаходяться у цьому класі.

### 1.3.1 Атрибути часових рядів

Часові дані містять два види атрибутів:

– контекстний атрибут;

– поведінковий атрибут.

Для випадку даних часового ряду контекстний атрибут відповідає виміру часу. Ці атрибути забезпечують контрольні точки, у яких виміряються поведінкові значення. Мітки часу могли відповідати фактичним тимчасовим вартостям, у яких виміряються точки даних, або вони могли відповідати індексам, у яких виміряються ці значення.

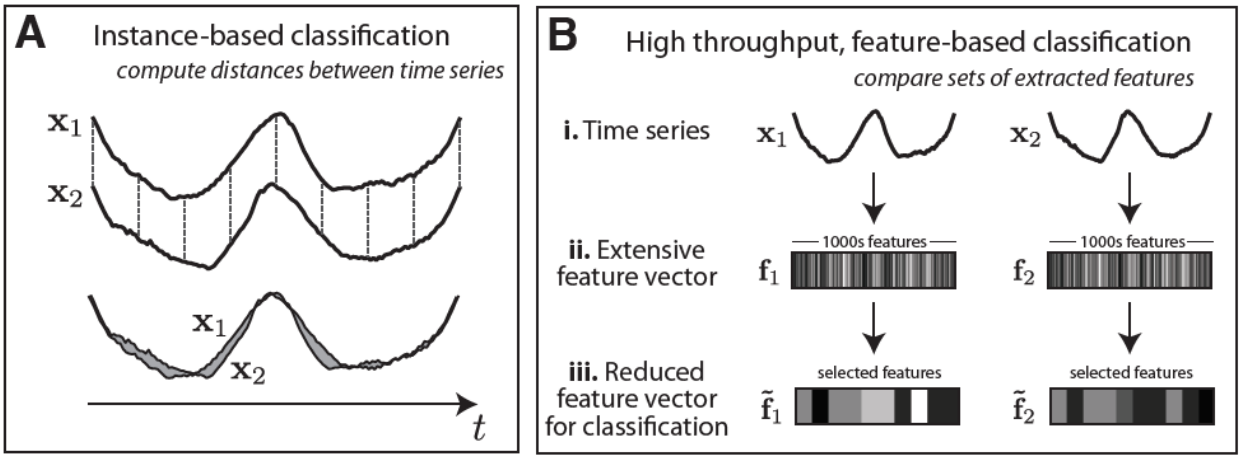



Рисунок 1.1 – Порівняння часових рядів на підставі контекстних і поведінкових атрибутів

Поведінковий атрибут може відповідати будь-якому виду поведінки, яка виміряється в контрольній точці. Деякі приклади включають значення біржевих даних, виміру датчика, такі як температура, інші медичні часові ряди тощо.

### 1.3.2 Вибір міри близькості кластеризації часових рядів

Визначення кластерів часового ряду надзвичайно складно через труднощі при визначенні близькості через різний часовий ряд, який може масштабуватися й перекладатися по-іншому й на часових й на поведінкових розмірах. Тому поняття близькості є дуже важливим для кластеризації даних часового ряду. Зверніть увагу, що, як тільки міра близькості була визначена для даних часового ряду, її можна розглядати як абстрактний об'єкт, на якім може використовуватися множина заснованих на подобі методів, таких як спектральні методи або методи поділу.

Дані часового ряду дозволяють різноманітні формулювання для процесу кластеризації, залежно від того, чи кластеризируються ряди на

основі їх кореляцій онлайн, або чи кластеризируються вони на основі їх форм. Перший звичайно виконується з підходом онлайн на основі маленького вікна минулої історії, тоді як останній звичайно виконується з офлайновим підходом до всього ряду.

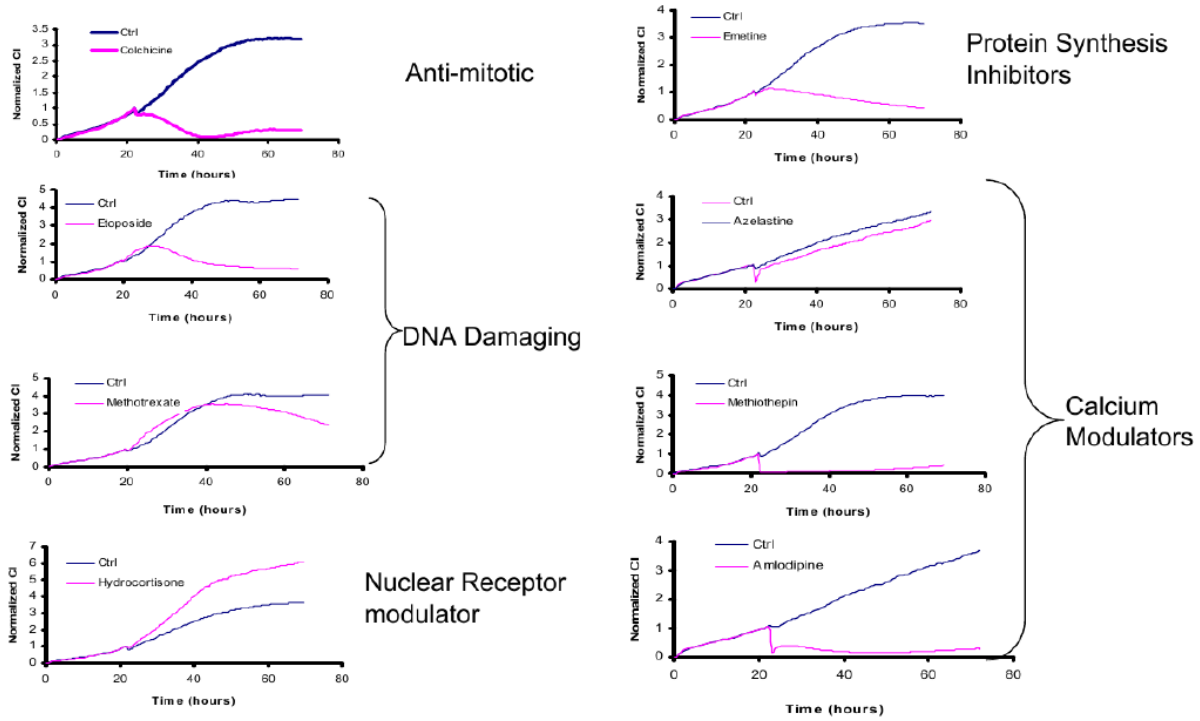



Рисунок 1.2 – Часові ряди в медицині

Однак деякі певні властивості, які є частиною природи даних часового ряду — такі, як висока розмірність, присутність шуму й висока кореляція — ставлять унікальні проблеми перед розробкою ефективних алгоритмів кластеризації. У часовому ряді надто важливо вирішити, яка близькість важлива для кластеризації:

– близькість у часі;

– близькість у формі;

– близькість у зміні даних.

Відповідно повинні бути обрані алгоритм, що відповідає кластеризації, й відповідна міра близькості. Наприклад, Евклідова відстань відображає подобу в часі, у той час як міра динамічної трансформації часової шкали (DTW) [18] відображає подобу у формі. Інші підходи, наприклад, засновані

на моделі прихованих марківських процесів (HMM), застосовуються, коли має значення близькість в зміні часових даних.

Значна різниця між кластеризацією даних часового ряду й кластеризацією простих об'єктів в Евклидовому просторі - те, що часовий ряд, який буде кластеризуватися, не завжди може мати рівні довжини послідовностей. Коли усі часові ряди мають рівну довжину, можуть бути застосовані стандартні методи кластеризації шляхом представлення кожного часового ряду як вектора й використання традиційної нормованої відстані. З таким підходом вчасно може бути використана тільки близькість у часі, у той час як близькість у формі й близькість в зміні ігноруються.

# 2 МАТЕМАТИЧНІ МОДЕЛІ МЕТОДІВ КЛАСТЕРНОГО АНАЛІЗУ

Обробка даних спрямована, як правило, на побудову математичної моделі об'єкта або явища, а також на одержання відповіді на запитання: «Чи достовірні наявні дані в межах необхідної точності або допусків?» [16].

Сама ж математична модель залежно від цілей (дослідження, управліннння, контроль) може бути використана для різних цілей: для предметно-значеннєвого аналізу об'єкта або явища, прогнозування їх стану в різних умовах функціонування, керування ними в конкретних ситуаціях, оптимізації окремих параметрів, а також для розв'язку якихось інших специфічних задач.

Кінцевою метою будь-якої обробки даних є висування гіпотез про клас і структурі математичної моделі досліджуваного явища, визначення складу й об'єму додаткових вимірів, вибір можливих методів наступної статистичної обробки й аналіз виконання основних передумов, що лежать у їхній основі.

## 2.1 Класифікація алгоритмів кластеризації

Алгоритм кластеризації - це функція $X_Q \to C$, яка для будь-якого об'єкта $x_i \in X_Q$ ставить у відповідність номер j кластера $c_j \in C$ [5)]

Множина кластерів $C = \{c_k\}_{k=1}^{g}$ (1.9) в деяких випадках відома заздалегідь, однак частіше ставиться завдання визначити оптимальне число кластерів, з точки зору того чи іншого критерію якості кластеризації.

Загальноприйнятою класифікації методів кластеризації не існує, але можна виділити ряд груп підходів [8)]. Можливі підходи до класифікації методів (алгоритмів) кластеризації:

1) Ієрархічний підхід. Передбачається наявність вкладених груп (кластерів різного порядку). Алгоритми в свою чергу поділяються на:

- агломеративні (об'єднавчі);
- дивізимні (ділильні). Виділяється ієрархічна дивізімна кластеризація або кількісна таксономія.

2) За кількістю ознак іноді виділяють методи класифікації:

- монотетичні;
- політетічні.

3) Імовірнісний підхід. Передбачається, що кожен даний об'єкт відноситься до одного з k класів. Деякі автори [9)] вважають, що дана група не відноситься до кластеризації і протиставляють її під назвою «дискримінація», тобто вибір віднесення об'єктів до однієї з відомих груп (навчальних вибірок). Імовірнісний підхід використовується в таких методах, як:

- K-середніх (k-means);
- K-медіан;
- EM-алгоритм (максимальної правдоподібності) [12)];
- алгоритми сімейства FOREL (формального елементу)[13)];
- дискримінантний аналіз.

4) Підходи на основі систем штучного інтелекту:

- метод нечіткої кластеризації C-середніх (C-means);
- нейронна мережа Кохонена;
- генетичний алгоритм.

5) Логічний підхід. Побудова дендрограми здійснюється за допомогою дерева рішень.

6) Теоретико-графічний підхід. Графові алгоритми кластеризації.

Існує багато (декілька сотень) конкретних методів, деякі з яких можна віднести відразу до декількох груп, такі як статистичні алгоритми кластеризації, використання ансамблю кластеризаторів, алгоритми сімейства KRAB, алгоритми, засновані на методі просіювання, алгоритм DBSCAN

тощо. Неієрархічні алгоритми використовують різноманітні цільові функції. Всі методи спираються на «гіпотезу компактності» [10)]: в просторі об'єктів все близькі об'єкти повинні належати до одного кластеру, а все різні об'єкти відповідно повинні знаходитися в різних кластерах.

Розглянемо деякі конкретні методи кластеризації.

## 2.2 Ієрархічні методи кластеризації

Ієрархічні алгоритми кластерного аналізу [14)] розділяють на агломеративні й дивізимні.

Ієрархічні алгоритми пов'язані з побудовою дендрограм і поділяються на:

а) агломеративні (об’єднавчі), які вирізняються послідовним об'єднанням вихідних елементів і відповідним зменшенням числа кластерів (побудова кластерів знизу нагору);

б) дивізимні (ділильні), у яких число кластерів зростає починаючи з один, у результаті чого утворюється послідовність груп, що розщеплюють (побудова кластерів зверху вниз).

### 2.2.1 Агломеративні методи

В ієрархічних агломеративних алгоритмах кластеризації вихідна множина об'єктів $Q = \{q_j\}_{j=1}^{n}$ на першому етапі представляється як множина $n$ кластерів $\{c_p\}_{p=1}^{n}$.

Ієрархічна агломераційна кластеризація полягає в об'єднанні обраної пари кластерів $c_i, c_j$ в кластер $c_l$ на кожному кроці відповідно до виразу:

$$c_l = \arg\min_{i \neq j} r(c_i, c_j), \tag{2.1}$$

де $r(c_i, c_j)$ – міра близькості між кластерами попереднього кроку.

Таким чином, на першому кроці алгоритму маємо множину кластерів:

$$\{c_p\}_{p=1}^{n} = \{c_1 = \{q_1\}, c_2 = \{q_2\}, \dots, c_n = \{q_n\}\}. \tag{2.2}$$

На другому кроці алгоритму, використовуючи обрану міру близькості $r(c_i, c_j)$ знаходять кластери з найменшим віддаленням одне від одного й здійснюють злиття двох найбільше близьких одне до одного кластерів $c_i, c_j$ у спільний кластер $c_l$.

Нова множина, що складається вже з $(n-1)$ кластерів, буде містити:

$$\left\{c_1 = \{q_1\}, c_2 = \{q_2\}, \dots, c_l = \{q_i, q_j\}, \dots, c_n = \{q_n\}\right\}. \tag{2.3}$$

Процес пошуку кластерів з найменшим віддаленням і їх злиття повторюють. У результаті формуються множини кластерів, кількість яких зменшується на 1 на кожному кроці.

Наприкінці процедури залишиться кластер $c_g$, що складається з усіх об'єктів і співпадає з первісною множиною $Q$:

$$c_g = Q = \{q_j\}_{j=1}^{n}. \tag{2.4}$$

Таким чином, до об'єднання усіх початкових кластерів в один може бути $(n-1)$ кроків. З метою прискорення обчислень у процедурах ієрархічної класифікації застосовують систему порогів, яка дозволяє

поєднувати класи, міра близькості яких не перевершує значення порога $\sigma$. Процес об'єднання – формування агломеративних кластерів може бути в будь-який час зупинене при перевищенні порогу близькості $\sigma$.

Для визначення відстані між кластерами можна вибрати різні способи. Залежно від цього одержують алгоритми з різними властивостями.

Перерахування відстані між кластером $c_k$ і кластером $c_l = \{c_i, c_j\}$ виконують за формулою:

$$d_{k,l} = \alpha_i d_{i,l} + \alpha_j d_{j,l} + \beta d_{i,j} + \gamma \left| d_{i,l} - d_{j,l} \right|, \qquad (2.5)$$

де $d$ – міри близькості між відповідними кластерами;

$\alpha_i, \alpha_j, \beta, \gamma$ - вагові коефіцієнти.

Існує кілька методів перерахування відстаней з використанням старих значень відстаней для поєднуваних кластерів, що вирізняються коефіцієнтами у формулі (2.5).

Застосування того або іншого методу залежить від способу визначення відстані між кластерами.

Наприклад, поширеним є застосування наступних коефіцієнтів:

$$\alpha_i = \alpha_j = 0{,}5 \quad \beta = 0{,}25 \quad \gamma = 0. \qquad (2.6)$$

Зручно сформувати матрицю відмінності (відстаней або мір близькості) $D$ (1.10) для початкових кластерів, яку потрібно розширювати за формулою (2.6) на кожному кроці алгоритму.

Після кожного кроку в таблиці переоцінюються мінімальні відстані між кластерами з урахуванням вибуття кластерів, об'єднаних на попередньому кроці, і з урахуванням нового кластеру, створеного замість тих, які вибули.

Перераховані відстані між усіма кластерами, включаючи новий, додаються в матрицю відстаней для забезпечення розрахунку наступного кроку.

Розроблена серія критеріїв для оцінки однорідності сформованих нових кластерів, наприклад, з застосування методів статистики.

Одним з найбільш універсальних критеріїв є статистичне відхилення кластера $c_l$:

$$\rho_l(c_l) = \sqrt{\frac{1}{n_l - 1} \sum_{q_j \in c_l} \left\| x_j - \bar{x} \right\|^2}, \tag{2.7}$$

де $n_l$ – число елементів в кластері $c_l$;

$\left\| x_j - \bar{x} \right\|^2$ - квадрат евклідової відстані між елементами кластеру;

$\bar{x}$ – центр кластеру, який визначається m-вимірним вектором середніх значень точок даних (1.3) множини даних $X_l$ атрибутів об’єктів $q_j \in c_l$, і компоненти якого вираховуються аналогічно (1.4):

$$\bar{x} = \frac{1}{n_l} \sum_{q_j \in c_l} x_j. \tag{2.8}$$

Мірою подібності між кластерами $c_i, c_j$ може бути приріст статистичного відхилення при їхньому об'єднанні:

$$\rho(c_i, c_j) = \rho(c_i, c_j) - \rho(c_i) - \rho(c_j). \tag{2.9}$$

На основі критеріїв будуються функціонали якості розбивки, наприклад, кореляційне відношення, що застосовуються як умова завершення кластеризації.

Іншими критеріями завершення кластеризації може бути досягнення мінімального статистичного відхилення свого порогового значення

$$\min_{i \neq j} \rho(c_i, c_j) > \sigma \tag{2.10}$$

або досягнення заданої кількості кластерів $g$.

### 2.2.2 Дивізимні методи

Цей підхід звичайно застосовують, коли необхідно розділити всю множину об'єктів $Q$ на відносно невелику кількість кластерів.

У дивізимних алгоритмах вихідна множина на першому етапі представляється як єдиний кластер:

$$c_1 = Q = \{q_j\}_{j=1}^{n}. \tag{2.11}$$

Кількість об’єктів $n_1$ в кластері $c_1$ дорівнює кількості об’єктів в вихідній множині $n_1 = n$.

На другому кроці алгоритму вибирається об'єкт $q_i$ який найбільш віддалений від інших об'єктів у цьому кластері. Віддалення об'єкта $q_i$ визначається як найбільша середня відстань до інших об'єктів кластера й може розраховуватися за формулою:

$$d_i = \frac{1}{n_1}\sum_{j=1}^{n} d(x_i, x_j) \forall x_j \in c_1. \tag{2.12}$$

Формується новий кластер $c_2$. Обраний найбільш віддалений об'єкт $q_i$ видаляється із кластера $c_1$ і міститься в кластер $c_2$. На наступних кроках алгоритму об'єкти із кластера $c_1$, у яких різниця значень між середньою відстанню до об'єктів в $c_1$ і середньою відстанню до об'єктів в $c_2$ найбільша, переносяться в $c_2$. Перенесення об'єктів триває доти, поки різниці середніх відстаней не стануть негативними. У результаті виконання послідовності кроків формуються два кластери.

До одного зі сформованих кластерів застосовують розглянуту вище процедуру поділу. Таким чином, на кожному кроці алгоритму один з

існуючих кластерів рекурсивно ділиться на два дочірні. У такий спосіб ітераційно утворюються кластери зверху вниз.

Поділ кластерів проводиться доти, поки всі члени одного кластера не будуть відповідати вимозі близькості або всі кластери не будуть містити по одному об'єкту.

Вибір кластера для поділу може здійснюватися на основі оцінки діаметрів кластерів. Оцінка діаметра кластерів може виконуватися із застосуванням формули:

$$D_k = \max d(x_i, x_j) \, \forall x_j \in c_k, \tag{2.13}$$

де $k = \overline{1, g}$, $g$ – кількість наявних кластерів.

Таким чином, реалізація ієрархічної кластеризації залежить від вибраного алгоритму (агломеративний або дивізимний), вибраної міри близькості між кластерами та встановленого порогу.

## 2.3 Вибір моделі кластеризації часових рядів для реалізації програмного додатку

При кластеризації часових рядів суттєвим є врахування того, що медицинські часові ряди, які використовуються в даній роботі, можуть бути майже однакові, але один з них може бути незначно зміщений у часі (уздовж осі часу).

Тому доцільно використати метрику, що реалізована в алгоритмі динамічної трансформації часової шкали (DTW) (2.14)

$$DTW(X_i, X_j) = \min\left(\frac{\sum_{k=1}^{K} d(w_k)}{K}\right). \quad (2.14)$$

і в його модифікації soft-DTW [15] при різних значеннях параметра згладжування γ.

$$DTW_\gamma(X_i, X_j) = -\gamma \log \sum_{k=1}^{K} e^{\left(\frac{d(w_k)}{K \cdot \gamma}\right)}. \quad (2.15)$$

Для порівняння застосуємо також кластеризацію методом $k$-середніх (K-means) [15] з використання евклідової відстані (1.11)

$$d_{E1}(x_i, x_j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}. \quad (2.16)$$

між часовими рядами і з використанням відстані між «центрами ваги» груп часових рядів (1.24):

$$d_{ct}(x_i, x_j) = d(\overline{x_i}, \overline{x_j}). \quad (2.17)$$

Останній метод в літературі називається також методом DBA-k-means (DTW Barycenter Averaging).

# 3 КОМП'ЮТЕРНА МОДЕЛЬ КЛАСТЕРНОГО АНАЛІЗУ

## 3.1 Методи візуалізації кластерного аналізу

Результатом кластерного аналізу є набір кластерів, що містять елементи вихідної множини. Кластерна модель повинна описувати як самі кластери, так і приналежність кожного об'єкта до одному з них.

Дуже часто кластеризація виступає першим кроком при аналізі даних і її ієрархічна важливість, безсумнівно, важлива. При цьому дуже цінним допоміжним інструментом може служити візуалізація, особливо в разі багатовимірних даних. Така візуалізація дозволяє наочно побачити групи об'єктів, які потім можна об'єднати в кластери (cluster), в тому числі і візуально.

Іноді кластери достатньо відобразити у вигляді графа.

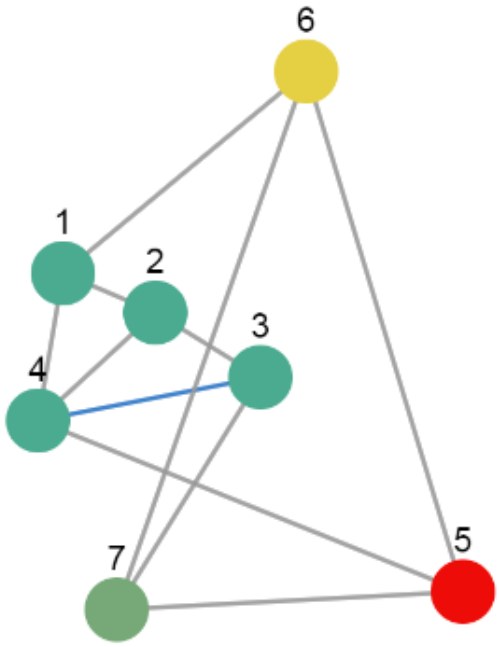



Рисунок 3.1 – Візуалізація кластерів у вигляді графа

У випадку невеликого числа об'єктів, що характеризуються двома змінними, результати можна зобразити за допомогою елементарних фігур

(трикутників, чотирикутників), відповідних до об'єктів, і множини прямих ліній.

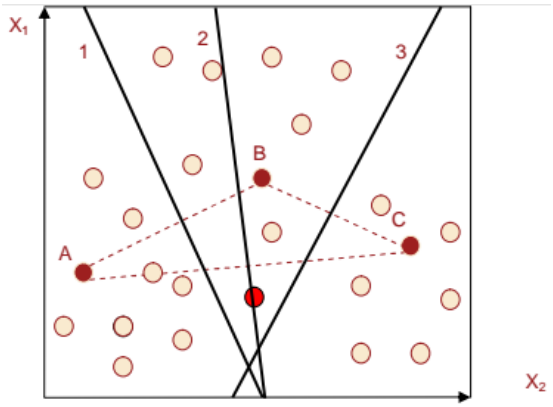



Рисунок 3.2 – Проста графічна візуалізація кластеризації

Візуалізація залежить також від вибраної кількості кластерів.

*k=3* *k=14*

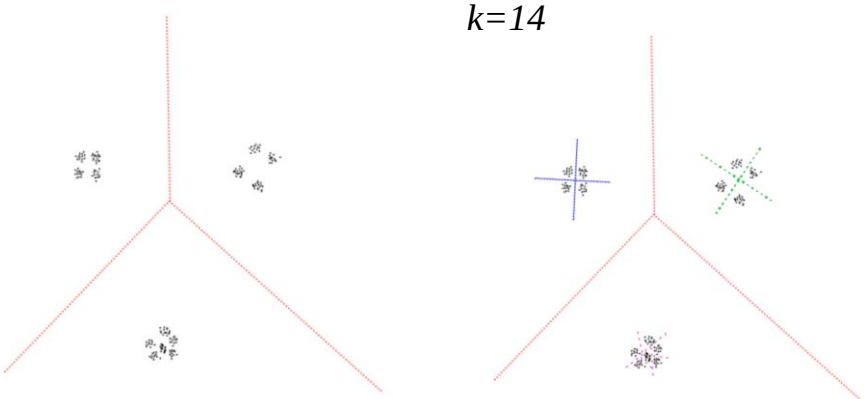


Рисунок 3.3 – Залежність візуалізації від вибраної кількості кластерів

Якщо кластери не можна розділити прямими лініями, то межі кластерів зображуються із застосуванням ламаних ліній. Приналежність об'єкта до декільком кластерів можна зобразити із застосуванням діаграм Венна.

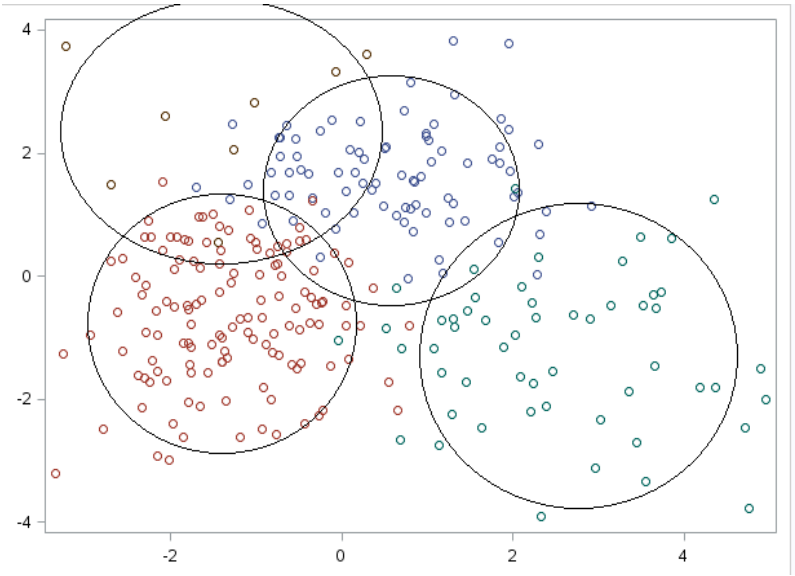



Рисунок 3.4 – Застосування діаграм Венна при кластеризації

Як при підготовці вихідних даних, так і при їх аналізі та оформленні результатів кластеризації можуть використовуватися карти самоорганізації (SOM або карти Кохонена) [17].

Ціль карт Кохонена - проектувати наявні дані в простір меншої розмірності зі збереженням відповідності відстані між наявними точками (об'єктами).

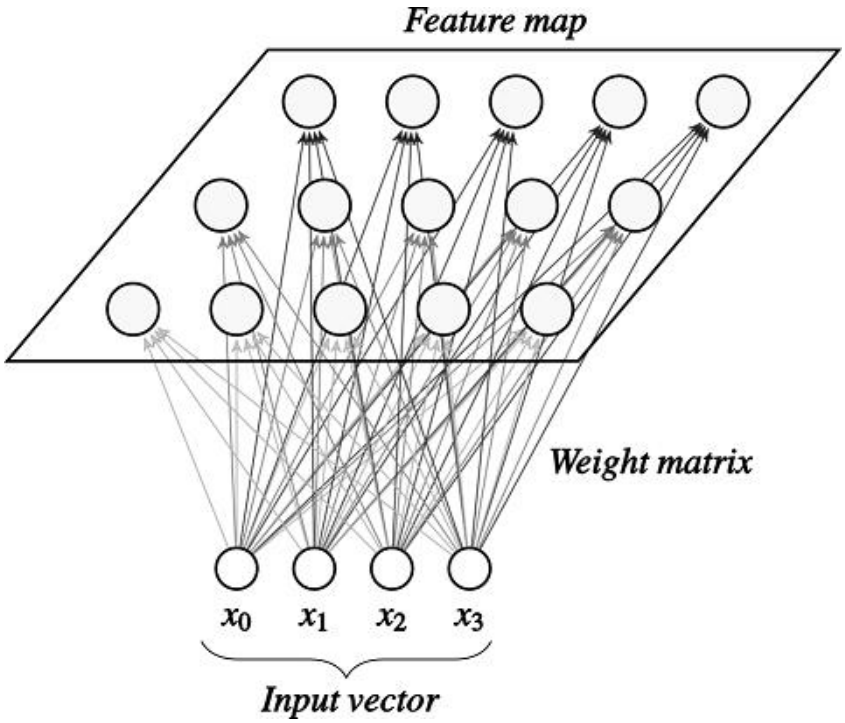



Рисунок 3.5 – Представлення карт самоорганізації (SOM)

Карти Кохонена роблять картографію від багатомірного входу на 1-вимірні або 2-вимірні ґрати вузлів з використанням ідеології нейронних мереж. Важливої особливістю цієї картографії є збереження топології, між вихідними даними й нейронами, на які вони проектуються. Ідея SOM заснована на організації функціонування головного мозку - активізація нейрона при сприйнятті інформації приводить до порушення ділянок у сусідніх частинах мозку.

У випадку нечіткої кластеризації приналежність об'єкта до кластера оцінюють імовірністю приналежності або ступенем приналежності. У цьому випадку результат можна представити у вигляді таблиці, у якій рядка відповідають об'єктам, стовпці – кластерам. У комірках таблиці вказується ймовірність або ступінь.

Таблиця 3.1 – Візуалізація нечіткої кластеризації таблицею

| Кластер | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0,7278 | 0,0157 | 0,0321 | 0,0428 | 0,1817 |
| 2 | 0,7078 | 0,0753 | 0,1399 | 0,0107 | 0,0664 |
| 3 | 0,7519 | 0,0188 | 0,0350 | 0,0525 | 0,1417 |
| 4 | 0,0694 | 0,6814 | 0,1464 | 0,0131 | 0,0896 |
| 5 | 0,0674 | 0,7204 | 0,0366 | 0,0628 | 0,1129 |
| 6 | 0,0247 | 0,8116 | 0,0984 | 0,0157 | 0,0496 |
| 7 | 0,0387 | 0,0228 | 0,5960 | 0,0753 | 0,2672 |
| 8 | 0,0142 | 0,0968 | 0,7596 | 0,0188 | 0,1106 |
| 9 | 0,0036 | 0,0242 | 0,8649 | 0,0814 | 0,0259 |
| 10 | 0,0173 | 0,1284 | 0,7134 | 0,0204 | 0,1204 |
| 11 | 0,0043 | 0,0321 | 0,0784 | 0,7116 | 0,1736 |
| 12 | 0,0308 | 0,1399 | 0,0323 | 0,5279 | 0,2692 |
| 13 | 0,0077 | 0,0350 | 0,0808 | 0,8682 | 0,0083 |
| 14 | 0,0038 | 0,1464 | 0,0852 | 0,7421 | 0,0225 |
| 15 | 0,0094 | 0,0366 | 0,2131 | 0,6284 | 0,1125 |
| 16 | 0,0428 | 0,1984 | 0,0320 | 0,7211 | 0,0057 |
| 17 | 0,0107 | 0,0496 | 0,3300 | 0,7399 | -0,1301 |
| 18 | 0,0053 | 0,2596 | 0,0242 | 0,0350 | 0,6760 |
| 19 | 0,0013 | 0,0649 | 0,0456 | 0,1464 | 0,7417 |
| 20 | 0,0081 | 0,0054 | 0,0753 | 0,1103 | 0,8009 |

Кінцевою метою будь-якої обробки даних є висування гіпотез про клас і структурі математичної моделі досліджуваного явища, визначення складу й

об'єму додаткових вимірів, вибір можливих методів наступної статистичної обробки й аналіз виконання основних передумов, що лежать у їхній основі.

Ряд алгоритмів кластеризації будують ієрархічні структури кластерів. У таких структурах самий верхній рівень відповідає всій множині об'єктів, тобто одному-єдиному кластеру.

На наступному рівні він ділиться на кілька підкластерів. Кожний з них ділиться ще на кілька тощо. Побудова такої ієрархії може відбуватися доти, поки кластери не будуть відповідати окремим об'єктам. Такі діаграми називаються дендрограмами (dendrograms). Цей термін підкреслює деревоподібну структуру діаграм (від грецького dendron — дерево).

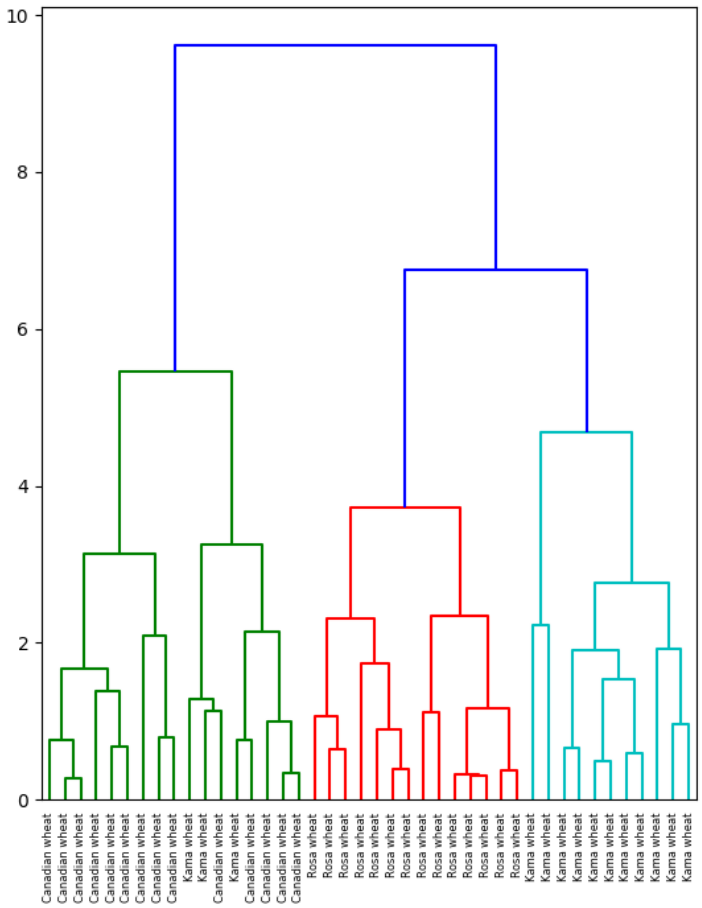



Рисунок 3.6 – Ієрархічна дендрограма для агломеративного алгоритму

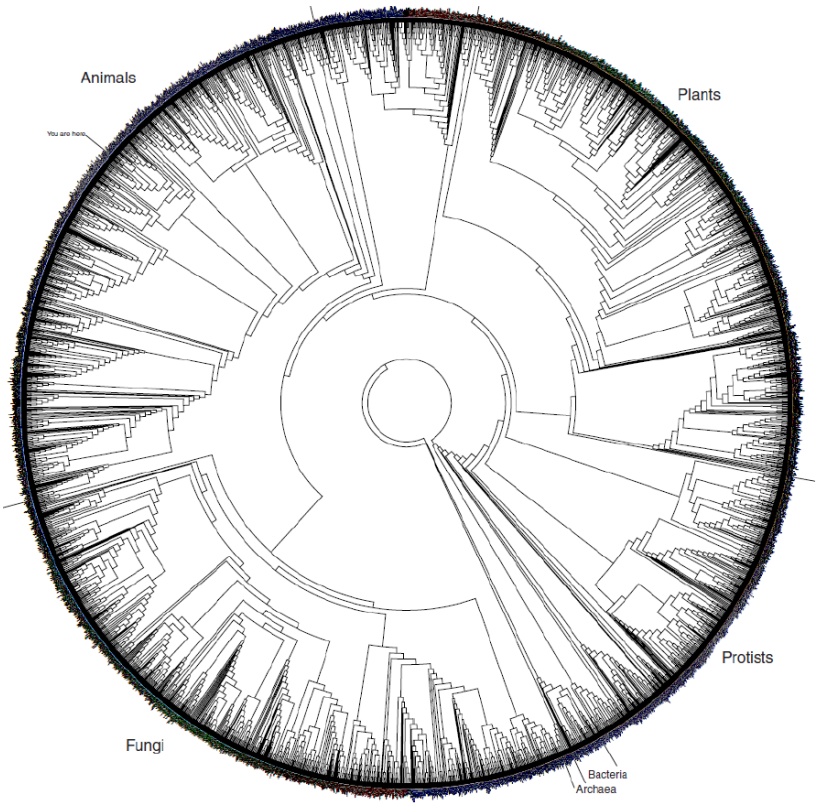



Рисунок 3.7 – Кругова ієрархічна дендрограма

Графічно процес групування в ієрархічній класифікації за агломеративним алгоритмом можна також відобразити організаційною діаграмою:

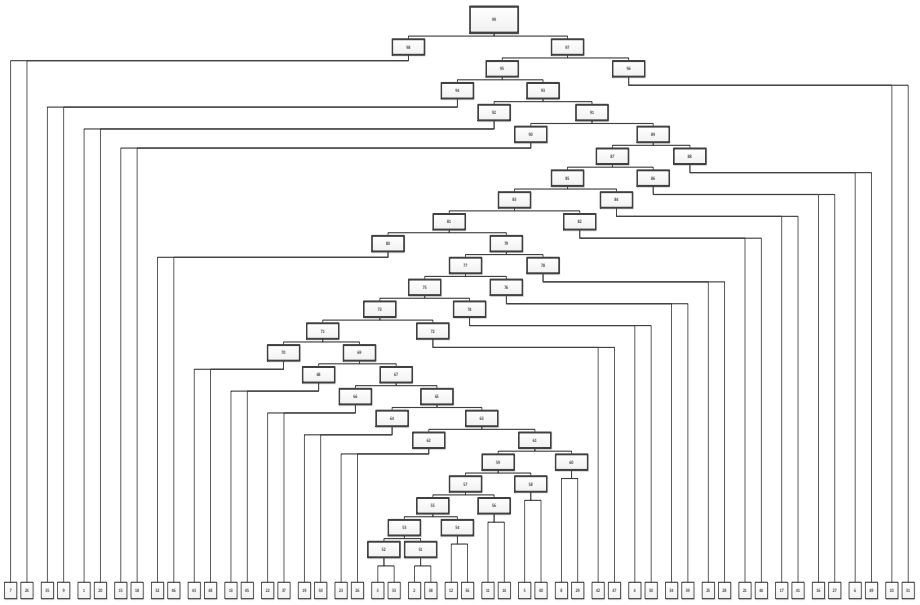


Рисунок 3.8 – Ієрархічна дендрограма у вигляді організаційної діаграми

Проблема візуалізації зростає при збільшенні вимірів векторів атрибутів об’єктів, які важко відобразити при кількості вимірів більше трьох.

Максимально можливим при графічній візуалізації є застосування тривимірних діаграм.

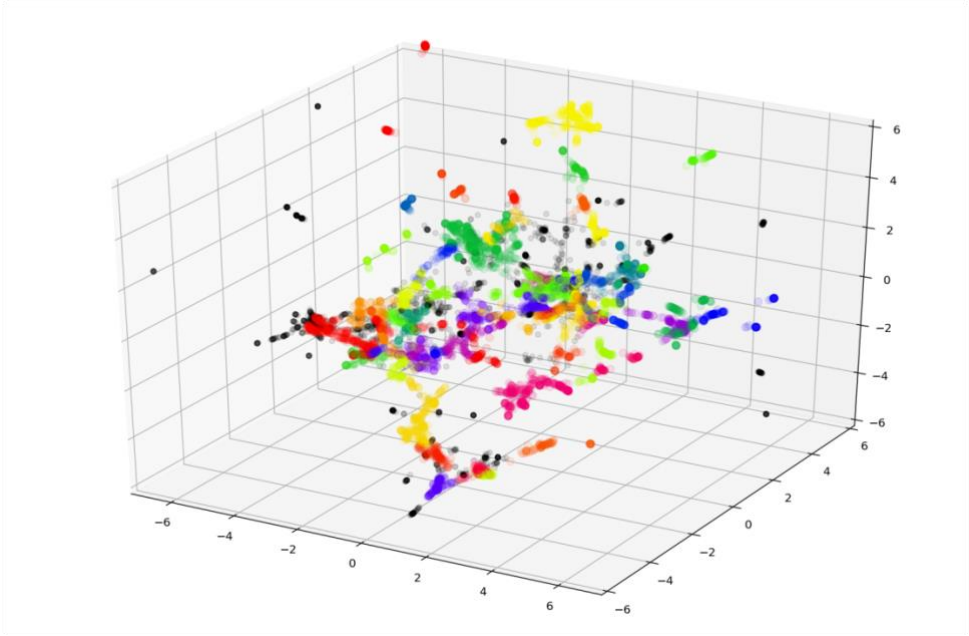



Рисунок 3.9 – Відображення кластеризації в тривимірному просторі

Візуалізація кластеризації часових рядів має свої особливості. Звичайно таку кластеризацію можна відобразити на графіках цих часових рядів.

Розглянемо, наприклад, графіки 100 часових рядів, розподілених по 4 кластерам. Конкретний приклад включає кластер 1 – 26 послідовностей, кластер 2 – 21 послідовність, кластер 3 – 22 послідовності, кластер 4 – 31 послідовность.

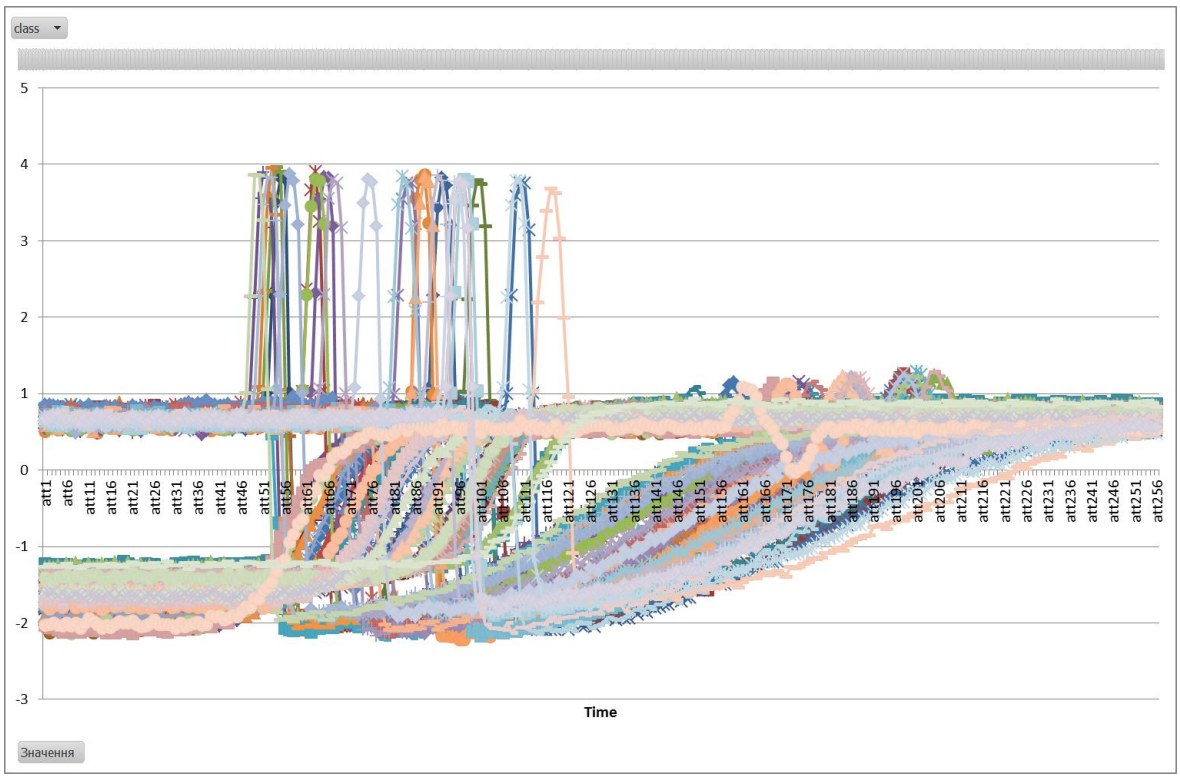



Рисунок 3.10 – Графік набору 100 часових рядів до проведення кластеризації

З графіків часових рядів видно, які особливості форми рядів сприяли їх відношенню до конкретних кластерів, а також як враховуються випадки, якщо два часових ряди однакові, але один з них незначно зміщений у часі (уздовж осі часу).

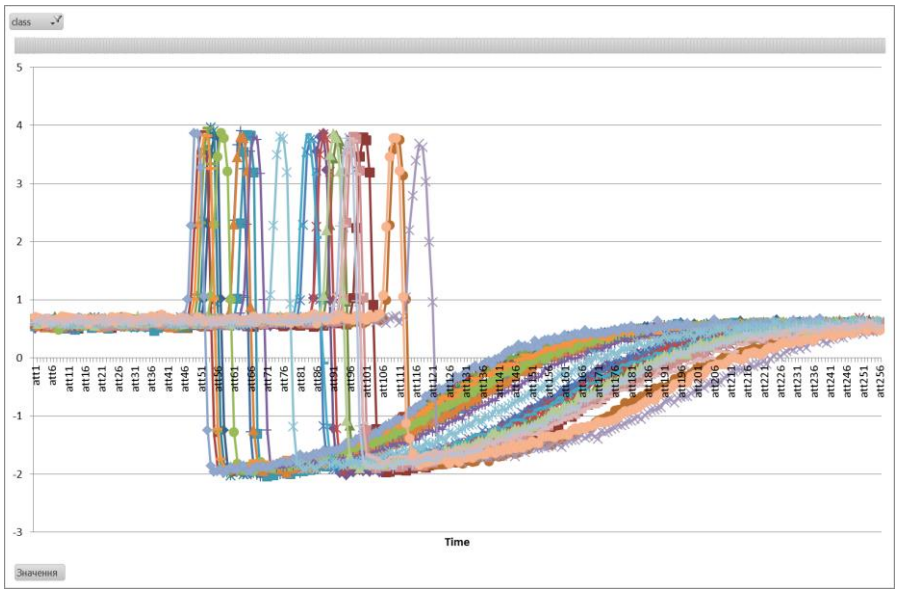



Рисунок 3.11 – Кластер №1 набору часових рядів

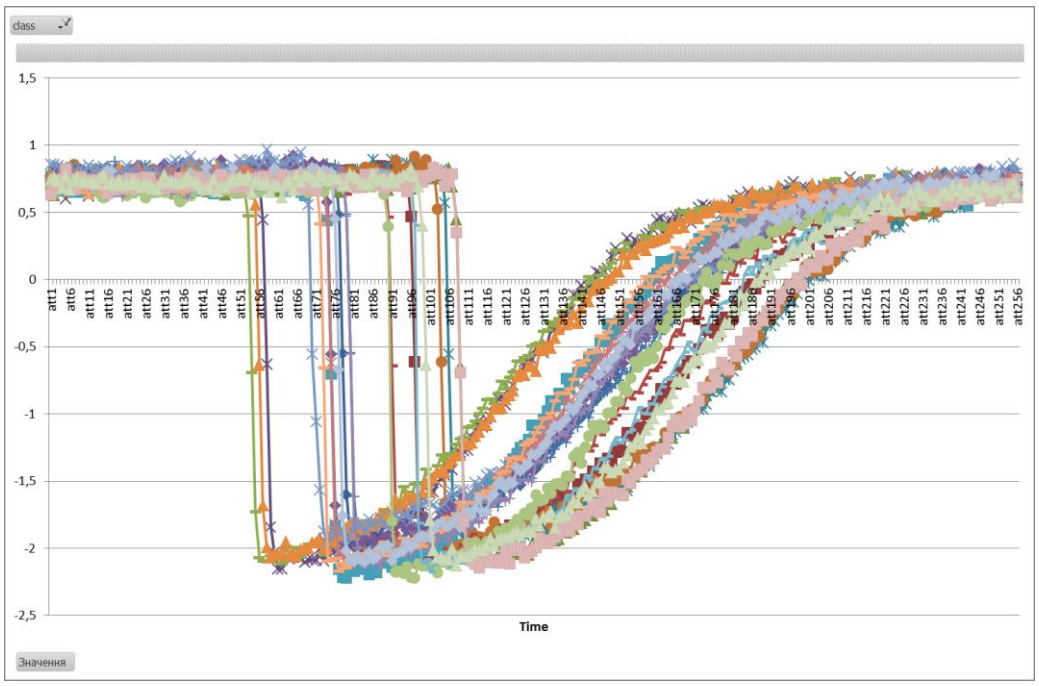



Рисунок 3.12 – Кластер №2 набору часових рядів

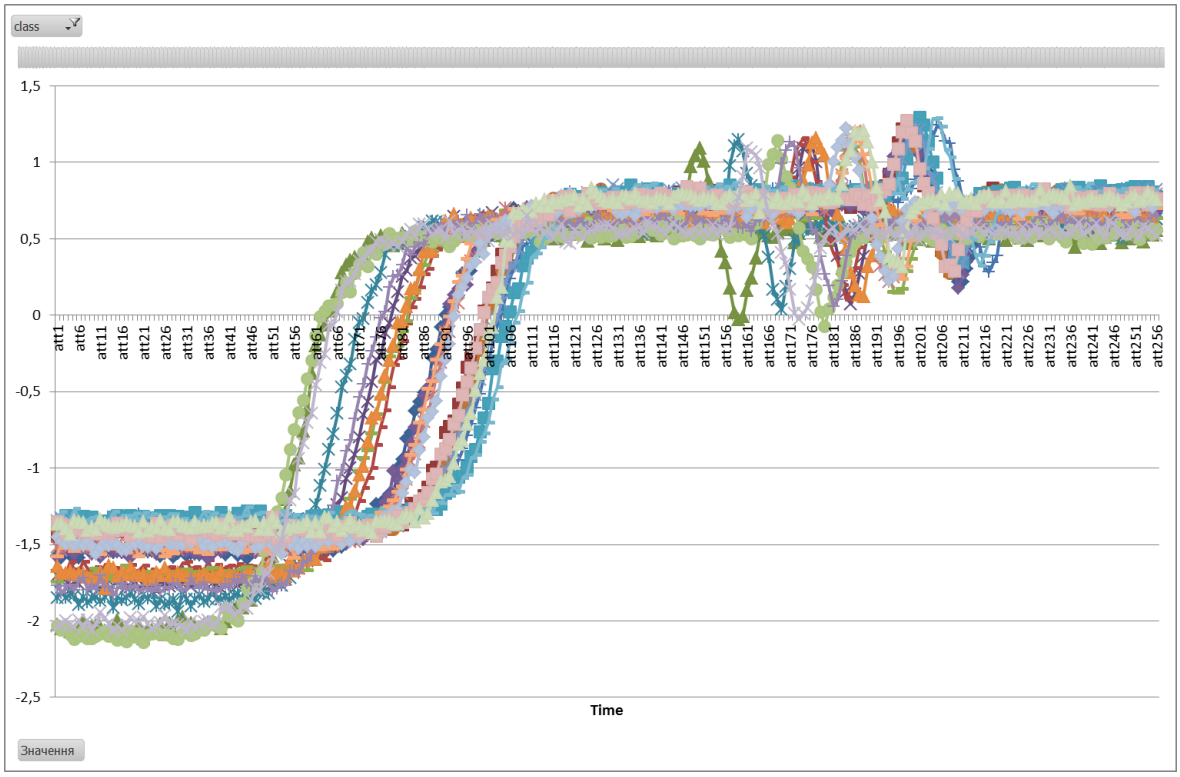



Рисунок 3.13 – Кластер №3 набору часових рядів

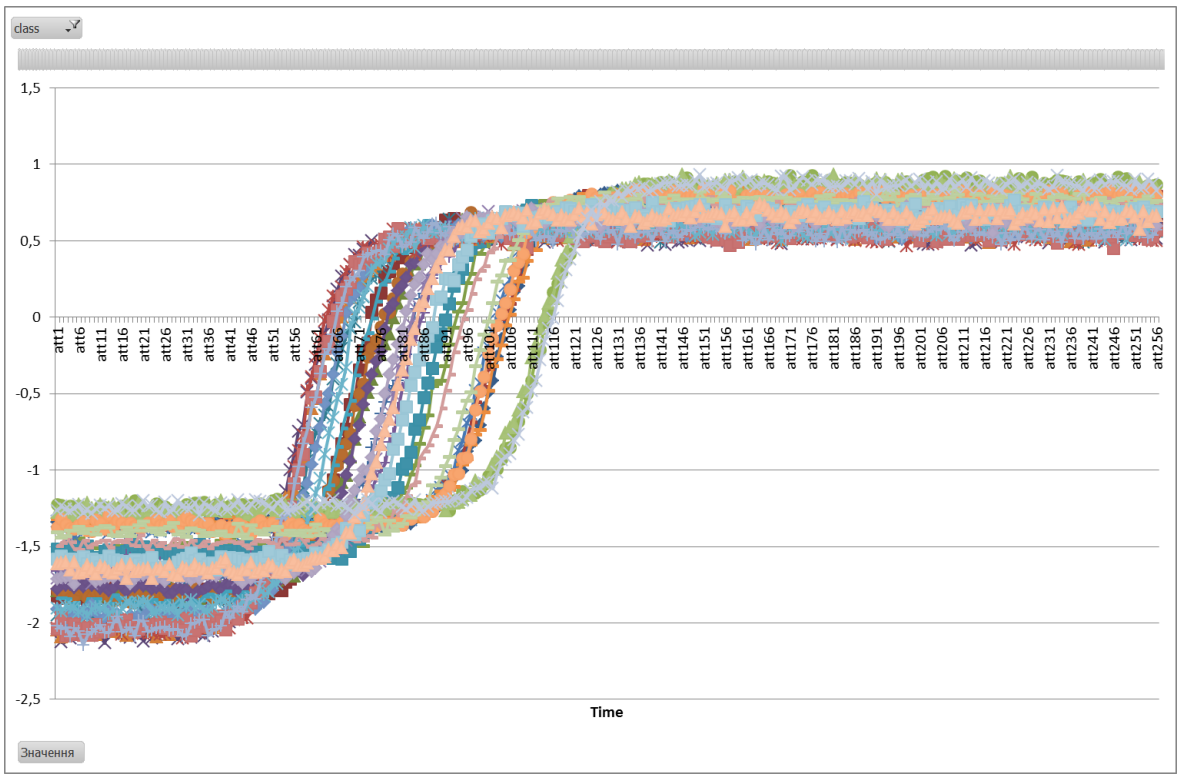



Рисунок 3.14 – Кластер №4 набору часових рядів

## 3.2 Характеристика вхідного набору даних часових рядів

Для реалізації кластерного аналізу часових рядів використаємо медицинський набір даних часових послідовностей електрокардіограм (ЕКГ) серцебиття.

Цей набір даних складається з колекції сигналів серцебиття, отриманих з бази даних діагностичних ЕКГ MIT-BIH Arrhythmia.

Сигнали відповідають формам електрокардіограми (ЕКГ) серцевих скорочень для нормального випадку і випадків ураження різними аритміями та інфарктом міокарда. Ці сигнали попередньо оброблялися і сегментувалисяся, причому кожен сегмент відповідає одному серцевому удару.

Числові характеристики набору даних:

– кількість зразків (часових послідовностей): 87554;

– кількість часових відліків: 187;

– кількість категорій (класів), визначених при формуванні бази даних ЕКГ: 5;

– частота дискретизації: 125 Гц;

– джерело даних: Фізионетичний MIT-BIH аритмічний набір даних;

– класи: [N: 0, S: 1, V: 2, F: 3, Q: 4].

В останній 4-й клас Q були включені всі часові послідовності, не віднесені до інших класів.

Цей набір даних складається з файлу CSV, який містить матрицю значень часових відліків, причому кожен рядок є прикладом часового ряду серцевого удару. Останній елемент кожного рядка позначає клас, визначений при формуванні бази даних ЕКГ, до якого належить цей часовий ряд.

Ця база даних описана в [19] і практично використовувалася при виявленні випадків інфаркту міокарда та для фундаментальних досліджень динаміки серця на більш ніж 500 сайтах по всьому світу.

Цей набір даних також використовувався при вивченні класифікації серцевих скорочень за допомогою глибоких архітектур нейронних мереж і спостереження за деякими можливостями навчання на ньому.

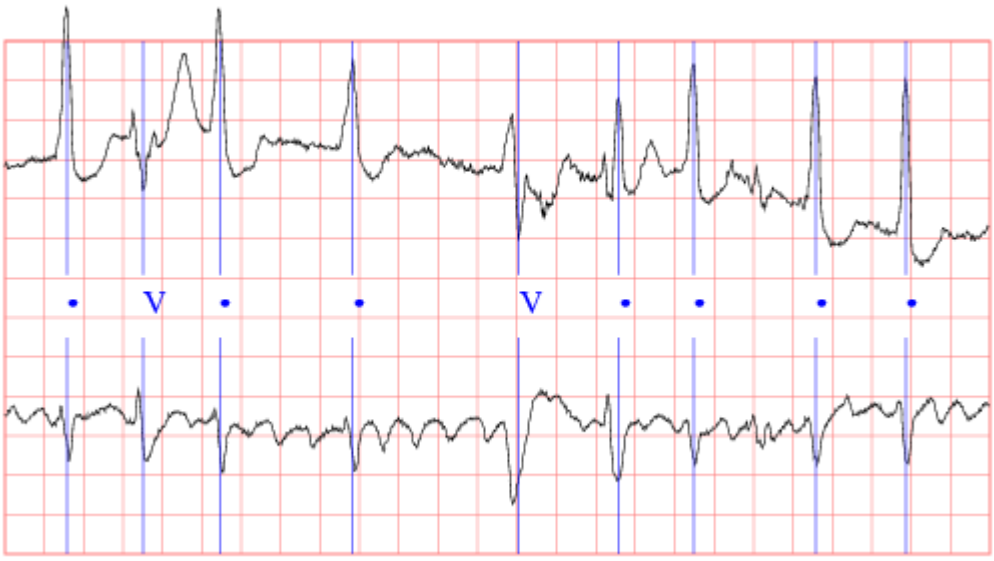



Рисунок 3.15 – Поділ часових послідовностей ЕКГ на сегменти

Сигнали ЕКГ збиралися, починаючи з 1975 року, лабораторією в Бостонській лікарні Beth Israel (медичний центр Beth Israel Deaconess).

Джерелом ЕКГ, включених до бази даних аритмії MIT-BIH, є набір з більш ніж 4000 довгострокових записів, які були отримані в Бостонській лікарні Beth Israel (медичний центр Beth Israel Deaconess) в період з 1975 по 1979 рік. Ці записи було отримано від змішаної популяції пацієнтів (близько 60%) та амбулаторних пацієнтів (близько 40%).

Пацієнтами були 25 чоловіків у віці від 32 до 89 років і 22 жінки у віці від 23 до 89 років.

Записи попередньо відбиралися, щоб

– включити різноманітність рідкісних, але клінічно важливих явищ;

– служити репрезентативним зразком різноманітних форм сигналів і артефактів, з якими може зіткнутися детектор аритмії при звичайному клінічному використанні;

– включити комплексні шлуночкові, вузлові та надшлуночкові аритмії та порушення провідності.

Деякі з цих записів були обрані тому, що особливості ритму, зміна морфології QRS або якість сигналу можуть спричинити значні труднощі для детекторів аритмії; ці записи отримали значну популярність серед користувачів бази даних.

Оригінальні аналогові записи були зроблені і відтворювалися з використанням дев'яти двоканальних аналогових магнітофонів, які для зменшення перешкод живилися від постійного струму.

Аналогові сигнали з виходу блоку відтворення були відфільтровані і нормалізовані для обмеження насичення аналого-цифрового перетворювача (АЦП) і для згладжування з використанням фільтру з смугою пропускання від 0,1 до 100 Гц відносно реального часу, що значно виходить за межі найнижчих і найвищих частот записів.

Записи були оцифровані при використанні частоти квантування 360 відліків в секунду на канал з 11-бітною роздільною здатністю в діапазоні ±10 мВ.

Сигнали, пропущені через фільтр, були оцифровані при використанні частоти дискретизації 360 відліків в секунду (360 Гц) на канал з 11-бітною роздільною здатністю в діапазоні ±10 мВ.

Частота дискретизації була обрана для полегшення реалізації цифрових фільтрів на 60 Гц (частоти мережі в США) в детекторах аритмії.

Для опису рядів серцевих ударів був визначений стандартний набір кодів анотацій для ЕКГ [20].

Кожен екземпляр анотації може мати до шести атрибутів:

– **time** - час в межах запису (записується у файлі анотації як номер зразка вибірки, до якого вказує анотація);

– **anntyp** - числовий код анотації;

– **subtyp**, **chan**, **num** - три малі цілі числа (від -128 до 127), які визначають атрибути, залежні від контексту;

– **aux** - вільний текстовий рядок.

Більшість медицинських баз даних, зокрема PhysioBank, використовують коди приміток (анотацій), зазначені в таблиці нижче:

Таблиця 3.1 – Коди анотацій серцевих ударів

| Код | Опис |
|---|---|
| 1 | 2 |
| N | Звичайний удар (відображається як " · " в базах даних PhysioBank, LightWAVE, pschart і psfd) |
| L | Білий блок гілок лівого пучка |
| R | Правий блок розгалужується |
| B | Переміщення блоку розгалужень пакета (не вказано) |
| A | Передсердя передчасно б'ють |
| a | Аберіальний передчасний збиток передсердя |
| J | Нодальний (сполучний) передчасний удар |
| S | Надшлуночкові передчасні або ектопічні удари (передсердні або вузлові) |
| V | Передчасне скорочення шлуночків |
| r | R-на-T передчасне скорочення шлуночків |
| F | Злиття шлуночків і нормального биття |
| i | Втеча з передсердя |

Продовження таблиці 3.2

| 1 | 2 |
|---|---|
| j | Nodal (junctional) перебіг удару |
| n | Надшлуночковий бічний потік (атріальний або вузловий) |
| i | Шлуночкова бійка |
| / | Темп збився |
| f | Злиття темпів і нормального удару |
| Q | Некласифікований удар |
| ? | Удар не класифікується під час навчання |

Анотації можна читати за допомогою додатків C, C++, Fortran і Python, що використовують функцію ***getann*** , і вони можуть бути записані за допомогою функцій ***putann*** , визначених у бібліотеці WFDB - waveform-database package.

Програми Matlab і Octave можуть читати і писати анотації за допомогою m-файлів. Також анотації можна читати програмами мови сценаріїв, що використовують функцію ***rdann*** , і вони можуть бути записані з використанням програм ***wrann*** , що належать до пакету програм WFDB.

Для уточнення інформації про окремі часові відліки використовуються додаткові коди анотацій і коди ритм-анотації:

Таблиця 3.2 – Коди додаткових анотацій серцевих ударів

| Код | Опис |
|---|---|
| [ | Початок шлуночкового тріпотіння / фібриляції |
| ! | Хвиля шлуночків |
| ] | Кінець шлуночкового тріпотіння / фібриляції |
| x | Непроведена P-хвиля (блокований АПК) |
| ( | Формування сигналу |
| ) | Кінець сигналу |
| p | Пік P-хвилі |
| t | Пік T-хвилі |
| u | Пік U-хвилі |
| ` | Перехрестя PQ |
| ' | J-точка |
| ^ | Артефакт кардіостимулятора |
| \| | Ізольований артефакт типу QRS |
| ~ | Зміна якості сигналу |
| + | Зміна ритму |
| s | Зміна сегмента ST |
| T | Зміна T-хвилі |
| * | Систола |

Продовження таблиці 3.2

| 1 | 2 |
|---|---|
| D | Діастола |
| = | Анотація вимірювань |
| " | Згідно до анотації |
| @ | Посилання на зовнішні дані |

Таблиця 3.3 – Коди ритм-анотацій

| Код | Опис |
|---|---|
| (AB | Передсердний бигемин |
| (AFIB | Миготлива аритмія |
| (AFL | Тріпотіння передсердь |
| (B | Шлуночкові великиміни |
| (BII | 2 серцевий блок |
| (IVR | Ідіовентрикулярний ритм |
| (N | Нормальний синусовий ритм |
| (NOD | Нодальний (AV-сполучний) ритм |
| (P | Темп ритму |
| (PREX | Попереднє збудження (WPW) |
| (SBR | Синусова брадикардія |
| (SVTA | Надшлуночкова тахіаритмія |
| (T | Шлуночкові тригемины |
| (VFL | Шлуночкові тріпотіння |
| (VT | Шлуночкова тахікардія |

Поле ***aux*** анотації звичайно містить URL-посилання (уніфікований локатор ресурсу, у формі ***http://machine.name/some/data*** ), придатний для переходу до веб-браузера. Якщо це можливо, текст посилання відображається підкресленим і синім кольором.

Анотації посилань можна використовувати для поєднання розширеного тексту, зображень або інших даних з файлом анотацій.

## 3.3 Обґрунтування вибору середовища програмної реалізації

У рамках дипломної роботи для реалізації програмного засобу кластеризації часових рядів була обрана мова Python.

Python - інтерпретована об'єктно-орієнтована мова програмування високого рівня з динамічною семантикою. Python підтримує модулі та пакети модулів, що сприяє модульності та повторному використанню коду. Серед основних його переваг можна назвати наступні:

– чистий синтаксис (для виділення блоків слід використовувати відступи); переносимість програм (що властиво більшості інтерпретованих мов);

– стандартний дистрибутив має велику кількість корисних модулів (включаючи модулем для розробки графічного інтерфейсу);

– можливість використання Python в діалоговому режимі (дуже корисно для експериментування та рішення простих задач);

– стандартний дистрибутив має просте, але разом з тим досить потужне середовище розробки, яка називається IDLE і яке написано на мові Python;

– зручний для вирішення математичних проблем (має засоби роботи з комплексними числами, може оперувати з цілими числами довільної величини, в діалоговому режимі може використовуватися як потужний калькулятор).

При цьому, на відміну від багатьох портованих систем, для всіх основних платформ Python має підтримку характерних для даної платформи технологій. В середньому програма, написана на Python, в 2-4 рази компактніше, ніж її аналог на C ++ або Java. Збереження байт-коду (файли .pyc і .pyo) дозволяє інтерпретатору не витрачати зайвий час на перекомпіляцію коду модулів при кожному запуску, на відміну, наприклад, від мови Perl.

Для обробки вхідних даних було обрано бібліотеку numpy.

NumPy є основним пакетом для наукових обчислень з Python. Він містить, серед іншого:

– модуль обробки потужних об'єктів N-розмірних масивів;

– спеціалізовані функції;

– інструменти для інтеграції C / C ++ і коду Fortran;

– корисні функції лінійної алгебри, перетворення Фур'є і можливості обробки статистичних послідовностей.

Крім очевидного наукового використання, NumPy також може використовуватися як ефективний багатовимірний контейнер загальних даних. Можуть бути визначені довільні типи даних. Це дозволяє NumPy легко і швидко інтегруватися з широким спектром баз даних.

NumPy ліцензується за ліцензією BSD, дозволяючи повторне використання з невеликими обмеженнями.

Для реалізації алгоритмів кластеризації вхідних даних було обрано бібліотеку tslearn.

tslearn є пакетом Python, який надає засоби машинного навчання для аналізу часових рядів.

Для реалізації інтерфейсу було обрано бібліотеку matplotlib.

matplotlib — бібліотека на мові програмування Python для візуалізації даних двовимірною 2D графікою (3D графіка також підтримується). Отримувані зображення можуть бути використані як ілюстрації в публікаціях.

matplotlib є гнучким, легко конфігурованим пакетом, який разом з NumPy, SciPy і IPython надає можливості, подібні до MATLAB. В даний час пакет працює з декількома графічними бібліотеками, включаючи wxWindows і PyGTK.

Пакет підтримує багато видів графіків і діаграм:

– графіки (line plot);

– діаграми розсіювання (scatter plot);

– стовпчасті діаграми (bar chart) і гістограми (histogram);

– секторні діаграми (pie chart);

– діаграми «Стовбур-листя» (stem plot);

– контурні графіки (contour plot);

– поля градієнтів (quiver);

– спектральні діаграми (spectrogram).

Користувач може вказати осі координат, сітку, додати підписи і пояснення, використовувати логарифмічну шкалу або полярні координати.

Для використання matplotlib необхідно створити об’єкт фігури та визначити його властивості (заголовок, підписи, осі координат, тип графіку, тощо).

### 3.4 Програмна реалізація

Програмний засіб реалізований у вигляді застосунку Python, який збережений під назвою cluster.py.

Застосунок приймає вхідні параметри:

– **input** ІМ’Я ФАЙЛУ;

– **n** – цільова кількість кластерів;

– **dba** – кластеризація методом DBA-k-means (DTW Barycenter Averaging);

– **softdtw** – кластеризація методом soft-DTW;

– **euclidean** – кластеризація методом $k$-середніх з використанням евклідової відстані;

– **gamma** – параметр згладжування при використанні методу soft-DTW.

Вхідним файлом може бути файл CSV (англ. comma-separated values) - файловий формат, котрий є відмежовувальним форматом для представлення табличних даних, у якому поля відокремлюються символом коми та переходу на новий рядок. Структура файлу приймається стандартною для медицинських баз даних часових рядів.

Результати кластеризації візуалізуються у вигляді графіків часових рядів, що віднесені до відповідних кластерів, і у вигляді переліку номерів відповідних часових рядів у кожному кластері.

### 3.5 Інструкція користувача

Для виконання кластеризації необхідно запустити будь-яким способом, наприклад з командного рядка, програмний застосунок cluster.py і зчитати результати проведеної кластеризації. Лінія кластер-центру виділяється червоним кольором.

Результати роботи програми можна зберегти у вигляді графічних файлів або таблиць номерів відповідних часових рядів у кожному кластері стандартними засобами Windows.

### 3.6 Тестування розробленого програмного засобу

Програмний засіб тестувався на вхідному наборі даних медицинських часових рядів, що складається з колекції сигналів серцебиття, отриманих з бази даних діагностичних ЕКГ, і який був описаний в розділі 3.2.

Набір з 87554 часових послідовностей розбивався на 4 кластери.

Результати роботи програми відображені на рисунках:

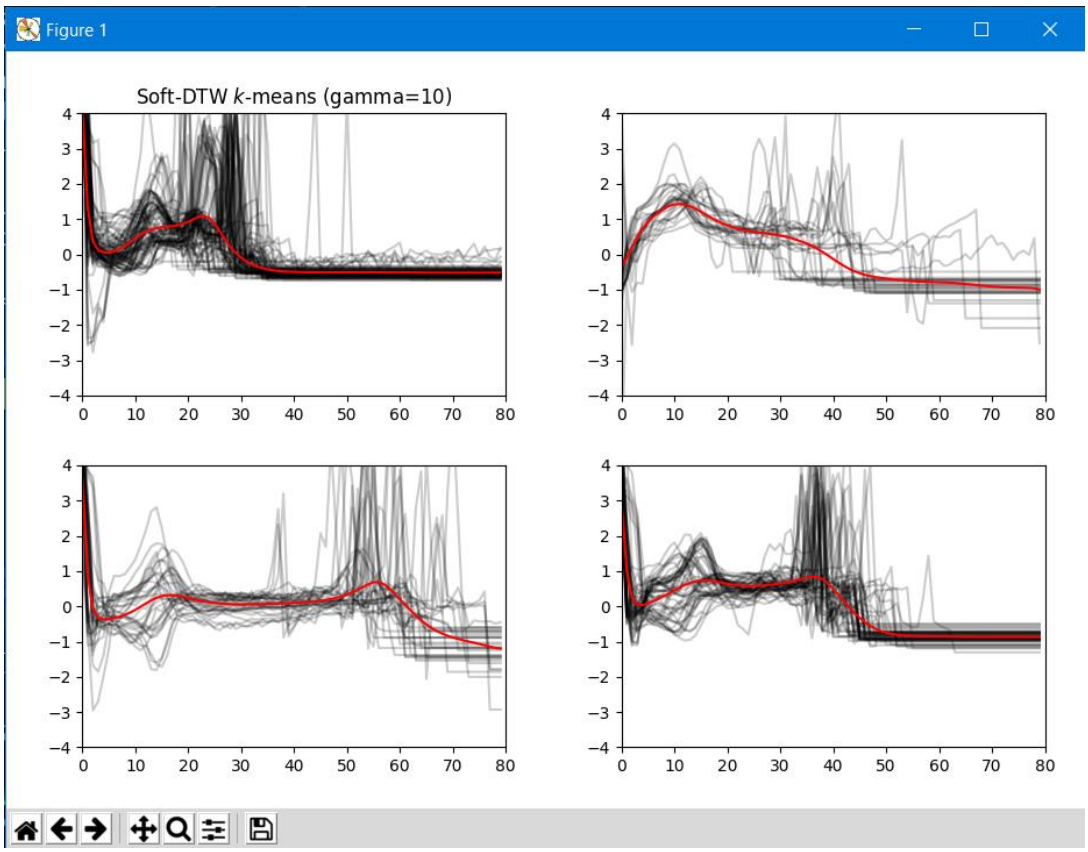



Рисунок 3.16 – Результати кластеризації методом soft-DTW при γ=10

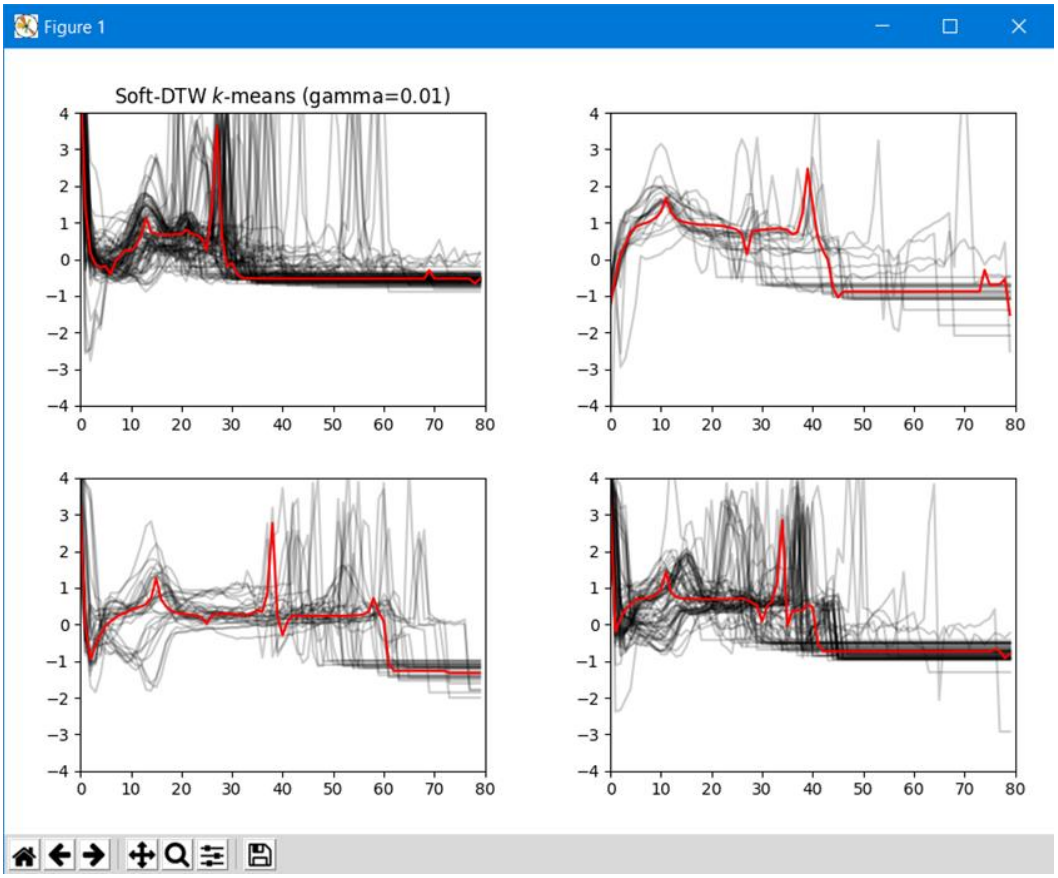



Рисунок 3.17 – Результати кластеризації методом soft-DTW при γ=0,01

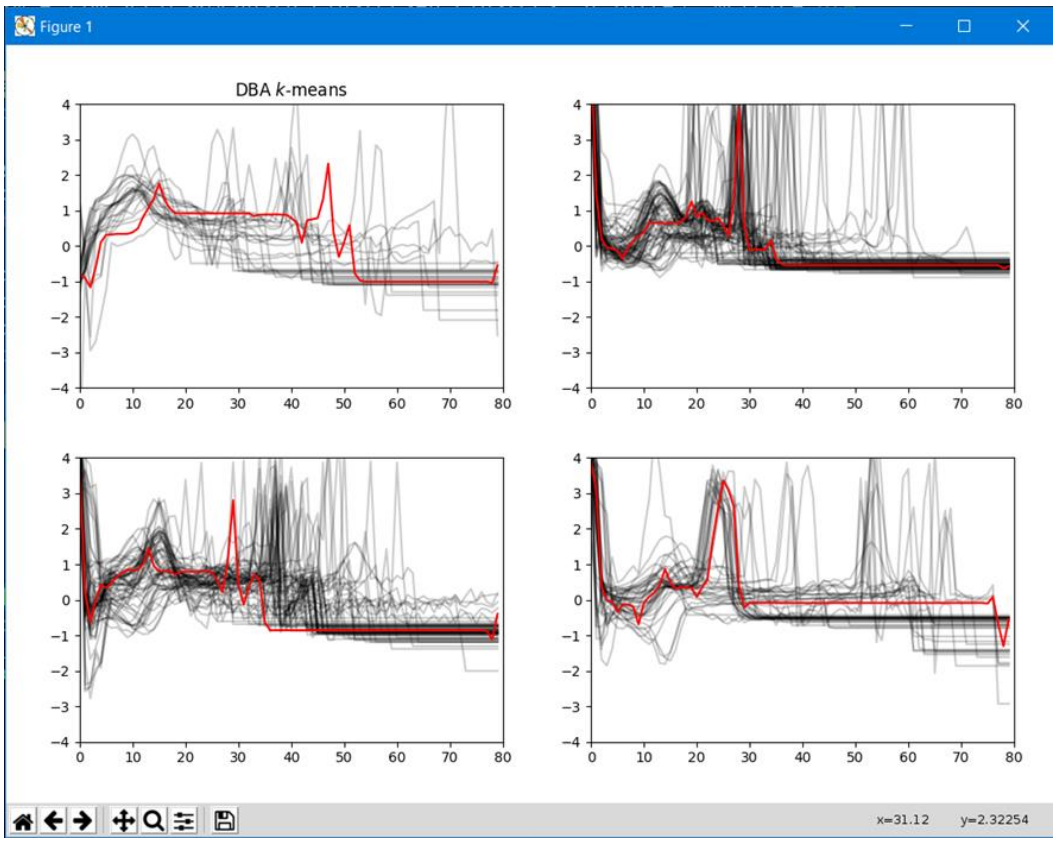



Рисунок 3.18 – Результати кластеризації методом DBA-k-means

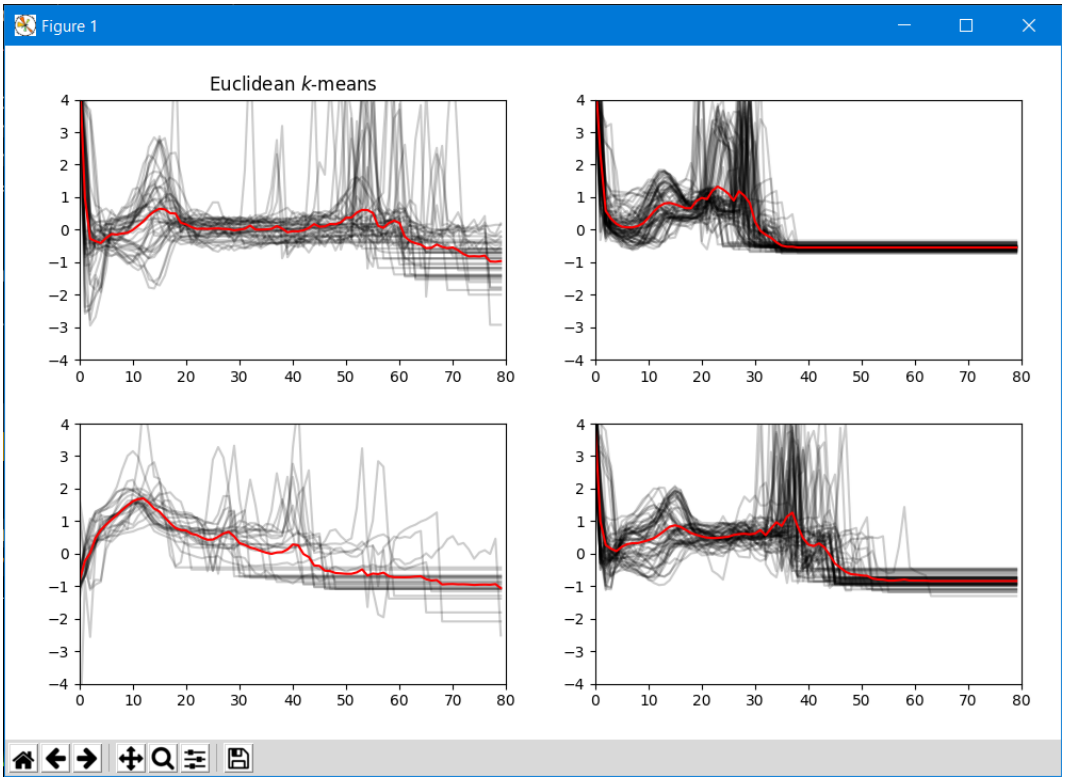



Рисунок 3.19 – Результати кластеризації методом k-means з використання евклідової відстані

Для порівняння результатів кластеризації різними методами зведемо результати співпадіння цих результатів з даними бази даних в таблицю для вхідних послідовностей, які на етапі підготовки бази даних MIT-BIH Arrhythmia були віднесені до класів [N: 0, S: 1, V: 2, F: 3]. Клас [Q: 4] був виключений з порівняння згідно його визначення в табл.3.4.

Таблиця 3.4 – Порівняння методів кластеризації

| Метод | Кластер N | Кластер S | Кластер V | Кластер F | Загальний % співпадіння |
|---|---|---|---|---|---|
| soft-DTW при γ=10 | 39,62% | 35,85% | 52,00% | 52,22% | 47,50% |
| soft-DTW при γ=0,01 | 30,00% | 34,92% | 59,09% | 43,53% | 44,00% |
| DBA-k-means | 43,59% | 37,70% | 60,87% | 53,25% | 47,50% |
| Euclidean k-means | 41,67% | 31,58% | 62,50% | 58,97% | 49,00% |

З таблиці видно, що метод кластеризації часових рядів для даної вибірки вхідних даних дають приблизно однаковий результат, тому що медична часова послідовність попередньо ділилася на сегменти до одного серцевого удару. Це дало змогу нівелювати незначні відхилення відліків по осі часу.

# 4 ОХОРОНА ПРАЦІ

## 4.1 Аналіз потенційних небезпечних і шкідливих виробничих чинників проектованого об’єкту, що мають вплив на персонал

У даному дипломному проекті розробляється програмне забезпечення.

Розроблене програмне забезпечення орієнтоване на роботу з персональним комп'ютером. Експлуатовані для вирішення внутрішньовиробничих завдань ПЕОМ типу IBM PC мають наступні характеристики:

| | |
|---|---|
| споживана потужність | 220 Вт; |
| робоча напруга | 220 В; |
| напруга джерел живлення | +12 В; - 12 В; +5 В; |
| робоча частота | 50 Гц. |

Виходячи з приведених характеристик, вочевидь, що для людини існує небезпека поразки електричним струмом, унаслідок недбалого поводження з комп'ютером і порушення правил експлуатації, залишення частин ПЕОМ, що знаходяться під напругою, відкритими або знятих для ремонту вузлів.

Відповідно до ДСН 3.3.6.042-99 [21] до легкої фізичної роботи відносяться всі види діяльності, виконувані сидячи і ті, що не потребують фізичної напруги. Робота користувача ПК відноситься до категорії 1а.

При роботі на ПЕОМ користувач піддається ряду потенційних небезпек. Унаслідок недотримання правил техніки безпеки при роботі з машиною (невиконання огляду відкритих частин ПЕОМ, що знаходяться під напругою або знятих для ремонту вузлів) для користувача існує небезпека поразки електричним струмом.

Джерелами підвищеної небезпеки можуть служити наступні елементи:

– розподільний щит;

– джерела живлення;

– блоки ПЕОМ і друку, що знаходяться в ремонті.

Ще одна проблема полягає у тому, що спектр випромінювання комп'ютерного монітора включає рентгенівську, ультрафіолетову і інфрачервону області, а також широкий діапазон хвиль інших частот. Небезпека рентгенівського проміння мала, оскільки цей вид випромінювання поглинається речовиною екрану. Проте велику увагу слід приділяти біологічним ефектам низькочастотних електромагнітних полів (аж до порушення ДНК).

Відповідно до НПАОП 0.00-7.15-18 [22], при обслуговуванні ПЕОМ мають місце фізичні і психофізичні небезпечні, а також шкідливі виробничі чинники:

– підвищене значення напруги в електричному ланцюзі, замикання якої може відбутися через тіло людини;

– підвищений рівень статичної електрики;

– підвищений рівень електромагнітних випромінювань;

– підвищена або знижена температура повітря робочої зони;

– підвищений або знижений рух повітря;

– підвищена або знижена вологість повітря;

– відсутність або недостатність природного світла;

– підвищена пульсація світлового потоку;

– недостатня освітленість робочого місця;

– підвищений рівень шуму на робочому місці;

– розумове перенапруження;

– емоційні навантаження;

– монотонність праці.

## 4.2 Заходи щодо техніки безпеки

Основним небезпечним чинником при роботі з ЕОМ є небезпека поразки людини електричним струмом, яка посилюється тим, що органи чуття людини не можуть на відстані знайти наявності електричної напруги на устаткуванні.

Проходячи через тіло людини, електричний струм чинить на нього складну дію, що є сукупністю термічної (нагрів тканин і біологічних середовищ), електролітичної (розкладання крові і плазми) і біологічної (роздратування і збудження нервових волокон і інших органів тканин організму) дій.

Тяжкість поразки людини електричним струмом залежить від цілого ряду чинників:

– значення сили струму;

– електричного опору тіла людини і тривалості протікання через нього струму;

– роду і частоти струму;

– індивідуальних властивостей людини і навколишнього середовища.

Розроблений дипломний проект передбачає наступні технічні способи і засоби, що застерігають людину від ураження електричним струмом:

– заземлення електроустановок;

– занулення;

– захисне відключення;

– електричне розділення мережі;

– використовування малої напруги;

– ізоляція частин, що проводять струм;

– огорожа електроустановок.

Занулення зменшує напругу дотику і обмежує години, протягом яких людина, ткнувшись до корпусу, може потрапити під дію напруги.

Струм однофазного короткого замикання визначається по наближеній формулі:

$$I_K = \frac{\text{Uф}}{\text{Zп} + \frac{\text{Zт}}{3}}\ , \qquad (4.1)$$

де Uф - номінальна фазна напруга мережі, В;

Zп - повний опір петлі, створене фазними і нульовими дротами, Ом;

$Z_T$ - повний опір струму короткого замикання на корпус, Ом.

Згідно таблиці 4 [23]: $Z_T/3 = 0{,}1$ Ом.

Для провідників і жил кабелю для розрахунку повного опору петлі використовуємо формулу(4.2.) :

$$Z\text{п} = \sqrt{\text{Rп}^2 + \text{Xп}^2}\ , \qquad (4.2)$$

де Rп = Rф + Ro - сумарний активний опір фазного Rф і нульового Ro дротів, Ом;

Xп - індуктивний опір паяння дротів, Ом.

Перетин 1 км мідного дроту S = 2.5 мм, тоді згідно таблицям 5 і 6 [23], має такий опір:

Xп = 0,11 Ом;

Rф = 7,55 Ом;

Ro = 7,55 Ом.

Отже, Rп = 7,55 + 7,55 = 15,1 Ом.

Тоді по формулі (4.2) знаходимо повний опір петлі :

$$Zп = \sqrt{15{,}1^2 + 0{,}11^2} \approx 15{,}1 \text{ (Ом).}$$

Струм однофазного короткого замикання рівний:

$$Iк = \frac{220}{15{,}1 + 0{,}1} = 14{,}47 \text{ (А).}$$

Дія плавкої вставки на ПЕОМ забезпечується, якщо виконується співвідношення:

$$Iк \geq k * Iн, \quad (4.3)$$

де Iн - номінальний струм спрацьовування плавкої вставки, А;

k - коефіцієнт кратності нелінійного струму Iн, А.

Коефіцієнт кратності нелінійного струму Iн розраховується по формулі (4.4.) :

$$Iн = P / U, \quad (4.4)$$

де P = 220 Вт - споживана потужність;

U = 220 В - робоча напруга;

k = 3 А - для плавких вставок.

Отже, Iн = 220 / 220 = 1 А.

Підставивши значення у вираз (4.3), одержимо:

14,47> 3*1.

Таким чином, доведено, що апарат забезпечить спрацьовування(і захист) при підвищенні номінального струму.

## 4.3 Заходи, що забезпечують виробничу санітарію і гігієну праці

Вимоги до виробничих приміщень встановлюються ДСН 3.3.6.042-99[21], ДБН, відповідними ГОСТами і ОСТамі з урахуванням небезпечних і шкідливих чинників, що утворюються в процесі експлуатації електроустаткування.

Підвищення працездатності людини і збереження її здоров'я забезпечується стабільними метеорологічними умовами.

Мікроклімат виробничих приміщень визначається діючими на організм людини поєднаннями температури, вологості і швидкості руху повітря, а також температури навколишніх поверхонь. Значне коливання параметрів мікроклімату приводити до порушення систем кровообігу, нервової і потовидільної, що може викликати підвищення або пониження температури тіла, слабкість, запаморочення і навіть непритомність.

Відповідно до ДСН 3.3.6.042-99 [21] встановлюють оптимальну і допустиму температуру, відносну вологість і швидкість руху повітря в робочій зоні . За відсутності надмірного тепла, вологи, шкідливих речовин в приміщенні досить природної вентиляції.

У приміщенні для виконання робіт операторського типу(категорія 1а), пов'язаних з нервово-емоційною напругою, проектом передбачається дотримання наступних нормованих величин параметрів мікроклімату (табл. 4.1).

Таблиця 4.1 - Санітарні норми мікроклімату робочої зони приміщень для робіт категорії 1а.

| Пора року | Температура, С | Відносна вологість, % | Швидкість руху повітря, м/с |
|---|---|---|---|
| Холодна | 22…24 | 40…60 | 0,1 |
| Тепло | 23…25 | 40…60 | 0,1 |

У приміщенні, де знаходиться ПЕОМ, повітрообмін реалізується за допомогою природної організованої вентиляції (з пристроєм вентиляційних каналів в перекриттях будівлі і вертикальних шахт) й установленого промислового кондиціонера фірми Mitsubishi, який дозволяє вирішити переважну більшість завдань по створінню та підтримці необхідних параметрів повітряного середовища. Цей метод забезпечує приток потрібної кількості свіжого повітря, визначеного в ДБН (30 $м^3$ в годину на одного працівника).

Шум на виробництві має шкідливу дію на організм людини. Стомлення операторів через шум збільшує число помилок при роботі, призводить до виникнення травм. Для оператора ПЕОМ джерелом шуму є робота принтера. Щоб усунути це джерело шуму, використовують наступні методи. При покупці принтера слід вибирати найбільш шумозахисні матричні принтери або з великою швидкістю роботи(струменеві, лазерні). Рекомендується принтер поміщати в найбільш віддалене місце від персоналу, або застосувати звукоізоляцію та звукопоглинання(під принтер підкладають демпфуючі підкладки з пористих звукопоглинальних матеріалів з листів тонкої повсті, поролону, пеноплену).

При роботі на ПЕОМ, проектом передбачені наступні методи захисту від електромагнітного випромінювання : обмеження часом, відстанню, властивостями екрану.

Обмеження годині роботи на ПЕОМ складає 3,5-4,5 години. Захист відстанню передбачає розміщення монітора на відстані 0,4-0,5 м від

оператора. Передбачений монітор 20" TFT, Samsung 2043BW відповідає вимогам стандарту TCO'03.

TCO'03 пред’являє жорсткі вимоги в таких областях: ергономіка (фізична, візуальна і зручність користування), енергія, випромінювання (електричних і магнітних полів), навколишнє середовище і екологія, а також пожежна та електрична безпека, які відповідають всім вимогам [24].

Для зниження стомлюваності та підвищення продуктивності праці обслуговуючого персоналу в колірній композиції інтер'єру приміщень для ПЕОМ дипломним проектом пропонується використовувати спокійні колірні поєднання і покриття, що не дають відблисків.

У проекті передбачається використання сумісного освітлення. У світлий час доби приміщення освітлюватиметься через віконні отвори, в решту часу використовуватиметься штучне освітлення.

Як штучне освітлення необхідно використовувати штучне робоче загальне освітлення. Для загального освітлення необхідно використовувати люмінесцентні лампи. Вони володіють наступними перевагами: високою світловою віддачею, тривалим терміном служби, хоча мають і недоліки: високу пульсацію світлового потоку.

При експлуатації ПЕОМ виробляється зорова робота. Відповідно до ДБН В.2.5-28-2006 [27] ця робота відноситься до розряду 5а. При цьому нормоване освітлення на робочому місці(Ен) при загальному освітленні рівна 200 лк.

Приміщення завдовжки 12 м, шириною 10 м, заввишки 4 м обладнується світильниками типу ЛП02П, оснащеними лампами типу ЛБ зі світловим потоком 3120 лм кожна.

Виконаємо розрахунок кількості світильників в робочому приміщенні завдовжки а=12 м, шириною b=10 м, заввишки з=4 м, використовуючи формулу (4.5) розрахунку штучного освітлення при горизонтальній робочій поверхні методом світлового потоку:

$$n = (E \cdot S \cdot Z \cdot k)/(F \cdot U \cdot M), \quad (4.5)$$

де F - світловий потік = 3120 лм;

E - максимально допустима освітленість робочих поверхонь = 200 лк;

S - площа підлоги = 120 м2;

Z - поправочний коефіцієнт світильника = 1,2;

k - коефіцієнт запасу, що враховує зниження освітленості в процесі експлуатації світильників = 1,5;

n - кількість світильників;

U - коефіцієнт використовування освітлювальної установки = 0,6;

M - кількість ламп у світильнику =2.

З формули (4.5) виразимо n (4.6) і визначимо кількість світильників для даного приміщення:

$$n = (E \cdot S \cdot Z \cdot k)/(F \cdot U \cdot M), \quad (4.6)$$

Отже, $n = (200 \cdot 120 \cdot 1{,}2 \cdot 1{,}5)/(3120 \cdot 0{,}6 \cdot 2) = 12$

Виходячи з цього, рекомендується використовувати 12 світильників. Світильники слід розміщувати рядами, бажано паралельно стіні з вікнами. Схема розташування світильників зображена на рис. 4.1.

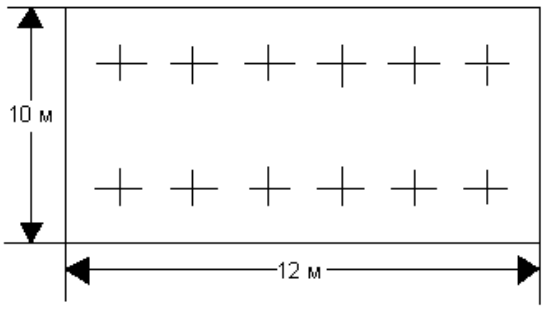



Рисунок 4.1 - Схема розташування світильників

## 4.4 Рекомендації по пожежній безпеці

Пожежі в приміщеннях, де встановлена обчислювальна техніка, представляють небезпеку для життя людини. Пожежі також пов’язані як з матеріальними втратами, так і з відмовою засобів обчислювальної техніки, що у свою чергу спричиняє за собою порушення ходу технологічного процесу.

Пожежа може виникнути при наявності горючої речовини та внесення джерела запалювання в горюче середовище. Пальними матеріалами в приміщеннях, де розташовані ПЕОМ, є:

– поліамід - матеріал корпусу мікросхеми, горюча речовина, температура самозаймання аерогелю 420 °С ;

– полівінілхлорид - ізоляційний матеріал, горюча речовина, температура запалювання 335 °С, температура самозаймання 530 °С, кількість енергії, що виділяється при згоранні - 18000 - 20700 кДж/кг;

– стеклотекстоліт ДЦ - матеріал друкарських плат, важкозаймистий матеріал, показник горючості 1.74, не схильний до температурного самозаймання;

– пластика кабельний №489 - матеріал ізоляції кабелю, горючий матеріал, показник горючості більш 2.1;

– деревина - будівельний і обробний матеріал, матеріал з якого виготовлені меблі, горючий матеріал, показник горючості більше 2.1, теплота згорання 18731 - 20853 кДж/кг, температура запалювання 399 °С, схильна до самозаймання.

Згідно ДСТУ Б В.1.1-36-2016 [28] приміщення відносяться до категорії В (пожежовибухонебезпечним) і згідно правилам побудови електроустановок простір усередині приміщення відноситься до вогненебезпечної зони класу П - IIа (зони, розташовані в приміщеннях, в яких зберігаються тверді горючі речовини).

Потенційними джерелами запалення при роботі ПЕОМ є:

– іскри при замиканні і розмиканні ланцюгів;

– іскри і дуги коротких замикань;

– перегріви від тривалого перевантаження і наявності перехідного опору.

Продуктами згорання, що виділяються при пожежі, є : оксид вуглецю, сірчистий газ, оксид азоту, синильна кислота, акропеін, фосген, хлор та ін. При горінні пластмас, окрім звичайних продуктів згорання, виділяються різні продукти термічного розкладання: хлорангідрідні кислоти, формальдегіди, хлористий водень, фосген, синильна кислота, аміак, фенол, ацетон, стирол та ін., що шкідливо впливають на організм людини.

Для захисту персоналу від дії небезпечних і шкідливих чинників пожежі проектом передбачається застосування промислового протигаза з коробкою марки В(жовта).

Пожежна безпека об’єктів народного господарства регламентується і забезпечується системами запобігання пожежам і протипожежному захисту[28]. Для успішного гасіння пожеж вирішальне значення має швидке виявлення пожежі і своєчасний виклик пожежних підрозділів до місця пожежі.

Зменшити горюче навантаження не представляється можливим, тому проектом передбачається застосувати наступні способи і їх комбінації для запобігання утворенню(внесення) джерел запалення :

– застосування устаткування, що задовольняє вимогам електростатичної безпеки;

– застосування в конструкції швидкодіючих засобів захисного відключення можливих джерел запалення;

– виключення можливості появи іскрового заряду статичної електрики в горючому середовищі з енергією, рівної і вище мінімальної енергії запалення;

– підтримка температури нагріву поверхні машин, механізмів, устаткування, пристроїв, речовин і матеріалів, які можуть увійти до контакту з пальним середовищем, нижче гранично допустимої, становить 80% як найменшої температури самозаймання пального.

– заміна небезпечних технологічних операцій більш безпечними;

– ізольоване розташування небезпечних технологічних установок і устаткування;

– зменшення кількості пальних і вибухонебезпечних речовин, що знаходяться у виробничих приміщеннях;

– запобігання можливості утворення пальних сумішей на лінії, вентиляційних системах і ін.;

– механізація, автоматизація та справність(потокова) виробництва;

– суворе дотримання стандартів і точне виконання встановленого технологічного режиму;

– запобігання можливості появи в небезпечних місцях джерел запалення;

– запобігання розповсюдженню пожеж і вибухів;

– використовування устаткування і пристроїв, при роботі яких не виникає джерел запалення;

– виконання вимог сумісного зберігання речовин і матеріалів;

– наявність громовідводу;

– ліквідація можливості самозаймання речовин і матеріалів .

Для запобігання пожежі в обчислювальних центрах проектом пропонується виконання наступних вимог :

– електроживлення ЕОМ повинно мати автоматичне блокування відключення електроенергії на випадок зупинки системи охолоджування і кондиціонування;

– система вентиляції обчислювальних центрів повинна бути обладнана блокуючими пристроями, що забезпечують її відключення на випадок пожежі;

– робочі місця повинні бути оснащені пожежними щитами, сигналізацією, засобами для сповіщення про пожежну небезпеку (телефонами), медичними аптечками для надання першої медичної допомоги, розробленим планом евакуації.

Для зниження пожежної небезпеки в приміщеннях використовуються первинні засоби гасіння пожеж, а також система автоматичної пожежної сигналізації, яка дозволяє знайти початкову стадію загоряння, швидко і точно оповістити службу пожежної охорони про час і місце виникнення пожежі.

Відповідно до правил пожежної безпеки для промислових підприємств приміщення категорії В підлягають устаткуванню системами автоматичної пожежної сигналізації. Проектом передбачається застосування датчика типу ІДФ - 1(димовий фотоелектричний датчик), оскільки специфікою пожеж обчислювальної техніки і радіоапаратури є, в першу чергу, виділення диму, а потім - підвищення температури.

При виникненні пожежі в робочому приміщенні обслуговуючий персонал зобов'язаний негайно вжити заходи по ліквідації пожежі. Для ліквідації пожежі використовують вогнегасники (пінні для повітря ОП-5, ОП-6, ОП-9, вуглекислотні ОУ-5), пісок, пожежний інвентар (сокири, ломи, багри, шерстяну або азбестову ковдри). Як засіб індивідуального захисту проектом передбачається використовування промислового протигаза з маскою, фільтруючої коробки В.

В якості організаційно-технічних заходів рекомендується проводити навчання робочого персоналу правилам пожежної безпеки.

# ВИСНОВКИ

У рамках даної роботи проведено огляд поширених методів кластеризації. Найбільш релевантними на даний час є методи кластеризації часових рядів. Кластеризація даних часового ряду, як і кластеризація для всіх типів даних, має мету створення кластерів з високою подобою всередені кластеру й низькою міжкластерною подобою. А саме, об'єкти, що належать тому ж кластеру, повинні показати високу подобу один одному, у той час як об'єкти, що належать різним кластерам, повинні показати низьку подобу, тобто високу відстань друг від друга.

Для експериментальних досліджень було обрано набір даних, що сформовані з бази даних ЕКГ серцевих ударів. Задача кластеризації та класифікації цих даних допомагає в обробці та виявленні аномалій у людей для діагностики проблем серцево-судинної системи.

У рамках дипломної роботи був розроблений і реалізований програмний засіб кластерного аналізу часових рядів. При проведені аналізу усі методи дали приблизно однаковий результат з невеликою перевагою методу к-середніх з використанням евклідової відстані. Найбільш перспективними методами кластеризації є DTW та Soft-DTW, які є стійкими до часових зміщень.

У розділі «Охорона праці» виконано аналіз потенційних небезпек при роботі із засобами обчислювальної техніки і механізмами, розроблені заходи щодо техніки безпеки, заходи, які забезпечують виробничу санітарію і гігієну праці, розраховане штучне освітлення, виконані рекомендації по пожежній безпеці.

# ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1) Tryon R.C. Cluster analysis. — London: Ann Arbor Edwards Bros, 1939. — 139 p.

2) Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: исследование зависимостей. — М.: Финансы и статистика, 1985.

3) Мандель И.Д. Кластерный анализ. ¬– М.: финансы и статистика. 1988. – 176 с.

4) Барсегян А.А. Методы и модели анализа данных: OLAP и Data Mining / Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. – СПб.: БХВ-Петербург, 2004. – 336 с.

5) Олдендерфер М. С., Блэшфилд Р. К. Кластерный анализ / Факторный, дискриминантный и кластерный анализ: пер. с англ.; Под. ред. И. С. Енюкова. — М.: Финансы и статистика, 1989. — 215 с.

6) Microsoft SQL Server 2008: Data mining – интеллектуальный анализ данных. Пер. с англ. / Дж. Макленнен, Чж. Танг, Б. Криват. – БХВ-Петербург. 2009. – 720 с.

7) Бураго Д. Ю., Бураго Ю. Д., Иванов С. В. Курс метрической геометрии. — 2004.

8) Бериков В. С., Лбов Г. С. Современные тенденции в кластерном анализе // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы», 2008. — 26 с.

9) Орлов А. И. Прикладная статистика. Учебник для вузов. — М.: Экзамен, 2006. — 672 с

10) Вятченин Д. А. Нечёткие методы автоматической классификации. — Минск: Технопринт, 2004. — 219 с.

11) MacQueen J. (1967). Some methods for classification and analysis of multivariate observations. In Proc. 5th Berkeley Symp. on Math. Statistics and Probability, pages 281—297.

12) Neal, Radford; Hinton, Geoffrey (1999). Michael I. Jordan, ed. “A view of the EM algorithm that justifies incremental, sparse, and other variants” (PDF). Learning in Graphical Models. Cambridge, MA: MIT Press: 355—368. ISBN 0262600323. Дата обращения 2009-03-22.

13) Воронцов К. В. Лекции по алгоритмам кластеризации и многомерного шкалирования, МГУ, 2007

14) Жамбю М. Иерархический кластер-анализ и соответствия. — М.: Финансы и статистика, 1988. — 345 с.

15) Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96) / Evangelos Simoudis, Jiawei Han, Usama M. Fayyad. — AAAI Press, 1996. — С. 226–231.

16) Сёмкин Б. И. Эквивалентность мер близости и иерархическая классификация многомерных данных // Иерархические классификационные построения в географической экологии и систематике. Владивосток: ДВНЦ АН СССР, 1979. С. 97—112.

17) Паршутин С.В. Кластеризация временных рядов с применением карт самоорганизации. Рижский Технический университет, Conference Paper, May 2007

18) Clustering Algorithms and Applications. Edited by Charu C. Aggarwal, Chandan K. Reddy. Taylor & Francis Group, LLC, 2014.

19) Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. IEEE Eng in Med and Biol 20(3):45-50 (May-June 2001). (PMID: 11446209).

20) PhysioBank Annotations – PhysioNet. Updated Wednesday, 6 July 2016 at 13:41 EDT. https://www.physionet.org/physiobank/annotations.shtml

21) ДСН 3.3.6.042-99 Державні санітарні норми мікроклімату виробничих приміщень. Режим доступу: www. URL: https://zakon.rada.gov.ua/rada/show/va042282-99

22) НПАОП 0.00-7.15-18 Вимоги щодо безпеки та захисту здоров`я працівників під час роботи з екранними пристроями. Режим доступу: www. URL: https://zakon.rada.gov.ua/laws/show/z0508-18

23) ДСТУ 7237:2011 Національний стандарт України. Система стандартів безпеки праці. Електробезпека. Загальні вимоги та

24) Номенклатура видів захисту. Режим доступу: www. URL: https://zakon.rada.gov.ua/rada/show/ru/v0037831-11

25) ДСанПіН 3.3.2.007-98. Державні санітарні правила і норми. Гігієнічні вимоги до організації роботи з візуальними дисплейними терміналами електронно-обчислювальних машин. Режим доступу: www. URL: https://zakon.rada.gov.ua/rada/show/v0007282-98

26) ДБН В.2.5-67:2013. Опалення вентиляція та кондиціонування. Режим доступу: www. URL: https://zakon.rada.gov.ua/rada/show/v0024858-13

27) ДБН В.2.5-28-2006. Природне і штучне освітлення. Режим доступу: www. URL: https://zakon.rada.gov.ua/rada/show/v0168667-06

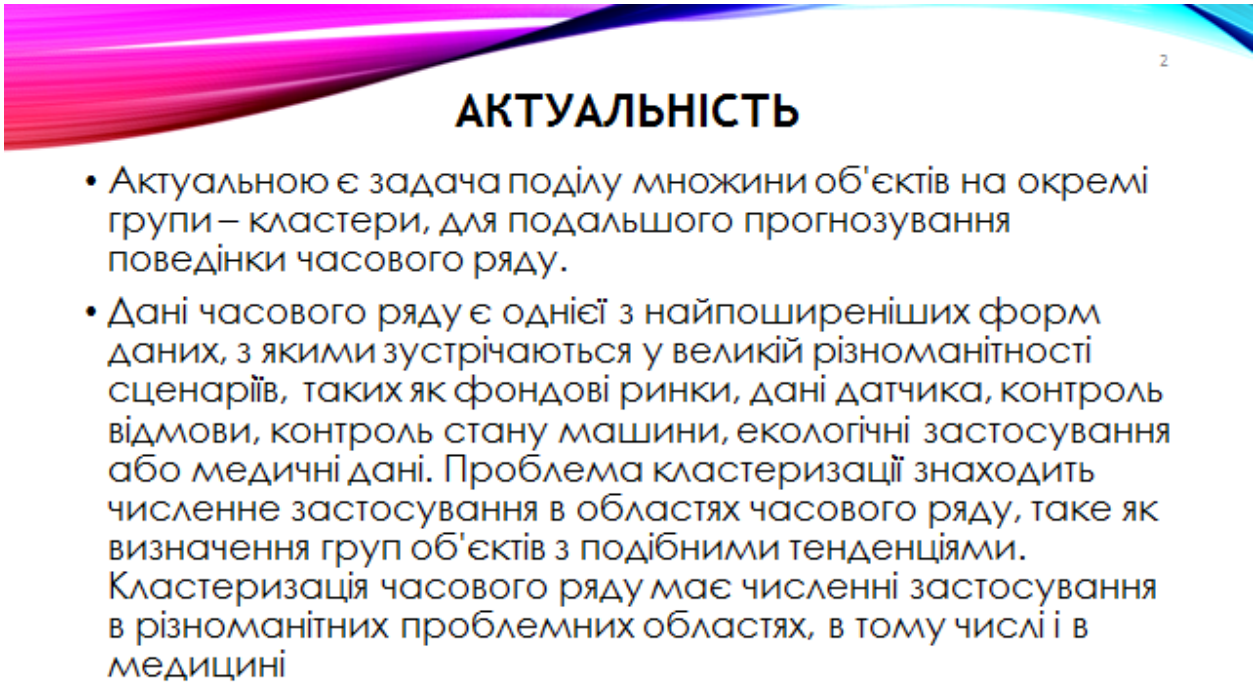
28) ДСТУ Б В.1.1-36-2016. Визначення категорії приміщень, будинків та зовнішніх установок за вибухопожежною та пожежною безпекою. Режим доступу: www. URL: http://online.budstandart.com/ua/catalog/doc-page.html?id_doc=65419

29) ДСП 173-96. Державні санітарні правила планування та забудови населених пунктів Режим доступу: www. URL: https://zakon.rada.gov.ua/laws/show/z0379-96

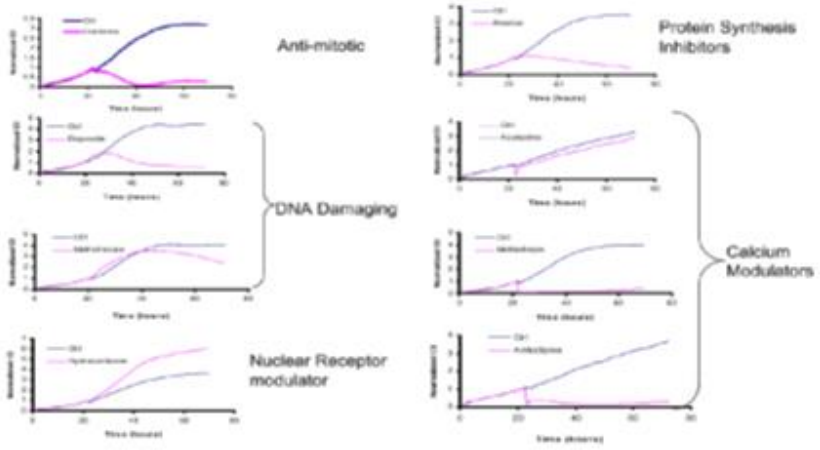
30) Симметрон. Электронные компоненты. Каталог 2002, 2002г. – 192с.

**Додаток А**

**Комп’ютерна презентація**

# ЧАСОВІ РЯДИ В МЕДИЦИНІ




# ХАРАКТЕРИСТИКА ВХІДНОГО НАБОРУ ДАНИХ ЧАСОВИХ РЯДІВ

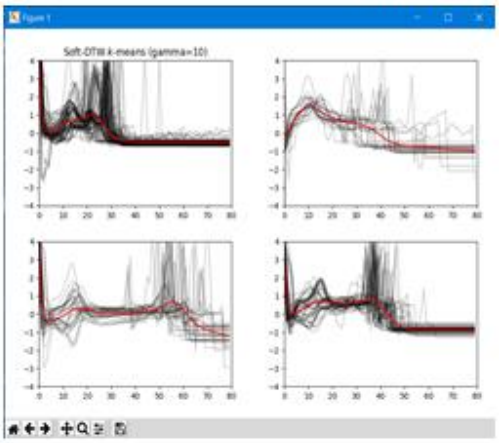
# КОДИ АНОТАЦІЙ СЕРЦЕВИХ УДАРІВ

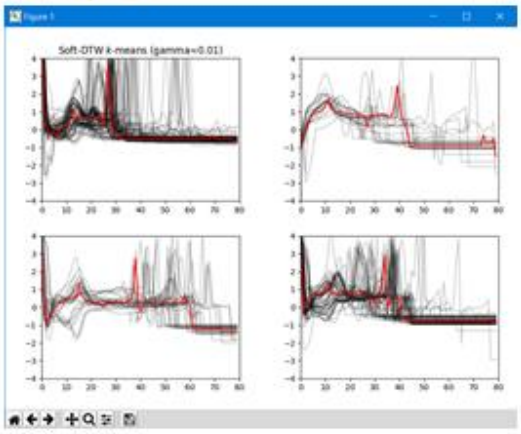
| Код | Опис |
|---|---|
| N | Звичайний удар (відображається як "·" в базах даних PhysioBank, LightWAVE, pschart і psfd) |
| L | Білий блок гілок лівого пучка |
| R | Правий блок розгалужується |
| B | Переміщення блоку розгалужень пакета (не вказано) |
| A | Передсердя передчасно б'ють |
| a | Аберіальний передчасний збиток передсердя |
| J | Нодальний (сполучний) передчасний удар |
| S | Надшлуночкові передчасні або ектопічні удари (передсердні або вузлові) |
| V | Передчасне скорочення шлуночків |
| r | R-на-T передчасне скорочення шлуночків |
| F | Злиття шлуночків і нормального биття |
| i | Втеча з передсердя |
| j | Nodal (junctional) перебіг удару |
| n | Надшлуночковий бічний потік (атріальний або вузловий) |
| i | Шлуночкова бійка |
| / | Темп збився |
| f | Злиття темпів і нормального удару |
| Q | Некласифікований удар |
| ? | Удар не класифікується під час навчання |

# ВХІДНІ ПАРАМЕТРИ

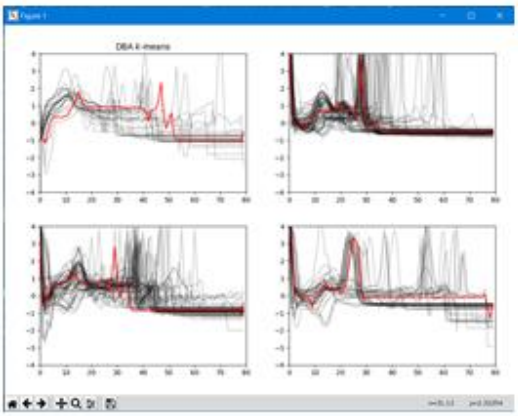
**input** ІМ'Я ФАЙЛУ;
**n** – цільова кількість кластерів;
**dba** – кластеризація методом DBA-k-means (DTW Barycenter Averaging);
**softdtw** – кластеризація методом soft-DTW;
**euclidean** – кластеризація методом $k$-середніх з використанням евклідової відстані;
**gamma** – параметр згладжування при використанні методу soft-DTW.

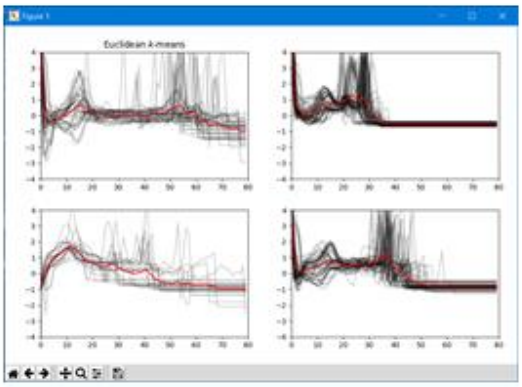
## РЕЗУЛЬТАТИ КЛАСТЕРИЗАЦІЇ МЕТОДОМ SOFT-DTW ПРИ У=10 ТА У=0,01




## РЕЗУЛЬТАТИ КЛАСТЕРИЗАЦІЇ МЕТОДОМ DBA-K-MEANS ТА K-MEANS З ВИКОРИСТАННЯ ЕВКЛІДОВОЇ ВІДСТАНІ



# ПОРІВНЯННЯ МЕТОДІВ КЛАСТЕРИЗАЦІЇ

| Метод | Кластер N | Кластер S | Кластер V | Кластер F | Загальний % співпадіння |
|---|---|---|---|---|---|
| soft-DTW при γ=10 | 39,62% | 35,85% | 52,00% | 52,22% | 47,50% |
| soft-DTW при γ=0,01 | 30,00% | 34,92% | 59,09% | 43,53% | 44,00% |
| DBA-k-means | 43,59% | 37,70% | 60,87% | 53,25% | 47,50% |
| Euclidean k-means | 41,67% | 31,58% | 62,50% | 58,97% | 49,00% |

# ВИСНОВКИ

- Для експериментальних досліджень було обрано набір даних, що сформовані з бази даних ЕКГ серцевих ударів. Задача кластеризації та класифікації цих даних допомагає в обробці та виявленні аномалій у людей для діагностики проблем серцево-судинної системи.
- У рамках дипломної роботи був розроблений і реалізований програмний засіб кластерного аналізу часових рядів. При проведені аналізу усі методи дали приблизно однаковий результат з невеликою перевагою методу к-середніх з використанням евклідової відстані. Найбільш перспективними методами кластеризації є DTW та Soft-DTW, які є стійкими до часових зміщень.