

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ  
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»  
Факультет інформатики та обчислювальної техніки  
Кафедра технічної кібернетики

«На правах рукопису»  
УДК 004.93'14

«До захисту допущено»

Завідувач кафедри

\_\_\_\_\_ М. М. Ткач  
(підпис) (ініціали, прізвище)

“ ” \_\_\_\_\_ 2015 р.

## Магістерська дисертація

зі спеціальності: 8.05020102 Комп'ютеризовані та робото технічні системи  
(код та назва спеціальності)

на тему: Дослідження методів критеріального групування в задачах  
кластеризації даних пацієнтів

Виконала: студентка 6 курсу,

групи ІК-31м  
(шифр групи)

Демчук Леся Юріївна

(прізвище, ім'я, по батькові)

(підпис)

Науковий керівник к.н.т., доцент Савицький А.Й.

(посада, науковий ступінь, вчене звання, прізвище та ініціали)

(підпис)

Консультант охорона  
праці  
(назва розділу)

к.н.т., доцент Праховнік Н.А.

(науковий ступінь, вчене звання, прізвище, ініціали)

(підпис)

Рецензент

(посада, науковий ступінь, вчене звання, науковий ступінь, прізвище, ініціали)

(підпис)

Засвідчую, що у цій магістерській дисертації  
немає запозичень з праць інших авторів без  
відповідних посилань.

Студент \_\_\_\_\_  
(підпис)

Київ – 2015 року

**ПОЯСНЮВАЛЬНА ЗАПИСКА ДО  
МАГІСТЕРСЬКОЇ ДИСЕРТАЦІЇ**

на тему: Дослідження методів критеріального групування в задачах  
кластеризації медичних даних пацієнтів

**НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ**  
**“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ”**

Факультет інформатики та обчислювальної техніки  
Кафедра технічної кібернетики

Освітньо-кваліфікаційний рівень «магістр»

Напрямок підготовки 6.050201 «Системна інженерія»

Спеціальність 8.05020102 «Комп’ютеризовані та робототехнічні системи»

ЗАТВЕРДЖУЮ

Завідувач кафедри

\_\_\_\_\_ М. М. Ткач  
(підпис) (ініціали, прізвище)

«\_\_» \_\_\_\_\_ 2015 р.

**ЗАВДАННЯ**

на магістерську дисертацію студентці

Демчук Лесі Юріївні

(прізвище, ім’я, по батькові)

1. Тема дисертації: Дослідження методів критеріального групування в задачах кластеризації даних пацієнтів

науковий керівник дисертації Савицький Артем Йосипович, к.т.н., доцент  
(прізвище, ім’я, по батькові, науковий ступінь, вчене звання)

затверджені наказом по університету від «10» березня 2015р. №292/2С

2. Строк подання студентом дисертації 02.06.2015р.

3. Об’єкт дослідження Методи критеріального групування

4. Предмет дослідження Правильність отриманого результату кластеризації методів критеріального групування

5. Перелік завдань, які потрібно розробити

- Огляд проблеми аналізу даних пацієнтів;
- Аналіз існуючих систем інтелектуального аналізу даних;
- Формування вибірки даних;
- Аналіз існуючих методів кластеризації;
- Розробка програмного продукту;

6. Орієнтовний перелік ілюстративного матеріалу      Блок-схема роботи ПЗ,  
структурна схема, функціональна схема, діаграма класів, копії екранних форм,  
діаграма послідовностей

7. Орієнтовний перелік публікацій      Дослідження методів критеріального  
групування в задачах кластеризації даних пацієнтів

8. Консультанти розділів дисертації

| Розділ        | Прізвище, ініціали та посада консультанта | Підпис, дата   |                  |
|---------------|-------------------------------------------|----------------|------------------|
|               |                                           | завдання видав | завдання прийняв |
| Охорона праці | Праховнік Н.А., к.н.т., доцент            |                |                  |
| ПЗ            | Пасько В.П., к.н.т., доцент               |                |                  |
|               |                                           |                |                  |
|               |                                           |                |                  |
|               |                                           |                |                  |
|               |                                           |                |                  |

9. Дата видачі завдання: 01.09.2014р.

## Календарний план

| № з/п | Назва етапів виконання магістерської дисертації                              | Строк виконання етапів | Примітка |
|-------|------------------------------------------------------------------------------|------------------------|----------|
| 1     | Аналіз проблеми накопичення даних пацієнтів                                  | 15.10.2014             |          |
| 2     | Огляд існуючих рішень вирішення проблеми аналізу даних пацієнтів             | 01.12.2014             |          |
| 3     | Формування вхідної вибірки даних                                             | 13.12.2014             |          |
| 4     | Огляд та аналіз методів кластеризації                                        | 20.02.2015             |          |
| 5     | Розробка програмного забезпечення для кластеризації та класифікації об'єктів | 18.03.2015             |          |
| 6     | Оцінка результатів кластеризації                                             | 24.04.2015             |          |
| 7     | Оформлення основної частини ПЗ                                               | 10.05.2015             |          |
| 8     | Оформлення розділу «Охорони праці та безпеки в надзвичайних ситуаціях»       | 18.05.2015             |          |
| 9     | Розробка графічних матеріалів та презентації                                 | 25.05.2015             |          |

Студент

\_\_\_\_\_

(підпис)

\_\_\_\_\_

(ініціали, прізвище)

Науковий керівник дисертації

\_\_\_\_\_

(підпис)

\_\_\_\_\_

(ініціали, прізвище)

## РЕФЕРАТ

Магістерська дисертація: 88 с., 29 рис., 14 табл., 2 додатки, 32 формули, 35 джерела.

**Актуальність теми.** З огляду на стрімкий розвиток інформаційних технологій, які в свою чергу використовуються в різних установах, зокрема і в медичних закладах, і дозволяють накопичувати велику кількість даних, можна зробити висновок, що найближчим часом може виникнути проблема в знаходженні шляху отримання корисної інформації з великого набору даних, а також використанні отриманої інформації в подальшому.

**Мета роботи і задачі дослідження.** Мета дисертаційної роботи – дослідити методи кластеризації з точки зору їх використання для вирішення задачі аналізу даних пацієнтів (об'єктів), а також показати можливість подальшого використання отриманих результатів для класифікації нових об'єктів на основі накопичених даних про стан пацієнтів. Для досягнення поставленої мети потрібно вирішити наступні задачі:

- проаналізувати існуючі системи аналізу даних;
- підготувати та сформулювати вибірку даних для критеріального групування;
- дослідити існуючі методи критеріального групування;
- розробити програмне забезпечення, що автоматизує процес критеріального групування, а також подальшої класифікації об'єкта;
- здійснити оцінку правильності отриманих результатів.

**Об'єкт дослідження** – методи критеріального групування.

**Предмет дослідження** – правильність отриманого результату кластеризації методів критеріального групування.

**Методи дослідження.** Для дослідження методів критеріального групування використовується технологія Data Mining, а саме, одна із складової технології Data Mining – процес кластеризація даних.

**Наукова новизна отриманих результатів.**

Проведено аналіз існуючих методів критеріального групування з точки зору можливості їх використання при вирішенні задачі кластеризації даних пацієнтів. Запропоновано підхід до вирішення задачі.

DATA MINING, КЛАСТЕРИЗАЦІЯ, МЕТОДИ КЛАСТЕРИЗАЦІЇ,  
ВИБІРКА, ПРАВИЛЬНІСТЬ КЛАСТЕРИЗАЦІЇ

## ABSTRACT

**Background.** Given the rapid development of information technology, which in turn are used in various institutions, particularly in hospitals, and allow to accumulate large amounts of data, we can conclude that soon may be a problem to find a way to get useful information from large data sets and use the information received in the future.

**Purpose and research tasks.** The aim of the thesis - research methods clustering by providing them similar signs for further classification based on accumulated data on the patients. To achieve this goal you need to solve the following problems:

- to define the set of criteria for which division will be held on clusters;
- to prepare and to form the input of data to the grouping criterion;
- to examine the existing methods of grouping criterion;
- to develop software that automates the process of grouping criterion, and further classification of the object;
- to analyze of the results.

**Object of research** - methods of criterion grouping.

**Subject of investigation** - the correctness of the result of clustering methods grouping criterion.

**Methods.** To study methods of grouping criterion used technology Data Mining, namely, one of the component technologies of Data Mining - Data clustering process.

**Scientific novelty of the results.**

The analysis of existing methods of grouping criterion in problems of clustering of patients and made findings on the use of a method in problems of this type. The approach to solving the problem of clustering these patients medical institutions.

DATA MINING, CLUSTERING, CLUSTERING METHODS, SAMPLING, ACCURACY

## ЗМІСТ

|                                                                   |    |
|-------------------------------------------------------------------|----|
| ПЕРЕЛІК СКОРОЧЕНЬ.....                                            | 13 |
| ВСТУП .....                                                       | 14 |
| 1    ОГЛЯД ПРОБЛЕМИ АНАЛІЗУ ДАНИХ ПАЦІЄНТІВ .....                 | 15 |
| 1.1    Постановка задачі кластеризації даних пацієнтів .....      | 16 |
| ВИСНОВКИ ДО РОЗДІЛУ .....                                         | 18 |
| 2    ОГЛЯД СИСТЕМ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ .....            | 19 |
| 2.1    Огляд технології інтелектуального аналізу даних .....      | 19 |
| 2.2    Огляд існуючих систем інтелектуального аналізу даних ..... | 21 |
| 2.2.1    Statistica.....                                          | 21 |
| 2.2.2    Weka .....                                               | 21 |
| 2.2.3    RapidMiner .....                                         | 22 |
| 2.3    Аналіз існуючих систем інтелектуального аналізу даних..... | 23 |
| ВИСНОВКИ ДО РОЗДІЛУ .....                                         | 23 |
| 3    ФОРМУВАННЯ ВХІДНОЇ ВИБІРКИ ДАНИХ.....                        | 24 |
| 3.1    Опис атрибутів кластеризації .....                         | 24 |
| 4    ОГЛЯД МЕТОДІВ КЛАСТЕРИЗАЦІЇ.....                             | 26 |
| 4.1    Ієрархічні методи .....                                    | 27 |
| 4.1.1    Правила об'єднання кластерів.....                        | 29 |
| 4.2    Алгоритм К-середніх .....                                  | 33 |
| 4.3    ЕМ алгоритм.....                                           | 35 |
| 4.4    Алгоритм DBSCAN .....                                      | 38 |
| 4.5    Алгоритм COBWEB.....                                       | 40 |

|       |                                                                                                                   |    |
|-------|-------------------------------------------------------------------------------------------------------------------|----|
| 4.6   | Міри подібності.....                                                                                              | 41 |
| 4.6.1 | Евклідова відстань .....                                                                                          | 41 |
| 4.6.2 | Квадрат евклідової відстані. ....                                                                                 | 42 |
| 4.6.3 | Виважена Евклідова відстань .....                                                                                 | 42 |
| 4.6.4 | Хеммінгова відстань.....                                                                                          | 42 |
| 4.6.5 | Відстань Чебишева .....                                                                                           | 43 |
| 4.6.6 | Показникова відстань .....                                                                                        | 43 |
| 4.6.7 | Відстань Канберра .....                                                                                           | 43 |
| 4.6.8 | Відстань Брея-Картіса .....                                                                                       | 44 |
| 4.7   | Вибір алгоритмів кластеризації, міри близькості та методики оцінки результату кластеризації для дослідження ..... | 45 |
|       | ВИСНОВКИ ДО РОЗДІЛУ .....                                                                                         | 46 |
| 5     | ОПИС ПРОГРАМНОГО ПРОДУКТУ .....                                                                                   | 47 |
| 5.1   | Засоби розробки .....                                                                                             | 47 |
| 5.1.1 | Бібліотека Extreme Optimization Numerical for .NET .....                                                          | 47 |
| 5.2   | Вимоги до технічного забезпечення .....                                                                           | 49 |
| 5.3   | Архітектура програмного забезпечення .....                                                                        | 49 |
| 5.3.1 | Опис класів програмного забезпечення .....                                                                        | 50 |
| 5.3.2 | Специфікація функцій розробленого програмного забезпечення .....                                                  | 50 |
| 5.4   | Опис роботи з програмним забезпеченням .....                                                                      | 52 |
|       | ВИСНОВКИ ДО РОЗДІЛУ .....                                                                                         | 54 |
| 6     | ДОСЛІДЖЕННЯ ПРАВИЛЬНОСТІ РОБОТИ АЛГОРИМІВ .....                                                                   | 55 |
| 6.1   | Алгоритм k-means .....                                                                                            | 55 |

|                           |                                                                             |    |
|---------------------------|-----------------------------------------------------------------------------|----|
| 6.2                       | Ієрархічний алгоритм .....                                                  | 58 |
| 6.3                       | Узагальнення по результатах роботи алгоритмів.....                          | 63 |
| 6.4                       | Застосування отриманих результатів для розпізнавання нових<br>об'єктів..... | 63 |
| ВИСНОВКИ ДО РОЗДІЛУ ..... |                                                                             | 66 |
| 7                         | ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ                           | 67 |
| 7.1                       | Загальні положення.....                                                     | 67 |
| 7.2                       | Аналіз шкідливих та небезпечних факторів .....                              | 69 |
| 7.2.1                     | Аналіз виробничого шуму .....                                               | 69 |
| 7.2.2                     | Аналіз мікроклімату приміщення .....                                        | 70 |
| 7.2.3                     | Аналіз випромінювання .....                                                 | 71 |
| 7.2.4                     | Аналіз природнього та штучного освітлення .....                             | 71 |
| 7.2.5                     | Аналіз електробезпеки .....                                                 | 71 |
| 7.3                       | Інженерне рішення.....                                                      | 72 |
| 7.4                       | Пожежна безпека .....                                                       | 76 |
| 7.4.1                     | Категорія приміщення за вибухопожежною небезпекою ...                       | 77 |
| 7.4.2                     | Пожежна сигналізація .....                                                  | 77 |
| 7.4.3                     | Засоби гасіння пожежі.....                                                  | 78 |
| 7.5                       | Техніка безпеки .....                                                       | 79 |
| 7.6                       | План ліквідації надзвичайної ситуації .....                                 | 80 |
| 7.6.1                     | Можливі аварійні ситуації .....                                             | 80 |
| 7.6.2                     | Оцінка радіаційної обстановки в зоні радіаційного<br>забруднення .....      | 81 |
| ВИСНОВКИ ДО РОЗДІЛУ ..... |                                                                             | 83 |
| ВИСНОВКИ.....             |                                                                             | 84 |

|                        |    |
|------------------------|----|
| ПЕРЕЛІК ПОСИЛАНЬ ..... | 86 |
| ДОДАТКИ.....           | 83 |

## ПЕРЕЛІК СКОРОЧЕНЬ

БД – база даних;

ДП – дані пацієнтів;

XML – Extensible Markup Language;

CSV – Comma-Separated Values;

КВКО – кількість вірно кластеризованих об'єктів;

CLR – Common Language Runtime;

ПЗ- програмне забезпечення;

ЕОМ – електронна обчислювальна машина;

ПЕОМ – персональна електронна обчислювальна машина;

АЕС – атомна електростанція.

## ВСТУП

На даний час глобальний процес автоматизації зачепив практично всі сфери людської діяльності, включаючи медичну галузь. Сучасні інформаційні технології дозволяють накопичувати досить велику кількість інформації, що є характерним, зокрема і для медичних закладів. Дані про стан пацієнтів накопичуються роками та досить часто є різномірними(невпорядкованими). Варто зазначити, що особливістю таких даних є багато розмірність, що ускладнює процес аналізу даних. Також дані можуть бути представлені не тільки в числовому вигляді, але і у категорійному вигляді. На сьогоднішній день досить відомою технологією аналізу даних Data mining.

Data mining – це процес виявлення у необроблених даних раніше невідомих нетривіальних, практично корисних і доступних інтерпретацій знань, необхідних для прийняття рішень у різних сферах діяльності. Data mining вирішує ряд задач, які пов'язані з інтелектуальним аналізом даних. Однією із задач, які вирішує Data mining є кластеризація - процес, що дозволяє групувати деякий набір об'єктів за схожим набором параметрів[1]. Іншими словами, з'являється можливість розподілити накопиченні дані медичних закладів по певних групах, за деякими схожими критеріями.

На разі існує велика кількість алгоритмів кластеризації, які відрізняються, щонайменше, набором вхідних параметрів і в залежності від обраних параметрів результати кластеризації можуть відрізнятись. Виникає складність в оцінці результату кластеризації. Один із способів оцінити результат кластеризації – розташувати об'єкти та центри кластерів на координатній площині, але тут виникає інша проблема - у випадку, коли дані багаторозмірні відпадає можливість розташування їх на координатній площині. В такому випадку необхідна деяка методика оцінки результатів кластеризації.

## 1 ОГЛЯД ПРОБЛЕМИ АНАЛІЗУ ДАНИХ ПАЦІЄНТІВ

Сучасний рівень розвитку інформаційних технологій дозволяє різним установам накопичувати та зберігати дані у базах даних, де медичні заклади не є виключенням. З часом кількість накопичених даних зростає та досягає великих розмірів. Через достатньо велику кількість інформації – лікар може ознайомитись лише з деякою її частиною, тому є необхідним знайти спосіб опрацювання інформації значно ефективнішим шляхом. З іншого боку, маючи накопичену інформацію(знання про стан пацієнтів), чому б не використовувати її для виявлення у певного (нового) пацієнта симптомів, які уже спостерігались у пацієнта, дані якого містяться у БД. Також варто зазначити, що особливістю медичних даних є представлення їх не завжди в числовому вигляді, що ускладнює процес обробки даних машиною.

На даний час для отримання корисної інформації з великого набору даних використовують технологію – Data mining. Data mining - це технологія, яка вивчає процес отримання нових, реальних та раніше невідомих знань в базах даних. Однією із задач технології Data mining є процес кластеризації. Процес кластеризації дозволяє розглядати великі набори даних та різко скорочувати їх, робити їх компактними та наглядними. Задача кластеризації полягає в розподіленні деякої множини об'єктів на групи схожих об'єктів, що називаються кластерами. Результатом процесу кластеризації являється набір кластерів, які містять в собі схожі елементи вхідної множини елементів[1].

На разі існує велика кількість алгоритмів кластеризації, які відрізняються набором вхідних параметрів і в залежності від обраних параметрів результати кластеризації можуть відрізнятись. В контексті задачі кластеризації даних пацієнтів виникає проблема, яка полягає у виборі алгоритму кластеризації: алгоритм кластеризації необхідно вибрати таким чином, щоб розбиття на кластери було якомога правильним.

Однією із найпоширеніших проблем алгоритмів кластеризації є те, що більшість алгоритмів кластеризації очікують вхідним параметром число

кластерів і зазвичай кількість кластерів заздалегідь невідома[2]. І тому потрібно використовувати деякі емпіричні правила, для вибору оптимального числа кластерів. Що ж стосується задачі кластеризації даних пацієнтів, то тут більшості випадків кількість кластерів заздалегідь відома, оскільки лікар точно знає як класифікується та чи інша хвороба.

Задача кластеризації набуває значно вагомішого характеру, якщо застосувати кластери не тільки для наглядного представлення об'єктів, але і для розпізнавання нових. Кожен новий об'єкт (пацієнт) відноситься до того кластеру (ступінь важкості хвороби), приєднання до якого найкращим чином буде задовольняти критерій якості кластеризації. Тому, можемо припустити, що поведінка об'єкту може бути схожою із поведінкою інших об'єктів, які знаходяться в кластері.

### 1.1 Постановка задачі кластеризації даних пацієнтів

Нехай відома множина об'єктів ДП  $I$ , кожен із яких представлений деяким набором атрибутів. Потрібно побудувати множину кластерів  $C$ , та відображення  $F$  множини  $I$  на множину  $C$ , тобто  $F: I \rightarrow C$ . При цьому об'єкти, що належать одному кластеру – повинні бути схожі між собою, а ті що належать до різних кластерів – повинні істотно відрізнятись. Відображення  $F$ , задає модель даних, яка є рішенням[2]. Якість рішення задачі визначається кількістю вірно кластеризованих об'єктів.

Визначимо множину наступним чином

$$I = \{i_1, i_2, \dots, i_n\} \quad (1.1)$$

де  $i_n$  – об'єкт, що досліджується.

В свою чергу кожен із досліджуваних об'єктів характеризується набором параметрів:

$$i_n = \{x_1, x_2, \dots, x_m\}. \quad (1.2)$$

Кожна змінна в свою чергу  $x_m$  – може приймати значення з деякої множини:

$$x_m = \{v_m^1, v_m^2, \dots\}. \quad (1.3)$$

Задача кластеризації полягає в побудові множини кластерів:

$$C = \{c_1, c_2, \dots, c_g\}, \quad (1.4)$$

$c_g$  – кластер, який містить схожі об'єкти із множини  $I$ :

$$c_g = \{i_n, i_p | i_n \in I, i_p \in I \wedge d(i_n, i_p) < \sigma\}, \quad (1.5)$$

де  $\sigma$  - величина, що визначає міру близькості для включення об'єктів в один кластер;  $d(i_n, i_p)$  – міра близькості між об'єктами, яка називається відстанню.

Невід'ємне значення  $d(i_n, i_p)$  – називається відстанню між елементами  $i_n$  та  $i_p$  якщо виконуються наступні умови:

$$- d(i_n, i_p) \geq 0, \text{ для всіх } i_n \text{ та } i_p, \quad (1.6)$$

$$- d(i_n, i_p) = 0, \text{ тоді і тільки тоді коли } i_n = i_p, \quad (1.7)$$

$$- d(i_n, i_p) = d(i_p, i_n), \quad (1.8)$$

$$- d(i_n, i_p) \leq d(i_p, i_r) + d(i_r, i_p). \quad (1.9)$$

Якщо відстань  $d(i_n, i_p)$  менше деякого значення  $\sigma$ , то можна сказати, що елементи розташовані близько один до одного та знаходяться в одному кластері. В іншому випадку елементи знаходяться в різних кластерах. Рішенням задачі є розбиття елементів на кластери, яке повинно задовольняти

деякому критерію. Цей критерій може являти собою деяких функціонал, який називають цільовою функцією.

## ВИСНОВКИ ДО РОЗДІЛУ

1. Здійснено огляд проблеми аналізу даних пацієнтів. Для вирішення даної проблеми обрано технологію Data Mining, а саме, одна із складової даної технології – кластеризація даних.
2. Виявлено основні проблеми кластеризації даних пацієнтів: багаторозмірність даних, можливе категоріальне представлення даних, вибір методу кластеризації та оцінка результату.
3. Здійснено постановку задачі кластеризації даних пацієнтів.

## 2 ОГЛЯД СИСТЕМ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

### 2.1 Огляд технології інтелектуального аналізу даних

Інтелектуальний аналіз даних (data mining) – це процес отримання та подальше застосування знань або раніше невідомої інформації з уже наявних доступних даних [3].

Швидкий розвиток і активне впровадження технологій інтелектуального аналізу даних пов'язано, в першу чергу, з необхідністю аналітичної обробки надвеликих обсягів інформації, що накопичується в сучасних сховищах даних.

Великий обсяг інформації з однієї сторони, дозволяє отримати більш точні розрахунки та аналіз, з іншої – перетворює пошук рішення в складне завдання. В результаті з'явився цілий клас систем, які дозволяють виконати аналіз всього обсягу інформації та спростити процес прийняття рішення.

Складність і різноманітність методів Data Mining вимагають створення спеціалізованих засобів для вирішення типових завдань аналізу інформації в конкретних галузях. Оскільки ці засоби використовуються в складі складних багатофункціональних систем підтримки прийняття рішень, вони повинні легко інтегруватися в подібні системи.

Data Mining є мультидисциплінарною галуззю, що виникла і розвивається на базі досягнень прикладної статистики, розпізнавання образів, методів штучного інтелекту, теорії баз даних тощо, тобто методи Data Mining знаходяться на стику різних напрямків інформаційних технологій, що представлені на рис. 1.1. Саме тому на сьогоднішній день існує така велика кількість методів і алгоритмів, реалізованих у різних системах Data Mining [2,5-6]. Багато з таких систем інтегрують в собі відразу кілька підходів.



Рис 1.1. Мультидисциплінарність технології Data mining

Задачі, що вирішує Data mining можна поділити на описові та передбачувальні. В описових задачах найголовніше – це дати наглядний опис наявних прихованих закономірностей, в той час як в задачах передбачення на першому плані стоїть питання передбачення для тих випадків, для яких даних ще немає.

До описових задач відносять:

- пошук асоціативних правил або паттернів(зразків)
- групування об'єктів за схожими ознаками
- побудова регресійної моделі

До передбачувальних задач відносять:

- класифікація об'єктів(для попередньо заданих класів)
- регресійний аналіз

## 2.2 Огляд існуючих систем інтелектуального аналізу даних

На сьогодні найвідоміші зарубіжні системи інтелектуального аналізу: Statistica, Weka, RapidMiner.

### 2.2.1 Statistica

Statistica - пакет для статистичного аналізу, розроблений компанією StatSoft[7]. В пакеті STATISTICA реалізовані процедури для аналізу даних (data analysis), управління даними (data management), добування даних (data mining), візуалізації даних (data visualization). Система STATISTICA складається з окремих модулів, кожен з яких є повноцінним Windows-застосунком. Можна швидко і зручно переключатися з одного модуля в інший, клацаючи мишею на значках модулів на робочому столі чи активізуючи відповідне вікно застосунку (якщо воно вже було відкрите) або вибираючи модулі в діалоговому вікні. Інтерфейс системи може бути вбудований у конкретний проект користувача. Найсильнішою стороною пакета є графіка і засоби редагування графічних матеріалів. Існує можливість розробити свій дизайн графіка і додати його до меню. Засоби керування графіками містять у собі роботу одночасно з декількома графіками, зміну розмірів складних об'єктів, розширені можливості малювання з додаванням художньої перспективи і спеціальних ефектів, розбивку сторінок. Наприклад, трьохмірні графіки можна обертати, накладати один на одного, стискувати або збільшувати. Крім того можна побачити на графіках, які фрагменти там змінилися під впливом змін в одній з перемінних.

### 2.2.2 Weka

Weka — програмне забезпечення для аналізу даних, написане на Java в університеті Уайкато (Нова Зеландія), розповсюджується за ліцензією GNU GPL.

Переваги використання Weka:

- вільний доступ під ліцензією GNU General Public License

- портативність, оскільки Weka повністю реалізована на мові програмування Java і, отже, працює практично на будь-якій сучасній платформі
- повна колекція попередньої обробки даних і методи моделювання
- простота використання завдяки своїм графічним інтерфейсом користувача.

Weka підтримує кілька стандартних задач інтелектуального аналізу даних, зокрема, попередню обробку даних, кластеризації, класифікації, регресії, візуалізації[8].

### 2.2.3 RapidMiner

RapidMiner - це програмна платформа, яка забезпечує інтегроване середовище для машинного навчання, інтелектуального аналізу даних, аналіз тексту, прогностичної аналітики та бізнес-аналітики, написано на мові програмування Java. Він використовується для ділових і промислових додатків, а також для наукових досліджень, освіти, професійної підготовки, швидкого прототипування і розробки додатків і підтримує всі етапи процесу видобутку даних, включаючи результати візуалізації, перевірки та оптимізації[9]. Starter Edition доступний для безкоштовного скачування, Personal Edition пропонується за \$ 999, Professional Edition € \$ 2999 і розцінки на Enterprise Edition доступний від розробника.

RapidMiner використовує модель клієнт-сервер. RapidMiner забезпечує аналіз даних і процедури машинного навчання, включаючи: завантаження даних і перетворення (перетворення з категорійних даних у числові), даних первинної обробки та візуалізації, прогностного аналізу і статистичного моделювання, оцінки та розгортання.

RapidMiner надає зручний графічний інтерфейс для розробки і виконання аналітичних процесів.

## 2.3 Аналіз існуючих систем інтелектуального аналізу даних

Розглянуті вище системи вирішують задачу кластеризації приблизно однаковими методами: k-means, ієрархічний алгоритм, DBSCAN та ін. Вагомою перевагою над іншими двома системами має RapidMiner, оскільки ця система не тільки здійснює аналіз даних, але й надає можливість попередньо перетворити дані із одного стану в інший (наприклад, із категоріального виду в цифровий), що є важливим коли дані подані не в цифровому вигляді.

Розглядаючи імовірність використання даних систем для проведення дослідження виявлено, що в описаних вище системах немає можливості оцінити правильність результату кластеризації при використанні того чи іншого алгоритму (оскільки правильність результату в межах поставленої задачі має велике значення). Тут користувач отримує лише інформацію про приналежність об'єктів до кластерів, не маючи змоги порівняти отриманні результати з еталонними. Тому, прийнято рішення щодо розробки програмного забезпечення, яке надаватиме можливість оцінити правильність результату кластеризації, а також використати отримані результати кластеризації для того, щоб здійснювати класифікацію об'єктів.

## ВИСНОВКИ ДО РОЗДІЛУ

1. Здійснено огляд задач, які вирішує технологія інтелектуального аналізу даних Data mining. Дана технологія вирішує задачі, що які необхідно вирішити в ході дослідження: задачу кластеризації.
2. Проаналізовано сучасний стан систем інтелектуального аналізу. Виділено найбільш відомі зарубіжні системи: Statistica, Weka, RapidMiner. Здійснено аналіз даних систем в контексті задачі кластеризації даних пацієнтів.
3. Виявлено необхідність в написанні програмного забезпечення, яке даватиме можливість аналізувати правильність отриманих результатів кластеризації.

### 3 ФОРМУВАННЯ ВХІДНОЇ ВИБІРКИ ДАНИХ

На теперішній час є доволі багато безкоштовних та вільно доступних БД в мережі Інтернет. Один із найбільших у наш час - це Machine Learning Repository. Machine Learning Repository - це репозиторій, де зберігається близько 311 БД з даними різного напрямку. Існує можливість завантажувати власні файли БД до репозиторію, які в подальшому будуть доступні користувачам репозиторію.

#### 3.1 Опис атрибутів кластеризації

Для проведення дослідження було обрано наступну вибірку: БД пацієнтів післяопераційного стану[10]. БД складається із 88 об'єктів та має наступні атрибути:

- **L-CORE** (внутрішня температура пацієнта):
  - висока ( $> 37$ )
  - середня ( $\geq 36$  and  $\leq 37$ )
  - низька ( $< 36$ )
- **L-SURF** (температура тіла пацієнта):
  - висока ( $> 36.5$ )
  - середня ( $\geq 36.5$  та  $\leq 35$ )
  - низька ( $< 35$ )
- **L-O2** (насичення киснем у %):
  - чудове ( $\geq 98$ )
  - добре ( $\geq 90$  та  $< 98$ )
  - середнє ( $\geq 80$  та  $< 90$ )
  - мале ( $< 80$ )
- **L-BP** (останнє вимірювання кров'яного тиску):
  - високий ( $> 130/90$ )
  - середній ( $\leq 130/90$  та  $\geq 90/70$ )
  - низький ( $< 90/70$ )

- **SURF-STBL** (стабільність температури тіла пацієнта):
  - стабільна
  - середньо стабільна
  - нестабільна
- **CORE-STBL** (стабільність внутрішньої температури пацієнта)
  - стабільна
  - середньо стабільна
  - нестабільна
- **BP-STBL** (стабільність кров'яного тиску)
  - стабільна
  - середньо стабільна
  - нестабільна
- **COMFORT** (коефіцієнт комфорту пацієнта)
  - число 0-20
- **ADM-DECS** (рішення) – даний атрибут показує приналежність об'єкту до однієї із категорій:
  - I (пацієнт відправляється у відділення інтенсивної терапії)
  - S (пацієнт відправляється додому)
  - A (пацієнт направляється на огляд)

#### 4 ОГЛЯД МЕТОДІВ КЛАСТЕРИЗАЦІЇ

Кластеризація – це розбиття елементів деякої множини (об’єкти, дані, вектора характеристик) на групи (кластери) за деяким принципом схожості [11]. Кластерний аналіз – це сукупність методів, що дозволяють класифікувати багатовимірні спостереження, кожне з яких описується набором вихідних змінних  $X_1, X_2 \dots X_n$  [12]. Метою кластерного аналізу є утворення груп схожих між собою об’єктів, які прийнято називати кластерами.

Методи кластерного аналізу дозволяють вирішувати наступні завдання:

- проведення класифікації об’єктів з урахуванням ознак, що віддзеркалюють сутність, природу об’єктів;
- побудова нових класифікацій для маловивчених явищ, коли необхідно встановити наявність зв’язків усередині сукупності та спробувати привнести до неї структуру.

В задачі кластерного аналізу формою подання вихідних даних служить прямокутна таблиця, кожний рядок якої представляє результат виміру  $k$ -ознак, що розглядаються, по одному з об’єктів:

$$\begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \quad (4.1)$$

Кластеризація включає в себе наступні етапи, які зображені на рис. 4.1:

- виділення характеристик;
- визначення метрики;
- розбиття об’єктів на групи;
- представлення результатів.



Рис. 4.1. Загальна схема кластеризації

На рис. 4.2 представлена класифікація алгоритмів та методів кластерного аналізу.

#### 4.1 Ієрархічні методи

Сутність ієрархічних агломеративних методів у тому, що на першому кроці кожний об'єкт вибірки розглядається як окремий кластер. Процес об'єднання кластерів відбувається послідовно: на підставі матриці відстаней або матриці подібності поєднуються найбільш близькі об'єкти. Послідовність об'єднання легко піддається геометричній інтерпретації й може бути представлена у вигляді графа-дерева (дендрограми) [13 – 15].

Основною передумовою ієрархічних дивізімних методів є те, що спочатку всі об'єкти належать до одного кластеру. У процесі класифікації за певними правилами поступово від цього кластера відділяються групи схожих між собою об'єктів. Таким чином, на кожному кроці кількість кластерів зростає, а міра відстані між кластерами зменшується.

При використанні ієрархічних методів існує можливість досить легко ідентифікувати викиди в наборі даних і, в результаті, підвищити якість даних.

Велика кількість методів ієрархічного кластерного аналізу різняться не тільки використовуваними мірами подібності (розходження), але й методами об'єднання об'єктів у кластери.

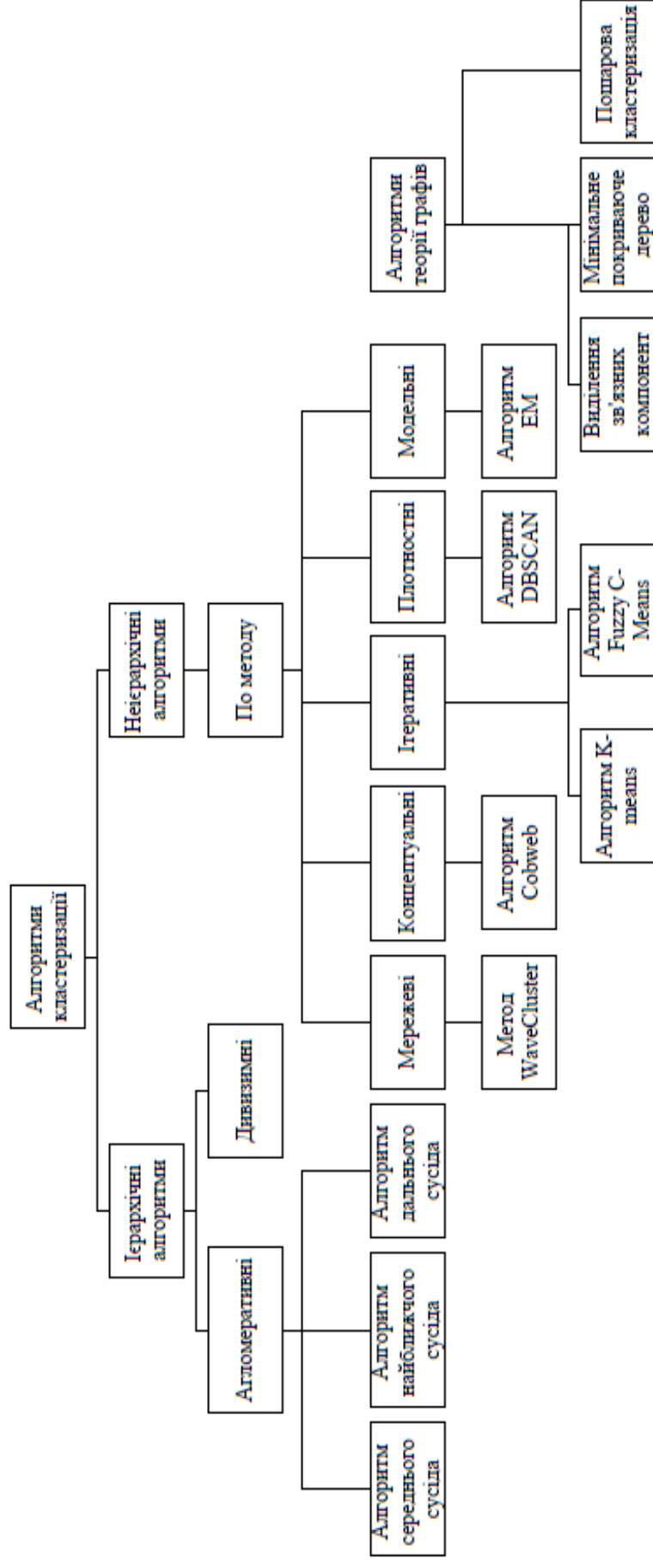


Рис.4.2. Класифікація методів кластерного аналізу

Розглянемо принцип роботи групи ієрархічних методів рис. 4.3.

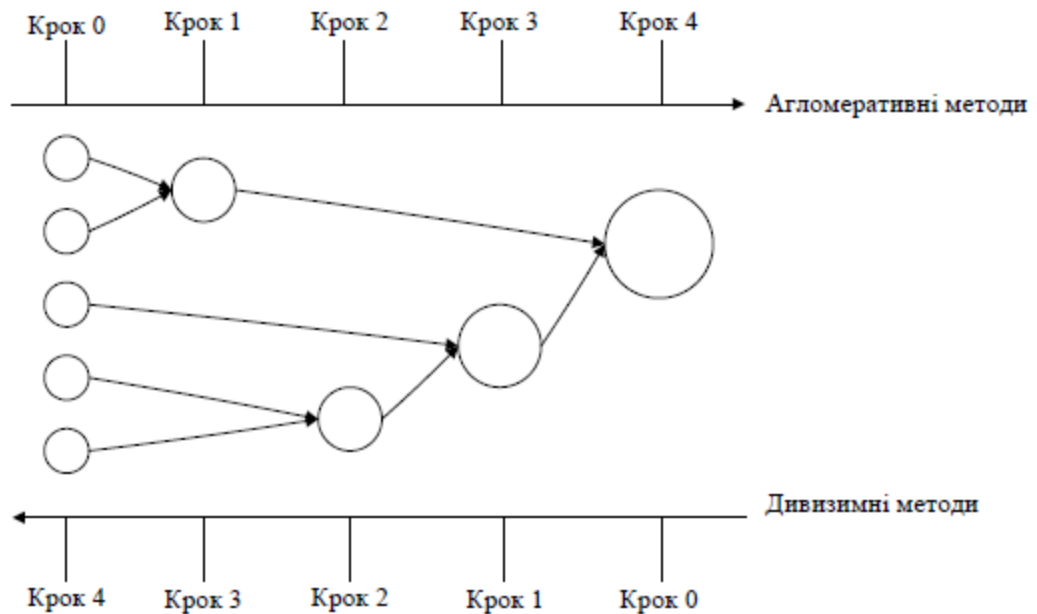


Рис.4.3 Принцип роботи ієрархічних методів

#### 4.1.1 Правила об'єднання кластерів

На першому кроці, коли кожний об'єкт являє собою окремий кластер, відстані між цими об'єктами визначаються обраною мірою. Однак коли зв'язуються разом кілька об'єктів, виникає питання, як слід визначити відстані між кластерами. Іншими словами, необхідно правило об'єднання або зв'язку для двох кластерів.

Існує декілька методів об'єднання кластерів [3]:

- 1) одиночний зв'язок (метод найближчого сусіда). У цьому методі відстань між двома кластерами визначається відстанню між двома найбільш близькими об'єктами (найближчими сусідами) у різних кластерах (рис. 4.4).

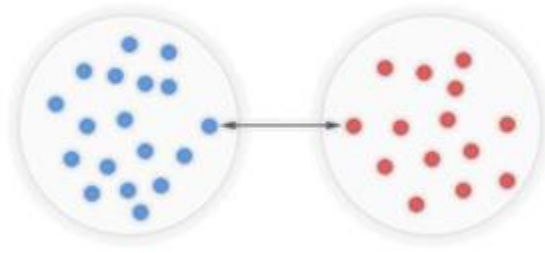


Рис. 4.4. Метод найближчого сусіда

Переваги методу:

- простота використання отриманих результатів;
- рішення не унікальні для конкретної ситуації, можливо їх використання для інших випадків;
- метою пошуку є не гарантовано вірне рішення, а краще з можливих.

Недоліки методу:

- даний метод не створює будь-яких моделей або правил, узагальнюючих попередній досвід, у виборі рішення вони ґрунтуються на всьому масиві доступних історичних даних, тому неможливо сказати, на якій підставі будуються відповіді;
- складність вибору міри «близькості» (метрики). Від цієї міри головним чином залежить обсяг множини записів, які потрібно зберігати в пам'яті для досягнення задовільної класифікації або прогнозу. Також існує висока залежність результатів класифікації від обраної метрики;
- при використанні методу виникає необхідність повного перебору навчальної вибірки при розпізнаванні, наслідок цього – обчислювальна трудомісткість;
- типові завдання даного методу – це завдання невеликої розмірності за кількістю класів і змінних.

2) повний зв'язок (метод найбільш віддалених сусідів). У цьому методі відстані між кластерами визначаються найбільшою відстанню між будь-якими двома об'єктами в різних кластерах (рис. 2.6).

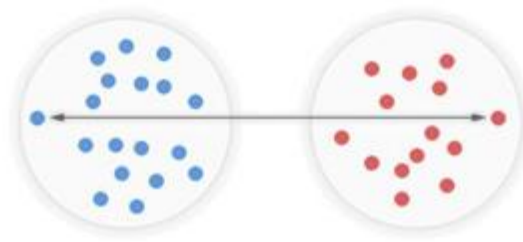


Рис. 4.5. Метод найвіддаленішого сусіда

Переваги методу:

- формуються відносно компактні гіперсферичні кластери, що складаються з об'єктів з великою схожістю.

Недоліки методу:

- непридатність методу в разі подовжених форм кластерів.

3) незважене попарне середнє. У цьому методі відстань між двома різними кластерами обчислюється як середня відстань між усіма парами об'єктів в них.

Переваги методу:

- метод ефективний, коли об'єкти формують кластери різних розмірів, проте він працює однаково добре і у випадках протяжних («ланцюгового» типу) кластерів.

4) виважене попарне середнє. Метод ідентичний до методу не виваженого попарного середнього, за винятком того, що при обчисленнях розмір відповідних кластерів (тобто число об'єктів, що містяться в них) використовується в якості вагового коефіцієнта. Тому запропонований метод повинен бути використаний, коли передбачаються нерівні розміри кластерів.

Переваги методу:

- ефективний у разі коли передбачаються нерівні розміри кластерів.

5) невважений центроїдний метод (рис. 4.6). У цьому методі відстань між двома кластерами визначається як відстань між їхніми центрами тяжіння.

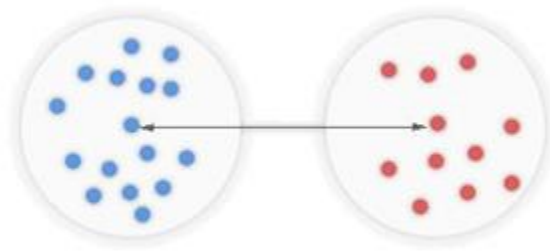


Рис.4.6. Невиважений центроїд

Недоліки методу:

- у разі об'єднання двох кластерів, в одному з яких багато елементів, а в іншому – мало, характеристики останнього практично ігноруються.
- 6) зважений центроїдний метод (медіана). Цей метод ідентичний попередньому, за винятком того, що при обчисленнях використовуються ваги для обліку різниці між розмірами кластерів (тобто числами об'єктів в них). Тому, якщо є (або підозрюються) значні відмінності в розмірах кластерів, варто віддати перевагу цьому методу.

Переваги методу:

- ефективний якщо є значні відмінності в розмірах кластерів.
- 7) метод Варда (рис. 4.7). Цей метод відрізняється від усіх інших методів, оскільки він використовує методи дисперсійного аналізу для оцінки відстаней між кластерами. Метод мінімізує суму квадратів (SS) для будь-яких двох (гіпотетичних) кластерів, які можуть бути сформовані на кожному кроці.

При використанні методу Варда дендрограма виходить з глибоко розділеними, більш «виразними» та найбільш однорідними в статистичному сенсі кластерами. У цілому метод видається дуже ефективним. При використанні цього методу створюються кластери приблизно рівних розмірів і які мають форму гіперсфер, що відповідає критерію якості кластеризації.

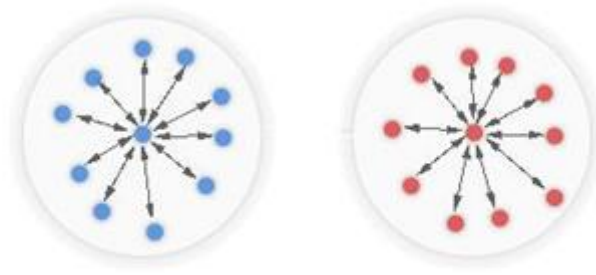


Рис.4.7. Метод Варда

Особливості методу:

- при використанні цього методу прагнення створюються кластери малого розміру.

#### 4.2 Алгоритм К-середніх

Найбільш поширений серед неієрархічних методів - алгоритм к-середніх. Для використання цього методу необхідно мати заздалегідь відому кількість кластерів[2].

Алгоритм к-середніх будує  $k$  кластерів, розташованих на можливо великих відстанях один від одного. Загальна ідея алгоритму: задане фіксоване число  $k$  кластерів спостереження зіставляються кластерам так, що середні в кластері (для всіх змінних) максимально можливо відрізняються один від одного.

Нехай є  $n$  спостережень, кожне з яких характеризується  $p$  ознаками  $x_1, x_2 \dots x_n$ . Ці спостереження необхідно розбити на  $k$  кластерів:

- з  $n$  точок досліджуваної сукупності відбираються випадковим чином або задаються дослідником виходячи з яких-небудь апріорних міркувань  $k$  точок (об'єктів) (ці точки приймаються за еталони);
- кожному еталону привласнюється порядковий номер, що є одночасно номером кластера;
- із  $n - k$  об'єктів, що залишилися, вибирається точка  $X_i$  з координатами  $(x_{i2}, \dots, x_{ip})$  і перевіряється, до якого з еталонів

- (центрів) вона знаходиться ближче всього. Для цього використовується одна з метрик;
- об'єкт, що перевіряється, приєднується до того центра (еталону), якому відповідає  $d_{ij}$ , ( $i = 1 \dots k$ );
  - еталон замінюється новим(перерахованим) з урахуванням приєднаної точки, і вага його (кількість об'єктів, що входять до даного кластера) збільшується на одиницю. Якщо зустрічаються дві або більше мінімальних відстаней, то  $i$ -й об'єкт приєднують до центра з найменшим порядковим номером;
  - вибирається наступна точка  $X_i$  і для неї повторюються всі процедури.

Таким чином, через  $(n - k)$  кроків всі точки (об'єкти) сукупності виявляються віднесеними до одного з  $k$  кластерів, але на цьому процес розбивки не закінчується. Для того, щоб домогтися стабільності розбивки по тому ж правилу, всі точки  $x_1, x_2 \dots x_n$  знову приєднуються до отриманих кластерів, при цьому ваги продовжують накопичуватися. Нова розбивка порівнюється з попередньою. Якщо вони збігаються, то робота алгоритму завершується, інакше цикл повторюється. Остаточна розбивка має центри ваги, які не збігаються з еталонами, їх можна позначити  $c_1, c_2, \dots c_k$ . При цьому кожна точка  $x_i$ ,  $i = 1..n$  буде відноситися до такого кластера (класу), для якого відстань мінімальна. Робота алгоритму показана на рис.4.8.

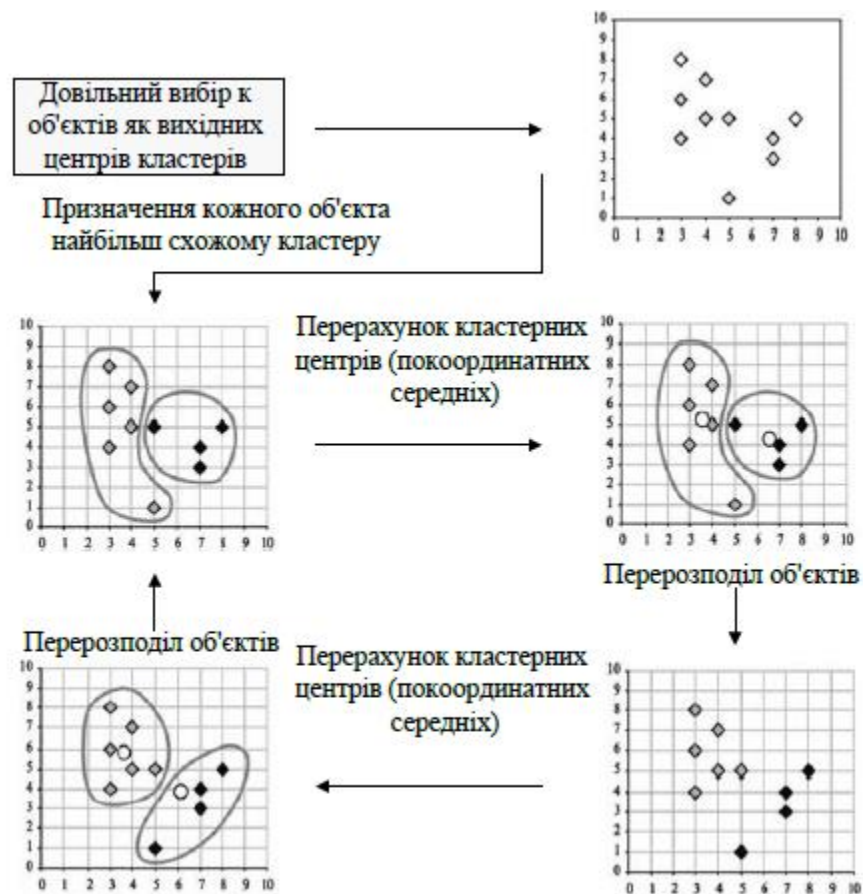


Рис. 4.8. Робота алгоритму  $k$ -середніх

Переваги алгоритму  $k$ -середніх:

- простота використання;
- швидкість використання;
- зрозумілість і прозорість алгоритму.

Недоліки алгоритму  $k$ -середніх:

- алгоритм дуже чутливий до викидів, які можуть спотворювати середнє.
- алгоритм може повільно працювати на великих базах даних.

Можливим рішенням цієї проблеми є використання вибірки даних.

### 4.3 ЕМ алгоритм

ЕМ алгоритм, що використовується в математичній статистиці для знаходження оцінок максимальної схожості параметрів ймовірних моделей, у випадку, коли модель залежить від деяких прихованих змінних. Кожна

ітерація алгоритму складається з двох кроків. На Е-кроці (expectation) вираховується очікуване значення функції правдоподібності, при цьому приховані змінні розглядаються як спостережувані. На М-кроці (maximization) вираховується оцінка максимальної схожості, таким чином збільшується очікувана схожість, вирахована на Е-кроку. Потім це значення використовується для Е-кроку на наступній ітерації. Алгоритм виконується до того часу, поки не буде виконана умова збіжності[24].

Загальна ідея алгоритму: нехай  $X$  — деяке з значень спостережуваних змінних, а  $T$  — приховані змінні. Разом  $X$  і  $T$  утворюють повний набір даних. Взагалі,  $T$  може бути деякою підказкою, яка полегшує рішення проблеми у випадку, якщо вона відома. Наприклад, якщо є суміш розподілів, то функція правдоподібності легко виражається через параметри відокремлених розподілів суміші.

Покладемо  $p$ - густину імовірності (в безперервному випадку) або функція ймовірностей (в дискретному випадку) повного набору даних з параметрами  $\theta$ :  $p(X, T|\theta)$ . Цю функцію можна розуміти як правдоподібність всієї моделі, якщо розглядати її як функцію параметрів  $\theta$ . Зауважимо, що умовний розподіл прихованої компоненти при деякому спостереженні та фіксованому наборі параметрів може бути виражене так (використовуючи розширену формулу Байеса і формулу повної ймовірності):

$$p(T|X, \theta) = \frac{p(X, T|\theta)}{p(X|\theta)} = \frac{p(X|T, \theta)p(T|\theta)}{\int p(X, \hat{T}|\theta)p(\hat{T}|\theta)d\hat{T}} \quad (4.1)$$

Таким чином, нам необхідно знати тільки розподіл спостережуваної компоненти при фіксованій прихованій  $p(X|T, \theta)$  і ймовірності прихованих даних  $p(T|\theta)$ .

ЕМ-алгоритм ітеративно покращує початкову оцінку  $\theta_0$ , обчислюючи нові значення оцінок  $\theta_1, \theta_2$  і так далі. На кожному кроці перехід від попереднього до наступного виконується таким чином:

$$Q(\theta) = E_T[\log p(X, T|\theta)|X] \quad (4.2)$$

де  $Q(\theta)$  — математичне очікування логарифма правдоподібності. Іншими словами, ми не можемо відразу обчислити точну правдоподібність, але за відомими даними  $X$  ми можемо знайти апостеріорну оцінку ймовірностей для різних значень прихованих змінних  $T$ . Для кожного набору значень  $T$  і параметрів  $\theta$  ми можемо обчислити математичне очікування функції правдоподібності з даного набору  $X$ . Воно залежить від попереднього значення  $\theta$ , бо це значення впливає на ймовірності прихованих змінних  $T$ .

$Q(\theta)$  — обчислюється наступним чином:

$$\theta_{n+1} = \operatorname{argmax} Q(\theta) \quad (4.3)$$

іншими словами, умовне математичне сподівання  $\log p(X, T|\theta)$  при умові  $\theta$ .

Інакше кажучи,  $\theta_{n+1}$  - це значення, максимізуючи (М) умовне математичне сподівання (Е) логарифма правдоподібності при даних значеннях спостережуваних змінних і попередньому значенні параметрів.

За певних обставин зручно розглядати ЕМ-алгоритм як два кроки максимізації, що чергуються.

Кроки ЕМ-алгоритму можна показати як:

**E(xpectation) крок:** Вибираємо  $q$ , щоб максимізувати  $F$ :

$$q^t = \operatorname{argmax} F(q, \theta^t) \quad (4.4)$$

**M(aximization) крок:** Вибираємо  $\theta$ , щоб максимізувати  $F$ :

$$\theta^{t+1} = \operatorname{argmax} F(q^t, \theta) \quad (4.5)$$

#### 4.4 Алгоритм DBSCAN

Алгоритм DBSCAN - [25] це алгоритм кластеризації, оснований на понятті щільності елементів. Він знаходить кількість кластерів, починаючи з деякою розрахунковою щільності розподілу відповідних вузлів. DBSCAN є одним з найбільш поширених алгоритмів кластеризації, а також найбільш цитованих в науковій літературі.

Ідея, покладена в основу алгоритму, полягає в тому, що всередині кожного кластера спостерігається типова щільність точок (об'єктів), яка помітно вище, ніж щільність зовні кластера, а також щільність в областях з шумом нижче щільності будь-якого з кластерів. Ще точніше, що для кожної точки кластера її сусідство заданого радіуса повинно містити не менше деякого числа точок, це число точок задається граничним значенням. Для детального викладу принципових особливостей алгоритму необхідно ввести ряд визначень - ознайомитися з якими більш докладно можна в [26].

Вхідні параметри алгоритму: *eps* – окіл, *D* – набір даних, *MinPts* – мінімальне число точок для формування щільнісного регіону. На рис.4.9. приведено реалізацію алгоритму:

```

DBSCAN(D, eps, MinPts)
    C = 0
    for each unvisited point P in dataset D
        mark P as visited
        NeighborPts = regionQuery(P, eps)
        if sizeof(NeighborPts) < MinPts
            mark P as NOISE
        else
            C = next cluster
            expandCluster(P, NeighborPts, C, eps, MinPts)

expandCluster(P, NeighborPts, C, eps, MinPts)
    add P to cluster C
    for each point P' in NeighborPts
        if P' is not visited
            mark P' as visited
            NeighborPts' = regionQuery(P', eps)
            if sizeof(NeighborPts') >= MinPts
                NeighborPts = NeighborPts joined with NeighborPts'
    if P' is not yet member of any cluster
        add P' to cluster C

regionQuery(P, eps)
    return all points within P's eps-neighborhood (including P)

```

Рис.4.9. Реалізація алгоритму DBSCAN

### Переваги

- не потрібно вказувати кількість кластерів(хоча для даного дослідження це не є перевагою);
- форми кластерів мають довільну форму;
- алгоритм стійкий до шумів;
- нечутливий до впорядкованості точок в БД.

### Недоліки

- алгоритм погано працює, коли кластери мають різну щільність;
- ускладнюється робота з даними, коли дані багатовимірні, оскільки процес визначення щільності стає набагато складніше;
- використання DBSCAN може бути затратним, коли для обчислення найближчих сусідів знадобиться обчислення всіх попарних відстаней, як це зазвичай буває для даних великої розмірності.

## 4.5 Алгоритм COBWEB

Алгоритм COBWEB [27] – класичний метод інкрементальної кластеризації. Він створює ієрархічну кластеризацію у вигляді дерева класифікації: кожен вузол цього дерева посилається на концепт і містить імовірнісний опис цього концепту, яке включає в себе ймовірність приналежності концепту до даного вузла і умовні ймовірності виду:

$$P(A_i = v_{ij} | C_k), \quad (4.6)$$

де  $A_i = v_{ij}$  - пара атрибут-значення,  $C_k$  - клас концепту.

Вузли, що знаходиться на певному рівні дерева класифікації, називають зрізом. Алгоритм використовує для побудови дерева класифікації евристичну міру оцінки, звану корисністю категорії - приріст очікуваного числа коректних припущень про значення атрибутів при знанні про їх належність до певної категорії щодо очікуваного числа коректних припущень про значення атрибутів без цього знання. Щоб додати новий об'єкт в дерево класифікації, алгоритм COBWEB ітеративно проходить все дерево в пошуках «кращого» вузла, до якого віднести цей об'єкт. Вибір вузла здійснюється на основі переміщення об'єкта в кожен вузол і обчислення корисності категорії отриманого зрізу. Також обчислюється корисність категорії для випадку, коли об'єкт відноситься до новостворюваного вузла. У підсумку об'єкт відноситься до того вузла, для якого корисність категорії більше.

### Недоліки

- припускається, що розподіл ймовірностей значень різних атрибутів статистично незалежні один від одного. Однак це припущення не завжди вірно, тому як часто між значеннями атрибутів існує кореляція;
- імовірнісний розподіл кластерів робить дуже складним їх оновлення, особливо в тому випадку, коли атрибути мають велике число можливих значень. Це викликано тим, що складність алгоритму

залежить не тільки від кількості атрибутів, але і від кількості їхніх можливих значень.

#### 4.6 Міри подібності

Для проведення кластеризації необхідно ввести поняття подібності об'єктів по спостережуваним змінним. У кожний кластер повинні потрапити об'єкти, що мають подібні характеристики.

Ключовим моментом в кластерному аналізі даних вважається вибір метрики (або міри близькості об'єктів). Подібність або розходження між об'єктами класифікації встановлюється в залежності від обраної метричної відстані між ними. Якщо кожен об'єкт описується  $i$  властивостями (ознаками), то він може бути представлений як точка в  $i$ -мірному просторі, і схожість з іншими об'єктами буде визначатися як відповідна відстань. Розглянута задача кластеризації зводиться до задачі визначення функції близькості між об'єктами класів – вибору міри відстані між об'єктами.

Використовуються різні міри відстані [28-31]. Розглянемо деякі з них.

##### 4.6.1 Евклідова відстань

Найбільш поширена відстань. Вона є географічною відстанню в багатовимірному просторі і обчислюється наступним чином:

$$d(X, Y) = \sqrt{\sum_{i=1}^m (X_i - Y_i)^2}, \quad (4.7)$$

де  $d(X, Y)$  – відстань між об'єктами  $X, Y$ ;

$X_i$  - значення  $i$ - властивості об'єкта  $X$ ;

$Y_i$  - значення  $i$ - властивості об'єкта  $Y$ .

З геометричної точки зору, евклідова міра відстані може виявитися безглуздою, якщо ознаки виміряні в різних одиницях. Щоб виправити становище, вдаються до нормування кожної ознаки.

Застосування евклідової відстані у якості міри виправдано в наступних випадках [32]:

- властивості (ознаки) об'єкта однорідні за фізичним змістом і однаково важливі для класифікації;
- простір ознак збігається з геометричним простором.

#### 4.6.2 Квадрат евклідової відстані.

Дана міра відстані використовується в тих випадках, коли потрібно надати більше значення більш віддаленим один від одного об'єктам. Це відстань обчислюється таким чином:

$$d(X, Y) = \sum_{i=1}^m (X_i - Y_i) \quad (4.8)$$

#### 4.6.3 Виважена Евклідова відстань

Застосовується в тих випадках, коли кожній  $i$ - властивості вдається приписати певну «вагу», пропорційно ступеню важливості ознаки в задачі кластеризації

$$d(X, Y) = \sum_{i=1}^m w_i (X_i - Y_i), \quad (4.9)$$

Визначення ваг, як правило, пов'язано з додатковими дослідженнями, наприклад, організацією опитування експертів і обробкою їх думок.

#### 4.6.4 Хеммінгова відстань

Також називається Манхеттенською, сіті-блок відстані або відстань міських кварталів. Це відстань є різницею по координатах. У більшості випадків ця міра відстані призводить до таких же результатів, як і для

звичайної відстані Евкліда. Проте вплив окремих великих різниць (викидів) зменшується (так як вони не зводяться в квадрат). Хеммінгова відстань обчислюється за формулою:

$$d(X, Y) = \sum_{i=1}^m (|X_i| - |Y_i|) \quad (4.10)$$

#### 4.6.5 Відстань Чебишева

Цю відстань використовують, коли необхідно визначити два об'єкти як «різні», якщо вони різняться з якої-небудь однієї координати (яким-небудь одним виміром). Приймає значення найбільшого модуля різниці між значеннями відповідних властивостей (ознак) об'єктів:

$$d(X, Y) = \max |X_i - Y_i| \quad (4.11)$$

#### 4.6.6 Показникова відстань

Для прогресивного збільшення або зменшення ваг, що відносяться до розмірності, для якої відповідні об'єкти сильно відрізняються. Це може бути досягнуто з використанням показникової відстані:

$$d(X, Y) = \left( \sum_{i=1}^m |X_i - Y_i|^p \right)^{1/p} \quad (4.12)$$

#### 4.6.7 Відстань Канберра

Канберрова відстань лежить між 0 і 1, але володіє нечутливістю до сильно асиметричних даними. Однак застосування цього методу при наявності нульових і негативних значень небажано, тому рекомендується попереднє перетворення даних.

Формула для обчислення відстані Канберра:

$$d(X, Y) = \sum_{i=1}^m \frac{|X_i - Y_i|}{X_i + Y_i} \quad (4.13)$$

#### 4.6.8 Відстань Брея-Картіса

Відстань Брея-Картіса має значення між 0 і 1, проте змінні з великими значеннями надають більший вплив на результат:

$$d(X, Y) = \frac{\sum_{i=1}^m |X_i - Y_i|}{\sum_{i=1}^m X_i + \sum_{i=1}^m Y_i} \quad (4.14)$$

Вибір міри відстані і ваг – дуже важливий етап, тому що від цих процедур залежать склад і кількість формованих класів, а також ступінь подібності об'єктів всередині класів.

Оцінка подібності між об'єктами сильно залежить від абсолютного значення ознаки й від ступеня його варіації в сукупності. Щоб усунути подібний вплив на процедуру класифікації, можна значення вихідних змінних нормувати одним із наступних способів:

$$z_{ij} = \frac{x_{ij} - \bar{x}_{ij}}{\delta_j}, z_{ij} = \frac{x_{ij}}{x_{maxj}}, z_{ij} = \frac{x_{ij}}{\bar{x}_{ij}}, z_{ij} = \frac{x_{ij}}{x_{minj}} \quad (4.15)$$

Візьмемо за критерій оптимальності мінімальну суму квадратів відстаней до прямої від спостережуваних точок, обчислених вздовж осі координат (методом найменших квадратів). В якості міри можна використовувати мінімум суми квадратів відстаней від точок до прямої, обчислених уздовж осі абсцис; мінімум суми квадратів відстаней довжин перпендикулярів, опущених з точок на пряму і т.д.

#### 4.7 Вибір алгоритмів кластеризації, міри близькості та методики оцінки результату кластеризації для дослідження

**Вибір методів кластеризації.** В контексті задачі кластеризації медичних даних пацієнтів дослідник в області медицини може висунути припущення про те, що пацієнти в основному попадають в фіксовану кількість категорій. На прикладі даних післяопераційного стану пацієнтів [10] бачимо, що пацієнт після операції може перебувати у трьох визначених станах. З цього слідує, що варто розглядати ті алгоритми кластеризації, в яких вхідним параметром задається кількість кластерів, оскільки існує чітко визначена кількість видів(ступенів) тієї чи іншої хвороби.

Серед вище розглянутих алгоритмів, такими алгоритмами, які вхідним параметром очікують число кластерів є наступні:

- K-means алгоритм
- Ієрархічний алгоритми

**Вибір міри близькості.** Щодо вибору конкретної міри близькості - не можна однозначно сказати, яку саме використовувати для дослідження, оскільки при різних метриках можна отримати різні розбиття вибірки на кластери. Тому, вирішено провести розбиття на кластери використовуючи усі можливі метрики, і в результаті з'явиться можливість сказати яка з метрик є найкращою для даної задачі.

**Методика оцінки результату кластеризації.** Зазвичай, в області медицини, дані є багаторозмірними. Багаторозмірність даних спричиняє складнощі при оцінюванні якості роботи того чи іншого алгоритму кластеризації, оскільки не можливо оцінити наскільки точно алгоритм опрацював дані. У випадку коли дані дво або трьох- вимірні оцінити точність роботи алгоритму можна проаналізувавши розташування об'єктів та кластерів на координатній.

В рамках вирішення поставленої задачі вирішено взяти за основу для порівняння алгоритмів взятю той факт, що тестова вибірка містить атрибут,

який вказує на приналежність об'єкту до певного класу. Вибір такої вибірки обумовлений тим, що є можливість визначити точність роботи кожного із алгоритмів кластеризації шляхом порівняння еталонної вибірки та отриманої в результаті кластеризації[33].

Для оцінки правильності кластеризації кожного із алгоритмів скористаємось наступною методикою.

Нехай:

- $N_0$ - загальна кількість об'єктів
- $N_1$ - кількість вірно кластеризованих об'єктів

Тоді правильність кластеризації:

$$A(Accuracy) = \frac{N_1}{N_0} \quad (4.18)$$

## ВИСНОВКИ ДО РОЗДІЛУ

1. Здійснено огляд основних типів алгоритмів кластеризації. Проаналізовано методи кластеризації даних та їх вхідні параметри. Також розглянуто основні методи об'єднання об'єктів у кластери та міри близькості об'єктів.
2. Виділено переваги та недоліки кожного із методів. Обрано наступні методи: k-means та ієрархічний.
3. Здійснено аналіз існуючих мір близькості та методів об'єднання об'єктів у кластери.
4. Проведено оцінку результатів кластеризації. Для оцінки розбиття об'єктів на кластери використовується поняття правильності.

## 5 ОПИС ПРОГРАМНОГО ПРОДУКТУ

Під час проведення дослідження розроблено ПЗ, яке автоматизує процес кластеризації та класифікації при різних параметрах.

### 5.1 Засоби розробки

Для розробки ПЗ обрано мову програмування C#. C# - це мова програмування створена спеціально для роботи у середовищі Microsoft .NET Framework.

Мова C# була розроблена з урахуванням сильних і слабких особливостей інших мов, зокрема Java і C++. Специфікація мови C# була написана Андерсом Гейлсбергом, Скотом Вілтанутом та Пітером Гольде.

Ключові особливості мови C#:

- Компонентна орієнтованість
- Код зібраний воєдино (декларації і реалізації об'єднані разом)
- Уніфікована система типів і їх безпечність
- Автоматична і мануальна робота за пам'яттю
- Використання єдиної бібліотеки класів - CLR [34]

Для проектування системи кластеризації даних обрано використовувати бібліотеку Extreme Optimization Numerical Libraries for .NET.

#### 5.1.1 Бібліотека Extreme Optimization Numerical Libraries for .NET

Бібліотека Extreme Optimization Numerical Libraries for .NET[35] представляє собою набір засобів загального призначення, яка має математичні і статистичні класи та побудована для платформи Microsoft .NET. Дана бібліотека забезпечує першу повну платформу для технічних і статистичних розрахунків побудована для Microsoft .NET платформи версії 2.0 та вище. Вона поєднує в собі математичну бібліотеку, векторну і матричну бібліотеки, а також бібліотеку статистики. Підтримує розробку додатків на мовах програмування C#, Visual Basic .NET, F #, C ++ / CLI, IronPython або будь-яку іншу .NET Framework мову.

Підтримує широкий спектр статистичних методів, доступних для платформи .NET. Спираючись на бібліотеки з математики, вона забезпечує міцні і надійні інструменти для аналізу та моделювання даних.

Extreme Optimization Numerical Libraries for .NET робить доступними широкий спектр статистичних методів на платформі .NET. Ґрунтуючись на математичних бібліотеках, Extreme Optimization Numerical Libraries for .NET забезпечує стійкі і надійні інструменти для аналізу даних, симуляції та моделювання.

Статистична бібліотека Extreme Optimization Numerical Libraries for .NET містить класи для використання широкого кола статистичних методів. Основні класи статистичної бібліотеки наведені в табл.5.1:

Таблиця 5.1. Основні класи бібліотеки

| Простір імен                           | Опис                                                                                                                                                              |
|----------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Extreme.Statistics                     | Містить базові класи для статистичних об'єктів, такі як змінні, лінійні моделі, ANOVA таблиці, а також часто використовувані статистичні типи даних і виключення. |
| Extreme.Statistics.Distributions       | Містить класи, які реалізують розподіл ймовірностей.                                                                                                              |
| Extreme.Statistics.Multivariate        | Містить класи, що реалізують методи багатовимірного статистичного аналізу, в тому числі кластерного аналізу та аналізу головних компонент.                        |
| Extreme.Statistics.Random              | Містить класи для роботи з випадковими номерами квазі-випадкових послідовностей.                                                                                  |
| Extreme.Statistics.TimeSeries-Analysis | Містить класи для роботи з тимчасовими рядами.                                                                                                                    |

## 5.2 Вимоги до технічного забезпечення

Загальні вимоги до технічного забезпечення представлено у вигляді необхідної конфігурації для комп'ютера інженера.

Системні вимоги до комп'ютера користувача наведені в табл.5.2.

Таблиця 5.2. Вимоги до комп'ютера

|                                   |                       |
|-----------------------------------|-----------------------|
| Процесор                          | 1GHz та більше        |
| Операційна система                | Windows XP/Vista/7/8  |
| Оперативна пам'ять                | 128Mb RAM та більше   |
| Відеоадаптер і монітор            | VGA (800x600) та вище |
| Вільне місце на диску             | 1Mb мінімум           |
| Пристрій взаємодії з користувачем | Клавіатура і миша     |

## 5.3 Архітектура програмного забезпечення

В якості архітектури програмного забезпечення було обрано принцип монолітного програмного забезпечення.

До переваг монолітного програмного забезпечення слід віднести:

- виконання всіх операцій на комп'ютері користувача, відсутність необхідності тримати в активному стані сервер програмного забезпечення;
- простота розробки програмного забезпечення;
- невразливість до атак злочинців різних типів;
- система збереження завантаження даних надає можливість обміну та повторного використання;
- відпадає необхідність у авторизації користувачів програмного забезпечення.

### 5.3.1 Опис класів програмного забезпечення

Основним класом програмного забезпечення являється Program. Цей клас є нащадком базового класу Object. При своєму створенні він викликає клас ClusteringForm.

Клас ClusteringForm відповідає за графічний інтерфейс програмного продукту, а також виклик інших класів, що відповідають за отримання параметрів кластеризації.

Клас ClusteringObject відповідає за об'єкт кластеризації.

Клас ConvertingHelper - перетворення даних в потрібний формат.

Нумератор Distances - задання міри близькості (відстані).

Нумератор LinkageMethods відповідає за метод об'єднання об'єктів.

Клас FileHelper відповідає за роботу з файлами – збереження, завантаження.

### 5.3.2 Специфікація функцій розробленого програмного забезпечення

Базовим класом розробленого програмного забезпечення є клас Program. Цей клас містить лише конструктор, в якому відбувається виклик графічного інтерфейсу програмного засобу.

Клас ClusteringForm окрім конструктору містить наступні методи:

Form1\_Load – ініціалізація та завантаження компонент інтерфейсу.

performAnalys\_Click – виконує аналіз даних при певних обраних параметрах.

SaveResultToFile – виконує збереження в файл.

Також в цьому класі міститься декілька приватних методів, які реагують на певні дії користувача.

Таблиця 5.3. Члени класу ClusteringObject

| Назва функції             | Вхідні параметри  | Вихідні параметри | Опис функції   |
|---------------------------|-------------------|-------------------|----------------|
| public ClusteringObject() | int objectId, int | -                 | Створює об'єкт |

|                                                           |                                                                                                                |                                                   |                                                                                                                                               |
|-----------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|---------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------|
|                                                           | clusterName                                                                                                    |                                                   | кластеризації з вказаним унікальним ідентифікатором та ім'ям кластеру                                                                         |
| public static int<br>GetTrueClusteringObjects             | List<ClusteringObject><br>listOfEnteringObjects,<br>List<ClusteringObject><br>listOfClusteringObjects          | int count                                         | Повертає кількість вірно кластеризованих об'єктів                                                                                             |
| public<br>Dictionary<int,double><br>GetDistanceToClusters | int [] clusteringObject,<br>double[] clusterCenters,<br>Distances distance,<br>LinkageMethods<br>linkageMethod | Dictionary<int,<br>double>()<br>distancesToCenter | Дана функція обраховує та повертає словник, де у якості ключа виступає номер кластеру та у якості значення – відстань від об'єкта до кластера |
| public int<br>GetObjectBelonging                          | Dictionary<int,double><br>distancesToClusters                                                                  | int<br>clusterName                                | Функція повертає приналежність вхідного об'єкта до одного із кластерів.                                                                       |

Клас FileHelper має методи, що виконують роботу з файлами.

Метод GetObjectsBelongsFromCsvFile – приймає на вхід шлях до файлу, в якому містяться об'єкти в форматі CSV та повертає список об'єктів ClusteringObject.

Метод GetDataTableFromCsvFile - приймає на вхід шлях до файлу, в якому містяться об'єкти в форматі CSV та повертає дані у вигляді таблиці для представлення їх на головній формі.

Метод SaveFile – зберігає отримані результати за вказаним шляхом.

Лістинг програми наведений у додатку А.

#### 5.4 Опис роботи з програмним забезпеченням

Для роботи з даним програмним забезпеченням (далі – ПЗ) достатньо знань на рівні користувача ПК. ПЗ має інтуїтивно зрозумілий віконний інтерфейс.

Головне вікно ПЗ складається з трьох панелей. Перша панель містить вхідні дані, друга панель – містить реалізацію основної функціональності – кластеризацію, класифікацію, аналіз результатів кластеризації, третя панель – вибір параметрів кластеризації. На рис. 5.1 показано головне вікно програми.

Дані для роботи програми завантажуються з файлу на ПК (файл з розширенням .cvs).

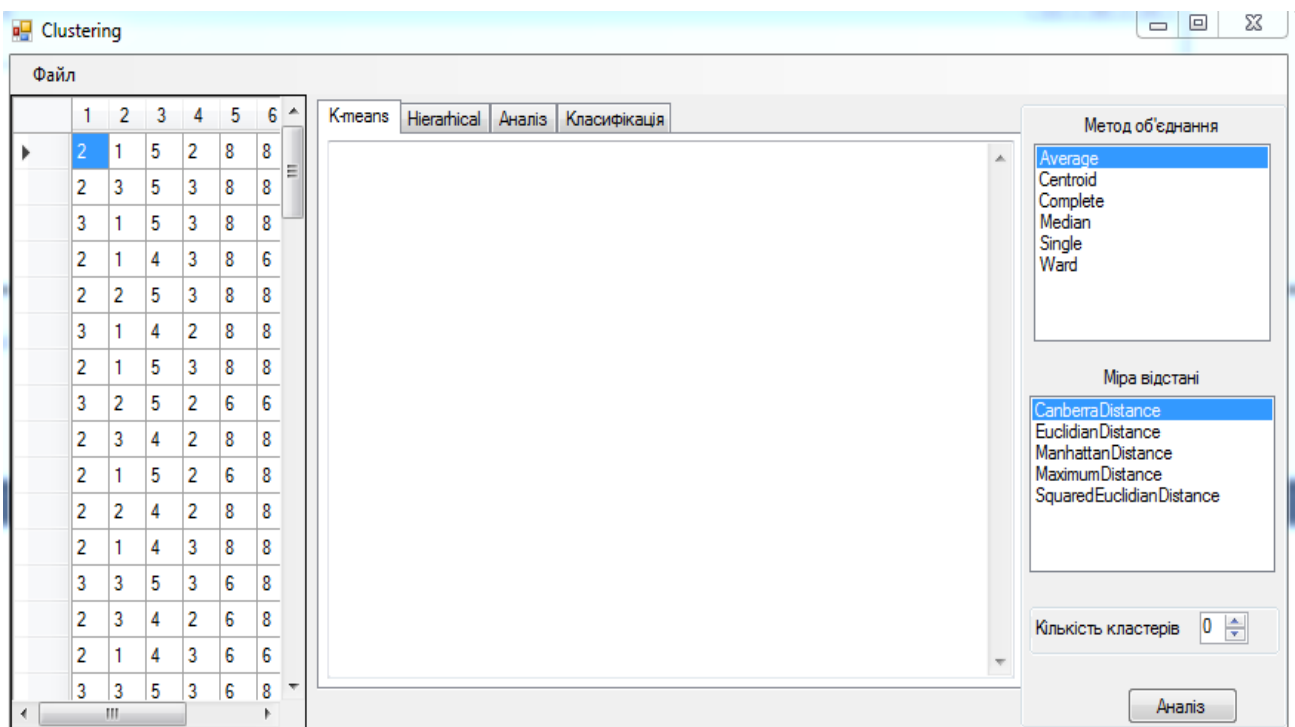


Рис. 5.1. Головне вікно програми

Після того, як дані завантажено та обрано параметри при яких необхідно провести кластеризацію та класифікацію на вкладках «K-means» та «Hierarchical» можна подивитись результат кластеризації по кожному об'єкту, а також на вкладці «Аналіз» можна переглянути порівняння роботи алгоритмів.

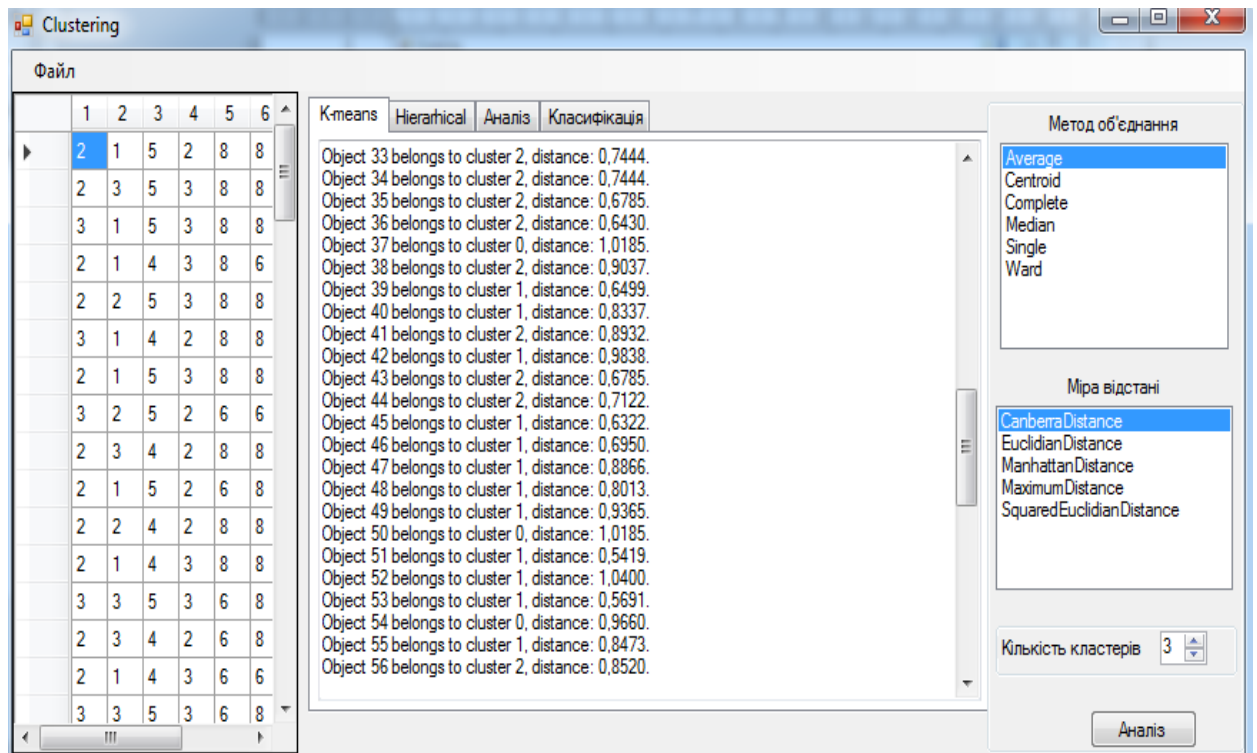


Рис. 5.2. Вигляд вкладки результату кластеризації

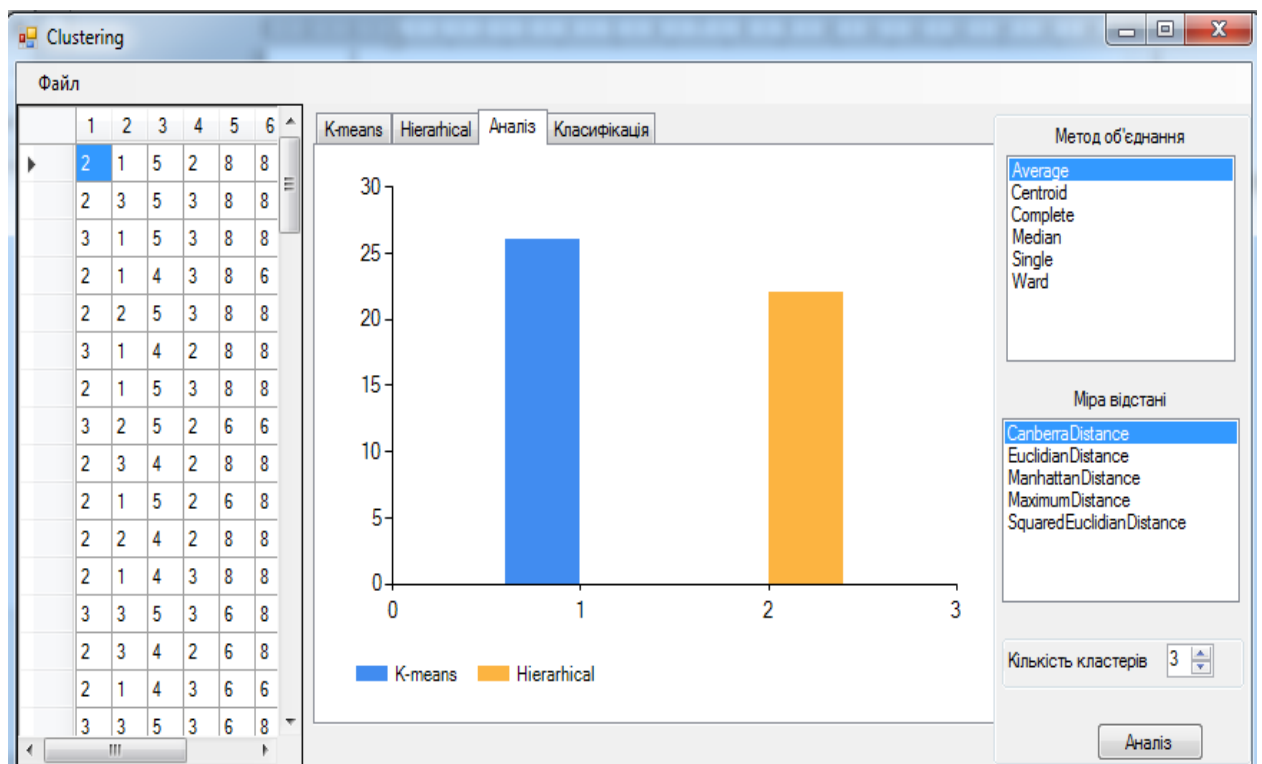


Рис. 5.3. Вигляд вкладки «Аналіз»

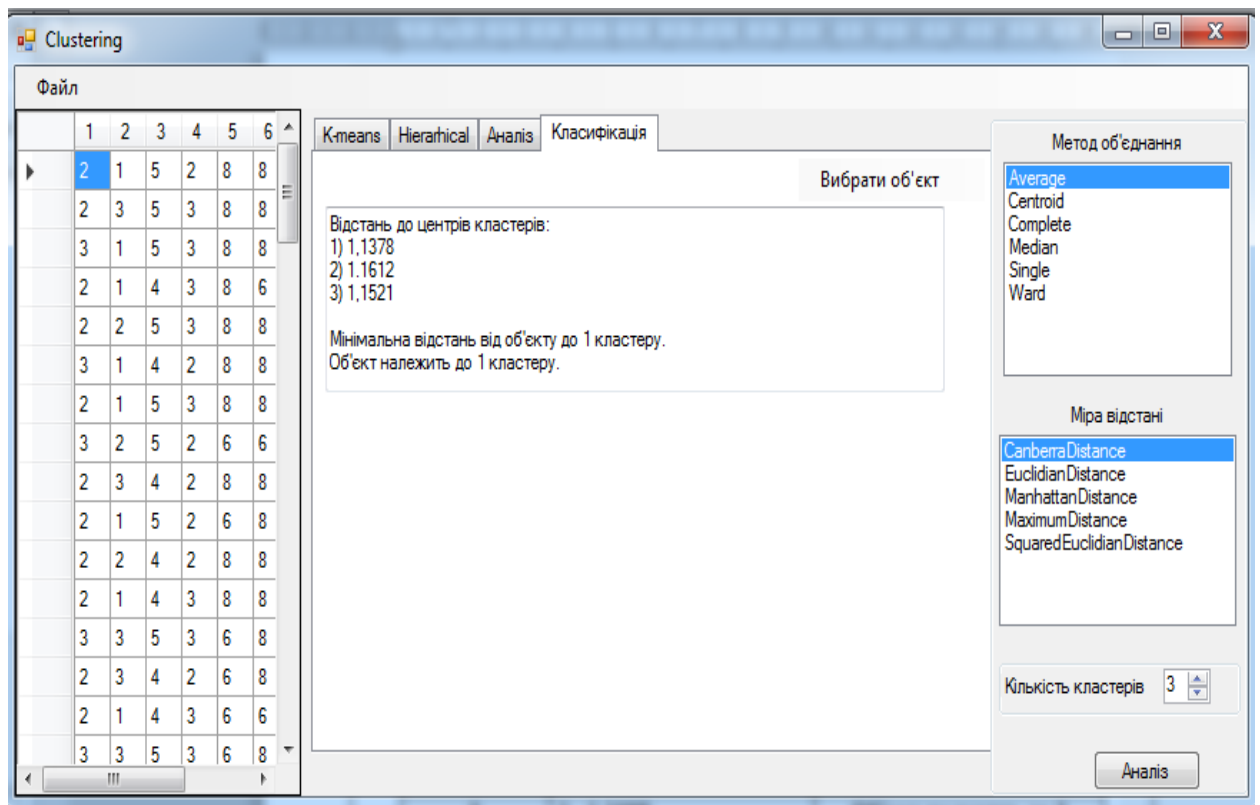


Рис. 5.4. Вигляд вкладки «Класифікація»

На вкладці «Класифікація» проводиться класифікація обраного об'єкту, який також завантажується з файлу.

## ВИСНОВКИ ДО РОЗДІЛУ

1. Описано основні можливості розробленого програмного забезпечення. Надано опис обраного засобу розробки, а саме мови програмування C#, вимог до технічного забезпечення та специфікація функцій.
2. Приведено опис роботи користувача з програмою.

## 6 ДОСЛІДЖЕННЯ ПРАВИЛЬНОСТІ РОБОТИ АЛГОРИМІВ

Для проведення дослідження розроблено програмне забезпечення, яке проводить кластеризацію по кожному із вище згаданих алгоритмів з певним набором вхідних параметрів. В рамках даного дослідження використано наступні загальні параметри для кожного із алгоритмів:

- розмір тестової вибірки – 88;
- кількість кластерів – 3.

Потрібно зазначити, що результат роботи кожного із алгоритмів залежить і від інших параметрів, наприклад, міри близькості. Розглянемо поділ на кластери тестової вибірки кожного із вище згаданих алгоритмів.

### 6.1 Алгоритм k-means

Особливістю даного алгоритму є те, що він чутливий до вибору міри близькості, тому результати кластеризації розглянуто в залежності від різних значень міри близькості.

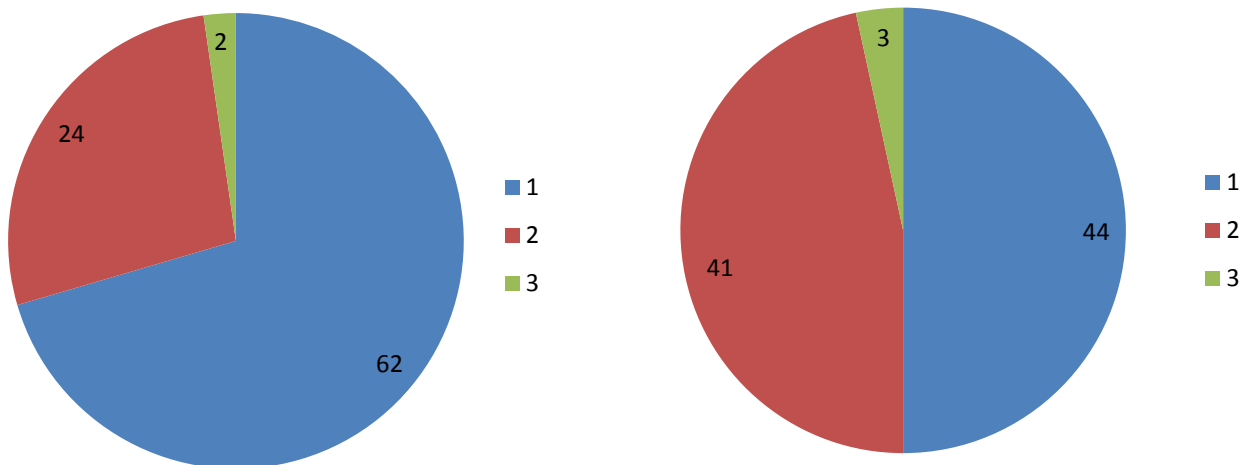


Рис.6.1. Порівняння результату кластеризації при виборі у якості міри близькості відстань Кенбера

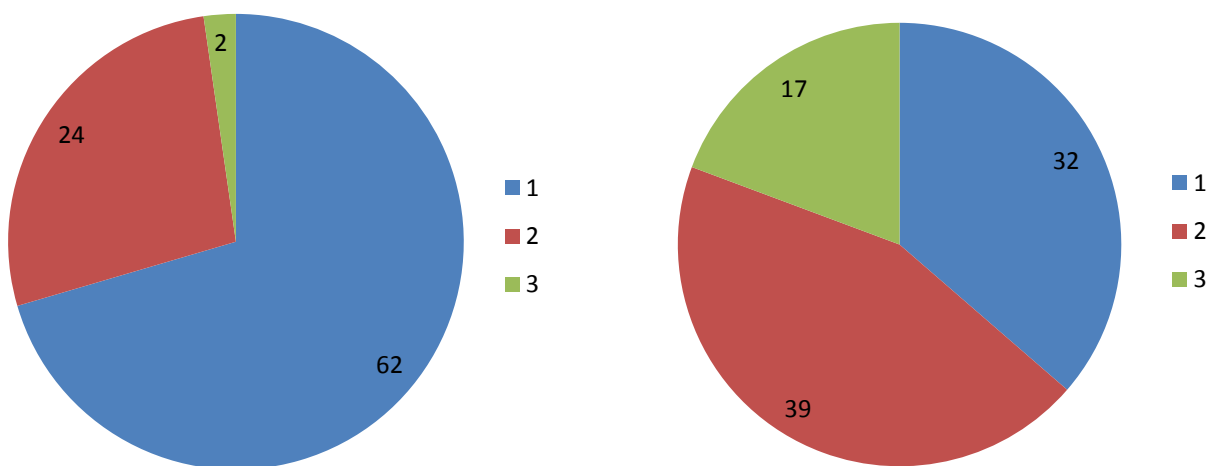


Рис.6.2. Порівняння результату кластеризації при виборі у якості міри близькості Евклідової відстані

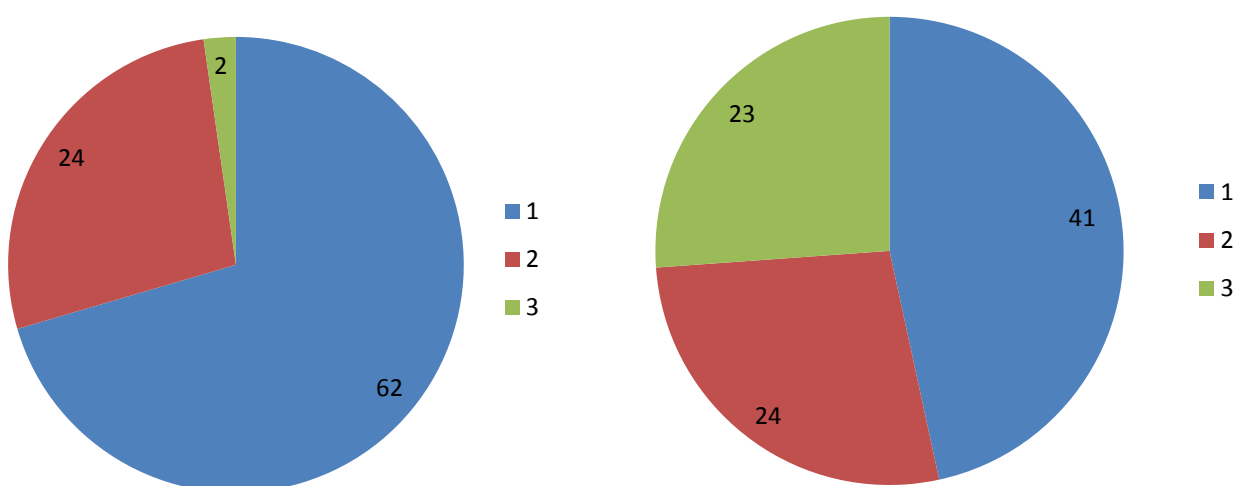


Рис.6.3. Порівняння результату кластеризації при виборі у якості міри близькості Хемінгова відстань

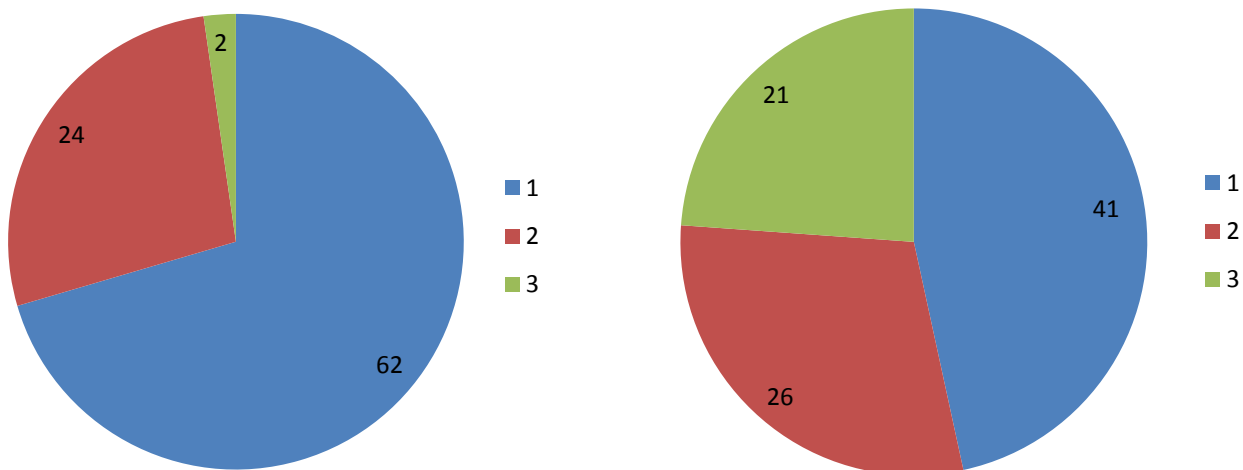


Рис.6.4. Порівняння результату кластеризації при виборі у якості міри близькості відстань Чебишева

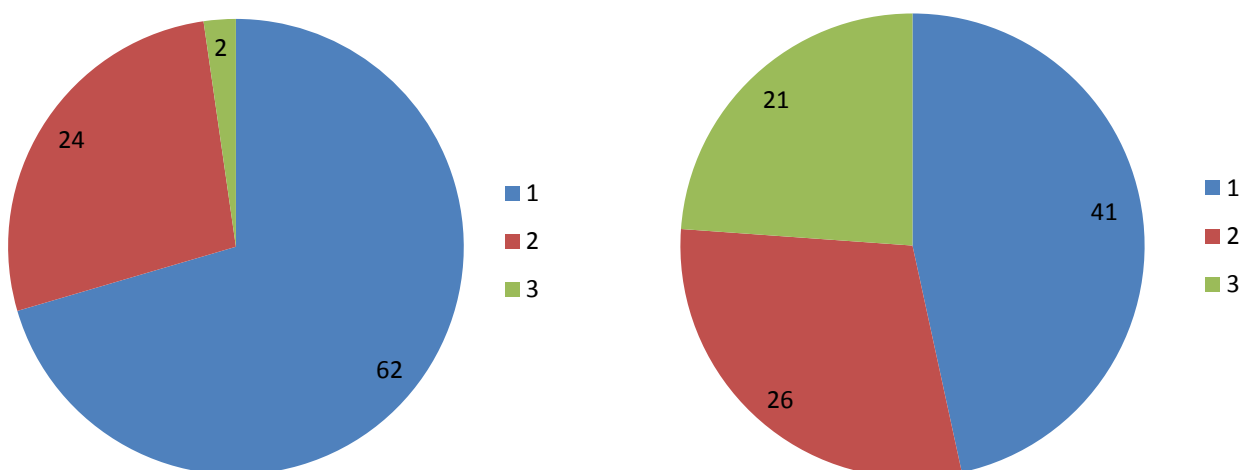


Рис.6.5. Порівняння результату кластеризації при виборі у якості міри близькості квадрат Евклідової відстані

Визначимо правильність роботи алгоритму для кожної з відстаней. Правильність по кожній метриці наведено в табл.6.1.

Таблиця 6.1. Правильність k - means

| Відстань                | КВКО | Правильність |
|-------------------------|------|--------------|
| Кенбера                 | 44   | 0,5          |
| Евклідова               | 36   | 0,41         |
| Хемінгова(Манхетенська) | 36   | 0,41         |
| Чебишева                | 42   | 0,47         |
| Квадрат Евклідової      | 36   | 0,41         |

З таблиці видно, що найточніший розподіл вибірки на кластери отримано з використанням у якості міри близькості – відстань Кенбера.

## 6.2 Ієрархічний алгоритм

Особливістю даного алгоритму в порівнянні з k-means є те, що він чутливий не тільки до вибору міри близькості, але і до методу з'єднання об'єктів у кластери. Оскільки попередній алгоритм розглянуто в контексті 5 різних мір близькості, і кластеризація з використанням даного алгоритму залежить від вибору одного з 6 методів з'єднання, то отримаємо 30 різних варіантів розбиття тестової вибірки на кластери.

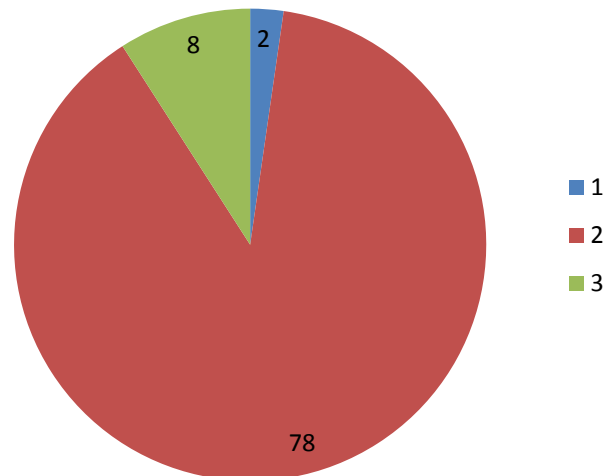
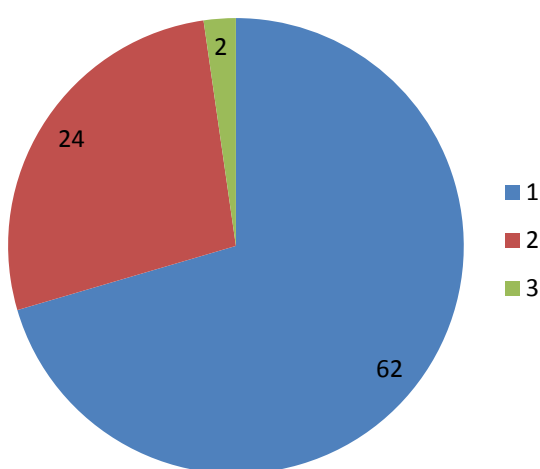


Рис.6.6. Порівняння результату кластеризації при виборі у якості міри близькості відстань Кенбера та усереднений метод зв'язку

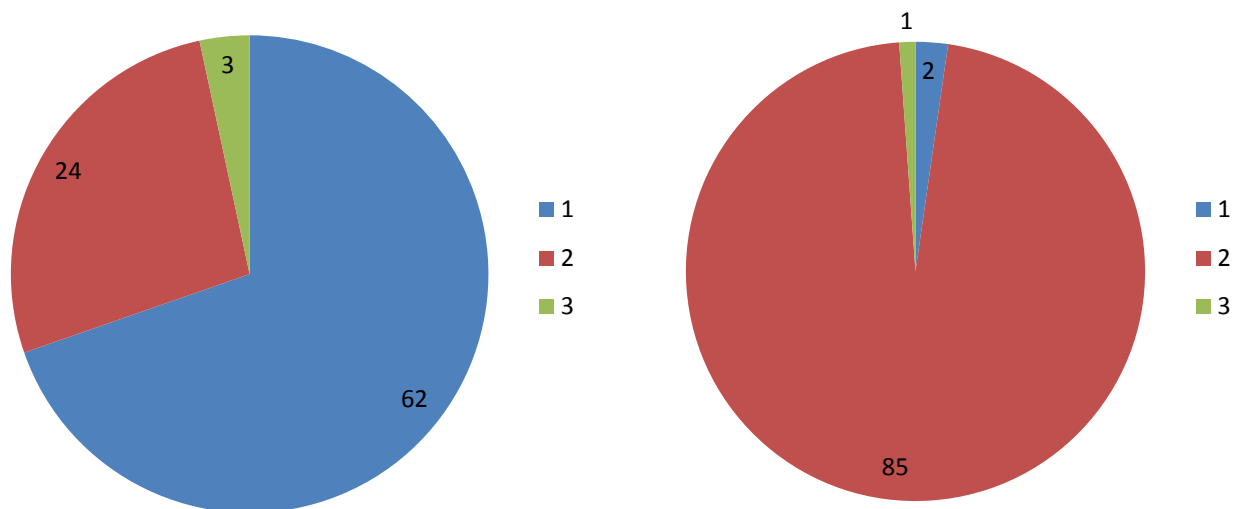


Рис.6.7. Порівняння результату кластеризації при виборі у якості міри близькості відстань Кенбера та центроїдний метод зв'язку

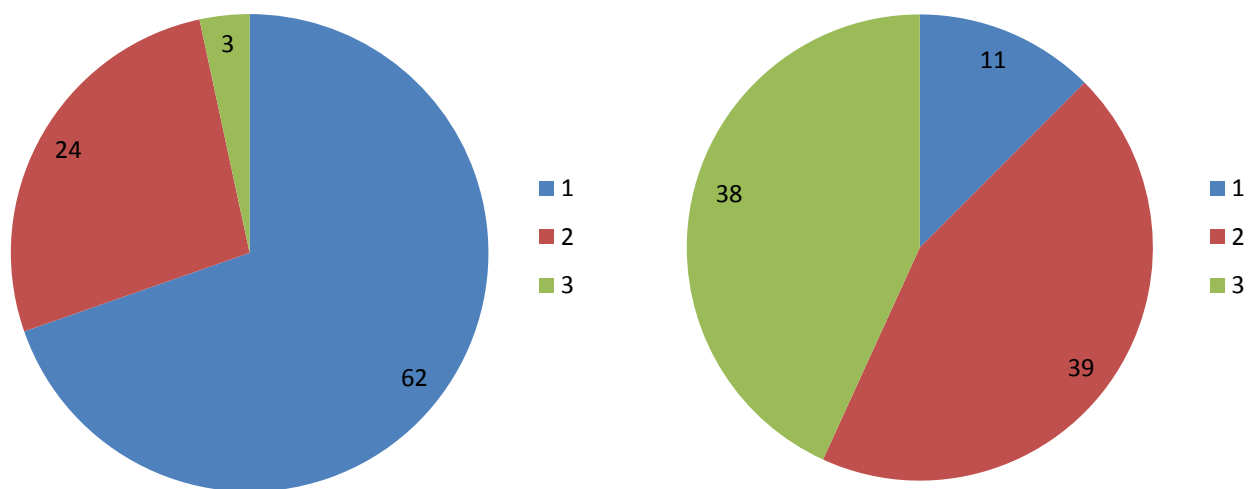


Рис.6.8. Порівняння результату кластеризації при виборі у якості міри близькості відстань Кенбера та повний метод зв'язку

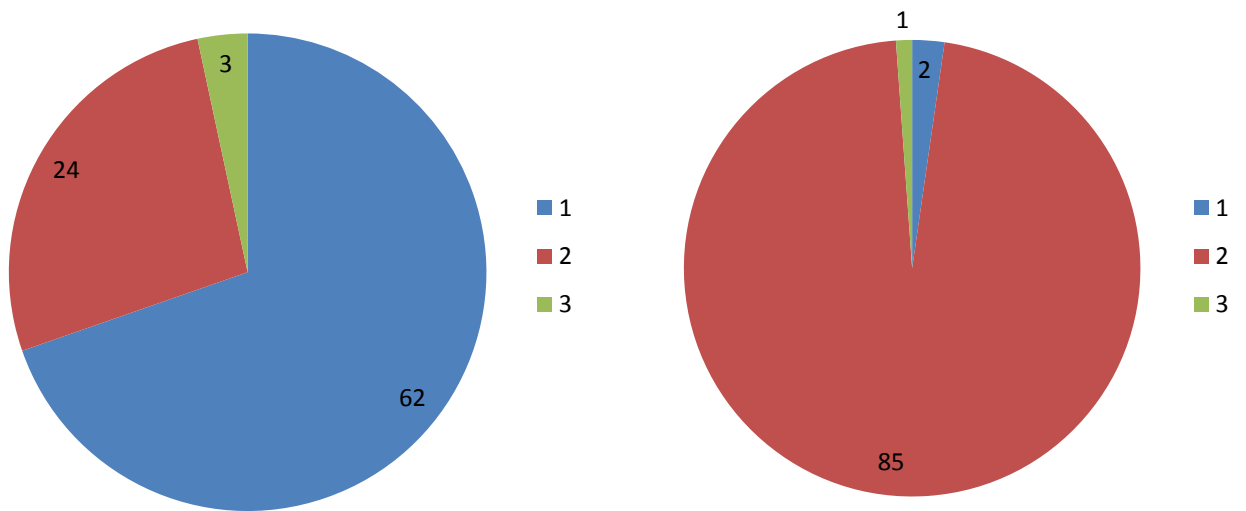


Рис.6.9 Порівняння результату кластеризації при виборі у якості міри близькості відстань Кенбера та медіанний метод зв'язку

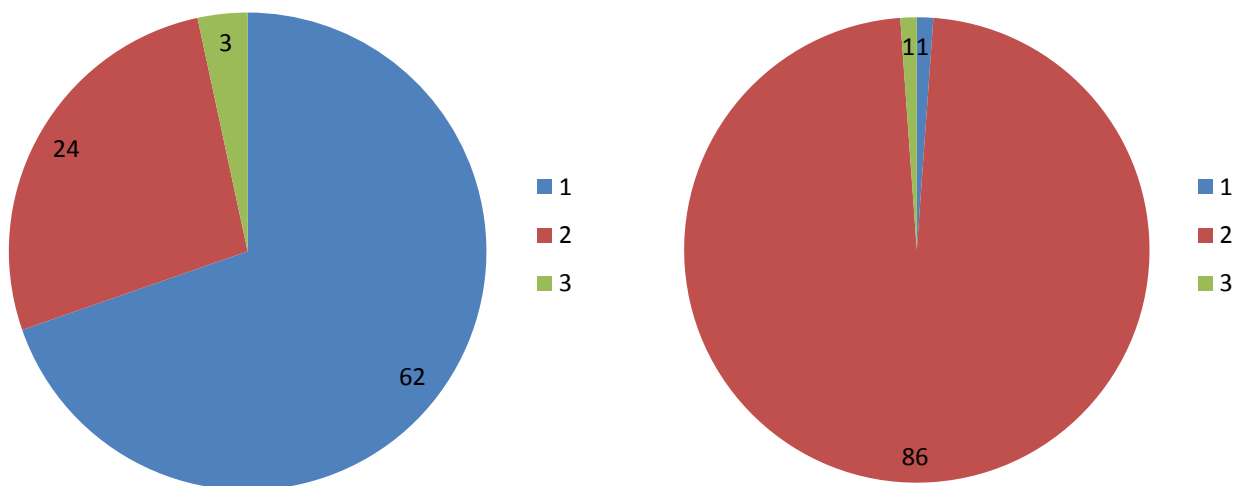


Рис.6.10. Порівняння результату кластеризації при виборі у якості міри близькості відстань Кенбера та одиночний метод зв'язку

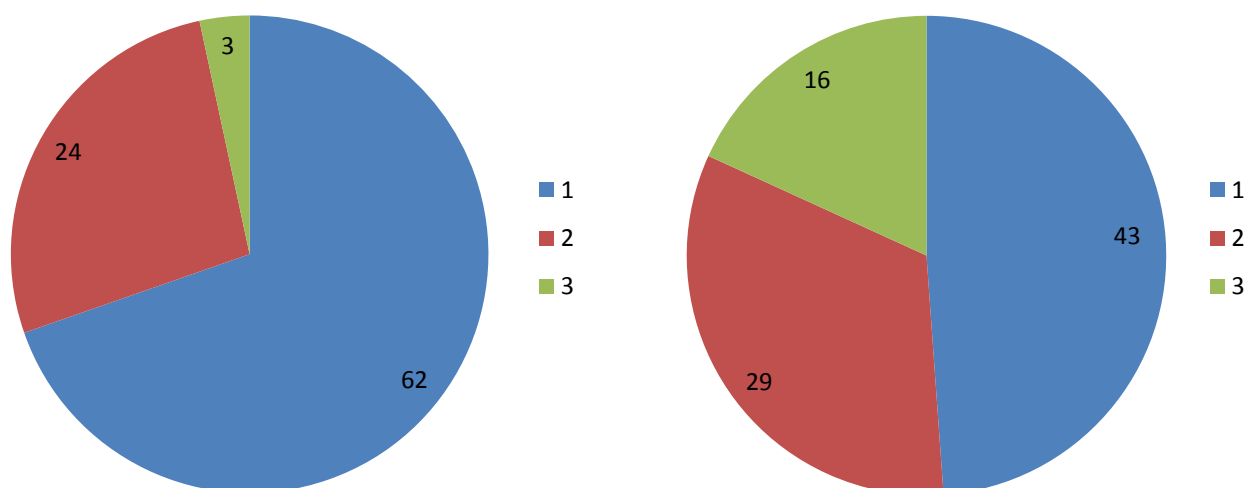


Рис.6.11. Порівняння результату кластеризації при виборі у якості міри близькості відстань Кенбера та метод зв'язку Варда

Подальші результати кластеризації подано у табл.6.2.

Таблиця 6.2. Розбиття по кластерам

| Міра близькості       | Метод з'єднання |             |        |           |           |             |                 |
|-----------------------|-----------------|-------------|--------|-----------|-----------|-------------|-----------------|
|                       | Середній        | Центральний | Повний | Медіанний | Одиночний | Метод Варда | Вхідне розбиття |
| Відстань Кенбера      | 1-2             | 1-2         | 1-11   | 1-2       | 1-1       | 1-43        |                 |
|                       | 2-78            | 2-85        | 2-39   | 2-85      | 2-86      | 2-29        |                 |
|                       | 3-8             | 3-1         | 3-39   | 3-1       | 3-1       | 3-16        |                 |
| Відстань Евклідова    | 1-5             | 1-2         | 1-5    | 1-1       | 1-2       | 1-35        |                 |
|                       | 2-82            | 2-83        | 2-21   | 2-86      | 2-83      | 2-48        |                 |
|                       | 3-1             | 3-3         | 3-62   | 3-1       | 3-3       | 3-5         |                 |
| Відстань Манхетенська | 1-83            | 1-2         | 1-46   | 1-86      | 1-2       | 1-32        | 1-62            |
|                       | 2-3             | 2-83        | 2-37   | 2-1       | 2-85      | 2-20        | 2-24            |

|                             |      |      |      |      |      |      |     |
|-----------------------------|------|------|------|------|------|------|-----|
|                             | 3-2  | 3-3  | 3-5  | 3-1  | 3-1  | 3-36 | 3-2 |
| Відстань Чебишева           | 1-5  | 1-5  | 1-38 | 1-5  | 1-2  | 1-6  |     |
|                             | 2-81 | 2-82 | 2-45 | 2-82 | 2-69 | 2-63 |     |
|                             | 3-2  | 3-1  | 3-5  | 3-1  | 3-17 | 3-19 |     |
| квадрат Евклідової відстані | 1-5  | 1-83 | 1-5  | 1-5  | 1-2  | 1-5  |     |
|                             | 2-82 | 2-3  | 2-21 | 2-82 | 2-83 | 2-61 |     |
|                             | 3-1  | 3-2  | 3-62 | 3-1  | 3-3  | 3-22 |     |

Визначимо правильність роботи алгоритму для кожного випадку, що приведено вище.

Таблиця 6.3.Правильність алгоритму

| Міра близькості       | Метод з'єднання |              |             |              |        |              |           |              |           |              |             |              |
|-----------------------|-----------------|--------------|-------------|--------------|--------|--------------|-----------|--------------|-----------|--------------|-------------|--------------|
|                       | Середній        |              | Центроїдний |              | Повний |              | Медіанний |              | Одиночний |              | Метод Варда |              |
|                       | КВКО            | Правильність | КВКО        | Правильність | КВКО   | Правильність | КВКО      | Правильність | КВКО      | Правильність | КВКО        | Правильність |
| Відстань Кенбера      | 22              | 0,25         | 24          | 0,27         | 21     | 0,24         | 24        | 0,27         | 24        | 0,27         | 41          | 0,47         |
| Відстань Евклідова    | 24              | 0,27         | 22          | 0,25         | 10     | 0,11         | 23        | 0,26         | 22        | 0,25         | 37          | 0,42         |
| Відстань Манхетенська | 62              | 0,7          | 22          | 0,25         | 39     | 0,44         | 63        | 0,71         | 25        | 0,28         | 30          | 0,34         |

|                             |    |      |    |      |    |      |    |      |    |      |    |      |
|-----------------------------|----|------|----|------|----|------|----|------|----|------|----|------|
| Відстань Чебишева           | 23 | 0,26 | 24 | 0,27 | 37 | 0,42 | 22 | 0,25 | 20 | 0,23 | 19 | 0,22 |
| квадрат Евклідової відстані | 24 | 0,27 | 62 | 0,7  | 10 | 0,11 | 24 | 0,27 | 22 | 0,25 | 17 | 0,19 |

З табл.6.3 видно, що найточніший розподіл вибірки на кластери отримано при наступних параметрах:

- Манхетенська відстань та усереднений метод зв'язку;
- квадрат Евклідової відстані та центроїдний метод зв'язку;
- Манхетенська відстань та медіанний метод зв'язку.

### 6.3 Узагальнення по результатах роботи алгоритмів

Після аналізу роботи обох алгоритмів отримано наступні значення:

Таблиця 6.4. Результати роботи обох алгоритмів

| Параметри алгоритму                                      | Правильність |
|----------------------------------------------------------|--------------|
| Алгоритм k means та відстань Кенбера                     | 0,5          |
| Манхетенська відстань та усереднений метод зв'язку       | 0,70         |
| Квадрат Евклідової відстані та центроїдний метод зв'язку | 0,70         |
| Манхетенська відстань та медіанний метод зв'язку         | 0,71         |

### 6.4 Застосування отриманих результатів для розпізнавання нових об'єктів

Задача кластеризації даних пацієнтів важлива не тільки з точки зору групування наявних об'єктів в БД за певними ознаками в кластери(групи), але й тим, що з'являється можливість класифікувати новий об'єкт до деякого

кластеру. Оскільки найточніший результат кластеризації отримано при використанні алгоритму ієрархічного алгоритму та при виборі у якості міри близькості – Манхетенську відстань та медіанний метод зв'язку, то класифікація нового об'єкта до одного із отриманих кластерів проходить наступним чином:

- на вхід в систему подається новий об'єкт, який має такі ж атрибути, що і об'єкти в вхідній вибірці;
- обраховується відстань від вхідного об'єкта до кожного із центрів кластерів;
- об'єкт належить до того кластеру, відстань до якого найменша.

Здійснити перевірку правильності та коректності даного підходу можна використовуючи об'єкти, що уже попередньо класифіковані і порівняти отримані результати з вхідними параметрами об'єктів. Як зазначалось вище, в рамках даного дослідження розроблено програмне забезпечення, яке автоматизує процес кластеризації та класифікації об'єкта. Нехай потрібно класифікувати об'єкти з наступними параметрами:

Таблиця 6.5.Атрибути вхідних об'єктів

| Показник                                       | Значення |      |      |    |
|------------------------------------------------|----------|------|------|----|
|                                                | I        | II   | III  | IV |
| <b>L-CORE</b> (внутрішня температура пацієнта) | 36,6     | 36,6 | 35,5 | 37 |
| <b>L-SURF</b> (температура тіла пацієнта)      | 34,8     | 36   | 35   | 36 |
| <b>L-O2</b> (насичення киснем у %)             | 94       | 98   | 99   | 98 |

|                                                                  |                                |                              |                                |                             |
|------------------------------------------------------------------|--------------------------------|------------------------------|--------------------------------|-----------------------------|
| <b>L-BP</b> (останнє вимірювання кров'яного тиску)               | 88/65                          | 120/80                       | 118/82                         | 120/80                      |
| <b>SURF-STBL</b> (стабільність температури тіла пацієнта)        | стабільна                      | стабільна                    | стабільна                      | нестабільна                 |
| <b>CORE-STBL</b> (стабільність внутрішньої температури пацієнта) | стабільна                      | стабільна                    | стабільна                      | стабільна                   |
| <b>BP-STBL</b> (стабільність кров'яного тиску)                   | нестабільна                    | стабільна                    | стабільна                      | нестабільна                 |
| <b>COMFORT</b> (коефіцієнт комфорту пацієнта)                    | 10                             | 15                           | 10                             | 5                           |
| <b>ADM-DECS</b> (рішення)                                        | Пацієнт відправляє на огляд(2) | Пацієнт відправляє додому(1) | Пацієнт відправляє на огляд(1) | Пацієнт відправляє у ВІТ(3) |

Класифікацію об'єктів здійснено наступним чином:

Таблиця 6.6.Результати класифікації

| № об'єкту | Відстань до центрів                    | Мінімальна відстань – приналежність кластеру |
|-----------|----------------------------------------|----------------------------------------------|
| 1         | 1 - 1.1612<br>2 - 1.1378<br>3 - 1.1521 | Об'єкт належить до 2 кластеру                |
| 2         | 1 - 1.0396<br>2 - 1.0845               | Об'єкт належить до 1 кластеру                |

|   |                                        |                               |
|---|----------------------------------------|-------------------------------|
|   | 3 - 1.0452                             |                               |
| 3 | 1 - 1.1098<br>2 - 1.1045<br>3 - 1.1198 | Об'єкт належить до 2 кластеру |
| 4 | 1 - 1.1078<br>2 - 1.1162<br>3 - 1.0989 | Об'єкт належить до 3 кластеру |

## ВИСНОВКИ ДО РОЗДІЛУ

1. Проведено кластеризацію кожним із обраних методів з використанням різних варіантів вибору міри близькості та методу з'єднання. Запропоновано підхід до оцінки правильності роботи кожного із алгоритмів.
2. Здійснено порівняння результатів кластеризації та виявлено, що в даному випадку найбільшу кількість вірно кластеризованих об'єктів отримали при використанні ієрархічного методу з наступними параметрами: кількість кластерів – 3, міра близькості - Манхетенська відстань, метод зв'язку - медіанний метод.
3. Показано можливість застосування кластеризації даних для прогнозування поведінки іншого об'єкта. Проведено класифікацію нового об'єкта на основі результатів отриманих при кластеризації.

## 7 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

### 7.1 Загальні положення

Далі буде описано один з варіантів приміщення, в якому буде використовуватись розроблене програмне забезпечення. Використання програмного забезпечення проводитиметься у кабінеті. Несуча конструкція приміщення виконана із цегли. Параметри приміщення вказано у табл. 7.1:

Таблиця 7.1. Параметри приміщення

| Параметр              | Значення | Од. виміру |
|-----------------------|----------|------------|
| Довжина (L)           | 5        | м.         |
| Ширина (D)            | 4        | м.         |
| Висота (H)            | 3        | м.         |
| Площа (S)             | 20       | кв. м.     |
| Об'єм (V)             | 60       | куб. м.    |
| Кількість працівників | 1        | чол.       |

Приміщення відповідає вимогам [16], згідно з якими, площа на одне робоче місце повинна бути не менша 6 кв. м., а об'єм не менше ніж 20 куб. м. План приміщення зображено на рис. 7.1.



Приміщення, у якому ведеться робота розміщена на першому поверсі. Вікна у кімнаті спроектовані так, щоб протягом робочого дня в кімнату потрапляло багато світла. Поряд з даною будівлею знаходяться споруди висотою з 5-ти поверхові будинки, які частково перекривають потрапляння сонячного проміння в офісне приміщенням.

Категорія робіт, які пов'язані з використанням проекту, належать до роботи категорії «легка-1а», що виконуються сидячи і не потребують фізичного навантаження (категорія 1а охоплює види діяльності, при яких витрати енергію не перевищують 90-120 ккал/год).

## 7.2 Аналіз шкідливих та небезпечних факторів

Для аналізу шкідливих та небезпечних факторів виділимо п'ять основних факторів, які можуть впливати на працівників: виробничий шум, мікроклімат приміщення, випромінювання ВДТ, природне та штучне освітлення та електробезпека.

### 7.2.1 Аналіз виробничого шуму

Джерелом шуму в приміщенні виступають принтер та ЕОМ, а саме їхні компоненти: вентилятор охолодження, CD-дисковод.

У цьому випадку рівень шуму охолоджувачів у комп'ютері становить 35 дБА та дисководів – 46 дБА. Принтер застосовується в роботі для друку вихідних документів. Струменевий принтер має рівень шуму приблизно 42 дБА.

Максимально принтер працює по 2 години за робочий день. Час роботи комп'ютера складає 8 годин.

Джерела шуму мають мінливий режим роботи. Критерієм гігієнічної оцінки нестійкого шуму є еквівалентний (щодо енергії) рівень звуку широкосмугового, стійкого та не імпульсивного шуму, який чинить на людину такий же вплив, як і нестійкий шум ( $L_{EKB}$ , дБА) [24]. Цей шум розраховується за формулою:

$$L_{EKB} = 10 \lg \frac{1}{T} \sum_{i=1}^n t_i 10^{0.1 L_i}, \quad (7.1)$$

де  $L_i$  – рівень звуку класу  $i$ ;  $t_i$  – відносний час впливу шуму класу  $L_i$ , % від часу вимірювання  $T$  (8 годин).

$$L_{EKB} = 10 \lg \frac{1}{8} (1 \cdot 10^{0.1 \cdot 35} + 1 \cdot 10^{0.1 \cdot 46} + 0,25 \cdot 10^{0.1 \cdot 42}) = 46,71 \quad (7.2)$$

Загальний рівень шуму складає 46,71 дБА. Допустимий еквівалентний рівень звуку згідно [17] наступні: для програміста ЕОМ нормований звуковий тиск повинен бути не вище 50 дБА. Тобто наявний рівень шуму відповідає вимогам.

#### 7.2.2 Аналіз мікроклімату приміщення

Допустимі значення мікроклімату приміщення наведені в табл. 7.3, потрібно перевірити відповідність необхідних значень параметрів з їх фактичними значеннями.

Таблиця 7.3. Допустимі значення параметрів мікроклімату

| Період року | Категорія робіт | Максимальна температура повітря, °С | Відносна вологість повітря, % | Швидкість руху повітря, м/с |
|-------------|-----------------|-------------------------------------|-------------------------------|-----------------------------|
| Холодний    | Легка-1а        | 22-24                               | 40-60                         | 0,1                         |
| Теплий      | Легка-1а        | 23-25                               | 40-60                         | 0,1                         |

У даному приміщенні використовується центральне водяне опалення низького тиску, що забезпечує температуру в приміщенні в холодну пору року в межах 22-24 градусів. Температуру також регулює канална загально-обмінна припливна та витяжна вентиляція, вона дає змогу тримати температуру повітря у приміщенні, в теплі пори року в межах 23-25 градусів. Відносна вологість повітря у приміщенні складає 45%,

швидкість руху повітря менша 0,1 м/с. Отже, розглядуване приміщення відповідає нормам.

#### 7.2.3 Аналіз випромінювання

ПК є значним джерелом електромагнітного випромінювання. Відповідно до [16], припустима інтенсивність потоку енергії 10Вт/м<sup>2</sup>, а напруженість електромагнітного поля по електричній складовій на відстані 0,5м від екрану – 10 В/м.

Сучасні комп'ютерні дисплеї, як наприклад Dell U2312NM, мають інтенсивність потоку не більшу за 1 Вт/м<sup>2</sup>, що відповідає нормам.

#### 7.2.4 Аналіз природнього та штучного освітлення

В даному приміщенні виконується робота з об'єктами розпізнавання 0,3 мм – 0,5 мм, тому за [18] робота у лабораторії класифікується як дуже високої точності. Рівень освітленості на робочому столі в зоні розташування документів має бути в межах 300-500 лк. Освітлення у кабінеті представлено природнім та штучним загальним освітленням.

#### 7.2.5 Аналіз електробезпеки

Згідно з [19] за ступенем небезпеки враження електрострумом приміщення відноситься до приміщення без підвищеної небезпеки. У приміщення проведено 3-х фазне електроживлення напругою 220 В, частотою 50 Гц та з максимальним струмом у 32 А.

Електроустановки підключені до групового щитка, що знаходиться в приміщенні. Використовуються кабелі та проводи з мідними жилами перерізом не менше 2,5 мм<sup>2</sup>. Застосована відкрита електропроводка в пластмасових коробах із важкогорючих матеріалів з помірною димоутворювальною здатністю.

Технічні захисні заходи електробезпеки: виконана робоча ізоляція струмопровідних частин (поліетилен, пластмаса), а також застосовані

пластмасові короби з важкогорючих матеріалів з помірною димоутворювальною здатністю. Електромагнітна блокування безпеки перемикачів виконана за допомогою стержневого магніту.

Електричний роз'їм (розетка, вилка) у разі підключення до електричної мережі гарантує з'єднання корпусу з заземленням, а в разі відключення гарантує від'єднання корпусу від заземлювача в останню чергу.

У разі аварії передбачені запобіжники в приладах і системі занулення, що забезпечує захист людини від ураження електричним струмом.

До роботи в даному приміщенні не допускаються особи, які не пройшли інструктаж з електробезпеки. Таким чином, в приміщенні заходи електробезпеки відповідають всім вимогам та нормам.

### 7.3 Інженерне рішення

Освітлення, відповідно до [20] приміщення для роботи з ВДТ повинні мати природне та штучне освітлення.

Природне освітлення. В даному приміщенні виконується робота з об'єктами розпізнавання 0,3 мм – 0,5 мм, тому за [18] робота у лабораторії класифікується як дуже високої точності. Рівень освітленості на робочому столі в зоні розташування документів має бути в межах 300-500 лк. Освітлення у кабінеті представлено природнім та штучним загальним освітленням.

Система природного освітлення: бокове однобічне, що здійснюється через світлові отвори (вікна) в зовнішніх стінах. Сумарна площа вікон становить 6 м<sup>2</sup>. Вікно орієнтоване на північ. Згідно [16] площа вікон задовольняє нормативним значенням.

Стіни вкриті матовою фарбою, світло-блакитного кольору, що полегшує роботу зору під час використання ЕОМ.

Результати роботи програми DIALux для оцінки природного освітлення зображено на рис. 7.2. Як видно, середня освітленість робочого місця становить 400 Лм, що відповідає нормам.

Штучне освітлення. В даному приміщенні встановлено чотири офісних освітлювальні установки DEXTRA Duet T5 Surface DUET5 139 LB та один світильник DEXTRA Axant Lay-In RAXTT5214 для додаткового освітлення робочого місця.

Основний світильник DEXTRA Duet T5 Surface DUET5 139 LB має такі характеристики:

- світловий потік (Світильник): 2550 lm;
- світловий потік (Лампи): 3100 lm;
- потужність світильників: 41.0 W;
- класифікація світильників по CIE: 100;
- CIE Flux Code: 57 91 98 100 82;
- комплектація: 1 x T5 HO 39w C84.

Крива розподілу світла основного світильника зображена на рис. 7.3.

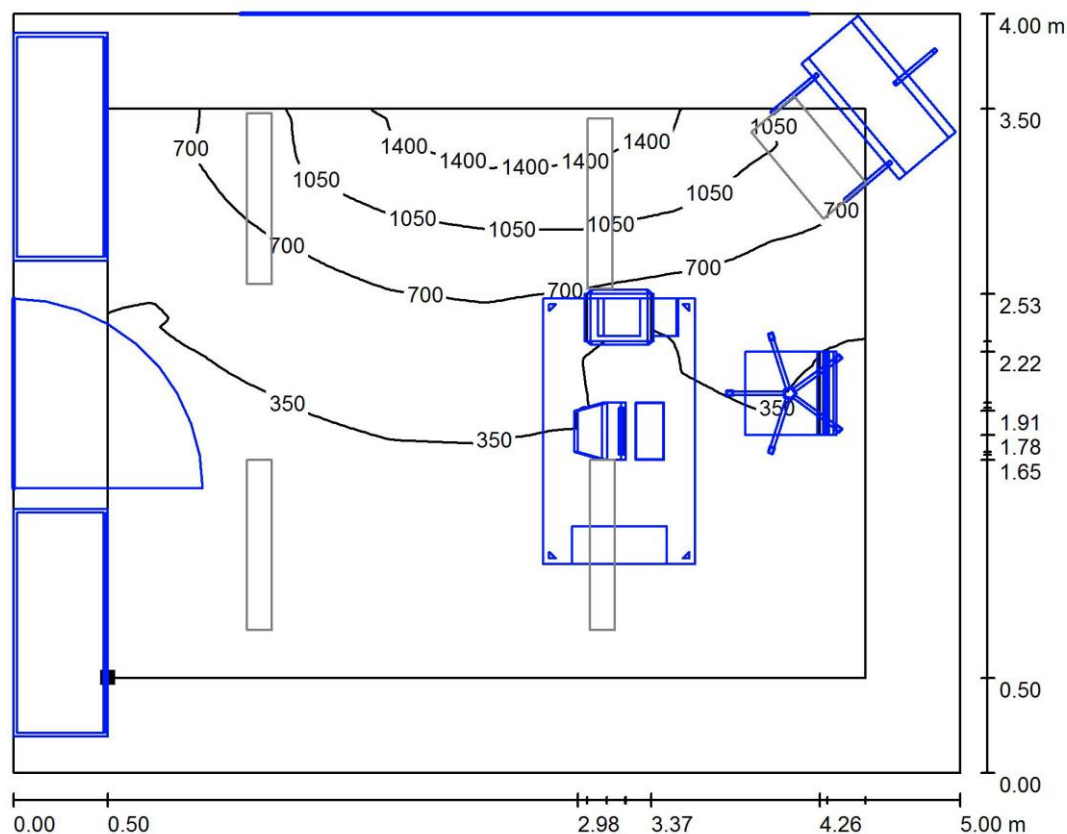


Рис. 7.2. Схема розподілення світла при природньому освітленні

Додатковий світильник DEXTRA Axant Lay-In RAXTT5214 має наступні характеристики:

- світловий потік (Світильник): 1591 lm;
- світловий потік (Лампи): 2500 lm;
- потужність світильників: 31.0 W;
- класифікація світильників по CIE: 100;
- CIE Flux Code: 53 80 96 100 64;
- комплектація: 2 x T5 HE 14w C84.

Результат роботи програми DIALux для оцінки штучного освітлення зображено на рис. 7.4. Як видно, освітлення відповідає нормі, тому що середня освітленість робочої поверхні складає 360 люмен, що знаходиться в допустимому діапазоні 300-500 лк.

Організація робочого місця. Робоче місце розташоване біля вікна. Прохід біля столу – 1м, що задовольняє вимогам [20]. Стіл має габаритні розміри 140x80x72 см. Стілець має можливість регулювання по висоті та

можливість зміни куту нахилу спинки. У якості відео-дисплейного терміналу використовується монітор з рідкокристалічним дисплеєм Dell U2312NM, з номінальним значенням освітлення 350лк. В кімнаті розташований стелаж для паперів. На добре помітному місці розташована аптечка першої медичної допомоги та вогнегасник. Детальну схему приміщення див на рис. 7.1.

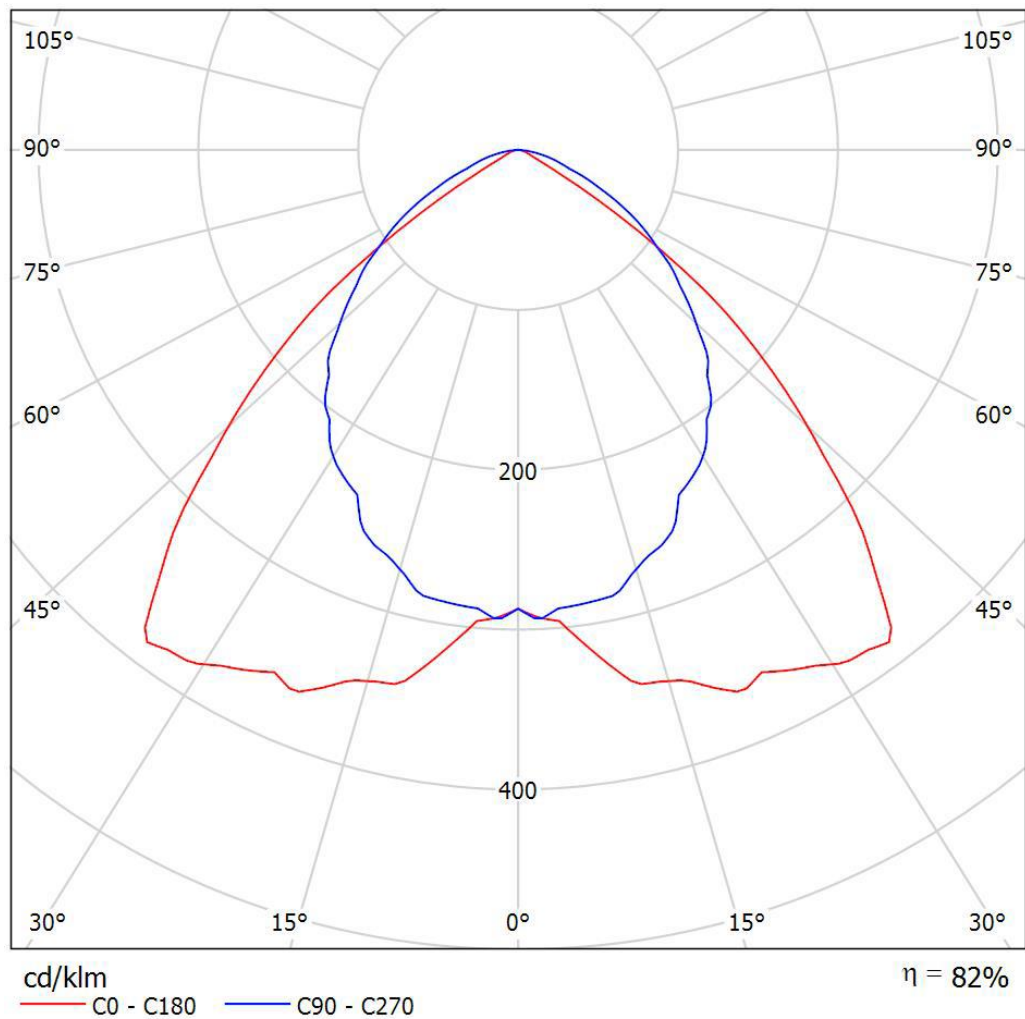


Рис. 7.2. Схема розподілу світла світильника DEXTRA Duet T5 Surface DUET5

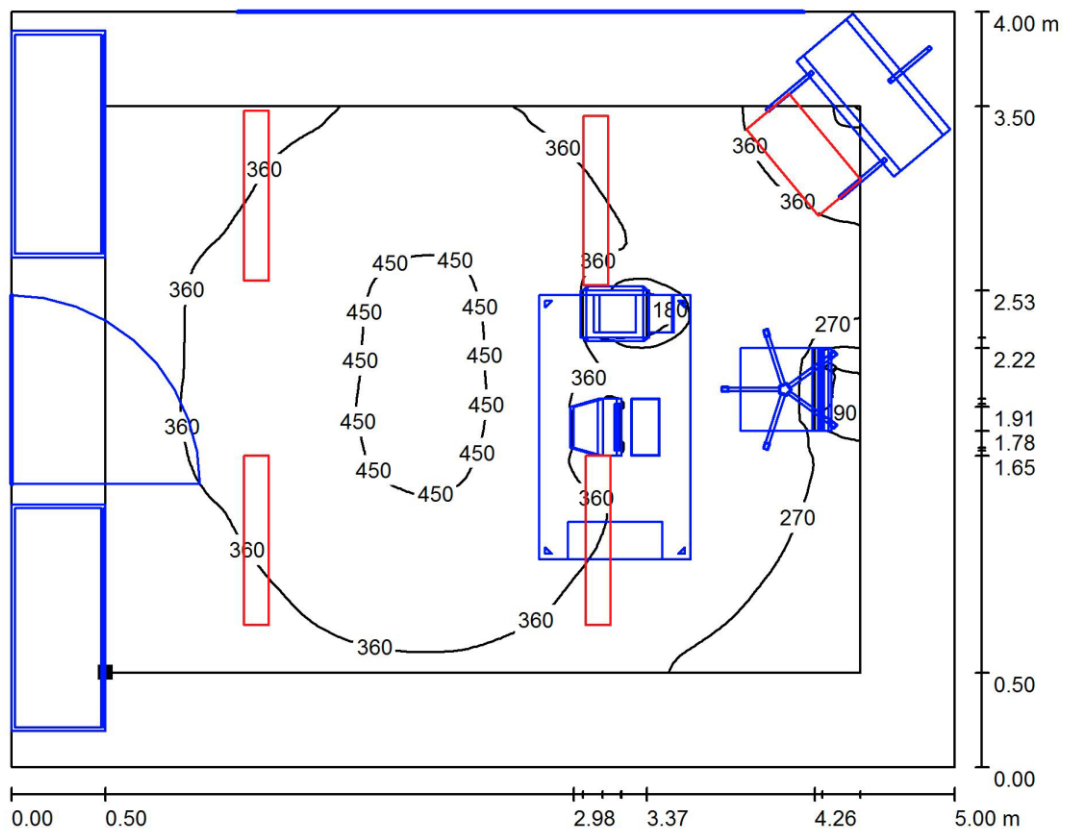


Рис. 7.3. Схема розподілу світла при штучному освітленні кабінету

#### 7.4 Пожежна безпека

Основними причинами, які можуть призвести до виникнення пожежі в приміщенні лабораторії, є: несправності в електрообладнанні – порушення електроізоляції; несправності, що виникли внаслідок механічних пошкоджень і т.п.; несправності в обчислювальній техніці, наприклад, коротке замикання, порушення протипожежного режиму.

Предмети та матеріали, які можуть горіти: меблі – столи, стільці, шафи і т.д.; папір – документація, конструктивні елементи приміщення – покриття підлоги, двері, віконні рами, елементи конструкції ПЕОМ та периферійних пристроїв – корпуси принтера, моніторів, клавіатури і т.п.

Відповідно до техніки безпеки користувачу ЕОМ, після закінчення роботи персональний комп'ютер і периферійні пристрої повинні бути відключені від електричної мережі. У разі виникнення аварійної ситуації необхідно негайно відключити персональний комп'ютер і периферійні пристрої від електричної мережі.

Не допускається:

- виконувати обслуговування, ремонт та налагодження персонального комп'ютеру та периферійних пристроїв безпосередньо на робочому місці оператора;
- зберігати біля персонального комп'ютеру та периферійних пристроїв папір, будь-які носії інформації (диски, флешки тощо), запасні блоки, деталі тощо, якщо вони не використовуються для поточної роботи;
- відключати захисні пристрої, самочинно проводити зміни у конструкції та складі персонального комп'ютеру та периферійних пристроїв або їх технічне налагодження;
- працювати з персональним комп'ютером, у яких під час роботи з'являються нехарактерні сигнали, нестабільне зображення на моніторі тощо.

#### 7.4.1 Категорія приміщення за вибухопожежною небезпекою

Дане приміщення відноситься до категорії В. Згідно з [21], приміщення, що містять ЕОМ, відносяться до категорії В, як приміщення, де перебувають спалимі та важко спалимі рідини, тверді спалимі та важко спалимі речовини та матеріали (в тому числі пил та волокна), а також речовини і матеріали, які здатні при взаємодії з водою, киснем повітря та одне з одним тільки горіти.

#### 7.4.2 Пожежна сигналізація

В даному приміщенні встановлена система пожежної сигналізації System Sensor. Обладнання сигналізації складається з трьох основних частин наведених нижче:

- прилад приймально-контрольний пожежний АРТОН-04П. (призначений для організації централізованої і автономної охорони різних об'єктів від пожеж, шляхом цілодобового контролю стану до 4-х шлейфів пожежної сигналізації);
- телефонний комунікатор БСКТ-1 (є блоком розширення приладів приймально-контрольних пожежних «Артон-04П», «Артон-08П» і призначений для автоматичного дозвону по 8 запрограмованим номерами і передачі кодованих повідомлень про зміну стану ППКП на ПЦН або у вигляді голосових повідомлень, а також для віддаленого моніторингу та управління ППКП);
- двох сповіщувачів пожежний димових оптичних точкових СПД-3 (призначені для виявлення спалахів в закритих приміщеннях різних будівель і споруд, що супроводжуються появою диму і передачі сигналу «ПОЖЕЖА» на ППК).

Вибір пожежних сповіщувачів здійснюється в залежності від характерних приміщень, виробництв, технологічних процесів відповідно Додатка К до [22] "Пожежна автоматика будинків і споруд". Згідно з цим, приміщення обладнане адресованими оповіщувачами.

#### 7.4.3 Засоби гасіння пожежі

Оскільки основним можливим джерелом загоряння у приміщенні є електрообладнання, використовувати воду для пожежогасіння категорично заборонено. Натомість варто використовувати порошкове пожежогасіння, яке працює разом з системою автоматичного виявлення та сповіщення про пожежу. Встановлена установка модульного типу, розташована безпосередньо у приміщенні.

Оскільки площа приміщення дорівнює 20 кв. м., можна використати 1 вогнегасник порошковий ВП-3 з масою заряду 3-кг з урахуванням 3 кг. на 20 м. кв.

Вогнегасники розміщено біля входних дверей та прикріплено до стіни.

## 7.5 Техніка безпеки

Напруга живлення офісної техніки (220 В) є небезпечною для життя людини. Тому, незважаючи на те що в конструкції комп'ютера та принтера передбачена достатня ізоляція від струмопровідних ділянок, необхідно знати та чітко виконувати ряд правил техніки безпеки.

Забороняється:

- знаходитися на робочому місці у верхньому одязі;
- класти одяг і сумки на стіл;
- зберігати на робочому місці напої та їжу;
- працювати на комп'ютері у вологому одязі та вологими руками;
- розташовуватися збоку або ззаду від включеного монітора;
- торкатись екрана, тильного боку дисплея, проводів живлення, заземлення, з'єднувальних кабелів
- пересувати комп'ютери і монітори, знімати кришку корпуса системного блоку;
- самостійно намагатися усунути будь-які неполадки в роботі комп'ютера, незалежно від того, коли і з чийої вини вони сталися;
- перекривати вентиляційні отвори на системному блоці та моніторі;
- класти речі на клавіатуру, монітор і системний блок.

Під час роботи на комп'ютері необхідно:

- забезпечити відстань від екрану до очей – 60-70 см. (відстань витягнутої руки);
- вертикально пряма спина, плечі опущені і розслаблені, ноги на підлозі і не схрещені, лікті, зап'ястя і кисті рук на одному рівні, ліктьові, тазостегнові, колінні, гомілковостопні суглоби під прямим кутом.

## 7.6 План ліквідації надзвичайної ситуації

У ході роботи потрібно створити план локалізації та ліквідації аварійних ситуацій, які можуть виникнути під час роботи на підприємстві.

### 7.6.1 Можливі аварійні ситуації

В залежності від масштабності наслідків аварійної ситуації розрізняють наступні рівні:

Надзвичайна ситуація загальнодержавного рівня – це надзвичайна ситуація, яка розвивається на території двох та більше областей (Автономної республіки Крим, міст Києва та Севастополя) або загрожує транскордонним перенесенням, а також у разі, коли для її ліквідації необхідні матеріали і технічні ресурси в обсягах, що перевищують власні можливості окремої області (Автономної республіки Крим, міст Києва та Севастополя), але не менше 1% обсягів видатків відповідного бюджету.

Надзвичайна ситуація регіонального рівня – це надзвичайна ситуація, яка розвивається на території двох або більше адміністративних районів (міст обласного значення), Автономної республіки Крим, областей, міст Києва та Севастополя або загрожує перенесенням на територію суміжної області, а також у разі, коли для її ліквідації необхідні матеріальні і технічні ресурси в обсягах, що перевищують власні можливості окремого району, але не менше 1% обсягів видатків відповідного бюджету.

Надзвичайна ситуація місцевого рівня – це надзвичайна ситуація, яка виходить за межі потенційно-небезпечного об'єкта, загрожує поширенням самої ситуації або її вторинних наслідків на довкілля, сусідні населені пункти, інженерні споруди, а також у разі, коли для її ліквідації необхідні матеріальні і технічні ресурси в обсягах, що перевищують власні можливості об'єкта. До місцевого рівня також належать всі надзвичайні ситуації, які виникають на об'єктах житлово-комунальної сфери та інших, що не входять до затверджених переліків потенційно-небезпечних об'єктів.

Надзвичайна ситуація об'єктового рівня – це надзвичайна ситуація, яка не підпадає під зазначені вище визначення, тобто така, що розгортається на території об'єкта або на самому об'єкті, її наслідки не виходять за межі об'єкта або його санітарно-захисної зони.

З усіх можливих аварій на РНО найбільш небезпечними є радіаційні аварії на атомних електростанціях з викидом радіоактивних речовин у навколишнє середовище.

В нашому випадку існує можливість радіаційного забруднення під час аварійної ситуації комунального рівня на атомній електростанції.

В Україні діють 5 атомних електростанцій з 16 енергетичними ядерними реакторами, 2 дослідних ядерних реактори та більше 8 тис. підприємств і організацій, які використовують у виробництві, науково-дослідній роботі та медичній практиці різноманітні радіоактивні речовини, а також зберігають і переробляють радіоактивні відходи. Найближчою АЕС до міста Київ є Чорнобильська АЕС з реакторами типу РБМК. На даний момент вона повністю виведена з експлуатації, проте на території Росії до цього часу використовується 11 реакторів цього типу.

#### 7.6.2 Оцінка радіаційної обстановки в зоні радіаційного забруднення

Потрібно розрахувати радіаційну обстановку в зоні радіаційного забруднення в результаті аварії на реакторі типу РБМК. Оцінку будемо проводити згідно з інструкціями в [23].

Аварія сталась о 3:00. Вимірний рівень радіації на 3:30 становить 45 р/год. Астрономічний час початку робіт щодо ліквідації наслідків аварії – 5:00. Тривалість робіт оцінюється в 3.5 години. Допустима доза радіації становить 20 радіан. Коефіцієнт ослаблення радіації – 3.

Всі розрахунки для оцінки радіаційного стану прийнято проводити у відносному часі. Точкою відліку вважається момент аварії:  $t_{\text{астрон}} = 3:00$ ,  $t_{\text{відн}} = 0$ . З огляду на це, відносний час початку роботи є наступним:  $t_{\text{пастрон}} = 5:00$ ,  $t_{\text{пвідн}} = 2$ . Слід також розрахувати час кінця роботи:

$$t_{\text{Квідн}} = t_{\text{Пвідн}} + T = 2 + 3,5 = 5,5 \quad (7.3)$$

Для наступних розрахунків слід вирахувати рівень радіації на першу годину після аварії. Для цього маючи коефіцієнт для перерахунку рівнів радіації на рівний час після аварії на АЕС  $K_{0.5\text{п}} = 0,81$  та виміряний на цей час рівень радіації – 45 Р/год по формулі 7.4 знайдемо шукане значення.

$$P_1 = P_{\text{вим}} \cdot K_{t_{\text{вим}}} \quad (7.4)$$

За формулою 7.4 маємо, що рівень радіації на першу годину після аварії – 36,45 Р/год.

Далі розрахуємо дозу отриманої радіації під час роботи. Розрахунки будемо проводити за формулою 7.5.

$$D = \frac{P_{\text{ср}} \cdot t_{\text{р}}}{K_{\text{осл}}} \quad (7.5)$$

Для використання формули 7.4 слід розрахувати значення середнього випромінювання, яке вимірюється як середнє арифметичне рівня радіації на початок робіт та на кінець робіт. Рівень радіації на певний момент розраховується за формулою 7.6.

$$P_t = \frac{P_1}{K_{t_{\text{н}}}} \quad (7.6)$$

На початок ліквідаційних робіт рівень радіації – 21,96 Р/год, на кінець ліквідаційних робіт – 6,63 Р/год. Середній рівень радіації – 14,29 Р/год.

Провівши необхідні обчислення за формулою 7.4 отримаємо, що доза отриманого при роботі випромінювання рівна 16,67 Р.

Далі слід розрахувати допустимий час роботи на зараженому об'єкті. Для цього слід розрахувати коефіцієнт  $\alpha$ . Для цього скористаємось формулою 7.7.

$$\alpha = \frac{P_1}{K_{\text{осл}} \cdot D_{\text{доп}}} \quad (7.7)$$

Вирахувана величина  $\alpha$  рівна 0,6. Маючи її значення та час початку роботи з ліквідації  $t_{\text{Пвідн}} = 2$  скориставшись графіком на рис. 7.5 виміряємо допустимий час роботи на реакторі.

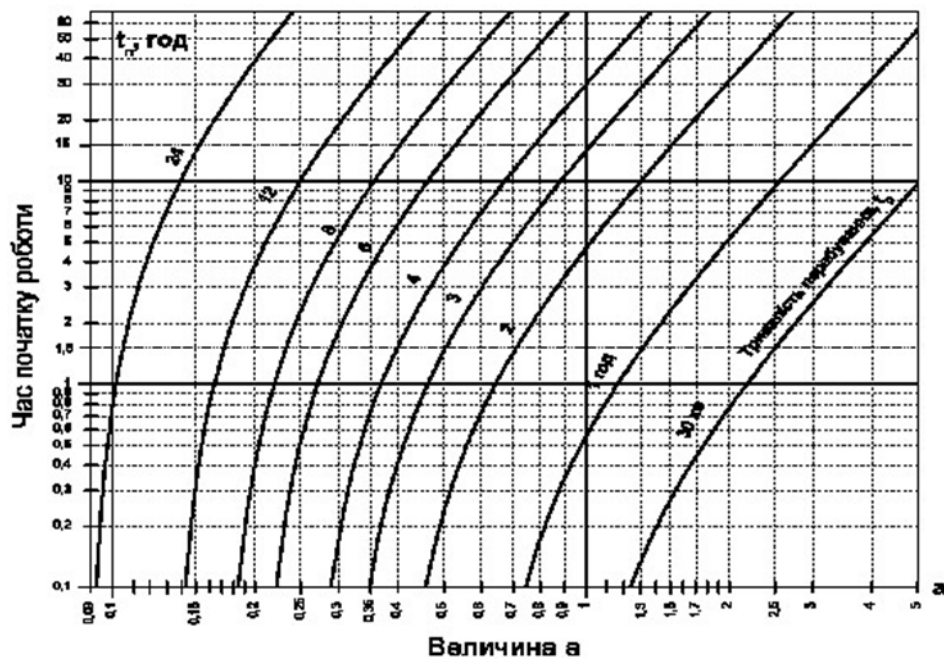


Рис. 7.4. Графік визначення тривалості перебування в зоні радіоактивного зараження при аварії на АЕС з реактором РБМК-1000

За графіком бачимо, що в даних умовах допустимий час роботи рівний 2 години 20 хвилин.

Отже, при роботі на даній АЕС слід обов'язково проводити інструктаж з техніки безпеки та завжди мати наготові службу екстреного реагування на випадок аварійної ситуації. Також забезпечити весь особовий склад необхідною захисною екіпіровкою.

## ВИСНОВКИ ДО РОЗДІЛУ

1. Проведено аналіз підприємства на базі якого розробляється проект. Розглянуто шкідливі та небезпечні фактори на робочому місці: рівень шуму, освітлення на робочому місці, шкідливих речовин у повітрі, випромінювання від електроприладів, які знаходяться в межах кімнати, електробезпеки, пожежної сигналізації та систем пожежогасіння встановлених у робочій кімнаті, а також проведено оцінку радіаційної обстановки в разі перебування підприємства у зоні радіаційного забруднення.
2. Проаналізовані фактори відповідають чинним нормативним вимогам.

## ВИСНОВКИ

Здійснено огляд проблеми аналізу даних пацієнтів та виявлено основні проблеми кластеризації даних пацієнтів: багаторозмірність даних, можливе категоріальне представлення даних, вибір методу кластеризації та оцінка результату. Сформовано задачу кластеризації даних пацієнтів.

Розглянуто сучасні системи аналізу даних з точки зору вирішення задачі кластеризації даних пацієнтів. Виявлено необхідність в написанні програмного забезпечення, яке даватиме можливість аналізувати правильність отриманих результатів кластеризації.

Проаналізовані існуючі методи кластеризації з точки зору їх використання для вирішення поставленої задачі. Також проаналізовано міри близькості та методи з'єднання об'єднання об'єктів у кластери. Для оцінки розбиття об'єктів на кластери використовується поняття правильності.

Описані завдання розробленого програмного забезпечення. Надано опис обраних засобів розробки та вимог до технічного забезпечення та специфікація функцій. Наведено покроковий опис роботи з програмою.

Проведено кластеризацію кожним із обраних методів з використанням різних варіантів вибору міри близькості та методу з'єднання. Здійснено порівняння результатів кластеризації та виявлено, що в даному випадку найбільшу кількість вірно кластеризованих об'єктів отримали при використанні ієрархічного методу з наступними параметрами: кількість кластерів – 3, міра близькості - Манхетенська відстань, метод зв'язку - медіанний метод. Показано можливість застосування кластеризації даних для прогнозування поведінки іншого об'єкта. Проведено класифікацію нового об'єкта на основі результатів отриманих при кластеризації.

Проаналізовано приміщення в якому може використовуватись розроблена технологія, з точки зору відповідності нормативам в галузі охорони праці та безпеки в надзвичайних ситуаціях. Встановлено, що приміщення відповідає чинним нормам. Описані правила техніки безпеки та

виконано розрахунок, що характеризують обстановку при виникненні надзвичайної ситуації і потрапляння підприємства у зону радіаційного забруднення, вироблені рекомендації для мінімізації можливих негативних наслідків.

## ПЕРЕЛІК ПОСИЛАНЬ

1. Data mining – интеллектуальный анализ данных [Электронный ресурс]. – Режим доступа к ресурсу:  
<http://www.inftech.webservis.ru/it/database/datamining/ar2.html>.
2. Барсегян А.А. Технология анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / А.А. Барсегян. – СПб.: БХВ-Петербург, 2007. – 384 с.
3. Data Mining – интеллектуальный анализ данных [Электронный ресурс]. – Режим доступа к ресурсу:  
<http://www.inftech.webservis.ru/it/database/datamining/ar2.html>.
4. Айвазян С.А. Прикладная статистика: классификация и снижение размерности. / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин — М.: Финансы и статистика, 2003. – 471 с.
5. Data Mining [Электронный ресурс]. – Режим доступа к ресурсу:  
[http://ru.wikipedia.org/wiki/Data\\_mining](http://ru.wikipedia.org/wiki/Data_mining).
6. Задачи Data Mining. Классификация и кластеризация [Электронный ресурс]. – Режим доступа к ресурсу: <http://www.intuit.ru/department/database/datamining/5/1.html>.
7. Statistica [Электронный ресурс]. – Режим доступа к ресурсу:  
<http://uk.wikipedia.org/wiki/Statistica>
8. Weka [Электронный ресурс]. – Режим доступа к ресурсу:  
<http://uk.wikipedia.org/wiki/Weka>
9. RapidMiner [Электронный ресурс]. – Режим доступа к ресурсу:  
<http://uk.wikipedia.org/wiki/RapidMiner>
10. Post-Operative Patient Data Set [Электронный ресурс]. – Режим доступа к ресурсу: <https://archive.ics.uci.edu/ml/datasets/Post-Operative+Patient>
11. Кластерный анализ [Электронный ресурс]. – Режим доступа к ресурсу:  
[http://ru.wikipedia.org/wiki/Кластерный\\_анализ](http://ru.wikipedia.org/wiki/Кластерный_анализ).
12. Электронный учебник по статистике. М. StatSoft.WEB [Электронный ресурс]. – Режим доступа к ресурсу:  
<http://www.statsoft.ru/home/textbook/default.htm>.

13. Meesad P. A hybrid intelligent system for medical diagnosis / P. Meesad, G. Yen // Proc. of IJCNN'01. Washington. 2001. P. 2558 – 2563
14. Меры сходства и различия [Электронный ресурс]. – Режим доступа к ресурсу: [http://www.forekc.ru/op2/index-mery\\_shodstva\\_i\\_razlichiya\\_4.htm](http://www.forekc.ru/op2/index-mery_shodstva_i_razlichiya_4.htm)
15. Кластерный анализ [Электронный ресурс]. - Режим доступа к ресурсу: <http://www.statsoft.ru/home/textbook/modules/stcluan.html>
16. НПАОП 0.00-1.28-10. Правила охорони праці під час експлуатації електронно-обчислювальних машин : затверджені наказом Держкомітету України з промислової безпеки, охорони праці та гірничого нагляду від 26.03.2010 № 65 [Електронний ресурс]. – Режим доступу: <http://dnop.com.ua>.
17. ДНС 3.3.6.037-99. Санітарні норми виробничого шуму, ультразвуку та інфразвуку МОЗ України [Електронний ресурс]. – Режим доступу : <http://dnop.com.ua/dnaop/act4878.htm>.
18. ДБН.В2.5.-28-2006. Штучне та природне освітлення [Електронний ресурс]. – Режим доступу : <http://dnop.com.ua/dnaop/act8425.htm>.
19. Правила улаштування електроустановок вид. 3-тє, перероб. і доп. – 736 с.
20. ДНАОП 0.00-1 31-99 Правила охорони праці під час експлуатації електронно-обчислювальних машин – Київ, 1999 р. – 30 с.
21. НАПБ Б.03.002-2007. Норми визначення категорій приміщень, будинків та зовнішніх установок за вибухопожежною та пожежною небезпекою : Затвердж. наказом МНС від 03.12.2007 № 833.
22. ДБН В.2.5-56-2010 «Інженерне обладнання будинків і споруд. Системи протипожежного захисту» [Електронний ресурс]. – Режим доступу : [www.mns.gov.ua/files/2012/2/9/87\\_\\_56\\_.doc](http://www.mns.gov.ua/files/2012/2/9/87__56_.doc).
23. Стеблюк М.І. Цивільна оборона та цивільний захист: Підручник. – 2-ге вид., переробл. – К.: Знання, 2009 р. – 487 с.
24. ЕМ – алгоритм [Электронный ресурс]. – Режим доступа к ресурсу: <https://ru.wikipedia.org/wiki/ЕМ-алгоритм>

- 25.DBSCAN [Электронный ресурс]. – Режим доступа к ресурсу:  
<http://en.wikipedia.org/wiki/DBSCAN>
26. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.
- 27.COBWEB [Электронный ресурс]. – Режим доступа к ресурсу:  
[http://en.wikipedia.org/wiki/Cobweb\\_%28clustering%29](http://en.wikipedia.org/wiki/Cobweb_%28clustering%29)
- 28.A Tutorial on Clustering Algorithms [Электронный ресурс]. – Режим доступа к ресурсу: [http://home.dei.polimi.it/matteucc/Clustering/tutorial\\_html/kmeans.html](http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html).
- 29.Milenova B. Clustering large databases with numeric and nominal values using orthogonal projections, Oracle Data Mining Technologies / B. Milenova, M. Campos, 2002. – p 317.
- 30.Guha S. CURE: An Efficient Clustering Algorithm for Large Databases, / S. Guha, R. Rastogi, K.Shim //Proc. ACM SIGMOD Int'l Conf. Management of Data, ACM Press, New York.
- 31.Bradley P. Scaling Clustering Algorithms to Large Databases / P. Bradley, U. Fayyad, C. Reina // Proc. 4th Int'l Conf. Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, Calif., 1998. – 180 с.
- 32.Ганти В. Добыча данных в сверхбольших базах данных / В. Ганти, Й. Герке, Р. Рамакришнан // Открытые системы, №9 – 10.
- 33.Метод контроля качества кластеризации мультиспектрального изображения [Электронный ресурс]. – Режим доступа к ресурсу: <http://izv-tn.tti.sfedu.ru/wp-content/uploads/2010/7/31.pdf>
- 34.Рихтер Дж. CLR via C#. Программирование на платформе Microsoft .NET Framework 4.5 на языке C#. 4-е изд. — СПб.: Питер, 2013. — 896 с.: ил. — (Серия «Мастер-класс»).
- 35.Extreme Optimization Numerical Libraries for .NET [Электронный ресурс]. – Режим доступа к ресурсу: <http://www.extremeoptimization.com/>

## ДОДАТКИ

## Додаток А. Графічний матеріал

Додаток Б. Компакт-диск з вихідним кодом