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

До захисту допускається
Т.в.о. завідувача кафедри
_____ Сафонова С.О.
«____»_____ 20__ р.

МАГІСТЕРСЬКА РОБОТА

НА ТЕМУ:

Інформаційна технологія кластерного аналізу слабоструктурованної інформації

Освітній рівень"Магістр"
Спеціальність 122 "Комп'ютерні науки"

Науковий керівник роботи:	_____ (підпис)	О.І.Рязанцев (ініціали, прізвище)
Консультант з охорони праці:	_____ (підпис)	Я.О.Критська (ініціали, прізвище)
Студент:	_____ (підпис)	О.О.Черкасов (ініціали, прізвище)
Група:		КН-18дм

Сєвєродонецьк 2020

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
СХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВОЛОДИМИРА ДАЛЯ

Факультет Інформаційних технологій та електроніки
Кафедра Комп'ютерних наук та інженерії
Освітній рівень магістр
Напрям підготовки _____
(шифр і назва)
Спеціальність 122 "Комп'ютерні науки"
(шифр і назва)

ЗАТВЕРДЖУЮ:

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

З А В Д А Н Н Я
НА МАГІСТЕРСЬКУ РОБОТУ СТУДЕНТУ

Черкасову Олександру Олександровичу
(прізвище, ім'я, по батькові)

1. Тема роботи Інформаційна технологія кластерного аналізу слабоструктурованної інформації

керівник проекту (роботи) Рязанцев Олександр Іванович, д.т.н., проф..
(прізвище, м.'я, по батькові, науковий ступінь, вчене звання)
затверджені наказом вищого навчального закладу від «11» 10 2019 р. № 135/15.15
2. Строк подання студентом роботи 10.01.2020
3. Вихідні дані до роботи Матеріали науково-дослідної практики, математичні моделі аналізу слабоструктурованих даних, теоретичні відомості про методи кластеризації даних.

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити) Сучасний стан питання кластерного аналізу слабоструктурованої інформації, математична модель кластерного аналізу, комп'ютерна модель кластерного аналізу, охорона праці та безпека в надзвичайних ситуаціях, висновки

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

6. Консультанти розділів проекту (роботи)

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Охорона праці та безпека в надзвичайних ситуаціях	Критська Я.О. ст. викл. кафедри КНІ		

7. Дата видачі завдання 14.10.2019

Керівник _____
(підпис)

Завдання прийняв до виконання _____
(підпис)

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів дипломного проекту (роботи)	Строк виконання етапів проекту (роботи)	Примітка
1	Отримання завдання до магістерської роботи	02.09.2019-15.09.2019	
2	Критичний аналіз літератури з досліджуваної проблеми	16.09.2019-22.09.2019	
3	Аналіз технічних засобів	23.09.2019-25.09.2019	
4	Розробка методу	26.09.2019-06.10.2019	
5	Програмна реалізація	07.10.2019-25.11.2019	
6	Розробка частини проекту "Охорона праці та безпеки в надзвичайних ситуаціях"	26.11.2019-1.01.2020	
7	Оформлення пояснювальної записки, автореферату та презентації	2.01.2020-9.01.2020	

Студент _____ О.О.Черкасов
(підпис) (прізвище та ініціали)

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

АНОТАЦІЯ

Черкасов О.О. Інформаційна технологія кластерного аналізу слабоструктурованної інформації.

Метою роботи є розробка методики відновлення слабоструктурованих медичних показників. Реалізація методики в комп'ютерній технологій дозволяє отримувати результати в темпі отримання експериментальних даних (на прикладі медичних даних). Об’єктом дослідження атестаційної роботи є медичні данні функціонального стану серцево-судинної системи. Предметом дослідження є методи і алгоритми кластеризації, що дозволяють відновити загублені медичні дані.

Використано методи аналітичного обґрунтування. Проведено дослідження методів кластерізації даних та їх аналіз. Запропоновано і описано методику, яка призначена для кластеризації слабоструктурованих медичних показників при втраті за допомогою Fuzzy C-Means-алгоритму. Результати дослідження показали, що кластерний аналіз кардіологічних даних може бути використаний для визначення кардіологічного статусу здоров'я людей.

Ключові слова: кластерізація, метод, слабоструктуровані дані, часові ряди, Fuzzy c-means, K-means, Python

ABSTRACT

Cherkasov O.O. Information technology of cluster analysis of poorly structured information.

The object of the study is the sequence of medical data of functional state of the cardiovascular system.

The subject of the study are clustering methods and algorithms, which allows to restore lost medical data.

The goal is to develop a technique of restoring of semi-structured medical values. The implementation of that technique in computer area allows to retrieve the results with the same speed as the experimental data will come (for example: medical data).

Methods of analytical explanations were used. Research of different data clustering methods and its analyzing has been done. Special technique has been proposed and described, which was developed for clustering of restored semi-structured medical data with help of Fuzzy C-Means algorithm. The results of that research showed that cluster analysis of data of the cardiovascular system can be used for definition of people’s heart health status.

Keywords: clustering, semi-structured data, time series, fuzzy c-means, k-means, python.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

FBC – Feature Based Clustering

MFS – Maximal Frequent Sequences

FCM – Fuzzy C-means

ООП – Об’єктно Орієнтоване Програмування

ВСТУП

Аналіз неструктурованих або слабоструктурованих природньомовних текстових даних, є дуже важливим і актуальним завданням з точки зору пошуку прихованих закономірностей у великих обсягах інформації, створення прогностичних моделей різних процесів (в тому числі соціальних і економічних) і вилучення знань. При аналізі великих обсягів неструктурованих або слабоструктурованих текстових даних і пошуку в них прихованих закономірностей виникає проблема відбору ознак суттєвої для аналізу інформації. З одного боку, необхідно, щоб виконання пошуку відбувалося за прийнятний час, а з іншого боку необхідно, щоб суттєві дані не були втрачені. Відбір великої кількості різноманітних ознак призводить до збільшення обчислювальної складності моделі, і як наслідок до необхідності використання значних обчислювальних ресурсів і витрат часу. Однак відкидання ознак, які здаються несуттєвими, або проявляються на рівні шуму, може привести до втрати інформації, значимої для пошуку прихованих закономірностей. Таким чином, апріорі є невідомим, яка частина наявних даних може знадобитися, а яка – ні.

Слід зазначити, що завдання смислової кластеризації документів є однією з ключових в аналізі великих обсягів неструктурованих або слабоструктурованих природньо-мовних текстових даних, і інформаційному пошуку. Зокрема, кластеризація є єдиним рішенням завдання, якщо немає точного уявлення про склад і структури даних. При використанні кластеризації на основі лексичного підходу виникає ряд проблем, пов'язаних з: вибором первинних кластерів, залежністю якості кластеризації від довжини тексту, визначенням загальної кількості кластерів, відсутність зв'язку між близькими за змістом текстами, в яких використовується різна лексика, тощо. Для подолання проблеми, пов'язаної з різницею в лексиці необхідно використовувати методи, засновані не тільки на основі лексичного подібності, а ще й на основі семантичної суміжності або асоціативності.

Метою роботи є розробка методики відновлення слабоструктурованих медичних показників. Реалізація методики в комп'ютерній технологій дозволяє отримувати результати в темпі отримання експериментальних даних (на прикладі медичних даних).

Для досягнення поставленої мети необхідно вирішити ряд задач:

- провести аналіз існуючих методів кластеризації слабоструктурованих даних;
- провести вибір математичної моделі кластерного аналізу;
- розробити методику відновлення слабоструктурованих медичних показників;
- реалізувати комп'ютерну модель для аналізу даних.

1 СУЧАСНИЙ СТАН ПИТАННЯ КЛАСТЕРНОГО АНАЛІЗУ СЛАБОСТРУКТУРОВАНОЇ ІНФОРМАЦІЇ

1.1 Загальний огляд кластерного аналізу

Кластерний аналіз об’єднує об’єкти даних основаних на інформації знайденій тільки в даних, що описують самі об’єкти та їх взаємозв’язок. Мета полягає в тому, що об’єкти всередині групи повинні бути однаковими (або пов’язаними) між собою та відрізнятися (або не бути пов’язаними) з об’єктами в інших групах. Чим більша подібність (чи однорідність) всередині групи та чим більша різниця між групами, тим краще або чіткіша кластеризація.

У багатьох програмах поняття кластера не визначено. Для того, щоб краще зрозуміти складність вирішення того, що являє собою кластер, на рисунку 1.1 відображено 20 точок та 3 різних способи їх кластеризації. Форми маркерів відображають їх приналежність до кластеру. Рисунки 1.1 (б) та 1.1 (г) розділяють дані в 2 та 6 частин відповідно. Однак явний поділ кожного з двох більших кластерів в 3 підкластери може бути просто артефактом людського зору. Також можливо сказати, що не дуже доцільно розділяти дані в чотири кластери, як це відображено на рисунку 1.1 (в). Цей рисунок показує, що визначення кластера може бути неточним та те, що найкраще визначення залежить від характеру даних та бажаних результатів.

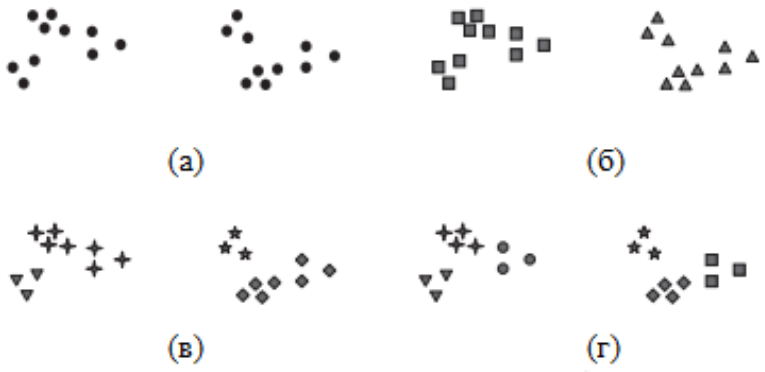


Рисунок 1.1 – Три різних способи кластеризації одних і тих саме даних

Кластерний аналіз пов’язаний з іншими методами, які використовуються для поділення даних в групи. Наприклад, кластеризацію можна розглядати як форму класифікації в тому, що вона створює маркування об’єктів класовими (кластерними) мітками. З цієї причини кластерний аналіз іноді називається безконтрольною класифікацією.

Щоб визначати «схожість» об'єктів, для початку потрібно скласти вектор характеристик для кожного об'єкта – як правило, це набір числових значень, наприклад,

зростання-вага людини. Однак існують також методи, що працюють з якісними (категорійними) характеристиками.

Після того, як ми визначили вектор характеристик, можна провести нормалізації, щоб всі компоненти давали однаковий внесок при розрахунку «відстані». У процесі нормалізації все значення приводяться до деякого діапазону, наприклад, [1, -1] або [0, 1].

Нарешті, для кожної пари об'єктів вимірюється «відстань» між ними - ступінь схожості.

Існує безліч метрик, ось лише основні з них:

1) Евклідова відстань

Найбільш поширена функція відстані. Являє собою геометричним відстанню в багатовимірному просторі:

$$\rho(x, x') = \sqrt{\sum_{i}^{n}(x_i - x_i^{'})^2} \quad . \tag{1.1}$$

2) Квадрат евклідової відстані

Застосовується для додання більшої ваги більш віддаленим один від одного об'єктів.

Це відстань обчислюється таким чином:

$$\rho(x, x') = \sum_{i}^{n}(x_i - x_i^{'})^2 \quad . \tag{1.2}$$

3) Відстань міських кварталів (Манхеттенська відстань)

Це відстань є середнім різниць по координатах. У більшості випадків ця міра відстані приводить до таких же результатів, як і для звичайного відстані Евкліда. Однак для цього заходу вплив окремих великих різниць (викидів) зменшується (тому що вони не зводяться в квадрат). Формула для розрахунку манхеттенської відстані:

$$\rho(x, x') = \sum_{i}^{n}\left|x_i - x_i^{'}\right| \quad . \tag{1.3}$$

4) Відстань Чебишева

Це відстань може виявитися корисним, коли потрібно визначити два об'єкти як «різні», якщо вони розрізняються за якоюсь однією координаті. Відстань Чебишева обчислюється за формулою:

$$\rho(x, x') = \max(|x_i - x_i'|). \quad (1.4)$$

5) Ступінна відстань

Застосовується в разі, коли необхідно збільшити або зменшити вагу, що відноситься до розмірності, для якої відповідні об'єкти сильно відрізняються. Ступінна відстань обчислюється за такою формулою:

$$\rho(x, x') = \sqrt[r]{\sum_{i}^{n}(x_i - x_i')^p}, \quad (1.5)$$

де r і p – параметри, що визначаються користувачем. Параметр p відповідальний за поступове зважування різниць за окремими координатами, параметр r відповідальний за прогресивне зважування великих відстаней між об'єктами. Якщо обидва параметри – r і p – дорівнюють двом, то це відстань збігається з відстанню Евкліда.

Вибір метрики повністю лежить на дослідника, оскільки результати кластеризації можуть істотно відрізнятися при використанні різних заходів.

1.2 Загальний огляд слабоструктурованої інформації

Неефективне управління інформацією веде до збільшення ризиків для різних форм бізнесу: зберігання персональних даних та іншої конфіденційної інформації на загальнодоступних інформаційних ресурсах, поява підозрілих призначених для користувача зашифрованих архівів, порушення політик доступу до важливої інформації та ін.

За цих обставин вміння якісно аналізувати інформацію і оперативно реагувати на будь-які невідповідності її зберігання політикам і вимогам бізнесу є ключовим показником зрілості інформаційної стратегії організації.

Слабоструктурована інформація – це дані, для яких визначені деякі правила і формати, але в найзагальнішому вигляді. Наприклад, рядок з адресою, рядок в прайс-листі, ПІБ і т. п. На відміну від неструктурованих, такі дані з меншим зусиллям перетворюються до структурованої форми, однак без процедури перетворення вони теж непридатні для аналізу.

Слабоструктуровані дані стають важливим об’єктом для досліджень по наступним причинам [1]:

– к таким джерелам даних, як Веб, зручно звертатися як до баз даних, але Веб неможливо «укласти» в якусь схему даних;

– бажано мати достатньо гнучкий формат для обміну даними між різними базами;

– навіть при роботі зі структурованими даними може бути зручно представляти їх у вигляді слабоструктурованих даних з метою навігації по ним.

Аналіз неструктурованих або слабоструктурованих текстових даних є дуже важливим і актуальним завданням з точки зору пошуку прихованих закономірностей у великих обсягах інформації, створюючи Прогнозні моделі різних процесів (у тому числі соціально-економічних) та видобування знань.

При аналізі великих обсягів неструктурованих або слабоструктурованих текстових даних і пошуку прихованих моделей в них, є проблема відбору характеристик, необхідних для аналізу інформації. З одного боку, необхідно, щоб пошук проходив протягом розумного часу, а з іншого боку, необхідні дані не повинні бути втрачені. Підбір великої кількості різних рис призводить до збільшення обчислювальної складності моделі, і як наслідок необхідності використання значних обчислювальних ресурсів і часу. Однак, відкидаючи знаки, які здаються незначними або які з'являються на рівні шуму може призвести до втрати інформації, яка є значущою для пошуку прихованих моделей. Таким чином, апріорі невідомо, яка частина доступних даних може знадобитися, а яка ні. Це завжди проблема для визначення взаємозв'язку ознак шуму з цільовою змінною завдання, яке буде вирішена.

При роботі з великими обсягами слабоструктурованих текстових даних є необхідність виконання пошуку інформації (у тому числі пошук прихованих моделей), вилучення знань і класифікація, яка вимагає створення відповідних моделей для автоматизації цих процесів. Слід зазначити, що в цьому випадку немає фізично вимірних значень, за допомогою яких можна виконувати математичні операції і створювати формальні моделі.

1.3 Огляд існуючих методів кластерного аналізу

Вагомий внесок у розвиток загальної теорії кластерного аналізу зробили Moore A.W., Dubes R.C. (алогритми та методи кластерізації); Ball G.H., Hall D.J., MacQueen J., Lloyd Stuart P. (метод k-середніх); Jordan M.I.; Friedman J. (ієрархічні методи); Hardin R.H., Sloane N.J.A., Smith W.D., Sokal R.R., Sneath, P.H. (центроїдний метод) та інші. Вагомий внесок у розвиток методів кластерного аналізу склали вітчизняні науковці: Дорофеюк А.А., Мучник И.Б., Растригин ДА., Загоруйко Н.Г та інші. Питання одночасного обробки тексту, графіки та звуку в рамках єдиної моделі представлення даних було оброблено вітчизняним дослідником Харламов А.А.

Розроблені методики не враховують можливості одночасної обробки графічних і текстових розділів документів. У той же час, підходи, які використовують аналіз зображень в багатьох документах може надати значну підтримку для пошукових систем. Незважаючи на різні характер текстів і образів, багато методів їх аналізу є спільними.

Умовно методи кластеризації діляться на два класи – ієрархічні і неієрархічні (Рис. 1.2).

Рисунок 1.2 – Методи кластерного аналізу

У неієрархічних алгоритмах є умова зупинки і кількості кластерів. Основою цих алгоритмів є гіпотеза про відносно невелику кількість прихованих чинників, що визначають структуру комунікації між рисами. Ієрархічні алгоритми не прив'язують до кількості кластерів. Ця характеристика визначається динамікою об'єднання і розщеплення кластерів при будівництві вкладеної кластерної дерева.

1.3.1 Огляд ієрархічних методів

Ієрархічні алгоритми поділяються на агломератові, побудовані за допомогою суміщення елементів, тобто зменшення кількості кластерів і дивизивні, що основані на поділу існуючих груп (кластерів).

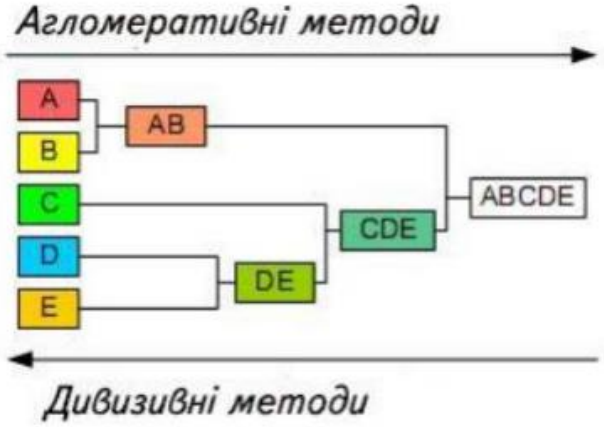


Рисунок 1.3 – Ієрархічні методи кластерного аналізу

В агломератових алгоритмах нові кластери утворюються шляхом об’єднання більш менших кластерів та, таким чином, дерево створюється від листків до стволу. В цих методах використовується підхід «знизу-вгору», в якому ми починаємо з окремих точок та поєднуємо їх в кластери для того, щоб утворити схожу на дерево структуру. Велика кількість варіантів можлива з точки зору об'єднання цих кластерів, що надає різні компромісні варіанти між якістю та ефективністю. Деякими з цих варіантів є методи одиночного, середнього та центроїдного зв’язку. У методі одиночного зв’язку використовується найкоротша відстань між будь-якою парою точок у двох кластері. У методі середнього зв’язку використовується середня відстань серед усіх пар. У методі центроїдного зв'язку використовується відстань між центроїдами. Деякі варіації цих методів мають недоліки в утворюванні ланцюгів, в яких більші кластери природньо зміщенні уперед для того, щоб вони були ближче до інших точок і, отже, залучать послідовно більшу кількість точок. Метод одиночного зв'язку є особливо чутливими до цього явища.

Розділювальні алгоритми формують нові кластери шляхом поділу більших кластерів на більш менші та, таким чином, дерево створюється від стволу до листів. В таких методах використовується підхід «зверху-вниз». Такий метод надає більше гнучкості в ієрархічній структурі та рівню балансу в різних кластерах.

Зазвичай результатом роботи ієрархічного алгоритму є дендрограмма. Дендрограма (dendrogram, від грец. dendron – «дерево») – деревоподібна діаграма, що містить n рівнів, кожний з яких відповідає одному із кроків процесу послідовного укрупнення кластерів. На рисунках 1.4, та 1.5 є можливість побачити приклад формування дендрограмми з сирих даних.

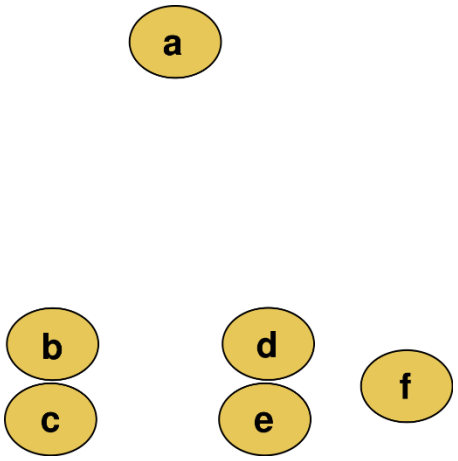


Рисунок 1.4 – Приклад відображення даних до формування дедрограмми

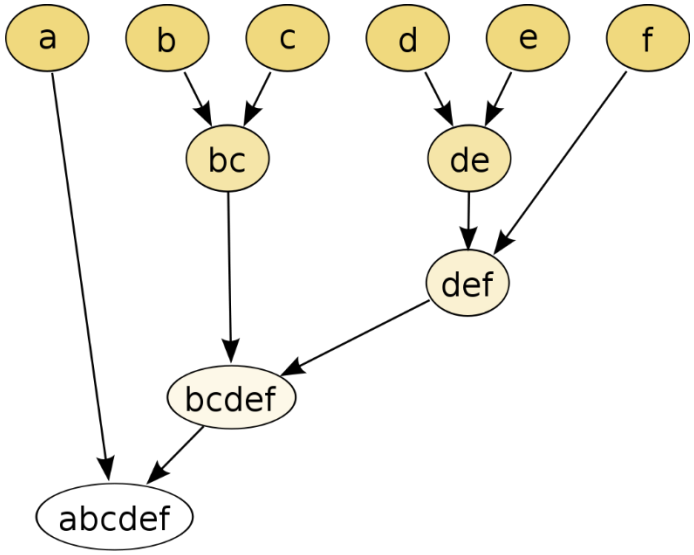


Рисунок 1.5 – Приклад дендрограмми

1.3.2 Огляд неієрархічних методів

Процес неієрархічної кластеризації завжди є ітеративним. У цьому випадку дані поділяються на кілька кластерів за один раз, зазвичай з використанням представників розділу. Вибір представника розділу та функції відстані є вирішальним і регулює поведінку основного алгоритму. У кожній ітерації точок даних присвоюються їхні найближчі представники розбиття, а потім представники налаштовуються згідно з наведеними вказівками, присвоєні класу. Існує інструкція, яка полягає в тому, щоб поєднувати це з ітераційним характером алгоритму EM, в якому виконані м'які завдання на E-кроці, а також параметри моделі (представники аналогічні кластеру) регулюються в M-кроці. Деякі загальні методи створення розділів виглядають таким чином:

– k-середніх: у цих методах представники розбиття відповідають значенню кожного кластера. Зверніть увагу, що представник розділу не відтворюються з вихідного набору даних, але створюється як функція базових даних. Евклідова відстань використовується для того, щоб розраховувати відстань між точками. Метод к-середніх вважається одним з найпростіших і класичних методів кластеризації даних, а також, можливо, один з найбільш широко використовуваних методів в практичній реалізації через свою простоту. Вхідна множина розділяється на К груп, при цьому мінімізується функція, що визначає відстані як суми квадратів помилок:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left|x_i - s_j\right|^2 \tag{1.6}$$

– - k-медіан: у цих методах медіана по кожному вимірюванню, а не середнє значення, використовується для створення представника розбиття. Як і у випадку з підходом k-середніх, представники розбиття не відтворюються з вихідного набору даних. Підхід k-медіан є більш стійким до шуму та перевищенням, оскільки медіана набору значень, як правило, менш чутлива до екстремальних значень даних.

Після отримання результатів кластерного аналізу методом k-середніх потрібно перевірити правильність кластеризації (тобто оцінити, наскільки кластери відрізняються один від одного). Для цього потрібно розрахувати середнє значення для кожного кластера. При хорошій кластеризації ці значення повинні дуже сильно відрізнятися для усіх розрахунків, або хоча б для більшої їх частини. Не дивлячись на простоту використання цього алгоритму, він має і недоліки. Цей алгоритм дуже чутливий до піків, що можуть сильно змінит и середнє значення. Цей недолік може бути виправленим завдяки різноманітним модифікаціям цього алгоритму, наприклад k-медіан. Також, k-середніх може повільно працювати на великих базах даних. Але, це також можливо частково вирішити, якщо використовувати вибірку з бази.

1.3.3 Огляд інших методів

В роботі [3] існують три алгоритми для кластеризації текстових документів: з виділенням функцій, з функцією ваги схеми і динамічного зменшення розмірів. Генетичний алгоритм (GA), гармонійний алгоритм пошуку (HS) і алгоритм оптимізації (PSO) є найбільш успішними методами вибору характеристик, встановлених за допомогою нової схеми

зважування, а саме: вага довжини (LFW), залежить від частоти і зовнішнього вигляду знаків в інших документах. У статті також пропонується новий метод зменшення динамічних вимірів (DDR) для зменшення кількості функцій, що використовуються в кластеризації, і таким чином підвищити ефективність алгоритмів. Алгоритм K-середніх використовується для кластера набору текстових документів на основі термінів (або рис), отриманих динамічним скороченням. Оцінено сім тестових наборів текстових даних різного розміру та складності. Аналіз результатів показує, що оптимізація з вагою довжини і динамічного скорочення дає оптимальні результати для майже всіх наборів даних, що перевіряються.

В роботі [4] впроваджується високомасштабований, вдосконалений алгоритм кластеризації на основі використання n-грам для зменшення високого виміру та отримання високоякісних кластерів тестових документів. Також у статті є порівняльний аналіз, що показує, що для зразків текстових даних наборів з видаленням стоп-слів, запропонований алгоритм працює краще, ніж без видалення стоп-слів.

Автори статті [5] окреслили методи удосконалення роботи системи з видобування заходів з соціальної інформації. На початку роботи, повідомлення оброблювались індивідуально, що викликало багато безглуздих подій у зв'язку з відсутніми деталями, що були розкидані в мільйонах текстових сегментів. Крім того, було проаналізовано значну кількість непотрібних текстів, які підвищували час обробки і зменшували продуктивність системи. У цій статті, була запропонована кластеризація для групування семантично пов'язаних сегментів тексту, фільтрації шуму, зменшення обсягу даних для обробки і просування тільки відповідні сегменти тексту в конвеєрі для вилучення інформації. Алгоритм кластеризації, що був портований в інфраструктуру обробки потоків, називається Storm.

В роботі [6] було розглянуто питання про зростання цифрових новин, що зберігаються в базах даних, і необхідність їх моделювати, щоб полегшити їх розуміння та можливість витягувати з них важливу інформацію. Для спрощення обробки інформації в базі даних, вам необхідна модель і специфічний метод кластеризації новин, заснованих на близькості і характеристиках цифрових новин. Автори пропонують використовувати моделі і методи бази даних графів, алгоритм графів MCL (алгоритм кластеризації Марквова), що дозволяє спростити обробку інформації шляхом визначення характеристик кожної вершини на діаграмі. Отже, при кластеризації цифрових новин процес завищення матриці дуже впливає на час алгоритму MCL, а процес наддування матриці впливає на утворення числа кластерів.

В роботі [7] підкреслюється важливість спільної кластеризації, особливо при розгляді розріджених великих даних. У статті представлена нова генеративна модель Sparse Poisson Latent Block Model (SPLBM), заснована на розподілі Пуассона, яке виникає для таблиць, таких як матриця документів-термів. Автори стверджують, що SPLBM має дві великі переваги. По-перше, це сувора статистична модель. По-друге, алгоритм був розроблений з

нуля, щоб впоратися з проблемами розрідженості даних. Як наслідок, крім пошуку однорідних блоків, як і інших доступних алгоритмів, він також відфільтровує однорідні, але галасливі через розрідженості дані. Експерименти з різних наборів даних різного розміру і структури показують, що алгоритм, заснований на SPLBM, явно перевершує сучасні алгоритми. Зокрема, представлений тут алгоритм на основі SPLBM процвітає в отриманні природною кластерної структури складних незбалансованих наборів даних, які інші відомі алгоритми не можуть ефективно обробляти.

У статті [8] представлений новий метод автоматичного виявлення та відстеження тем новин з новинного архіву мультимодального ТВ. Автори пропонують тему і графік (MT-AOG) для спільного подання текстових і візуальних елементів новин і їх прихованих тематичних структур. MT-AOG використовує граматику, що залежить від контексту, яка може описувати ієрархічний склад тем новин по семантичним елементам, що стосуються залучених людей, пов'язаних з ними місць і того, що сталося, і моделювати контекстуальні відносини між елементами в ієрархії. У статті описано як за допомогою процесу кластеризації, який об'єднує історії про близькі події, були виділені новинні теми. Спочатку новинні документи кластеризуються, потім виявлені теми безперервно відслідковуються і оновлюються вхідними потоками новин. Генеруються траєкторії теми, щоб показати, як теми з'являються, еволюціонують і зникають з плином часу.

Автори [9] розглядають способи кластеризації документів за темами. Вони провели ряд експериментів на різних реальних і штучних наборах даних, таких як NEWS 20, Reuters, електронні листи, дослідні роботи з різних тем. Використовується алгоритм TFI-DF з Fuzzy K-Means. Спочатку експерименти були виконані на невеликому наборі новинних даних. Далі, найкращий алгоритм був застосований до розширеного набору даних. Поряд з різними кластерами відповідних документів представлений наведений коефіцієнт, ентропія і тренд f мер, щоб показати поведінку алгоритму для кожного набору даних.

У статті [10] пропонується новий паралельний дизайн недавно з'явилася евристики для неконтрольованої і жорсткої кластеризації. Вихідний метод знаходить кластери шляхом розбиття категоріальних великих наборів даних відповідно до лінійної математичної моделлю, відомої як реляційний аналіз. Автори поліпшили цей метод для обробки кластеризації документів з ітеративним використанням моделі програмування MapReduce. Ефективність запропонованого методу оцінюється з використанням набору даних BBC, наданого сайтом BBC News. І експерименти показали, що цей метод дає результати якості з низькою обчислювальної вартістю.

В роботі [12] вирішується так звана «нова проблема користувача». Вона полягає в необхідності отримати деякі дані про нового користувача, щоб почати робити персоналізовані пропозиції. Автори намагаються вирішити нову проблему користувачів за допомогою унікальної персоналізованої стратегії і вивчають можливе поліпшення

результатів кластеризації документів і, зокрема, кластеризацію новинних статей з Інтернету при використанні текстових n-грам під час фази вилучення ключових слів. У статті представляється і оцінюється автоматизований підхід, який об'єднує кластеризацію новинних статей, отриманих з Інтернету, використовуючи n-грами. Потім цей метод порівнюється з методом «мішок слів», яке раніше був використаний авторами в якості алгоритму кластеризації (W-kmeans). Експерименти показують, що шляхом точного налаштування вагових параметрів між ключовим словом і n-грамами, а також самого значення n можна домогтися значного поліпшення щодо показників результатів кластеризації.

Автори роботи [13] пропонує ідею застосування кластеризації документів на основі алгоритму Feature Based Clustering (FBC), тому що вважають, що буде простіше використовувати величезну кількість існуючих документів, якщо вони об'єднані в кілька тем. FBC для кластеризації функцій послідовних даних використовує алгоритм K-Means. Особливості текстового документа можуть бути представлені як послідовність слів. Для обробки у вигляді послідовних даних функції повинні бути вилучені з колекції неструктурованих текстових документів. Тому авторами ставляться завдання попередньої обробки даних, щоб надати для кластеризації відповідні форми документів. У статті розглядається два типи послідовного шаблону з використанням простої форми: часта послідовність слів (FWS) і максимальна часта послідовність (MFS). Обидва типи підходять для текстових даних. Різниця полягає в застосуванні принципу максимуму в MFS. Отже, сума MFS з текстового документа буде менше суми його FWS. У цьому дослідженні автори вибирають максимальні частотні послідовності (MFS) як уявлення функції і пропонують структуру для проведення FBC з використанням MFS в якості функцій. Структура тестується на наборі даних кластерів, який є підмножиною текстових даних групи новин. Результат показує, що на точність результату кластеризації впливає значення параметра, набір даних і кількість цільового кластера.

1.4 Постановка завдання дослідження

Одним з актуальних завдань математичного моделювання на основі аналізу даних є задача кластеризації об'єктів, інформація про які представлена у вигляді багатовимірних різноманітних часових рядів. Під рядами різних типів розуміють ряди, які містять набори для виміру реальних (кількісних), порядкових, номінальних або логічних (якісних) змінних. Ці види завдань виникають при побудові моделей об'єктів у важкоформних зонах досліджень, наприклад, у медицині, коли необхідно описати типові групи пацієнтів з аналогічною

динамікою розвитку захворювань на основі даних про зміни клінічних показників і діагностичних рис. Типізація пацієнтів дозволяє, зокрема, розробляти методи лікування, оптимальні для кожної групи.

У завданні кластерного аналізу потрібно розбивати багато об'єктів, описаних набором деяких змінних (або матрицею попарних відстаней), до відносно невеликої кількості кластерів, так що критерій якості групування набуває найкращого значення. Кількість кластерів може бути попередньо відібрана або не встановлена (в останньому випадку оптимальну кількість груп потрібно визначати автоматично). Під критерієм якості зазвичай розуміють деякий функціонал залежний від розбросу об'єктів в межах групи та відстаней між групами.

Метою роботи є розробка методики відновлення слабоструктурованих медичних показників. Реалізація методики в комп'ютерній технологій дозволяє отримувати результати в темпі отримання експериментальних даних (на прикладі медичних даних).

Об’єктом дослідження атестаційної роботи є медичні данні функціонального стану серцево-судинної системи.

Предметом дослідження є методи і алгоритми кластеризації, що дозволяють відновити загублені медичні дані.

Для досягнення поставленої мети необхідно вирішити ряд задач:

- провести аналіз проблеми за тематикою роботи;
- провести огляд літературних джерел за тематикою роботи;
- провести аналіз існуючих методів кластеризації слабоструктурованих даних;
- провести вибір математичної моделі кластерного аналізу;
- розробити методику відновлення слабоструктурованих медичних показників;
- реалізувати комп'ютерну модель для аналізу даних.

2. МАТЕМАТИЧНА МОДЕЛЬ КЛАСТЕРНОГО АНАЛІЗУ

Аналіз часових рядів – сукупність математико-статистичних методів аналізу, призначених для виявлення структури часових рядів і для їх прогнозування. Сюди належать, зокрема, методи регресійного аналізу. Виявлення структури часового ряду необхідно для того, щоб побудувати математичну модель того явища, яке є джерелом аналізованого часового ряду. Прогноз майбутніх значень часового ряду використовується для ефективного прийняття рішень.

Метод аналізу часового ряду визначається, з одного боку, цілями аналізу, а з іншого боку, ймовірнісної природою формування його значень.

При практичному аналізі часових рядів послідовно проходять такі етапи [19]:

1) Опис поведінки часового ряду;

2) Виділення та видалення закономірних складових часового ряду, що залежать від часу: тренда, сезонних і циклічних складових;

3) Виділення та видалення низько- або високочастотних складових процесу (фільтрація);

4) Дослідження випадкової складової часового ряду, що залишилася після видалення перерахованих вище складових;

5) Побудова (підбір) математичної моделі для опису випадкової складової і перевірка її адекватності;

6) Прогнозування майбутнього розвитку процесу, представленого часовим рядом.

Це і є основною проблемою, оскільки забагато трудомістких етапів, для отримання результату. Що ускладнює роботу, тому існує сенс спростити цю процедуру за допомогою комп′ютеризування процесу. Адже більшу частину роботи виконують фахівці вручну, збір інформації займає декілька місяців.

Існує така проблема як те що, добре підібрати математичну модель вдається не для всякого часового ряду. Нерідко буває і так, що для опису підходять відразу декілька моделей. Неоднозначність вибору моделі може спостерігатися як на етапі виділення детермінованого компонента ряду, так і при виборі структури ряду залишків. Тому досить часто розробляють декілька прогнозів, зроблених за допомогою різних моделей. Що знову ж таки збільшує трудоємкість, та збільшує кількість часу роботи над результатом.

Вибираючи між ієрархічними й неієрархічними методами, потрібно враховувати їхні особливості. Неієрархічні методи виявляють вищу стійкість щодо шумів і викидів, некоректного вибору метрики, введення незначущих змінних у набір, що бере участь у кластеризації. Фахівець повинен заздалегідь визначити кількість кластерів, кількість ітерацій, правило зупинки, а також деякі інші параметри кластеризації. Ієрархічні методи,

відмовляються від визначення кількості кластерів, а будують повне дерево вкладених кластерів. Вони порівняно з неієрархічними методами демонструють наочність і можливість одержати детальне подання структури даних. Ієрархічні методи не можуть працювати з великими наборами даних. Результати кластеризації можуть не мати достатнього статистичного обґрунтування.

Для вирішення поставленої задачі використаємо неієрархічні алгоритми, тому що вони базуються на певному завданні оптимізації, тобто, групування початкового набору об'єктів в кластерах є вирішенням деяких крайніх проблем.

Розглянемо деякі з найпопулярніших методів:

– k-середніх;

– FCM.

2.1 Математична модель кластерного аналізу методом k-середніх

Одним з найпопулярніших методів чисельного аналізу є метод найменших квадратів. Для завдання кластеризації він виглядає наступним чином:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left|x_i - s_j\right|^2 \to \min \quad . \tag{2.1}$$

Числова реалізація цієї задачі називається методом k-середніх.

Ідея методу полягає в тому, що спочатку обирається k довільних початкових центрів. Потім всі об'єкти діляться на k груп, що найближче усього знаходяться до відповідного центру. Наступний крок обчислює центри знайдених кластерів. Процедура повторюється ітеративно, доки стабілізуються центри кластера. Алгоритм розділення об'єктів $x_i(i = 0,1,...,n)$ базується на мінімізації відстані між кластерами, у випадку, якщо в якості відстані використовується Евклідова відстань, то цільовою функцією можна вважати:

$$S = \sum_{j=1}^{k}\sum\left\{\left|x_i - \mu_j\right|^2, x_i \in c_j\right\}, \tag{2.2}$$

де x_i – і-й об’єкт, а c_j являє собою j-й кластер з центром μ_j.

2.1.1 Метод k-середніх

Алгоритм k-середніх прагне звести до мінімуму загальне квадратичне відхилення об'єктів кластера від центрів цих кластерів, визначених формулою 2.2.

Досягнення глобальної мінімальної S не гарантується, оскільки алгоритм може завершити свою роботу в одному з локальних мінімумів. Результат кластеризації залежить від вибору початкових центрів кластера. Необхідно заздалегідь знати кількість кластерів.

Кроки виконання алгоритму наступні:

Алгоритм починається з відбору первинних центрів кластерів. Вони можуть бути обрані будь-яким способом, наприклад, випадково або на основі аналізу вхідних даних.

Потім кожна ітерація розбиває об'єкти на кластери в співвідношенні із тим, який об'єкт був ближче до центру маси на основі обраної метрики, а потім для кожного кластера переобчислюється центр мас.

Алгоритм закінчується, коли при переобчисленні центрів кластерів пвони не змінюються.

2.1.2 Реалізація методу кластеризації k-середніх

Почнемо розгляд методу з функції, яка власне і запускає сам метод кластеризації – k-середніх.

Вхідними параметрами для неї є масив об’єктів для кластеризації (в нашому випадку медичні дані), кількість кластерів. Значення, що повертається – масив об’єктів класу Cluster.

Алгоритм буде завершено тоді, коли центри кластерів після перерахунку перестануть змінюватись.

Кожен крок методу такий:

1) Розраховуємо відстані об’єкту до центрів кластерів та додаємо об’єкт до кластеру відстань до якого буде найменшою (лістинг 2.1).

```
for point in data:
    distances = dist(point, centers)
    cluster_index = np.argmin(distances)
    clusters[cluster_index].points.append(point)
```

Рисунок 2.1 – Лістинг приєднання об’єктів до кластерів.

2) Перераховуємо центри кластерів на основі нових точок (лістинг 2.2). Також розраховуємо відстань між попереднім центром кластера та новим. Це необхідно для того, щоб закінчити виконання алгоритму.

```
def recalculate_center(cluster):
    if np.size(cluster.points) == 0:
        return
    new_center = np.mean(cluster.points, axis=0)
    cluster.error = dist(cluster.center, new_center, None)
    cluster.center = new_center
```

Рисунок 2.2 – Лістинг розрахунку нових центрів.

3) Складемо усі результати відстаней між попередніми центрами та новими центрами кластерів. Необхідно, щоб отриманий результат дорівнював 0 для того, щоб було можливо закінчити виконання алгоритму (лістинг 2.3).

```
result = sum(map(lambda it: it.error, clusters))
if result == 0.:
    break
```

Рисунок 2.3 – Лістинг кінця алгоритму

Демонстрація результатів роботи методу k-середніх, що складається з 1080 точок, та поділу його на 3 кластера представлена на рисунку 2.4.

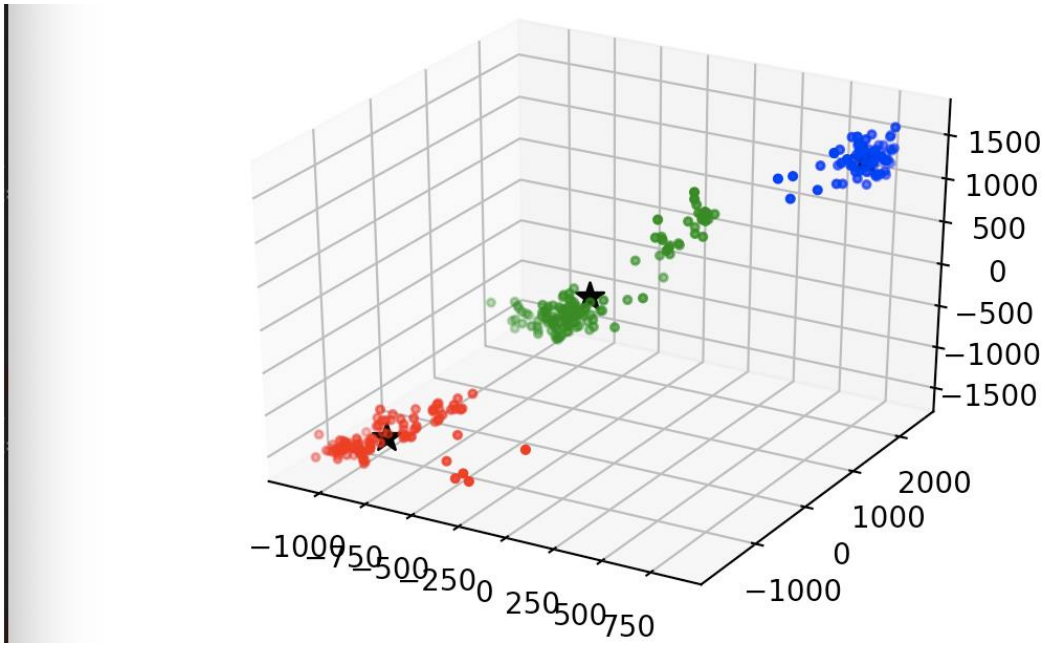


Рисунок 2.4 – приклад виконання алгоритму k-середніх

2.2 Математична модель кластерного аналізу методом FCM

Метод нечіткої кластеризації називають FCM-методом (Fuzzy Classifier Means, Fuzzy C-Means). Метою FCM-методу кластеризації є автоматична класифікація безлічі об'єктів, які задаються векторами ознак в просторі ознак. Іншими словами, такий метод визначає кластери і відповідно класифікує об'єкти. Кластери представляються нечіткими множинами, і, крім того, кордони між кластерами також є нечіткими.

FCM-метод кластеризації передбачає, що об'єкти належать всім кластерам з певною приналежністю. Ступінь приналежності визначається відстанню від об'єкта до відповідних кластерних центрів. Даний метод ітераційно обчислює центри кластерів і нові ступені приналежності об'єктів.

Для заданої множини До вхідних векторів x_k і N виділяються кластерів c_j передбачається, що будь-який x_k належить будь-якому c_j з приналежністю μ_{jk} інтервалу [0,1], де j - номер кластера, а k - номер вхідного вектора.

Беруться до уваги наступні умови нормування для μ_{jk}:

$$\sum_{j=1}^{N}\mu_{jk}=1,\forall k=1,...,K \quad ; \tag{2.3}$$

$$0<\sum_{k=1}^{N}\mu_{jk}\le K,\forall j=1,...,N\,. \tag{2.4}$$

Мета методу – мінімізація суми всіх зважених відстаней $\left\|x_k-c_j\right\|$:

$$\sum_{j=1}^{N}\sum_{k=1}^{K}(\mu_{jk})^q \parallel x_k-c_j \parallel \rightarrow min \quad , \tag{2.5}$$

де q – фіксований параметр, що задається перед ітераціями.

Для досягнення вищевказаної мети необхідно вирішити наступну систему рівнянь:

$$\partial/\partial\mu_{jk}\left(\sum_{j=1}^{N}\sum_{k=1}^{K}(\mu_{jk})^q \parallel x_k-c_j \parallel\right)=0, \tag{2.6}$$

$$\partial / \partial c_j \left(\sum_{j=1}^{N} \sum_{k=1}^{K} (\mu_{jk})^q \| x_k - c_j \| \right) = 0. \quad (2.7)$$

Спільно з умовами нормування μ_{jk} дана система диференціальних рівнянь має наступне рішення:

$$c_j = \frac{\sum_{j=1}^{N} (\mu_{jk})^q x_k}{\sum_{k=1}^{N} (\mu_{jk})^q}, \quad (2.8)$$

(зважений центр гравітації) і

$$\mu_{jk} = \frac{1 / \left\| x_k - c_j \right\|^{1/(q-1)}}{\sum_{j=1}^{N} \left(1 / \left\| x_k - c_j \right\|^{1/(q-1)} \right)}. \quad (2.9)$$

2.2.1 Метод FCM

Крок 1. Ініціалізація.

Вибираються наступні параметри:

– необхідну кількість кластерів *N, 2 <N <K*;

– міра відстаней, як Евклідова відстань;

– фіксований параметр *q* (зазвичай 2);

– початкова (на нульовій ітерації) матриця приналежності $U^{(0)} = (\mu_{jk})^{(0)}$ об'єктів x_k з урахуванням заданих початкових центрів кластерів c_j.

Крок 2. Регулювання позицій $c_j^{(t)}$ центрів кластерів.

На *t*-м ітераційному кроці при відомій матриці $\mu_{jk}^{(t)}$ обчислюється $c_j^{(t)}$ (формула 2.8).

Крок 3. Коригування значень приналежності μ_{jk}.

З огляду на відомі $c_j^{(t)}$, обчислюються $\mu_{jk}^{(t)}$ (формула 2.9).

Крок 4. Зупинка методу

Метод нечіткої кластеризації зупиняється при виконанні наступної умови:

$$\left\| U^{(t+1)} - U^{(t)} \right\| \leq \varepsilon, \tag{2.10}$$

де || || - матрична норма (наприклад, Евклідова норма);

e – заздалегідь задається рівень точності.

2.2.2 Реалізація методу кластеризації FCM

Почнемо розгляд методу з функції, яка власне і запускає сам метод кластеризації – FCM.

Вхідними параметрами для неї є масив об’єктів для кластеризації (в нашому випадку медичні дані), кількість кластерів, коефіцієнт невизначеності, максимальна кількість ітерацій та мінімальна помилка. Значення, що повертається – масив об’єктів класу Cluster.

Кроки по знаходженню нових центрів кластерів і перерахунку матриці приналежності обмежені максимальною кількістю ітерацій, що за замовчуванням дорівнює 100 та мінімальним значенням помилки, що за замовчуванням дорівнює 0,01.

Кожен крок методу такий:

1) розраховуємо базові значення для центрів кластерів та u_matrix (лістинг 2.6).

```
for i in range(k):
    clusters.append(Cluster(data.max(), data.shape[1]))
centers = np.array(list(map(lambda it: it.center, clusters)))
print(centers)
u_matrix = np.random.rand(data.shape[0], k)
u_matrix = u_matrix / np.sum(u_matrix, axis=0, keepdims=True)
```

Рисунок 2.6 – Лістинг створення масиву кластерів та u_matrix

2) Перераховуємо центри кластерів для і-того кроку (лістинг 2.7).

```
for [i, cluster] in enumerate(clusters):
    tmp = u_matrix[:, i]
    multiplication = data.T * (tmp ** m)
```

```
sum1 = np.sum(multiplication, axis=1)
sum2 = np.sum(tmp ** m, axis=0)
cluster.center = sum1 / sum2
```

Рисунок 2.7 – Лістинг розрахунку центрів кластерів.

3) Розраховуємо ваги для кожної точки в залежності від нових центрів кластерів (лістинг 2.8).

```
for [i, point] in enumerate(data):
    for [j, cluster] in enumerate(clusters):
        def calculate(n):
            delta1 = dist(point, cluster.center, None)
            delta2 = dist(point, clusters[n].center, None)
            delta = (delta1 / delta2) ** (2. / (m - 1))
            return delta
        tmp = sum(
            [calculate(n) for n
             in
             range(len(clusters))]
        )
        u_matrix[i][j] = 1 / tmp
```

Рисунок 2.8 – Лістинг розрахунку ваг для кожної точки на і-тому кроці.

4) Розраховуємо помилку на даному кроці та перевіряємо, чи необхідно закінчувати виконання алгоритму та розподіляти точки поміж кластерів в залежності від їх ваг (лістинг 2.9).

```
error = dist(u_matrix, u_matrix_old, None)
    print(f"iteration {iterations}, error = {error}")
    if error < minimal_error:
        break
    iterations = iterations + 1
for [i, point] in enumerate(data):
    clusters[np.argmax(u_matrix[i])].points.append(point)
```

Рисунок 2.9 – Лістинг кінця алгоритму

Демонстрація результатів роботи методу на масиві даних, що складається з 300 точок, та поділу його на 3 кластери представлена на рисунку 2.10.

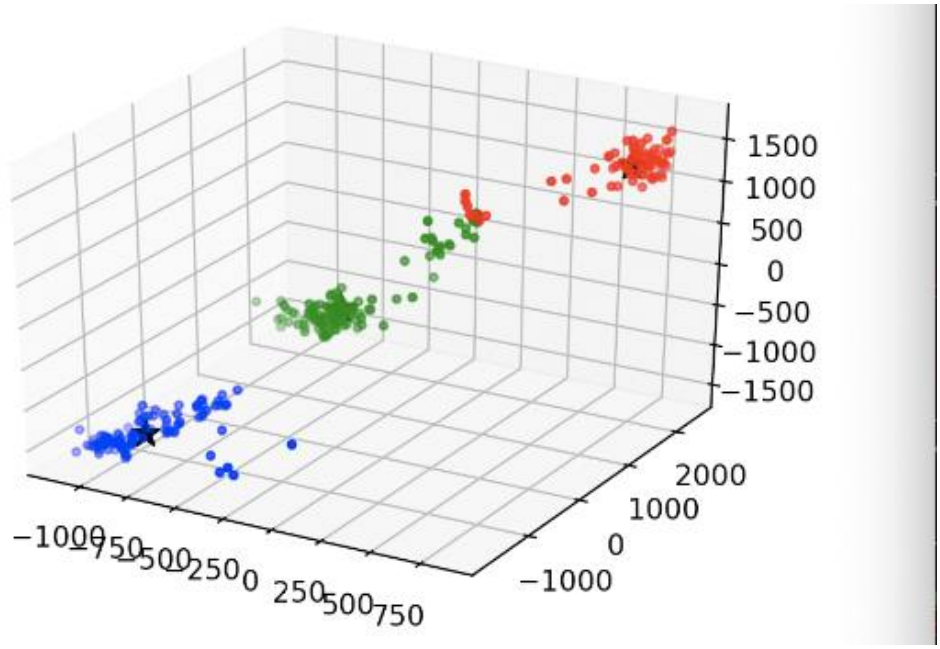


Рисунок 2.10 – приклад виконання алгоритму FCM

2.2.3 Перевірка результатів

Для перевірки правильності виконання методу використаємо готову реалізацію цього алгоритму з пакету skfuzzy [14]. Демонстрація результату представлена на рисунку 2.11.

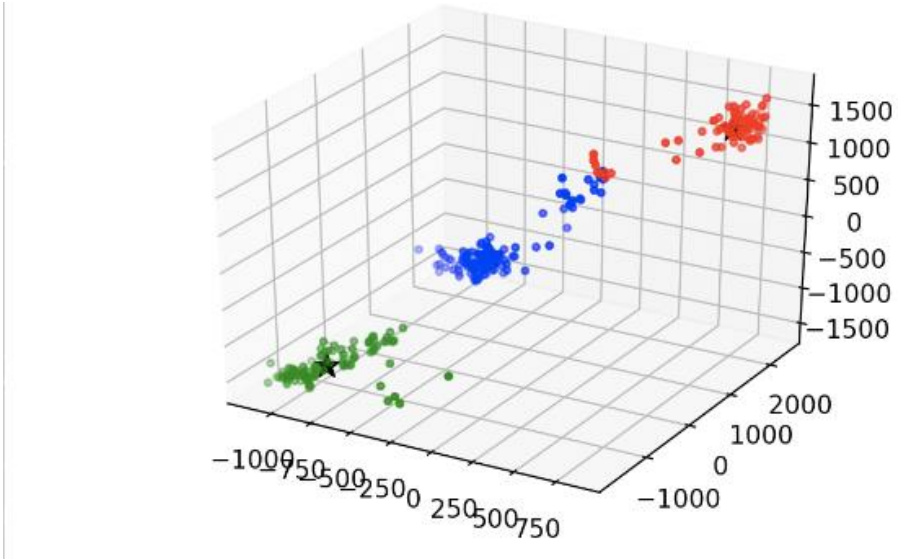


Рисунок 2.11 – приклад виконання алгоритму FCM за допомогою бібліотеки

Згідно з рисунком 2.11 видно, що метод відпрацював вірно та кластери були сформовані однаково.

2.3 Оцінка кластеризації

Для оцінки результатів існує кілька індексів оцінки якості кластеризації, нижче приведені деякі з них.

2.3.1 Індекс оцінки силуета (Silhouette index)

«Силует» кожного кластеру можна визначити наступним чином: припустимо елемент x_j належить кластеру c_p. Позначимо середню відстань від цього об'єкта до інших об'єктів з цього ж кластера c_p через a_{pj}. Тепер позначимо середню відстань від x_j до об'єктів з іншого кластера c_q, $q \neq p$ через $d_{q,j}$. Покладемо $b_{pj} = \min_{q \neq p} d_{qj}$. Сенс цієї величини можна визначити як міру несхожості окремого елемента з елементами найближчого кластера. Таким чином, «силует» кожного окремого елемента визначається як

$$S_{xj} = \frac{b_{pj} - a_{pj}}{\max(a_{pj}, b_{pj})}. \quad (2.11)$$

Знаменник введений з метою нормалізації. Очевидно, що високе значення показника S_{xj} характеризує собою «кращу» приналежність елемента x_j до кластеру p. Оцінка для всієї кластерної структури досягається усередненням показника за елементами:

$$SWC = \frac{1}{N} \sum_{j=1}^{N} S_{xj}. \quad (2.12)$$

Найкраще розбиття характеризується максимальним SWC, що досягається коли відстані всередині кластера a_{pj} мало, а відстань між елементами сусідніх кластерів b_{pj} велика. Також на практиці використовуються варіації індексу силуету: спрощений силует (Simplified Silhouette) і альтернативний силует (Alternative Silhouette). У першому випадку при визначенні a_{pj} і b_{pj} використовуються відстані не між одним елементом і всіма іншими елементами кластерів, а між елементом і центроїдом відповідного кластера. У другому

випадку $S_{xj} = \frac{b_{pj}}{a_{pj} + \varepsilon}$ – мала константа, введена для того щоб уникнути поділу на 0, якщо $a_{pj} = 0$.

2.3.3 Calinski-Harabasz індекс

Нехай $\overline{d}^2$ – середній квадрат відстані між елементами в кластеризуємій безлічі і $\overline{d}_{ci}^2$ – середній квадрат відстані між елементами в кластері c_i. Тоді сума відстаней всередині груп

$$WGSS = \frac{1}{2}\sum_{i=1}^{c}(n_{ci} - 1)\overline{d}_{ci}^2, \tag{2.13}$$

І сума відстаней між групами

$$BGSS = \frac{1}{2}((c-1)\overline{d}^2 + (N-c)A_c), \tag{2.14}$$

Де $A_c = \frac{1}{N-c}\sum_{i=1}^{c}(n_{ci} - 1)(\overline{d}^2 - \overline{d}_{ci}^2)$ – зважена середня різниця відстаней між центрами кластерів і загальним центром безлічі. Тоді визначимо формулу індексу як

$$VRC = \frac{\frac{BGSS}{c-1}}{\frac{WGSS}{N-c}} = \frac{\overline{d}^2 + \frac{N-c}{c-1}A_c}{\overline{d}^2 - A_c} = \frac{1 + \frac{N-c}{c-1}a_c}{1 - a_c}, \tag{2.15}$$

Де $a_c = \frac{A_c}{\overline{d}^2}$. Легко бачити, що якщо все відстані між точками однакові, то $a_c = 0$ і VRC = 1. $a_c = 1$ виключно для «ідеальної» кластеризації, коли в кластерах немає ніякого відхилення. При нормальному розподілі даних a_c повільно, але невпинно зростає при збільшенні c. Однак, VRC убуває при постійному a_c і зростаючому c, що трошки балансує зростання a_c в разі нормального розподілу. Максимальне значення індексу VRC відповідає оптимальній структурі кластерів.

2.3.3 Індекс Девіда-Болдуіна

Нехай $S_i = \left\{ \frac{1}{n_{c_i}} \sum_{x \in c_i} \left\| x - v_i \right\|^q \right\}^{1/q}$ – міра розподілу всередині кластера c_i і $d_{ij} = \left\{ \sum_{k=1}^{d} (v_i^k - v_j^k)^p \right\}^{1/p}$ – міра відмінності між кластерами (dissimilarity measure), тоді мірою схожості між кластерами c_i і c_j може бути будь-яка функція R_{ij}, яка задовільняє такі умови:

1) $R_{ij} \geq 0$;
2) $R_{ij} = R_{ji}$;
3) при $S_i = 0$ та $S_j = 0$ $R_{ij} = 0$;
4) при $S_j > S_k$ та $d_{ij} = d_{ik}$ $R_{ij} > R_{ik}$;
5) при $S_j = S_k$ та $d_{ij} < d_{ik}$ $R_{ij} > R_{ik}$.

Самі автори пропонують наступний варіант підрахунку схожості

$$R_{ij} = \frac{S_i + S_j}{d_{ij}}, \tag{2.16}$$

Тоді сам індекс обчислюється за формулою

$$DB = \frac{1}{c} \sum_{i=1}^{c} R_i, \tag{2.17}$$

де $R_i = \max_{i, j \in \{1 \ldots c\}, i \neq j} (R_{ij})$. Як видно з визначення, DB індекс визначає середню схожість між кластером c_i і найбільш близьким до нього кластером. Оскільки мається на увазі, що кластери в структурі значно відрізняються один від одного, найкращою буде структура з мінімальним DB. Існують також інші варіанти підрахунку розкиду всередині кластера, в часності, із застосуванням такої ж методики на гра фаз, як і для індексу Данна.

2.4 Розробка методики відновлення втрачених даних

Для початку розробки методики необхідно проаналізувати вихідні дані. Вхідними параметрами для кластеризації є масив об’єктів, в нашому випадку медичні дані, що представлені у вигляді часових рядів, які є слабоструктурованою інформацією.

Другим кроком є аналіз методів кластеризації для запропонованих даних. Так як масив даних представлений у вигляді часових рядів необхідно використовувати неієрархічні алгоритми.

На основі аналізу методів кластеризації необхідно реалізувати відібрані алгоритми та відібрати той, що найкраще кластеризує данні. Для реалізації доцільніше використовувати методи FCM та k-means.

В представленому випадку частина вихідних даних була втрачена, проте для того щоб поставити діагноз необхідно мати усі дані.

Отже, запропонована наступна методика відновлення втрачених даних:

1) Реалізувати метод FCM для кластеризації даних;
2) Проаналізувати втрачені данні;
3) Програмно викинути max та min значення;
4) Усереднити значення, що залишилися;
5) Відобразити отримані результати.

3 КОМП'ЮТЕРНА МОДЕЛЬ КЛАСТЕРНОГО АНАЛІЗУ

3.1 Обґрунтування вибору середовища програмної реалізації

Найпоширеніші мови програмування для реалізації алгоритмів кластерного аналізу наступні:

- мова програмування Matlab;
- мова програмування R;
- мова програмування Python.

3.1.1 Мова програмування Matlab

MATLAB — це високорівнева мова та інтерактивне середовище для програмування, чисельних розрахунків та візуалізації результатів. Використовуючи MATLAB, ви можете аналізувати дані, розробляти алгоритми, створювати моделі та додатки [18].

MATLAB широко використовується в таких областях, як:

- обробка та комунікації сигналів;
- обробка зображень і відео,
- систем управління,
- автоматизація тестування та вимірювання,
- фінансова інженерія,
- обчислювальна біологія та ін.

В середовищі математичного моделювання MATLAB, на додаток до вбудованих функцій кластеризації, ви можете реалізувати свій власний алгоритм за допомогою високорівневої об'єктно-орієнтованої мови програмування MATLAB [15].

Середовище математичного моделювання MATLAB знаходить широке застосування як в науковій, так і в виховній діяльності. Наприклад, [18] обговорює аспекти програмування в C і в системі MATLAB.

3.1.2 Мова програмування R

R є мовою програмування для статистичної обробки даних і графіки, але в той же час це безкоштовне програмне забезпечення з відкритим сирцевим кодом розроблений в рамках проекту GNU.

R використовується там, де потрібно працювати з даними. Це не тільки статистика в вузькому розумінні слова, але і "первинний" аналіз (графіки, таблиці спряженості), та передове математичне моделювання. R може бути використаний без особливих проблем в тих місцях, де й комерційні програми аналізу рівня MatLab/Octave. З іншого боку, цілком природньо, що основна обчислювальна потужність R найкраще проявляється у статистичному аналізі: від обчислення середніх значень до вейвлет-перетворень часових рядів [21].

Середовище розробки R та основні базові пакети надають можливість провести кластерний аналіз за допомогою алгоритму k-means, а графічні можливості візуалізують відображення результатів згідно принципам аналізу.

Результатом задачі кластеризації за допомогою алгоритму k-means є векторні набори даних, візуалізація яких графічно з вказівкою центрів k, які були знайдені в ході вирішення задач.

Отже, мова та середовище R дає широкий спектр послуг, інструментів та пакетів для аналізу даних.

3.1.3 Мова програмування Python

Python – це високорівнева інтерпретована мова програмування, яка фокусується на вдосконаленні продуктивності розробників. Історія Python сходить до 1980-х, коли член центру математики та інформатики в Нідерландах Гвідо ван Россум створив першу версію мови [18].

Python є об'єктно-орієнтованою мовою програмування та підтримує такі поняття, як поліморфізм, перевантаження операторів і множинне спадкування. Об'єктно-орієнтована природа Python, є могутнім засобом структурування програмного коду багаторазового користування, крім того, робить цю мову ідеальним інструментом підтримки сценаріїв для об'єктно-орієнтованих мов, таких як C++ і Java.

Python може використовуватися і розповсюджуватися абсолютно безкоштовно. Немає ніяких обмежень на його копіювання, вбудовування в свої системи або поширення в складі ваших продуктів.

Стандартна реалізація мови Python написана на переносимому ANSI З, завдяки чому він компілюється і працює практично на всіх основних платформах. Наприклад, програми на мові Python можуть виконуватися на найширшому спектрі пристроїв, починаючи від наладонних комп'ютерів (PDA) і закінчуючи суперкомп'ютерами.

Крім самого інтерпретатора мови в складі Python поширюється стандартна бібліотека модулів, яка також реалізована належним шляхом. Крім того, програми на мові Python компілюються в байт-код, який однаково добре працює на будь-яких платформах, де встановлена сумісна версія Python. Все це означає, що програми на мові Python, що використовують основні можливості мови і стандартні бібліотеки, будуть працювати однаково і в Linux, і в Windows, і в будь-яких інших операційних системах, де встановлений інтерпретатор Python.

З точки зору функціональних можливостей Python можна назвати гібридом. Його інструментальні засоби вкладаються в діапазон між традиційними мовами сценаріїв (такими як Tel, Scheme і Perl) і мовами розробки програмних систем (такими як C, C ++ і Java). Python забезпечує простоту і невимушеність мови сценаріїв і міць, яку зазвичай можна знайти в компільованих мовами. Перевищуючи можливості інших мов сценаріїв, така комбінація робить Python зручним засобом розробки великомасштабних проектів. Для попереднього ознайомлення нижче наводиться список основних можливостей, які є в арсеналі Python.

Python автоматично розподіляє пам'ять під об'єкти і звільняє її, коли об'єкти стають непотрібними.

Для створення великих систем Python надає такі можливості, як модулі, класи і виключення. Вони дозволяють розбити систему на складові, застосовувати ООП для створення програмного коду багаторазового користування і обробляти події та помилки, що виникають.

Python надає найбільш типові структури даних, такі як списки, словники і рядки, у вигляді особливостей, властивих самій мові програмування.

Щоб запустити програму на мові Python, досить просто ввести її ім'я. Не потрібно виконувати проміжну компіляцію і зв'язування, як це робиться в мовах програмування, подібних C або C ++. Інтерпретатор Python негайно виконує програму, що дозволяє програмувати в інтерактивному режимі і отримувати результати відразу ж після внесення змін - у більшості випадків ви зможете спостерігати ефект зміни програми з тією швидкістю, з якою ви вводите зміни з клавіатури.

Отже, мова Python має такі переваги:

– вільне програмне забезпечення;

- наявність великої кількості бібліотек;
- можливість роботи з XML/HTML файлами;
- програмування математичних та наукових розрахунків;
- можливість роботи з відео-, аудіо-та графічними файлами;
- динамічна типізація;
- підтримки об'єктно-орієнтованого програмування;
- інтеграція з C/C++ якщо можливості Python недостатньо;
- зрозумілий і лаконічний синтаксис;
- величезна кількість модулів;
- крос-платформенність.

Numpy – розширення мови Python, що додає підтримку великих багатовимірних масивів і матриць, разом з великою бібліотекою високорівневих математичних функцій для операцій з цими масивами. Попередник Numpy, Numeric, був спочатку створений Jim Hugunin. Numpy — відкрите програмне забезпечення і має багато розробників. NumPy можна розглядати як гарну вільну альтернативу MATLAB, оскільки мова програмування MATLAB зовні нагадує NumPy: обидві вони інтерпретовані, і обидві дозволяють користувачам писати швидкі програми поки більшість операцій проводяться над масивами або матрицями, а не над скалярами.

Бібліотеки Python – це файли з шаблонами коду. Вони придумані, щоб люди не повинні були повторно вводити той же код кожного разу: вони просто мають відкрити файл, вставити свої дані і отримати бажаний результат. У цій статті ми даємо Опис бібліотек, які найчастіше використовуються:

Pandas. Бібліотека надає структури даних та засоби аналізу. Добре обробляє неповні, неврегульовані та непозначені дані [20].

Pandas дозволяє замінювати досить складні операції з даними одною або двома командами. Містить багато готових до виконання методів групування, фільтрування, об'єднання даних і можливості розпізнавання різних типів джерел. До можливостей бібліотеки відносять:

- DataFrame — тип данних з інтегрованою індексацією, створений для маніпуляцій над данними;
- інструменти для зчитування та запису данних таких розширень, як: CSV, Excel, JSON, SAS, FWF;
- обробка відсутніх даних;
- підтримка операцій типу Group by;
- вставка та видалення строк/стовпців;
- злиття та приєднання датасетів;

– ієрархічна індексація осей для роботи з високорозмірними даними в низькорозмірних структурах даних;

– функціональність для роботи з часовими рядами.

Matplotlib – бібліотека на мові програмування Python для візуалізації даних двовимірною 2D графікою (3D графіка також підтримується). Пакет підтримує багато видів графіків і діаграм:

– графіки (line plot);

– діаграми розсіювання (scatter plot);

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

– секторні діаграми (pie chart);

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

– контурні графіки (contour plot);

– поля градієнтів (quiver);

– спектральні діаграми (spectrogram).

Нескладні тривимірні графіки можна будувати з допомогою набору інструментів (toolkit) mplot3d.

3.2 Тип вхідних даних CSV

CSV – текстовий файл, що містить інформацію. Кожен рядок є окремим рядком у таблиці, а стовпці розділяються спеціальними символами – роздільниками (наприклад, кома). Останнім часом роздільником може бути не тільки кома, але й іншими символами (пробіл, крапка з комою і т. д.).

РОзробнику не потрібен спеціальний інструмент для роботи з файлом CSV. Щоб відкрити файл достатньо навіть простого блокноту або іншої програми, яка може читати тексти. А для автоматичного перетворення даних у файл Excel або аналогічні програми будуть придатні. Наприклад, безкоштовний пакет Open Office.

3.3 Програмна реалізація

Для реалізації алгоритмів кластеризації було застосовано мову програмування Python з цілою низкою доповнень та бібліотек, зокрема: Pandas, Numpy, Matplotlib. Перші дві бібліотеки застосовувалися переважно для зручного представлення вхідних даних. Остання з бібліотек стала незамінною для візуалізації результатів, а саме, при побудові діаграм та графіків. Для реалізації був обраний метод кластеризації FCM.

Візьмемо для аналізу 1080 записів з набору і перші 3 характеристики. На рисунку 3.1 зображено часткову вибірку таких даних.

syst	diast	med
90	40	60
80	65	45
80	40	60
90	35	45
70	40	55
90	50	45
90	60	60
70	45	45
80	35	40
70	45	55
90	50	65
80	35	45
90	40	60
90	45	45

Рисунок 3.1 – Приклад вибіркових медичних даних

Розглянемо приклад, коли частина даних була загублена. Спробуємо відновити ці данні та подивитися, наскільки відрізняться кластеризація.

Для відновлення даних виконаємо декілька кроків:

1) Необхідно зібрати інформацію зі всіх листів з позиції втрачених даних;

2) Сформуємо масив та видалимо максимальні та мінімальні значення в ньому;

3) Порахуємо середне арифметичне значення отриманого масиву даних.

4) Проведемо кластеризацію даних обраним методом, який був описаний у другому розділі.

В ході експерименту розглянемо такі варіанти втрати даних у кількості 15, 30, 45, 60. Демонстрація результатів роботи методу на масиві даних, що складається з 1080 точок, та поділу його на 12 кластерів із втратою даних представлена на рисунках 3.2-3.5.

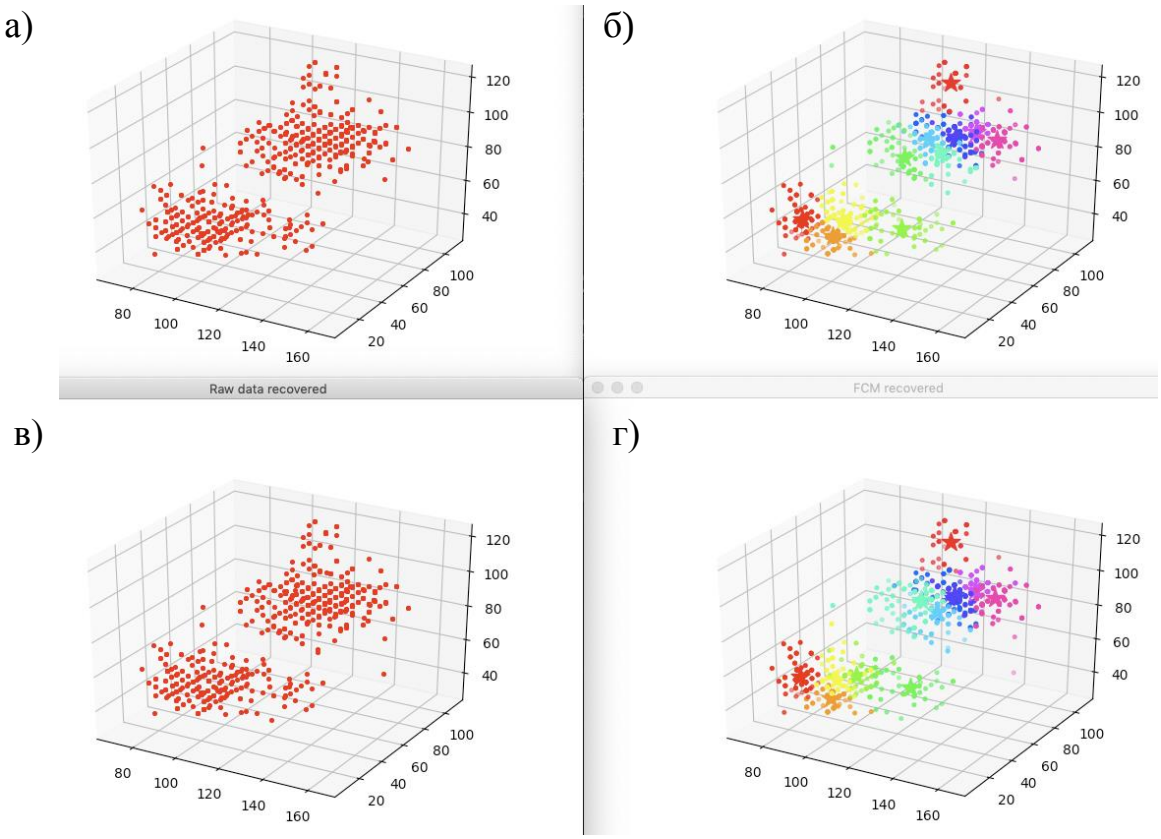


Рисунок 3.2 – Відновлення при втраті даних у кількості 15

(а) – вхідні данні; б) – кластерний аналіз вхідних даних; в) – вхідні дані з відновленням; г) – кластерний аналіз з відновленням даних)

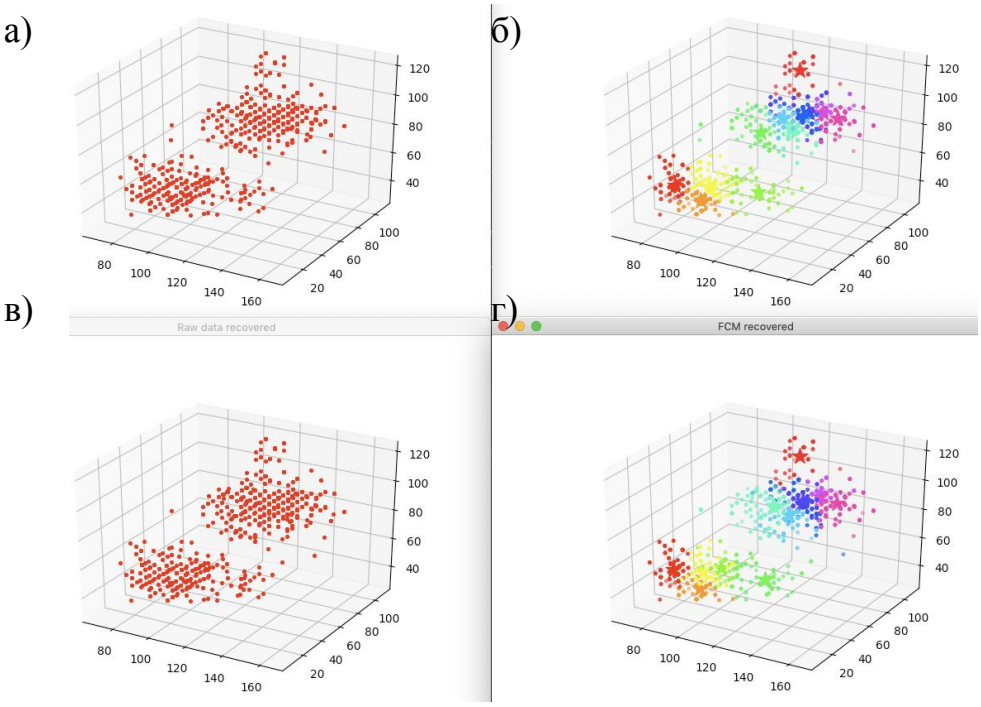


Рисунок 3.3 – Відновлення при втраті даних у кількості 30

(а) – вхідні данні; б) – кластерний аналіз вхідних даних; в) – вхідні дані з відновленням; г) – кластерний аналіз з відновленням даних)

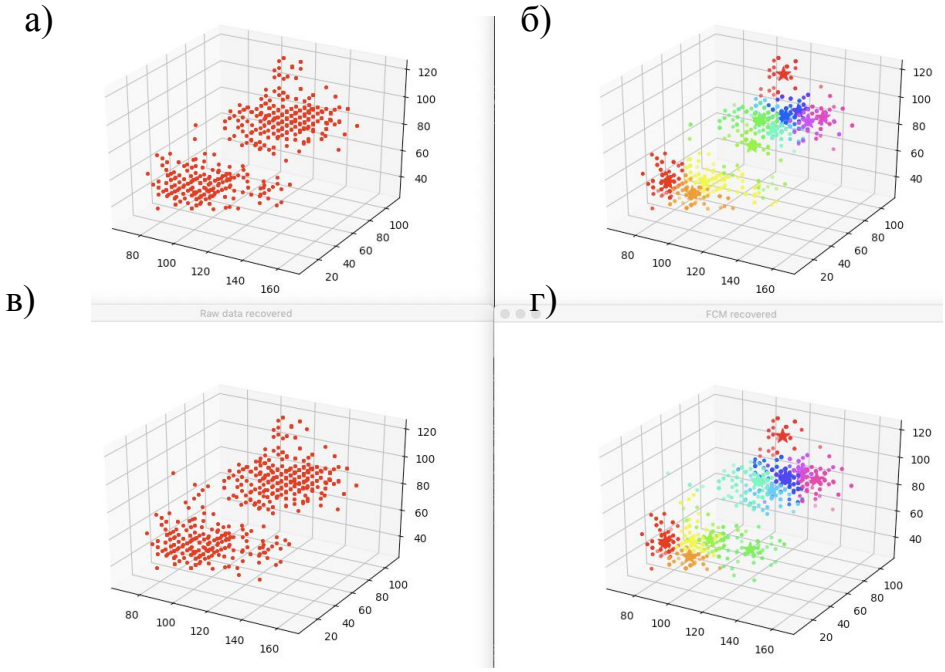


Рисунок 3.4 – Відновлення при втраті даних у кількості 45

(а) – вхідні данні; б) – кластерний аналіз вхідних даних; в) – вхідні дані з відновленням; г) – кластерний аналіз з відновленням даних)

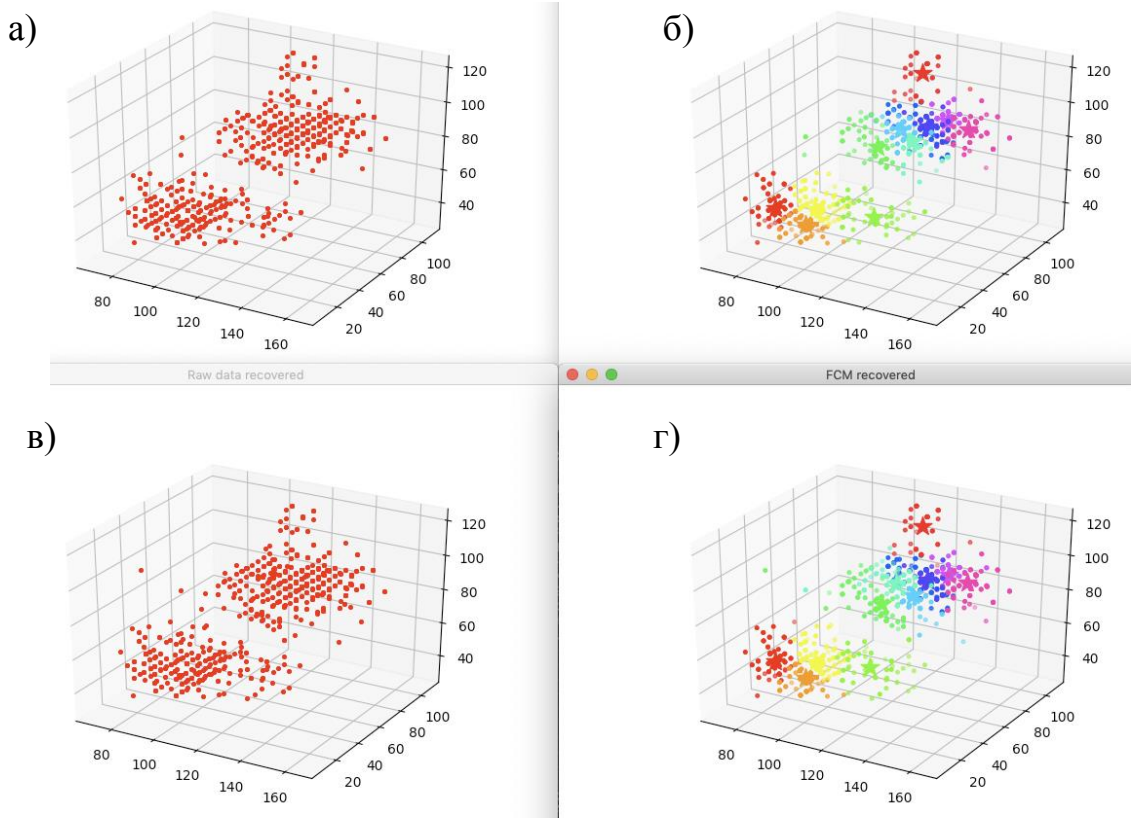


Рисунок 3.5 – Відновлення при втраті даних у кількості 60

(а) – вхідні данні; б) – кластерний аналіз вхідних даних; в) – вхідні дані з відновленням; г) – кластерний аналіз з відновленням даних)

3.4 Оцінка якості кластеризації

Вибір оптимального рішення будемо засновувати на понятті якості кластеризації. Якістю кластеризації назвемо ступінь наближення результату кластеризації до ідеального рішення.

Оскільки ідеальне рішення задачі кластеризації невідомо, то оцінити якість можна двома способами – експертним і формальним.

Експертна вибір найкращого рішення задачі полягає в оцінці рішення фахівцями в даній предметній області. Але експертна оцінка часто об'єктивно неможлива через велику обсягу та складності даних. Тому важливу роль відіграють формальні критерії оцінки якості кластеризації. Формальні критерії оцінюють якість кластеризації за деяким показником, обчисленому на підставі результатів кластеризації.

Оцінка кластеризації, яка вимірює відповідність результатів кластеризації, розглядається як одна з важливих проблем, важливих для успіху додатків кластеризації. Компактність і поділ є двома важливими параметрами для оцінки результатів кластеризації.

Компактність: елементи кластера повинні бути якомога ближче один до одного. Це властивість можна виразити через відстані між елементами кластера, щільністю всередині кластера або ж обсягом, займаним кластером в багатовимірному просторі.

Роздільність: відстань між різними кластерами повинно бути якомога більше. Відстань між кластерами зазвичай вимірюється одним з трьох наступних способів: 1) як відстань між найближчими елементами кластерів, 2) як відстань між найбільш віддаленими один від одного елементами кластерів і 3) як відстань між центрами кластерів.

В якості індексів для тестування були обрані наступні: індекс силуета, індекс Девіда-Болдуіна, Calinski-Harabasz індекс.

Для розрахунку значень, що необхідні для оцінки якості кластеризації було застосовано мову програмування Python. Код для реалізації представлено на рисунку 3.6.

```
score = sk.silhouette_score(data, labels, 'euclidean')
    print(f"Silhouette score: {score}")
    score = sk.calinski_harabaz_score(data, labels)
    print(f"Calinski Harabaz score: {score}")
    score = sk.davies_bouldin_score(data, labels)
    print(f"Davies Bouldin score: {score}")
```

Рисунок 3.6 – Розрахунок значень для оцінки якості кластеризації

Результати розрахунку представлені у таблиці 3.1.

Таблиця 3.1 – Результати розрахунку значень для оцінки якості кластеризації

Індекси / кількість втрат данних	0	15	30	45	60
індекс силуета	0.2306	0.2354	0.2229	0.2185	0.2334
Calinski-Harabasz індекс	926.0828	986.3909	862.5809	870.9748	789.6379
індекс Девіда-Болдуіна	1.2131	1.2345	1.2329	1.2378	1.2073

Отже, індекс силуета показує, наскільки середня відстань до об'єктів свого кластера відрізняється від середньої відстані до об'єктів інших кластерів. Дана величина лежить в діапазоні [-1,1]. Значення, близькі до -1, відповідають поганим (розрізненим) кластеризації, значення, близькі до нуля, кажуть про те, що кластери перетинаються і накладаються один на одного, значення, близькі до 1, відповідають «щільним» чітко виділеним кластерам. Таким чином, чим більше силует, тим чіткіше виділені кластери, і вони є компактними, щільно згруповані хмари точок. Як можна побачити з індексу силуету метод відновлення данних працює досі добре.

Чим вище значення у індексі Calinski-Harabasz тим кращим є рішення.

У індексі Девіда-Болдуіна значення, близькі до нуля, вказують на кращий розділ. Як можна побачити майже при всіх втратах данних розподіл гарний, олтже метод добре відпрацював.

4 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

В даному розділі проведено аналіз потенційних небезпечних та шкідливих виробничих факторів, причин пожеж. Розглянуті заходи, які дозволяють забезпечити гігієну праці і виробничу санітарію. На підставі аналізу розроблені заходи з техніки безпеки та рекомендації з пожежної профілактики.

Завданням даної магістерської роботи був: дослідження роботи та моделювання бездротових сенсорних мереж у контексті IoT.

4.1 Аналіз стану умов праці

4.1.1 Вимоги до приміщень

Геометричні розміри приміщення зазначені в табл. 4.1.

Таблиця 4.1 - Розміри приміщення.

Найменування	Значення
Довжина, м	4,4
Ширина, м	2,8
Висота, м	2,5
Площа, м2	12,32
Об’єм, м3	30,8

Згідно з ДСН 3.3.6.042-99 [21] розмір площі для одного робочого місця оператора персонального комп’ютера має бути не менше 6 кв. м, а об’єм — не менше 20 куб. м. Отже, дане приміщення цілком відповідає зазначеним нормам.

Для забезпечення потрібного рівного освітленості кімната має вікно та систему загального рівномірного освітлення, що встановлена на стелі. Для дотримання вимог пожежної безпеки встановлено порошковий вогнегасник та систему автоматичної пожежної сигналізації.

4.1.2 Вимоги до організації місця праці

При порівнянні відповідності характеристик робочого місця нормативним основні вимоги до організації робочого місця за ДСанПіН 3.3.2.007-98 [22] (табл. 4.2) і відповідними фактичними значеннями для робочого місця, констатуємо повну відповідність.

Таблиця 4.2 - Характеристики робочого місця

Найменування параметра	Фактичне Значення	Нормативне Значення
Висота робочої поверхні, мм	700	680 ÷ 800
Висота простору для ніг, мм	650	не менше 600
Ширина простору для ніг, мм	540	не менше 500
Глибина простору для ніг, мм	660	не менше 650
Висота поверхні сидіння, мм	420	400 ÷ 500
Ширина сидіння, мм	410	не менше 400
Глибина сидіння, мм	420	не менше 400
Висота поверхні спинки, мм	500	не менше 300
Ширина опорної поверхні спинки, мм	400	не менше 380
Радіус кривини спинки в горизонтальній площині, мм	400	400
Відстань від очей до екрану дисплея, мм	750	700 ÷ 800

4.2 Навантаження та напруженість процесу праці

Під час виконання робіт використовують ПК та периферійні пристрої, що призводить до навантаження на окремі системи організму. Такі перекоси у напруженні різних систем організму, що трапляються під час роботи з ПК, зокрема, значна напруженість зорового аналізатора і довготривале малорухоме положення перед екраном, не тільки не зменшують загального напруження, а навпаки, призводять до його посилення і появи стресових реакцій.

Найбільшому ризику виникнення різноманітних порушень піддаються: органи зору, м’язово-скелетна система, нервово-психічна діяльність, репродуктивна функція у жінок.

Тобто наявне психофізіологічні небезпечні та шкідливі фактори:

а) фізичного перевантаження:

– статичного;

– динамічного;

б) нервово-психічного перевантаження:

– розумового перенапруження;

– монотонності праці;

– перенапруження аналізаторів;

– емоційних перевантажень.

Роботу за дипломним проектом визнано, таку, що займає 50% часу робочого дня та за восьмигодинної робочої зміни рекомендовано встановити додаткові регламентовані перерви: для розробників програм тривалістю 15 хв через кожну годину роботи.

4.3 Виробнича санітарія

На підставі аналізу небезпечних та шкідливих факторів при виробництві (експлуатації), пожежної безпеки можуть бути надалі вирішені питання необхідності забезпечення працюючих достатньою кількістю освітлення, вентиляції повітря, організації заземлення, тощо.

4.3.1 Аналіз небезпечних та шкідливих факторів при виробництві (експлуатації) виробу

Аналіз небезпечних та шкідливих виробничих факторів виконується у табличній формі (табл. 4.3). Роботу, пов'язану з ЕОП з ВДТ, у тому числі на тих, які мають робочі місця, обладнані ЕОМ з ВДТ і ПП, виконують із забезпеченням виконання НПАОП 0.00-7.15-18 [25], яке встановлюють вимоги безпеки до обладнання робочих місць, до роботи із застосуванням ЕОМ з ВДТ і ПП. Переважно роботи за проектами виконують у кабінетах чи інших приміщеннях, де використовують різноманітне електрообладнання, зокрема персональні комп’ютери (ПК) та периферійні пристрої. Основними робочими характеристиками персонального комп'ютера є:

– робоча напруга U=+220В +-5%;

– робочий струм I=2А;

– споживана потужність P=350 Вт.

Таблиця 4.3 – Аналіз небезпечних і шкідливих виробничих факторів

Небезпечні і шкідливі виробничі фактори	Джерела факторів (види робіт)	Кількісна Оцінка	Нормативні Документи
Фізичні:			
підвищена або знижена вологість повітря	експлуатація ЕОМ	3	[21]
підвищений рівень напруги електричної мережі, замикання якої може відбутися через тіло людини	-//-	3	[23] [24]
Психофізіологічні:			
нервово-психічна перевантаження (розумове, перенапруження аналізаторів-зорових)	- формулювання теми; - пошук інформацію про предметну область; - пошук інформації про наявні аналоги; - проектування структур та алгоритмів; - виконання роботи; - оформлення записки.	4	[24] [25]
фізичні (статичне - сидіння)	порушення умов організації робочого часу (безперервна робота)	2	[22] [25]

4.3.2 Пожежна безпека

Висока щільність елементів в електронних схемах призводить до значного підвищення температури окремих вузлів (80…100℃). При проходженні електричного струму по провідниках і деталей виділяється тепло, що в умовах їх високої щільності може привести до перегріву, і може служити причиною запалювання ізоляційних матеріалів. Слабкий опір ізоляційних матеріалів дії температури може викликати порушення ізоляції і привести до короткого замикання між струмоведучими частинами обладнання (шини, електроди).

Заземлені конструкції, що знаходяться в приміщеннях, де розміщені робочі місця (батареї опалення, водопровідні труби, кабелі із заземленим відкритим екраном), надійно захищені діелектричними щитками та/або сітками з метою недопущення потрапляння працівника під напругу.

В приміщенні наявна затверджена «План-схема евакуації з кабінету (приміщення)».

Горючими матеріалами в приміщенні, де розташовані ЕОМ, є:

1) поліамід - матеріал корпусу мікросхем, горюча речовина, температура самозаймання 420℃;

2) полівінілхлорид - ізоляційний матеріал, горюча речовина, температура запалювання 335℃, температура самозаймання 530℃;

3) склотекстоліт ДЦ - матеріал друкарських плат, важкогорючий матеріал, показник горючості 1.74, не схильний до температурного самозаймання;

4) пластикат кабельний №489 - матеріал ізоляції кабелів, горючий матеріал, показник горючості більше 2.1;

5) деревина - будівельний і обробний матеріал, з якого виготовлені меблі, горючий матеріал, показник горючості більше 2.1, температура запалювання 255℃, температура самозаймання 399℃.

Простори усередині приміщень в межах, яких можуть утворюватися або знаходиться пожежонебезпечні речовини і матеріали відповідно до ДСТУ Б В.1.1-36:2016 [26] відносяться до пожежонебезпечної зони класу П-IIа. Це обумовлено тим, що в приміщенні знаходяться тверді горючі та важкозаймисті речовини та матеріали. Приміщенню, у якому розташоване робоче місце, присвоюється II ступень вогнестійкості.

Причинами можливого загоряння і пожежі можуть бути:

1) несправність електроустановки;

2) конструктивні недоліки устаткування;

3) коротке замикання в електричних мережах;

4) запалювання горючих матеріалів, що знаходяться в безпосередній близькості від електроустановки.

Продуктами згорання, що виділяються на пожежі, є: окис вуглецю; сірчистий газ; окис азоту; синильна кислота; акромін; фосген; хлор і ін. При горінні пластмас, окрім звичних продуктів згорання, виділяються різні продукти термічного розкладання: хлорангідридні кислоти, формальдегіди, хлористий водень, фосген, синильна кислота, аміак, фенол, ацетон, стирол.

4.3.3 Електробезпека

На робочому місці виконуються наступні вимоги електробезпеки: ПК, периферійні пристрої та устаткування для обслуговування, електропроводи і кабелі за виконанням та ступенем захисту відповідають класу зони за ПУЕ, мають апаратуру захисту від струму короткого замикання та інших аварійних режимів. Лінія електромережі для живлення ПК,

периферійних пристроїв і устаткування для обслуговування, виконана як окрема групова три- провідна мережа, шляхом прокладання фазового, нульового робочого та нульового захисного провідників. Нульовий захисний провідник використовується для заземлення електроприймачів. Штепсельні з'єднання та електророзетки крім контактів фазового та нульового робочого провідників мають спеціальні контакти для підключення нульового захисного провідника. Електромережа штепсельних розеток для живлення персональних ПК, укладено по підлозі поруч зі стінами відповідно до затвердженого плану розміщення обладнання та технічних характеристик обладнання. Металеві труби та гнучкі металеві рукави заземлені. Захисне заземлення включає в себе заземлюючих пристроїв і провідник, який з'єднує заземлюючий пристрій з обладнанням, яке заземлюється - заземлюючий провідник.

4.4 Гігієнічні вимоги до параметрів виробничого середовища

4.4.1 Мікроклімат

Мікроклімат робочих приміщень - це клімат внутрішнього середовища цих приміщень, що визначається діючої на організм людини з'єднанням температури, вологості, швидкості переміщення повітря. В даному приміщенні проводяться роботи, що виконуються сидячи і не потребують динамічного фізичного напруження, то для нього відповідає категорія робіт 1а. Отже оптимальні значення для температури, відносної вологості й рухливості повітря для зазначеного робочого місця відповідають ДСН 3.3.6.042-99 [21] і наведені в табл. А.4:

Таблиця 4.4 – Норми мікроклімату робочої зони об’єкту

Період Року	Категорія Робіт	Температура C^0	Відносна вологість %	Швидкість руху повітря, м/с
Холодна	Легка-1а	22-24	40-60	0,1
Тепла	Легка-1а	23-25	40-60	0,1

4.4.2 Освітлення

Для виробничих та адміністративних приміщень світловий коефіцієнт приймається не менше -1/8, в побутових - 1/10:

$$S_b = \left(\frac{1}{5} / \frac{1}{10}\right) * S_n \quad (4.1)$$

де Sb – площа віконних прорізів, м2;

Sn – площа підлоги, м2 .

Sn = a·b = 4,4·2,8 = 12,32 м2 ,

S = 1/10·12,32 = 1,232 м2 .

Приймаємо 1 вікно площею S=1,2 м2.

Світильники загального освітлення розташовуються над робочими поверхнями в рівномірно-прямокутному порядку. Для організації освітлення в темний час доби передбачається обладнати приміщення, довжина якого складає 4,4 м, ширина 2,8 м, світильниками ЛПО2П, оснащеними лампою типа ЛБ (одна - 80 Вт) з світловим потоком 5400 лм. Розрахунок штучного освітлення виробляється по коефіцієнтах використання світлового потоку, яким визначається потік, необхідний для створення заданої освітленості при загальному рівномірному освітленні. Розрахунок кількості світильників n виробляється по формулі (4.2):

$$n = \frac{E * S * Z * K}{F * U * M} \quad (4.2)$$

де E - нормована освітленість робочої поверхні, визначається нормами – 300 лк;

S - освітлювана площа, м2 ; S = 12,32 м2 ;

Z - поправочний коефіцієнт світильника (Z = 1,15 для ламп розжарювання та ДРЛ; Z = 1,1 для люмінесцентних ламп) приймаємо рівним 1,1;

K - коефіцієнт запасу, що враховує зниження освітленості в процесі експлуатації – 1,5;

U - коефіцієнт використання, залежний від типу світильника, показника індексу приміщення і т.п.

- 0,575 M - число люмінесцентних ламп в світильнику - 1;

F - світловий потік лампи - 5400лм (для ЛБ-80).

Підставивши числові значення у формулу (4.2), отримуємо:

$$n = \frac{300 * 12{,}32 * 1{,}15 * 1{,}5}{5400 * 0{,}575 * 1} \approx 2{,}0$$

Приймаємо освітлювальну установку, яка складається з 2-х світильників, оснащених лампами типа ЛБ (одна - 80 Вт) зі світловим потоком 5400 лм.

4.5 Вентилювання

У приміщенні, де знаходяться ЕОМ, повітрообмін реалізується за допомогою природної організованої вентиляції (вентиляційні шахти) і установки в віконному отворі автономного кондиціонера БК-2000. Цей метод забезпечує приток потрібної кількості свіжого повітря (30 м3 на годину на одного працюючого).

Також має здійснюватися провітрювання приміщення, в залежності від погодних умов, тривалість повинна бути не менше 10 хв. Найкращий обмін повітря здійснюється при наскрізному провітрюванні.

4.6 Заходи з організації виробничого середовища та попередження виникнення надзвичайних ситуацій

Відповідно до санітарно-гігієнічних нормативів та правил експлуатації обладнання наводимо приклади деяких заходів безпеки.

1) Заходи безпеки під час експлуатації персонального комп’ютера та периферійних пристроїв передбачають:

– правильне організування місця праці та дотримання оптимальних режимів праці та відпочинку під час роботи з ПК;

– експлуатацію сертифікованого обладнання;

– дотримання заходів електробезпеки;

– забезпечення оптимальних параметрів мікроклімату;

– забезпечення раціонального освітлення місця праці (освітленість робочого місця не перевищувала 2/3 нормальної освітленості приміщення);

– облаштовуючи приміщення для роботи з ПК, потрібно передбачити припливно-витяжну вентиляцію або кондиціювання повітря:

а) якщо об’єм приміщення 20 м3, то потрібно подати не менш як 30 м3/год повітря;

б) якщо об’єм приміщення у межах від 20 до 40 м3, то потрібно подати не менш як 20 м3/год повітря;

в) якщо об’єм приміщення становить понад 40 м3, допускається природна вентиляція, у випадку, коли немає виділення шкідливих речовин.

– зниження рівня шуму та вібрації:

а) у джерелі виникнення, шляхом застосування раціональних конструкцій, нових матеріалів і технологічних процесів;

б) звукоізолювання устаткування за допомогою глушників, резонаторів, кожухів, захисних конструкцій, оздоблення стін, стелі, підлоги тощо;

в) використання засобів індивідуального захисту).

Розрахунок захисного заземлення (забезпечення електробезпеки будівлі).

Загальний опір захисного заземлення визначається за формулою:

$$R_{ззп} = \frac{R_з \cdot R_п}{R_п \cdot n \cdot \eta_з + R_з \cdot \eta_п}, \quad (4.3)$$

де Rз - опір заземлення, якими когут бать труби, опори, кути і т.п., Ом;

Rш - опір опори, яке з'єднує заземлювачі, Ом;

п - кількість заземлювачів;

ηз - коефіцієнт екранування заземлювача; приймається в межах 0,2 ÷ 0,9; ηз = 0,7

ηш - коефіцієнт екранування сполучної стійки; приймається в межах 0,1 ÷ 0,7; ηш = 0,5;

Опір заземлення визначається за формулою:

$$R_з = \frac{\rho}{2\pi \cdot l} \cdot \left(\ln \frac{2 \cdot l}{d} + \frac{1}{2} \ln \frac{4 \cdot t + l}{4 \cdot t - l} \right), \quad (4.4)$$

де ρ - питомий опір ґрунту, залежить від типу грунту, Ом·м;

для піску - 400 ÷ 700 Ом·м; приймаємо ρ = 400 Ом·м;

l - довжина заземлювача, м; для труб - 2-3 м; l = 3 м;

d - діаметр заземлювача, м; для труб - 0,03-0,05 м; d = 0,05 м;

t - відстань від середини забитого в ґрунт заземлювача до рівня землі, м; t = 2 м.

$$R_з = \frac{400}{2 \cdot 3{,}14 \cdot 3} \left(\ln \frac{2 \cdot 3}{0{,}05} + \frac{1}{2} \ln \frac{4 \cdot 2 + 3}{4 \cdot 2 - 3} \right) = 110 \text{ , Ом}$$

Опір смуги, що з'єднує заземлювачі, визначається за формулою:

$$R_ш = \frac{\rho}{2\pi \cdot L} \cdot \ln \frac{2 \cdot L^2}{b \cdot t^1}, \quad (4.5)$$

де L - довжина смуги, що з'єднує заземлювачі (м) і приблизно дорівнює периметру будівлі: Пбуд. = 42·2 + 38·2 = 160 м; L = 160 м;

b - ширина смуги, м; b = 0,03 м;

t1 - глибина заземлення від рівня землі, м; t1 = 0,5 м.

$$R_n = \frac{400}{2 \cdot 3{,}14 \cdot 160} \cdot \ln \frac{2 \cdot 160^2}{0{,}03 \cdot 0{,}5} = 5{,}99 \text{ , Ом}$$

Кількість заземлювачів захисного заземлення визначається за формулою:

$$n = \frac{2 \cdot R_з}{4 \cdot \eta_з} \text{ ,} \qquad (4.6)$$

де 4 - допустимий загальний опір, Ом;

2 - коефіцієнт сезонності.

Визначаємо загальний опір захисного заземлення:

$$R_{ззп} = \frac{110 \cdot 5{,}99}{5{,}99 \cdot 79 \cdot 0{,}7 + 110 \cdot 0{,}5} = 1{,}7 \text{ Ом}$$

Висновок: дане захисне заземлення буде забезпечувати електробезпеку будівлі, так як виконується умова: Rззп < 4 Ом.

3) При виникненню пожеж при роботі на ПЕОМ від таких можливими джерел запалювання як:

– іскри і дуги коротких замикань;

– перегрів провідників, резисторів та інших радіодеталей ПЕОМ, від тривалої перевантаження та наявність перехідного опору;

– іскри при розмиканні і розмиканні ланцюгів;

– розряди статичної електрики;

– необережному поводженню з вогнем, а також вибухи газо-повітряних і паро-повітряних сумішей.

4.7 Охорона навколишнього природного середовища

Діяльність за темою магістерської роботи, а саме: оптимізація запитів до бази даних в процесі її виконання впливає на навколишнє природнє середовище і регламентується нормами діючого законодавства: Законом України «Про охорону навколишнього природного середовища» [28], Законом України «Про забезпечення санітарного та епідемічного благополуччя населення» [29], Законом України «Про відходи» [30].

В процесі роботи виникають процеси поводження з відходами ІТ галузі. Нижче надано перелік відходів, що утворюються в процесі роботи:

- Відпрацьовані люмінесцентні лампи - I клас небезпеки
- Змінні носії інформації - IV клас небезпеки
- Відходи друкуючих пристроїв - IV клас небезпеки
- Макулатура - IV клас небезпеки
- Побутові відходи - IV клас небезпеки

ВИСНОВКИ

В атестаційній роботі були розглянуті можливості кластерного аналізу слабоструктурованих даних на прикладі медичних даних функціонального стану серцево-судинної системи.

Для досягнення мети дослідження було проведено аналіз проблеми та огляд літературних джерел за тематикою роботи.

На основі проведених досліджень стало можливим зробити обґрунтований вибір найбільш ефективних для поставленої задачі алгоритмів: неієрархічні алгоритми k-means та c-means. Даний вибір обґрунтовується ітераційною структурою алгоритмів, і обчислення для кожного об’єкта можна проводити незалежно від інших, що є основою масово паралельних обчислень на графічних процесора.

Запропоновано і описано методику, яка призначена для кластеризації слабоструктурованих медичних показників при втраті за допомогою Fuzzy C-Means-алгоритму. Така кластеризація використовується в деякому просторі показників, орієнтуючись на ступінь схожості одного об’єкта з іншим. Система дозволяє виділяти необхідну кількість кластерів (груп) даних в залежності від потреб користувача. В результаті роботи системи користувач отримує розподіл даних на кластери у вигляді малюнків.

Запропонований алгоритм реалізований на мові програмування Python. Засобами програмних модулів досліджувалися реальні медичні дані.

Для експерименту були розглянуті ситуації із втратою даних, адже для діагностики хвороб серця важливо бачити повну картину даних. В результаті були отримані кластеризовані дані.

Для оцінки якості кластеризації отримані дані були порінянні із еталоном за допомогою індекса силуета, індекса Девіда-Болдуіна, Calinski-Harabasz індекса, що показали, що алгоритм добре відпрацював.

Наукова новизна отриманих результатів полягає в розробці та реалізації методики за допомогою відновлення різної кількості даних.

Таким чином усі поставленні завдання даної роботи були виконанні і була досягнута поставлена мета.

Результати дослідження показали, що кластерний аналіз кардіологічних даних може бути використаний для визначення кардіологічного статусу здоров'я людей.

В результаті проведеної роботи було зроблено аналіз умов праці, шкідливих та небезпечних чинників, з якими стикається робітник. Було визначено параметри і певні характеристики приміщення для роботи над запропонованим проектом написаному в дипломній роботі, описано, які заходи потрібно зробити для того, щоб дане приміщення

відповідало необхідним нормам і було комфортним і безпечним для робітника. Приведені рекомендації щодо організації робочого місця, а також важливу інформацію щодо пожежної та електробезпеки. Були наведені розміри приміщення та значення температури, вологості й рухливості повітря, необхідна кількість і потужність ламп та інші параметри, значення яких впливає на умови праці робітника, а також – наведені інструкції з охорони праці, техніки безпеки при роботі на комп’ютері.

А також визначені основні екологічні аспекти впливу на навколишнє природнє середовище та зазначені заходи щодо поводження з ними.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1) Buneman, P. Semistructured data [Текст] / P. Buneman // Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems (May 11-15, 1997). – Tucson, Arizona, United States, 1997. – P. 117-121

2) Semi-Structured Data [Електронний ресурс] // Datamation. – Режим доступу: www/URL: https://www.datamation.com/big-data/semi-structured-data.html. – 22.11.2018. – Загол. з екрану.

3) Abualigah, L.M. Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering [Текст] / L.M. Abualigah, A.T. Khader, M.A. Al-Betar, O.A. Alomari // Expert Systems with Applications. – 2017. – 84. – P. 24-36.

4) Kanimozhi, K.V. A novel map-reduce based augmented clustering algorithm for big text datasets [Текст] / K.V. Kanimozhi, M. Venkatesan // Advances in Intelligent Systems and Computing. – 2018. – Vol. 542. – P. 427-436.

5) Jenhani, F. Social stream clustering to improve events extraction [Текст] / F. Jenhani, M.S. Gouider, L.B. Said // Smart Innovation, Systems and Technologies. – 2018. – Vol. 73. – P. 319-329.

6) Al-Fath, A.M.U. Implementation of MCL algorithm in clustering digital news with graph representation [Текст] / A.M.U. Al-Fath, W.K.R. Saleh, S. Sa'Adah // 4th International Conference on Information and Communication Technology (25-27 May, 2016). – Bandung; Indonesia.

7) Ailem, M. Sparse Poisson Latent Block Model for Document Clustering [Текст] / M. Ailem, F. Role, M. Nadif // IEEE Transactions on Knowledge and Data Engineering. – 2017, №29 (7). – P. 1563-1576.

8) Joo, J. Joint Image-Text News Topic Detection and Tracking by Multimodal Topic And-Or Graph [Текст] / W. Li, J. Joo, H. Qi, S.-C. Zhu // IEEE Transactions on Multimedia. – 2017. – Vol. 19, Issue 2, 19(2). – P. 367-381.

9) Bafna, P. Document clustering: TF-IDF approach [Текст] / P. Bafna, D. Pramod, A. Vaidya // International Conference on Electrical, Electronics, and Optimization Techniques, ICEEOT. – 2016. – P. 61-66.

10) Lamari, Y. Parallel document clustering using iterative mapreduce [Текст] / Y. Lamari, S.C. Slaoui // International Conference on Big Data and Advanced Wireless Technologies (10-11 November 2016). – Blagoevgrad; Bulgaria.

11) Patil, H. Document clustering: A summarized survey (Book Chapter) [Текст] / H. Patil, R.S. Thakur // Pattern and Data Analysis in Healthcare Settings. – 2016. – P. 264-281.

12) Le, T.M.V. Semantic visualization with neighborhood graph regularization [Текст] / T.M.V. Le, H.W. Lauw // Journal of Artificial Intelligence Research. – 2016. – Vol. 55. – P. 1091-1133.

13) Rahmawati, D. Document clustering using sequential pattern (SP): Maximal frequent sequences (MFS) as SP representation [Текст] / D. Rahmawati, G.A. Putri Saptawati, Y. Widyani, // Proceedings of 2015 International Conference on Data and Software Engineering, ICODSE 2015. – 2016. – P. 98-102.

14) Tan, P.-N. Cluster Analysis: Basic Concepts and Algorithms. Introduction to Data Mining [Електроний ресурс] / P.-N. Tan, M. Steinbach, K. Vipin. – Режим доступу: www/URL: https://www-users.cs.umn.edu/~kumar001/dmbook/ch8.pdf. – 22.11.2018. – Загол. з екрану.

15) MATLAB [Електронний ресурс] // MATLAB. – Режим доступу: www/URL: https://matlab.ru/products/matlab. – 20.12.2018. – Загол. з екрану.

16) Афонин, В.В. О структурировании лабораторно-практических занятий при изучении дисциплин программирования [Електроний ресурс] / В.В. Афонин, С.А. Федосин. – Режим доступу: www/URL: https://cyberleninka.ru/article/v/o-strukturirovanii-... – 22.11.2018. – Загол. з екрану.

17) Анализ данных с R [Електронний ресурс] // Анализ данных с R. – Режим доступу: www/URL: https://www.soc.univ.kiev.ua/sites/default/files/ course/materials/r1.pdf. – 21.12.2018. – Загол. з екрану.

18) Язык программирования Python 3 [Електронний ресурс] // Python 3 для начинающих. – Режим доступу: www/URL: https://pythonworld.ru/. – 23.12.2018. – Загол. з екрану.

19) Аналіз часових рядів [Електронний ресурс] // Аналіз часових рядів. – Режим доступу: www/URL: https://www.rusnauka.com/45_PWMN_2016/ Economics/14_220211.doc.htm. – 24.12.2018. – Загол. з екрану.

20) Полезные библиотеки Python для data science [Електронний ресурс] // Наука о данных (Data Science). – Режим доступу: www/URL: https://blog.skillfactory.ru/python/biblioteki-python-dlya-data-science/. – 26.12.2018. – Загол. з екрану.

21) Державні санітарні норми. ДСН 3.3.6.042-99. «Санітарні норми мікроклімату виробничих приміщень. Міністерство охорони здоров'я України (МОЗ)» - Режим доступу: https://zakon.rada.gov.ua/rada/show/va042282-99 - 01.02.1999 р.

22) Державні санітарні правила і норми. ДСанПІН 3.3.2.007-98. «Державні санітарні правила і норми роботи з візуальними дисплейними терміналами електронно-обчислювальних машин. Міністерство охорони здоров'я України (МОЗ)» - Режим доступу: https://zakon.rada.gov.ua/rada/show/v0007282-98 - 10.12.1998 р.

23) Державний стандарт України. ДСТУ Б В.2.5-82:2016 «Електробезпека в будівлях і спорудах. Вимоги до захисних заходів від ураження електричним струмом» - Режим доступу: http://epicentre.co.ua/dstu/doc28522.html - 01.07.2016 р.

24) Міждержавний стандарт. ГОСТ 13109-97. «Норми якості електричної енергії в системах електропостачання загального призначення» - Режим доступу: https://dnaop.com/html/42313/doc-ГОСТ_13109-97 - 21.11.1997 р.

25) НПАОП 0.00-7.15-18. «Вимоги безпеки і захисту здоров'я працівників при роботі з екранними пристроями» - Режим доступу: https://zakon.rada.gov.ua/laws/show/z0508-18 - 14.02.2018 р.

26) Державний стандарт України. ДСТУ Б В.1.1-36:2016. «Визначення категорій приміщень, будинків та зовнішніх установок за вибухопожежною та пожежною небезпекою» - Режим доступу: https://zakon.rada.gov.ua/rada/show/v0158858-16 - 15.06.2016 р.

27) Міждержавний стандарт. ГОСТ 12.1.044-89. «Система стандартів безпеки праці. Вогнестійкість. Номенклатура показників і методи їх визначення (ICO 4589-84)» - Режим доступу: http://helpnik.college.ks.ua/standart/gost/Catalog/Index/4/4085.htm - 12.12.1989 р.

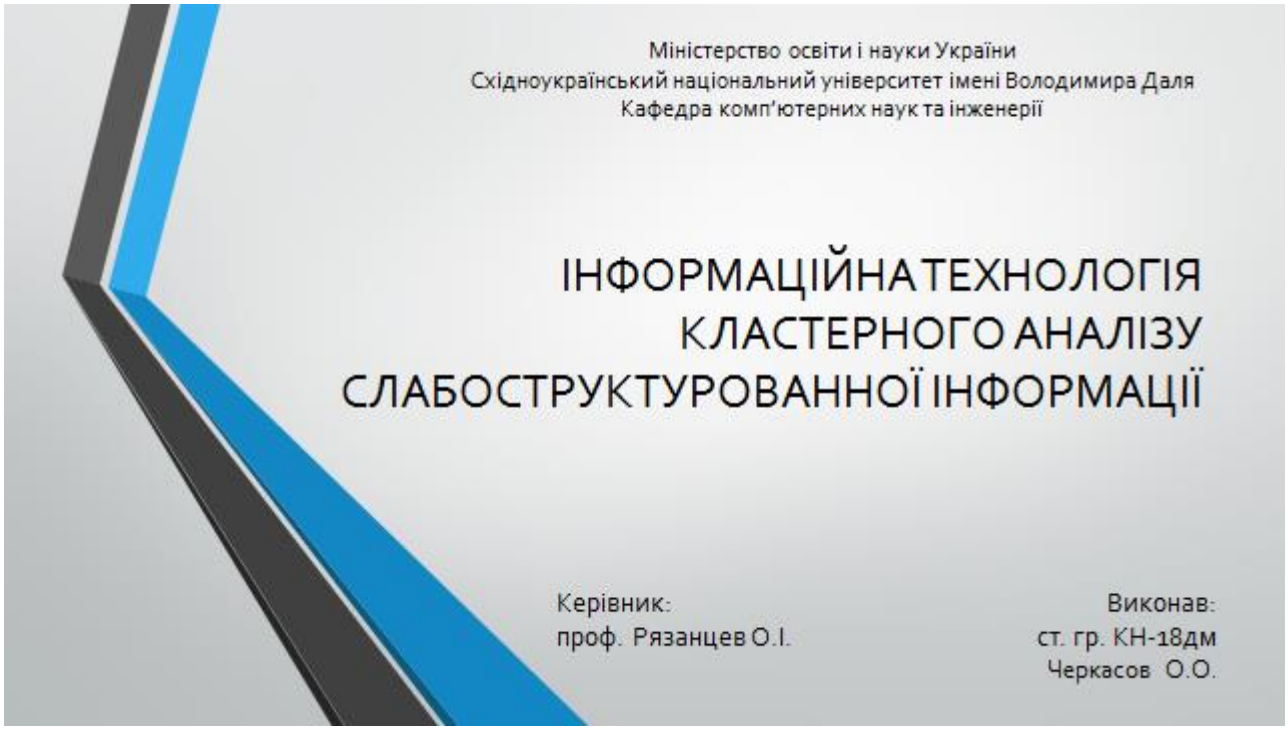
28) Закон України «Про охорону навколишнього природного середовища» - Режим доступу: https://zakon.rada.gov.ua/laws/show/1264-12 - 26.06.1991 р.

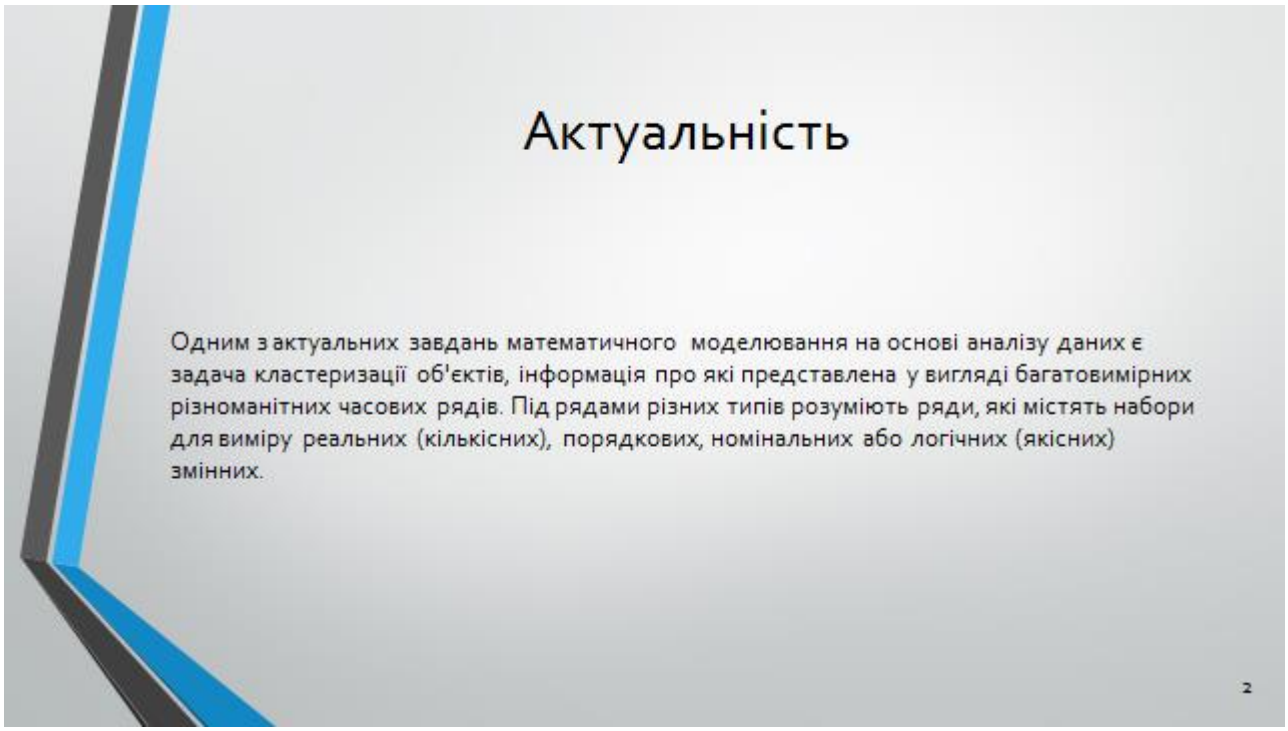
29) Закони України «Про охорону навколишнього природного середовища» - Режим доступу - https://zakon.rada.gov.ua/laws/show/4004-12 - 24.02.1994 р.

30) Закон України «Про відходи» - Режим доступу: https://zakon.rada.gov.ua/laws/show/187/98-%D0%B2%D1%80 – 05.03.1998 р.

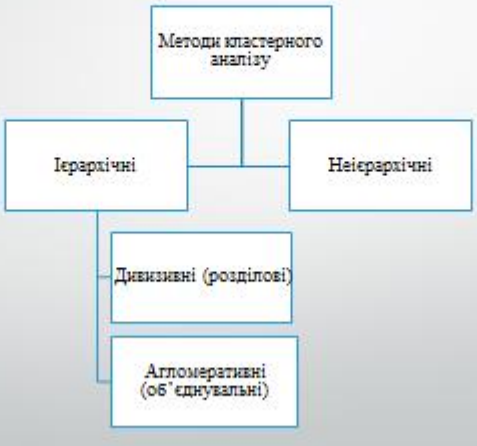
ДОДАТОК А.

Електронні плакати





Огляд існуючих методів кластерного аналізу



3

Постановка завдання дослідження

Об'єкт дослідження

- медичні данні функціонального стану серцево-судинної системи.

Предмет дослідження

- методи і алгоритми кластеризації, що дозволяють відновити загублені медичні дані.

Мета роботи

- розробка методики відновлення слабоструктурованих медичних показників.

Задачі дослідження

- провести аналіз проблеми за тематикою роботи;
- провести огляд літературних джерел за тематикою роботи;
- провести аналіз існуючих методів кластеризації слабоструктурованих даних;
- провести вибір математичної моделі кластерного аналізу;
- розробити методику відновлення слабоструктурованих медичних показників;
- реалізувати комп'ютерну модель для аналізу даних

4

Математична модель кластерного аналізу методом k-середніх та його реалізація

Завдання кластеризації методом k-середніх можна виразити наступним чином:

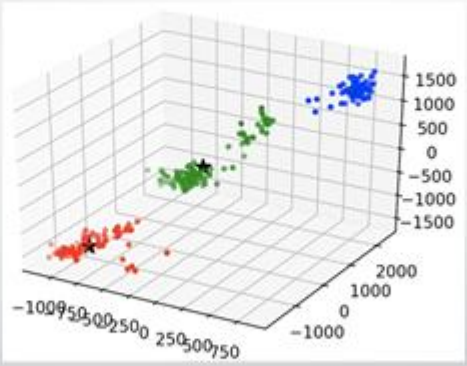
$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left\|x_i - s_j\right\|^2 \to \min$$

Кроки виконання алгоритму:

1. Алгоритм починається з відбору первинних центрів кластерів. Вони можуть бути обрані будь-яким способом, наприклад, випадково або на основі аналізу вхідних даних.
2. Потім кожна ітерація розбиває об'єкти на кластери в співвідношенні із тим, який об'єкт був ближче до центру на основі обраної метрики, а потім для кожного кластера переобчислюється центр.
3. Алгоритм закінчується тоді, коли при переобчисленні центрів кластерів, вони не змінюються.

5

Математична модель кластерного аналізу методом k-середніх та його реалізація



Результат роботи алгоритму k-середніх

6

Математична модель кластерного аналізу методом FCM та його реалізація

Кроки виконання алгоритму:

1. Ініціалізація

Вибираються наступні параметри:
- необхідна кількість кластерів N, 2 <N <K;
- міра відстаней, наприклад Евклідова відстань;
- фіксований параметр q (зазвичай 2);
- початкова (на нульовій ітерації) матриця приналежності $U^{(0)} = (\mu_{ik})^{(0)}$ об'єктів з урахуванням заданих початкових центрів кластерів.

2. Регулювання позицій центрів кластерів

На t-м ітераційному кроці при відомій матриці $U^{(t)}$ обчислюється

$$c_j = \frac{\sum_{i=1}^{N} (\mu_{ij})^q x_i}{\sum_{k=1}^{N} (\mu_{ik})^q}$$

3. Коригування значень приналежності $\mu_{ik}^{(t)}$

З огляду на відомі $c_j^{(t)}$, обчислюються $\mu_{ik}^{(t)}$

$$\mu_{ik} = \frac{1 / \|x_k - c_i\|^{2/(q-1)}}{\sum_{j=1}^{N} \left(1 / \|x_k - c_j\|^{2/(q-1)} \right)}$$

4. Зупинка методу

Метод нечіткої кластеризації зупиняється при виконанні наступної умови:

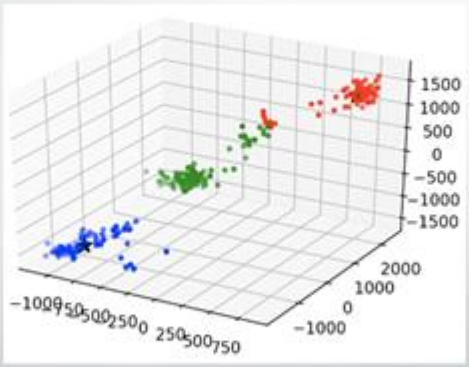
$$\|U^{(t+1)} - U^{(t)}\| \le \varepsilon,$$

де $\|\|$ - матрична норма (наприклад, Евклідова норма);

ε - заздалегідь задається рівень точності

7

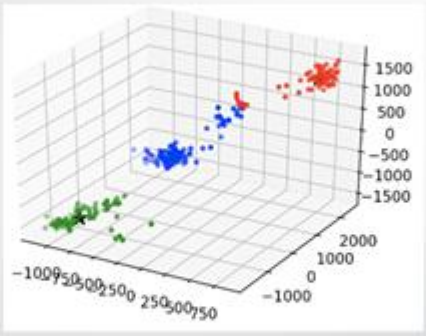
Математична модель кластерного аналізу методом FCM та його реалізація



Результат роботи алгоритму FCM

8

Перевірка результатів



Для перевірки правильності виконання методу використаємо готову реалізацію цього алгоритму з пакету skfuzzy

9

Обґрунтування вибору середовища програмної реалізації

Найпоширеніші мови програмування для реалізації алгоритмів кластерного аналізу

	Matlab	R	Python
інтеграція з C/C++	+	+	+
програмування математичних та наукових розрахунків	+	+	+
Синтаксис	традиційний	вузькоспеціалізований	традиційний
Наявність пакетів для аналізу даних	+	+	+
Розповсюдження	платне	безоплатне	безоплатне

10

Комп'ютерна модель кластерного аналізу при втраті даних

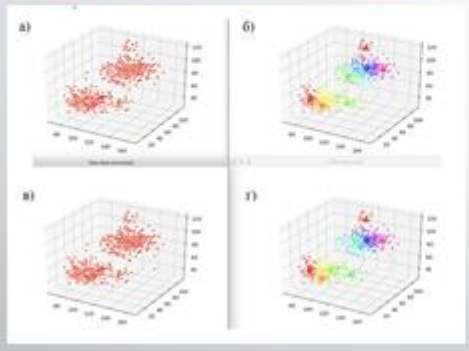
Для відновлення даних виконаємо декілька кроків:

1. Необхідно зібрати інформацію зі всіх листів з позиції втрачених даних;
2. Сформуємо масив та видалимо максимальні та мінімальні значення в ньому;
3. Порахуємо середнє арифметичне значення отриманого масиву даних;
4. Проведемо кластеризацію даних обраним методом.

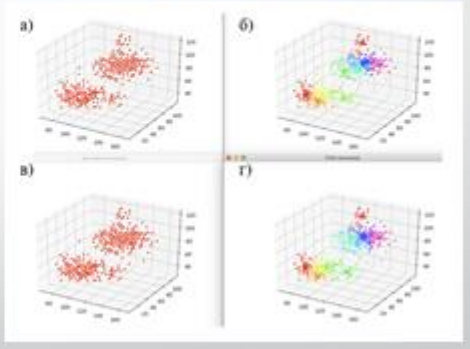
11

Демонстрація результатів роботи методу на масиві даних, що складається з 1080 точок, та поділу його на 12 кластерів із втратою даних

Відновлення при втраті даних у кількості 15



Відновлення при втраті даних у кількості 30

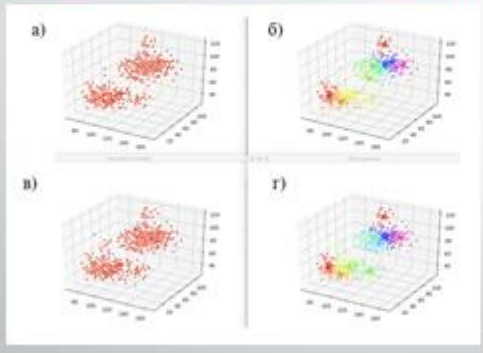


а) – вхідні данні;
б) – кластерний аналіз вхідних даних;
в) – вхідні дані з відновленням;
г) – кластерний аналіз з відновленням даних.

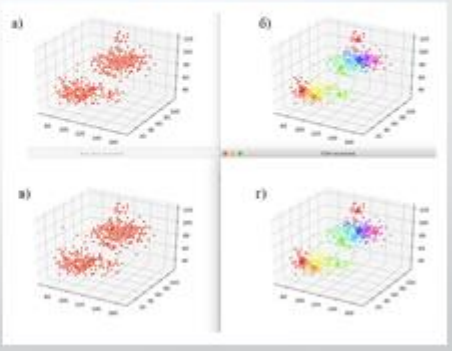
12

Демонстрація результатів роботи методу на масиві даних, що складається з 1080 точок, та поділу його на 12 кластерів із втратою даних

Відновлення при втраті даних у кількості 45



Відновлення при втраті даних у кількості 60



а) – вхідні данні;
б) – кластерний аналіз вхідних даних;
в) – вхідні дані з відновленням;
г) – кластерний аналіз з відновленням даних.

13

Оцінка якості кластеризації

Індекс оцінки силуета (Silhouette index)

- Оцінка для всієї кластерної структури досягається усередненням показника за елементами:

$$SWC = \frac{1}{N}\sum_{j=1}^{N} S_{ij}$$

де

$$S_{ij} = \frac{b_{ij} - a_{ij}}{\max(a_{ij}, b_{ij})}$$

- Найкраще розбиття характеризується максимальним SWC

Calinski-Harabasz індекс

- визначимо формулу індексу як

$$VRC = \frac{\frac{BGSS}{c-1}}{\frac{WGSS}{N-c}} = \frac{\bar{d}^2 + \frac{N-c}{c-1}A_c}{\bar{d}^2 - A_c} = \frac{1 + \frac{N-c}{c-1}a_c}{1 - a_c},$$

- Максимальне значення індексу VRC відповідає оптимальній структурі кластерів

Індекс Девіда-Болдуіна

- індекс обчислюється за формулою

$$DB = \frac{1}{c}\sum_{i=1}^{c} R_i$$

де

$$R_i = \max_{1 \le j \le c, j \ne i}(R_{ij})$$

- Оскільки мається на увазі, що кластери в структурі значно відрізняються один від одного, найкращою буде структура з мінімальним DB

14

Оцінка якості кластеризації

Результати розрахунку значень для оцінки якості кластеризації

Індекси / кількість втрат данних	0	15	30	45	60
індекс силуета	0.2306	0.2354	0.2229	0.2185	0.2334
Calinski-Harabasz індекс	926.0828	986.3909	862.5809	870.9748	789.6379
індекс Девіда-Болдуіна	1.2131	1.2345	1.2329	1.2378	1.2073

15

Висновки

- В атестаційній роботі були розглянуті можливості кластерного аналізу слабоструктурованих даних на прикладі медичних даних функціонального стану серцево-судинної системи.
- Запропоновано методику, яка призначена для кластеризації слабоструктурованих медичних показників при втраті за допомогою Fuzzy C-Means-алгоритму. Така кластеризація використовується в деякому просторі показників, орієнтуючись на ступінь схожості одного об'єкта з іншим. Система дозволяє виділяти необхідну кількість кластерів (груп) даних в залежності від потреб користувача. В результаті роботи системи користувач отримує розподіл даних на кластери у вигляді малюнків.
- Запропонований алгоритм реалізований на мові програмування Python. Засобами програмних модулів досліджувалися реальні медичні дані.
- Для експерименту були розглянуті ситуації із втратою даних, адже для діагностики хвороб серця важливо бачити повну картину даних. В результаті були отримані откластеризовані данні.
- Для оцінки якості кластеризації отримані дані були порінянні із еталоном за допомогою індекса силуета, індекса Девіда-Болдуіна, Calinski-Harabasz індекса, що показали, що алгоритм добре відпрацював.
- Результати дослідження показали, що кластерний аналіз кардіологічних даних може бути використаний для визначення кардіологічного статусу здоров'я людей.

16