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

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

**МАГІСТЕРСЬКА РОБОТА**

НА ТЕМУ:

# Методи та спеціалізовані комп'ютерні засоби для класифікації багатовимірних об'єктів

Освітній рівень"Магістр"
Спеціальність 123 "Комп'ютерна інженерія"

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

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

Студент: _____ (підпис) М.М. Баранов (ініціали, прізвище)

Група: КІ-17зм

Сєвєродонецьк 2019

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 123 "Комп'ютерна інженерія"
(шифр і назва)

**ЗАТВЕРДЖУЮ:**

Завідувач кафедри _____
_____ І.С. Скарга-Бандурова
«_____» _____20____р.

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

Баранову Михайлу Миколайовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Методи та спеціалізовані комп'ютерні засоби для класифікації багатовимірних об'єктів

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 «18» 10 2018 р. № 221/48
2. Строк подання студентом роботи 10.01.2018
3. Вихідні дані до роботи Матеріали науково-дослідної практики, алгоритми класу Forel, агломеративні методи кластерного аналізу, література для роботи з багатовимірними об'єктами

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити) Огляд задач кластеризації та постановка завдання, алгоритми класифікації, теорія розпливчастих множин, розробка та програмна реалізація методу класифікації багатовимірних об'єктів, 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 18.10.2018

Керівник _____
(підпис)

Завдання прийняв до виконання _____
(підпис)

**КАЛЕНДАРНИЙ ПЛАН**

| № з/п | Назва етапів дипломного проекту (роботи) | Строк виконання етапів проекту ( роботи ) | Примітка |
|---|---|---|---|
| 1 | Розробка технічного завдання | 10.09.2018-15.09.2018 | |
| 2 | Огляд методів класифікації | 16.09.2018-22.09.2018 | |
| 3 | Застосуванян теорії розпливчатих множин | 23.09.2018-25.09.2018 | |
| 4 | Розробкак та програмна реалізація методу клафікації | 26.09.2018-06.10.2018 | |
| 5 | Аналіз результатів дослідження | 07.10.2018-25.11.2018 | |
| 6 | Розробка частини проекту "Охорона праці та безпеки в надзвичайних ситуаціях" | 26.11.2018-1.12.2018 | |
| 7 | Оформлення пояснювальної записки, автореферату та презентації | 2.12.2018-09.01.2019 | |
| | | | |
| | | | |

**Студент** _____ Баранов М.М.
( підпис ) (прізвище та ініціали)

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

# АНОТАЦІЯ

Баранов М.М. Методи та спеціалізовані комп’ютерні засоби для класифікації багатовимірних об'єктів

Магістерська робота присвячена дослідженню методу класифікації багатовимірних об'єктів з застосуванням теорії розмитих множин. Був побудований алгоритм класифікації. Здійснена програмна реалізація даного алгоритму. Проведено аналіз результатів класифікації. У результаті аналізу можна зробити висновок про доцільність застосування розробленого алгоритму для класифікації багатовимірних об'єктів.

**Ключові слова:** класифікація, таксономія, класифікація, forel.

# ABSTRACT

Baranov M.M. Methods and specialized computer tools for the classification of multidimensional objects.

The master's work is devoted to the classification of multi-dimensional objects using fuzzy set theory. Was built by the classification algorithm. Implemented a software implementation of the algorithm. The analysis of the results of the classification. As a result of analysis, we can conclude the feasibility of the developed algorithms for classification of multidimensional objects.

**Keywords:** clustering, taxonomy, classification, forel.

# АННОТАЦИЯ

Магистреская работа посвящена исследованию классификации многомерных объектов с использованием теории размытых множеств. Был построен алгоритм классификации. Осуществлена программная реализация данного алгоритма. Проведен анализ результатов классификации. В результате анализа можно сделать вывод о целесообразности применения разработанного алгоритма для классификации многомерных объектов.

**Ключевые слова:** кластеризация, таксономия, классификация, forel.

# ЗМІСТ

# ВСТУП

Рішення фундаментальної завдачі кластеризації та таксономії в будь-якій області знань породило величезну кількість алгоритмів і підходів до її вирішення. У нашому випадку для автоматичної системи розпізнавання образів, яка орієнтована для роботи з непідготовленими, з точки зору математики, фахівцям необхідно надати допомогу у виборі алгоритму. Якість роботи застосовуваного алгоритму або методу, як відомо, залежить насамперед від вхідних даних і характеру їх розподілу. Вибір того чи іншого алгоритму в конкретній ситуації дуже непросте завдання. Необхідно проводити аналіз вхідних даних, і грунтуючись на результатах приймати рішення у виборі методу розпізнавання.

Як показує досвід аналізу масових джерел, число об'єктів може досягати багатьох десятків і сотень, кількість функцій і може становити до декількох десятків. Очевидно, що аналіз матриці даних з великою кількістю об'єктів і атрибутів практично неефективний - можна тільки визначити окремі характеристики досліджуваної структури, витягти ілюстративні конкретні приклади. Це піднімає проблему інтеграції, концентрації вихідних даних, тобто побудови узагальнених характеристик множини ознак і множини об'єктів. Ці проблеми можуть бути вирішені за допомогою методів багатовимірного аналізу. Методи, спрямовані на аналіз структури множини ознак і виявлення узагальнених чинників.

На даний момент реально існуючі програми класифікації працюють з незначною кількістю параметрів, що обумовлено суттєвим зростанням складності при їх зростанні. Одним із прикладів зростання складності задачі класифікації може служити сімейство алгоритмів FOREL, де велика кількість операцій пов'язано з перерахунком значень параметрів однієї координатної системи в іншу, або алгоритми кластерного аналізу, які теж вимагають аналогічних операцій перерахунку пов'язаних зі зменшенням числа параметрів з подальшим їх елементом розгортання.

# 1 ОГЛЯД ЗАДАЧ КЛАСТЕРИЗАЦІЇ ТА ПОСТАНОВКА ЗАВДАННЯ

## 1.1 Вступ до задач кластеризації

Задача кластеризації схожа з задачами класифікації, є її логічним продовженням, але її відмінність в тому, що класи досліджуваного набору даних заздалегідь не визначені. Синонімами терміну кластеризація є автоматична класифікація, навчання без вчителя і таксономія.

Кластеризація призначена для розбиття сукупності об'єктів на однорідні групи (кластери або класи). Якщо дані вибірки представити як точки в просторі ознак, то завдання кластеризації зводиться до визначення згущувань точок. Мета кластеризації - пошук існуючих структур.

Кластеризація є описовою процедурою, вона не робить ніяких статистичних висновків, але дає можливість провести розвідувальний аналіз і вивчити структуру даних. Саме поняття кластер визначено неоднозначно: в кожному дослідженні свої кластери. Перекладається поняття кластер як скупчення, гроно. Кластер можна охарактеризувати як групу об'єктів, що мають загальні властивості.

Характеристиками кластера можна назвати дві ознаки:

– внутрішня однорідність;

– зовнішня ізольованість.

Питання, що задається аналітиками при вирішенні багатьох завдань, полягає в тому, як організувати дані в наочні структури, тобто розгорнути таксономії.

Найбільше застосування кластеризація спочатку отримала в таких науках як біологія, антропологія, психологія. Для вирішення економічних завдань кластеризація тривалий час мало використовувалася через специфіку економічних даних і явищ.

У таблиці 1.1 наведено порівняння деяких параметрів задач класифікації та кластеризації.

Таблиця 1.1 - Порівняння класифікації та кластеризації

| Характеристика | Класифікація | Кластеризація |
|---|---|---|
| 1 | 2 | 3 |
| Контрольованість навчання | Контрольоване навчання | Неконтрольоване навчання |
| Стратегія | Навчання з вчителем | Навчання без вчителя |
| Наявність мітки класу | Навчальна множина супроводжується міткою, що вказує клас, до якого відноситься спостереження | Мітки класу навчальної множини невідомі |

Продовження табл.1.1

| 1 | 2 | 3 |
|---|---|---|
| Підстава для класифікації | Нові дані класифікуються на підставі навчальної множини | Дано безліч даних з метою встановлення існування класів або кластерів даних |

На рисунку 1.1 схематично представлені завдання класифікації і кластеризації.

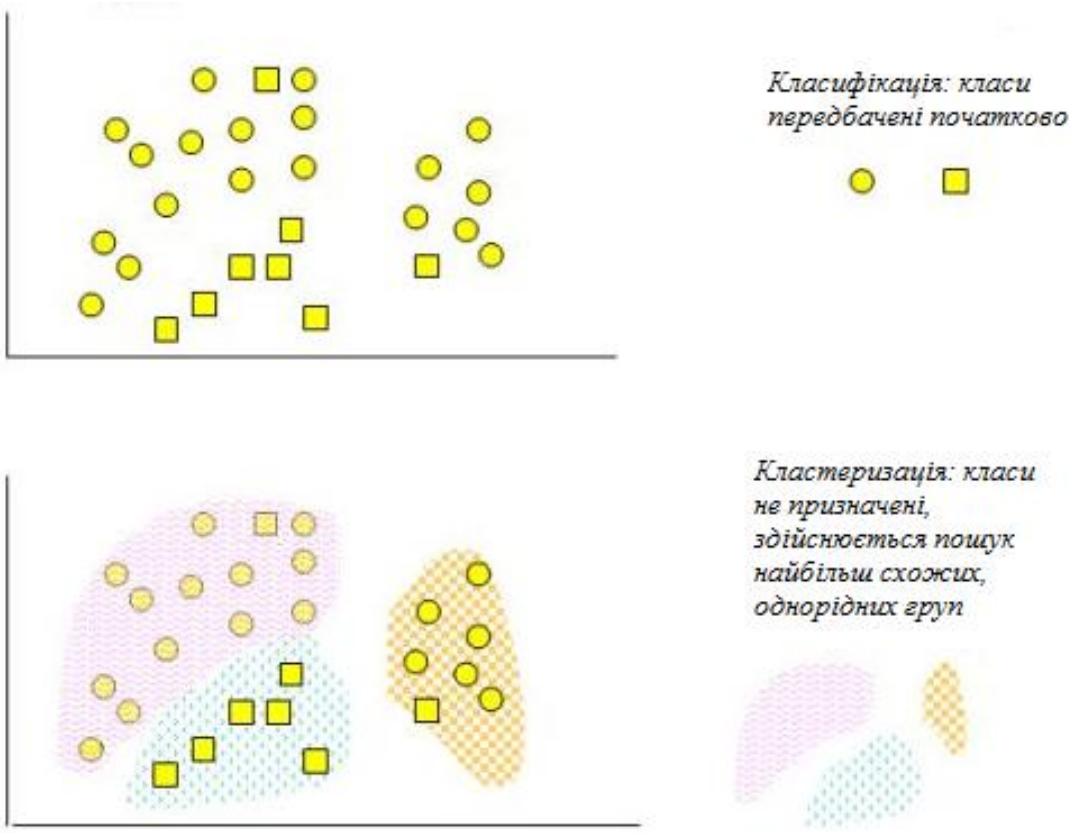



Рисунок 1.1- Порівняння задач класифікації та кластеризації

Кластери можуть бути непересічними, або ексклюзивними. Схематичне зображення непересічних і пересічних кластерів дано на рисунку 1.2.

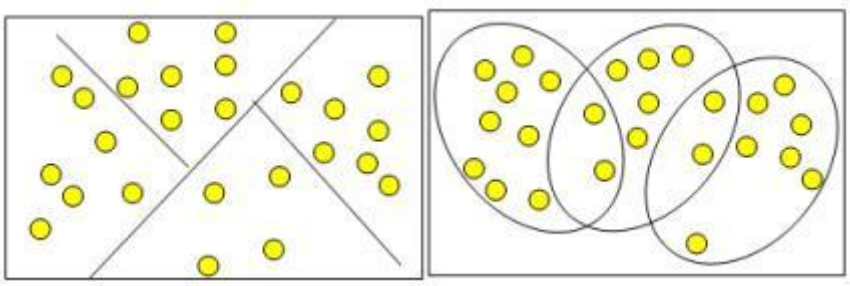


Рисунок 1.2 - Непересічні і пересічні кластери

Слід зазначити, що в результаті застосування різних методів кластерного аналізу можуть бути отримані кластери різної форми. Наприклад, можливі кластери цепочного типу, коли кластери представлені довгими ланцюжками, кластери подовженої форми, а деякі методи можуть створювати кластери довільної форми.

Різні методи можуть прагнути створювати кластери певних розмірів (наприклад, малих або великих) або припускати в наборі даних наявність кластерів різного розміру. Деякі методи кластерного аналізу особливо чутливі до шумів або викидам, інші - менш.

В результаті застосування різних методів кластеризації можуть бути отримані неоднакові результати, це нормально і є особливістю роботи того чи іншого алгоритму. Дані особливості слід враховувати при виборі методу кластеризації.

На сьогоднішній день розроблено більше сотні різних алгоритмів кластеризації.

### 1.2 Процес кластеризації і застосування кластерного аналізу

Процес кластеризації залежить від обраного методу і майже завжди є ітеративним. Він може стати захоплюючим процесом і включати безліч експериментів з вибору різноманітних параметрів, наприклад, заходи відстані, типу стандартизації змінних, кількості кластерів. Однак експерименти не повинні бути самоціллю - адже кінцевою метою кластеризації є отримання змістовних відомостей про структуру досліджуваних даних. Отримані результати потребують подальшої інтерпретації, дослідження та вивчення властивостей і характеристик об'єктів для можливості точного опису сформованих кластерів.

Кластерний аналіз застосовується в різних областях. Він корисний, коли потрібно класифікувати велику кількість інформації. Огляд багатьох опублікованих досліджень, що проводяться за допомогою кластерного аналізу, дав Хартіган, 1975. Так, в медицині використовується кластеризація захворювань, лікування захворювань і їх симптомів, а також таксономія пацієнтів, препаратів. В археології встановлюються таксономії кам'яних споруд і древніх об'єктів. У маркетингу це може бути задача сегментації конкурентів і споживачів. У менеджменті прикладом задачі кластеризації буде розбиття персоналу на різні групи, класифікація споживачів і постачальників, виявлення схожих виробничих ситуацій, при яких виникає брак. У медицині - класифікація симптомів. У соціології завдання кластеризації - розбиття респондентів на однорідні групи. Ефективність алгоритмів, насамперед, залежить від розподілу вхідних даних і типу цих даних.

Багаторазові спроби класифікації методів кластерного аналізу призводять до десяткам, а то й сотням різноманітних класів. Таке різноманіття породжується великою кількістю можливих способів обчислення відстані між окремими спостереженнями, чи не меншою кількістю методів обчислення відстані між окремими кластерами в процесі кластеризації і різноманітними оцінками оптимальності кінцевої кластерної структури. Найбільшого поширення набули дві групи алгоритмів кластерного аналізу: ієрархічні (агломеративні) методи і ітеративні методи угруповання.

### 1.3 Агломеративні методи кластерного аналізу

У агломеративного-ієрархічних методах, спочатку всі об'єкти спостереження розглядаються як окремі, самостійні кластери, що складаються всього лише з одного елемента. Якщо прийняти, що обсяг вибірки дорівнює N, то в цьому випадку можна використовуючи ту чи іншу метрику, обчислити відстані між всіма можливими парами об'єктів. Таких відстаней буде N * N.

Далі, з урахуванням того, що в реальних даних використовуються не дві ознаки, а десятки, а іноді й сотні, можна уявити який великий обсяг обчислень необхідно виконати навіть для цієї найпростішої операції. Очевидно, що без використання потужної обчислювальної техніки реалізація кластерного аналізу даних вельми проблематична.

Матриця відстаней може бути отримана за допомогою різноманітних метрик: евклідової, Махаланобіса, сімейства метрик Маньківського. Вибір метрики обирається самим дослідником. Після обчислення матриці відстаней починається процес агломерації, що проходить послідовно крок за кроком. На першому кроці цього процесу два вихідних спостереження монокластера, між якими саме мінімальне відстань, об'єднуються в один кластер, що складається вже з двох об'єктів спостережень. Таким чином, замість колишніх N монокластеров (кластерів, що складаються з одного об'єкта) після першого кроку залишиться N-1 кластерів, з яких один кластер буде містити два об'єкти спостереження, а N-2 кластерів будуть як і раніше складатися всього лише з одного об'єкта. Відзначимо, що на другому кроці можливі різні методи об'єднання між собою N-2 кластерів. Це викликано тим, що один з цих кластерів вже містить два об'єкти. З цієї причини виникає два основних питання:

– як обчислювати координати такого кластера з двох (а далі і більше двох) об'єктів;

– як обчислювати відстань до таких поліоб'єктних кластерів від монокластеров і між поліоб'єктнимі кластерами.

Це і визначає остаточну структуру підсумкових кластерів (під структурою кластерів мається на увазі склад окремих кластерів і їх взаємне розташування в багатовимірному просторі). Різноманітні комбінації метрик і методів обчислення координат і взаємних відстаней кластерів і породжують те різноманіття методів кластерного аналізу, про який було сказано вище.

На другому кроці в залежності від обраних методів обчислення координат кластера складається з декількох об'єктів і способу обчислення межкластерних відстаней можливе або повторне об'єднання двох окремих спостережень в новий кластер, або приєднання одного нового спостереження до кластеру, що складається з двох об'єктів. Для зручності більшість програм агломеративного-ієрархічних методів по закінченні роботи можуть надати для перегляду два основних графіка. Перший графік називається дендрограмма, що відображає процес агломерації, злиття окремих спостережень в єдиний остаточний кластер. Цей графік схематично нагадує дерево, за що і отримав таку назву. Нижче наведено рисунок з такою дендрограммою для навчального прикладу, що складається з 5 спостережень за двома змінним.

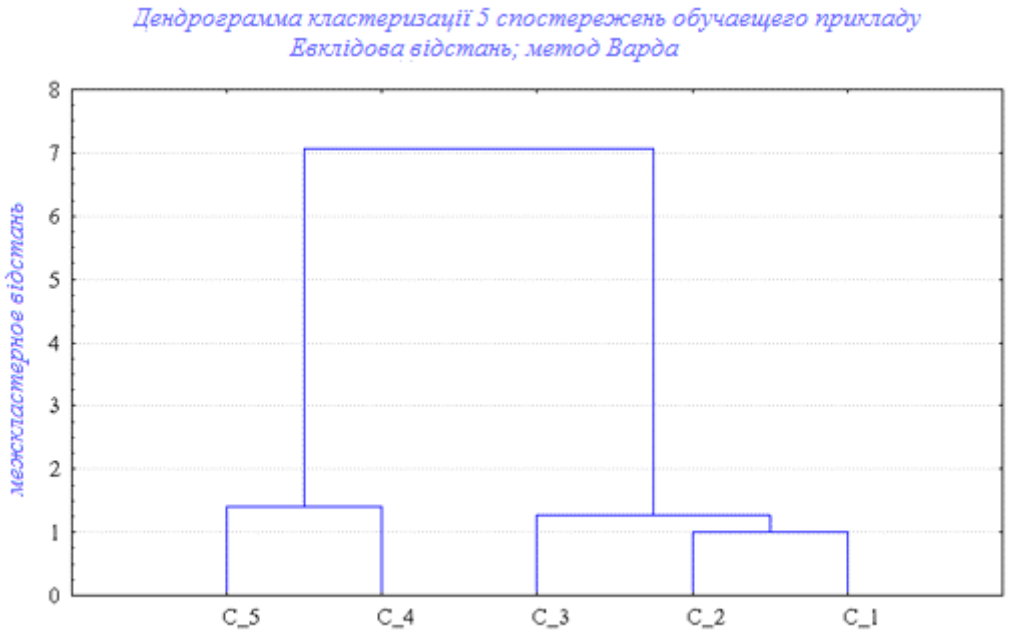



Рисунок 1.3 – Діаграма кластеризації 5 спостережень навчального прикладу

Вертикальна вісь такого графіка являє собою вісь між кластерного відстані, а по горизонтальній осі відзначені номера об'єктів - випадків використаних в аналізі. З цієї дендрограмми видно, що спочатку об'єднуються в один кластер об'єкти №1 і №2, оскільки

відстань між ними саме мінімальне і дорівнює 1. Це злиття відображається на графіку горизонтальною лінією з'єднує вертикальні відрізки виходять з точок позначених як С_1 і С_2. Звернемо увагу на те, що сама горизонтальна лінія проходить точно на рівні межкластерного відстані рівного 1.

Далі на другому кроці до цього кластеру, що включає в себе вже два об'єкти, приєднується об'єкт №3, позначений як С_3. На наступному кроці відбувається об'єднання об'єктів №4 та №5, відстань між якими дорівнює 1,41. І на останньому кроці відбувається об'єднання кластера з об'єктів 1, 2 і 3 з кластером з об'єктів 4 і 5. На графіку видно, що відстань між цими двома передостанніми кластерами (останній кластер включає в себе всі 5 об'єктів) більше 5, але менше 6, оскільки верхня горизонтальна лінія, що з'єднує два передостанніх кластера, проходить на рівні приблизно рівному 7, а рівень з'єднання об'єктів 4 і 5 дорівнює 1,41.

Розташована нижче дендрограмма отримана при аналізі реального масиву даних складається з 70 об'єктів, кожен з яких характеризувався 12 ознаками.

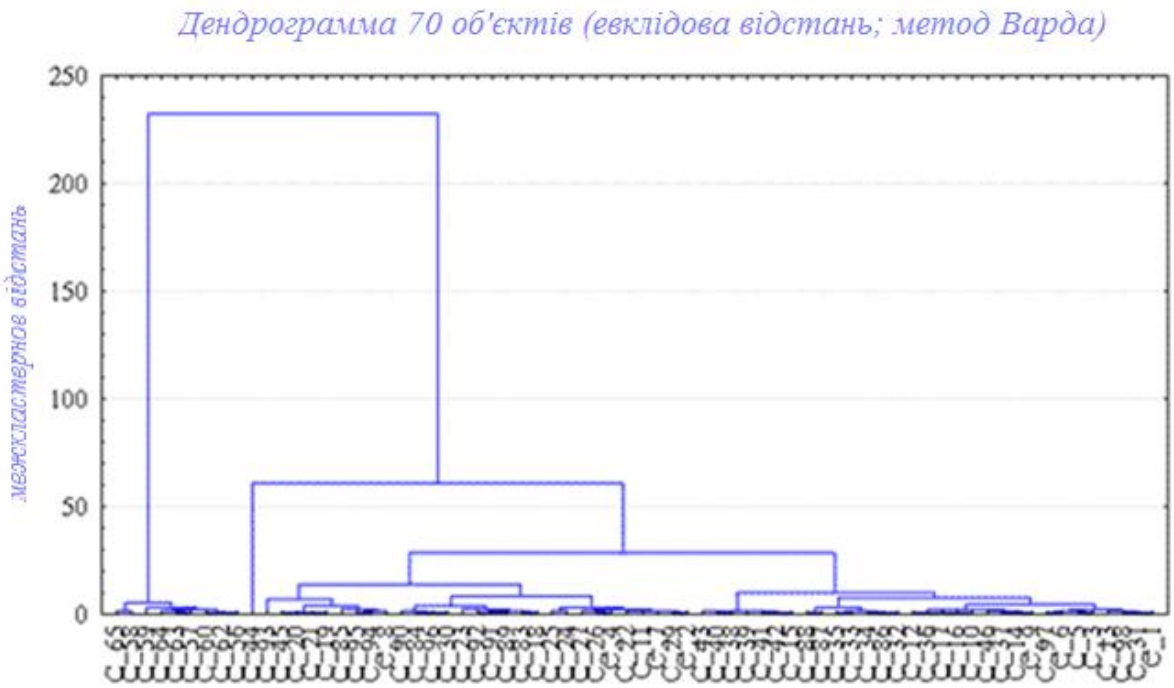



Рисунок 1.4 – Дендрограмма 70 об'єктів (метод Варда)

З графіка видно, що на останньому кроці, коли відбулося злиття двох останніх кластерів, відстань між ними близько 200 одиниць. Видно, що перший кластер (домовимося, що він розташований зліва) включає в себе набагато менше об'єктів (9), ніж другий кластер (розташований праворуч). Оскільки всього в аналізі використано 70 об'єктів, то в другому кластері 61 об'єкт.

Нижче наведено збільшений ділянку дендрограмми на якому досить чітко видно номери спостережень, що позначаються як C_65, C_58 і т.д. (зліва направо): 65, 58, 59, 64, 63, 57, 60, 62, 56, 44, 94 і т.д.

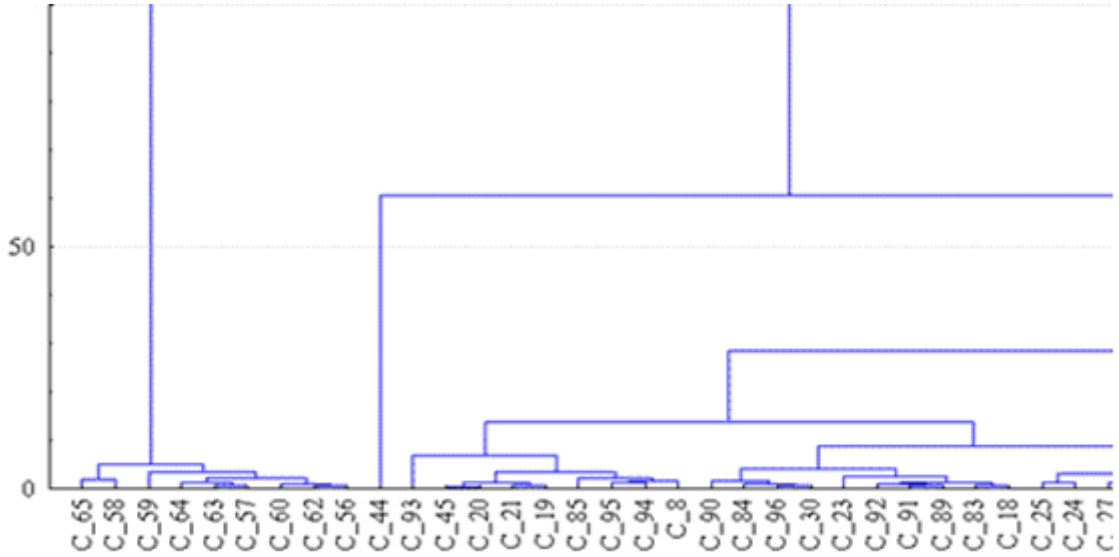



Рисунок 1.5 – Збільшена ділянка дендрограмми 70 об'єктів

Видно, що об'єкт 44 являє собою монокластер об'єднується на передостанньому кроці з правим кластером і потім вже на останньому кроці всі спостереження об'єднуються в один кластер.

Другий графік, який будується в таких процедурах - це графік зміни межкластерних відстаней на кожному кроці об'єднання. Нижче наведено подібний графік для наведеної вище дендрограмми.

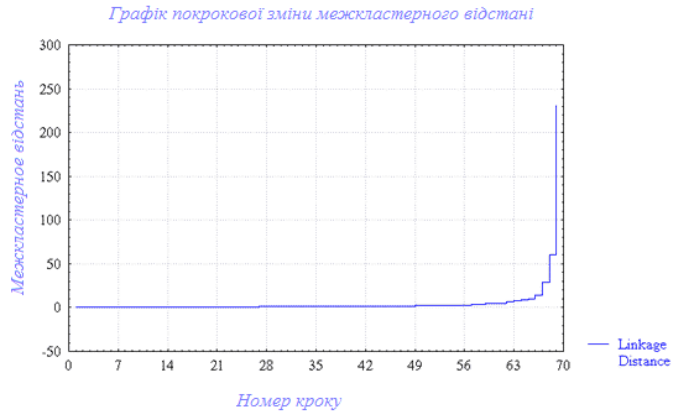



Рисунок 1.6 – Графік покрокової зміни межкластерної відстані

Можна вивести в табличному вигляді результати об'єднання об'єктів на кожному кроці кластеризації. У більшості таких таблиць щоб уникнути плутанини використовується різна термінологія для позначення вихідних спостережень - монокластеров, і власне кластерів складаються з двох і більше спостережень. Нижче наводитися початкова та кінцева частини такої таблиці для наведених вище графіків, отриманих при кластеризації 70 об'єктів.

Таблиця 1.2 - Результати об'єднання на кожному кроці кластеризації

| Відстань між кластерами. | Об'єкт номер 1. | Об'єкт номер 12. | Об'єкт номер 13 | Об'єкт номер 14 | Об'єкт номер 15 |
|---|---|---|---|---|---|
| 4358891 | C_20 | C_45 | На першому кроці об'єдналися N 20 і N45 | | |
| 5337617 | C_19 | C_21 | На другому кроці об'єдналися N 19 і N21 | | |
| 5514541 | C_16 | C_17 | На третьому кроці об'єдналися N 16 і N17 | | |
| 5886407 | C_14 | C_37 | На четвертому кроці об'єдналися N 14 і N37 | | |
| 6134596 | C_1 | C_31 | На п'ятому кроці об'єдналися N1 і N31 | | |
| 6265001 | C_30 | C_96 | На шостому кроці об'єдналися N 30 і N96 | | |
| 6297126 | C_10 | C_16 | C_17 | На цьому кроці до кластеру що складається з 16 і 17 об'єктів приєднався об'єкт під номером 10. Видно, що на кожному кроці межкластерное відстань збільшується. | |
| 6434321 | C_18 | C_83 | | | |
| 6493070 | C_15 | C_42 | | | |
| 7051605 | C_38 | C_40 | | | |
| 7265998 | C_15 | C_42 | C_41 | | |
| 7267050 | C_4 | C_26 | | | |
| 7689960 | C_9 | C_14 | C_37 | | |

Нижня частина цієї таблиці показує, що на передостанньому кроці приєднався об'єкт номер 44, а потім відбулося злиття 9 спостережень лівого кластера і 61 спостереження лівого кластера в один загальний кластер.приєднався об'єкт номер 44, а потім відбулося злиття 9 спостережень лівого кластера і 61 спостереження лівого кластера в один загальний кластер.

Для того щоб продемонструвати залежність кластерної структури від вибору метрики і вибору алгоритму об'єднання кластерів, наведемо нижче дендрограмма відповідає алгоритму повного зв'язку. І тут ми бачимо, що об'єкт №44 об'єднується у всій решті вибіркою на самому останньому кроці.

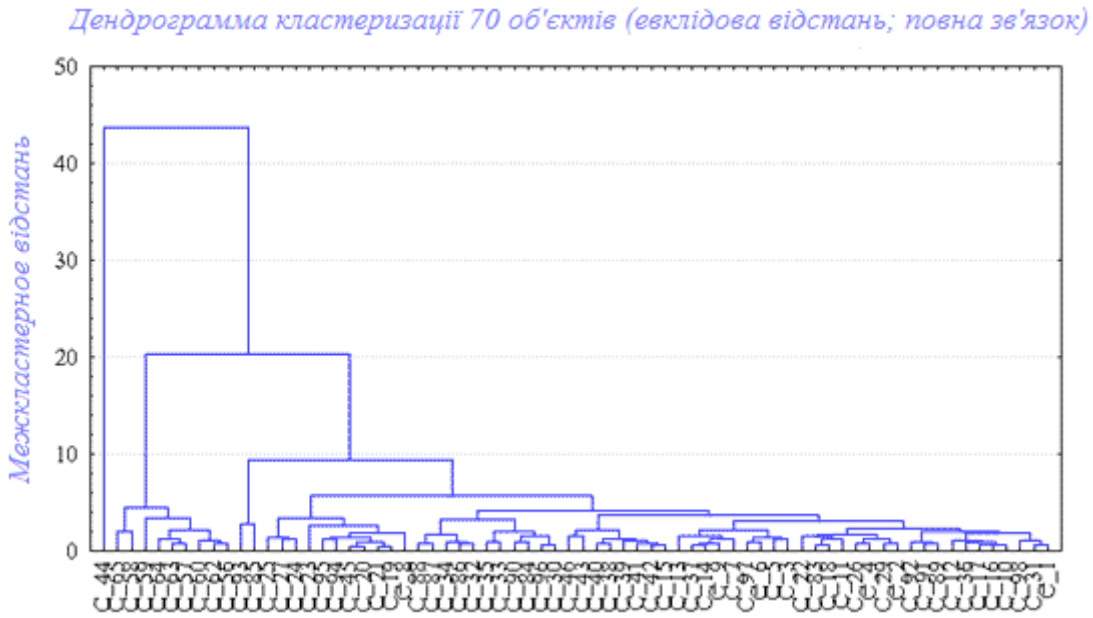



Рисунок 1.7 – Дендрограмма кластеризації 70 об'єктів (повна зв'язок)

А тепер порівняємо її з іншого дендрограмма, отриманої при використанні методу одиночній зв'язку до тих же самим даними. На відміну від методу повної зв'язку, видно, що цей метод породжує довгі ланцюжки послідовно приєднуються один до одного об'єктів. Однак у всіх трьох випадках можна говорити про те, що виділяється дві основні угруповання об'єктів.

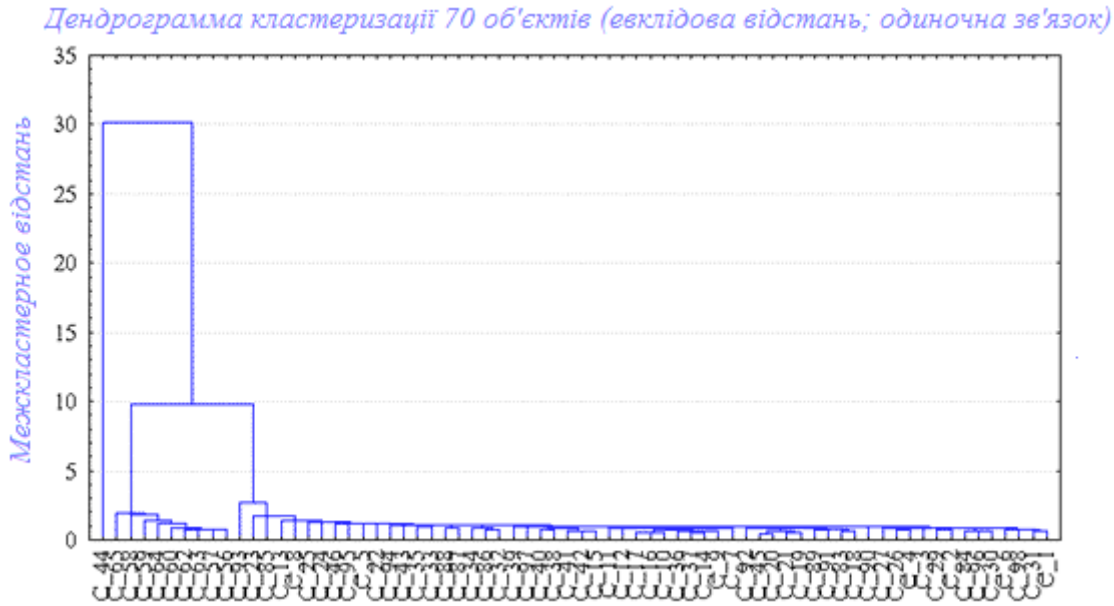



Рисунок 1.8 – Дендрограмма кластеризації 70 об'єктів (одинарна зв'язок)

Звернемо також увагу на те, що у всіх трьох випадках об'єкт №44 приєднується як монокластер, хоча і на різних етапах процесу кластеризації. Виділення таких монокластеров є непоганим засобом виявлення аномальних спостережень, які називаються в літературі також викидами. Порівняємо відстань між двома кластерами, що об'єдналися на передостанньому кроці, зі стрибком цієї відстані на останньому кроці. Як бачимо, відстань між лівим і правим кластерами близько 7-8 одиниць, що особливо добре видно на графіку трохи нижче. Тоді як відстань між об'єктом №44 і кластером що складається з всіх інших об'єктів вибірки становить порядку 20 одиниць. Це також є ще одним підтвердження аномальність цього об'єкта. Подальше вивчення особливостей цього об'єкта, можливо, дозволить встановити причини такої аномальність.

Видалимо цей підозрілий об'єкт №44 і знову проведемо кластеризацию. Отримана при цьому дендрограмма наведена нижче.

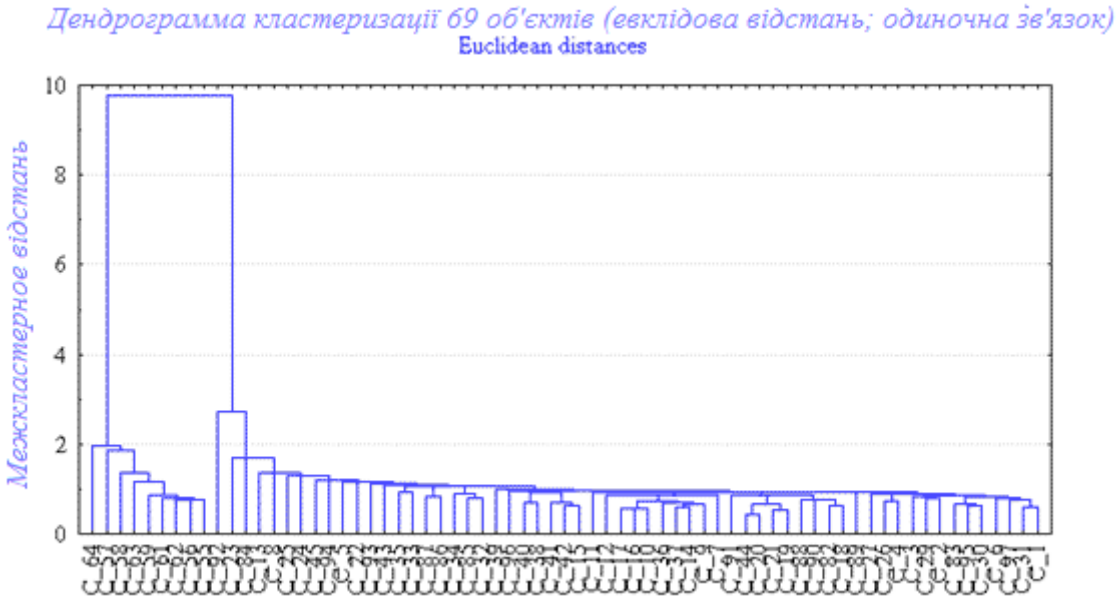



Рисунок 1.9 – Дендрограмма кластеризації 69 об'єктів

Видно, що цепочечний ефект зберігся, як збереглося і розбиття на дві локальні групи спостережень. Відзначимо, що агломеративні (об'єднуючі) методи на останньому кроці об'єднують всі спостереження в одні кластер. Тому використовувати побудовану дендрограмма для виділення тієї чи іншої кількості окремих кластерів можна шляхом розрізання цієї дендрограмми на певному значенні межкластерного відстані. Фактично це означає, що ми проводимо горизонтальну лінію, розсікаючи дерево зв'язків в тому місці, де спостерігається максимальний стрибок у зміні межкластерного відстані.

Крім об'єднують методів ієрархічної кластеризації існують і протилежні методи - дівізімние, в яких на початковому етапі вся вибірка розглядається як єдиний кластер, а потім вже починається процес його поділу на складові частини. Процес поділу продовжується до тих пір, поки кожне спостереження чи не перетвориться в окремий кластер. У свою чергу дівізімние алгоритми поділяються на монотетіческіе і політетіческіе. У монотетіческой класифікації розподіл проводиться на підставі єдиного ознаки, що має максимальну інформативність. У політетіческіх ж алгоритмах враховуються всі ознаки. Оскільки дані алгоритми оперують відстанями між спостереженнями, то в деяких програмах передбачена можливість роботи нема з вихідною матрицею об'єкт - ознака, а з симетричною матрицею відстаней між спостереженнями.

### 1.4 Ітераційні методи кластеризації

Серед ітераційних методів найбільш популярним методом є метод k-середніх Мак-Кіна. На відміну від ієрархічних методів у більшості реалізацій цього методу сам користувач повинен задати шукане число кінцевих кластерів, яке зазвичай позначається як k. Як і в ієрархічних методах кластеризації, користувач при цьому може вибрати той чи інший тип метрики. Різні алгоритми методу k-середніх відрізняються і способом вибору початкових центрів задаються кластерів. У деяких варіантах методу сам користувач може (або повинен) задати такі початкові точки, або вибравши їх із реальних спостережень, або задавши координати цих точок по кожній із змінних. В інших реалізаціях цього методу вибір заданого числа k початкових точок проводиться випадковим чином, причому ці початкові точки можуть надалі уточнюватися в кілька етапів. Можна виділити 4 основних етапи таких методів:

– вибираються або призначаються k спостережень, які будуть первинними центрами кластерів;

– при необхідності формуються проміжні кластери приписуванням кожного спостереження до найближчих заданих кластерним центрам;

– після призначення всіх спостережень окремим кластерам проводиться заміна первинних кластерних центрів на кластерні середні;

– попередня ітерація повторюється до тих пір, поки зміни координат кластерних центрів не стануть мінімальними.

У деяких варіантах цього методу ми можемо задати числове значення критерію, трактують як мінімальна відстань для відбору нових центрів кластерів. Спостереження не розглядатиметься як претендент на новий центр кластера, якщо його відстань до замінного центру кластера перевищує задане число. Такий параметр в ряді програм називається радіусом. Крім цього параметра можливе завдання і максимального числа ітерацій або досягнення певного, зазвичай достатньо малого, числа, з яким порівнюється зміна відстані для всіх кластерних центрів. Цей параметр зазвичай називається конвергенцією, тому відображає збіжність ітераційного процесу кластеризації. Нижче ми наведемо частину результатів, які отримані при використанні методу k-середніх Мак-Кіна до попередніх даних. Число шуканих кластерів задавалося спочатку рівним 3, а потім 2. Перша їх частина містить результати однофакторного дисперсійного аналізу, в якому в якості групують фактора виступає номер кластера. У першому стовпці список 12 змінних, далі йдуть суми квадратів (SS) і ступеня свободи (df), потім F-критерій Фішера і в останньому стовпці - досягнутий рівень значимості p.

Таблиця 1.3 - Кластеризація методом Мак-Кіна

| | Between | | Within | | | |
|---|---|---|---|---|---|---|
| Змінні | SS | df | SS | df | F | p |
| X1 | 1606,203 | 1 | 165,2964 | 68 | 660,7634 | 0,000000 |
| X2 | 621,964 | 1 | 916,1421 | 68 | 46,1648 | 0,000000 |
| X3 | 0,305 | 1 | 3,0978 | 68 | 6,6914 | 0,011832 |
| X4 | 0,146 | 1 | 3,2248 | 68 | 3,0697 | 0,084272 |
| X5 | 30,464 | 1 | 65,9877 | 68 | 31,3934 | 0,000000 |
| X6 | 6,936 | 1 | 17,2187 | 68 | 27,3910 | 0,000002 |
| X7 | 18,213 | 1 | 70,8901 | 68 | 17,4706 | 0,000085 |
| X8 | 0,160 | 1 | 0,6721 | 68 | 16,1832 | 0,000147 |
| X9 | 7,981 | 1 | 11,2471 | 68 | 48,2525 | 0,000000 |
| X10 | 6,943 | 1 | 8,6925 | 68 | 54,3172 | 0,000000 |
| X11 | 8,598 | 1 | 5,4052 | 68 | 108,1661 | 0,000000 |
| X12 | 7,673 | 1 | 3,6936 | 68 | 141,2533 | 0,000000 |

Як видно з цієї таблиці, нульова гіпотеза про рівність середніх значень в трьох групах відкидається. Нижче наведено графік середніх значень всіх змінних по окремих кластерам. Ці ж кластерні середні змінних наведені далі у вигляді таблиці.

Таблиця 1.4 - Таблиця змінних по окремих кластерам

| Змінна | Кластер №1 | Кластер №2 | Кластер №3 |
|---|---|---|---|
| X1 | 46,62000 | 33,78334 | 48,11867 |
| X2 | 51,00000 | 89,04000 | 80,62035 |
| X3 | 1,75000 | 0,37856 | 0,55613 |
| X4 | 1,25000 | 0,36733 | 0,49113 |
| X5 | 12,75000 | 3,25667 | 5,10217 |
| X6 | 5,00000 | 0,83222 | 1,71883 |
| X7 | 12,25000 | 3,68889 | 5,09550 |
| X8 | 0,80000 | 0,05556 | 0,18833 |
| X9 | 4,75000 | 0,82222 | 1,78233 |
| X10 | 4,50000 | 0,97778 | 1,87567 |
| X11 | 3,25000 | 0,35444 | 1,37067 |
| X12 | 2,75000 | 0,22222 | 1,18567 |

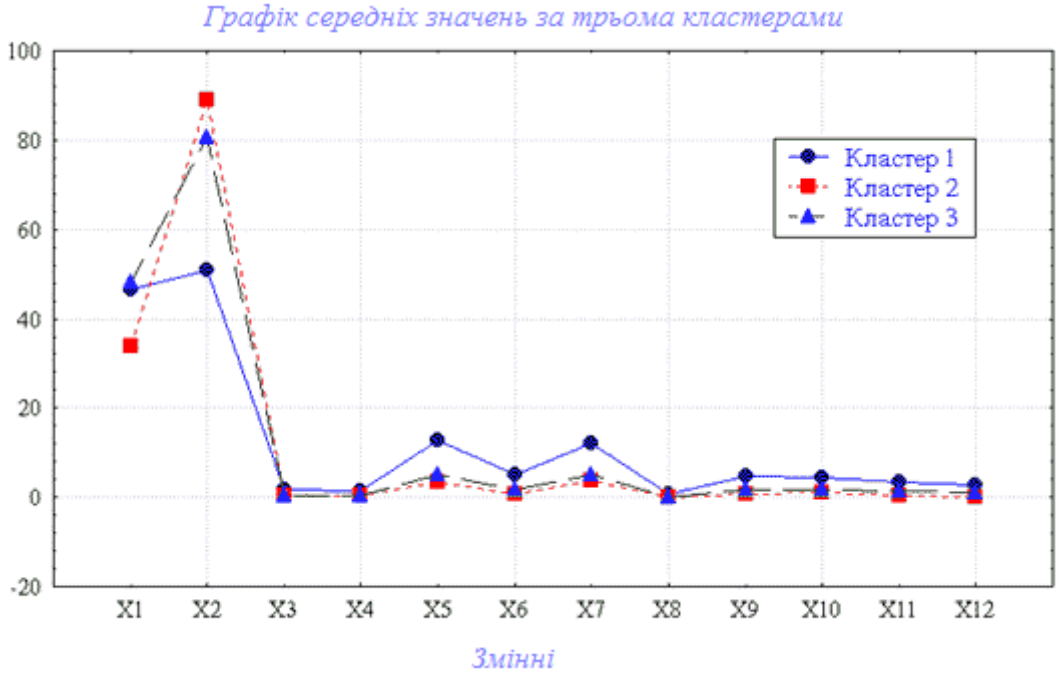



Рисунок 1.10 – Графік середніх значення за трьома кластерами

Аналіз середніх значень змінних для кожного кластера дозволяє зробити висновок про те, що за ознакою X1 кластери 1 і 3 мають близькі значення, тоді як кластер 2 має середнє значення набагато менший, ніж в інших двох кластерах. Навпаки, за ознакою X2 перший кластер має саме мінімальне значення, тоді як 2-й і 3-й кластери мають більш високі і близькі між собою середні значення. Для ознак X3-X12 середні значення в кластері 1 значно вище, ніж в кластерах 2 і 3.

Наступна таблиця дисперсійного аналізу результатів кластеризації на два кластери також показує необхідність відхилення нульової гіпотези про рівність групових середніх

майже по всіх 12 ознаками, за винятком змінної X4, для якої досягнутий рівень значимості виявився більше 5%.

Таблиця 1.5 - Кластеризація на два кластери

| | Between | | Within | | | |
|---|---|---|---|---|---|---|
| Змінні | SS | df | SS | df | F | p |
| X1 | 1606,203 | 1 | 165,2964 | 68 | 660,7634 | 0,000000 |
| X2 | 621,964 | 1 | 916,1421 | 68 | 46,1648 | 0,000000 |
| X3 | 0,305 | 1 | 3,0978 | 68 | 6,6914 | 0,011832 |
| X4 | 0,146 | 1 | 3,2248 | 68 | 3,0697 | 0,084272 |
| X5 | 30,464 | 1 | 65,9877 | 68 | 31,3934 | 0,000000 |
| X6 | 6,936 | 1 | 17,2187 | 68 | 27,3910 | 0,000002 |
| X7 | 18,213 | 1 | 70,8901 | 68 | 17,4706 | 0,000085 |
| X8 | 0,160 | 1 | 0,6721 | 68 | 16,1832 | 0,000147 |
| X9 | 7,981 | 1 | 11,2471 | 68 | 48,2525 | 0,000000 |
| X10 | 6,943 | 1 | 8,6925 | 68 | 54,3172 | 0,000000 |
| X11 | 8,598 | 1 | 5,4052 | 68 | 108,1661 | 0,000000 |
| X12 | 7,673 | 1 | 3,6936 | 68 | 141,2533 | 0,000000 |

Нижче наведені графіки таблиця групових середніх для випадку кластеризації на два кластери.

Таблиця 1.6 - Середні значення при кластеризації на два кластери

| Переменные | Кластер №1 | Кластер № 2 |
|---|---|---|
| X1 | 33,78334 | 48,09410 |
| X2 | 89,04000 | 80,13477 |
| X3 | 0,37856 | 0,57570 |
| X4 | 0,36733 | 0,50357 |
| X5 | 3,25667 | 5,22754 |
| X6 | 0,83222 | 1,77262 |
| X7 | 3,68889 | 5,21279 |
| X8 | 0,05556 | 0,19836 |
| X9 | 0,82222 | 1,83098 |
| X10 | 0,97778 | 1,91869 |
| X11 | 0,35444 | 1,40148 |
| X12 | 0,22222 | 1,21131 |


Рисунок 1.11 – Графік середніх значення з двох кластерами

У тому випадку, коли ми не маємо можливості заздалегідь визначитися з найбільш імовірним числом кластерів, ми змушений повторити розрахунки, задаючи різне їх число, подібно до того, як це було зроблено вище. А потім, порівнюючи отримані результати між собою, зупинитися на одному з найбільш прийнятних варіантів кластеризації.

### 1.5 Представлення результатів кластеризації

Крім тих результатів кластерного аналізу, про які вже йшлося вище (середні по кластерам, дендрограмми, дисперсійний аналіз тощо), наводиться і така важлива інформація, як середня відстань до центру кластера (для кожного з кластерів), максимальне і мінімальне відстань і, відповідно, найбільш віддалене і найбільш близьке до центру кластера спостереження (типовий, еталонний представник даного кластера), а також частка дисперсії відстані пояснюється кластерним розбиттям (коефіцієнт детермінації R2). Не менш важливою інформацією є і приналежність конкретного спостереження до того чи іншого кластеру.

У ряді програм ієрархічного кластерного аналізу всі вузли дендрограмми ідентифікуються послідовними номерами.

Далі для кожного з цих номерів цих вузлів можна отримати докладний список вихідних об'єктів - спостережень, що входять як складові елементи в даний кластер. Аналогічну ідентифікацію приналежності вихідних спостережень до даної кластеру виробляють і програми, виконують кластерний аналіз за алгоритмом k-середніх Мак-Кіна. У цьому випадку до вихідних ознаками додається ще одна ознака, в який поміщається номер того кластера, в який включено дане спостереження. У тих ітераційних програмах кластерного аналізу, де є вибір оптимизирующего функціоналу, як правило, є і можливість бачити або на кожній ітерації, або наприкінці аналізу, числове значення цього функціоналу. Це дозволяє в разі перебору декількох варіантів поєднань просторової метрики та інших параметрів алгоритму зробити вибір більш оптимальної класифікації. Іншою важливою інформацією є Матриця відстаней, в якій зберігається матриця взаємних відстаней між об'єктами, обчислена в обраній користувачем метриці. Дана матриця відстаней може бути збережена окремо і згодом використана самостійно в інших статистичних процедурах і методах. Однак найбільш цікавою можливістю є представлення отриманих кластерів в осях спеціальних змінних, в яких вдалість отриманої класифікації можна оцінити візуально.

### 1.6 Стійкість і якість кластеризації

Очевидно, що було б неправильно ставити питання про те, наскільки абсолютна та чи інша класифікація, отримана за допомогою методів кластерного аналізу. Все в світі відносно. Тому в застосуванні процедур кластерного аналізу важливим аспектом є стійкість структури кластерів, що відображає реальну об'єктивність класифікації. У прикладах вище ми бачили, що при зміні методу кластеризації подібна стійкість виявлялася в тому, що на дендрограмма досить чітко проглядалися два кластери.

В якості одного з можливих способів перевірки стійкості результатів кластерного аналізу може бути використаний метод порівняння результатів отриманих для різних алгоритмів кластеризації. Інші шляхи, це так званий бутстреп-метод запропонований Б. Ефрона в 1977р., Методи складного ножа і ковзного контролю. Найбільш простий засіб перевірки стійкості кластерного рішення може полягати в тому, щоб вихідну вибірку випадковим чином розділити на дві приблизно рівні частини, провести кластеризацію обох частин і потім порівняти отримані результати. Більш трудомісткий шлях передбачає послідовне виключення спочатку першого об'єкта і кластеризацию залишилися (N - 1)

об'єктів. Далі послідовно проводячи цю процедуру з виключенням другого, третього. об'єктів аналізується структура всіх N отриманих кластерів. Інший алгоритм перевірки стійкості передбачає багаторазове розмноження, дублювання вихідної вибірки з N об'єктів, потім об'єднання всіх дубльованих вибірок в одну велику вибірку (псевдогенеральную сукупність) і випадкове витяг з неї нової вибірки з N об'єктів. Після цього проводиться кластеризація цієї вибірки, далі витягується нова випадкова вибірка і знову проводиться кластеризація. Очевидно, що це також досить трудомісткий шлях.

Не менше проблем і при оцінці якості кластеризації. Як ми вже говорили вище, відомо досить багато алгоритмів оптимізації кластерних рішень. Перші роботи, які містили формулювання критерію мінімізації внутрікластерной дисперсії і алгоритм (типу k-середніх) пошуку оптимального рішення з'явилися в 50-х роках. У 1963р. у статті Дж. Уорда також викладався подібний оптимізаційний ієрархічний алгоритм. Все це говорить про те, що не існує універсального критерію оптимізації кластерного рішення. Все це ускладнює вибір дослідником оптимального рішення. У такій ситуації найкращим способом утвердитися в тому, що знайдене кластерне рішення є на даному етапі дослідження оптимальним, є тільки узгодженість цього рішення з висновками, отриманими за допомогою інших методів багатовимірної статистики.

На користь виведення про оптимальність кластеризації служать також і позитивні результати перевірки пророкують моментів отриманого рішення вже на інших об'єктах дослідження. При використанні ієрархічних методів кластерного аналізу можна рекомендувати порівняння між собою кількох графіків покрокового зміни межкластерного відстані. При цьому перевагу слід віддати тому варіанту, для якого спостерігається плоска лінія такого збільшення від першого кроку до декількох передостанніх кроків з різким вертикальним підйомом цього графіка на останніх 1-2 кроках кластеризації.

### 1.7 Постановка завдання дослідження

Метою роботи є дослідження методу класифікації багатовимірних об'єктів з застосуванням теорії розмитих множин. Для досягнення мети нам необхідно виконати наступні кроки:

– провести аналіз алгоритмів класу FOREL, SKAT, NTPP, KOLLAPS, BIGFOR, DINA, ROST, ваг ознак і перевірку інформативності ознак, методи порівняння алгоритмів таксономії;

– провести аналіз теорії розпливчастих множин, функції приналежності розпливчато мети, багатокрокових процесів прийняття рішення;

– розробити алгоритм класифікації, який буде працювати більш з чим 5-параметрами даних.

# 2 АЛГОРИТМИ КЛАСИФІКАЦІЇ

## 2.1 Алгоритми класу FOREL

Людина, групуючи об'єкти, керується деяким критерієм (F). Отже, гіпотеза сильніша, ніж гіпотеза компактності (H), повинна бути сформульована з урахуванням цього критерію. Тестовий алгоритм повинен допускати тільки таку угруповання, яке задовольняє критерію F. Іншими словами, слід конкретизувати поняття подібності, схожості об'єктів.

Вважаємо, що ознаки об'єктів задані в сильних шкалах, і ми можемо працювати в метричних просторах. Зокрема, можемо в евклідовому багатовимірному просторі ознак ввести відстань між точками.

Нехай $C_j$ = ($x_{j1}$, $x_{j2}$, .., $x_{ji}$, .., $x_{jn}$) - координати центру ваги j-го таксона,

Нагадаю, що координати центру ваги системи матеріальних точок визначаються за формулою:

$$x_{i,j} = \frac{\sum_{k-1}^{N_j} m_{j,k} x'_{j,i,k}}{\sum_{k-1}^{N_j} m_{j,i,k}}, \tag{2.1}$$

де – і - я координата k-ої точки j-го таксона, $N_j$ - число точок j-го таксону. У нашому випадку всі маси рівні 1, і отримуємо:

$$x_{i,j} = \frac{\sum_{k-1}^{N_j} x'_{j,i,k}}{N_j}. \tag{2.2}$$

$P_{ji}$- відстань між центром ваги і довільної точкою $a_i$,

Сума таких відстаней j-го таксона:

$$\rho_j = \sum_j P_{j,i}. \tag{2.3}$$

$$F = \sum_{j-1}^{k} \rho_j. \tag{2.4}$$

Сенс критерію схожості на центр полягає в тому, щоб знайти таке розбиття m об'єктів на k таксонів, при якому F мінімальна.

## 2.2 Алгоритм FOREL 1

Алгоритм роботи з об'єктами, описаними кількісними характеристиками, тобто ознаками, виміряними в шкалі інтервалів, різниць, відносин або абсолютною. Об'єкти, включені в один таксон, потрапляють в гіперсферу з певним центром С і радіусом R. Змінюючи радіус можна отримати різне число таксонів k.

При фіксованому R алгоритм працює таким чином. Центр $C^{(1)}$ гіперсфери радіуса R поміщається в будь-яку точку безлічі об'єктів. Визначаються точки, які опинилися всередині цієї сфери. Для цього обчислюються відстані ρ від точки $C^{(1)}$ до всіх М точок, і ті з них для яких $\rho \leq R$, вважаються внутрішніми. Обчислимо центр ваги, і центр сфери переміщається в цей центр ваги $C^{(2)}$. Для нового положення сфери перебувають знову внутрішні точки і їх центр тяжіння. Процедура зміщення гіперсфери радіуса R повторюється до тих пір поки не перестануть змінятся координати центру ваги $C^{(i)}$. Невеликі відхилення сфери від центру ваги привели б її до руху в бік цього центру, так що становище сфери в області локального згустку є стійким. Назвемо цю стійку сферу таксоном $S^{(1)}$. Внутрішні точки з подальшого розгляду виключаємо. Потім центр наступної такої ж гіперсфери поєднується з будь з решти точок і процедура повторюється. Очевидно, що кількість таксонів k тим більше, чим менше радіус таксона k. Бажане для користувача кількість таксонів k може бути знайдено відповідним підбором радіуса R. З цією метою радіус послідовно зменшується, з рівномірним кроком ΔR від Rmax, при якому всі використані точки об'єднуються в один таксон, до тих пір, поки не буде отримано кількість таксонів, Найближчим до заданого числа k. Змінюючи R, можна простежити генезис таксонів, побачити на які дрібні таксони розпадається той чи інший таксон. Алгоритм FOREL дає рішення за кінцеве число кроків, однак, очевидно, що рішення може залежати від порядку перегляду точок і тому буває не єдиним. Вибір одного рішення з багатьох можливих робиться за критерієм якості F, який для алгоритму FOREL виглядає так:

$$F = \sum_{l=1}^{k} \sum_{\alpha}^{m_\alpha} \rho_\alpha^2 (X_\alpha; C^l), \qquad (2.5)$$

де $\rho_\alpha^2 (X_\alpha; C^{(l)})$, - квадрат евклидова відстані від точки α з координатами $X_\alpha$ до центру свого таксона $C^{(l)}$, а $m_l$- число об'єктів в таксоні $S_l$. Кращий варіант таксономії → *min F*. Алгоритм FOREL закінчує свою роботу при k' ≥ k.

### 2.3 Алгоритм FOREL 2

Даний алгоритм працює при заданому k. Радіус R змінюється не з постійним кроком, а з перемінним: ΔR на кожному черговому кроці таксономії зменшується вдвічі, так що при пошуку і-го варіанта таксономії радіус сфери дорівнює:

$R_i=R_i \pm \Delta R_i$, де $\Delta R_i=Rmax/2^i$, а знак + або - вибирається залежно від того скільки таксонів було отримано на (i-1) кроці. Якщо $k^{(i-1)} \leq k$, то знак -, а при $k^{(i-1)}>k$ - +. Алгоритм працює доти, поки методом послідовних наближень радіуса що не отримана k таксонів при мінімальному значенні радіуса R, чи не буде вичерпаний ліміт ітерацій.

Функція якості таксономії в алгоритмі FOREL2 виглядає так:

$$F = f(k^i)\ \sum_{l=1}^{k} \sum_{\alpha}^{m_\alpha} \rho_\alpha^2\ (X_\alpha; C^l), \qquad (2.6)$$

де $f(k^i) = \begin{cases} \infty, \text{якщо } k^i \neq k, \\ 1, \text{якщо } k^i = k. \end{cases}$

### 2.4 Алгоритм FOREL 5

Даний алгоритм роботи з бінарними ознаками. Центри сфер переміщаються в дискретно просторі, і знаходиться завжди в одній з вершин одиничного N-мірного куба. Визначаються внутрішні точки, і для них визначається вершина гіперкуба, найближча до центру тяжіння цих точок. Для цього по кожній координаті $x_j$ вибирається значення $x_j=0$, якщо не менше половини внутрішніх точок має координату $x_j = 0$, і $x_j = 1$ - в іншому випадку. Параметрами, від яких залежить результат таксономії, служить бажане число таксонів (k) і крок зменшення радіусу таксона ΔR (зазвичай $0{,}1 \times R_{max}$, де $R_{max}$ - відстань від центру ваги всіх об'єктів до самої віддаленої точки. При цьому алгоритм FOREL5 мінімізує суму відстаней від усіх точок ($\alpha_i$) до центрів своїх таксонів ($C^l$).

### 2.5 Алгоритм SKAT

В результаті роботи алгоритму FOREL можуть утворитися нестійкі таксони, випадкові згустки, які тяжіють до одного з отриманих таксонів. Це може статися в результаті недостатньої коректності вихідних даних.

Один з евристичних прийомів для обліку нестійких таксонів реалізований в алгоритмі SKAT.

На вхід програми поступає вихідна множина об'єктів m і результат таксономії за алгоритмом FOREL S. Процес повторюється з тим же радіусом сфер, але на всіх m точках множини і з початковими точками, що збігаються з центрами таксонів.

### 2.6 Алгоритм NTPP

Алгоритм спрямованого таксономического пошуку ознак, виміряних в сильних шкалах, призначений для вибору заданого числа найбільш інформативних ознак. При цьому забезпечується максимально можлива незалежність вибраних ознак один від одного.

Також, як формується завдання угрупування схожих об'єктів (рядків матриці), приблизно однаково провідних себе в даному просторі ознак, ставиться завдання об'єднання ознак, які приблизно однаково проявляють себе на даних об'єктах.

Таксономія стовпців матриць аналогічна таксономії рядків. Отримані таксони інтерпретуються як групи властивостей, обумовлені деяким загальним фактором, так що програма може використовуватися для аналізу структури простору ознак.

Основне завдання, яке вирішується програмою NTPP, полягає у виборі заданого числа (N1) найбільш інформативних ознак з вихідного безлічі (N). Програма проводить таксономию N ознак на IK таксонів (N> IK> N1) за алгоритмом FOREL. В якості міри схожості між ознаками після їх нормування до інтервалу [0,1] вибрано евклідова відстань.

Так як в один таксон групуються ознаки по max подібністю, то між таксонами забезпечується max несхожість. Далі з кожного таксону вибирається по одному типовому (найближчого до середнього) представнику і перебором з IK по N1 формується $C_{IK}^{IN}$ різних підсистем з типових ознак. У даній версії програми кращої підсистеми вважається та, при якій виходить найкраще розпізнавання навчальної вибірки лінійним вирішальним

правилом. Передбачена можливість після перегляду чергових (КТ2) підсистем ознак видавати на друк інформацію про отримані до даного моменту (КР) кращих варіантах. У програмі передбачена можливість отримання декількох (КР) кращих поєднань ознак.

### 2.7 Алгоритм KOLLAPS

Алгоритм KOLLAPS застосовується в задачах виділення локальних згустків точок з фону. Наприклад, виділення яскравих сузір'їв на тлі неба.

На першому етапі вирішення завдання визначаються центри згустків, а на другому - перевіряється, чи дійсно ці точки є центрами стійких таксонів.

Насамперед задається деяке значення порога щільності - кількості точок $m_j$ в таксоні (d). Якщо $m_j$ > d, то запам'ятовується центр таксона, а його точки з подальшого розгляду виключаються. В іншому випадку центр не запам'ятовується, але точки таксона з розгляду виключаються. Потім центр переноситься в будь-яку з решти точок і процес продовжується до вичерпання усіх точок. Відновлюємо все безліч точок, вибираємо таксон з максимальним значенням d, поміщаємо сферу в його центр. Починаємо стискати сферу. На кожному кроці стиснення визначаємо число внутрішніх точок. Якщо початковий радіус був занадто великий, то зміна d буде повільним. У міру входження в більш щільні області, темп зміни буде збільшуватися, що і служить сигналом до зупинки стиснення. Фіксується число внутрішніх точок таксона $m'_j$. Процедура повторюється для всіх запомненних центрів таксонів. Вибирається k таксонів з найбільшою кількістю точок.

### 2.8 Алгоритм BIGFOR

Що робити, якщо масив з n точок дуже великий і не поміщається в оперативній пам'яті? При цьому витрати на поточечное читання координат з зовнішнього носія будуть неприйнятно великі.

1. Розбити вихідний масив на t = m / V подмасивов, V - число точок у подмасіва m. За допомогою FOREL-2 розділити кожен з подмасивов на $k'$ таксонів. Опис кожного j-го таксона містить координати його центру і кількість внутрішніх точок $m'_j$. В результаті отримаємо $q = tk'$ точок-центрів таксонів.

2. Знову використовуємо FOREL-2 для розбиття цих точок на k-таксонів. Тільки при розрахунку центрів тяжіння враховується вага (масу) $m'_j$.

3. Перерозподіляти точки між k таксонами.

### 2.9 Ієрархічна таксономія

Для ієрархічної таксономії за методом знизу-вгору використовується алгоритм BIGFOR, тільки з виключеною процедурою перерозподілу точок.

На першому кроці радіус встановлюється малим, що дає таксони нижнього рівня. На наступних кроках таксони укрупнюються, утворюючи вищележачі впусти ієрархії. Процес припиняється, коли в підсумковий таксон увійдуть всі точки вихідної множини.

Для кластеризації зверху-вниз використовується FOREL з послідовним зменшенням радіусу сфери.

1. Визначаємо мінімальний радіус гіперсфери R, що включає всі точки m. Ці точки становлять таксон верхнього рівня.

2. Зменшуючи від кроку до кроку радіус, визначаємо таксони i-х рівнів.

3. Процес завершується при числі таксонів нижнього рівня рівного m - по точці в таксоні.

### 2.10 Динамічна таксономія - алгоритм DINA

При залежності потужності кластеризованого множини від часу результат кластеризації може змінюватися з появою або зникненням точок.

Для таксономії об'єктів, що виникають по одному або невеликими групами застосовується алгоритм DINA.

Здається деякий радіус R. Перша з'явилася точка або група точок оголошується центром першого таксону. При появі нової точки проводиться перевірка, чи потрапляє точка всередину гіперсфери. Залежно від результату точка або включається до складу таксона, а центр гіперсфери зміщується в центр ваги внутрішніх точок, або нова точка оголошується центром нового таксона. Далі процес очевидний.

Можна стежити за тим, щоб таксони не переповнялися - містили по можливості однакову кількість точок. При переповненні таксон можна розбити на два з однаковим числом точок.

Перехід від опису вихідних об'єктів до опису таксонів еквівалентний переходу від даних до знань. Ієрархічна таксономія відображає структуру нашого знання про досліджуваному явище. Можна будувати ієрархії понять (зростаючі пірамідальні мережі) в процесі накопичення нових фактів. Можуть виникати таксони з надмірно великою кількістю об'єктів і тоді їх слід таксономіровать, що еквівалентно деталізації знань.

### 2.11 Таксономія з супер метою. Алгоритм ROST

Розглянуті вище алгоритми таксономії - це універсальні алгоритми, так би мовити, на всі випадки життя. Однак такий підхід у багатьох конкретних випадках може виявитися неоптимальним.

Наприклад, в задачах розпізнавання усного мовлення, зазвичай користуються не окремими фонемами, а їх групами (звукотипи). При цьому неприйнятно, якщо в одну групу потраплять дуже схожі за своїми спектральним характеристикам звуки т, к, п (ток, кіт, хто). Таксони на рівні звукотипи повинні будуватися з урахуванням супермету: крім того, що таксони повинні об'єднувати схожі елементи, кількість таксонів має бути мінімальним, але достатнім для прийняття рішень на більш високих рівнях.

Таким чином, ми чітко сформулювали мету таксономії і терміни типу самонавчання, навчання без учителя більш неприйнятні.

Таксономія з урахуванням супермету може бути отримана алгоритмом ROST.

Алгоритм ROST - це варіант ієрархічної таксономії методом знизу вгору. Спочатку ми застосовуємо FOREL з малим радіусом гіперсфери. Потім радіус збільшуємо і після кожного кроку робимо перевірку на відповідність супермету: чи не виникають помилки через укрупнення таксонів (наприклад, звукотипи). Якщо ні, то процес триває, в зворотному випадку, то точки таксона, що приводить до помилок піддаються повторній таксономії з меншим радіусом гіперсфери і з подальшого розгляду виключаються. Укрупнення продовжується до тих пір, поки з розгляду не будуть вилучені всі крапки.

### 2.12 Метод порівняння алгоритмів таксономії

Головне, що цікавить користувача - якість отриманих рішень. Для того щоб сформулювати критерій якості, згадаємо про те, що таксономія проводиться не тільки і не стільки для компактного перетворення безлічі m об'єктів в k таксонів. Надалі ці таксони або їх типові представники використовуються для короткого опису наявних об'єктів і, що більш важливо, для розпізнавання нових об'єктів генеральної сукупності. Кожен новий об'єкт відноситься до найбільш близькій таксону (образу).

Нехай M - потужність генеральної сукупності, m <M - потужність деякої вибірки з генеральної сукупності. На безлічі вибірки проведена таксономія деяким алгоритмом F. Якщо тепер пред'являти програмі класифікації решта M-m об'єктів і приєднувати їх до вже отриманих таксонам, отримаємо варіант таксономії генеральної сукупності S'. Якщо тим же алгоритмом F зробити таксономию об'єктів всієї генеральної сукупності, то отримаємо варіант S. Таксономії S і S' назвемо відповідно базової і вибірковою.

Якщо базова і вибіркова таксономії збігаються, то алгоритм F вдало вгадав структуру генеральної сукупності за випадковою вибіркою. Здатність по малих вибірках правильно вгадувати структурні закономірності генеральної сукупності і є основна характеристика якості (Q) таксономії.

Нехай p та q - об'єкти. Введемо поняття таксономічної відстані:

P = 1, якщо p, q належать різним таксонам;

P = 0, якщо p, q належать одному таксону.

Нехай у квадратній матриці розмірністю M×M стовпці і рядки відповідають об'єктам генеральної сукупності, а на перетині p-го рядка з q-м стовпцем знаходиться значення ρ (p, q). Матриця симетрична, а діагональні елементи рівні 0. Різниця R (S, S') між таксономії можна отримати, підсумувавши число елементів з незбіжними значеннями у двох матриць, що представляють результати S, S' таксономій. Розділивши отриману суму на максимально можливе число розбіжностей, яке дорівнює $M^2$-M, отримаємо відстань Хеммінга між матрицями, нормоване в діапазоні від 0 до 1:

$$R(S,S') = \sum_{p,q-1}^{M} \frac{\rho(p,q)-\rho'(p,q)}{M^2-M}. \quad (2.7)$$

Чим менше значення R (S, S'), тим краще алгоритм F вгадав структуру генеральної сукупності.

Для того щоб практично реалізувати перевірку якості, можна написати програму, яка генерувала б дані з заданим законом розподілу властивостей у просторі ознак, потужності генеральнихсукупностей і випадкових вибірок. Така програма (полігон Таксон) розроблена, і перевірка різних алгоритмів дала наступні результати.

По-перше, кращим алгоритмом виявився KRAB, на другому місці був SKAT і на третьому - FOREL. Але FOREL дає швидкі і прості рішення.

По-друге, з'ясувалося, що якість таксономії не залежить від розмірності простору ознак. Можна сказати, що по цій властивості машина значно перевершує людину, лише коли безпосередньо бачить поділюване множини. Інакше він переходить до поділу за кожною ознакою окремо.

### 2.13 Ваги ознак і перевірка інформативності ознак

У реальних задачах, коли ми не можемо точно визначити важливість тієї чи іншої признаковой координати, іншими словами, визначити розмірність простору ознак, слід зробити припущення висунути гіпотезу про важливість тієї чи іншої ознаки і формально врахувати це, привласнюючи ознаками деякі значення ваг. Тоді евклідова відстань між об'єктами в n-вимірному просторі ознак:

$$\rho(p,q) = \sum_{i-1}^{n} \sqrt{\gamma_i (x_{p,i} - x_{q,i})}. \quad (2.8)$$

В одній і тій же таблиці можуть зустрічатися властивості, виміряні в різних шкалах. При таксономії не використовуються всі властивості метричного простору. Потрібні лише все парні відстані між об'єктами. Для цього достатньо оцінити міру близькості або відмінності об'єктів по кожному з різнотипних властивостей окремо і потім знайти загальну міру їх близькості за всіма властивостями.

# 3 ТЕОРІЯ РОЗПЛИВЧАСТИХ МНОЖИН

## 3.1 Вступ до теорії розпливчастих множин

Нехай X = {x} - сукупність об'єктів (точок), що позначаються через x. Тоді розпливчасте множина A в X є сукупність упорядкованих пар A = {x, $\mu_A$, (x)}, x є A, де $\mu_A$(x) являє собою ступінь приналежності X до A, а $\mu_A$: → M – функція відображення X в простір M, називається простором приналежності. Коли M містить тільки дві точки 0 і 1, A є нераспливчатим множиною і його функція приналежності збігається з характеристичною функцією нераспливчатого множини.

Звичайно передбачається, що M є інтервал [0,1], причому 0 і 1 являють собою нижчу і вищу ступеня приналежності. У багатьох додатках функція приналежності повинна бути оцінена виходячи з частковою інформації (наприклад, значення, що приймаються нею на кінцевому множини опорних точок ($x_1$, ..., $x_N$). Коли A визначено в такий спосіб не повністю - і, отже, лише приблизно, ми можемо говорити, що воно частково визначено за допомогою поясняющего прикладу. Завдання оцінки $\mu_A$ за відомим множини пар ($x_1$, $\mu_A$, ($x_1$)), …, $x_N$, $\mu_N$, ($x_N$), тобто завдання абстрагування - завдання, що грає центральну роль в розпізнаванні образів.

## 3.2 Основні визначення

Нормальність - розмите множина A нормально тоді і тільки тоді, коли, $Sup_x\mu_A\ (x) = 1$, тобто супремум на X дорівнює 1.

Розпливчасте множина субнормально, якщо воно не є нормальним. Непорожнє субнормальне розпливчасте множина може бути нормалізовано поділом кожного $\mu_A$(x) на величину $Sup_x\mu_A\ (x)$. Розпливчасте множина порожньо тоді і тільки тоді, коли $\mu_A$(x) ≡ 0. Носій розпливчастого множини A є така множина S(A), що x є S(A) ↔ $\mu_A$(x) > 0, якщо $\mu_A$(x) = const на S(A), то A нераспливчато.

Відзначимо, що нераспливчатое множина може бути субнормальний. Дві розпливчасті множини рівні (A = B), тоді і тільки тоді, коли $\mu_A$(x) = $\mu_B$(x) для всіх x в X.

Розпливчасте множина A міститься в розпливчастому множині B, або є підмножиною B(A є B), тоді і тільки тоді, коли $\mu_A \leq \mu_B$. У цьому сенсі розпливчасте множина дуже великих чисел є підмножина розпливчастого множини великих чисел.

Кажуть, що A' є доповнення до A тоді і тільки тоді, коли $\mu_{A'} = 1-\mu_A$. Наприклад, розпливчасті множини A = {"Високі люди"} і A' = {"Невисокі люди"} є доповненням один до одного, якщо заперечення «НЕ» розуміється як операція, що замінює $\mu_A(x)$ на $1-\mu_A(x)$ для кожного x в X.

### 3.3 Операції над нечіткими множинами

Перетином A і B позначається $A \cap B$ і визначається як найбільша розпливчасте множина, що міститься як в A, так і в B. Функція приналежності $A \cap B$ визначається рівністю:

$$\mu_{A\cap B}(x) = \min\left(\mu_A(x), \mu_B(x)\right), x \in X, \quad (3.1)$$

де min(a, b) = a, якщо $a \leq b$, і min(a, b) = b, якщо a > b.

Якщо використовувати замість символу min знак кон'юнкції $\cap$ , можна переписати умову в більш простому вигляді:

$$\mu_{A\cap B} = \mu_A \cap \mu_B. \quad (3.2)$$

Поняття перетину має близьке відношення до поняття з'єднувального союзу «І». Так, якщо A - клас високих людей і B - клас повних людей то $A \cap B$ - клас людей, які одночасно високі і повні.

Слід зауважити, що «І» розуміється в «жорсткому» сенсі, тобто відсутня можливість будь-якої «компенсації» наявних значень $\mu_A(x)$ якими-небудь значеннями $\mu_B(x)$, і навпаки.

М'яка інтерпретація відповідає алгебраїчного добутку $\mu_A(x) * \mu_B(x)$.

Поняття об'єднання множин двояко поняттю перетину. Об'єднання A і B позначається $A \cup B$ і визначається, як найменше розпливчасте множина, що міститься як A, так і B. Функція приналежності $A \cup B$ визначається рівністю:

$$\mu_{A\cup B}(x) = max\left(\mu_A(x), \mu_B(x)\right), x \in X, \quad (3.3)$$

де max(a, b) = a, якщо a ≥ b, і max(a, b) = b, якщо a < b.

Якщо використовувати замість символу max знак диз'юнкції, можна переписати умову в більш простому вигляді:

$$\mu_{A\cup B} = \mu_A \cup \mu_B. \tag{3.4}$$

Операція об'єднання має близьке відношення до з'єднувального союзу «АБО». Так, якщо А - клас високих людей і В - клас повних людей то А ∪ В - {клас людей, які високі або повні}.

Можна розрізняти «АБО» в «жорсткому» сенсі, т.е $\mu_{A\cup B} = \mu_A \cup \mu_B$, від «АБО» в «м'якому» сенсі відповідного алгебраїчній сумі, що позначається як А ⊕ В.

$$\mu_{A\oplus B}(x) = \mu_A(x) + \mu_B(x) - \mu_A(x) * \mu_B(x). \tag{3.5}$$

Операція перетину та об'єднання пов'язані наступним тотожністю:

$$A \cup B = (A' \cap B')'. \tag{3.6}$$

Алгебраїчний добуток розпливчасті множин А і В позначається через АВ і визначається:

$$\mu_{AB}(x) = \mu_A(x) * \mu_B(x), x \in X. \tag{3.7}$$

Алгебраїчна сума розпливчасті множин А і В позначається через А ⊕ В і визначається рівністю:

$$\mu_{A\oplus B}(x) = \mu_A(x) + \mu_B(x) - \mu_A(x) * \mu_B(x). \tag{3.8}$$

Відзначимо, що операції ∪ і ∩ ассоціатіни і дистрибутивни по відношенню один до одного. Операції * (добутку) та ⊕ (суми) асоціативні, але не дистрибутивни. Операція * дистрибутивна по відношенню до об'єднання ∪ , але не навпаки. Взагалі, такою властивістю володіє будь-яка операція ***, монотонно спадна по кожному зі своїх аргументів. У символічній запису, якщо $b \geq b' \rightarrow a * b \geq a * b'$ і $a \geq a' \rightarrow a * b \geq a' * b$, то

більшість результатів залишається справедливо при заміні операції ∩ на операцію *, яка є асоціативної і дистрибутивної щодо операції ∪.

### 3.4 Властивості нечітких множин

α-розрізом нечіткого множини A ⊆ X, позначається як $A_\alpha$, називається наступне нечітке множина:

$$A_\alpha = \{x \in X;\ \mu_A(x) \geq \alpha\}, \tag{3.9}$$

то є множина, яке визначається наступною характеристичною функцією (функцією приналежності):

$$XA_\alpha = \begin{cases} 0, \mu_A(x) < \alpha, \\ 1, \mu_A(x) \geq \alpha. \end{cases} \tag{3.10}$$

Для α-розрізу нечіткого множини істинна імплікація:

$$\alpha_1 < \alpha_2 \Rightarrow A_{\alpha 1} \supset A_{\alpha 2}. \tag{3.11}$$

### 3.5 Основи розпливчастих множин

Нехай A - розпливчасте множина в просторі $X = R^n$. Тоді, A є опуклим розпливчастим множиною в тому і тільки тому випадку, якщо функція приналежності для кожної пари точок x, y з X задовольняє нерівності:

$$\mu_A(\lambda x + (1 - \lambda)y) \geq \min(\mu_A(x)\mu_A(yx)), \tag{3,12}$$

для всіх $0 \leq \lambda \leq 1$. Відповідно А є увігнутим, якщо його доповнення $A'$ опукло. Неважко показати, що якщо два розпливчастих множини А і В випуклі, їх перетин А ∩ В також опукло. З іншого боку, якщо А і В увігнуті, то увігнуті будуть і їх об'єднання А ∪ В.

Розпливчасте відношення на прямому добутку просторів:

$$X \times Y = \{(x, y), x \in X, y \in Y\}, \tag{3.13}$$

є розпливчасте множина в $X \times Y$, характеризуемое функцией принадлежности $\mu_R$, яка складає кожної впорядкованої парі (x, y) її ступінь приналежності $\mu_R(x, y)$ к R. У загальному випадку n-арное розпливчасте відношення на декартовому добутку:

$$X = X_1 \times X_2 \times \ldots \times X_n, \tag{3.14}$$

є розпливчасте множина в X, що описується залежної від n змінних функцій приналежності:

$$\mu_R(x_1, \ldots, x_n), x_i \in X_i, i = 1, \ldots, n. \tag{3.15}$$

Розпливчасті множини, породжувані відображеннями. Нехай f:X-›Y - відображення з X = {x} в Y = {y}, причому образ елемента x позначається через Y=f(x), і нехай - розпливчасте А множина в просторі X. Тоді відображення f породжує розпливчасте множина В в просторі Y з функцією приналежності, задається співвідношенням:

$$\mu_B(y) = \sup_{x \in f^{-1}(y)} \mu_A(x). \tag{3.16}$$

Причому супремум береться по всіх точках, що становить прообраз $f^{-1}(y)$ в X точки y.

Умовні розпливчасті множини. Розпливчасте множина B(x) у просторі Y={y} називається умовним по x, якщо функція приналежності залежить від змінної x як від параметра. Ця залежність виражається наступним чином:

$$\mu_{\mathrm{B}}(y|x). \tag{3.17}$$

Припустимо, що область зміни параметра x є простір X і при цьому кожному x з X відповідає розмите множина B(x) в Y. Таким чином, ми маємо справу з відображенням з X в простір розмитих множин в Y, характеризується функцією $\mu_{\mathrm{B}}(y|x)$. За допомогою цього відображення будь-який заданий розпливчасте множина A в X породжує розпливчасте множина B в Y, який визначається співвідношенням:

$$\mu_B(y) = \sup_x \min(\mu_A(x), \mu_{\mathrm{B}}(y|x)), \tag{3.18}$$

де $\mu_A$ і $\mu_{\mathrm{B}}$ - функції приналежності множин A і B відповідно. використовуючи операції ∪ і ∩, можна переписати умову:

$$\mu_B(y) = \cup_x \min(\mu_A(x) \cap \mu_{\mathrm{B}}(y|x)). \tag{3.19}$$

Розкладність. Нехай X = {x}, Y = {y} і нехай C - розпливчасте множина в просторі Z = $X \times Y$ з функцією приналежності $\mu_C(y, x)$. Тоді C називається разложимости в по X і Y в тому і тільки тому випадку, якщо C допускає представлення C = A ∩ B , або, що еквівалентно:

$$\mu_C(y, x) = \mu_A(x) \cap \mu_{\mathrm{B}}(y), \tag{3.20}$$

де A і B - розпливчасті множини з функціями приналежності $\mu_A(x)$ і $\mu_B(y)$ відповідно. (Таким чином, A і B циліндричні розпливчасті множини в Z.) Це визначення справедливо для розпливчастого множини, заданого в декартовом добутку будь-якого числа параметрів.

### 3.6 Розпливчасті цілі, обмеження та рішення

Головними елементами процесу прийняття рішення є:

– множина альтернатив;

– множина обмежень;

– функція перевагу, що ставить кожній альтернативі у відповідність виграш (або програш), який буде отриманий в результаті вибору цієї альтернативи.

Логічна схема ухвалення рішення в розпливчастих умовах інша. Її найважливішою рисою служить симетрія стосовно цілям і обмеженням. Ця симетрія усуває відмінності між цілями і обмеженнями і дозволяє досить просто сформувати на їх основі рішення.

Нехай X = {x} - заданий безліч альтернатив, Тоді розпливчаста мета G буде ототожнюватися з фіксованим розпливчастим множиною G в X.

Наприклад, якщо X = $R^1$ (дійсна пряма), а розпливчаста мета формується як x має бути значно більше 10 то її можна представити як розпливчасте множина в $R^1$ з функцією приналежності має, скажімо такий вигляд:

$$\mu_G(x) = \begin{cases} 0, \ x < 10, \\ (1 + (x - 10)^{-2})^{-1}, x \geq 10. \end{cases} \quad (3.21)$$

Аналогічно мета x повинно бути в околиці 15:

$$\mu_G(x) = (1 + (x - 15)^4)^{-1}, \quad (3.22)$$

Відзначимо, що обидва множини випуклі.

### 3.7 Функція приналежності розпливчатою мети

При звичайному підході функція приналежності, використовувана в процесі прийняття рішення, служить для встановлення лінійної впорядкованості на множині альтернатив. Очевидно, що функція приналежності розпливчатою мети виконує ту ж задачу (передбачається, що прийняті функцією значення утворюють лінійні впорядковані множини) і, звичайно, може бути отримана з функції приналежності за допомогою нормалізації, зберігає встановлену лінійну впорядкованість. По суті, така нормалізація призводить до спільного знаменника різні цілі і обмеження і дозволяє, таким чином, звертатися з ними однаковим чином:

– це є важливим аргументом на користь того, що в якості одного з основних компонентів в логічній схемі прийняття рішень в розпливчастих умовах поняття мети, а не функції перевагу;

– подібним же чином розпливчасте обмеження, або просто обмеження, C в просторі X визначається як деякий розпливчасте множина в X.

Наприклад, у випадку X = $R^1$ обмеження x знаходиться приблизно в діапазоні 2-10 може бути представлено розпливчастим множиною з функцією приналежності, скажімо вида:

$$\mu_C(x) = (1 + a(x - b)^m)^{-1}, \quad (3.23)$$

a - позитивне число, m - парне позитивне число. Важливо, що мета і обмеження розглядаються як розпливчасті множини в просторі альтернатив. Це дає можливість не робити між ними відмінності при формуванні рішення.

### 3.8 Розпливчасте рішення, оптимальне рішення

Розпливчасте рішення, слід визначити як розпливчасте множина в просторі альтернатив, що виходить в результаті перетину заданих цілей і обмежень.

Визначення. Нехай в просторі альтернатив X задані: розпливчаста мета G і розпливчасте обмеження C. Тоді розпливчасте множина D, утворене перетинанням G і C, називається рішенням. У символічній формі $D = G \cap C$ і відповідно:

$$\mu_D = \mu_G \cap \mu_C. \quad (3.24)$$

У загальному випадку:

$$D = G_1 \cap G_2 \cap \ldots \cap G_N \cap C_1 \cap C_2 \cap \ldots \cap C_M, \quad (3.25)$$

$$\mu_D = \mu_{G_1} \cap \mu_{G_2} \cap \ldots \cap \mu_{G_N} \cap \mu_{C_1} \cap \mu_{C_2} \cap \ldots \cap \mu_{C_M}. \quad (3.26)$$

Визначення рішення як перетину цілей і обмежень відповідає розумінню союзу «І» в «жорсткому» сенсі (і повні і високі).

Приймемо, що D - розпливчасте множина з функцією приналежності. Нехай K - множина точок в X, в яких досягає максимуму (якщо він існує). Тоді нераспливчатое, але, взагалі кажучи, субнормального підмножина $D^M$ з D, яке визначається умовами:

$$\mu_{D^M}(x) = \begin{cases} max\mu_D(x) \text{для } x \in K, \\ 0 \text{ для інших } x. \end{cases} \quad (3.27)$$

Буде називатися оптимальним рішенням, а для кожного x з носія $D^M$ - максимизирующим рішенням. Відзначимо, що в $R^n$ достатньою умовою єдиності максимизирующего рішення є сильна опуклість розпливчастого множини D, тобто опуклість D і наявність у нього унімодальної функції приналежності.

У тих випадках коли рішення D може бути виражене опуклою комбінацією цілей і обмежень з ваговими коефіцієнтами, що характеризують відносну важливість складових елементів:

$$\mu_D(x) = \sum_{i=1}^{n} \alpha_i(x)\, \mu_{G_i}(x) + \sum_{j=1}^{m} \beta_j(x) \mu_{G_j}(x), \quad (3.28)$$

де $\alpha_i$ і $\beta_j$ - функції приналежності такі, що

$$\sum_{i=1}^{n} \alpha_i(x) + \sum_{j=1}^{m} \beta_j(x) \equiv 1. \quad (3.29)$$

Відзначається, що формула нагадує відомий спосіб відомості векторного критерію до скалярному, за допомогою утворення лінійної комбінації компонент векторної функції мети.

### 3.9 Цілі і обмеження - розпливчасті множини в різних просторах

Раніше розглядався випадок, коли цілі і обмеження є розпливчатими множинами в просторі альтернатив X. Практичний інтерес представляє більш загальний випадок, коли цілі й обмеження - розпливчасті множини в різних просторах.

Нехай f - відображення X = {x} в Y = {y}, причому змінної x позначено вхідний вплив (причина), а змінною y - відповідає вихідна (слідство). Припустимо, що цілі задані як розпливчасті безлічі $G_1$, $G_2$, ..., $G_n$ в Y, в той час як ограніченія розпливчасті множини $C_1$, $C_2$, ..., $C_m$ в просторі X. Маючи розпливчасте множину $G_i$ в Y, можна знайти розпливчасте множину в X , яке індукує $G_i$ в Y. Функція приналежності $G_i^-$ задається рівністю:

$$\mu_{G_i^-} = \mu_{G_i}\big(f(x)\big), i = 1, \dots, n. \quad (3.30)$$

Після цього рішення D може бути виражене перетином $G_1^-$, $G_2^-$, …, $G_n^-$ і $C_1^-$, $C_2^-$, …, $C_m^-$ $\mu_D(x)$ в розгорнутому вигляді:

$$\mu_D(x) = \mu_{G_1}\big(f(x)\big) \cap \dots \cap \mu_{G_n}\big(f(x)\big) \cap \mu_{C_1}\big(f(x)\big) \cap \dots \cap \mu_{C_m}\big(f(x)\big), \quad (3.31)$$

де f:X-›Y, таким чином, випадок коли цілі і обмеження задаються як розпливчасті множини в різних просторах, може бути зведений до випадку, коли вони задаються в одному і тому ж просторі. Це співвідношення є досить корисним при аналізі багатокрокових процесів прийняття рішення.

### 3.10 Багатокрокові процеси прийняття рішень

Для простоти будемо припускати, що керуюча система А є інваріантної за часом детермінованою системою з кінцевим числом станів. Саме кожний стан $x_t$ в якому система А знаходиться в момент часу t, t = 0, 1, 2, ..., належить заданому кінцевому множини можливих станів X = $\{\delta_1, \dots, \delta_n\}$, при цьому має місце в момент t вхідний сигнал u(t) є елементом множини U = $\{\alpha_1, \dots, \alpha_m\}$.

Еволюція системи в часі описується рівнянням стану:

$$x_{t+1} = f(x_t u_t), t = 0, 1, 2, \dots, n, \quad (3.32)$$

в якому f - задає функція, що відображає $X \times U$ в X. Таким чином, $f(x_t u_t)$ являє собою подальший стан для $x_t$ при вхідному сигналі $u_t$. Якщо f є випадковою функцією, то А - стохастична система, стан якої в момент t + 1 характеризується розподілом ймовірності P ($x_{t+1} \mid x_t, u_t$) на X, умовним по $x_t$ і $u_t$.

Аналогічно якщо f - распливчатая функція, то А є розпливчастою системою, стан якої в момент t + 1 є умовне по $x_t$ і $u_t$ розпливчасте множина, що характеризується функцією приналежності $\mu_{(x_{t+1} \mid x_t, u_t)}$ (говорячи про розпливчастих умовах, ми маємо на увазі, що розпливчастими є цілі і (або) обмеження, однак це не означає, що керуюча система обов'язково є розпливчастою).

Далі функція f вважається нерасплівчатой, якщо це не обумовлено. Передбачається, що в кожен момент часу t на вхідну змінну накладено розпливчасте обмеження $C^t$, що є розпливчастим множиною U з функцією приналежності $\mu_t(u_t)$. Крім того вважається, що мета розпливчасте множина $G^N$ в X, визначається функцією приналежності $\mu_{G^N}(x_N)$, де N - час закінчення процесу. Ці пропозиції є спільними для більшості завдань.

# 4 РОЗРОБКА ТА ПРОГРАМНА РЕАЛІЗАЦІЯ МЕТОДУ КЛАСИФІКАЦІЇ БАГАТОВИМІРНИХ ОБ'ЄКТІВ

## 4.1 Вибір середовища розробки

У рамках даної дипломної роботи була розроблена програма для класифікації багатовимірних об'єктів. Для її реалізації було обрано середовище розробки C#.

C# – об'єктно-орієнтована мова програмування. Розроблена в компанії Microsoft як мова розробки додатків для платформи Microsoft. NET Framework і згодом був стандартизований як ECMA-334 і ISO / IEC 23270.

C# відноситься до мов з C-подібним синтаксисом, з них його синтаксис найбільш близький до C++ і Java. Мова має статичну типізацію, підтримує поліморфізм, перевантаження операторів, делегати, атрибути, події, властивості, узагальнені типи і методи, ітератори, анонімні функції з підтримкою замикань, LINQ, виключення, коментарі у форматі XML.

Переваги C#:

– C# створювалася паралельно з каркасом Framework.Net і повністю враховує всі можливості як FCL, так і CLR;

– C# цілком об'єктно-орієнтована мова з можливостями наступності, де навіть типи, вбудовані в мову, представлені класами;

– C# спадкоємець мов C / C ++, він зберігає кращі риси цих популярних мов програмування, з ними у нього загальний синтаксис, а знайомі оператори мови полегшують перехід програмістів від C ++ до C #.

## 4.2 Розробка алгоритму класифікації

1) Беремо стовпець з найбільшою дисперсією рівномірно розподіленим по всіх класах. Цей вибір обумовлений обраною стратегією користовича.

2) Всі об'єкти розбиваємо на $k^2$ класів.

3) Шукаємо функцію приналежності в кожному класі (беремо за значенням n розбиваємо на k класи).

4) Переразподіл об'єктів класів згідно значень функції приналежності.

5) Перевірка, перерахунок функції приналежності якщо вона буде близька до 1 то, тоді зупиняємо перерахунок функції прінадлженості.

6) Формування пріедікатов результату для кожного класу.

Таке рішення полягає у виборі деякої альтернативи з при нечітко поставленній мети - розбиття А на k підмножин, ототожнюється з фіксованим розмитим множиною $A_k$ в X, з урахуванням обмежень С в X при яких виконується еквіваленція виду:

$$A_i \leftrightarrow P_i\left(a_1^i(x_1), a_2^i(x_2), \dots, a_n^i(x_n)\right) \forall\, x \in A_i, \tag{4.1}$$

і задовольняє умовам:

$$\forall i P_i\left(a_1^i(x_1), \dots, a_n^i(x_n)\right) \wedge\ \forall j P_j\left(a_1^j(x_1), \dots, a_n^j(x_n)\right) = \\ = 0, .. i \neq j, \bigvee_{i=1}^{k} A_i = A. \tag{4.2}$$

Зміни функцій приналежності характерних параметрів (вказаний крок можна розглядати як коригування розбиття $X_j$ на k розмитих множин) таким чином, щоб:

$$inf\,\mu_{A'_{j1}}(x'_{j1}) < sup\mu_{A'_{j1}}(x'_{j1}) = inf\,\mu_{A'_{j2}}(x'_{j2}) < sup\mu_{A'_{j2}}(x'_{j2}) = \cdots \\ < sup\mu_{A'_{jk-1}}(x'_{jk-1}) = inf\,\mu_{A'_{jk}}(x'_{jk}) < sup\mu_{A'_{jk}}(x'_{jk}), \tag{4.3}$$

де

$$\mu_{A'_{ji}}(x'_{ji}) = \{inf A'_{ji}, sup(A'_{ji}/A'_{ji+1})\}. \tag{4.4}$$

Встановлюється відповідність між натуральними числами, з одного боку, та елементами підмножин, з іншого боку:

$$A'_{1i} \leftrightarrow n_1, A'_{2i} \leftrightarrow n_2, \dots, A'_{nn} \leftrightarrow n_n. \tag{4.5}$$

Вибір опорного вирішального правила розмитого множини, кількість елементів якого для кожної з підматриць визначається згідно виразу:

$$A'_{ji} \leftrightarrow n_i = min\{n_1, n_2, \dots, n_n\}. \quad (4.6)$$

Вираз написане для функцій приналежності з урахуванням (4.6) і (4.4), може бути представлено в наступному вигляді:

$$\mu_{A_{ji}}(x_{ji}) = \min\{\mu_{A'_{1i}}(x'_{1i}), \mu_{A'_{2i}}(x'_{2i}), \dots, \mu_{A'_{ni}}(x'_{ni})\}. \quad (4.7)$$

Таким чином, в результаті описаних дій проводиться визначення попереднього вирішального правила для кожної з підматриць.

### 4.3 Ілюстрація роботи програми

Розроблена для дослідження методу класифікації багатовимірних об'єктів програма реалізує наступні функції:

1) створює матрицю властивостей і пропонує ввести бажану кількість класів;
2) виводить матрицю властивостей;
3) розбиває її на передбачувані класи;
4) відображення результату розбиття матриці на класи.

Інтерфейс розробленої програми представлений на рисунку 4.1.

На формі програми ми бачимо поля для введення кількості даних, параметрів і кількості класів для створення матриці властивостей.

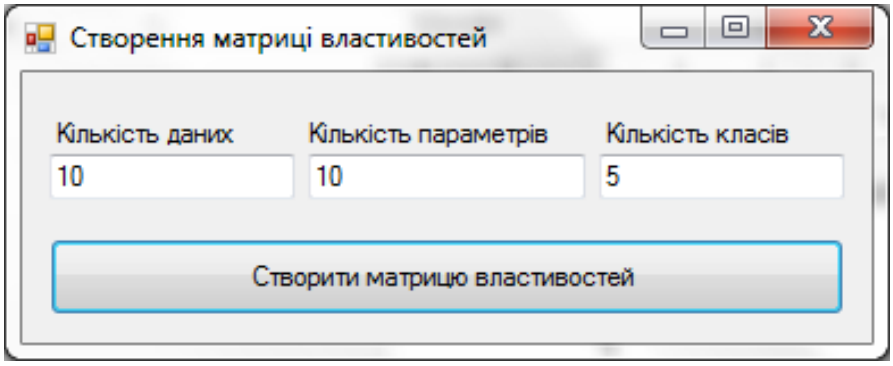



Рисунок 4.1 – Створення матриці властивостей

Після створення матриці властивостей, як показано на рисунку 4.2 ми повинні тепер її розбити на класи і побачити отриманий результат програми 4.3.

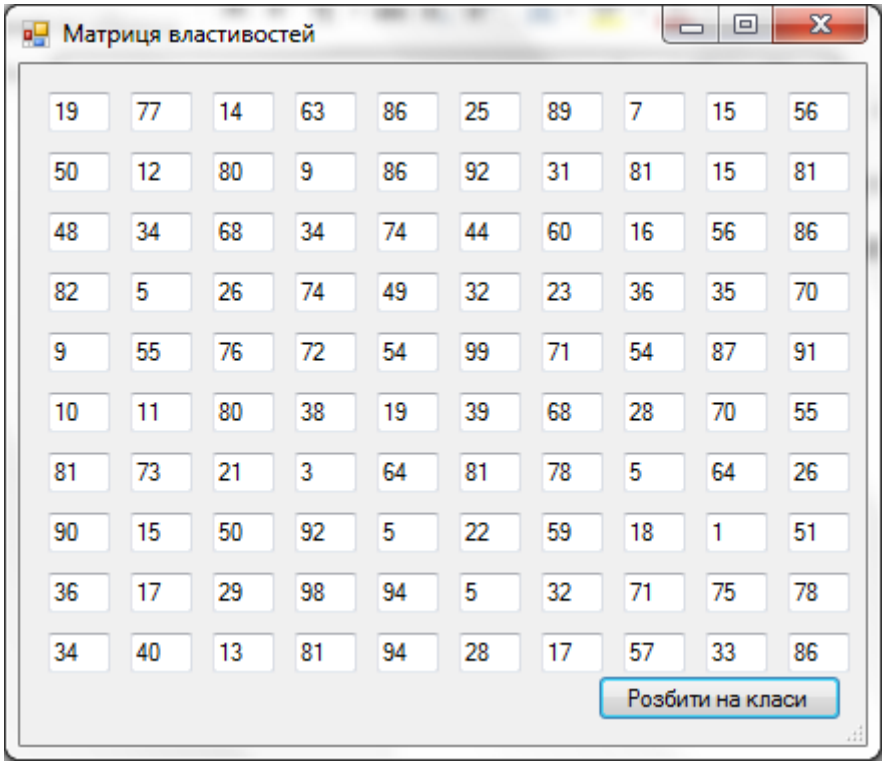



Рисунок 4.2 – Матриця властивостей

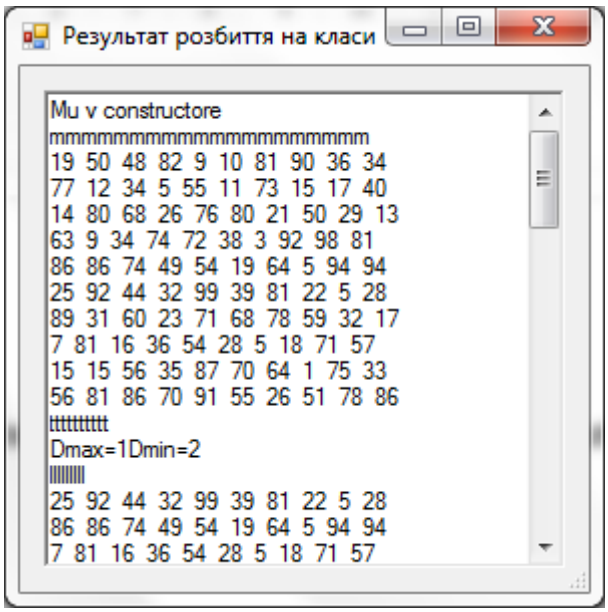



Рисунок 4.3 – Результат розбиття на класи

### 4.4 Аналіз результатів класифікації

Як було показано на рисунках 4.1 і 4.2 ми створили матрицю властивостей 10x10 і ввели розбиття на 5 класів, але програма порахувала, що 5 класів мало і за заданою їй логіці поділила їх на 7. Ми спостерігаємо за тим, що наш алгоритм класифікації добре справляється з поставленим завданням і показує досить не погані результати.

# 5 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ

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

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

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

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

Відповідно до [15] до легкої фізичної роботи відносяться всі види діяльності, виконувані сидячи і ті, що не потребують фізичної напруги. Робота користувача ПК відноситься до категорії 1а.

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

Джерелами підвищеної небезпеки можуть служити наступні елементи:

- розподільний щит;
- джерела живлення;
- блоки ПЕОМ і друку, що знаходяться в ремонті.

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

Відповідно до [16], при обслуговуванні ПЕОМ мають місце фізичні і психофізичні небезпечні, а також шкідливі виробничі чинники:

– підвищене значення напруги в електричному ланцюзі, замикання якої може відбутися через тіло людини;
– підвищений рівень статичної електрики;
– підвищений рівень електромагнітних випромінювань;
– підвищена або знижена температура повітря робочої зони;
– підвищений або знижений рух повітря;
– підвищена або знижена вологість повітря;
– відсутність або недостатність природного світла;
– підвищена пульсація світлового потоку;
– недостатня освітленість робочого місця;
– підвищений рівень шуму на робочому місці;
– розумове перенапруження;
– емоційні навантаження;
– монотонність праці.

### 5.2 Заходи щодо техніки безпеки

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

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

Тяжкість поразки людини електричним струмом залежить від цілого ряду чинників:

– значення сили струму;
– електричного опору тіла людини і тривалості протікання через нього струму;
– роду і частоти струму;
– індивідуальних властивостей людини і навколишнього середовища.

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

- заземлення електроустановок;
- занулення;
- захисне відключення;
- електричне розділення мережі;
- використовування малої напруги;
- ізоляція частин, що проводять струм;
- огорожа електроустановок.

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

Струм однофазного короткого замикання визначається по наближеній формулі:

$$I_K = \frac{\text{Uф}}{\text{Zп} + \frac{\text{Zт}}{3}}\ , \qquad (5.1)$$

де Uф - номінальна фазна напруга мережі, В;

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

Zт - повний опір струму короткого замикання на корпус, Ом.

Згідно таблиці 4 [17]: Zт /3 = 0,1 Ом.

Для провідників і жил кабелю для розрахунку повного опору петлі використовуємо формулу(5.2.) :

$$\text{Zп} = \sqrt{\text{Rп}^2 + \text{Xп}^2}\ , \qquad (5.2)$$

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

Xп - індуктивний опір паяння дротів, Ом.

Перетин 1 км мідного дроту S = 2.5 мм, тоді згідно таблицям 5 і 6 [17], має такий опір:

Xп = 0,11 Ом;

Rф = 7,55 Ом;

Ro = 7,55 Ом.

Отже, Rп = 7,55 + 7,55 = 15,1 Ом.

Тоді по формулі (5.2) знаходимо повний опір петлі :

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

Струм однофазного короткого замикання рівний:

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

Дія плавкої вставки на ПЕОМ забезпечується, якщо виконується співвідношення:

$$Iк \geq k * Iн, \quad (5.3)$$

де Iн - номінальний струм спрацьовування плавкої вставки, А;

k - коефіцієнт кратності нелінійного струму Iн, А.

Коефіцієнт кратності нелінійного струму Iн розраховується по формулі (5.4.) :

$$Iн = P / U, \quad (5.4)$$

де P = 220 Вт - споживана потужність;

U = 220 В - робоча напруга;

k = 3 А - для плавких вставок.

Отже, Iн = 220 / 220 = 1 А.

Підставивши значення у вираз (5.3), одержимо:

$$14,47 > 3*1.$$

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

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

Вимоги до виробничих приміщень встановлюються [25], ДБН, відповідними ГОСТами і ОСТамі з урахуванням небезпечних і шкідливих чинників, що утворюються в процесі експлуатації електроустаткування.

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

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

Відповідно до [15] встановлюють оптимальну і допустиму температуру, відносну вологість і швидкість руху повітря в робочій зоні . За відсутності надмірного тепла, вологи, шкідливих речовин в приміщенні досить природної вентиляції.

У приміщенні для виконання робіт операторського типу(категорія 1а), пов'язаних з нервово-емоційною напругою, проектом передбачається дотримання наступних нормованих величин параметрів мікроклімату (табл.5.1).

Таблиця 5.1 - Санітарні норми мікроклімату робочої зони приміщень для робіт категорії 1а.

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

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

Шум на виробництві має шкідливу дію на організм людини. Стомлення операторів через шум збільшує число помилок при роботі, призводить до виникнення травм. Для

оператора ПЕОМ джерелом шуму є робота принтера. Щоб усунути це джерело шуму, використовують наступні методи. При покупці принтера слід вибирати найбільш шумозахисні матричні принтери або з великою швидкістю роботи(струменеві, лазерні). Рекомендується принтер поміщати в найбільш віддалене місце від персоналу, або застосувати звукоізоляцію та звукопоглинання(під принтер підкладають демпфуючі підкладки з пористих звукопоглинальних матеріалів з листів тонкої повсті, поролону, пеноплену).

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

Обмеження годині роботи на ПЕОМ складає 3,5-4,5 години. Захист відстанню передбачає розміщення монітора на відстані 0,4-0,5 м від оператора. Передбачений монітор 20" TFT, Samsung 2043BW відповідає вимогам стандарту TCO'03.

TCO'03 пред’являє жорсткі вимоги в таких областях: ергономіка (фізична, візуальна і зручність користування), енергія, випромінювання (електричних і магнітних полів), навколишнє середовище і екологія, а також пожежна та електрична безпека, які відповідають всім вимогам [18].

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

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

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

При експлуатації ПЕОМ виробляється зорова робота. Відповідно до [22] ця робота відноситься до розряду 5а. При цьому нормоване освітлення на робочому місці(Ен) при загальному освітленні рівна 200 лк.

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

Виконаємо розрахунок кількості світильників в робочому приміщенні завдовжки a=12 м, шириною b=10 м, заввишки з=4 м, використовуючи формулу (5.5) розрахунку штучного освітлення при горизонтальній робочій поверхні методом світлового потоку:

$$n = (E \cdot S \cdot Z \cdot k)/(F \cdot U \cdot M), \quad (5.5)$$

де F - світловий потік = 3120 лм;

E - максимально допустима освітленість робочих поверхонь = 200 лк;

S - площа підлоги = 120 м2;

Z - поправочний коефіцієнт світильника = 1,2;

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

n - кількість світильників;

U - коефіцієнт використовування освітлювальної установки = 0,6;

M - кількість ламп у світильнику =2.

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

Виходячи з цього, рекомендується використовувати 12 світильників. Світильники слід розміщувати рядами, бажано паралельно стіні з вікнами. Схема розташування світильників зображена на рис. 5.1.

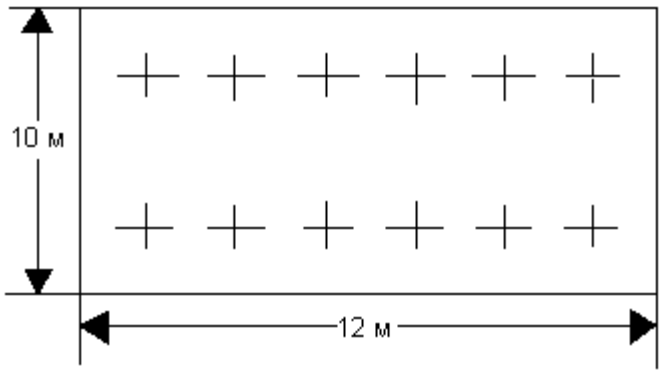



Рисунок 5.1 - Схема розташування світильників

### 5.4 Рекомендації по пожежній безпеці

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

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

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

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

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

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

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

Згідно [24] приміщення відносяться до категорії В (пожежовибухонебезпечним) і згідно правилам побудови електроустановок простір усередині приміщення відноситься до вогненебезпечної зони класу П - IIа (зони, розташовані в приміщеннях, в яких зберігаються тверді горючі речовини).

Потенційними джерелами запалення при роботі ПЕОМ є:

- іскри при замиканні і розмиканні ланцюгів;
- іскри і дуги коротких замикань;
- перегріви від тривалого перевантаження і наявності перехідного опору.

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

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

Пожежна безпека об’єктів народного господарства регламентується [19] і забезпечується системами запобігання пожежам і протипожежному захисту. Для успішного гасіння пожеж вирішальне значення має швидке виявлення пожежі і своєчасний виклик пожежних підрозділів до місця пожежі.

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

– застосування устаткування, що задовольняє вимогам електростатичної безпеки;

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

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

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

– заміна небезпечних технологічних операцій більш безпечними;

– ізольоване розташування небезпечних технологічних установок і устаткування;

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

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

– механізація, автоматизація та справність(потокова) виробництва;

– суворе дотримання стандартів і точне виконання встановленого технологічного режиму;

– запобігання можливості появи в небезпечних місцях джерел запалення;

– запобігання розповсюдженню пожеж і вибухів;

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

– виконання вимог сумісного зберігання речовин і матеріалів;

– наявність громовідводу;

– ліквідація можливості самозаймання речовин і матеріалів .

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

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

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

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

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

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

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

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

### 5.5 Вплив на навколишнє середовища

В даний час зростає кількість комп'ютерної техніки в усіх галузях діяльності людини. Багато користувачів і виробників помиляються, вважаючи, що зі зменшенням і удосконаленням комп'ютерів, зменшиться їх негативний вплив на навколишнє середовище.

На даний момент найбільш суворим з існуючих світових стандартів екологічності для комп'ютерної техніки є стандарт TCO-99. У порівнянні з попередніми він містить додаткові обмеження по частині екології, ергономіки, енергоспоживання і емісії пристроїв.

Організація по захисту навколишнього середовища Greenpeace з 2006 року оцінює виробників електроніки за кількістю важких металів і отруйних речовин, наприклад інгібіторів горіння, використовуваних ними при виробництві (інгібітор - речовина, присутність якого в невеликих кількостях призводить до запобігання або уповільнення процесів горіння або корозії; інгібітори знижують швидкість хімічних реакцій або пригнічують їх). Однак навіть оцінки такої організації, як Greenpeace, не можуть претендувати на об'єктивність. Адже в одних випадках вона використовує перевірену інформацію, що стосується, наприклад, заходів щодо утилізації відходів, а в інших спирається тільки на дані виробника. А якщо компанія не повідомляє ніяких відомостей, то автоматично опиняється на нижніх рядках рейтингу. Крім того, енергетичні витрати на виробництво і перевезення продукції також необхідно враховувати при оцінці екологічної ефективності. Адже часи, коли техніка виготовлялася тільки на одному заводі, давно пройшли. Сьогодні окремі комплектуючі закуповуються на різних підприємствах по всьому світу, після чого здійснюється складання пристроїв. Тому найчастіше навіть самі компанії не можуть знати, які шкідливі речовини потрапляють в атмосферу при виготовленні їх продукції і які саме метали або токсини в ній містяться.

ЖК-екрани - один з джерел парникових газів, які набагато шкідливіше діоксиду вуглецю. Рідкокристалічні монітори швидко знайшли популярність, прийшовши на зміну громіздким ЕПТ-моделям. І це не дивно, адже вони мають тонкі корпуса і споживають значно менше електроенергії. За іншим аспектам екологічної безпеки дисплеї на основі рідких кристалів також вважалися проривом, тому що в них не використовувався газ, що містить свинець. Досить довго ніхто не звертав уваги на застосовуваний для чищення РК-панелей трехфторістий азот (NF3), і тільки в середині 2008 року вченими було доведено наявність даної хімічної речовини в атмосфері. Відкриття було вражаючим: порівняно з діоксидом вуглецю (CO2) NF3 має в 17 000 разів більше активного парникового газу, а його атмосферний час напіврозпаду може складати від 550 до 740 світлових років (у CO2 - від 30 до 40 років). Закону, який обмежував би рівень викиду NF3, поки не існує.

Виявлення енерговитрат є таким же проблематичним процесом, як і визначення кількості матеріалів, придатних для вторинної переробки, і важких металів, що містяться в пристроях. Таким чином, надійним показником екологічності залишається тільки рівень енергоспоживання.

Полівінілхлорид, що позначається зазвичай абревіатурою ПВХ, - це різновид пластику, що застосовується в самих різних цілях. З нього зроблена зовнішня оболонка кабелів, якими з'єднуються пристрої, він оточує електричний провід портативного комп'ютера. Це дешевий, міцний і вельми поширений матеріал. Разом з тим, за словами IT- аналітика «Грінпіс» Кейсі Харрелл, «ПВХ - найгірший з пластиків». Він є причиною виникнення гормонального дисбалансу, проблем в репродуктивної сфері та різних форм раку. Полівінілхлорид практично неможливо правильно утилізувати. Внаслідок старий матеріал виявляється зазвичай на звалищі з відходами або, того гірше, спалюється з метою вилучення мідних жил і інших цінних компонентів. При його згорянні утворюється вкрай шкідливий канцерогенний діоксин. Звалища і хімічні поховання забруднюють джерела води. Єдиний спосіб правильно утилізувати ПВХ полягає в тому, щоб відправити його в центр небезпечних відходів.

Залишається лише сподіватися, що настане час, коли технології будуть допомагати людині, не завдаючи незворотної шкоди здоров'ю навколишнього середовища.

# ВИСНОВКИ

У результаті дипломної роботи був розроблений і програмно реалізований алгоритм класифікації. Для цього були вирішені наступні завдання:

– було досліджено алгоритми класу FOREL, SKAT, NTPP і інші алгоритми класифікації;

– була досліджена теорія розмитих множин, та визначені можливості її застосування в рішенні задач класифікації;

– розроблено та програмно реалізовано алгоритм класифікації багатовимірних об'єктів;

– був проведений аналіз алгоритму класифікації.

Практичні дослідження роботи показали доцільність алгоритму класифікації. Проте залишаються невирішеним ряд проблем. По-перше, це відсутність достатньої кількості верифікованих даних, край необхідних при налагодженні алгоритму та його подальшого удосконалення. По-друге, це давня проблема відображення отриманих класів в багатовимірному просторі. Іншими словами робота з великою кількістю даних дуже важка робота і вимагає складних обчислювальних операцій а так само великої кількісті ресурсів для обчислення алгоритмів класифікації багатовимірних об'єктів.

Реалізувати алгоритм який буде працювати з великою кількістю параметрів поки не вдається через складність, зумовлену істотним зростанням при зростанні кількості параметрів.

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

# ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1) Rad, A. Fast Circle Detection Using Gradient Pair Vectors [Text] / A. A. Rad, K. Faez, N. Qaragozlou. – Sydney, Australia: Proc. VIIth Digital Image Computing, 2003. – pp. 879-887.

2) Дюран, Б. Кластерный анализ [Текст]: учеб. / Б. Дюран, П. Оделл. – М.: Статистика, 1977. – 128 с.

3) Айвазян, С. А. Классификация многомерных наблюдений [Текст]: учеб. пособие / С. А. Айвазян, З. И. Бежаева, О. В. Староверов. – М.: Статистика, 1974. – 240 с.

4) Андерсон, Т. Введение в многомерный статистический анализ [Текст] / Т. Андерсон. – М.: Мир, 1963. – 499 с.

5) Загоруйко, Н. Г. Прикладные методы анализа данных и знаний [Текст] / Н. Г. Загоруйко. – Новосибирск: ИМ СО РАН, 1999. – 270 с.

6) Дуда, Р. Распознавание образов и анализ сцен [Текст] / Р. Дуда, П. Харт; перевод с англ. – М.: Мир, 1976. – 511 с.

7) Прэтт, У. К. Цифровая обработка изображений [Текст]: в 2 т. Т. 1 / У.К. Прэтт; перевод с англ. – М.: Мир, 1982. – 312 с.

8) Прэтт, У. К. Цифровая обработка изображений [Текст]: в 2 т. Т. 2 / У.К. Прэтт; перевод с англ. – М.: Мир, 1982. – 480 с.

9) Путятин, Е. П. Обработка изображений в робототехнике [Текст] / Е. П. Путятин, С. И. Аверин. – М.: Машиностроение, 1990. – 320 с.

10) Шилдт, Г. Язык программирования C#. Часть 1 [Текст] / Г. Шилдт; перевод с англ. – К.: ДиаСофт, 2006. – 264 с.

11) Шилдт, Г. Язык программирования C#. Часть 2 [Текст] / Г. Шилдт; перевод с англ. – К.: ДиаСофт, 2006. – 263 с.

12) Фор, А. Восприятие и распознавание образов [Текст] / А. Фор; перевод с фр. – М.: Машиностроение, 1989. – 271 с.

13) Форсайт, Д. А. Компьютерное зрение. Современный подход. [Текст] / Д. А. Форсайт, Д. Понс; перевод с англ. – К.: Вильямс, 2004. – 928 с.

14) Шапиро, Л. Компьютерное зрение [Текст] / Л. Шапиро, Дж. Стокман; перевод с англ. – М.: БИНОМ. Лаборатория знаний, 2006. – 752 с.

15) ГОСТ 12.1.005-88. Міждержавний стандарт. Система стандартів безпеки праці. Загальні санітарно-гігієнічні вимоги до повітря робочої зони

16) ГОСТ 12.0.003-74 Небезпечні і шкідливі виробничі фактори. Класифікація

17) ДСТУ 7237:2011 Національний стандарт України. Система стандартів безпеки праці. Електробезпека. Загальні вимоги та номенклатура видів захисту

18) ДСанПіН 3.3.2.007-98. Державні санітарні правила і норми. Гігієнічні вимоги до організації роботи з візуальними дисплейними терміналами електронно-обчислювальних машин.

19) ГОСТ 12.1.004-91. Пожежна безпека. Загальні вимоги .

20) ДБН В.2.5-67. Опалення вентиляція та кондиціонування.

21) ГОСТ 12.1.006-84. Електромагнітні поля радіочастот. Допустимі рівні на робочих місцях і вимоги до проведення контролю

22) ДБН В.2.5-28-2006. Природне і штучне освітлення.

23) ГОСТ 12.4.009-83. Пожежна техніка для захисту об'єктів. Основні види. Розміщення і обслуговування.

24) ДСТУ Б В.1.1-36-2016. Визначення категорії приміщень, будинків та зовнішніх установок за вибухопожежною та пожежною безпекою.

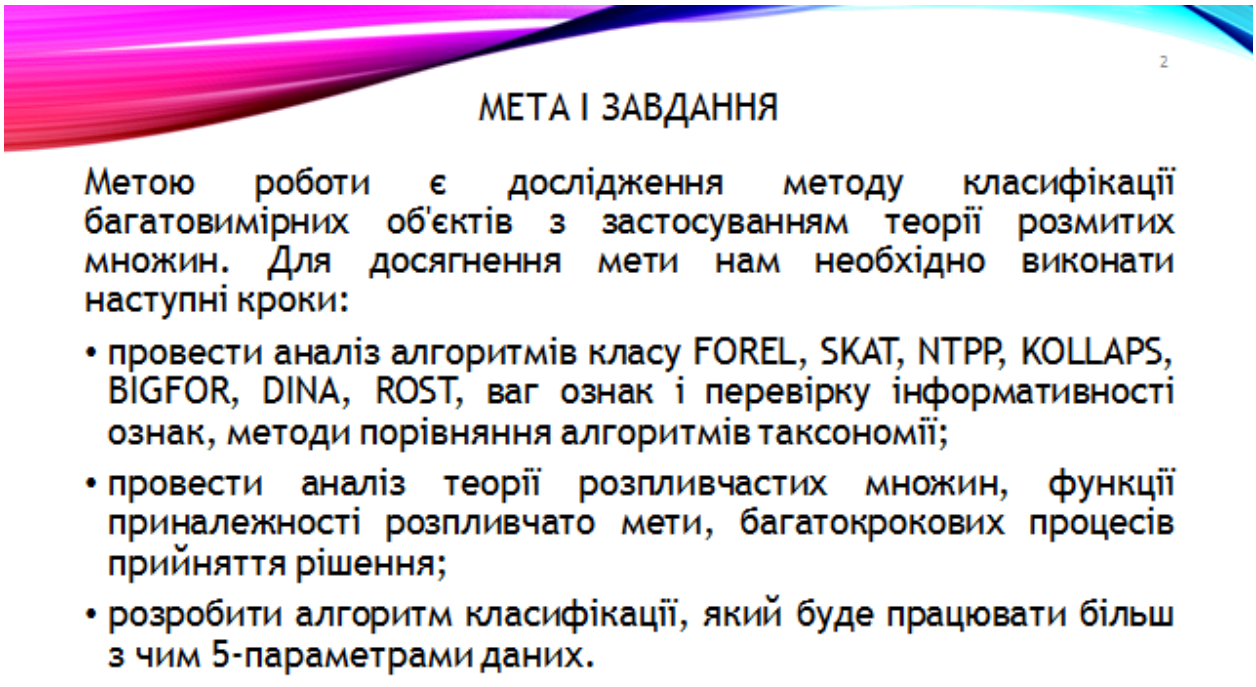
25) ДСП 173-96. Державні санітарні правила планування та забудови населених пунктів

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

27) Симметрон. Электронные компоненты. Каталог 2004, 2004г. – 192с.

# ДОДАТОК А

## Електронні плакати

## КЛАСИФІКАЦІЯ. ГІПОТЕЗА КОМПАКТНОСТІ

Під класифікацією розуміється система угруповання множини об'єктів, складена на основі врахування загальних ознак цих об'єктів і закономірних зв'язків між ними. Метою класифікації є утворення груп схожих між собою об'єктів, які прийнято називати класами або кластерами.

Методи класифікації, можна розділити на кілька груп. За способом завдання показника якості класифікації методи поділяються на евристичні та оптимізаційні. За способом об'єднання - на дівізімние, агломеративні і ітеративні. Евристичні алгоритми засновані на досвіді та інтуїції людини.

## КЛАСИФІКАЦІЯ. ГІПОТЕЗА КОМПАКТНОСТІ

Гіпотеза компактності дає на практиці гарні результати класифікації, якщо є достатня відповідність між змістом виділених ознак і побудованим геометричним простором.

Узагальненням гіпотези компактності є гіпотеза простої геометричної структури. Вона полягає в наступному: подібним в змістовному сенсі об'єктам класифікації відповідає проста структура в геометричному просторі ознак: схильність вздовж прямої, на колі, у сфері, по спіралі, на решітці.

У багатьох випадках, коли умови розв'язуваної задачі задають певну структуру об'єктів, застосування цієї гіпотези і відповідних алгоритмів класифікації призводять до гарних практичних результатів і добре узгоджуються з уявленнями людини про отримувані класах образів.

## КРИТЕРІЇ ПРИРОДНОСТІ КЛАСИФІКАЦІЇ

Якщо класи реальні, природні, існують насправді, чітко відокремлені один від одного, то будь-який алгоритм кластерного аналізу їх виділить. Отже, як критерій природності класифікації слід розглядати стійкість щодо вибору алгоритму кластерного аналізу.

Перевірити стійкість можна, застосувавши до даних кілька різних підходів. Якщо отримані результати змістовно близькі, то вони адекватні дійсності. В іншому випадку слід припустити, що природною класифікації не існує, завдання кластерного аналізу не має рішення, і можна проводити тільки угруповання.

## КРИТЕРІЇ ПРИРОДНОСТІ КЛАСИФІКАЦІЇ

Оскільки значення ознак завжди вимірюються з похибками, то "реальне" розбиття має бути стійке (тобто не змінюватися або змінюватися слабо) при малих відхиленнях вихідних даних.

Алгоритмів класифікації існує нескінченно багато, і "реальне" розбиття повинно бути стійке по відношенню до переходу до іншого алгоритму. Іншими словами, якщо "реальне" розбиття на діагностичні класи можливо, то воно знаходиться за допомогою будь-якого алгоритму автоматичної класифікація.

Отже, критерієм природності класифікації може служити збіг результатів роботи двох досить розрізняються алгоритмів, наприклад "найближчого сусіда" і "далекого сусіда".

## ПРОБЛЕМА ПОШУКУ ПРИРОДНОЇ КЛАСИФІКАЦІЇ

Існують різні точки зору на цю проблему. На Всесоюзній школі-семінарі «Використання математичних методів в задачах класифікації» (м Пущино, 1986 р), зокрема, були висловлені думки, що природна класифікація:

- закон природи;
- заснована на глибоких закономірності, тоді як штучна класифікація - на неглибоких;
- для конкретного індивіда та, яка найбільш швидко випливає з його тезауруса;
- задовольняє багатьом цілям; мета штучної класифікації задає чоловік;
- класифікація з точки зору споживача продукції;
- класифікація, що дозволяє робити прогнози;
- володіє критерієм стійкості.

## ПРОБЛЕМА ПОШУКУ ПРИРОДНОЇ КЛАСИФІКАЦІЇ

Можна виділити два критерії природності, з приводу яких є відносне згода:

- природна класифікація повинна бути реальною, відповідної дійсного світу, позбавленої внесеного дослідником суб'єктивізму;
- природна класифікація повинна бути важливою або з наукової точки зору (давати можливість прогнозу, передбачення нових властивостей, стиснення інформації), або з практичної.

## МАТРИЦЯ ДАНИХ

Багато об'єктів дослідження характеризуються множиною параметрів, і за результатами спостереження за їх функціонуванням формуються багатовимірні сукупності (матриці). Тоді матриця даних має вигляд:

$$\mathbf{X}=\begin{vmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ . & . & . & . \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{vmatrix}$$

## МАТРИЦЯ ДАНИХ

Рядки такої матриці відповідають результатам реєстрації всіх спостережуваних параметрів об'єкта в одному експерименті, а стовпці містять результати спостережень за одним параметром (фактором, варіант) у всіх експериментах. Позначимо кількість параметрів через m (m > 1), а кількість спостережень - через n.

У матриці елемент $x_{ij}$ відповідає значенню j-й варіанти в і-му спостереженні. Матриця, взагалі кажучи, може містити порожні значення деяких елементів, наприклад, через пропуски в реєстрації значень параметрів. У багатовимірному аналізі бажано усунути пропущені значення.

Для цього існують спеціальні прийоми, зокрема, викреслювання відповідних рядків матриці або занесення середніх значень замість відсутніх. Надалі будемо вважати, що матриця не містить порожніх елементів, а параметри об'єкта характеризуються безперервними випадковими величинами.

11

## РОЗРОБКА АЛГОРИТМУ КЛАСИФІКАЦІЇ

Беремо стовпець з найбільшою дисперсією рівномірно розподіленим по всіх класах. Цей вибір обумовлений обраною стратегією користувача

Всі об'єкти розбиваємо на $k^2$ класів

Шукаємо функцію приналежності в кожному класі (беремо за значенням n розбиваємо на k класи)

12

## РОЗРОБКА АЛГОРИТМУ КЛАСИФІКАЦІЇ

Перерозподіл об'єктів класів згідно значень функції прінадлежності

Перевірка, перерахунок функції прінадлежності якщо вона буде близька до 1 то, тоді зупиняємо перерахунок функції прінадлежності

Формування предикатов результату для кожного класу

## РОЗРОБКА АЛГОРИТМУ КЛАСИФІКАЦІЇ

Таке рішення полягає у виборі деякої альтернативи з при нечітко поставленні мети - розбиття A на k підмножин, ототожнюється з фіксованим розмитим множиною $A_k$ в X, з урахуванням обмежень C в X при яких виконується еквіваленція виду:

$$A_i \leftrightarrow P_i\left(a_1^i(x_1), a_2^i(x_2), \dots, a_n^i(x_n)\right) \forall\, x \in A_i$$

і задовольняє умовам:

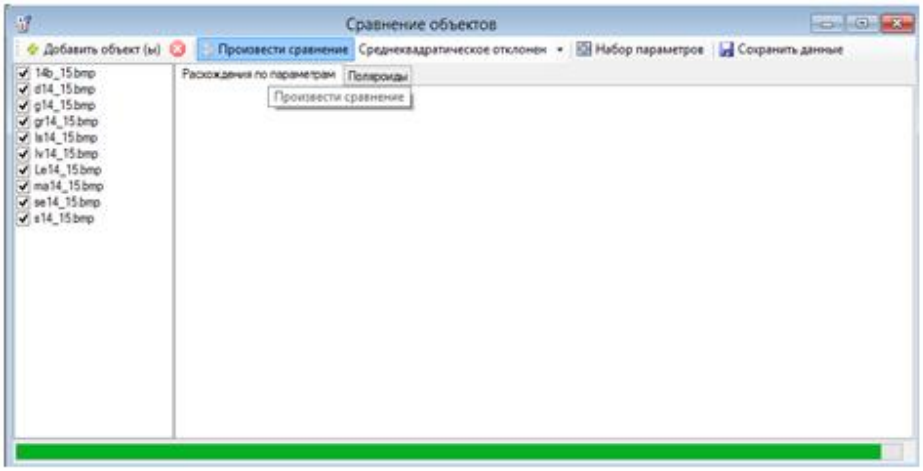
$$\forall i P_i\left(a_1^i(x_1), \dots, a_n^i(x_n)\right) \wedge\ \forall j P_j\left(a_1^j(x_1), \dots, a_n^j(x_n)\right) = \\ = 0, .. i \neq j, \vee_{i=1}^{k} A_i = A$$

## РОЗРОБКА АЛГОРИТМУ КЛАСИФІКАЦІЇ

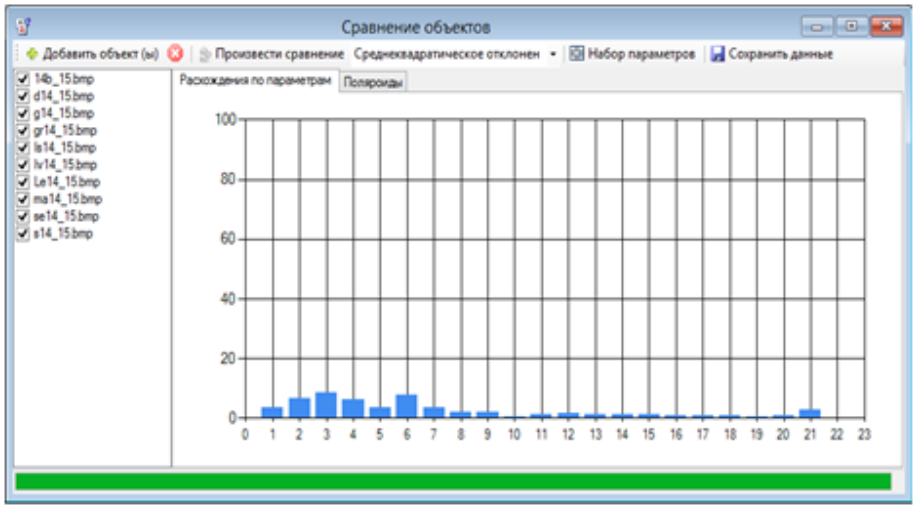
Зміни функцій приналежності характерних параметрів (вказаний крок можна розглядати як коригування розбиття $X_j$ на k розмитих множин) таким чином, щоб:

$$inf\mu_{A_{j1}^{'}}(x_{j1}^{'}) < sup\mu_{A_{j1}^{'}}(x_{j1}^{'}) = inf\mu_{A_{j2}^{'}}(x_{j2}^{'}) < sup\mu_{A_{j2}^{'}}(x_{j2}^{'}) = \cdots \\ < sup\mu_{A_{jk-1}^{'}}(x_{jk-1}) = inf\mu_{A_{jk}^{'}}(x_{jk}^{'}) < sup\mu_{A_{jk}^{'}}(x_{jk}^{'})$$

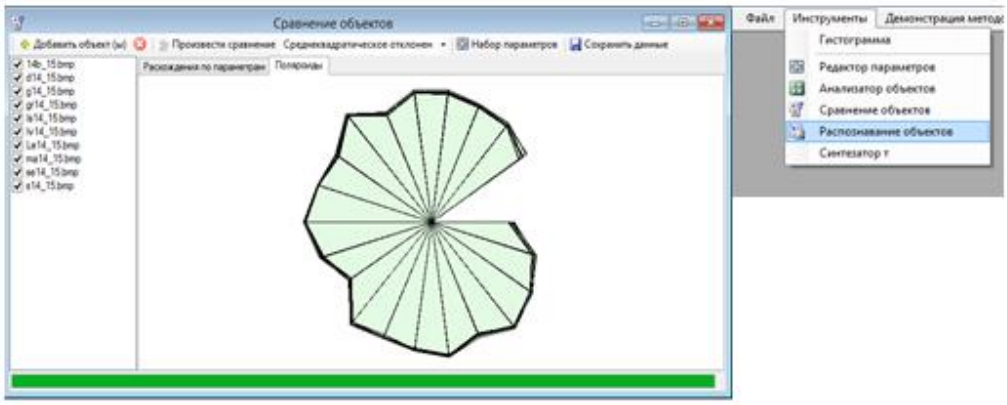
## ПРОГРАМНА РЕАЛІЗАЦІЯ



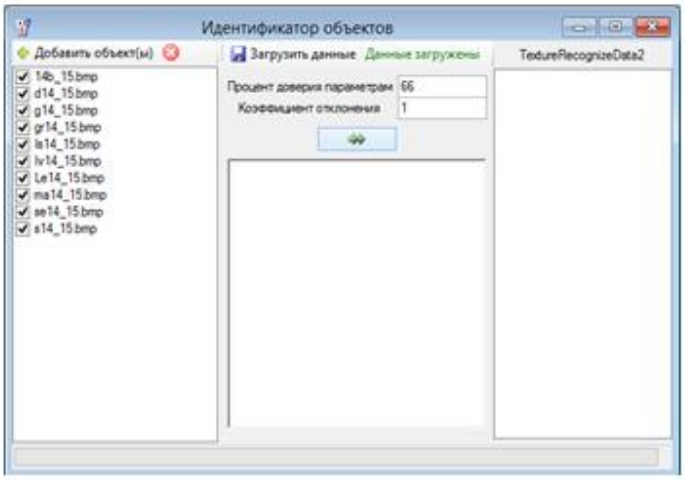

## ПРОГРАМНА РЕАЛІЗАЦІЯ



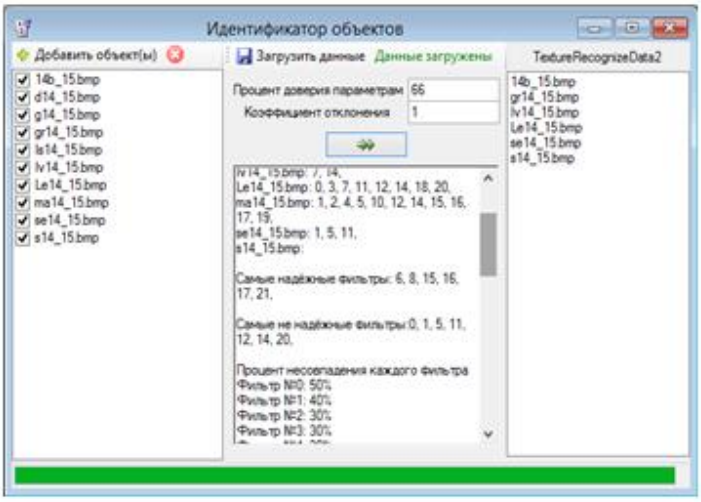
## ПРОГРАМНА РЕАЛІЗАЦІЯ



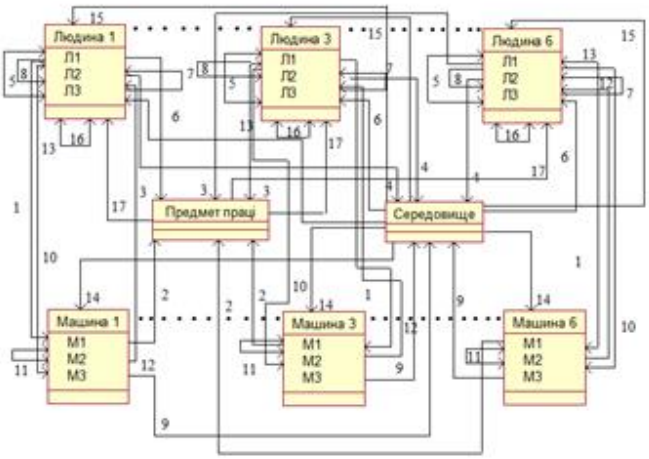

## ПРОГРАМНА РЕАЛІЗАЦІЯ



## ПРОГРАМНА РЕАЛІЗАЦІЯ




## ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ



21

## ВИСНОВКИ

У результату дипломної роботи був розроблений і програмно реалізований алгоритм класифікації. Практичні дослідження роботи показали доцільність алгоритму класифікації. Проте залишаються невирішеним ряд проблем:

- по-перше, це відсутність достатньої кількості верифікованих даних, край необхідних при налагодженні алгоритму та його подальшого удосконалення;
- по-друге, це давня проблема відображення отриманих класів в багатовимірному просторі.

Іншими словами робота з великою кількістю даних дуже важка робота і вимагає складних обчислювальних операцій а так само великої кількісті ресурсів для обчислення алгоритмів класифікації багатовимірних об'єктів.

22

# ДЯКУЮ ЗА УВАГУ