Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)
(рівень вищої освіти)

Використання методів ітераційної кластеризації складних образів
(тема)

Виконав:
студент 2 курсу, групи СШІм-18-1

Городовенко А.А.
(прізвище, ініціали)

Спеціальність 122 – Комп'ютерні науки
(код і повна назва спеціальності)

Тип програми Освітньо-професійна
(освітньо-професійна або освітньо -наукова)

Освітня програма Системи штучного інтелекту (СШІ)
(повна назва освітньої програми)

Керівник доц. Магдаліна І.В.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ В.О. Філатов
(підпис) (прізвище, ініціали)

2019 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук

Кафедра Штучного інтелекту

Рівень вищої освіти другий (магістерський)

Спеціальність 122 – Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо -наукова)

Освітня програма Системи штучного інтелекту (СШІ)
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри
_____
(підпис)
«____»_____ 20 ___ р.

ЗАВДАННЯ
НА АТЕСТАЦІЙНУ РОБОТУ

студентові Городовенку Анатолію Анатолійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Використання методів ітераційної кластеризації складних образів
_____
затверджена наказом по університету від 04.11.2019 р. №1623Ст
2. Термін подання студентом роботи до екзаменаційної комісії 19 грудня 2019 р.
3. Вихідні дані до роботи методи і моделі динамічної кластеризації лінійно неподільних експериментальних даних з різними характеристиками, вибірки з зашумленими даними, процес динамічної кластеризації лінійно-неподільних експериментальних даних з різноманітними характеристиками, алгоритм кластеризації Хамелеон, данні для початкової побудови математичної моделі, багаторівнева модель динамічної кластеризації даних
_____
4. Перелік питань, що потрібно опрацювати в роботі аналіз предметної області, дослідження алгоритму кластеризації Хамелеон і розробка модифікації даного алгоритму, критерії якості результатів кластеризації, аналіз і побудова мат моделей, тестування результатів на експериментальних і реальних даних
_____
_____
_____
_____
_____

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів)Рисунок 1 – Різні стадії алгоритму багаторівневого k-way поділу, Рисунок 2 – Схема роботи методу розподілу навпіл, Рисунок 3 – Процедура рекурсивної бісекції, Рисунок 4 – Два розбиття графа щодо осей координат, Рисунок 5 – 4-way поділ, виконаний за CND схемою, Рисунок 6 – 8-way поділ, виконаний за CND схемою, Рисунок 7 – Рекурсивно інерційний поділ, Рисунок 8 – Суцільною лінією показано розбиття, отримане LND-алгоритмом, Рисунок 9 – 3d фігури, отримані обертанням і зміщенням, Рисунок 10 – Графічне представлення опису даних математичною моделлю, Рисунок 11 – Графічне представлення залишків, Рисунок 12 – Залежність часу побудови асиметричного графа в залежності від кількості елементів вибірки для модифікованого і не модифікованого варіантів алгоритму, Рисунок 13 – Графічне представлення даних описаних математичною моделлю

6. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Основна частина	Доцент Магдаліна І.В.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Об'єктний аналіз поставленої задачі	04.11.19 – 06.11.19	Виконано
2	Розробка математичної моделі	07.11.19 – 15.11.19	Виконано
3	Реалізація алгоритму	16.11.19 – 20.11.19	Виконано
4	Тестування і налагодження програми	21.11.19 – 29.11.19	Виконано
5	Підготовка пояснювальної записки	30.11.19 – 03.12.19	Виконано
6	Підготовка презентації та доповіді	03.12.19 – 08.12.19	Виконано
7	Нормоконтроль, рецензування	10.12.19 – 17.12.19	Виконано

Дата видачі завдання ___04 листопада___2019 р.

Студент _____
(підпис)

Керівник роботи _____ _____
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Записка пояснювальна: 81 с., 16 рис., 8 табл., 24 дод., 14 джерел.

АЛГОРИТМ ХАМЕЛЕОН, ЗВ'ЯЗНІСТЬ, КЛАСТЕРИЗАЦІЯ, K–НАЙБЛИЖЧИХ СУСІДІВ, ПОБУДОВА ГРАФА

Об'єкт дослідження – це процес динамічної кластеризації лінійно нероздільних експериментальних даних з різними характеристиками.

Предмет дослідження – це методи та моделі динамічної кластеризації лінійно неподільних експериментальних даних з різними характеристиками.

Мета роботи – розробка універсальної математичної моделі залежності вибору комбінації алгоритмів на різних етапах ієрархічного алгоритму Хамелеон від вихідних характеристик аналізованого набору даних з метою поліпшення якості кластеризації.

Методи дослідження – для розробки математичної моделі вибору найкращого методу кластеризації зразків на основі характеристик вхідних даних були використані різні методи кластеризації та графіки, моделювання.

РЕФЕРАТ

Пояснительная записка: 81 с., 16 рис., 8 табл., 24 прил., 14 источников.

АЛГОРИТМ ХАМЕЛЕОН, КЛАСТЕРИЗАЦИЯ, K–БЛИЖАЙШИХ СОСЕДЕЙ, ПОСТРОЕНИЕ ГРАФА, СВЯЗНОСТЬ

Объект исследования – это процесс динамической кластеризации линейно неразделимых экспериментальных данных с разными характеристиками.

Предмет исследования – это методы и модели динамической кластеризации линейно неделимых экспериментальных данных с различными характеристиками.

Цель работы – разработка универсальной математической модели зависимости выбора комбинации алгоритмов на различных этапах иерархического алгоритма Хамелеон от исходных характеристик рассматриваемого набора данных с целью улучшения качества кластеризации.

Методы исследования – для разработки математической модели выбора наилучшего метода кластеризации образцов на основе характеристик входных данных были использованы различные методы кластеризации и графики, моделирования.

ABSTRACT

Explanatory note: 81 p., 16 fig., 8 tabl., 24 ann., 14 sources.

BUILDING GRAPH, CHAMELEON ALGORITHM, CLUSTERIZATION, CONNECTIVITY, K–NEAREST NEIGHBORS

The object of study: is the process of dynamic clustering of linearly indivisible experimental data with different characteristics.

The subject of the study: are methods and models of dynamic clustering of linearly indivisible experimental data with different characteristics.

The purpose of the work: is to develop a universal mathematical model for the dependence of the combination of algorithms at different stages of the chameleon hierarchical algorithm on the original characteristics of the analyzed data set in order to improve the quality of clustering.

Research Methods: To develop a mathematical model for selecting the best method for clustering samples based on the characteristics of the input data were used different clustering methods and graphs, modeling.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

CND – Coordinate Nested Dissection – покоординатне розбиття;

FCC – First Choice Coarsening – поєднання кращого (першого) вибору;

GGGP – Greedy Graph Growing Algorithm – алгоритм зростаючого графа з урахуванням вигод;

GGP – Graph Growth Partitioning – алгоритм зростаючого графа;

GR – Greedy refinement – послідовне відновлення;

HEM – Heavy Edge Matching – парування важких ребер;

HSM – Heaviest Schema Matching – поєднання важких схем;

HTM – Heavy-triangle matching – поєднання важких трикутників;

LEM – Light Edge Matching – парування легких ребер;

LND – Levelized Nested Dissection – рівневе осередкове розбиття;

MRGB – Modified Recursive Graph Bisection – модифікований рекурсивний поділ графа;

RGB – Recursive Graph Bisection – рекурсивний поділ графа;

RIB – Recursive Inertial Bisection – рекурсивно інерційний поділ;

RM – Random Matching – випадкове парування;

RNCB – Recursive Node Cluster Bisection – рекурсивний поділ вузлів в кластери;

SB – Spectral bisection – спектральна бісекція;

SFCT – Space-flling Curve Techniques – розподіл мережі з використанням кривих заповнюючих простір;

SO – Spectral octasection – спектральна октасекція;

SQ – Spectral quadrisection – спектральна квадрісекція;

TLP – Top-level Partitioning – високорівневий поділ;

ВСТУП

Актуальність теми. Аналіз даних набуває все більшого значення в сучасному світі. У різних сферах людської діяльності постійно виникає потреба у вирішенні проблем аналізу, прогнозування та діагностики, виявлення прихованих залежностей та підтримки прийняття оптимальних рішень. Актуальність цих завдань визначається швидким зростанням інформації, розвитком технологій її збору, зберігання та організації в базах даних та сховищах даних (у тому числі Інтернет-технологій), в результаті яких точні методи аналізу інформації та моделювання досліджувані об'єкти часто відстають від потреб реального життя. Наразі актуальною проблемою є розробка універсальних та надійних методів та підходів, придатних для обробки інформації з різних сфер. Технології та підходи математичної теорії розпізнавання та класифікації та математичного моделювання можуть послужити основою для таких досліджень.

Існує багато різних методів, які можна застосувати для вирішення проблеми. Існує також ряд проблем для доступних методів. Серед цих проблем можна виділити наступні:

а) проблема обґрунтування якості результатів аналізу. Для різних вибірок та даних різні методи оцінки результату можуть дати найкращий результат;

б) у багатьох областях, особливо в медицині, наявні дані галасливі;

в) проблема аналізу великої кількості різнорідних факторів;

г) не лінійність відносин; наявність упущень, помилок вимірювань змінних.

Оскільки для різних наборів даних різні методи показують найкращі результати, для кожного окремого набору даних необхідний певний критерій вибору найкращого методу.

Сформулюємо актуальність досліджень у цій галузі:

а) необхідність організації на єдиних принципах та синхронізації вибору методу кластеризації на основі даних аналізованої вибірки;

б) необхідність уніфікації технологій кластеризації і тим самим скорочення часу на вибір методу;

в) необхідність надання користувачам якісного рішення проблеми аналізу з різними даними, що вивчаються;

г) постійне збільшення обсягу надходить інформації та неоднорідність цієї інформації потребує розробки технологій аналізу цих даних;

д) окремі методи кластеризації добре працюють на відповідних зразках, але не є універсальними;

е) необхідність аналізу складних зразків із класами перекриття та перекриття.

В магістерській роботі були розроблені методи та моделі вибору найкращого методу кластеризації даних на основі таких характеристик вибірки:

а) підхід до кластеризації на основі алгоритму Хамелеона;

б) методи роботи з графіками на окремих етапах алгоритму Хамелеона;

в) математична модель вибору оптимального k при побудові графіка k-найближчих сусідів;

г) математична модель вибору найкращого методу кластеризації зразків на основі характеристик вхідних даних.

У роботі пропонуються модифікації алгоритму динамічної кластеризації Хамелеона. Цей алгоритм складається з наступних етапів: складання графіка, огрубіння графіка, поділ графіка, відновлення та вдосконалення графіка. На кожному з цих етапів були запропоновані та впроваджені набори методів, які покращують якість кластеризації для кожного запропонованого набору вихідних даних.

Окремі алгоритми були модифіковані з метою їх використання та покращення якості кластеризації побудованої моделі.

Була побудована математична модель залежності вибору k при побудові графіка k-найближчих сусідів на основі характеристик вибірки. Ця модель дозволяє прискорити процес побудови графіків. Математична модель будується із залежності вибору комбінації алгоритмів на різних етапах алгоритму Хамелеона від початкових характеристик аналізованого набору даних. Ця модель дозволяє визначити, який з алгоритмів слід використовувати для вибірки даних з деякими характеристиками, щоб отримати кластери найкращої якості.

Завдання дослідження:

а) проаналізувати та теоретично обґрунтувати процедуру підвищення точності кластеризації за допомогою ієрархічного алгоритму Хамелеона шляхом вибору найбільш ефективних алгоритмів на кожному етапі роботи;

б) визначити набір характеристик вихідних даних та критерії порівняння алгоритмів, що використовуються в рамках алгоритму Хамелеона;

в) розробити математичну модель вибору k при побудові графіка k-найближчих сусідів на основі характеристик вибірки для симетричних та асиметричних графіків;

г) розробити математичну модель вибору найкращого методу кластеризації зразків на основі характеристик вхідних даних;

Об'єкт дослідження – процес динамічної кластеризації лінійно невіддільних експериментальних даних з різними характеристиками.

Предмет дослідження – методи та моделі динамічного кластеризації лінійно невіддільних експериментальних даних з різними характеристиками.

Методи дослідження – у роботі були використані різні методи кластеризації та роботи з графіками, моделювання для розробки

математичної моделі вибору найкращого методу кластеризації зразків на основі характеристик вхідних даних.

Наукова новизна результатів така:

а) подальший розвиток отримав ієрархічний алгоритм Хамелеона, який відрізняється від існуючого інтеграцією існуючих алгоритмів кластеризації з ієрархічним алгоритмом Хамелеона; така модифікація алгоритму Хамелеона покращує якість кластеризації при роботі зі складними лінійно невіддільними зашумленими експериментальними даними;

б) удосконалено критерій порівняння алгоритмів та аналізу вихідних даних, що використовуються в рамках алгоритму Хамелеона;

Практична цінність результатів. Практична цінність роботи така:

а) була створена модель вибору k для алгоритму k-найближчих сусідів на основі характеристик, поданих для подальшого застосування в рамках алгоритму Хамелеона;

б) створена модель залежності якості кластеризації складних лінійно невіддільних шумових даних від характеристик даних та алгоритмів динамічної кластеризації.

1 АНАЛІЗ ПРЕДМЕТОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Дослідження та аналіз методів інтелектуального аналізу даних

Інтелектуальний аналіз даних (Data Mining) – це процес виявлення раніше невідомих, нетривіальних, практично корисних та доступних інтерпретацій знань у вихідних даних, необхідних для прийняття рішень у різних сферах людської діяльності.

Методи майнінгу даних поділяються на статистичні (описовий аналіз, кореляційний та регресійний аналіз, факторний аналіз, аналіз дисперсії, компонентний аналіз, дискримінантний аналіз, аналіз часових рядів) та кібернетичні (штучні нейронні мережі, еволюційне програмування, генетичні алгоритми, асоціативна пам’ять, нечітка логіка, дерева рішень, експертні системи обробки).

Аналіз кластерів (кластеризація даних) – завдання розбиття заданої вибірки об'єктів (ситуацій) на підмножини, що називаються кластерами, так що кожен кластер складається з подібних об'єктів, а об'єкти різних кластерів значно відрізняються. Завдання кластеризації стосується статистичної обробки, а також широкого класу навчальних завдань без вчителя. Кластерний аналіз – це багатовимірна статистична процедура, яка збирає дані, що містять інформацію про вибірку об'єктів, а потім організовує об'єкти у відносно однорідні групи (кластери) (Q-кластеризація, або Q-техніка, сам кластерний аналіз) [1].

Кластерний аналіз – це спосіб групування багатовимірних об'єктів, заснований на представленні результатів окремих спостережень за точками відповідного геометричного простору з подальшим виділенням груп як «скупчення» цих точок (скупчення, таксони). Кластер англійською мовою означає «згусток», «гроно винограду», «грона зірок» тощо. Цей метод дослідження був розроблений в останні роки у зв'язку з можливістю комп'ютерної обробки великих баз даних. Кластер – це група елементів,

що характеризується загальною властивістю, головна мета кластерного аналізу – знайти групи подібних об’єктів у вибірці [2].

Кластерний аналіз передбачає відбір компактних, віддалених одна від одної груп об'єктів, пошуку «природного» поділу населення на ділянки кластерних об’єктів. Він використовується, коли вихідні дані подаються у вигляді матриць близькості або відстані між об'єктами або у вигляді точок у багатовимірному просторі [3]. Найбільш поширеними є дані другого типу, для яких кластерний аналіз орієнтований на виявлення деяких геометрично віддалених груп, в межах яких об'єкти знаходяться поблизу [2].

Діапазон застосувань кластерного аналізу дуже широкий: він використовується в археології, медицині, психології, хімії, біології, державному управлінні, філології, антропології, маркетингу, соціології та інших дисциплінах. Однак універсальність програми призвела до появи великої кількості несумісних термінів, методів та підходів, що перешкоджають однозначному використанню та послідовній інтерпретації кластерного аналізу.

Завдання та умови.

Кластерний аналіз виконує такі основні завдання:

а) розробка типології чи класифікації;

б) вивчення корисних концептуальних схем для групування об’єктів;

в) генерування гіпотез на основі дослідження даних;

г) тестування гіпотез або досліджень, щоб визначити, чи типи (групи), визначені так чи інакше, насправді є у наявних даних.

Незалежно від предмета дослідження, застосування кластерного аналізу передбачає наступні кроки:

а) відбір вибірки для кластеризації;

б) визначення набору змінних, за якими будуть оцінюватися об’єкти у вибірці;

в) обчислення значень міри подібності між об'єктами;

г) застосування методу кластерного аналізу для створення груп подібних об'єктів;

д) перевірка результатів кластерного рішення.

Кластерний аналіз має такі вимоги до даних:

а) показники не повинні співвідносити один одного;

б) показники повинні бути безрозмірними;

в) їх розподіл має бути близьким до нормального;

г) показники повинні відповідати вимозі «стійкості», що розуміється як відсутність впливу випадкових факторів на їх значення;

д) зразок повинен бути однорідним, не містити «залишків».

Якщо перед кластерним аналізом передує факторний аналіз, то вибірку не потрібно «відновлювати» – заявлені вимоги дотримуються автоматично самою процедурою моделювання факторів (є ще одна перевага – z-стандартизація без негативних наслідків для вибірки; якщо він проводиться безпосередньо для кластерного аналізу, це може призвести до зниження чіткості поділу груп). В іншому випадку вибір потрібно скорегувати.

Цілі кластеризації:

а) розуміння даних шляхом ідентифікації структури кластера. Поділ вибірки на групи подібних об'єктів дає можливість спростити подальшу обробку даних та прийняття рішень шляхом застосування різного методу аналізу до кожного кластеру (стратегія поділу та підкорення);

б) стиснення даних. Якщо початковий зразок надмірно великий, то ви можете зменшити його, залишивши одного з найбільш типових представників кожного кластеру;

в) виявлення новинок. Вибираються нетипові об'єкти, які не можна приєднати до жодного кластеру;

У першому випадку вони намагаються зробити кількість кластерів меншими. У другому випадку важливіше забезпечити високу схожість об’єктів всередині кожного кластеру, і може бути якомога більше

кластерів. У третьому випадку найбільший інтерес представляють окремі об'єкти, які не входять до жодного кластеру.

У всіх цих випадках може застосовуватися ієрархічна кластеризація, коли великі кластери розбиваються на менші, які в свою чергу розбиваються ще меншими і т. Д. Такі завдання називаються проблемами систематики.

Результатом таксономії є деревоподібна ієрархічна структура. Більше того, кожен об’єкт характеризується переліком усіх кластерів, до яких він належить, як правило, від великих до малих.

Кластеризація (навчання без вчителя) відрізняється від класифікації (навчання з викладачем) тим, що мітки вихідних об'єктів спочатку не встановлені, і навіть сам набір може бути не відомий.

Рішення проблеми кластеризації принципово неоднозначне, і для цього є кілька причин:

а) не існує однозначно найкращого критерію якості кластеризації. Відомо низку евристичних критеріїв, а також ряд алгоритмів, які не мають чітко визначеного критерію, але реалізують досить обґрунтовану кластеризацію «за конструкцією». Усі вони можуть дати різні результати;

б) кількість кластерів, як правило, не відома заздалегідь і встановлюється відповідно до якогось суб'єктивного критерію;

в) результат кластеризації істотно залежить від метрики, вибір якої, як правило, також суб'єктивний і визначається експертом.

Завдання кластерного аналізу (або навчання без вчителя) полягає в наступному. Існує навчальний зразок $X\ell = \{x1, ..., x\ell\} \in X$ і функція відстані між об'єктами $\rho(x, x')$. Необхідно розбити вибірку на роз'єднані підмножини, що називаються кластерами, так що кожен кластер складається з об'єктів, близьких за метрикою ρ, а об'єкти різних кластерів значно відрізняються. У цьому випадку мітка (номер) кластера уі присвоюється кожному об'єкту хі є $X\ell$. Алгоритм кластеризації – це функція а: $X \to Y$, яка пов'язує мітку кластера у є Y з будь-яким об'єктом х

є X. Набір міток Y відомий в деяких випадках, але частіше завдання полягає у визначенні оптимальної кількості кластерів з точки зору того чи іншого критерію якості кластеризації [2].

Рішення проблеми кластерного аналізу – це розділ, який задовольняє деяку умову оптимальності. Цей критерій може представляти деякий функціонал, що виражає рівні бажаності різних розділів і груп. Ця функціональність часто називається об'єктивною функцією. Завдання кластерного аналізу – це проблема оптимізації, тобто пошук мінімуму цільової функції для заданого набору обмежень. Прикладом цільової функції є, зокрема, сума квадратів внутрішньо групових відхилень для всіх кластерів [2].

Можна виділити наступні основні етапи кластерного аналізу.

Формування системи змінних. Часто необхідно заздалегідь вибрати з початкового набору змінних найефективнішу підсистему (в зарубіжній літературі цей процес називається «вибір функції»). Крім того, у деяких завданнях доцільно трансформувати вихідні змінні таким чином, щоб сформувати нові, більш інформативні показники («вилучення функції»). Щоб уникнути «домінування» змінних з великою шкалою вимірювання, проводиться попередня нормалізація початкових змінних.

Визначення способу обчислення відстані між об'єктами або групами об'єктів. Цей метод повинен відображати специфіку застосованої проблеми. Наприклад, у випадку безперервних змінних може бути вказана евклідова відстань. Для усунення ефекту сильних лінійних кореляцій між змінними використовується відстань Махалонобіс. Для номінальних змінних можна використовувати відстань Хеммінга. Для груп об’єктів метод знаходження відстані також визначається, наприклад, за принципом «далекий сусід», «найближчий сусід» тощо. Принцип «віддаленого сусіда» виправданий, коли є апріорна інформація що таксони мають компактну сферичну форму. Принцип «близького сусіда» має сенс застосовувати,

якщо відомо, що таксони можуть мати «витягнуту» форму або розташовані концентрично.

Групування об’єктів. На цьому кроці створення груп об’єктів. Поділ на групи може бути «важким» (формується розділ початкового набору об'єктів) і може бути «нечітким» (розраховується ступінь належності кожного об'єкта до груп). Існує велика різноманітність алгоритмів групування.

Представлення результатів. Потрібний простий та інформативний опис отриманих кластерів. Часто для такого опису вибирається «типовий об'єкт» або визначається набір показників, усереднених по групі показників. Опис також використовується у вигляді набору таксонів. Під таксоном маємо на увазі піддомен простору змінних мінімального обсягу, що мають деяку задану форму і містять точки відповідної групи.

Визначення якості отриманого групування. Після завершення кластеризації необхідно переконатися, що сформовані групи справді відображають внутрішні закономірності, характерні для вирішуваної проблеми, сприяють досягненню цілей аналізу та допомагають виявити нові властивості досліджуваних об’єктів. Існують також більш формальні методи контролю якості, пов'язані з знаходженням ймовірності випадкового утворення груп, які можна обчислити в рамках певної моделі розподілу (з верифікацією статистичних гіпотез щодо однорідності спостережень різних класів); методом завантаження; з розрахунком різних показників якості (внутрішньо груповий розкид, індекс Гудмана-Крускала; Ранд; С-індекс тощо) [4].

1.2 Основні підходи, що використовуються в кластерному аналізі

В даний час існує декілька підходів до вирішення проблеми кластерного аналізу, які базуються на різних уявленнях про проблему, використанні додаткової інформації, характерної для кожної предметної

області тощо. Коротко перерахуйте найбільш часто використовувані підходи. Зауважте, що описана нижче класифікація не є чіткою; деякі методи можуть бути розроблені на основі поєднання різних підходів:

а) імовірнісний підхід. Передбачається, що кожен об'єкт у популяції належить до одного з класів К, однак номери класів безпосередньо не визначаються. Об'єкти вибираються випадковим чином і незалежно від населення; тому змінні, що описують об'єкти, є випадковими. Для кожного класу визначається розподіл ймовірності даної родини; Параметри розподілу невідомі. Існуюча вибірка спостережень – це реалізація суміші розподілів;

б) підхід з використанням аналогії з центром ваги. Для кожної групи визначається вектор середніх значень показників, інтерпретується як «центр ваги» групи. Використовується критерій внутрішньо групового розсіювання: де координата «центру ваги» k-го кластера в змінній Xj, j = 1,2, ..., n, k = 1,2, ..., K. Оптимальне групування для даного К відповідає мінімальному значенню критерію;

в) підхід, заснований на теорії графів. Найвідоміший алгоритм цього сімейства – це найкоротший алгоритм відкритого шляху. Попередньо побудовано мінімальне прольотове дерево графа, у якому вершини відповідають об'єктам, а краї мають довжину, рівну відстані між відповідними об'єктами. Для утворення скупчень краї максимальної довжини видаляються із побудованого дерева;

г) ієрархічний підхід. Цей напрям також пов'язаний з графо-теоретичним підходом. Результати групування представлені у вигляді дерева групування (дендрограма). Алгоритми, засновані на такому підході, можна поділити на агломеративні (поступово поєднуючи найближчі групи чи об’єкти) та подільні (у яких початкова група поділяється на найбільш віддалені підгрупи за ступенями; ті, у свою чергу, також поділяються на підгрупи тощо). Рішення щодо групування – це вкладена ієрархія підгруп;

д) підхід, заснований на концепції найближчого сусіда. Групування здійснюється послідовно шляхом присвоєння об'єкта кластеру, в якому знаходиться найближчий об'єкт, за умови, що відстань до об'єкта не перевищує заданого порогу. Існують різні варіанти визначення відстані; при визначенні міри близькості також може враховуватися розташування інших сусідніх точок;

е) алгоритми нечіткого аналізу кластерів. Використовуючи такий підхід, передбачається, що кожен кластер – це нечітка сукупність об'єктів;

є) підхід із використанням штучних нейронних мереж заснований на аналогії з процесами, що відбуваються в біологічних нейронних системах. Відомо велику кількість алгоритмів цього сімейства. Типова архітектура – це одношарова мережа, в якій кожному нейрону відповідає певний кластер. У процесі навчання мережі відбувається ітеративна зміна передавальних ваг між вхідними та вихідними вузлами мережі; тим самим шукаючи оптимальне значення критерію групування. Нейронні мережі дозволяють ефективно використовувати паралельні обчислювальні методи;

ж) еволюційний (генетичний) підхід. Алгоритми цього сімейства побудовані на аналогії з природною еволюцією. Вони використовують поняття популяції - набір різних варіантів групування (їх також називають хромосомами, за аналогією з відповідними біологічними об'єктами), а еволюційні оператори - процедури, які дозволяють отримати одну або кілька хромосом нащадків з однієї або декількох батьківських хромосом. Ці процедури: відбір, рекомбінація та мутація. Генетичний алгоритм здатний шукати рішення, яке забезпечує глобальний мінімум критерію якості групування.

Існують й інші підходи до вирішення проблеми глобальної оптимізації, які можна застосувати в кластерному аналізі. Наприклад, відомий підхід моделювання відпалу, який також ґрунтується на ідеї моделювання природних процесів.

1.3 Вимоги до алгоритму кластеризації

Кластеризація є перспективним полем для досліджень, де сфери потенційного застосування диктують додаткові вимоги:

а) масштабованість. Багато алгоритмів кластеризації добре працюють на невеликих вибірках даних, які складаються з менш ніж 200 об'єктів; однак великі бази даних можуть містити мільйони об'єктів. Кластеризація великої вибірки може призвести до упереджених результатів. Потрібні добре масштабовані алгоритми;

б) здатність працювати з характеристиками різних типів. Багато алгоритмів розроблені для кластеризації числових даних. Тим не менш, може бути затребувана робота з іншими типами даних: двійкові, категоричні (номінальні) та порядкові дані або суміш цих типів даних;

в) розпізнавання скупчень довільної форми. Багато алгоритмів кластеризації визначають кластери на основі евклідової та манхеттенської відстаней. Алгоритми, засновані на таких типах відстаней, як правило, знаходять сферичні скупчення з однаковою щільністю та однаковими розмірами. Однак скупчення може бути будь-якої форми. Важливо розробити алгоритм пошуку кластерів довільної форми;

г) мінімальні вимоги до галузі дослідження для визначення вхідних параметрів. Багато алгоритмів кластеризації вимагають певних вхідних параметрів (наприклад, таких як кількість кластерів). Результати кластеризації можуть бути досить чутливими до вхідних параметрів. Параметри часто важко визначити, особливо якщо вхідний набір складається з багатовимірних об'єктів. Це не тільки вимагає втручання користувача, але й ускладнює моніторинг якості кластеризації;

д) вміння працювати з галасливими даними. Більшість реальних баз даних містять втрачені, невідомі або помилкові дані. Деякі алгоритми кластеризації чутливі до таких даних, що може стосуватися низької якості кластеризації;

е) нечутливість до порядку вхідних записів. Деякі алгоритми кластеризації чутливі до порядку, в цьому випадку для одного і того ж набору, коли об’єкти представлені в іншому порядку, будуть отримані абсолютно різні класи;

є) високий розмір. База даних або сховище даних може мати кілька вимірів або ознак. Багато алгоритмів хороші для роботи з низькомірними даними;

ж) обмежене кластеризація. Реальні програми можуть накладати на кластеризацію різні типи обмежень. Перспективним завданням є пошук груп даних з хорошою кластеризацією, яка відповідає висунутим обмеженням;

з) простота використання та інтерпретація. Користувач очікує зрозумілих, порівняних та корисних результатів кластеризації;

и) кількість переглядів бази даних. Наявна пам'ять повинна бути достатньою для обробки великих наборів даних високих розмірів [3], [5], [6].

1.4 Актуальні проблеми кластерного аналізу

Незважаючи на велику кількість досліджень у галузі кластерного аналізу, у цій галузі існує низка нагальних проблем. Перерахуємо основні проблеми:

а) проблема обґрунтування якості результатів аналізу. Відомо, що процес групування багато в чому суб’єктивний. Це виражається, зокрема, у тому, що один і той же набір об'єктів можна класифікувати по-різному залежно від області застосування, ступеня повноти знань про об’єкти дослідження тощо. Тому необхідно розробити методи, щоб повністю прийняти враховувати наявні експертні знання, а також розробляти відповідні критерії якості групування;

б) для багатьох напрямків дослідження, які важко формалізувати, характерна відсутність знань про досліджувані об'єкти, що ускладнює формулювання їх математичних моделей;

в) проблема аналізу великої кількості різнорідних (кількісних чи якісних) факторів. У випадку гетерогенного простору виникає методологічна проблема визначення метрики в ньому. З іншого боку, навіть у просторі однотипних (кількісних) змінних із збільшенням їх кількості посилюється «прокляття виміру», що може призвести до майже повної нерозрізненості балів. Отже, відстань від будь-якої точки до її «найближчого сусіда» для деяких типів відстаней може практично збігатися (з урахуванням точності машини) з відстані до її «далекого сусіда». Візуальні аналогії, що мають відношення до простору малого розміру, стають абсолютно неприйнятними у просторі великого розміру;

г) нелінійність відносин; наявність упущень, помилок вимірювань змінних. Класичні методи зменшення розмірності (метод основних компонентів; метод незалежних компонентів), що використовуються в кластерному аналізі, в основному зосереджені на лінійних зв’язках між змінними. Для виявлення більш складних зв’язків такі алгоритми, як нелінійні (ядерні) методи основних компонентів тощо;

д) необхідність подання результатів аналізу у формі, зрозумілій фахівцям прикладної галузі. Окрім хорошої здатності прогнозування будь-якого алгоритму аналізу даних, важливо, наскільки зрозумілі та інтерпретовані його результати. Для поліпшення інтерпретації рішень можна використовувати логічні моделі. Такі моделі використовуються для вирішення завдань розпізнавання образів та прогнозування кількісних показників;

е) проблема пошуку глобального екстремуму в критерії групування якості. Критерій якості - це функція, яка залежить від великої кількості факторів, нелінійних, з багатьма локальними екстремумами. Для пошуку кластерів необхідно вирішити складну комбінаторну задачу пошуку

оптимального варіанту класифікації. Як відомо, кількість різних варіантів поділу N об'єктів на K групи становить:

$$M(N,K)=\frac{1}{K!}\sum_{i=1}^{K}(-1)^{i}\binom{K}{i}(K-i)^{N}; \qquad (1.1)$$

є) проблема стабільності групування рішень. У класичних алгоритмах для вирішення задач кластерного аналізу результати групування можуть сильно відрізнятися залежно від вибору початкових умов, порядку об'єктів, параметрів алгоритмів тощо [4].

Тому алгоритм повного перебору варіантів має складність, яка експоненціально залежить від розмірності. Якщо кількість груп не відома заздалегідь, то вичерпне завдання стає ще складніше. Таким чином, у міру збільшення розмірності таблиць даних відбувається «комбінаторний вибух». Алгоритми класичного кластерного аналізу виконують спрямований пошук у відносно невеликому підмножині простору рішення, використовуючи різні апріорні обмеження. Більше того, пошук строго оптимального рішення не гарантується. Для пошуку оптимального рішення використовуються більш складні методи, такі як генетичні (еволюційні) алгоритми, нейронні мережі тощо. Існують експериментальні дослідження, що підтверджують переваги таких алгоритмів над класичними алгоритмами;

На основі аналізу сформулюємо цілі наукових досліджень:

а) дослідження та розробка модифікованого підходу до кластеризації на основі алгоритму Хамелеона;

б) модифікація методів роботи з графіками на окремих етапах алгоритму Хамелеона;

в) побудова математичної моделі вибору оптимального k при побудові графіка k-найближчих сусідів та найкращого методу кластеризації зразків на основі характеристик вхідних даних.

2 ДОСЛІДЖЕННЯ АЛГОРИТМУ КЛАСТЕРИЗЦІ ХАМЕЛЕОН ТА РОЗРОБКА МОДИФІКАЦІЇ ЦЬОГО АЛГОРИТМУ

2.1 Аналіз та опис основних етапів алгоритму Хамелеон

Проблема розділення графа – це поділ вершин цього графа на p приблизно рівних частин, так що кількість ребер між вершинами різних класів є мінімальним. Це завдання використовується в багатьох різних сферах, включаючи паралельні наукові обчислення або планування завдань. Проблема відокремлення є NP-повною. Тим не менш, велика кількість розроблених алгоритмів знаходить досить хороший поділ. Проблема розділення k-шляху найчастіше вирішується методом рекурсивної бісекції. Останнім часом з'явився високоефективний метод розділення графіка k-way – багаторівнева рекурсивна бісекція (MLRB). Основна структура багаторівневої рекурсивної бісекції дуже проста. На початку графік G закручується до декількох сотень вершин, потім отриманий зменшений графік ділиться навпіл, а потім цей поділ проектується назад на початковий графік шляхом періодичного відновлення поділу.

Багаторівнева парадигма також може бути застосована для побудови поділу на k-шлях безпосередньо на вихідному графіку (рисунок 2.1) [7]. Графік розміщений в послідовному порядку, як і в попередній схемі, але тепер грубий графік одразу ділиться на k частини, і цей k поділ послідовно відновлюється до початкового. Існує ряд переваг негайного виконання k поділу. По-перше, грубість потрібно робити лише один раз, що зменшує складність алгоритму та час виконання. По-друге, добре відомо, що багаторівнева рекурсивна бісекція може працювати гірше, ніж k-розщеплення. Таким чином, метод досягнення негайно k-поділу може відокремити краще. Слід зауважити, що негайно обчислити хороше розділення в k-напрямку важче, ніж виконати хороший розбір. З цієї

причини найпоширеніше рішення проблеми поділу k-шляху виконується за допомогою рекурсивної бісекції [8].

На стадії огрубіння розмір графіка поступово зменшується; на етапі початкового розбиття виконується k-шлях поділу зменшеного графіка (6-ти напрямний в цьому випадку); на етапі відновлення виконується проекція поділу на початковий графік.

Наприклад, найпростішим методом обчислення початкового поділу в контексті багаторівневого алгоритму є грубування графу до k вершин. Однак у фазі вдосконалення необхідно вдосконалити поділ k-шляху, що набагато складніше, ніж поліпшення бісекції. Навіть для 8-ти напрямного розділення час виконання для цієї схеми досить високий. Щоб покращити поділ k-шляху для k> 8, час виконання стає надмірно довгим.

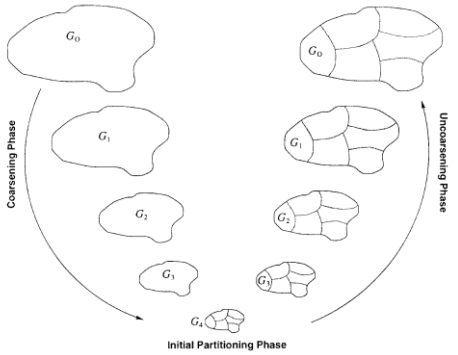


Рисунок 2.1 – Різні стадії алгоритму багаторівневого k-way поділу

Хамелеон – це новий ієрархічний алгоритм, який долає обмеження існуючих алгоритмів кластеризації. Цей алгоритм враховує динамічне моделювання в ієрархічній кластеризації. Ключовим моментом алгоритму Хамелеона є те, що він враховує як взаємопов'язаність, так і близькість при визначенні однакових пар кластерів. Саме це дозволяє подолати обмеження. Хамелеон використовує новий підхід при визначенні ступеня взаємопов'язаності та близькості між парами кластерів. При такому підході алгоритм сам обчислює внутрішні характеристики кластерів, тому вони не

залежать від статичних моделей, встановлених користувачем, і можуть автоматично підлаштовуватися під внутрішні характеристики комбінованих кластерів.

Хамелеон знаходить кластери, використовуючи двофазний алгоритм. На першому кроці Хамелеон використовує алгоритм розподілу графіків для кластеризації набору в досить невеликі підкласи. Другий крок використовує алгоритм пошуку природних кластерів шляхом послідовного поєднання отриманих малих підкласів.

Хамелеон представляє об'єкти через загальновживаний графік k-найближчого сусіда. Ця презентація графіків дозволяє масштабувати великі обсяги даних. Кожна вершина в цьому графі представляє один об'єкт даних. Між вершинами є ребро, якщо один об'єкт є одним з k-найближчих сусідів другого об'єкта. Графік k-найближчих сусідів містить поняття про те, що радіус суміжності об'єкта визначається щільністю області, в якій знаходиться даний об’єкт. Це дозволяє ідентифікувати природні скупчення.

Наступним кроком є побудова черги послідовно зменшених гіперграфів – фаза грубості (Coarsening Phase). Для огрублення графіків можна застосувати кілька існуючих алгоритмів. На кожному рівні грубозернистості огрубіння закінчується, як тільки розмір отриманого грубого графіка зменшується в 1,7 рази.

На третьому етапі поділ грубого графіка k-напрямку виконується таким чином, що виконується обмеження балансу та оптимізується функція розділення (мінкет).

Четвертий крок – відновити графік. Поділ грубого графіка проектується на наступний рівень вихідного графіка і алгоритм уточнення розділення виконується для вдосконалення цільової функції без порушення обмеження балансу.

При останній ітерації Хамелеона визначається індекс подібності між кожною парою кластерів з урахуванням їх відносної зв’язності та відносної

близькості. Це дозволяє вибирати кластери, які добре пов’язані та досить близькі. Вибираючи кластери та виходячи з цих двох критеріїв, Хамелеон долає обмеження існуючих алгоритмів, які оцінюють взаємозв'язок чи близькість.

Тому можна виділити наступні етапи:

а) побудова графіка. Граф може бути побудований симетричним або асиметричним. У побудові графіка можуть бути застосовані різні типи відстаней: Евклідова, Манхеттенська, Мінковська, Сквеклійська;

б) огрубіння графіка (Coarsening). Згрупування графіка може бути виконано наступними методами: Випадкове узгодження (RM), Узгодження важких країв (HEM), Збіг легких країв (LEM);

в) початковий поділ графіка (початкове розділення). Існує кілька підходів до поділу графіків: графічні методи, комбінаторні методи та спектральні методи. Алгоритми також можна виконувати як частину рекурсивної бісекції, оскільки більшість методів виконують поділ графів навпіл;

г) відновлення графіків (Uncoarsening) та покращення поділу графіків (уточнення) Для поліпшення поділу графів використовуються такі алгоритми: Керніган – Лін (KL), Межа KL, Fiduccia-Mattheyses (FM), Межа FM. Ті ж алгоритми можуть застосовуватися на етапі поділу, приймаючи для початкового випадкового поділу грубого графіка;

д) поєднання подібних класів для отримання остаточного розділу.

2.2 Аналіз, опис та модифікація початкового розділення графіка в рамках алгоритму Хамелеона (Initial partitioning)

Для пошуку впорядкування використовується рекурсивна бісекція, що зменшує заповнення при розкладанні розрідженої матриці. Ці алгоритми зазвичай називають вкладеними алгоритмами розділів. Вкладені секції рекурсивно ділять графік на майже рівні половини, видаляючи вузли

роздільника, поки не буде отримано потрібну кількість розділів. Один із способів отримати вузловий роздільник – спочатку отримати ділення на половину графіка, а потім обчислити вузловий роздільник від крайового роздільника. Вузли графіка нумеруються так, що на кожному рівні рекурсії вузли сепаратора нумеруються після вузлів у половинах графіка. Ефективність та складність алгоритму вкладеного розділу залежить від алгоритму розрахунку роздільника. Загалом невеликі сепаратори призводять до меншої наповнення [9].

Для вирішення проблеми розподілу графів може застосовуватися рекурсивно метод бінарного поділу, при якому на першій ітерації графік ділиться на дві рівні частини, потім на другому кроці кожна з отриманих частин також ділиться на дві частини тощо У випадку, коли необхідна кількість розділів k не є потужністю двох, кожен поділ навпіл повинен бути виконаний у відповідному відношенні [10].

Рекурсивна процедура ділення служить для поділу графіка, що містить n вершин, на довільну кількість доменів k. Представлений результат ділення 100 вершин графа на 7 частин приблизно однакового розміру (рисунок 2.3). У кожному з коренів дерева розділів вказана кількість вершин відповідного підграфа, два числа, розділені косою рисою, вказують на частку, в якій вершини підграфа поділяються на дві частини [11].

Схема операції способу поділу навпіл на прикладі поділу графіка на 5 частин можна описати наступним чином (рисунок 2.2). Спочатку графік слід розділити на 2 частини у співвідношенні 2: 3 (суцільна лінія), потім праву частину перегородки у співвідношенні 1: 3 (пунктирна лінія), після чого залишається розділити 2 крайні підрегіони на ліворуч та праворуч у співвідношенні 1: 1 (пунктирна лінія з крапкою) [10].

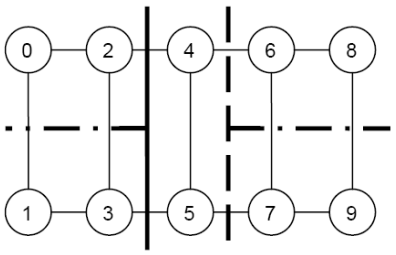


Рисунок 2.2 – Схема роботи методу розподілу навпіл

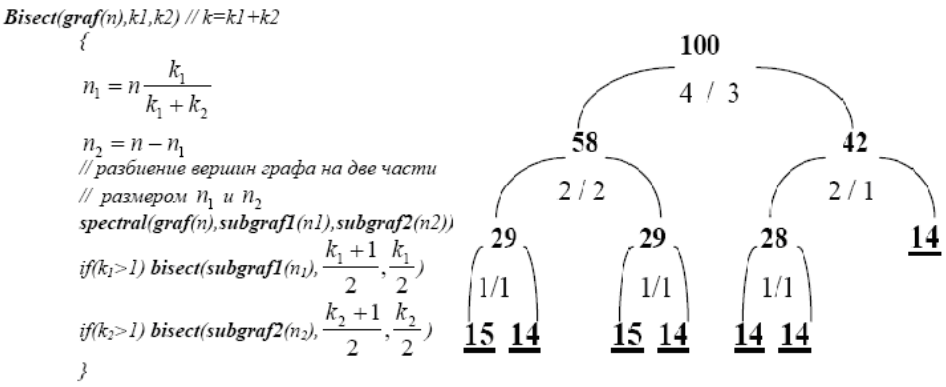


Рисунок 2.3 – Процедура рекурсивної бісекції

2.2.1 Геометричні алгоритми

Цей клас методів використовує геометричну інформацію про графік, щоб знайти хороший роздільник. Геометричні алгоритми обчислюють розділення лише на основі координат елементів множини, а не на зв’язності елементів. Оскільки ці технології не враховують зв'язок між елементами набору, зазвичай використовується мінімізація відповідних метрик, наприклад, кількість елементів, що прилягають до зовнішніх елементів (тобто розмір межі підкласу) [1].

Алгоритми геометричного поділу, як правило, швидкі, але часто виробляють поділи, якість яких гірша, ніж інші методи. Однак через випадковий характер цих алгоритмів часто потрібні багаторазові повтори (від 5 до 50), щоб отримати рішення, порівнянні за якістю з

спектральними методами. Багаторазові повторення збільшують загальний час роботи, але він залишається істотно меншим, ніж у спектральних методів. Алгоритми розділення геометричних графіків застосовні лише в тому випадку, якщо відомі координати вузлів графіків. У багатьох предметних областях (наприклад, лінійне програмування, VLSI-дизайн) немає геометричної інформації, пов'язаної з графіком [9], [12].

Координатне вкладене розсічення (CND). Це метод, заснований на рекурсивному поділі мережі по найдовшій стороні. В якості ілюстрації показаний приклад мережі, і два можливі перегородки, вироблені цим методом, розташовані вздовж осі x та вздовж осі y. Правильним у цьому випадку буде перегородка вздовж осі x.

Загальна схема методу така. Спочатку обчислюються центри маси елементів мережі. Отримані точки проектуються на вісь, що відповідає найбільшій стороні спільної мережі. Таким чином, ми отримуємо упорядкований список усіх елементів мережі. Розділивши список навпіл (можливо, у правильній пропорції), ми отримаємо необхідну бісекцію. Аналогічним чином отримані фрагменти перегородки рекурсивно діляться на потрібну кількість частин.

Метод розподілу координат вкладений дуже швидко і вимагає невеликої кількості оперативної пам’яті. Однак отриманий розділ за якістю поступається більш складним і обчислювально трудомістким методам. Крім того, у випадку складної структури мережі алгоритм може бути розділений на непоєднані підмережі (рисунок 2.4) [10].

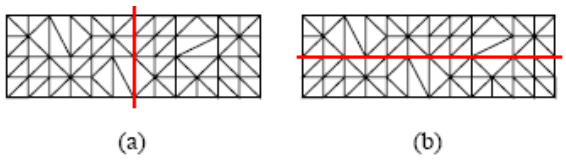


Рисунок 2.4 – Два розбиття графа щодо осей координат: (а) граф розділений щодо осі x; (б) щодо осі y

Спочатку було виконано, поділ показаний суцільною лінією. Після поділу показано пунктирною лінією для кожної підмножини. Верхня та нижня ліві підмножини не пов’язані між собою (рисунок 2.5) [13].

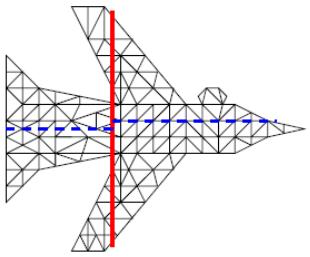


Рисунок 2.5 – 4-way поділ, виконаний за CND схемою

Алгоритм CND працює наступним чином. Обчислюються центри маси елементів набору. Після цього виконується проекція центрів маси на вісь координат, що відповідає максимальному значенню в наборі. Ця операція впорядковує елементи набору. Щоб поділити множину, отриманий упорядкований список ділиться навпіл. Кожен із отриманих підмножин, у свою чергу, можна поділити аналогічно.

8-сторонне розділення, виконане за описаним способом. Показані центри маси елементів множини та показані лінії рекурсивної бісекції. На першому кроці весь набір було розбито; на малюнку ця секція показана суцільною лінією. Потім поділ проводили для кожної з отриманих підмножин (рисунок 2.6).

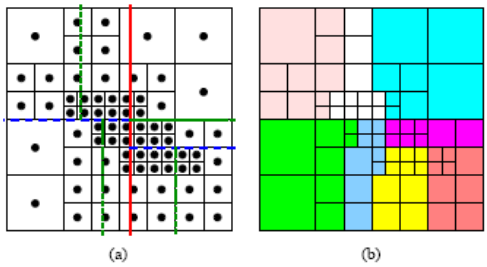


Рисунок 2.6 – 8-way поділ, виконаний за CND схемою

Алгоритм CND працює дуже швидко, не потребує великого обсягу пам’яті і його можна легко паралелізувати. Крім описаних переваг, слід додати легкість та компактність опису виконаного поділу, для цього достатньо вказати роздільники, що використовуються на кожному з вузлів дерева рекурсивної бісекції. Однак поділ, виконаний алгоритмом CND, є низькою якістю. Крім того, для складних геометричних фігур в результаті поділу можна отримати непоєднані підмножини [12].

Мережевий поділ за допомогою кривих заповнення простору (Методи кривих пробілів у просторі (SFCT)). Одним із недоліків графічних методів є те, що для кожної бісекції ці методи враховують лише один вимір. Таким чином, схеми, що враховують більше розмірів, можуть забезпечити кращий розподіл [12].

Один із цих методів розташовує елементи відповідно до положення їх центрів маси вздовж кривих. Криві заповнення простору – це криві, які повністю заповнюють форми великих розмірів (наприклад, квадрат або куб). Використання таких кривих забезпечує близькість точок фігури, відповідних точкам, близьким до кривої. Після отримання списку мережевих елементів, розташованих відповідно до місця розташування на кривій, досить розділити список на необхідну кількість деталей у встановленому порядку. Метод, отриманий в результаті такого підходу, в літературі називається технікою кривої заповнення простору [10].

Криві, що заповнюють простір, бувають декількох форм: Peano, Hilbert, Z-order. Криві між собою відрізняються порядком з'єднання осередків [13].

Алгоритм містить наступні етапи:

а) ініціалізація. На етапі ініціалізації вибирається квадрат, в який потрапляють усі наявні елементи набору. Це нульовий рівень декомпозиції простору. 4 квадратні площі поділу позначені від I до IV. I відповідає початку кривої, а IV до кінця;

б) ітеративний крок 1. Для створення наступного рівня розкладання квадрат ділиться на 4 квадратні площі. Проводиться крива через центри отриманих областей. Як і етап ініціалізації, кожен із результатів квадратів позначений мітками;

в) ітеративний крок 2. На цьому етапі мітки розміщуються в районах нових квадратів і кожен з 4 ділянок кривої змінюється. Замість відрізка, що з'єднує центр квадрата, отриманого на попередньому кроці, з сусіднім, крива буде проведена в центрах під квадратами;

d) поділ триває до тих пір, поки кожна окремо мічена область не містить не більше 1 елемента вихідного набору;

г) вирішальний крок. На основі отриманої кривої складається сортований список елементів набору. Далі список ділиться на необхідну кількість класів.

Ділення мережі за допомогою кривих, що заповнюють простір, є досить швидким методом, і отримане розділення трохи краще, ніж попередній алгоритм. Найкращий результат поділу отримується при розділенні класів, в яких залежності між вузлами відповідають їх близькості в просторі [12].

Рекурсивна поділка графіка (RGB). Для обчислення відстані використовується діаметр – значення, протилежне евклідовій відстані. Тому відстань між двома вузлами n1 і n2 d (n1, n2) дорівнює кількості ребер найкоротшого шляху, що з'єднує n1 і n2.

На першому кроці алгоритму обчислюється діаметр, потім вузли сортуються відповідно до відстані до граничного вузла. Половина вершин, які знаходяться ближче до граничного вузла, залишається в одному класі, решта належать іншому класу [14].

Рекурсивна інерціальна бісекція (RIB). Рекурсивно-інерційний метод удвічі будує основну інерційну вісь, розглядаючи елементи мережі як точкові маси. Лінія бісекцій, ортогональна до осі, дає границю найменшої довжини і є лінією розділення класу (рисунок 2.7) [14].

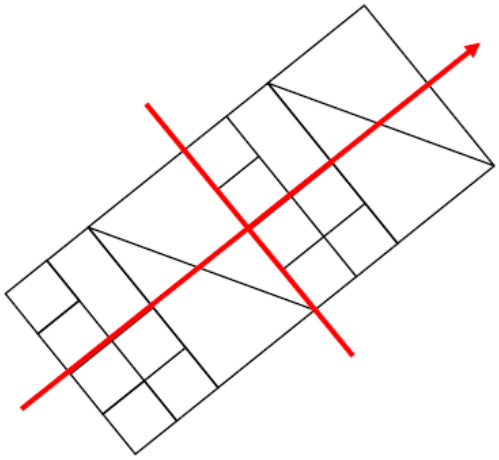


Рисунок 2.7 – Рекурсивно інерційний поділ

Модифікований рекурсивний поділ графіка (MRGB). Цей метод є модифікацією методу RGB. Це покращує оригінальний метод, виробляючи взаємопов'язані підмножини в результаті поділу.

У запропонованому способі кожен поділ на два починається з пошуку двох відносно крайніх вузлів графіка, а потім навколо них будується поділ, утворюючи два набори. По-перше, кожен набір складається з тих вузлів, які розташовані на відстані одного з'єднання від крайнього вузла, потім два і т. Д. Додавання вузлів триває до тих пір, поки один із підмножин не містить половину всіх вузлів, а решта вузлів додаються до менших один, і немає вузлів, які потрібно додати. Якщо є вільні вузли, які не пов'язані з меншим набором, то вони додаються до більшого [14].

Рекурсивна поділка кластера вузла (RNCB). Цей метод поєднує підходи модифікованого рекурсивного поділу та рекурсивної спектральної бісекції. В її основі лежить ідея вузлового кластера, який є сполученими елементами набору, об’єднаних у кластер. Такі кластери будуть вузлами в матриці Лаплаціа. Тому матриця буде меншою і метод спектральної бісекції буде працювати швидше. Основна проблема - баланс остаточного розділу.

2.2.2 Комбінаторні методи

Геометричні алгоритми з попереднього розділу мають дві стандартні вразливості: перша полягає в тому, що кожна вершина графіка має геометричні координати, і вони не завжди доступні для всіх графіків. Друга полягає в тому, що у графіку не приділяється увага структурі відносин, а перевага надається просторовій близькості, яка передбачає невелику відстань між вершинами. Хоча це і є розумне припущення, для багатьох графіків існують суперечності. Наприклад, на графіку на рисунку 2.4 вершини з різних сторін крил розташовані близько в просторі, але вільно з'єднані. Недоліки геометричних алгоритмів враховуються в структурних або комбінаторних методах [12].

Комбінаторні методи, навпаки, прагнуть групувати вершини, пов'язані між собою, не надаючи значення відстані між ними в просторі. Таким чином, ці алгоритми виконують розділення на основі лише інформації про суміжність, а не на вершинних координатах. Тому виконувані розділення мають менший роздільник і менш схильні до наявності розділених підмножин порівняно з геометричними алгоритмами. Однак комбінаторні методи є більш повільними, ніж геометричні алгоритми, і їх не можна паралелізувати [15].

Алгоритм розподілу приріст графіка (GGP). Ще один простий спосіб розділити графік навпіл – це починати з одного вузла і динамічно збільшувати область навколо цього першого вузла до тих пір, поки не буде включена половина вузлів (або половина загальної ваги вузлів). Якість алгоритму зростаючого графіка чутливий до вибору вузла, з якого починається зростання графіку – різні початкові вузли призводять до роздільників різної ваги. Щоб частково вирішити цю проблему, можна випадковим чином вибрати 10 вузлів і виростити 10 різних областей. Результат із меншою вагою сепаратора вибирається в якості остаточного поділу. Цей поділ може подаватися на вхід алгоритму KL як початкове

наближення до вдосконалення. І знову, оскільки Gm дуже малий, цей крок займає невеликий відсоток від загального часу виконання [9].

Жадібний алгоритм зростаючого графіка (GGGP). Алгоритм висхідного графа, описаний вище, збільшує розділ випадковим чином. Однак для кожного вузла v можна визначити коефіцієнт посилення ваги сепаратора, отриманого з вставки v в області росту. Таким чином, можна впорядкувати граничні вузли графіка в не зменшуваному порядку відповідно до їх вигоди. Спочатку вставляється одиниця з найбільшим зменшенням (або найменшим збільшенням) ваги сепаратора. Коли вузол вставляється у зростаючий відділ, переваги суміжних його вузлів, розташованих на кордоні, оновлюються, і вони можуть залишати кордон. Структури даних, необхідні для реалізації цієї схеми, по суті такі ж, як і ті, що вимагаються алгоритмом KL. Єдина відмінність полягає в тому, що замість обчислення всіх переваг для всіх вузлів ми виконуємо розрахунки лише для вузлів, які знаходяться на кордоні. Цей вигідний алгоритм також чутливий до вибору пускового вузла, але менше, ніж GGP. У нашій реалізації ми випадковим чином вибираємо 5-6 вузлів як вихідні точки алгоритму і вибираємо розділення з меншою вагою сепаратора. Експерименти показують, що GGGP вимагає трохи менше часу, ніж GGP, щоб розділити приблизний графік (тому що йому потрібно менше повторень), а початковий розщеплення, знайдене відповідно до цієї схеми, краще, ніж те, яке знайдено GGP [9].

Нівельована розрізана розріз (LND) Як правило, перегородка має кілька ріжучих країв, якщо сусідні вершини знаходяться в одному підмножині. LND намагається поєднати з'єднані вершини разом, починаючи з підмножини, яка містить лише одну вершину, а потім збільшує підмножину, додаючи сусідні вершини. Більш детально алгоритм виглядає наступним чином:

а) початкова вершина v0 вибирається і позначається цифрою 0, переважно периферійною;

б) для кожної вершини відстань до v0 обчислюється за допомогою алгоритму пошуку вперше в широті (BFS), починаючи з вершини v0;

в) коли половина вершин уже позначена, графік ділиться на два підграфа.

Щоб знизити ризик поганого вибору початкової вершини v0, можна зробити кілька спроб для різних початкових вершин (рисунок 2.8) [15].

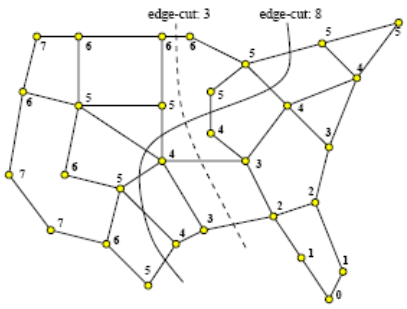


Рисунок 2.8 – Суцільною лінією показано розбиття, отримане LND-алгоритмом

Недоліком цього алгоритму є те, що отриманий таким чином розділ сильно залежить від початкової обраної вершини. Тому можна зробити висновок, що цей алгоритм слід доповнити методами знаходження початкової вершини [12].

Розсічення насінництва (SGB). На початку цього алгоритму з сукупності вершин графіка вибираються 2 суміжних рівних підмножини. Вибрані підмножини будуть вихідними підмножинами насіння. Далі, поступово від решти вершин, по одному додається по черзі до кожної їх підмножини. На кожному кроці для додавання вибирається вершина, найближча до підмножини.

Алгоритм SGB не ефективний, але оскільки він приблизно в 5 разів швидший за алгоритм KL, його можна викликати в кілька разів і успішно застосовувати як частину глобального алгоритму.

3 КРИТЕРІЇ ЯКОСТІ РЕЗУЛЬТАТІВ КЛАСТЕРИЗАЦІЇ

3.1 Опис досліджуваних статистичних характеристик вибірки вхідних даних

Вибірка або вибіркова сукупність – це сукупність випадків (предметів, об'єктів, подій, зразків), використовуючи конкретну процедуру, обрану із загальної сукупності для участі у дослідженні.

Основні з них – очікування, дисперсія, асиметрія та надлишок, режим та медіана.

Нехай X – випадкова величина, (x1, x2, ..., xn) можливі значення випадкової величини (p1, p2, ..., pn), відповідні їм ймовірності. Тоді $\gamma_k = \sum_i x_i^k p_i$, де k = 1, 2, ..., n, називається початковим моментом, а число $\mu_k = \sum_i (x_i - \xi)^k p_i$ – центральним моментом випадкової величини X – центр розподілу.

Перший початковий момент називається математичним очікуванням випадкової величини і зазвичай позначається як M (x):

$$\gamma_1 = \xi = \sum_i x_i p_i = M(x). \tag{3.1}$$

Математичне очікування визначає положення центру розподілу сукупності – тобто деяке середнє значення, навколо якого зосереджені всі можливі значення випадкової величини.

На практиці при роботі з обмеженими масивами даних замість математичного очікування використовується середнє значення вибірки.

В якості однієї з числових характеристик статистичної вибірки використовується вектор математичних очікувань ознак, що дозволяє отримати центрированний статистичний зразок Xo з елементами:

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij},\ j \equiv \overline{1,m}, \tag{3.2}$$

$$x_{ij}^0 = x_{ij} - \overline{x}_j, i = \overline{1,n},\ j = \overline{i,m}. \tag{3.3}$$

Другий центральний момент називається дисперсією випадкової величини і служить мірою її розсіювання для вибірок:

$$\mu_2 = D = \sum\nolimits_i (x_i - \xi)^2 p_i, \tag{3.4}$$

$$\mu_2 = D = \sum\nolimits_i (x_i - \overline{x})^2 p_i. \tag{3.5}$$

Дисперсія характеризує поширення випадкової величини поблизу математичного очікування.

Дисперсія має розмірність квадрата розмірності випадкової величини; Тому для отримання характеристики розсіювання з розмірністю випадкової величини часто використовується не дисперсія як показник дисперсії, а стандартне відхилення σ, обчислене як квадратний корінь дисперсії:

$$\sigma = \sqrt{\mu_2} = \sqrt{D} = \sqrt{\sum\nolimits_i (x_i - \overline{x})^2 p_i}. \tag{3.6}$$

Третій центральний момент служить в якості характеристики асиметрії розподілу:

$$\mu_3 = D = \sum\nolimits_i (x_i - \xi)^3 p_i. \tag{3.7}$$

Щоб отримати безрозмірну величину, замість μ3 вводять коефіцієнт асиметрії:

$$As = \frac{\mu_2}{\sigma^3} = \frac{\sum_i (x_i - \xi)^2 p_i}{\sigma^3}. \quad (3.8)$$

Асиметрія характеризує неоднакову повторюваність метеорологічного значення відносно середнього значення або нерівномірність часових інтервалів, протягом яких це метеорологічне значення вище середнього або нижче середнього.

Четвертий центральний момент використовується для оцінки крутості кривої розподілу порівняно із звичайною кривою розподілу. Коефіцієнт Ексцесу:

$$As = \frac{\mu_4}{\sigma^4} == \frac{\sum_i (x_i - \xi)^4 p_i}{\sigma^4}. \quad (3.9)$$

Ексцес характеризує гостровершинності (крутість, "пікоподібне") розподілу. Розмах – це різниця між найбільшим і найменшим значеннями ряду даних. Вибірковий коефіцієнт кореляції знаходиться за формулою:

$$r(X,Y) = \frac{k(X,Y)}{\sigma_x^4 \cdot \sigma_y^4} = \frac{\sum n_{xy} xy - x^4 y^4}{n\sigma_x^4 \cdot \sigma_y^4}. \quad (3.10)$$

де σ_x^4, σ_y^4 – вибіркові середні квадратичні відхилення величин X іY. Для розрахунків вручну використовується перетворена формула:

$$r_{xy} = \frac{n\sum(x_i \cdot y_i) - \sum x_i \cdot \sum y_i}{\sqrt{(n\sum x_i^2 - (\sum x_i)^2) \cdot (n\sum y_i^2 - (\sum y_i)^2)}}. \quad (3.11)$$

Коефіцієнт кореляції вибірки r (X, Y) показує щільність лінійної залежності між X і Y: чим ближче | r (X, Y) | до єдності, тим сильніша лінійна залежність між X і Y [16].

Під час експерименту була додана обчислена характеристика, яка представляє максимальну відстань між з'єднаними компонентами та кількістю елементів у з'єднаному компоненті. Ця характеристика обчислюється так:

$$SetDist = \max\left(\frac{dist(avComponent_i, avComponent_j)}{\max\left(\frac{\max ComponentOstovEdge_{ij}}{ComponentVertexNum_{ij}}\right)}\right), \quad (3.12)$$

де avComponent є центроїдом підключеного компонента;

ComponentOstovEdge – це край, що з'єднує вершини, що належать до одного компонента;

ComponentVertexNum – кількість вершин у компоненті.

Ця характеристика не є трудомісткою в обчисленні і існує взаємозв'язок між нею і значенням k при побудові графіка найближчих сусідів.

Оскільки побудова математичної моделі вимагає характеристики всієї вибірки, а не лише окремих ознак, середня, максимальна та мінімальна величини для вибірки будуть обчислюватися для кожної статистичної характеристики.

3.2 Критерії оцінки якості кластеризації

Основними аспектами оцінки є ефективність, надійність, простота та ефективність. Розрахунок часу проводився на 1,9 ГГц Intel Core i3 подвійний процесор з 6 Гб пам'яті.

Час виконання різних графіків може бути використаний для вивчення масштабованості алгоритму. Продуктивність залежить від кількості вершин і ребер.

Оцінка результатів огрублення графіка. Оцінка результатів базується на зменшенні коефіцієнта якості огрубіння та відновлення. Тільки в цій комбінації можна оцінити якість застосування алгоритмів грубого згортання в різних комбінаціях з алгоритмами відновлення.

Оцінка результатів кластеризації. Процес оцінки результатів алгоритму кластеризації називається оцінкою достовірності кластера. Існує два критерії вимірювання якості кластеризації, оцінки та вибору оптимальної схеми кластеризації:

а) компактність: елементи в кожній групі повинні бути максимально наближені один до одного. Мірилом компактності є дисперсність;

б) розділення: кластери повинні бути максимально віддалені один від одного. Існує три загальні підходи до вимірювання відстані між двома різними кластерами: відстань між найближчими членами кластера, відстань між найвіддаленішими членами та відстань між центрами кластерів.

Існує три різні методи оцінки результатів алгоритмів кластеризації:

а) зовнішній критерій (зовнішні критерії);

б) внутрішній критерій (внутрішні критерії);

в) відносний критерій (відносні критерії).

І внутрішні, і зовнішні критерії базуються на статистичних методах, і вони мають високе розрахункове значення. Зовнішні методи оцінки ефектів кластеризації базуються на певній інтуїції користувача. Внутрішні критерії базуються на показниках, заснованих на наборі даних та схемі кластеризації. Основним недоліком цих двох методів є їх обчислювальна складність.

Основою відносних критеріїв є порівняння різних схем кластеризації. Один або кілька алгоритмів кластеризації виконуються кілька разів з різними вхідними параметрами в одному наборі даних. Мета відносних критеріїв – вибрати найкращу схему кластеризації з різних

результатів. Основою для порівняння є показник валідності. Існує кілька індексів дійсності.

3.3 Створення експериментальних зразків

Для перевірки працездатності методу потрібна велика кількість зразків з різним розподілом у загальній сукупності. Відсутність реального джерела даних необхідного обсягу, різноманітності та якості змушує звернутися до альтернативного джерела [17]. Оскільки використання різних вхідних даних з певними статистичними характеристиками, продуктивність та якість кластеризації можуть сильно відрізнятися, необхідно проаналізувати синтетичні вибірки, створені спеціально для цього завдання [18]. Досліджень у цій галузі мало, і все надзвичайно специфічно для розглянутих проблем.

Існує ряд методів генерації експериментальних даних, які дозволяють аналізувати кластеризацію систематично та послідовно. Такі генератори використовують параметризовані моделі, що створюють реалістичні дані. Ці генератори навчаються за реальними даними.

Helmers and Bunke (2003) розроблені для роботи зі зразками рукописного тексту. Бейрд (2000) та Беард (1993) працювали з образами. Роджерс та ін. (2003) працював з 2D зображеннями білка. Давидов та ін. (2004) отримав набори даних, що позначають текстовий вміст від WWW. (Srikant, 1999), GSTD (Theodoridis et al., 1999) та Jeske et al. (2005) також займався створенням синтетичних даних. GSTD імітує броунівський рух. Існує ряд прихованих моделей Маркова (HMM) на основі генераторів даних. Рачковський та Куссул (1998) продемонстрували більш загальний алгоритм для формування шаблонів з особливостей у просторі, включаючи фоновий шум. Пей і Заяне (2006). Займається отриманням даних для безконтрольного навчання та виявлення поза. Van der Walt та

Bernard (2007) демонструють корисність синтетичних генераторів набору даних на основі різної щільності [19].

Жоден із перерахованих методів не дозволяє генерувати набори візуальних даних із певними статистичними характеристиками вибірки.

У статті Vineet Chaoji, Mohammad Al Hasan, Saeed Salem та Mohammed J. Zaki «SPARCL: Ефективна та ефективна кластерна форма» [20] для тестів на масштабованість, а також для створення 3D-даних, заснованого на власному генераторі кластерів на фігурах. Щоб створити фігуру в 2D, точки на полотні були вибрані випадковим чином і додані точки, що утворюють потрібні фігури. Опорною точкою для всіх фігур була точка (0,0).

Для отримання складних фігур використовувались фігури, отримані за допомогою обертання та зміщення (коло, прямокутник, еліпс, кругові смуги тощо). Генерація 3d фігур базується на 2d фігурах. Такий підхід дозволяє побудувати правдоподібну 3d фігуру, а не лише кілька шарів 2d фігури. Як і у випадку 2d, обертання та зміщення поєднуються для отримання 3d фігури. Після створення всіх форм шум додається випадковим чином (від 1% до 2%). Набір даних, показаний на малюнку 3d, має 100 000 точок та 10 кластерів (рисунок 3.1) [20].

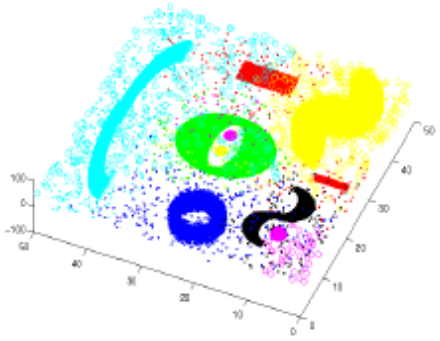


Рисунок 3.1 – 3d фігури, отримані обертанням і зміщенням

У цій роботі створення 3D фігур виконується за допомогою 3d s max studio. Ця програма дозволяє генерувати тривимірну фігуру необхідної щільності та з необхідною кількістю балів. Далі фігуру можна експортувати. Статистична характеристика отриманої вибірки буде залежати від характеру фігур, їх розміру, щільності та місця розташування. Ці параметри вибираються при створенні фігур. Шум додають до зразка безпосередньо перед аналізом.

Для проведення експерименту за цим методом було створено 130 зразків з різними статистичними характеристиками та відсотком шуму.

Згенеровані зразки є досить простим і точним способом проведення експерименту над великою кількістю зразків з відомими структурними характеристиками. Серед недоліків можна перерахувати:

а) залежність згенерованих зразків від програми генератора. Щільність і кількість вершин можуть бути різними навіть при однакових параметрах для генерації вибірки;

б) цей спосіб створення вибірки дозволяє створити набір з певними характеристиками, але такі зразки не завжди можуть мати необхідну структуру. У деяких предметних областях все ще важко відтворити структуру реальних даних під час вибірки.

Велика кількість реальних зразків була використана в різних роботах. Багато зразків не підходять для дослідження в цій роботі за кількістю ознак або іншими характеристиками (таблиця 3.1).

Таблиця 3.1 – Посилання на ресурси з наборами даних для кластеризації

Назва ресурсу	Посилання на ресурс
1	2
People	http://people.sc.fsu.edu/~jburkardt/datasets/datasets.html
Weka	http://weka.wikispaces.com/Datasets

Продовження таблиці 3.1

1	2
Cologne University	http://www.uni-koeln.de/themen/statistik/data/index.e.html
Standard datasets	http://cs.joensuu.fi/sipu/datasets/
D Star	http://uisacad2.uis.edu/dstar/data/clusteringdata.html
UCI KDD	http://kdd.ics.uci.edu/

У третьому розділі розроблена модель даних для аналізу вхідних параметрів вибірки до запуску алгоритму. Наведено загальні положення моделі даних. Виділено та описано характеристики зразків вхідних даних, які необхідно враховувати при побудові моделі алгоритму кластеризації, і як вхідні параметри моделі під час кластеризації за допомогою модифікованого алгоритму Хамелеона. Огляд можливостей отримання даних для аналізу.

4 АНАЛІЗ ТА ПОБУДОВА МАТЕМАТИЧНИХ МОДЕЛЕЙ. РЕЗУЛЬТАТИ ТЕСТУВАННЯ НА ЕКСПЕРИМЕНТАЛЬНИХ ТА РЕАЛЬНИХ ДАНИХ

4.1 Побудова математичної моделі

Математичне моделювання – це метод вивчення процесів чи явищ шляхом побудови їх математичних моделей та дослідження цих моделей. Цей вид досліджень поділяється на:

а) аналітичні дослідження впливу змін зовнішніх умов на вихідні ефекти об’єкта дослідження, коли можна явно записати залежність $\bar{Y} = \mathrm{f}(\bar{y}, \bar{v}, \mathrm{t})$, що описує взаємозв'язок зовнішніх умов та контрольні дії з результатами об'єкта дослідження, а також за допомогою аналітичних перетворень (у тому числі з використанням чисельних методів) шукається потрібне рішення;

б) імітаційне математичне моделювання, коли експериментальні дослідження математичної моделі об’єкта дослідження виконуються з відтворенням бажаних режимів роботи об’єкта шляхом імітації сигналів зовнішніх впливів.

Математичною моделлю називають формальну систему, яка являє собою скінченну сукупність символів і абсолютно точні правила роботи з цими символами в поєднанні з інтерпретацією властивостей певного об'єкта певними символами, відношеннями і константами. «Математична модель – це система математичних зв’язків, що описують процес чи явище, що вивчається» [17].

Структура моделі – це тип залежності між вхідними та вихідними ефектами об'єкта дослідження. Параметри моделі - коефіцієнти залежності. Математичні моделі класифікуються наступним чином (таблиця 4.1).

Таблиця 4.1 – Класифікація математичних моделей

Ознаки класифікації	Види матмоделей	
За залежностю структури і параметрів моделі від величини вхідних і вихідних об'єктів	Лінійні	Нелінійні
За наявністю випадкових неконтрольованих факторів	Детерміновані	Стохастичні
За характером зміни виходів об'єкта вивчення при зміні його входів	Статичні	Динамічні
За пристосованістю до умов, що змінюються властивостями об'єкта	Неадаптивні	Адаптивні

Також іноді математики класифікуються за методом побудови:

а) теоретичні або моделі внутрішнього механізму (засновані на фундаментальних законах природи);

б) функціональні або кібернетичні (описують лише зовнішню, поведінкову сторону функціонування досліджуваного явища, об'єкта і ґрунтуються на результатах експериментальних досліджень об’єктів);

в) комбіновані поєднують теоретичні кібернетичні моделі.

Аналіз ситуації дозволяє виявити основні типи параметрів, які описують стан системи: керований, цільовий та некерований.

Керовані параметри – це бажані параметри і їх значення визначають стратегію.

Цілі необхідні для опису ваших цілей. Значення цільових параметрів залежать від керованих параметрів.

Значення неконтрольованих параметрів управління не може змінюватись, залишаючись постійними, відомими повністю або частково.

Ми вважаємо, що залежності між параметрами задаються у вигляді наступного набору функцій:

$$W_i = F(X1,X2,...,Xn, a1,a2,...,ak),\ i=(1,m), \qquad (4.1)$$

де W – позначення цільових параметрів;

X – позначення керованих параметрів, а позначення некерованих параметрів;

m – кількість цільових параметрів;

n – кількість керованих параметрів;

k – кількість неконтрольованих параметрів.

4.2 Математична модель залежності вибору параметра k при побудові графіка k-nn від початкових характеристик вибірки

Для оптимізації вибору вихідного параметра k при побудові графіка k-nn необхідно побудувати математичну модель залежності k від характеристик оброблюваної вибірки. Математична модель буде побудована на основі вивчення вищезазначених зразків.

Побудова математичної моделі проводилася на наборі експериментальних зразків. Набір зразків складається з 132 зразків, серед них 33 унікальні зразки та 3 варіації кожного з них, отримані додаванням шуму 20%, 40% та 60%. Експеримент також проводився на наборах експериментальних та реальних зразків, отриманих на ресурсах обміну наборами даних.

Метою цих експериментів був вибір контрольованих параметрів цієї моделі залежності, які можуть відображати необхідні характеристики вибірки даних. У рамках роботи було проведено 3 експерименти для вибору контрольованих параметрів:

а) у першому експерименті були проаналізовані наступні характеристики: кількість об'єктів у вибірці, мінімальні та максимальні значення очікування, дисперсії та дисперсії. Залежності між цими параметрами та значенням k не виявлені;

б) у другому експерименті в якості контрольованого параметра були обрані довжина найбільшого обертового краю повністю пов'язаного

графіка та середня довжина всіх інших ребер скелета. Ці характеристики показують залежність, але використання цього підходу не є доцільним через складність побудови скелета повністю пов'язаного графіка;

в) в третьому експерименті в якості характеристики використовували кількість з'єднаних компонентів, максимальну відстань між з'єднаними компонентами та кількість елементів у з'єднаному компоненті. В результаті дослідження була побудована математична модель для оптимізації вибору початкового значення k при побудові асиметричного графіка k-nn. Модель асиметричного графіка k-nn має такий вигляд:

$$k = a + b \cdot x_1 + c \cdot x_2 + d \cdot x_1^2 + e \cdot x_2^2 + f \cdot x_1 \cdot x_2 + g \cdot x_1^3 + h \cdot x_2^3 + i \cdot x_1 \cdot x_2^2 + j \cdot x_1^2 \cdot x_2 \ , \qquad (4.2)$$

де x1 – коефіцієнт відстані;

x2 – кількість підключених компонентів. Значення коефіцієнтів представлені (таблиця 4.2).

Таблиця 4.2 – Значення коефіцієнтів моделі для визначення k у графі k-nn.

№ п/п	Коефіцієнт	№ п/п	Коефіцієнт
α	4,963024	f	4,18E-04
b	2,33E-02	g	1,05E-08
c	0,42939	h	1,14E-05
d	-4,45E-05	i	1,19E-05
e	-3,86E-03	j	-4,73E-07

Якість побудованої моделі можна судити на основі таких характеристик:

а) стандартна похибка оцінки становить 11,2986020522291;

б) коефіцієнт кратного визначення дорівнює 0,6452864929;

в) статистика Дубліна-Уотсона становить 1,224157318003058.

Графічне представлення опису даних математичною моделлю асиметричного графіка k-nn представлена (рисунок 4.1).

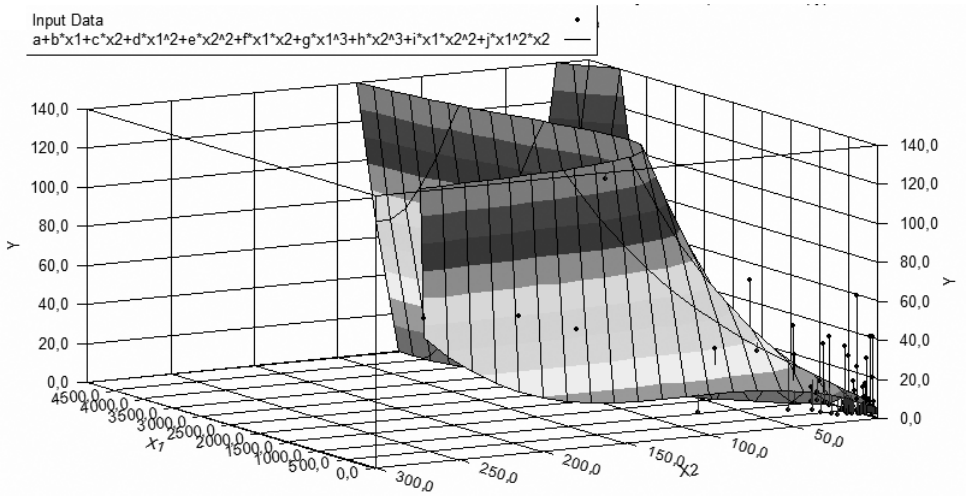


Рисунок 4.1 – Графічне представлення опису даних математичною моделлю

Залишки під час побудови цієї моделі представлені (рисунок 4.2).

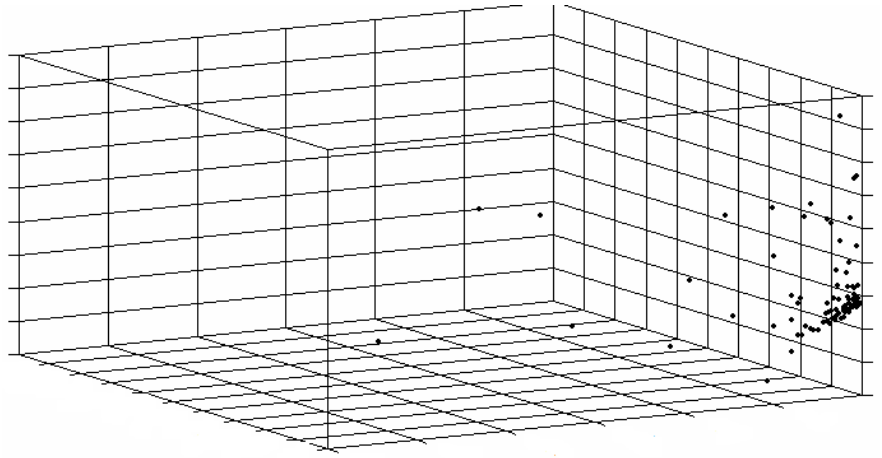


Рисунок 4.2 – Графічне представлення залишків

Оцінка та статистика якості цієї моделі не є залишковими показниками ефективності застосування отриманої моделі, оскільки

модель є лише одним із етапів вибору k. Застосування підходу було вивчено у 285 пробах. Використання цієї моделі покращило час виконання етапу побудови графіку в 62,45% випадків. У 37,55% випадків час свинцю погіршився.

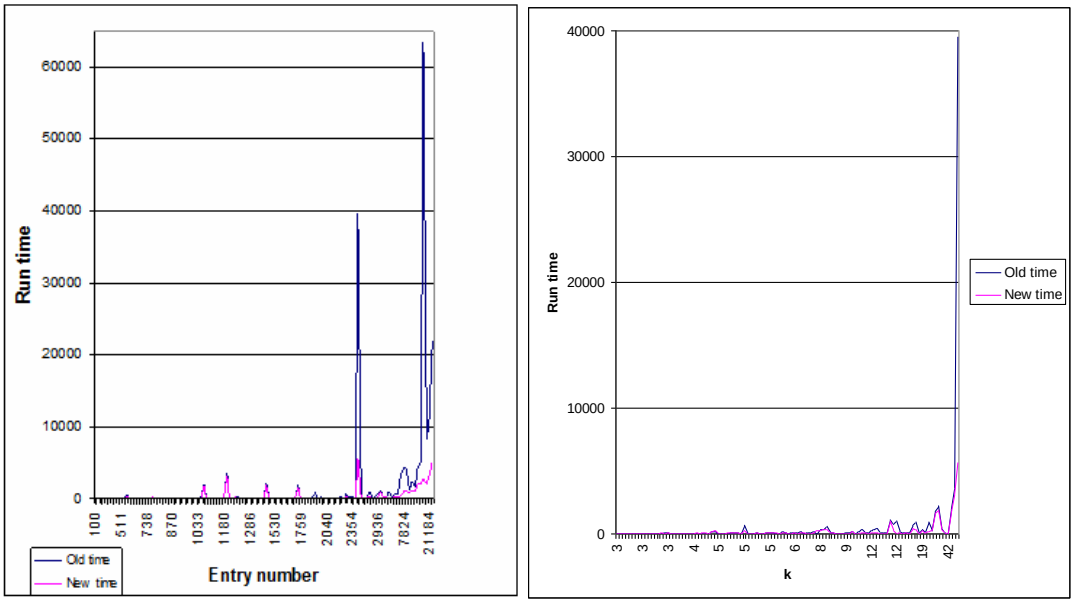
Час виконання погіршився лише у тих випадках, коли k було менше або дорівнює 3, а час виконання невеликий, тому погіршення показника часу істотно не впливає на ефективність методу в цілому. Негативний результат застосування моделі був отриманий у 7,71% випадків. В середньому тривалість виконання покращилася на 161%. Негативний результат вважається тоді, коли k значно перевищує мінімум, необхідний для задоволення умови з'єднання, навіть якщо час побудови графіка скоротився.

Порівняння часу проведення до та після нанесення моделі залежно від кількості елементів оброблюваної вибірки та від отриманого значення k показано (рисунок 4.3).

Також в результаті дослідження була побудована математична модель для оптимізації вибору початкового значення k при побудові симетричного графіка k-nn. Модель асиметричного графіка k-nn має такий вигляд:

$$k = a + b \cdot x_1 + c \cdot x_1^2 + d \cdot x_1^3 + e \cdot x_2 + f \cdot x_2^2 + g \cdot x_2^3 + h \cdot x_2^4 + i \cdot x_2^5, \quad (4.3)$$

де x1 – коефіцієнт відстані; x2 – кількість підключених компонентів. Значення коефіцієнтів представлені (таблиця 4.3).



а) б)

Рисунок 4.3 – Залежність часу побудови асиметричного графа в залежності: а) від кількості елементів вибірки для модифікованого і не модифікованого варіантів алгоритму; б) від отриманого значення k для модифікованого і не модифікованого варіантів алгоритму

Таблиця 4.3 – Значення коефіцієнтів моделі для визначення k в k-nn графі

№ п/п	Коефіцієнт	№ п/п	Коефіцієнт
α	-0,547360564	f	-3,09E-02
b	-7,46E-14	g	1,55E-04
c	1,51E-29	h	-3,34E-07
d	-6,56E-48	i	2,61E-10
e	2,323285358		

Про якість побудованої моделі можна судити, виходячи з таких характеристик:

а) стандартна помилка оцінки дорівнює +42,8805641130193;

б) коефіцієнт множинної детермінації дорівнює 0,15118817;

в) статистика Дубліна-Ватсона становить +1,26055939255469.

Побудована математична модель для оптимізації вибору початкового значення k при побудові симетричного графіка k-nn. Модель асиметричного графіка k-nn зображена (рисунок 4.5).

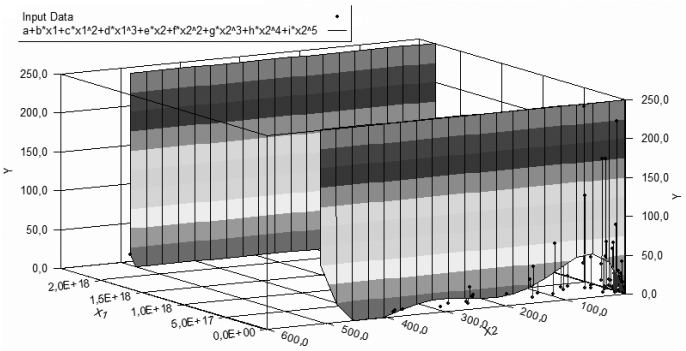


Рисунок 4.5 – Графічне представлення даних описаних математичною моделлю

Залишки при побудові даної моделі представлені (рисунок 4.6).

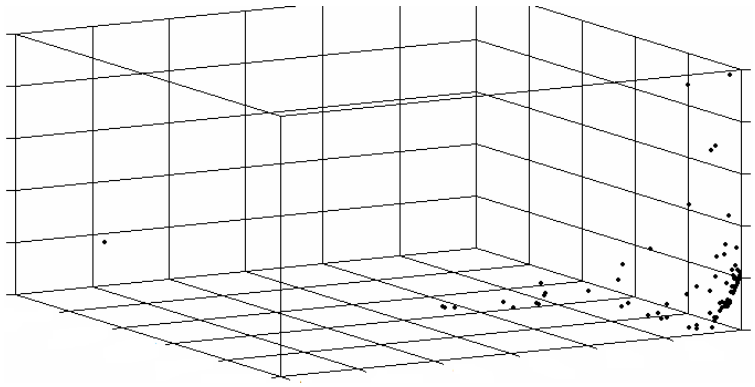
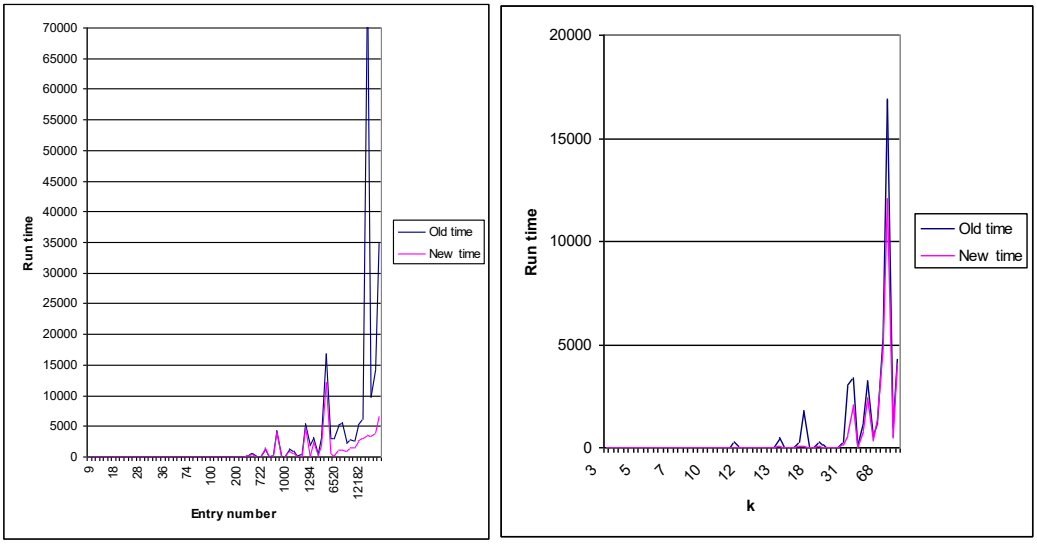


Рисунок 4.6 – Графічне представлення остатків

Застосування цієї моделі покращило час виконання етапу побудови графіку в 69,23% випадків. У 20,51% випадків час свинцю погіршився.

Негативний результат застосування моделі був отриманий у 5,12% випадків. В середньому тривалість виконання покращилася на 169%.

Порівняння часу проведення до та після нанесення моделі, залежно від кількості елементів у зразку, що обробляється, та залежно від отриманого значення k показане (рисунок 4.7).



а) б)

Рисунок 4.7 – Залежність часу побудови симетричного графа в залежності: а) від кількості елементів вибірки для модифікованого і не модифікованого варіантів алгоритму; б) від отриманого значення k для модифікованого і не модифікованого варіантів алгоритму

Використання моделі особливо важливо для великих зразків. Результати будуть використані для подальших досліджень та модифікацій алгоритму Хамелеона.

4.3 Математична модель вибору алгоритмів у рамках модифікованого алгоритму Хамелеона залежно від початкових характеристик вхідних даних

Контрольовані параметри – характеристики вибірки, такі як максимальні та мінімальні значення очікування, дисперсії, розсіювання та обчислений параметр, представлений раніше.

Складовими цільового параметра є:

а) алгоритм побудови графіка;

б) міра відстані;

в) алгоритм огрублення графіка;

г) алгоритм початкового поділу графіка;

д) міра схожості класів;

е) алгоритм відновлення графіка.

На основі доступних алгоритмів для аналізу було складено 14784 комбінацій.

Математична модель, отримана на основі цих даних, має вигляд:

$$Y = a*x1+b*x2+c*x3+d*x4+e*x5+f*x6+g*x7+h*x8+i*x9+j, \quad (4.4)$$

де x1 - x9 відповідають характеристикам зразків, на основі яких будується модель. Значення коефіцієнтів представлені в таблиці 4.4.

Таблиця 4.4 – Значення коефіцієнтів моделі вибору алгоритму в рамках модифікованого алгоритму Хамелеона

№ п/п	Коефіцієнт	№ п/п	Коефіцієнт
1	2	3	4
a	-1,23342597062584E-03	f	-3,36507290538207E-08

Продовження таблиці 4.4

1	2	3	4
b	9,6134072197013E-03	g	-4,33939843885126E-07
c	0,1397927311073	h	0,014864961283673
d	-5,23830679510672E-02	i	0,137284071634882
e	-1,14102909947611	j	2,04545274263206

Слід зазначити, що під час дослідження було виявлено недоцільним використання алгоритму побудови симетричної графіки, а асиметричний алгоритм буде використаний у подальших експериментах. Побудова моделі здійснювалася на основі результатів експериментів з розділення експериментальних зразків з використанням різних комбінацій алгоритмів в рамках модифікованого алгоритму Хамелеона.

ВИСНОВКИ

Атестаційна робота забезпечує вирішення проблеми кластеризації великих обсягів лінійно невіддільних галасливих даних, а також зразків з різними статистичними характеристиками. Вирішення цієї проблеми полягає у застосуванні модифікації алгоритму Хамелеона із поєднанням методів та алгоритмів на кожному етапі алгоритму, який найбільш підходить для вхідної вибірки на основі статистичних характеристик вибірки.

Розроблена модифікація методу дозволяє вирішити задачу кластеризації на основі індивідуального підходу до кожної вибірки і в той же час є універсальною для великої кількості різних вибірок. Створені моделі можуть скоротити час кластеризації, уніфікувати технологію кластеризації, а в деяких випадках покращити якість оцінки результатів на основі декількох методів. Усі ці результати мають важливе наукове та практичне значення в галузі кластеризації даних та обміну даними.

У ході атестаційної роботи проводилися дослідження та аналіз методів, інструментів та технологій, що використовуються в роботі з даними для кластеризації та класифікації.

У процесі дослідження в першому розділі роботи виявлено, що на даний момент існує дуже велика кількість методів кластеризації. Існуючі методи дуже різноманітні, мають велику кількість обмежень, накладених на зразки для обробки. Наведено класифікацію методів та відображено основні характеристики підходів, переваг та недоліків. Алгоритми BIRCH, CURE, CHAMELEON і ROCK вважаються найбільш актуальними при вирішенні проблеми кластеризації великих обсягів галасливих лінійно невіддільних даних. Зроблено висновок, що метод Хамелеона можна модифікувати, завдяки чому швидкість кластеризації даних зростатиме, а алгоритм стане більш універсальним.

Метою роботи була розробка універсальної математичної моделі залежності вибору комбінації алгоритмів на різних етапах алгоритму Хамелеона від початкових характеристик аналізованих наборів даних з метою підвищення якості кластеризації. Цілі дослідження були сформульовані:

а) дослідження та розробка модифікації алгоритму «Хамелеон» для кластеризації різних зразків даних наведено у другому розділі роботи бакалавра;

б) удосконалення методів роботи з графіками на окремих етапах алгоритму Хамелеона, подано у другому розділі дисертації;

в) розробка математичної моделі вибору k при побудові графіка k-найближчих сусідів на основі характеристик вибірки наведена в третьому розділі статті;

г) розробка математичної моделі вибору найкращого методу кластеризації зразків на основі характеристик вхідних даних наведена у третьому розділі статті;

д) розробка та створення програмної системи для кластеризації різних наборів даних на основі модифікованих методів наведено у четвертому розділі роботи.

Другий розділ вивчає та аналізує алгоритм Хамелеона. На основі цього методу розроблений модифікований метод кластеризації даних. Висвітлено та проаналізовано основні етапи алгоритму. Було виявлено, що цей алгоритм потрібно модифікувати, щоб можна було обробляти різні зразки різними методами на кожному етапі алгоритму, алгоритм легко розширюється, і на кожному етапі можна застосовувати різні методи, не впливаючи на інші стадії. Проаналізовано етап алгоритму Хамелеона – побудова графіка. У підрозділі описані запропоновані алгоритми модифікації цього етапу. Описані можливі параметри. Визначається область для введення нового методу визначення k при побудові графіка k-nn, що дозволяє скоротити час і витрати при виборі цього параметра.

Описані запропоновані способи модифікації грубої графіка в рамках алгоритму Хамелеона. Описано 12 запропонованих алгоритмів модифікації цього етапу. Проаналізовано етап алгоритму Хамелеона – розділення графіків. Запропоновано та описано 16 алгоритмів модифікації цього етапу. Досліджено етап відновлення та вдосконалення графіка. Модифікація алгоритму 8 пропонується додатковими методами. Проводиться аналіз об'єднання подібних класів для отримання остаточного розділу як завершального етапу модифікованого алгоритму.

У третьому розділі розроблена модель даних для аналізу вхідних параметрів вибірки до запуску алгоритму. Наведено загальні положення моделі даних. Виділено та описано характеристики зразків вхідних даних, які необхідно враховувати при побудові моделі алгоритму кластеризації, і як вхідні параметри моделі під час кластеризації за допомогою модифікованого алгоритму Хамелеона. Огляд можливостей отримання даних для аналізу.

Також були вивчені та проаналізовані методи оцінки якості поділу, отримані в результаті алгоритму. Розроблено загальну структуру підходу до оцінки результатів кластеризації на основі модифікованого алгоритму Хамелеона. Відібрано, досліджено та описано різні методи оцінки результатів кластеризації. Поєднання методів дозволяє оцінити якість поділу, виробленого алгоритмом для різних вибірок з різними вхідними характеристиками. Аналізуються вхідні дані, необхідні для експерименту та оцінки якості модифікованого алгоритму Хамелеона.

Здійснено аналіз можливих підходів до генерації експериментальних зразків. Запропоновано модель побудови 3D-зразків з необхідними характеристиками та розташуванням предметів у просторі. Проаналізовані джерела існуючі реальні та експериментальні дані. Для аналізу були відібрані існуючі зразки та для експерименту було створено набір 3d-зразків.

У четвертому розділі аналізується загальна схема побудови математичної моделі. Математична модель побудована для вибору k при побудові графіка k-nn в рамках модифікованого алгоритму Хамелеона. Представлені результати експериментів за допомогою даної моделі. Розроблена математична модель вибору алгоритмів в рамках модифікованого алгоритму Хамелеона. Зроблено висновки щодо доцільності використання окремих алгоритмів та методів. Представлені результати застосування розроблених методів до реальних даних.

В результаті роботи магістра отримано такі нові наукові результати:

а) вперше була побудована модель залежності залежності кластеризації великих обсягів складних лінійних невіддільних галасливих даних від характеристик даних та алгоритмів динамічного кластеризації;

б) була розроблена багаторівнева модель динамічного кластеризації даних, яка відрізняється від існуючої інтеграцією зі стадіями алгоритму Хамелеона. Така інтеграція дозволяє підвищити якість моделі з точки зору роботи зі складними лінійно невіддільними галасливими експериментальними даними;

в) отримав подальший розвиток метод Хамелеона, який, на відміну від існуючого, використовує різні алгоритми на різних етапах кластеризації, що дозволяє враховувати різницю даних та використовувати методи, які краще працюють на конкретних даних;

г) удосконалений метод побудови графіків, що дозволяє прискорити процес побудови графіка шляхом вибору оптимального k, виходячи з характеристик даних, що аналізуються.

Практична цінність роботи така:

а) модель вибору k для алгоритму k-найближчих сусідів створюється на основі даних даних;

б) створена модель залежності залежності кластеризації великих обсягів складних лінійно невіддільних галасливих даних від характеристик даних та алгоритмів динамічного кластеризації.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1. Ляховец А.В. Экспериментальные результаты исследования качества кластеризации разнообразных наборов данных с помощью модифицированного алгоритма Хамелеон // Вестник запорожского национального университета. 2011. №2. С. 73–86.

2. Ляховец А.В. Характеристики выборки данных для выбора k при построении графа к-ближайших соседей // Сучасні проблеми і досягнення в галузі радіотехніки, телекомунікацій та інформаційних технологій: матеріали наук.-практ. конф. з міжнар. участю. Запорожье, 2012. С. 168–169.

3. Ляховец А.В. Характеристики выборки данных для выбора k при построении графа к-ближайших соседей // Проблемы информатики и моделирования ПИМ 2012: матеріали наук.-практ. конф. з міжнар. участю. Ялта, 2012. С. 59.

4. Anil K. Jain: Data clustering: 50 years beyond K-means // Pattern Recognition Letters 31. 2010. P. 651–666.

5. Geisser M. Spinal canal size and clinical symptoms among persons diagnosed with lumbar spinal stenosis // The Clinical journal of pain. 2007. P. 780–785.

6. Dr. Thangadurai K., Uma M., Dr. Punithavalli M. A Study On Rough Clustering // Global Journal of Computer Science and Technology. 2010. № 10. P. 10–15.

7. Математические методы в археологических реконструкциях / А.П. Деревянко, Ю.П. Холюшкин, В.Т. Воронин, П.С. Ростовцев [та ін.]. Новосибирск: ИАЭТ СО РАН, 1995. 139 с.

8. Kovacs F., Legany C., Babos A. Cluster Validity Measurement Techniques // Artificial Intelligence, Knowledge Engineering, Databases: In 5th WSEAS International Conference. AIKED. 2006. P. 10.

9. Анализ данных и процессов: учеб. пособ. / А. А. Барсегян, М. С. Куприянов, И. И. Холод [и др.]. Санкт-Петербург, 2009. 512 с.

10. Мандель И. Д. Кластерный анализ / за ред. Л. В. Сергеевой. Москва, 1988. 176 с.

11. Дюран Б. Кластерный анализ / пер. с англ. Е. Демиденко. Москва: Статистика, 1999. 105 с.

12. Jain K. Algorithms for Clustering Data // Journal ACM Computing Surveys. 2012. P. 43–62.

13. Kaufman L., Rousseeuw P. J. An Introduction to Cluster Analysis // John Wiley & Sons. 2007. P. 215–266.

14. Андрейчиков А.В., Андрейчикова О.Н. Интеллектуальные информационные системы. Львов: Финансы и статистика, 2004. 424 с.