Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет _____Комп'ютерних наук_____
(повна назва)

Кафедра_____Програмної інженерії_____
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

_____другий (магістерський)_____
(рівень вищої освіти)

Дослідження методів кластеризації складних образів
(тема)

Виконав: студент 2 курсу, групи ІПЗм-17-1___
Спеціальності 121-Інженерія програмного забезпечення
(код і повна назва спеціальності)
Освітньо-наукової програми
Інженерія програмного забезпечення_____
(повна назва освітньої програми)

___Кривко Я.В._____
(прізвище, ініціали)
Керівник_____проф. Власенко Л.А._____
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри, проф. _____ З.В.Дудар

2019 р.

Харківський національний університет радіоелектроніки

Факультет _Комп'ютерних наук_____

Кафедра_Програмної інженерії_____

Рівень вищої освіти_другий (магістерський)_____

Спеціальність _121-Інженерія програмного забезпечення_____

(код і повна назва)

Освітньо-наукова програма_Інженерія програмного забезпечення_____

(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри _____

(підпис)

«____»_____ 20 ___ р.

ЗАВДАННЯ

НА АТЕСТАЦІЙНУ РОБОТУ

студентові_Кривку Ярославу Вікторовичу_____

(прізвище, ім'я, по батькові)

1. Тема роботи _Дослідження методів кластеризації складних образів_____
затверджена наказом по університету від "___" _____20 ___ р. № _____

заповнюється вручну після отримання наказу

2. Термін подання студентом роботи до екзаменаційної комісії
05 червня 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: *мета роботи, аналіз проблемної галузі і постановка задачі, 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, висновки.*

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів): *актуальність досліджень, постановка задачі, структурна схема модифікованого алгоритму Хамелеон, модифікація алгоритму Хамелеон, опис реальних і експериментальних вибірок, математична модель залежності вибору параметра k при побудові k-n графа, математична модель залежності вибору параметра k для асиметричного графа, математична модель залежності вибору параметра k для симетричного графа, висновки за експериментальними результатами роботи модифікованого алгоритму Хамелеон, висновки.*

Зворотний бік бланку завдання

6. Консультанти розділів роботи

Найменування розділу	Консультант (посада, прізвище, ім’я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Спецчастина	проф. Власенко Л.А.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка *
1.	Аналіз предметної галузі	19 квітня 2019 р.	
2.	Огляд існуючих методів	27 квітня 2019 р.	
3.	Модифікація алгоритму	15 травня 2019 р.	
4.	Підготовка пояснювальної записки	25 травня 2019 р.	
5.	Спецчастина	26 травня 2019 р.	
6.	Підготовка презентації та доповіді	28 травня 2019 р.	
7.	Попередній захист	30 травня 2019 р.	
8.	Нормоконтроль, рецензування	02 червня 2019 р.	
9.	Внесення диплому в електронний архів	03 червня 2019 р.	
10.	Допуск до захисту у зав. кафедри	04 червня 2019 р.	
* заповнюється вручну після виконання чергового пункту			

Дата видачі завдання _____ _____2019 р.

Студент_____
(підпис)

Керівник роботи_____ ___проф. Власенко Л.А._____
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ / ABSTRACT

Пояснювальна записка до атестаційної роботи: 90 с., 13 рис., 6 табл., 3 додатки, 20 джерел.

АЛГОРИТМ ХАМЕЛЕОН, ЗВ’ЯЗНІСТЬ, КЛАСТЕРИЗАЦІЯ, МЕТОД K-НАЙБЛИЖЧИХ СУСІДІВ, ПОБУДОВА ГРАФА.

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

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

Методи дослідження. У роботі використовувалися різні методи кластеризації та роботи з графами, моделювання для розробки математичної моделі вибору найкращого методу для кластеризації вибірок на підставі характеристик вхідних даних.

Практична цінність отриманих результатів. Практична цінність роботи полягає в наступному:

– створено модель вибору K для алгоритму K-найближчих сусідів, що заснована на характеристиках, даних для подальшого застосування в рамках алгоритму Хамелеон;
– створено модель залежності якості кластеризації складних лінійно нероздільних зашумленних даних від характеристик даних і алгоритмів динамічного кластеризації.

CHAMELEON ALGORITHM, CLUSTERING, CONNECTIVITY, GRAPH CONSTRUCTION, METHOD OF K-NEAREST NEIGHBORS.

Object of research – dynamic clustering process of experimental data with different characteristics.

Subject of research – methods and models of dynamic clustering of experimental data with different characteristics.

Research methods. There were used various clustering methods and methods of working with graphs, methods of modeling for the development of a mathematical model of choosing the best method for clustering of samples based on the characteristics of the input data.

The practical value of the results. The practical value of the work lies in the following:

– created a model of selecting K for K-nearest neighbors algorithm based on the characteristics of the data for further use with Chameleon algorithm;
– created a model of dependence of quality of clustering complex data on the characteristics of data and dynamic clustering algorithms.

ЗМІСТ

ВСТУП

Аналіз даних набуває все більшої і більшої значущості у сучасному світі. В різних областях людської діяльності, постійно виникає необхідність вирішення задач аналізу, прогнозу, діагностики, виявлення прихованих залежностей та підтримки прийняття оптимальних рішень. Актуальність даних завдань обумовлюється бурхливим зростанням обсягу інформації, розвитком технологій її збору, зберігання та організації в базах і сховищах даних (у тому числі інтернет-технологій), внаслідок чого точні методи аналізу інформації та моделювання досліджуваних об'єктів найчастіше відстають від потреб реального життя. На даний момент є актуальною проблема розробки універсальних і надійних методів і підходів, придатних для обробки інформації з різних областей. В якості базису для таких досліджень можуть служити технології і підходи математичної теорії розпізнавання та класифікації та математичне моделювання.

Існує безліч різних методів, які можуть бути застосовані для вирішення поставленої задачі. Також існує ряд проблем для наявних методів. Серед таких проблем можна виділити наступні:

- проблема обґрунтування якості результатів аналізу. Для різних вибірок і даних різні методи оцінювання результату можуть давати кращий результат;
- в багатьох областях, а надто в медицині, наявні дані зашумлені;
- проблема аналізу великого числа різнотипних факторів;
- нелінійність взаємозв'язків; наявність перепусток, похибок вимірювання змінних.

Так як для різних наборів даних різні методи показують найкращі результати, то для кожного окремого набору даних необхідний якийсь критерій вибору найкращого методу.

Сформулюємо актуальність досліджень в даній області:

– необхідність організувати на єдиних принципах і синхронізувати вибір методу кластеризації на підставі даних аналізованої вибірки;

– потреба уніфікувати технології кластеризації і за рахунок цього скоротити час на вибір методу;

– необхідність забезпечення користувачів якісним вирішенням завдання аналізу при різних досліджуваних даних;

– постійне збільшення обсягу надходить інформації та різнорідність цієї інформації вимагає розвитку технологій аналізу цих даних;

– окремі методи кластеризації добре працюють на відповідних вибірках, але не є універсальними;

– необхідність аналізу складних вибірок, з пересічними і класами що накладаються.

У пояснювальній записці розроблені методи і моделі вибору найкращого методу для кластеризації даних, засновані на наступних характеристиках вибірки:

– підхід до кластеризації на основі алгоритму Хамелеон;

– методи роботи з графами на окремих етапах алгоритму Хамелеон;

– математична модель вибору оптимального k при побудові графа K-найближчих сусідів;

– математична модель вибору найкращого методу для кластеризації вибірок на підставі характеристик вхідних даних.

В роботі запропоновані модифікації алгоритму динамічного кластеризації Хамелеон. Даний алгоритм складається з наступних кроків: побудова графа, огрубіння графа, поділ графа, відновлення і поліпшення графа. На кожному з даних етапів були запропоновані і реалізовані набори методів, які дозволяють покращити якість кластеризації для кожного запропонованого вихідного набору даних.

Окремі алгоритми були модифіковані для можливості їх використання і поліпшення якості кластеризації побудованої моделі.

Була побудована математична модель залежності вибору K при побудові графа K-найближчих сусідів на підставі характеристик вибірки. Дана модель дозволяє прискорити процес побудови графа. Також була побудована математична модель залежності вибору комбінації алгоритмів на різних етапах алгоритму Хамелеон від вихідних характеристик аналізованого набору даних. Ця модель дозволяє визначити, які з алгоритмів мають бути використані для вибірки даних що володіє деякими характеристиками для отримання найкращої якості кластеризації.

Сформулюємо завдання дослідження:

– проаналізувати і теоретично обґрунтувати процедуру підвищення точності кластеризації з використанням ієрархічного алгоритму Хамелеон, за рахунок вибору найбільш ефективних алгоритмів на кожному з етапів роботи;
– виявити набір характеристик вихідних даних та критеріїв порівняння алгоритмів, що вживаються в рамках алгоритму Хамелеон;
– розробити математичну модель вибору K при побудові графа K-найближчих сусідів на підставі характеристик вибірки для симетричного і асиметричного графів;
– розробити математичну модель вибору найкращого методу для кластеризації вибірок на підставі характеристик вхідних даних.

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

Предмет дослідження – методи і моделі динамічної кластеризації лінійно нероздільних експериментальних даних з різними характеристиками.

Методи дослідження. У роботі було використано різні методи кластеризації і роботи з графами, моделювання для розробки математичної моделі вибору

найкращого методу для кластеризації вибірок на підставі характеристик вхідних даних.

Наукова новизна отриманих результатів полягає в наступному:

– подальший розвиток отримав ієрархічний алгоритм Хамелеон, який відрізняється від існуючого інтеграцією існуючих алгоритмів кластеризації з ієрархічним алгоритмом Хамелеон, така модифікація алгоритму Хамелеон дозволяє поліпшити якість кластеризації при роботі зі складними лінійно нероздільними зашумленими експериментальними даними;
– удосконалено критерій порівняння алгоритмів і аналізу вихідних даних, що застосовуються в рамках алгоритму Хамелеон;

Практична цінність роботи полягає в наступному:

– створено модель вибору K для алгоритму K-найближчих сусідів, що заснована на характеристиках, даних для подальшого застосування в рамках алгоритму Хамелеон;
– створено модель залежності якості кластеризації складних лінійно нероздільних зашумленних даних від характеристик даних і алгоритмів динамічного кластеризації.

1 АНАЛІЗ ПРОБЛЕМНОЇ ГАЛУЗІ

1.1 Дослідження і аналіз методів інтелектуального аналізу даних

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

Методи Data Mining розділяються на статистичні (дескриптивний аналіз, кореляційний і регресивний аналіз, факторний аналіз, дисперсійний аналіз, компонентний аналіз, дискримінантний аналіз, аналіз тимчасових рядів) та кібернетичні (штучні нейронні мережі, еволюційне програмування, генетичні алгоритми, асоціативна пам'ять, нечітка логіка, дерева рішень, системи обробки експертних знань).

Кластерний аналіз (Data clustering) – завдання розбиття заданої вибірки об'єктів на підмножини, звані кластерами, так, щоб кожен кластер складався з схожих об'єктів, а об'єкти різних кластерів суттєво різнилися. Завдання кластеризації відноситься до статистичної обробки, а також до широкого класу завдань навчання без вчителя. Кластерний аналіз – це багатовимірна статистична процедура, що виконує збір даних, що містять інформацію щодо вибірки об'єктів, і потім впорядковує об'єкти в порівняно однорідні групи (Q кластеризація, або Q техніка, власне кластерний аналіз) [1].

Кластерний аналіз – це спосіб групування багатовимірних об'єктів, заснований на представленні результатів окремих спостережень точками підходящого геометричного простору з подальшим виділенням груп як "згустків" цих точок (кластерів, таксонів). Кластер (cluster) в англійській мові означає "згусток", "гроно винограду", "скупчення зірок". Даний метод дослідження отримав розвиток в останні роки у зв'язку з можливістю комп'ютерної обробки

великих баз даних. Кластер – група елементів, що характеризуються загальним властивістю, Головна мета кластерного аналізу – знаходження груп схожих об'єктів у вибірці [2].

Кластерний аналіз передбачає виділення компактних, віддалених один від одного груп об'єктів, відшукує "природне" розбиття сукупності на області скупчення об'єктів. Вона використовується, коли вихідні дані представлені у вигляді матриць близькості або відстаней між об'єктами або у вигляді точок у багатовимірному просторі [3]. Найбільш поширені дані другого виду, для яких кластерний аналіз орієнтований на виділення деяких геометрично видалених груп, всередині яких об'єкти близькі [2].

Спектр застосувань кластерного аналізу дуже широкий: його використовують в археології, медицині, психології, хімії, біології, державному управлінні, філології, антропології, маркетингу, соціології та інших дисциплінах. Однак універсальність застосування призвела до появи великої кількості несумісних термінів, методів і підходів, що утруднюють однозначне використання і несуперечливу інтерпретацію кластерного аналізу.

Можна виділити такі завдання кластерного аналізу:

– розробка типології або класифікації;
– дослідження корисних концептуальних схем групування об'єктів;
– породження гіпотез на основі дослідження даних;
– перевірка гіпотез або дослідження для визначення, чи дійсно типи (групи), виділені тим чи іншим способом, присутні в наявних даних.

Незалежно від предмета вивчення застосування кластерного аналізу передбачає наступні етапи:

– відбір вибірки для кластеризації;
– визначення безлічі змінних, за якими будуть оцінюватися об'єкти у вибірці;
– обчислення значень тієї чи іншої міри подібностей між об'єктами;

– застосування методу кластерного аналізу для створення груп схожих об'єктів;
– перевірка достовірності результатів кластерного рішення.

Кластерний аналіз пред'являє наступні вимоги до даних:

– показники не повинні корелювати між собою;
– показники мають бути безрозмірними;
– їх розподіл має бути близьким до нормального;
– показники повинні відповідати вимозі стійкості, під якою розуміється відсутність впливу на їх значення випадкових факторів;
– вибірка має бути однорідна.

Якщо кластерному аналізу передує Факторний аналіз, то вибірка не потребує втручання – викладені вимоги виконуються автоматично самою процедурою факторного моделювання (якщо проводити Z-стандартизацію без негативних наслідків для вибірки безпосередньо для кластерного аналізу, вона може спричинити зменшення чіткості поділу груп). В іншому випадку вибірку потрібно коригувати.

Мета кластеризації:

– розуміння даних шляхом виявлення кластерної структури. Розбиття вибірки на групи схожих об'єктів дозволяє спростити подальшу обробку даних та прийняття рішень, застосовуючи до кожного кластеру свій метод аналізу (стратегія розділяй і володарюй);
– стиснення даних. Якщо вихідна вибірка надлишково велика, то можна скоротити її, залишивши по одному найбільш типовому представнику від кожного кластеру;
– виявлення новизни (novelty detection). Виділяються нетипові об'єкти, які не вдається приєднати до жодного з кластерів.

В першому випадку число кластерів намагаються зробити меншим. У другому випадку важливіше забезпечити високу ступінь схожості об'єктів

всередині кожного кластеру, а кластерів може бути скільки завгодно. У третьому випадку найбільший інтерес представляють окремі об'єкти, які не вписуються ні в один з кластерів.

У всіх цих випадках може застосовуватися ієрархічна кластеризація, коли великі кластери дробляться на більш дрібні, ті в свою чергу дробляться ще на більш дрібні. Такі завдання називаються завданнями таксономії.

Результатом таксономії є деревоподібна ієрархічна структура. При цьому кожен об'єкт характеризується перерахуванням всіх кластерів, до яких він належить, зазвичай від великого до дрібного.

Кластеризація (навчання без вчителя) відрізняється від класифікації (навчання з учителем) тим, що мітки вихідних об'єктів спочатку не задані, і навіть може бути невідомою їх множина.

Вирішення завдання кластеризації принципово неоднозначно, і тому є кілька причин:

– не існує однозначно найкращого критерію якості кластеризації. Відомий цілий ряд евристичних критеріїв, а також ряд алгоритмів, які не мають чітко вираженого критерію, але які здійснюють достатньо розумну кластеризацію з побудови. Всі вони можуть давати різні результати;
– число кластерів, як правило, невідомо заздалегідь і встановлюється відповідно з деяким суб'єктивним критерієм;
– 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. Мається навчальна вибірка $X\ell = \{x_1, ..., x_\ell\} \in X$ і функція відстані між об'єктами $\rho(x, x')$. Потрібно розбити вибірку на непересічні підмножини, звані кластерами, так, щоб кожен кластер складався з об'єктів, близьких по метриці ρ, а об'єкти різних кластерів суттєво різнилися. При цьому кожному об'єкту $x_i \in X\ell$ приписується мітка (число) кластера y_i. Алгоритм кластеризації – це функція a: X

→ Y, яка будь-якому об'єкту x є X ставить у відповідність мітку кластера y є Y. Множина міток Y в деяких випадках відомо заздалегідь, однак частіше ставиться завдання визначити оптимальне число кластерів, з точки зору того чи іншого критерію якості кластеризації [2].

Вирішенням завдання кластерного аналізу є розбиття, що задовольняє деяку умову оптимальності. Цей критерій може являти собою деякий функціонал, що виражає рівні бажаності різних сегментів і угруповань. Цей функціонал часто називають цільовою функцією. Завданням кластерного аналізу є задача оптимізації, тобто знаходження мінімуму цільової функції при деякому заданому наборі обмежень. Прикладом цільової функції може служити, зокрема, сума квадратів внутрішньогрупових відхилень за всіма кластерами [2].

Можна виділити наступні основні етапи кластерного аналізу.

Формування системи змінних. Нерідко вимагається попередньо вибрати з вхідної множини змінних найбільш ефективну підсистему (у зарубіжній літературі цей процес називається feature selection). Крім того, у деяких завданнях доцільно трансформувати вихідні змінні так, щоб Утворити нові, більш інформативні показники (feature extraction). Щоб уникнути домінування змінних з більшим масштабом вимірювання, проводять попереднє нормування вихідних змінних.

Визначення способу обчислення відстані між об 'єктами або групами об'єктів. Цей спосіб повинен відображати специфіку розв'язуваної прикладної задачі. Наприклад, у випадку безперервних змінних може бути задано Евклідову відстань. Щоб виключити вплив сильних лінійних кореляцій між змінними, застосовують відстань Махаланобіса. Для номінальних змінних може використовуватися відстань Хеммінга. Для груп об'єктів також визначається спосіб знаходження відстані, наприклад, за принципом далекого сусіда, ближнього сусіда та інші. Принцип далекого сусіда виправданий в випадку, коли є завжди апріорна інформація про те, що таксони мають компактну сферичну форму. Принцип ближнього сусіда має сенс застосовувати, якщо відомо, що таксони можуть мати витягнуту форму або концентрично розташовані.

Групування шаблонів. На цьому кроці проводиться створення груп об'єктів. Розбиття на групи може бути жорстким (формується розбиття вихідного багатьох об'єктів), а може бути і нечітким (розрахувати ступінь приналежності кожного об'єкта до груп). Існує велика різноманіття алгоритмів угруповання.

Подання результатів. Потрібно дістати простий і інформативний опис отриманих кластерів. Часто для такого опису вибирається "типовий об'єкт" або визначається набір усереднених по групі показників. Використовується також опис у вигляді набору таксонів. Під таксоном будемо розуміти частину простору змінних мінімального обсягу, що має деяку задану форму і що містить точки відповідної групи.

Визначення якості отриманої угруповання. Після завершення кластеризації необхідно впевнитися в тому, що сформовані групи дійсно відображають внутрішні закономірності, характерні для розв'язуваного завдання, сприяють досягненню цілей аналізу, допомагають відкрити нові властивості що вивчаються об'єктів. Існують також більш формальні способи перевірки якості, пов'язані з перебуванням ймовірності випадкового освіти груп, яку можна обчислити в рамках тієї чи іншої моделі розподілу (з перевіркою статистичних гіпотез про однорідності спостережень різних класів), з bootstrap методом, з обчисленням різних показників якості (внутрішньогрупового розподілення, Гудмана та Крускала та інші [4].

1.2 Основні підходи, які використовуються у кластерному аналізі.

Зараз існують декілька підходів до вирішення завдання кластерного аналізу, які засновані на різних уявленнях про завдання, використанні специфічної для кожної предметної області додаткової інформації та інші. Коротко перелічимо підходи що використовуються найчастіше. Зауважимо, що описана нижче класифікація не є чіткою; деякі методи можуть бути розроблені на основі комбінації різних підходів:

– імовірнісний підхід. Передбачається, що кожен об'єкт Генеральної сукупності належить одному з K класів, Однак номера класів безпосередньо не визначаються. Об'єкти вибираються з генеральної сукупності випадково і незалежно, тому змінні, що описують об'єкти, випадкові. Для кожного класу визначено імовірнісний розподіл заданого сімейства; параметри розподілу невідомі. Наявна вибірка спостережень являє собою реалізацію суміші розподілів;
– підхід, який використовує аналогію з центром ваги. Для кожної групи визначається вектор середніх значень показників, інтерпретується як центр ваги групи. Використовується критерій внутрішньогрупового розподілення, де k – координата центру ваги К-го кластеру за змінною Xj, j=1,2,…,n, k=1,2,…,K. Оптимальне угруповання, при заданому K, відповідає мінімальному значенню критерію;
– підхід, заснований на теорії графів. Найбільш відомий алгоритм цього сімейства – алгоритм найкоротшого незамкнутого шляху. Попередньо будується мінімальне остове дерево графа, в якому вершини відповідають об'єктам, а ребра мають довжину, рівну відстані між відповідними об'єктами. Для утворення кластерів з побудованого дерева видаляються ребра максимальної довжини;

– ієрархічний підхід. Даний напрямок також має відношення до теоретико графів підходу. Результати угруповання представляються у вигляді дерева угруповання. Алгоритми, засновані на цьому підході, можна розділити на агломеративні (такі, що поетапно об'єднують найближчі групи чи об’єкти) та дивізимні (в яких поетапно здійснюється поділ вихідної групи на найбільш віддалені підгрупи; ті в свою чергу також розділяються на підгрупи). Групувальні рішення представляють собою вкладену ієрархію підгруп;
– підхід, заснований на понятті найближчого сусіда. Угруповання здійснюється послідовно шляхом приписування об'єкта кластеру, в якому знаходиться найближчий об’єкт, за умови, що відстань до об'єкта не перевищить заданий поріг. Існують різноманітні варіанти визначення відстані; при визначенні міри запобіжного близькості може враховуватися і розташування інших сусідніх точок;
– нечіткі алгоритми кластерного аналізу. При використанні даного підходу передбачається, що кожен кластер являє собою нечітке безліч об'єктів;
– підхід, що використовує штучні нейронні мережі, заснований на аналогією із процесами, що відбуваються в біологічних нейронних системах. Відомо велике число алгоритмів даного сімейства. Типова архітектура являє собою одношарову мережа, в якій кожен нейрон відповідає деякого кластеру. У процесі навчання мережі відбувається ітеративна зміна передавальних ваг між вхідними і вихідними вузлами мережі; тим самим здійснюється пошук оптимального значення критерію угруповання. Нейронні мережі дозволяють ефективно використовувати паралельні методи обчислень;
– еволюційний (генетичний) підхід. Алгоритми цього сімейства побудовані на аналогії з природною еволюцією. В них використовуються поняття популяції – набору різних варіантів угруповання (званих також хромосомами, за аналогією з відповідними біологічними об'єктами), і еволюційних операторів – процедур, що дозволяють з однієї або декількох

батьківських хромосом отримати одну або декілька хромосом нащадків. Цими процедурами є: селекція, рекомбінація і мутація. Генетичний алгоритм здатний здійснювати пошук рішення, що доправляє глобальний мінімум критерієм якості угруповання.

Існують також інші підходи до вирішення завдання глобальної оптимізації, які можуть застосовуватися у кластерному аналізі. Так, наприклад, відомий підхід імітації відпалу, також що базується на ідеї моделювання природних процесів.

1.3 Вимоги до алгоритму кластеризації

Кластеризація – це багатообіцяюче поле для досліджень, де області потенційного застосування диктують додаткові вимоги:

– масштабованість. Багато алгоритмів кластеризації добре працюють на маленьких вибірках даних, які складаються з менш ніж 200 об'єктів; тим не менш, великі бази даних можуть містити мільйони об'єктів. Кластеризація вибірки з великої множини може призвести до необ'єктивних результатів. Необхідні добре масштабовані алгоритми;
– можливість роботи з характеристиками різних типів. Багато алгоритми розроблені для кластеризації числових даних. Тим не менш, може бути затребувана робота з іншими типами даних: бінарних, категоріальних (номінальних) та порядкових даних або суміші цих типів даних;
– розпізнавання кластерів довільної форми. Багато алгоритми кластеризації визначають кластери, ґрунтуючись на евклідовій та мангеттенській відстанях. Алгоритми, засновані на таких видах відстаней, мають тенденцію до знаходження кластерів сферичної форми з однаковою

щільністю і однакового розміру. Однак кластер може бути будь-який форми. Важливо розробити алгоритм для знаходження кластерів довільної форми;

– мінімальні вимоги до області дослідження для визначення вхідних параметрів. Багатьом алгоритмам кластеризації необхідні певні вхідні параметри (наприклад, такі як кількість кластерів). Результати кластеризації можуть бути достатньо чутливі до вхідним параметрами. Часто параметри досить важко визначити, особливо якщо вхідна множина складається із багатовимірних об'єктів. Це не тільки вимагає втручання користувача, але і ускладнює контроль за якістю кластеризації;
– можливість роботи з зашумленими даними. Більшість реальних баз даних містить викиди або загублені, невідомі або неправдиві дані. Деякі алгоритми кластеризації чутливі до таких даних, що може призвести до поганої якості кластеризації;
– нечутливість до порядку що входять записів. Деякі алгоритми кластеризації чутливі до порядком, у такому випадку для одного і того ж безлічі, коли об'єкти представлені в різному порядку, будуть отримані зовсім різні класи;
– висока розмірність. База даних чи сховище даних може мати кілька вимірів чи атрибутів. Багато алгоритми гарні для роботи з низько razmernymi даними;
– кластеризація на основі обмежень. Реальні програми можуть накладати на кластеризацію різні типи обмежень. Багатообіцяючим завданням є знаходження груп даних з хорошим проходженням кластеризації, яка відповідає висунутим обмеженням;
– простота використання та інтерпретації. Користувач очікує інтерпретовані, порівняні і корисні результати кластеризації;
– кількість переглядів бази даних. Наявної пам'яті повинно вистачати на обробку великих множин даних високих розмірностей [3, 5, 6].

1.4 Актуальні проблеми кластерного аналізу

Незважаючи на велику кількість досліджень в області кластерного аналізу, в цій області існує низка актуальних проблем. Перерахуємо основні проблеми.

- проблема обґрунтування якості результатів аналізу. Відомо, що процес угруповання значною мірою носить суб'єктивний характер. Це виражається, зокрема, в тому, що один і той же набір об'єктів може класифікуватися по-різному залежно від прикладної області, ступеня повноти знань про об'єкти вивчення і т. Д. Тому необхідно розробляти методи, що дозволяють максимально повно враховувати наявні експертні знання, а також розробляти відповідні критерії якості угрупування;
- багато хто важко формалізуються галузі досліджень характеризуються недостатністю знань про що вивчаються об'єктах, що ускладнює формулювання їх математичних моделей;
- проблема аналізу великого числа різнотипних (кількісних або якісних) факторів. У разі різнотипного простору виникає методологічна проблема визначення в ньому метрики. З іншого боку, навіть у просторі однотипних (кількісних) змінних при збільшенні їх числа посилюється прокляття розмірності, що може призвести до майже повної непомітності точок. Так, відстань від будь-якої точки до її "найближчого сусіда" для деяких видів відстаней може практично збігатися (з урахуванням машинної точності) з відстанню до її "далекого сусіда". Зорові аналогією, доречні в просторі малої розмірності, стають абсолютно неприйнятними у просторі Великої розмірності;
- нелінійність взаємозв'язків; наявність перепусток, похибок вимірювання змінних. Класичні методи зниження розмірності (метод головних компонент; метод незалежних компонент), що використовуються в кластерному аналізі, в основному орієнтовані на лінійні залежності між

змінними. Для виявлення більш складних взаємозв'язків вимагаються такі алгоритми, як нелінійні (ядерні) методи головних компонент;

– необхідність подання результатів аналізу у формі, зрозумілою фахівцям прикладної області. Крім гарної прогнозуючої здібності для будь-якого алгоритму аналізу даних важливо, наскільки зрозумілими і інтерпретуються є його результати. Для поліпшення інтерпретуємо рішень можна використовувати логічні моделі. Такого роду моделі використовуються для вирішення завдань розпізнавання образів та прогнозування кількісних показників;
– проблема пошуку глобального екстремуму у критерію якості угруповання. Критерій якості, є функцією, залежною від Великої кількості факторів, нелінійним, таким, що володіє безліччю локальних екстремумів. Для знаходження кластерів необхідно вирішити складну комбінаторних завдання пошуку оптимального варіанту класифікації. В формулі 1.1 приведено формулу знаходження числа різних варіантів розбиття N об'єктів на K груп.

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

де N – кількість об'єктів для розподілення на групи;

K – кількість груп, на які необхідно виконати розділення.

– проблема стійкості групувальних рішень. В класичних алгоритмах вирішення завдань кластерного аналізу результати угруповання можуть сильно змінюватися залежно від вибору початкових умов, порядку об'єктів, параметрів роботи алгоритмів і т. Д. [4].

Тому алгоритм повного перебору варіантів має трудові затрати, експоненціально залежні від розмірності. Якщо число груп заздалегідь невідомо,

то завдання перебору стає ще складніше. Таким чином, при збільшенні розмірності таблиць даних відбувається "комбінаторний вибух". Класичні алгоритми кластерного аналізу здійснюють спрямований пошук у порівняно невеликому підмножині простору рішень, використовуючи різного роду апріорні обмеження. При цьому знаходження строго оптимального рішення не гарантується. Для пошуку оптимального рішення застосовуються більш складні методи, такі як генетичні (еволюційні) алгоритми, нейронні мережі. Існують експериментальні дослідження, що підтверджують переваги таких алгоритмів перед класичними алгоритмами;

На основі виконаного аналізу сформулюємо завдання наукового дослідження:

– дослідження та розробка модифікованого підходу до кластеризації на основі алгоритму Хамелеон;
– модифікація методів роботи з графами на окремих етапах алгоритму Хамелеон;
– побудова математичної моделі вибору оптимального k при побудові графа K найближчих сусідів;
– побудова математичної моделі вибору найкращого методу для кластеризації вибірок на підставі характеристик вхідних даних.

2 ПОСТАНОВКА ЗАДАЧІ ТА РОЗРОБКА МОДИФІКАЦІЇ АЛГОРИТМУ ХАМЕЛЕОН

2.1 Аналіз і опис основних етапів алгоритму Хамелеон

Проблема поділу графа – це поділ вершин цього графа на p приблизно рівних частин таким чином, щоб кількість ребер між вершинами із різних класів було мінімальним. Ця задача застосовується в багатьох різних областях, включаючи паралельні наукові обчислення або планування завдань. Проблема поділу є NP – повною. Тим не менш, велика кількість розроблених алгоритмів знаходять досить хороші результати поділу. Завдання K-Way поділу найчастіше вирішується методом рекурсивно навпіл. Останнім часом з 'явився високоефективний метод для K-Way поділу графа – багаторівнева рекурсивна бісекція (multilevel recursive bisection). Основна структура багаторівневої рекурсивної навпіл дуже проста. На початку граф G огрублюється до декількох сотень вершин, далі виконується поділ навпіл отриманого зменшеного графа, а потім цей поділ проектується назад на вихідний граф через періодичне відновлення поділу.

Багаторівнева парадигма також може бути застосована для побудови K-Way поділу прямо на вихідній множині, як показано на рисунку 2.1 [7]. Множина огрубляется послідовно, як і в попередній схемі, але тепер огрублена множина ділиться відразу на k частин і поділ послідовно відновлюється до вихідної множини. Існує ряд переваг виконання відразу k поділу. По-перше, огрубіння необхідно виробити лише одного разу, що зменшує складність алгоритму і час виконання. По-друге, добре відомо, що багаторівнева рекурсивна бісекція може працювати гірше, ніж K-Way поділ. Таким чином, метод досягнення відразу K-Way поділу може виконати завдання краще. Слід зауважити, що відразу розрахувати хороший K-Way поділ важче, ніж виконати хорошу бісекцію. Саме з цієї причини

найбільш поширене рішення задачі K-Way поділу виконується за допомогою рекурсивної бісекції [8].

На стадії огрубіння розмір множини послідовно зменшується, на стадії вихідного розбиття виконується K-Way поділ зменшеної множини (6-Way в даному випадку), на стадії відновлення виконується проектування поділу на початковий граф.

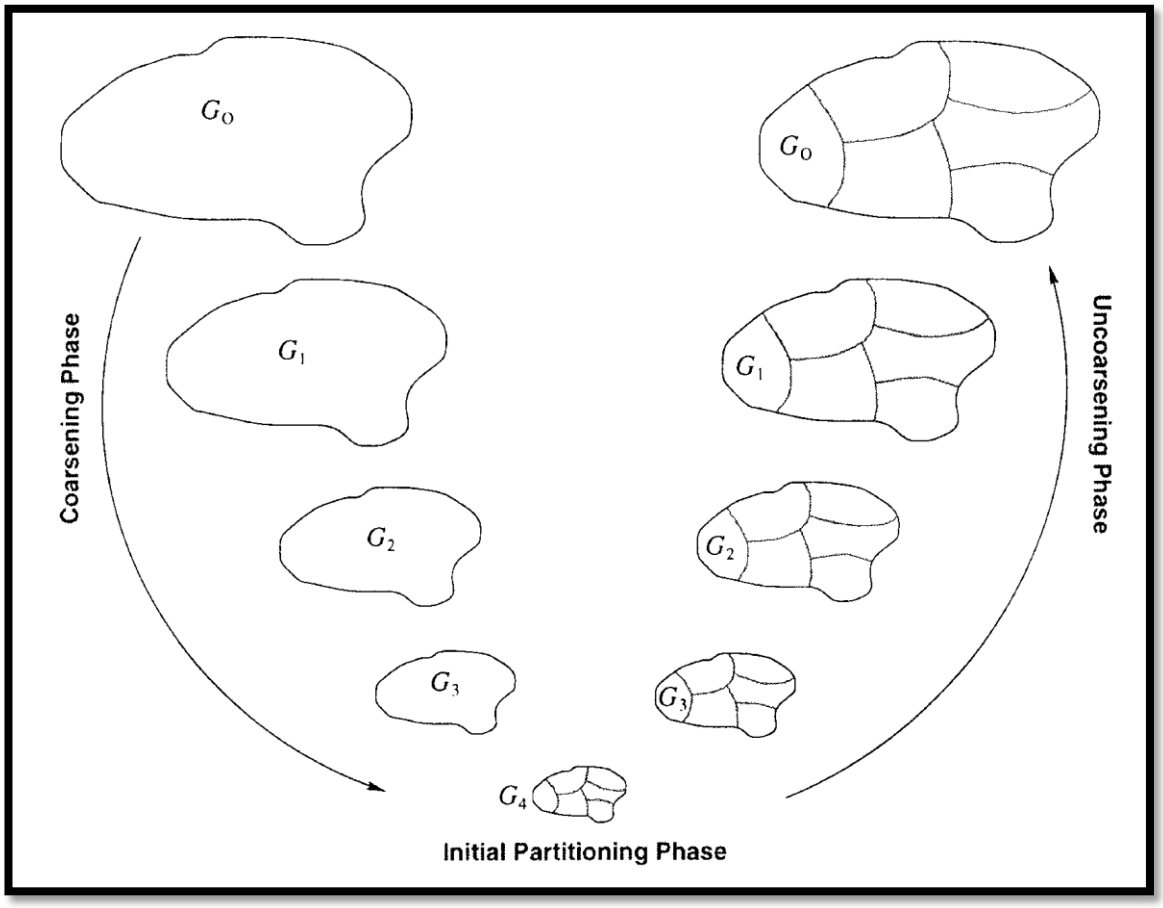

Рисунок 2.1 – Різні стадії алгоритму багаторівневого K-Way поділу

Наприклад, найпростіший метод для обчислення початкової поділу в контексті багаторівневого алгоритму це огрубіння графа до k вершин. Тим не менш, на фазі удосконалення необхідно удосконалити K-Way поділ, якій значно більш складний, ніж рекурсивна бісекція. Навіть для 8-Way поділу час, виконання

для даної схеми досить великий. Для удосконалення K-Way поділу для k > 8 час виконання стає надмірно великим.

Хамелеон – це новий ієрархічний алгоритм, який долає обмеження існуючих алгоритмів кластеризації. Даний алгоритм розглядає динамічне моделювання в ієрархічній кластеризації. Ключовим моментом в алгоритмі Хамелеон є те, що він враховує одночасно взаємопов'язаність і близькість при визначенні однакових пар кластерів. Саме це дозволяє подолати обмеження. Хамелеон використовує новий підхід при визначенні ступеня взаємопов'язаності і близькості між парами кластерів. При даному підході алгоритм сам прораховує внутрішні характеристики кластерів, Отже, вони не залежать від статичних, встановлених користувачем моделей, і можуть автоматично підлаштовуватися до внутрішніх характеристиками об 'єднаних кластерів.

Хамелеон знаходить кластери, використовуючи двофазний алгоритм. На першому кроці Хамелеон використовує алгоритм розбиття графи для кластеризації безлічі на досить маленькі підкласи. На другому кроці використовується алгоритм для знаходження природних кластерів за допомогою послідовного об 'єднання отриманих маленьких підкласів.

Хамелеон представляє об'єкти за допомогою часто використовуваного графа K найближчих сусідів (K nearest neighbor graph). Це подання даних у вигляді графа дозволяє змінити масштаб великих обсягів даних. Кожна вершина в даному графі представляє один об'єкт даних. Між вершинами існує ребро, якщо один об'єкт є одним з K найближчих сусідів другого об'єкта. Граф K найближчих сусідів містить концепцію, що радіус суміжності об’єкта визначається щільністю регіону, в якому даний об'єкт знаходиться. Воно дозволяє виявляти природничі кластери.

На наступному кроці будується черга з послідовно зменшених гіперграфів – стадія огрубіння (Coarsening Phase). Для огрубіння графів можуть бути застосовані декілька наявних алгоритмів. На кожному рівні огрубіння закінчується, як тільки розмір результуючого оргубленого графа зменшився в 1,7 раз.

На третій стадії виконується K-Way поділ оргубленого графа таким чином, щоб було задоволено обмеження балансу і оптимізована функція поділу.

На четвертому кроці виконується відновлення графа. Поділ оргубленого графа проектується на наступний рівень вихідного графа та виконується алгоритм поліпшення поділу (partitioning refinement algorithm) для поліпшення цільової функції, не порушуючи обмеження балансу.

На останній ітерації Хамелеона визначається показник схожості між кожною парою кластерів, беручи до уваги їх відносну зв'язність і відносну близькість. Це дозволяє вибрати для об 'єднання кластери, які добре пов'язані і досить близькі. Вибираючи кластери і ґрунтуючись на цих двох критеріях, Хамелеон долає обмеження існуючих алгоритмів, які оцінюють взаємозв'язок чи близькість.

Отже, можна виділити наступні стадії:

– побудова графа. Граф може бути побудований симетричний або асиметричний. При побудові графа можуть бути застосовані різні види відстаней;
– огрубіння графа. Огрубіння графа може бути виконано наступними методами: Random Matching (RM), Heavy Edge Matching (HEM), Light Edge Matching (LEM);
– початковий поділ графа. Існує кілька підходів до поділу графів: графічні методи, комбінаторні методи і спектральні методи. Також алгоритми можуть бути виконані в рамках рекурсивної навпіл, так як більшість методів виконує ділення графа навпіл;
– відновлення графа та удосконалення поділу графа Для поліпшення поділу графа застосовуються такі алгоритми: Kernighan–Lin (KL), Boundary KL, Fiduccia-Mattheyses (FM), Boundary FM. Ці ж алгоритми можуть бути застосовані на етапі поділу, взявши за початкову випадкове поділ оргубленого графа;
– об 'єднання схожих класів для отримання фінального розбиття.

2.2 Аналіз, опис і модифікація етапу початкового поділу графа

Рекурсивне ділення навпіл застосовується для пошуку перевпорядження, яке скорочує заповнення при розкладанні розрідженої матриці. Ці алгоритми зазвичай згадуються як алгоритми вкладених перерізів. Вкладені перерізи рекурсивно поділяють граф на майже рівні половини, видаляючи вузли сепаратора, поки не отримано бажане число поділів. Один шлях отримання вузлового сепаратора полягає в тому, щоб спочатку отримати ділення навпіл графа та потім обчислювати вузлової сепаратор з реберного сепаратора. Вузли графа нумеруються так, що на кожному рівні рекурсії вузли сепаратора нумеруються після вузлів в половинках графа. Ефективність і складність алгоритму вкладених перерізів залежить від алгоритму обчислення сепаратора. Взагалі, маленькі сепаратори призводять до меншого заповнення [9].

Для вирішення завдання розбиття графа можна рекурсивно застосувати метод бінарної ділення, при якому на першій ітерації граф розділяється на дві рівні частини, далі на другому кроці кожна з отриманих частин також розбивається на дві частини тощо У випадку, коли необхідну кількість сегментів k не є ступенем двійки, кожне ділення навпіл необхідно здійснювати у відповідному співвідношенні [10].

Процедура рекурсивного розподілу навпіл, служить для розбиття графа, що містить n вершин на довільне число доменів k. Схема роботи методу розподілу навпіл на 5 частин наведена на рисунку 2.2 і може бути описана наступним чином. Спочатку граф слід розділити на 2 частини у співвідношенні 2:3 (на схемі це безперервна лінія), потім праву частину розбиття стосовно 1:3 (на схемі це пунктирна лінія), після цього залишилося розділити 2 крайні підобласті зліва і справа у відношенні 1:1 (на схемі це пунктир з точкою) [10].

На рисунку 2.3 представлено результат розбиття 100 вершин графа на 7 частин приблизно рівного розміру. В кожному з коренів дерева розрізів зазначено число вершин відповідного підграфа, два числа, розділені знаком риски, вказують пропорцію, в якій вершини підграфа розбиваються на дві частини [11].

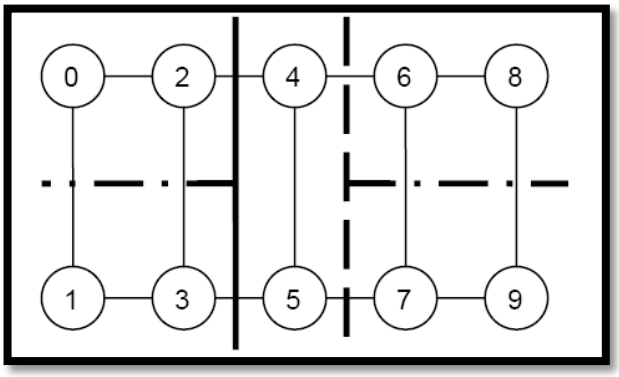

Рисунок 2.2 – Схема роботи методу розподілу навпіл

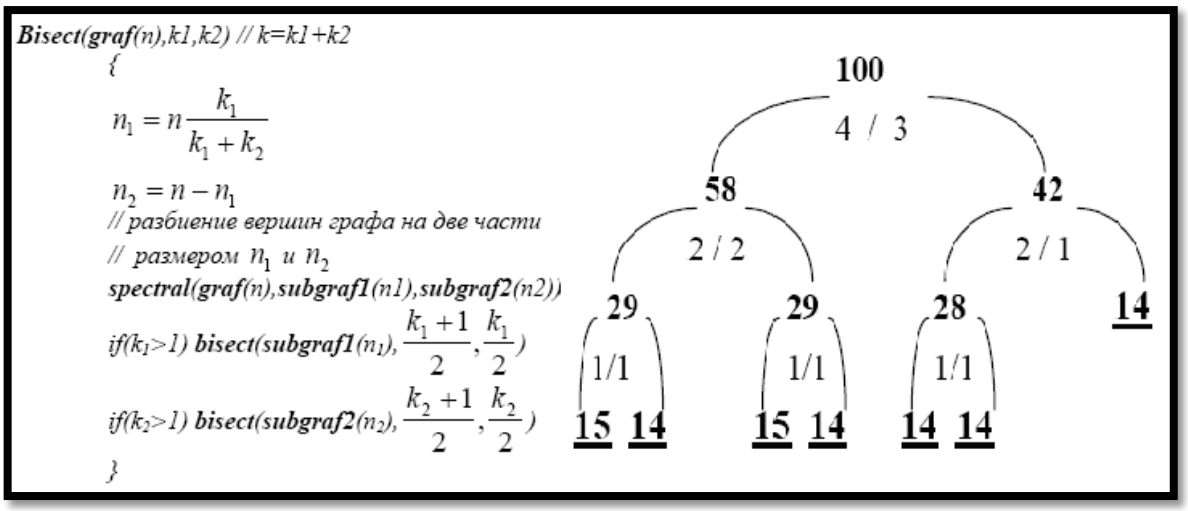

Рисунок 2.3 – Процедура рекурсивного розподілу множини на 7 частин.

Метод працює дуже швидко і вимагає невеликої кількості оперативної пам'яті. Однак, одержуване розбиття поступається за якістю більш складним і обчислювально трудомістким методам.

2.2.1 Геометричні алгоритми

Цей клас методів використовується для знаходження хорошого розподілення геометричної інформації про граф. Геометричні алгоритми обчислюють поділ, ґрунтуючись лише на даних про координати елементів множини, а не на пов'язаність елементів. Так як дані технології не розглядають зв'язність між елементами множини, зазвичай використовується мінімізація відповідних метрик, таких як кількість елементів суміжних із зовнішніми елементами (тобто розмір рамки підкласів) [1].

Геометричні алгоритми поділу мають тенденцію бути швидкими, але часто видають поділ, якість якого гірша, ніж у інших методів. Однак, через випадкову природу цих алгоритмів, часто потрібні багаторазові повторення (від 5 до 50), щоб отримати рішення, порівнянні за якістю зі спектральними методами. Багаторазові повторення збільшують загальний час, але воно залишається все ж істотно менше, ніж у спектральних методів. Геометричні алгоритми поділу графа застосовуються, тільки якщо відомі координати вузлів графа. У багатьох предметних областях (наприклад, лінійному програмуванні), немає ніякої геометричній інформації, пов'язаної з графом [9,12].

Покоординатне розбиття (Coordinate Nested Dissection, CND). Це метод, заснований на рекурсивному поділі навпіл множини по найбільш довгій стороні. Як ілюстрацію на рисунку 2.4 наведено зразок мережі, і два можливих розбиття, вироблених даним методом – по осі x та по осі y. Правильним в даному випадку буде розбиття по осі x.

Загальна схема виконання методу полягає у наступному. Спочатку обчислюються центри мас елементів мережі. Отримані точки проектуються на вісь, відповідну найбільшою стороні в спільній мережі. Таким чином, ми отримуємо впорядкований список всіх елементів мережі. Діленням списку навпіл (можливо, в

належній пропорції) ми отримуємо необхідну бісекцію. Аналогічним способом отримані фрагменти розбиття рекурсивно діляться на потрібне число частин.

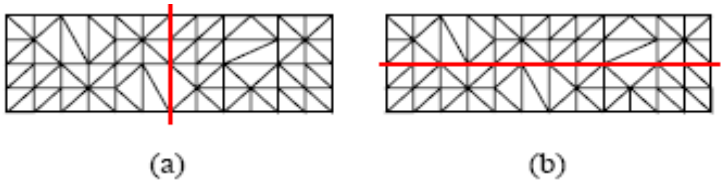

Рисунок 2.4 – Два CND-розбиття графа відносно осей координат. На малюнку (a) граф розділено по осі x, на малюнку (b) по осі y.

Метод координатного вкладеного розбиття працює. Однак, одержуване розбиття поступається за якістю більш складним і обчислювально трудомістким методам. Крім того, у випадку складної структури мережі алгоритм може отримувати розбиття з непов'язаними підмережами [10].

На рисунку 2.5 спочатку було виконано поділ, показаний суцільною лінією. Потім виконано поділи, показані пунктиром для кожної підмножини. Верхня та нижня ліві підмножини незв'язані [13].

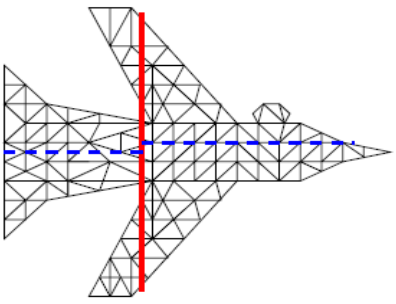

Рисунок 2.5 – 4-Way поділ, виконаний за CND схемою.

CND алгоритм працює наступним чином. Відповідно розраховуються центри мас елементів множини. Потім виконується проектування центрів мас на вісь

координат, якій відповідає максимальна величина у множині. Дана операція впорядковує елементи множини. Для поділу безлічі виробляється ділення навпіл отриманого упорядкованого списку. Кожна з отриманих підмножин, у свою чергу може бути поділена подібним чином. На рисунку 2.6 показано 8-Way поділ, виконаний за допомогою згаданого методу. Показані центри мас елементів множини і показано лінії рекурсивної бісекції. На першому кроці була проведена бісекція всієї множини, на малюнку дане розбиття показано суцільною лінією. Потім поділ було виконано для кожного з отриманих підмножин.

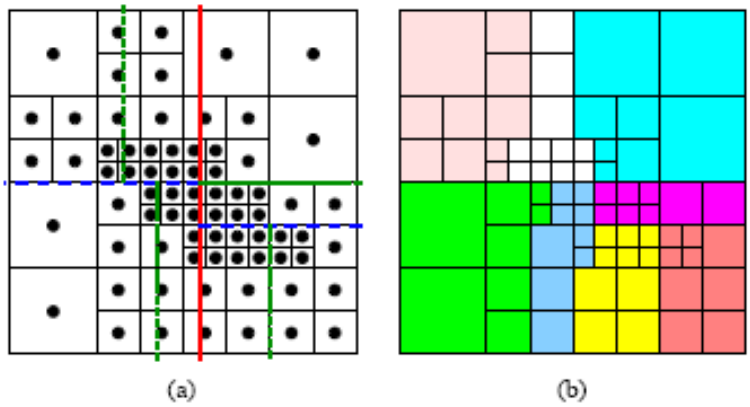

Рисунок 2.6 – 8-Way поділ, виконаний за допомогою CND алгоритму.

CND алгоритм працює дуже швидко, не потребує у великому обсязі пам 'яті і легко може бути розпаралелений. На додаток до описаних перевагам необхідно додати легкість і компактність опису виконаного поділу, для цього достатньо вказати роздільники, використані на кожному з вузлів дерева рекурсивної бісекції. Однак поділ, виконаний за CND алгоритмом, є низькоякісним. Крім того для складних геометричних фігур в результаті виконання поділу можуть бути отримані незв'язані підмножини [12].

Ділення множини з використанням кривих, які заповнюють простір (Space flling Curve Techniques, SFCT). Одним з важливих недоліків графічних методів є те, що при кожній бісекції ці методи враховують тільки одну розмірність. Таким

чином, схеми, що враховують більше розмірностей, можуть забезпечити краще розбиття [12].

Один з таких методів впорядковує елементи відповідно з позиціями центрів їхніх мас вздовж кривих. Криві, що заповнюють простір – це криві, що повністю заповнюють фігури великих розмірностей (наприклад, квадрат або куб). Застосування таких кривих забезпечує близькість точок фігури, відповідних точкам, близьким на кривої. Після отримання списків елементів мережі, упорядкованого відповідно з розташуванням на кривій, досить розділити список на необхідну кількість частин відповідно до встановленого порядку. Що отримується в результаті такого підходу метод носить в літературі найменування Space filling curve technique [10].

Криві, що заповнюють простір бувають декількох видів: Пеано, Гільберта, Z-порядку. Криві відрізняються між собою порядком з 'єднання комірок [13].

Алгоритм містить наступні кроки:

– ініціалізація. На етапі ініціалізації вибирається квадрат, в який потрапляють всі наявні елементи множини. Це нульовий рівень декомпозиції простору. 4 квадратні області поділу позначаються мітками з I по IV. I відповідає початку кривої, а IV закінченню;
– ітеративний крок 1. Для створення наступного рівня декомпозиції квадрат розділяється на 4 квадратні області. Через центри отриманих областей проводиться крива. Подібно до етапу ініціалізації кожен з отриманих квадратів розмічається мітками;
– ітеративний крок 2. На даному кроці відбувається виставлення міток в областях нових квадратів і зміна кожного із 4 дільниць кривої. Замість відрізка що з'єднує центр отриманого на попередньому кроці квадрата з сусіднім, крива буде проведена по центрам під квадратів;
– поділи тривають до тих пір, поки до кожної окремо поміченої області не буде утримуватися не більше 1 елемента вихідної множини;

– вирішальний крок. За отриманою кривою складається відсортований список елементів множини. Далі список ділиться на класи.

Найкращий результат поділу виходить при поділі класів, в яких залежності між вузлами відповідають їх близькості в просторі [12].

Рекурсивний поділ графа (Recursive Graph Bisection, RGB). Для розрахунку відстані використовується діаметр – величина, протилежна прямокутну евклідову відстанню. Отже, відстань між двома вузлами n1 і n2 d(n1, n2) дорівнює кількості ребер у найкоротшому шляху, що сполучає n1 і n2.

На першому кроці алгоритму розраховується діаметр, потім вузли сортуються відповідно до відстані до граничного вузла. Та половина вершин, яка ближче до граничного узлу, залишається в одному класі, інші належать до інших класів [14].

Рекурсивний інерційний поділ (Recursive Inertial Bisection, RIB) зображений на рисунку 2.7. Рекурсивно інерційний метод поділу навпіл будує головну інерційну вісь, вважаючи елементи мережі точковими масами. Лінія бісекції, ортогональна отриманій осі, дає кордон найменшої довжини та є лінією поділу класів [14].

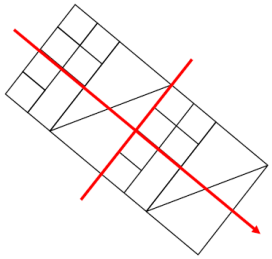

Рисунок 2.7 – Рекурсивний інерційний поділ.

Модифікований рекурсивний поділ (Modified Recursive Graph Bisection, MRGB). Даний метод є модифікацією RGB методу. Він покращує оригінальний метод за допомогою отримання взаємопов'язаних підмножин в результаті поділу.

У запропонованому методі кожний поділ надвоє починається з пошуку двох відносно екстремальних вузлів графа та потім будується поділ навколо них, формуючи дві множини. Спочатку кожна множина складається з тих вузлів, які знаходяться на відстані одного зв'язку від екстремального вузла, потім двох і так далі. Додавання вузлів триває до тих пір поки одна з підмножин не буде містити половину всіх вузлів, а до меншої буде додана решта вузлів і не залишиться вузлів для додавання. Якщо є вільні вузли, не пов'язані з меншою множиною, то вони додаються у більшу [14].

Рекурсивний поділ вузлів в кластери (Recursive Node Cluster Bisection, RNCB). Даний метод комбінує підходи модифікованого рекурсивного поділу та рекурсивної спектральної бісекції. Він ґрунтується на ідеї вузлового кластера, який являє собою сполучені елементи множини, об'єднані в кластер. Такі кластери будуть вузлами в матриці Лапласа. Отже матриця буде менше і метод спектральної бісекції спрацює швидше. Головною проблемою є збалансованість кінцевого розбиття.

2.2.2 комбінаторні методи

Геометричні алгоритми з попередньої секції мають дві стандартні уразливості: перша в тому, щоб кожна вершина графа мала геометричні координати, а вони не завжди присутні для всіх графів; друга в тому, що не приділяється увага структурі взаємозв'язків у графі, а віддається перевага просторовій близькості, яка передбачає невелику відстань між вершинами. Хоча це доцільна опущення, для багатьох графів існують протиріччя. Наприклад, у графі на рисунку 2.5 вершини з різних сторін крил близькі в просторі, але малопов’язані.

Недоліки геометричних алгоритмів враховані в структурних або комбінаторних методах [12].

Комбінаторні методи навпаки прагнуть згрупувати пов'язані між собою вершини, не надаючи значення відстані між ними в просторі. Отже, виконані поділи мають менший роздільник і менш схильні до наявності розділених підмножин порівняно з геометричними алгоритмами. Тим не менш, комбінаторні методи є повільнішими, ніж геометричні алгоритми, і не можуть бути розпаралелені [15].

Алгоритм зростаючого графа (Graph Growth Partitioning, GGP). Інший простий шлях ділення графа навпіл полягає в тому, щоб почати з одного вузла і нарощувати область навколо цього першого вузла динамічно, поки не буде включена половина вузлів (або половина загальної ваги вузлів). Якість алгоритму зростаючого графа чутливо до вибору вузла, з якого починається зростання графа, різні початкові вузли призводять до роздільниів різної ваги. Щоб частково вирішити цю проблему, можна безладно вибрати 10 вузлів і виростити 10 різних областей. Цей поділ може бути подано на вхід KL алгоритму як початкове наближення для поліпшення. І знову, оскільки підмножина дуже мала, цей крок займає маленький відсоток від повного часу роботи [9].

Алгоритм зростаючого графа з урахуванням вигоди (Greed Graph Growing Algorithm, GGGP). Алгоритм зростаючого графа, описаний вище, вирощує поділ випадковим чином. Однак, для кожного вузла V можна визначити вигоду в вазі сепаратора, отриману від вставки V в зростаючий регіон. Таким чином, можна упорядкувати граничні вузли графа в зростаючому порядку згідно їх вигоді. Вузол з найбільшим зменшенням (або найменшим збільшенням) у вазі сепаратора вставляється першим. Коли вузол вставляється в зростаючий поділ, то вигоди його суміжних вузлів, що знаходяться на кордоні, оновлюються, і ті, можливо, йдуть з кордону. Структури даних, що вимагаються для здійснення цієї схеми, по суті ті ж, що вимагаються KL алгоритмом. Єдина відмінність полягає в тому, що замість попереднього обчислення всіх вигод для всіх вузлів ми виробляємо обчислення

тільки для вузлів, які стоять на кордоні. Цей вигідний алгоритм також чутливий до вибору початкового вузла але менше ніж GGP. В нашій реалізації ми випадково вибираємо 5-6 вузлів як відправні точки алгоритму і вибираємо поділ з меншою вагою сепаратора. Експерименти показують, що GGGP потребує трохи меншої кількості часу, ніж GGP, для поділу грубого графа (тому що йому достатньо меншої кількості повторів), і початковий поділ, знайдений згідно з цією схеми, буде краще, ніж знайдений GGP [9].

Рівневе осередкове розбиття (Levelized Nested Dissection, LND). Зазвичай розбиття буде мати мало розрізаючих ребер, якщо суміжні вершини знаходяться в одній підмножині. LND намагається поміщати пов'язані вершини разом, починаючи з підмножини, яка містить лише одну вершину, і потім збільшує підмножину шляхом додавання суміжних вершин. На рисунку 2.8 показаний поділ отриманий LND-алгоритмом.

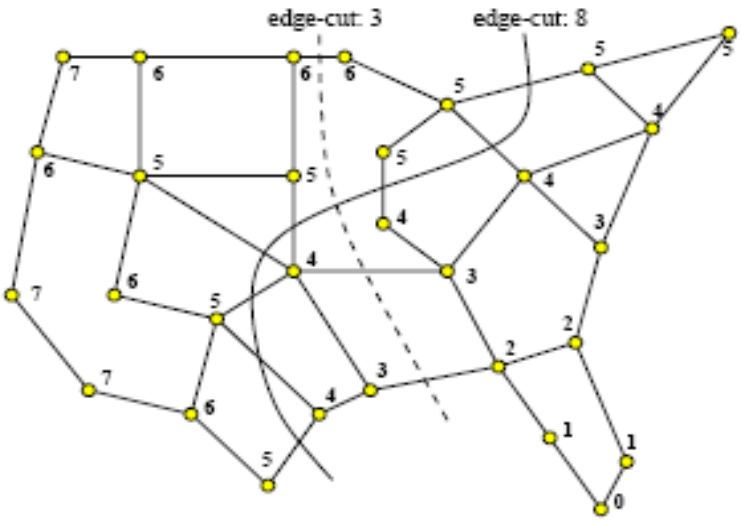

Рисунок 2.8 – Суцільною лінією показано розбиття, отримане LND.

Докладніше, алгоритм полягає в наступному:

– вибирається початкова вершина V_0 і позначається номером 0, переважно периферійна;

– для кожної вершини обчислюється відстань до V_0, використовуючи Breadth First search, BFS алгоритм, починаючи з вершини V_0;

– коли половина вершин вже позначена, граф розбивається на два підграфа.

Для зменшення ризику, що початкова вершина V_0 обрана погано, може бути виконано кілька спроб для різних початкових вершин [15].

До недоліку даного алгоритму можна віднести те, що отримане таким чином розбиття сильно залежить від початково обраної вершини. Тому можна зробити висновок, що даний алгоритм повинен бути доповнений методами пошуку початкової вершини [12].

Seed Growth Bisection (SGB). На початку роботи даного алгоритму вибирається 2 несуміжних рівних підмножини з безлічі вершин графа. Вибрані підмножини будуть стартовими підмножин поділу (seed). Далі, поступово з решти вершини по одній додаються почергово до кожного їх підмножин. На кожному кроці для додавання вибирається вершина, Найближча до підмножини. При додаванні вершини в підмножину X вибирається вершина, для якої поділ cut({a},Y) – cut({a},X) мінімальний. Інтуїтивно мінімізується число ребер, доданих до сепаратора що розділяє множину X та Y, при цьому збільшується число ребер, що обмежують від подальшого додавання до множини що розділяє.

Сам по собі алгоритм SGB не є результативним, але так як він приблизно в 5 разів швидше KL алгоритму, то він може викликатися кілька разів та бути успішно застосований як частина глобального алгоритму. Так само даний алгоритм може бути використаний для початкового розбиття для KL і FM алгоритмів.

3 КРИТЕРІЇ ЯКОСТІ РЕЗУЛЬТАТІВ КЛАСТЕРИЗАЦІЇ

3.1 Опис досліджуваних статистичних характеристик вибірки вхідних даних

Вибірка або вибіркова сукупність це безліч випадків (піддослідних, об'єктів, подій, зразків), за допомогою визначеної процедури вибраних з генеральної сукупності для участі у дослідженні.

Вибіркова сукупність це сукупність статистичних одиниць, відібраних за певними правилами з генеральної сукупності для проведення спостереження.

Можна сказати що вибіркова сукупність це частина генеральної сукупності, об'єкти якої виступають як основні об'єкти спостереження. Ця частина генеральної сукупності відбирається за спеціальними правилами так, щоб її характеристики відображали властивості всієї генеральної сукупності. Таким чином, досліджується лише частина генеральної сукупності, але так, що є можливість отримати повне уявлення про всю сукупність у цілому. Це, в свою чергу, дає економію часу, людських ресурсів і матеріальних витрат.

Основними з них є математичне очікування, дисперсія, асиметрія, ексцес, мода і медіана.

Нехай X є випадкова величина, (x_1, x_2,..., x_n) – можливі значення випадкової величини, (p_1, p_2,..., p_n) – відповідні їм ймовірності, тоді на формулі 3.1 зображено k-m початковий момент, де k = 1, 2,..., n.

$$\gamma_k = \sum\nolimits_i x_i^k p_i \tag{3.1}$$

де x_i – можливе значення випадкової величини;

p_i – відповідна величина ймовірності.

На формулі 3.2 зображено число, яке називається к-м центральним моментом випадкової величини X, тобто центром розподілу.

$$\mu_k = \sum_i (x_i - k)^k \, p_i \tag{3.2}$$

де x_i – можливе значення випадкової величини;

p_i – відповідна величина ймовірності;

k – 1, 2,..., n.

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

Перший початковий момент приведений у формулі 3.3 і називається математичним очікуванням випадкової величини і зазвичай позначається як M(x).

$$\gamma_1 = \xi = \sum_i x_i p_i = M(x). \tag{3.3}$$

де x_i – можливе значення випадкової величини;

p_i – відповідна величина ймовірності;

k – 1, 2,..., n.

На практиці, маючи справу з обмеженими масивами даних, замість математичного очікування використовують вибіркове середнє.

В якості однієї з числових характеристик статистичної вибірки використовується вектор математичних очікувань ознак, що дозволяє отримати відцентровану вибірку X_o з елементами, вона приведена на формулі 3.4.

$$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij},\ j \equiv \overline{1,m}, \tag{3.4}$$

де x_{ij} – можливе значення випадкової величини;

n – початковий момент;

Початкові моменти другого, третього і четвертого порядків так само як і умовні моменти самостійного значення не мають, а використовуються для спрощеного обчислення центральних моментів.

Другий центральний момент називається дисперсією випадкової величини і служить мірою її розсіювання для вибірки, він приведений на формулі 3.5.

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

де x_i – можливе значення випадкової величини;

n – початковий момент;

p_i – відповідна величина ймовірності.

Дисперсія характеризує розподілення випадкової величини щодо математичного очікування. Вона є мірою відхилення значень випадкової величини від середнього значення розподілу. Більші значення дисперсії свідчать про більші відхилення значень випадкової величини від центру розподілу.

Дисперсія має розмірність квадрата розмірності випадкової величини, тому, щоб отримати характеристику розсіювання з розмірністю випадкової величини, часто використовують у якості показника розсіювання не дисперсію, а середнє квадратичне відхилення σ яке обчислюється як корінь квадратний із дисперсії, квадратичне відхилення приведене на формулі 3.6. Використовують й інші формули для розрахунку коефіцієнта асиметрії.

$$\sigma = \sqrt{\mu_2} = \sqrt{D} = \sqrt{\sum_i (x_i - \bar{x})^2 p_i}. \quad (3.6)$$

де x_i – можливе значення випадкової величини;

p_i – відповідна величина ймовірності.

Третій центральний момент служить в якості характеристики асиметрії розподілу, він приведений на формулі 3.7. Центральний момент третього порядку дорівнює нулю в симетричному розподілі і використовується для визначення показника асиметрії (скошеності). Розрізняють також нормовані моменти, під якими розуміють відношення моменту к-го порядку до середнього квадратичного відхилення в к-му степені Відповідно до цього коефіцієнт скошеності можна розглядати як третій нормований центральний момент розподілу.

$$\mu_3 = D = \sum_i (x_i - \xi)^3 p_i. \quad (3.7)$$

де x_i – можливе значення випадкової величини;

p_i – відповідна величина ймовірності.

Щоб отримати безрозмірну величину, замість μ3 вводять коефіцієнт асиметрії, він приведений на формулі 3.8.

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

де x_i – можливе значення випадкової величини;

p_i – відповідна величина ймовірності;

σ^3 – куб середнього квадратичного відхилення.

Асиметрія характеризує неоднакову повторюваність метеорологічної величини відносно середнього значення або неоднаковість проміжків часу, протягом яких дана метеорологічна значення керує значення вище середнього чи нижче середньої

Четвертий центральний момент використовується для оцінки крутості кривої розподілу порівняно з нормальною кривої розподілу. В якості міри крутості використовують коефіцієнт ексцесу він приведений на формулі 3.9.

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

де x_i – можливе значення випадкової величини;

p_i – відповідна величина ймовірності;

σ^3 – куб середнього квадратичного відхилення.

Ексцес характеризує крутість розподілу. Коефіцієнт ексцесу це числова характеристика розподілу ймовірностей дійсної випадкової величини. Коефіцієнт ексцесу характеризує крутість, тобто, стрімкість підвищення кривої розподілу у порівнянні з нормальною кривою. В такий спосіб можна встановити відхилення заданого закону від нормального. Розмах – це різність між найбільшим і найменшим значеннями ряду даних. Вибірковий коефіцієнт кореляції знаходиться за допомогою формули 3.10, де σ_x^4, σ_y^4 вибіркові середні квадратичні відхилення величин X та У.

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

де x_i – можливе значення випадкової величини;

p_i – відповідна величина ймовірності;

σ^4_x – вибіркове квадратичне відхилення величини x;

σ^4_y – вибіркове квадратичне відхилення величини y.

Вибірковий коефіцієнт кореляції r(X, Y) показує тісноту лінійного зв 'язку між X і Y: чим ближче r(X, Y) до одиниці, тим сильніший лінійний зв'язок між X і Y [16].

В ході експерименту була додана обчислюється характеристика SetDist, що представляє максимальну відстань між компонентами зв'язності та кількість елементів в компоненті зв'язності. Ця характеристика обчислюється за формулою 3.11, на формулі зображено avComponent який є центроїдом компоненту зв'язності, ComponentOstovEdge – ребро, що з'єднує вершини які належать одному компоненту, ComponentVertexNum це кількість вершин у компоненті.

Дана характеристики не трудомістка в розрахунку і існує залежність між нею і значенням k при побудові графа найближчих сусідів.

$$SetDist = \max\left(\frac{dist(avComponent_i, avComponent_j)}{\max\left(\frac{\max ComponentOstovEdge_{ij}}{ComponentVertexNum_{ij}}\right)}\right) \quad (3.11)$$

де avComponent – центроїд компоненту зв'язності;

ComponentOstovEdge – ребро, що з'єднує вершини одного компоненту;

ComponentVertexNum –кількість вершин у компоненті;

Так як для побудови математичної моделі необхідні характеристики всієї вибірки, а не тільки окремих атрибутів, для кожної статистичної характеристики буде розрахована середня, максимальне і мінімальне значення за вибіркою.

3.2 Критерії оцінювання якості кластеризації

Ключові аспекти оцінювання це ефективність, надійність, простота і результативність. Розрахунок часу вироблявся на 3.5 GHz Intel (R) Pentium (R) Quad CPU з 8 GB пам'яті.

Час продуктивності для різних графів може бути використано для вивчення масштабованості алгоритму. Продуктивність залежить від кількості вершин і ребер.

Оцінювання результатів огрубіння графа. Оцінка результатів ґрунтується на зменшенні фактору якості огрубіння і відновлення. Тільки в даній комбінації можливо оцінити якість застосування алгоритмів огрубіння в різних комбінаціях з алгоритмами відновлення.

Оцінювання результатів кластеризації. Процес оцінки результатів алгоритму кластеризації називається оцінкою cluster validity assessment. Існує два критерії вимірювання якості кластеризації, оцінки та вибору оптимальної схеми кластеризації:

– компактність: елементи в кожній групі повинні бути як можна ближче один до одного, наскільки це можливо. Мірою компактності є дисперсія;
– поділ: кластери повинні бути максимально віддалені один від одного. Є три загальних підходу відстані, що підтримуються між двома різними кластерами: відстань між найближчими членами кластера, відстань між самими віддаленими членами і відстані між центрами кластерів.

Існують три різних методу оцінки результатів алгоритмів кластеризації:

– зовнішній критерій;
– внутрішній критерій;
– відносний критерій.

Як внутрішні, так і зовнішні критерії засновані на статистичних методах, і вони мають високу обчислену вартість. Зовнішні методи оцінки дії кластеризації засновані на якійсь інтуїції користувача. Внутрішні критерії засновані на метричних показниках, які засновані на наборі даних та схемі кластеризації. Основним недоліком цих двох методів є їх обчислювальна складність.

В основі відносних критеріїв лежить порівняння різних схем кластеризації. Один або кілька алгоритмів кластеризації виконуються кілька разів з різними вхідними параметрами на тому ж наборі даних. Мета відносних критеріїв – вибір кращої схеми кластеризації з різних результатів. Основою для порівняння є індекс валідності.

3.3 Створення експериментальних вибірок

Для перевірки працездатності методу необхідна велика кількість вибірок з різним розподілом в генеральній сукупності. Відсутність реального джерела даних необхідного обсягу, різноманітності і якості змушує звернутися до альтернативного джерела [17]. Так як, використання різних вхідних даних з певними статистичними характеристиками, продуктивність і якість кластеризації може сильно відрізнятися [18], необхідно проводити аналіз на синтетичних вибірках, створених спеціально для даної задачі. Досліджень в даній області небагато, і всі вкрай специфічні для розглянутих завдань.

Існує ряд методів генерації експериментальних даних, що дозволяють провести аналіз кластеризації систематично і послідовно. Такі генератори використовують параметризовані моделі, які створюють реалістичні дані. Ці генератори навчені на реальних даних.

Helmers і Bunke (2003) розроблено для роботи зі зразками почерку. Baird (1993) і Baird (2000) працювали з зображеннями. Rogers (2003) працював з 2d зображеннями протеїну. David і співавтори (2004) отримували набори даних помічаючи текстовий контент із WWW. (Srikant, 1999), GSTD (1999) і Jeske (2005) також займалися синтетичним створення даних. GSTD моделює броунівський рух. Існує ряд прихованих моделей Маркова на основі генераторів даних. Rachkovskij і Kussul (1998) продемонстрували більш загальний алгоритм генерації зразків з ознак в просторі, включаючи фоновий шум. Pei і Zaiane (2006) займалися отриманням даних неконтрольованого навчання і виявлення викидів. Van der Walt і Bernard (2007) демонструють корисність синтетичних генераторів набору даних на основі різних щільності [19].

Жоден з перерахованих методів не дозволяє генерувати набори наочних даних з певними статистичними характеристиками вибірки.

У статті Vineet Chaoji, Mohammad Al Hasan, Saeed Salem, і Mohammed J. Zaki "SPARCL: Efficient and Effective Shape based Clustering" [20] для тестів масштабованості, а також для створення 3d-даних, написано власний генератор кластерів, заснований на фігурах. Для створення фігури в 2d випадковим чином вибиралися точки на канві і додавалися точки, які формують бажані фігури. Точкою відліку для всіх фігур була початкова точка 0,0.

Щоб отримати складні фігури, використовувалися фігури, отримані обертанням і зміщенням (коло, прямокутник, еліпс, кругові смуги). Генерація 3d фігури побудована на 2d постатях. Такий підхід дозволяє побудувати правдоподібну 3d фігуру, а не тільки декілька шарів 2d фігури. Як і у випадку 2d, комбінується обертання і зміщення для отримання 3d фігури (Рис. 3,1). Як тільки створені всі фігури, випадковим чином додається шум (від 1% до 2%). Показаний на малюнку 3d набір даних має 100000 точок і 10 кластерів [20].

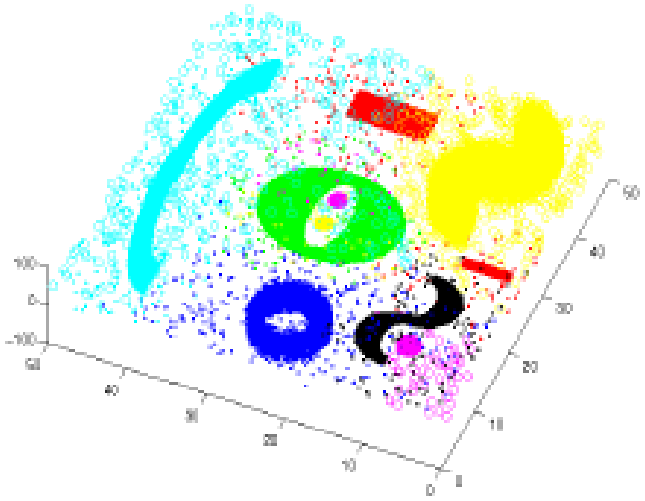

Рисунок 3.1 – 3d фігури, отримані обертанням та переміщенням

В даній роботі створення 3D фігур засновано на 3DS Max Studio. Даний Додаток дозволяє згенерувати тривимірну фігуру необхідної щільності та з

необхідною кількістю точок. Далі фігура може бути експортована. Статистичні характеристики отриманої вибірки будуть залежати від характеру фігур, їх розміру, щільності та розташування. Зміни значення цих параметрів підбираються при створенні фігур. Додавання шуму до вибірки проводиться безпосередньо перед проведенням аналізу.

Для проведення експерименту даним методом було згенеровано 130 вибірок з різними статистичними характеристиками і процентом шуму. Додавання шуму до вибірки проводиться безпосередньо перед проведенням аналізу.

Можливо, не всі вибірки однаково хороші для вивчення і порівняння алгоритмів кластеризації. Деякі можуть не мати добре виражену структуру кластерів, кластери можуть бути неоднорідні, розподілені різним чином. Багато які вибірки не підходять для дослідження у рамках даної роботи кількістю атрибутів чи іншими характеристиками. Всі ці фактори впливають на якість роботи алгоритмів.

Згенеровані вибірки – це досить простий і точний спосіб проведення експерименту на великій кількості вибірок з відомими структурними характеристиками, кластери можуть бути неоднорідні, розподілені різним чином Серед недоліків можна перерахувати:

– залежність згенерованих вибірок від програми генератора. Щільність і кількість вершин може бути різним навіть при однакових параметрах для генерації вибірки;
– даний спосіб створення вибірки дозволяє створити набір з певними характеристиками, але такі вибірки не завжди можуть мати необхідну структуру. В деяких предметних областях все ще важко відтворити структуру реальних даних при генерації вибірки.

Велика кількість реальних вибірок використовувати в різних роботах. Список спільних файлів з реальними вибірками представлений в таблиці 3,1. Багато які вибірки не підходять для дослідження у рамках даної роботи кількістю атрибутів

чи іншими характеристиками. В таблиці 3.1 приведений список посилань на ресурси з реальними вибірками даних.

Таблиця 3.1 – Посилання на ресурси з наборами даних для кластеризації.

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

Велика кількість реальних вибірок використовувати в різних роботах. Багато які вибірки не підходять для дослідження у рамках даної роботи кількістю атрибутів чи іншими характеристиками.

4 АНАЛІЗ ТА ПОБУДОВА МАТМОДЕЛЕЙ. ТЕСТУВАННЯ РЕЗУЛЬТАТІВ

4.1 Побудова математичної моделі

Математичне моделювання – метод дослідження процесів або явищ шляхом побудови їх математичних моделей і дослідження цих моделей. Математичні моделі досліджуються, як правило, за допомогою аналогових обчислювальних машин, або цифрових обчислювальних машин та комп'ютерів. Такого роду дослідження підрозділяються на:

– аналітичне дослідження впливу змін зовнішніх умов на вихідні впливу об'єкта навчання, коли є можливість в явному вигляді записати залежність $\bar{Y}$ = f($\bar{y}$, $\bar{v}$, t) яка описує взаємозв'язок зовнішніх умов і керуючих впливів з виходами об'єкта дослідження, і шляхом аналітичних перетворень (у тому числі з використанням чисельних методів) відшукується об’єкт потрібного пошуку рішення;
– імітаційне математичне моделювання, коли виконуються експериментальні дослідження математичної моделі об'єкта вивчення з відтворенням бажаних режимів функціонування об'єкта шляхом імітації сигналів зовнішніх впливів.

Математичної моделлю називається формальна система, що представляє собою кінцеве збори символів і абсолютно точних правил оперування з цими символами в сукупності з інтерпретацією властивостей певного об 'єкта деякими символами, відносинами та константами. Математична модель – система математичних співвідношень, що описують досліджуваний процес чи явище [17].

Структура моделі – це тип залежності між вхідними і вихідними впливами об'єкта вивчення. Параметри моделі – це коефіцієнти залежностей. Класифікація математичних моделей наведена в таблиці 4.1.

Таблиця 4.1 – Класифікація математичних моделей.

Ознаки класифікації	Види моделей	
За залежністю структури і параметрів моделі від величини вхідних та 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 змінюваних властивостей об'єкта	Неадаптивні	Адаптивні

Так само іноді математичні моделі класифікують за способом побудови:

– теоретичні або моделі внутрішнього механізму базуються на фундаментальних законах природи;
– функціональні або кібернетичні описують тільки зовнішню, поведінкову сторону функціонування досліджуваного явища, об'єкта і будуються за результатами експериментальних досліджень об'єктів;
– комбіновані об'єднують в собі теоретичні кібернетичні моделі.

Аналіз ситуації дозволяє виділити основні типи параметрів, що описують стан системи: керовані, цільові і некеровані.

Керовані параметри є шуканими та їх значення визначають стратегію.

Цільові параметри необхідні для опису поставлених цілей. Значення цільових властивостей залежать від керованих параметрів.

Некеровані параметри не можуть змінюватися керівництвом, залишаючись постійними, відомими повністю або частково.

Будемо вважати, що залежності між параметрами задаються у вигляді наступного набору функцій який зображено на формулі 4.1.

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

де X – позначення керованих параметрів;

a – позначення некерованих параметрів;

W – позначення цільових параметрів;

n – число керованих параметрів;

k – число некерованих параметрів;

m – число цільових параметрів.

W – позначення цільових параметрів, X – позначення керованих параметрів, a – позначення некерованих параметрів, m – число цільових параметрів, n – число керованих параметрів, k – число некерованих параметрів.

4.2 Математична модель залежності вибору параметра k

Для оптимізації вибору початкової параметра k при побудові k-nn графа необхідно побудувати математичну модель залежності k від характеристик

оброблюваної вибірки. Математична модель буде побудована на основі дослідження вищеописаних вибірок.

Побудова математичної моделі виконувалося на наборі експериментальних вибірок. Набір вибірок складається з 132 вибірок, серед них 33 унікальних вибірки і 3 варіацій кожної з них отриманої шляхом додавання 20%, 40% і 60% шуму. Експеримент так само проводився на наборах експериментальних та реальних вибірок отриманих на ресурсах обміну наборами даних. Структура моделі – це тип залежності між вхідними і вихідними впливами об'єкта вивчення.

Метою цих експериментів був вибір керованих параметрів даної моделі залежності, здатних відобразити необхідні характеристики вибірки даних. Структура моделі – це тип залежності між вхідними і вихідними впливами об'єкта вивчення. В рамках роботи було проведено 3 експерименти для вибору керованих параметрів:

– у першому експерименті аналізувалися такі характеристики: кількість об'єктів у вибірці, мінімальні та максимальні значення математичного очікування, дисперсії і розкиду. Залежності між даними параметрами і значенням k не виявлено;
– у другому експерименті в якості керованого параметра була обрана довжина найбільшого остовного ребра повнозв’язаного графа. Дана характеристика показує залежність, але використання даного підходу не є доцільним у зв'язку з трудомісткістю побудови остова повнозв’язаного графа;
– в третьому експерименті в якості характеристики використовувалась кількість компонентів зв'язності, максимальна відстань між компонентами зв'язності та кількість елементів в компоненті зв'язності. Друга характеристика обчислюється за формулою 3.9.

В результаті дослідження була побудована математична модель для оптимізації вибору початкового значення k при побудові асиметричного k-nn графа. Структура моделі – це тип залежності між вхідними і вихідними впливами

об'єкта вивчення. Параметри моделі – це коефіцієнти залежностей. Модель для асиметричного k-nn графа представлена на формулі 4.2 і зображена на рисунку 4.1.

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

де x1 – коефіцієнт відстані;

x2 – кількість компонентів зв'язності.

Значення коефіцієнтів a-j представлені в таблиці 4.2.

Таблиця 4.2 – Значення коефіцієнтів.

№ з/п	Коефіцієнт	№ з/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

Статистичні характеристики отриманої вибірки будуть залежати від характеру фігур, їх розміру, щільності та розташування.

Перевірити адекватність моделі значить встановити, наскільки добре модель описує реальні процеси, що відбуваються в системі, наскільки якісно властивості моделі (функції, параметри, характеристики) відповідають властивостям модельованого об'єкта. Найбільш поширеним способом перевірки моделей на адекватність є використання методів математичної статистики. висока міра точності моделі не є гарантією того, що модель правильно відбиває реальне середовище. Перевірка на точність моделі є лише першим етапом процедури

дослідження адекватності, а отже завжди слід здійснювати, як перевірку точності моделі, так і її адекватність.

Про якість побудованої моделі можна судити що стандартна помилка оцінки дорівнює 11,2986020522291.

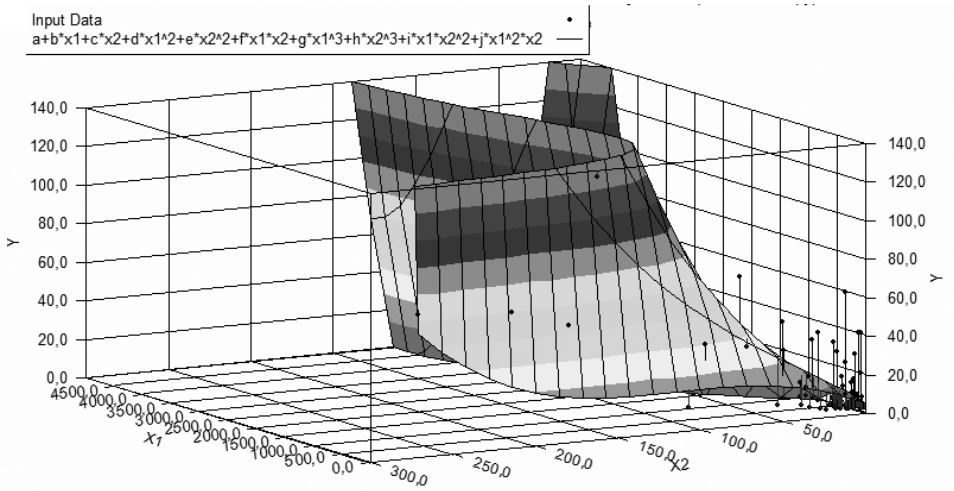

Рисунок 4.1 – Графічний вигляд опису даних математичної моделлю.

Перш ніж приступити до перевірки і встановлення адекватності, необхідно виробити критерій, який дозволив би зробити висновок про відповідність моделі і об'єкта. Вони базуються в основному на методах дисперсійного аналізу та аналізу залишків.

Отримавши математичну модель, конструктор системи повинен перевірити її за допомогою даних, не використаних при розробці моделі.

По-перше, сутність всіх методів є однаковою, та полягає у перевірці висунутої гіпотези (в даному випадку про адекватність моделі) на основі деяких статистичних критеріїв. При перевірці гіпотез методами математичної статистики необхідно мати наувазі, що статистичні критерії не можуть довести жодної гіпотези, вони можуть лише вказати на відсутність спростування. Відповідно до вище сказаного всі відомі способи перевірки на адекватність є однаково прийнятними, а при

застосуванні комплексного підходу можуть використовуватись різні статистичні гіпотези. Залишки при побудові даної моделі представлені на рисунку 4.2.

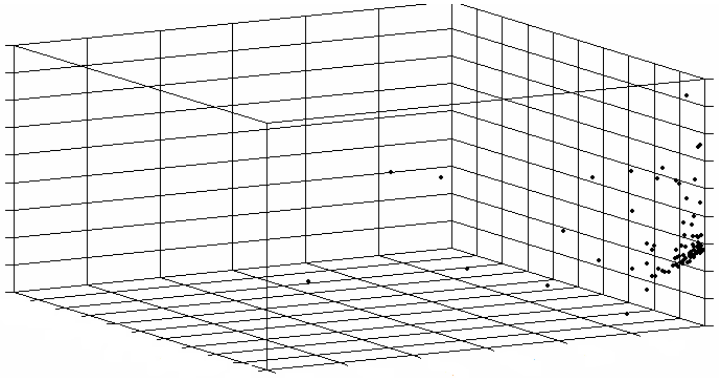

Рисунок 4.2 – Графічний вигляд залишку.

Оцінки і статистики якості даної моделі не є залишковими показниками ефективності застосування отриманої моделі, бо модель є лише одним з етапів вибору k. Застосування підходу досліджувалося на 285 вибірках. Застосування даної моделі поліпшили час виконання етапу побудови графа в 62,45% випадків. В 37,55% випадків час виконання погіршився.

Час виконання погіршився лише у тих випадках, коли k було менше або дорівнює 3 і час виконання мало, Отже, погіршення тимчасового показника несуттєво позначається на продуктивності методу в цілому. Негативний результат застосування моделі отриманий в 7,71% випадків. В середньому час виконання покращилося на 161%. Негативним результат вважається при отриманні k суттєво більшому мінімально необхідної для дотримання умови зв'язного мовлення, навіть якщо час побудови графа зменшилася.

Порівняння часу виконання до і після застосування моделі в залежності від кількості представлено на рисунку 4.3 і на рисунку 4.4.

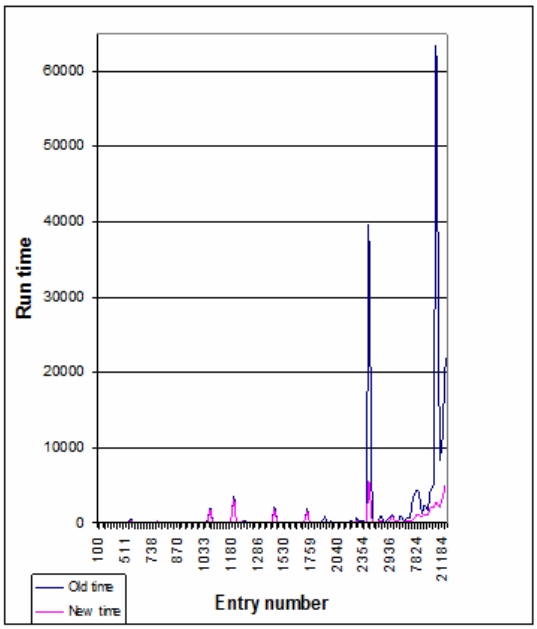

Рисунок 4.3 – Залежність часу побудови графа залежно від кількості елементів.

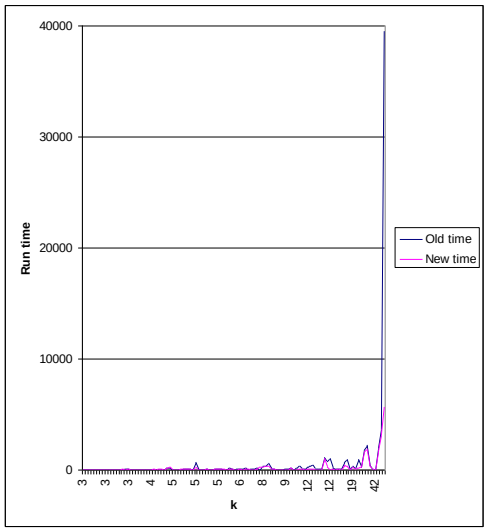

Рисунок 4.4 – Залежність часу побудови асиметричного графа залежно від отриманого значення k.

Так само в результаті дослідження була побудована математична модель для оптимізації вибору початкового значення k при побудові симетричного k-nn графа.

Модель для асиметричного k-nn графа має представлена на формулі 4.3 і зображена на рисунку 4.5.

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

де x1 – коефіцієнт відстані;

x2 – кількість компонентів зв'язності.

Значення коефіцієнтів a-j представлені в таблиці 4.3.

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 наступних характеристик:

– стандартна помилка оцінки дорівнює 42, 8805641130193;
– коефіцієнт множинної детермінації дорівнює 0,15118817;
– статистика Дубліна Ватсона становить 1, 26055939255469.

Гіпотезу про нормальний розподіл можна перевірити стандартними статистичними тестами.

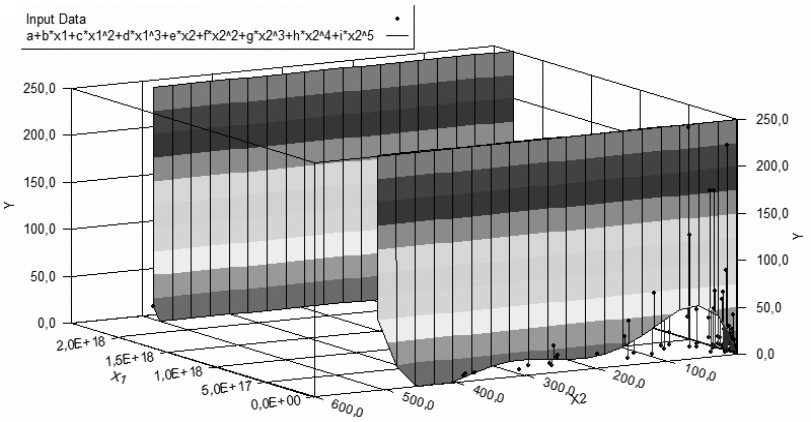

Рисунок 4.5 – Графічний вигляд даних описаних математичної моделлю

Залишки при побудові даної моделі представлені на рис. 4.6.

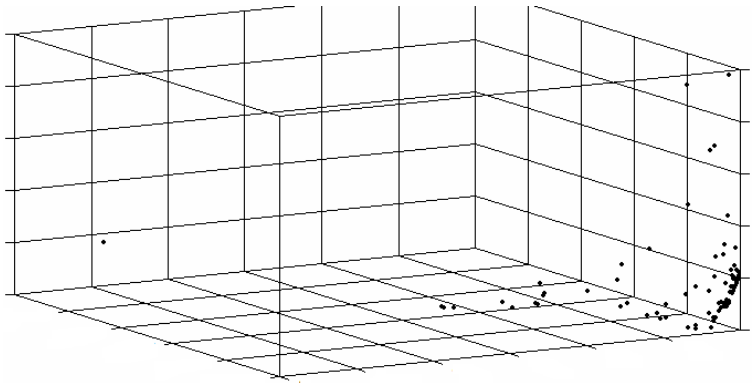

Рисунок 4.6 – Графічний вигляд залишку

Застосування даної моделі поліпшило час виконання етапу побудови графа в 69,23% випадків. В 20,51% випадків час виконання погіршився. Негативний результат застосування моделі отриманий в 5,12% випадків. В середньому час виконання покращилося на 169%.

Порівняння часу виконання до і після застосування моделі, в залежності від кількості елементів у оброблюваної вибіркою, представлено на рисунку 4.7, залежно від отриманого значення k представлено на рисунку 4.8.

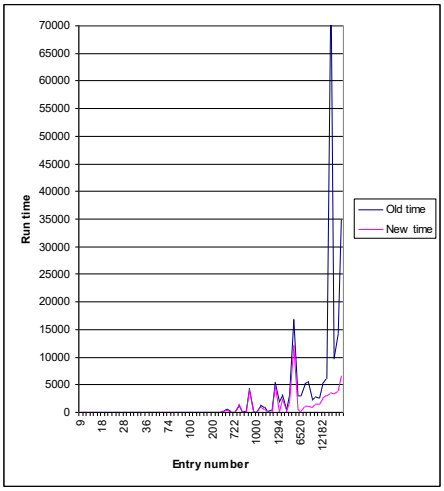

Рисунок 4.7 – Залежність часу побудови симетричного графа залежно від кількості елементів.

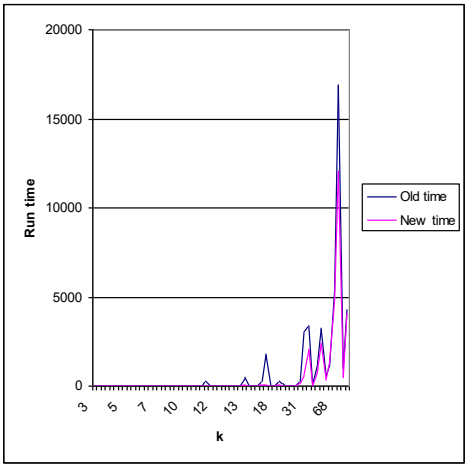

Рисунок 4.8 – Залежність часу побудови симетричного графа залежно від отриманого значення k.

Використання моделі особливо критично для великих вибірок. Отримані результати будуть використані для подальших досліджень і модифікацій алгоритму Хамелеон.

4.3 Математична модель вибору алгоритмів в рамках модифікованого алгоритму Хамелеон залежно від заданої властивості вхідних даних

Керовані параметри – характеристики вибірки такі як максимальні і мінімальні значення математичного очікування, дисперсії, розкиду і який вираховується параметр представлений раніше.

Складовими цільового параметра є:

– алгоритм побудови графа;
– міра відстані;
– алгоритм огрубіння графа;
– алгоритм початкового поділу графа;
– алгоритм відновлення графа.

На підставі наявних алгоритмів складено 14784 комбінацій для аналізу. Гіпотезу про нормальний розподіл можна перевірити

Математична модель, отримана на підставі цих даних представлена на формулі 4.4.

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

де x1-x9 – характеристики вибірок на підставі яких побудована модель.

Значення коефіцієнтів представлені в таблиці 4.4.

Перевірка на точність моделі є лише першим етапом процедури дослідження адекватності, а отже завжди слід здійснювати, як перевірку точності моделі, так і її адекватність.

Таблиця 4.4 – Значення коефіцієнтів моделі вибору алгоритмів в рамках модифікованого алгоритму Хамелеон.

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

В ході дослідження була встановлена недоцільність використання симетричного алгоритму побудови графа Побудова математичної моделі проводилась на підставі результатів експеримент поділу експериментальних вибірок за допомогою різних комбінацій алгоритмів в рамках модифікованого алгоритму Хамелеон.

ВИСНОВКИ

У пояснювальній записці приведено вирішення завдання кластеризації великих обсягів лінійно нероздільних зашумленних даних, а також вибірок з різними статистичними характеристиками. Рішення цього завдання полягає у застосуванні модифікації алгоритму Хамелеон з тією комбінацією методів і алгоритмів на кожному етапі алгоритму, яка максимально підходить для вхідних вибірки на основі статистичних характеристик вибірки. Розроблена модифікація методу дозволяє вирішити завдання кластеризації на підставі індивідуального підходу до кожної вибіркою і в той же час є універсальним для Великої кількості різних вибірок. Створених моделей дозволяють скоротити час на кластеризацію, уніфікувати технологію кластеризації, у ряді випадків підвищити якість оцінювання результатів, ґрунтуючись на кількох методах. Всі ці результати мають важливе наукове і практичне значення в області кластеризації даних і data mining.

У роботі проведено дослідження та аналіз методів, засобів і технологій, що застосовуються під час роботи з даними, для кластеризації і класифікації.

У процесі дослідження в першому розділі роботи виявлено, що на даний момент існує дуже велика кількість методів кластеризації. Існуючі методи вельми різноманітні, мають велику кількість обмежень, що накладаються на вибірки для обробки. Наведена класифікація методів і відображені основні характеристики підходів, переваги і недоліки. Розглянуті алгоритми CHAMELEON, CURE, BIRCH, і ROCK як найбільш актуальні при вирішенні завдання кластеризації великих обсягів зашумленних лінійно нероздільних даних. Зроблено висновок, що метод Хамелеон може бути модифіковано, завдяки чому підвищиться швидкість кластеризації даних та алгоритм стане більш універсальним.

Метою роботи є розробка універсальної математичної моделі залежності вибору комбінації алгоритмів на різних етапах алгоритму Хамелеон від вихідних

характеристик аналізованого набору даних з метою поліпшення якості кластеризації. Були сформульовані завдання дослідження:

– ослідження та розробка модифікації алгоритму Хамелеон для кластеризації різних вибірок даних, наводиться в другому розділі бакалаврської роботи;
– вдосконалення методів роботи з графами на окремих етапах алгоритму Хамелеон, наводиться в другому розділі дисертаційної роботи;
– розробка математичної моделі вибору k при побудові графа K найближчих сусідів на підставі характеристик вибірки, наводиться у третьому розділі роботи;
– розробка математичної моделі вибору найкращого методу для кластеризації вибірок на підставі характеристик вхідних даних, наводиться у третьому розділі роботи;
– розробка і створення програмної системи кластеризації різних наборів даних на підставі модифікованих методів, наводиться у четвертому розділі роботи.

В другому розділі вироблено дослідження та аналіз алгоритму Хамелеон. На базі даного методу розроблений модифікований метод кластеризації даних. Виділені і проаналізовані основні етапи алгоритму. Виявлено, що зазначене алгоритм потребує у модифікації для отримання можливості обробки різних вибірок різними методами на кожному з етапів алгоритму, алгоритм легко розширюється і на кожному з етапів можуть бути застосовані різні методи без впливу на інші етапи. Проаналізований етап алгоритму Хамелеон – побудова графа. У підрозділі описані запропоновані алгоритми для модифікації даного етапу. Описані можливі параметри. Визначено область для впровадження нового методу визначення k при побудові k-nn графа, що дозволяє скоротить час і витрати при виборі даного параметра. Описані запропоновані методи для модифікації огрубіння графа в рамках алгоритму Хамелеон. Описано 12 запропонованих алгоритмів для модифікації даного етапу. Проаналізований етап алгоритму Хамелеон – поділ

графа. Запропоновано та описано 16 алгоритмів для модифікації даного етапу. Досліджений етап відновлення і поліпшення графа. Запропонована модифікація алгоритму 8 додатковими методами. Проведено аналіз об 'єднання схожих класів для отримання фінального розбиття як фінального етапу модифікуємо алгоритму.

В третьому розділі розроблена модель даних для аналізу вхідних параметрів вибірки перед роботою алгоритму. Приведені загальні Положення моделі даних. Виділені і описані характеристики вибірок вхідних даних, які необхідно враховувати при побудові моделі алгоритму кластеризації, і як вхідні параметри моделі при проведенні кластеризації модифікованим алгоритмом Хамелеон. Проведено огляд можливостей отримання даних для аналізу. Також досліджені і проаналізовані методики оцінки якості поділу, отриманого в результаті роботи алгоритму. Розроблена Загальна структура підходу до оцінки результатів кластеризації на підставі модифікованого алгоритму Хамелеон. Обрані, досліджені і описані різні методи оцінки результатів кластеризації. Сукупність методів дозволяє оцінити якість поділу, вироблене алгоритмом для різних вибірок з різними вхідними характеристиками. Проаналізовані вхідні дані, необхідні для експерименту і оцінки якості роботи модифікованого алгоритму Хамелеон. Проведено аналіз можливих підходів до генерації експериментальних вибірок. Запропонована модель побудови 3d вибірок з необхідними характеристиками і розташуванням об'єктів у просторі. Проаналізовані джерела існуючих реальних і експериментальних даних. Обрані існуючі вибірки для аналізу та згенеровано сукупність 3d вибірок для експерименту.

В четвертому розділі проведено аналіз загальної схеми побудови математичної моделі. Побудована математична модель для вибору k при побудові k-nn графа в рамках модифікованого алгоритму Хамелеон. Наведено результати експериментів застосування даної моделі. Розроблена математична модель для вибору алгоритмів в рамках модифікованого алгоритму Хамелеон. Зроблені висновки щодо доцільності використання окремих алгоритмів і методів. Наведено результати застосування розроблених методів на реальних даних.

В результаті роботи отримані наступні нові наукові результати:

– вперше побудована модель залежності якості кластеризації великих обсягів складних лінійно нероздільних зашумленних даних від характеристик даних і алгоритмів динамічного кластеризації;
– подальший розвиток отримала багаторівнева модель динамічного кластеризації даних, яка відрізняється від існуючих інтеграцією з етапами алгоритму Хамелеон. Така інтеграція дозволяє поліпшити якість моделі в аспекті роботи зі складними лінійно нероздільними зашумлен експериментальними даними;
– подальший розвиток отримав метод Хамелеон, який на відміну від існуючого використовує різні алгоритми на різних етапах кластеризації, що дозволяє враховувати різницю у даних і використовувати методи, які працюють краще на конкретних даних;
– удосконалено метод побудови графа, який дозволяє прискорити процес побудови графа через вибір оптимального k, ґрунтуючись на характеристиках аналізованих даних.

Практична цінність роботи полягає в наступному:

– створено модель вибору k для алгоритму K найближчих сусідів, заснована на характеристиках даних;
– створено модель залежності якості кластеризації великих обсягів складних лінійно нероздільних зашумленних даних від характеристик даних і алгоритмів динамічного кластеризації.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Ляховець А.В. Экспериментальные результаты исследования качества кластеризации разнообразных наборов данных с помощью модифицированного алгоритма Хамелеон.//Вісник запорізького національного університету. 2011 №2. С. 86-73.

2. Ляховець А.В. Характеристики выборки данных для выбора k при построении графа к-ближайших соседей.//Тези доповіді VI Міжнародної науково-практичної конференції «Сучасні проблеми і досягнення в галузі радіотехніки, телекомунікацій та інформаційних технологій». 2012. С. 168-169.

3. Ляховець А.В. Характеристики выборки данных для выбора k при построении графа к-ближайших соседей.//Тези доповіді XXII Міжнародної науково-технычної конференції «Проблемы информатики и моделирования ПИМ 2012». 2012. С. 59.

4. Ляховець А.В. Кластеризация с помощью нейронной сети Кохонена и модифицированного алгоритма иерархической кластеризации Хамелеон в различных предметных областях.//Науково-технічний журнал «Регистрация ,хранение и обработка данных». 2013. С. 53.

5. Anil K. Jain. Data clustering: 50 years beyond K-means. – Мічіган: Pattern Recognition Letters, 2010. – 658 с.

6. Michael E Geisser, Andrew J Haig, Henry C Tong. Spinal canal size and clinical symptoms among persons diagnosed with lumbar spinal stenosis.//The Clinical journal of pain. 2007. С. 780-785.

7. S. Sumathi. Fundamentals of relational database management systems – Берлін: Springer-Verlag, 2007. – 359 с.

8. Dr.K.Thangadurai, M.Uma, Dr.M.Punithavalli. A Study On Rough Clustering. - Global Journal of Computer Science and Technology Vol. 10 Issue 5 Ver. 1.0 July 2010

9. Gan, Guojun, Chaoqun Ma, and Jianhong Wu, Data Clustering: Theory, Algorithms, and Applications, ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA, 2007.

10. Nitin Agarwal and Huan Liu."Modeling and Data Mining in Blogosphere". Synthesis Lectures on Data Mining and Knowledge Discovery No. 1, Morgan and Claypool Publishers, Robert Grossman (Editor), August 2009. Morgan and Claypool webiste; Amazon website.

11. F. Sha, G. Gardarin Gradual clustering algorithms for metric spaces In Proceedings: Proc. 15èmes Journées Bases de Données Avancées, *BDA'99* (October 1999)

12. S.Thirumuruganathan A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm Posted:17 may 2010 http://saravanaganathan.wordpress.com/2010/introduction-to-k-nearest-algorithm/

13. Саламов Асаф Ага Джавад оглы. Компьютерный анализ вторичной структуры глобулярных белков. – Новосибирск: Институт цитологии и генетики, 1991. – 568 с.

14. Соловйов В.В., Саламов А.А., Капитонов В.В. Факторы, определяющие формирование вторичной структуры глобулярных белков // Молекулярная биология. 1991. С. 78.

15. Secondary Structure by Combining Nearest-neighbor Algorithm and Multiple Sequence Alignments. J. Mol. Biol., 247:11-15, 1995.

16. Multilevel graph partitioning schemes / G.Karypis,V.Kumar. - Minneapolis(Mn) : 1995. : (UMSI research report) / Univ.of Minnesota;

17. Sung Kyu Lim Practical Problems in VLSI Physical Design Automation. Springer; 1 edition (September 10, 2008)

18. Paralel Dynamic Load-Balancin for Adaptive Distributive Memory PDE solvers Nasir Touhed. Submitted in accordance with the requirements for the degree of Doctor of Philosophy The university of Leeds 1998

19. Randolf Rotta A Multi-Level Algorithm for Modularity Graph Clustering Diploma thesis. Institut fur Informatik, Lehrstuhl Software-Systemtechnik

20. Деревянко А.П., Холюшкин Ю.П., Воронин В.Т., Ростовцев П.С. и др. Математические методы в археологических реконструкциях. - «СО РАН» Новосибирск, 1995.