Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту
(повна назва)
Кафедра Інформатики
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти перший (бакалаврський)

РОЗРОБКА ТА МОДЕЛЮВАННЯ МЕТОДУ ОНЛАЙН-КЛАСТЕРИЗАЦІЇ ДАНИХ ЗА УМОВ ПЕРЕТИННИХ КЛАСТЕРІВ
(тема)

Виконав:
студент 4 курсу, групи ІТІНФ-19-2

Фалько М.К.
(прізвище, ініціали)

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

Тип програми освітньо-професійна

Освітня програма Інформатика
(повна назва освітньої програми)

Керівник доц. Бєлова Н.В.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ Кобилін О.А.
(підпис) (прізвище, ініціали)

2023 р.

Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту
(повна назва)

Кафедра Інформатики
(повна назва)

Рівень вищої освіти перший (бакалаврський)

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

Тип програми освітньо-професійна

Освітня програма Інформатика
(повна назва освітньої програми)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 2023 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Фальку Михайлу Костянтиновичу
(прізвище, ім'я, по батькові)

1. Тема роботи Розробка та моделювання методу онлайн-кластеризації даних за умов перетинних кластерів

затверджена наказом університету від 15 травня 2023 року № 474 Ст

2. Термін подання студентом роботи до екзаменаційної комісії 22 травня 2023 р.

3. Вихідні дані до роботи науково-методична та науково-технічна література, матеріали конференцій, дані інтернет-мережі, UCI репозиторій, середовище розробки MATLAB.

4. Перелік питань, що потрібно опрацювати в роботі

1. Огляд та аналіз існуючих методів кластеризації за умов перетинних кластерів.
2. Розробка та моделювання методу онлайн-кластеризації даних.
3. Підбір програмних засобів та програмна реалізація методу.
4. Порівняльний аналіз роботи запропонованого методу з більш відомими класичними алгоритмами кластеризації.

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри) Аналіз предметної області, постановка задачі, онлайн-кластеризація даних за умов перетинних кластерів, аналіз отриманих результатів.

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

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Консультант з дотримання діючих стандартів та норм	Доцент Творошенко І.С.		

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	10.04.2023	
2	Аналіз завдання, підбір літератури	11.04.23-17.04.23	
3	Аналіз літератури з досліджуваної проблеми	18.04.23-20.04.23	
4	Аналіз технічних засобів	21.04.23-30.04.23	
5	Розробка методу онлайн-кластеризації даних за умов перетинних кластерів	01.05.23-14.05.23	
6	Програмна реалізація	15.05.23-23.05.23	
7	Оформлення пояснювальної записки	24.05.23-26.05.23	
8	Перевірка на плагіат	27.05.23	
9	Рецензування	28.05.23	
10	Підготовка презентації та доповіді	29.05.23-30.05.23	
11	Занесення роботи в електронний архів	31.05.23	
12	Попередній захист кваліфікаційної роботи	31.05.23	

Дата видачі завдання 10 квітня 2023 р.

Студент _____
(підпис)

Керівник роботи _____ _____доц. Бєлова Н.В._____
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ/ABSTRACT

Пояснювальна записка до кваліфікаційної роботи: 58 с., 3 табл., 11 рис., 39 джерел.

НЕЧІТКІ НЕЙРОННІ МЕРЕЖІ, МАШИННЕ НАВЧАННЯ, ДОСТОВІРНА НЕЧІТКА КЛАСТЕРИЗАЦІЯ, ПЕРЕТИННІ КЛАСТЕРИ, МІРА ПОДІБНОСТІ СПЕЦІАЛЬНОГО ТИПУ.

Об’єктом роботи є тестові набори даних Іриси Фішера та Вина із Uci репозиторію для методів кластеризації та класифікації даних.

Метою роботи є розробка та моделювання методу нечіткої кластеризації даних, що базуються на теорії достовірності та дозволяють кластеризувати данні в онлайн режимі за умов перетинних кластерів.

В якості альтернативи імовірнісним методам нечіткої кластеризації було використано теорію правдоподібності, що найкраще справляється із проблемою взаємного перетинання класів, які формуються в процесі аналізу даних.

В результаті роботи запропоновано метод достовірної нечіткої онлайн-кластеризації даних за умов перетинних кластерів.

FUZZY NEURAL NETWORKS, MACHINE LEARNING, CREDIBILISTIC FUZZY CLUSTERING, INTERSECTING CLUSTERS, SIMILARITY MEASURE OF SPECIAL TYPE.

The object of the work is the test Iris Fisher and Wine data sets from the Uci repository for data clustering and classification methods.

The purpose of the work is to develop and model a method of fuzzy clustering of data, based on the theory of reliability and allowing clustering of data online under the conditions of intersecting clusters.

As an alternative to the probabilistic methods of fuzzy clustering, the probability theory was used, which best copes with the problem of mutual crossing of classes that are formed in the process of data analysis.

As a result of the work, a method of reliable fuzzy online clustering of data under the conditions of intersecting clusters is proposed.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

FCM – Fuzzy *C*-Means (нечіткі *c*-середні)

СКМ – системи комп’ютерної математики

CHI – Calinsky-Harabas Index (індекс Калінскі-Харабаса)

DBI – Davis-Baldwin Index (індекс Девіса-Болдуіна)

SI – Silhouette Index (індекс силуету)

CFC – Credibilistic Fuzzy Clustering (достовірна нечітка кластеризація)

ВСТУП

Завдання кластеризації (класифікації без урахування) є складовою загальної проблеми Data Mining, на вирішення якої сьогодні розроблено безліч підходів, методів, алгоритмів. У межах цієї задачі особливе місце посідає задача нечіткої кластеризації, яка розглядає ситуацію, коли формовані класи взаємно перекриваються, тобто кожне спостереження може одночасно належати до кількох або всіх класів-кластерів. В рамках цієї підзадачі на сьогодні сформувалося два основні підходи: імовірнісний, коли для кожного спостереження оцінюється ймовірність його приналежності до кожного з можливих класів, і можливісний, де оцінюється можливість (не імовірнісна) приналежності до деяких класів. Обидва ці підходи пов'язані з вирішенням задачі оптимізації (нелінійного програмування) з адоптованою цільовою функцією і в загальному випадку можуть призводити до різних кінцевих результатів. Незважаючи на досить серйозну математичну основу цих підходів, вони страждають від ряду суттєвих недоліків: такий імовірнісний підхід дуже чутливий до «аномальних» спостережень, які практично «розмазуються» з однаковими рівнями належності всім кластерам.

Можливісний підхід, в свою чергу, пов’язаний із так званою «проблемою збіжності», коли деякі кластери зливаються разом, що взагалі не дозволяє розбити оброблювану вибірку на однорідні групи-кластери.

Обидва розглянуті підходи обробляють дані в пакетному режимі, тобто мається на увазі, що весь масив спостережень заданий апріорі і не змінюється у процесі аналізу. Якщо ж дані надходять в режимі онлайн (задачі Data Stream Mining), класичні ймовірнісні і можливі алгоритми нечіткої кластеризації стають непрацездатними. У цій ситуації першому плані виходять послідовні алгоритми, засновані на градієнтній оптимізації прийнятих цільових функцій. Також онлайн процедури були розроблені як у рамках імовірнісного, так і можливісного підходів та підтвердили свою працездатність.

У більшості завдань кластеризації, пов’язаних з обробкою реальних даних, вихідна інформація, як правило, спотворена аномальними викидами (перешкодами) і пропусками, причому кількість цих викидів і «дір» може бути порівнянна з обсягом «чистих» даних, при цьому можлива ситуація, коли усі дані є «брудними». Зрозуміло, що «класичні» методи у цій ситуації неефективні.

Для боротьби з аномальними викидами в задачах нечіткої кластеризації були запропоновані робасні методи, засновані на використанні як робасних цільових функцій спеціального виду, так і подібних заходів, нечутливих до викидів, і призначені для роботи як в пакетному, так і послідовному режимах.

Що ж до відсутніх спостережень, що містять пропуски, то тут також було розроблено низку методів (у рамках ймовірнісного та можливісного підходів) як пакетних, так і онлайн.

1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАВДАННЯ

Інтелектуальний аналіз даних представляє собою процес виявлення придатну для використання інформацію (закономінність) у великих наборах даних. В інтелектуальному аналізі даних застосовується математичний аналіз для виявлення закономірностей і тенденцій, існуючих в даних. Зазвичай такі закономірності не можна виявити традиційно при перегляді даних, оскільки зв’язок надто складний [1-4].

Ці закономірності та тренди можна зібрати разом і визначити як модель інтелектуального аналізу даних. Моделі інтелектуального аналізу даних можна застосовувати до конкретних сценаріїв, а саме:

– прогнозування: оцінка продажів, прогнозування навантаження на сервер або час простого сервера;

– ризик і ймовірність: вибір найбільш підходящих замовників для цільової розсилки, визначення точок рівноваги для ризикованих сценаріїв, визначення ймовірних діагнозів або інших результатів;

– рекомендації: визначення продуктів, які з високою ймовірністю можуть бути продані разом, створення рекомендацій;

– пошук послідовностей: аналіз вибору замовників під час вдосконалення покупок, прогнозування наступного можливого події;

– групування: розподіл замовників або подій на кластери, що містять елементи, аналіз і прогнозування загальних рис.

Побудова моделі інтелектуального аналізу даних є частиною більш масштабного процесу, в який входять усі завдання, від формулювання питань щодо даних і створення моделей для відповідей на ці питання до розвертання моделей у робочому середовищі. Цей процес можна представити як послідовність наступних шести базових кроків:

Крок 1. Постановка задачі.

Крок 2. Підготовка даних.

Крок 3. Вивчення даних.

Крок 4. Побудова моделей.

Крок 5. Дослідження та перевірка моделей.

Крок 6. Розгортання та оновлення моделей.

На рисунку 1.1 описані зв’язки між кожним етапом процесу та технологіями в Microsoft SQL Server, які можна використовувати для виконання кожного етапу.

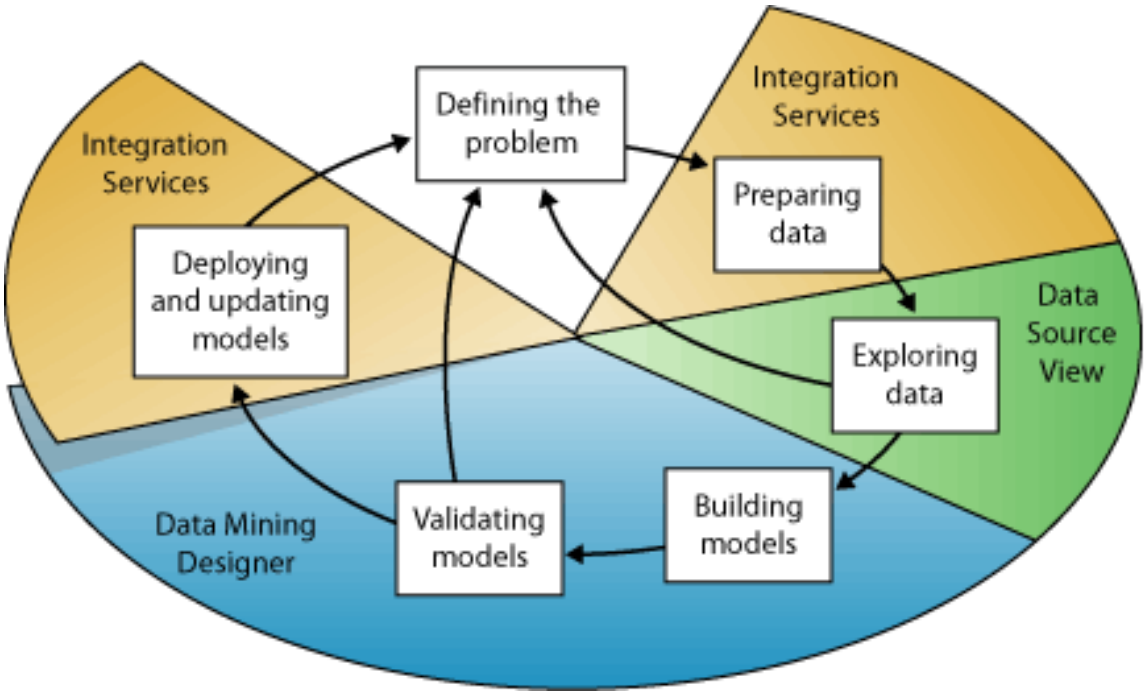


Рисунок 1.1 – Основні етапи інтелектуального аналізу даних

Процес, представлений на рисунку 1.1, є циклічним, тобто створення аналітичної моделі даних є динамічним і повторюваним процесом. Виконавши перегляд даних, користувач може виявити, що даних недостатньо для створення необхідних моделей інтелектуального аналізу даних, що веде до необхідності пошуку додаткових даних. Також може виникнути ситуація, коли після побудови кількох моделей виявиться, що вони не дають адекватну відповідь на поставлену задачу, і тому необхідно поставити задачу по-іншому. Може виникнути необхідність в оновленні вже розвернутих моделей для рахунку нових отриманих даних. Для створення хорошої моделі можна багаторазово повторити кожен крок процесу.

Для прийняття правильних рішень при створенні моделі інтелектуального аналізу даних необхідно приймати дані. Методи дослідження даних включають в себе розрахунок мінімальних і максимальних значень, вирахування середнього і стандартного відхилення і вивчення розподілу даних. Наприклад, за максимальним, мінімальним і середнім значенням можна зазначити, що вибір даних не є репрезентативним для наявних клієнтів або бізнес-процесів, і тому необхідно отримати більш збалансовані дані або змінити прогнози, що лежать в основі очікуваних результатів. Стандартні відхилення та інші характеристики розподілу можуть повідомити корисні відомості про стабільність і точність результатів. Більша величина стандартного відхилення може свідчити про те, що додавання нових даних допоможе вдосконалити модель. Дані, які сильно відхиляються від стандартного розподілу, можуть виявитися виявленими або показати точну картину реальної проблеми, яка робить складним підбір відповідної моделі для даних.

Вивчення даних у всіх власних представленнях про бізнес-проблеми може призвести до висновку про наявність помилок в наборі даних, а потім можна виробити стратегію для усунення проблеми або отримати більш глибоке представлення про моделі поведінки, характерних для бізнесу.

При обробці потоків даних класичні методи кластерного аналізу не пригодні. У таких випадках використовуються неієрархічні методи, засновані на розділенні, які представляють собою ітераційні методи уточнення вихідної сукупності. В процесі ділення нові кластери формуються до тих пор, поки не буде виконано правило встановлення.

Така неієрархічна кластеризація складається в розділенні набору даних на певну кількість окремих кластерів. Існує два підходи. Перший полягає у визначенні меж кластерів як найбільш щільних ділянок у багатомірному просторі вихідних даних, тобто визначення кластера там, де є велике «згущення точок». Другий підхід полягає в мінімізації заходів відмінності об’єктів.

На рисунку 1.2 продемонстровані групи методів кластерного аналізу.

Рисунок 1.2 – Групи методів кластерного аналізу

Задача кластеризації масивів багатовимірних даних, що задані у вигляді векторів ознак іноді досить великих розмірностей є невід'ємною частиною загальної проблеми інтелектуального аналізу даних (Data Mining), а для її вирішення на цей час запропонована велика кількість підходів, методів, алгоритмів від найпростіших типу *K*-середніх до досить складних, заснованих на аналізі щільностей розподілу даних високої розмірності у великих масивах.

Особливе місце тут займають методи нечіткої кластеризації, що синтезовані за припущень взаємного перетину класів, які формуються в процесі аналізу даних [1, 2, 5].

На цей час найбільше розповсюдження отримали алгоритми ймовірнісної нечіткої кластеризації, найбільш популярним з яких є метод нечітких *C*-середніх (FCM) Дж. Бездека [2], завдяки, перш за все, простоті чисельної реалізації. В той же час цей підхід не позбавлений і суттєвих недоліків, пов'язаних з жорсткими «ймовірнісними» обмеженнями на рівні

належностей та підвищеною чутливістю до аномальних викидів, що досить часто є присутніми у вихідних масивах даних.

Цих недоліків в значній мірі позбавлений метод можливісної нечіткої кластеризації [6] більш стійкий до впливу різного типу шумів та збурень у даних. У той же час цей підхід потерпає від, так званої, проблеми співпадіння, коли кластери, що формуються у процесі обчислень, починають «зливатися» у один великий клас.

Тому в якості альтернативи ймовірнісному та можливісному підходам було запропоновано, так званий, довірчий підхід [7-9], заснований на теорії довіри [10, 11], при цьому в процесі нечіткої кластеризації розраховуються не лише центроїди кластерів та рівні належності спостережень до цих кластерів, але й рівні довіри до отриманих результатів.

1.1 Кластерний аналіз

Кластерний аналіз (англ. Data clustering) – задача розбиття заданої вибірки об’єктів (ситуацій) на підмножини, які називаються кластерами, так, щоб кожен кластер складався з схожих об’єктів, а об’єкти різних кластерів істотно відрізнялися. Завдання кластеризації відноситься до статистичної обробки, а також до широкого класу завдань навчання без вчителя.

Кластерний аналіз – це не якийсь один алгоритм, а загальна задача, для розв’язання якої використовуються різні підходи. Зокрема, алгоритми побудови кластерів можуть суттєво відрізнятись у розумінні того, що відносити в один кластер і як їх ефективно шукати.

Серед популярних концепцій кластерів є групи з елементами, які утворюються ґрунтуючись на відстані між ними, на щільності ділянок у просторі даних, інтервалах або на конкретних статистичних розподілах. Тому кластеризація може бути сформульована як задача багатокритеріальної оптимізації.

Відповідний алгоритм кластеризації та вибору параметрів (включаючи такі параметри, як функція відстані, порогове значення щільності або кількість очікуваних кластерів) залежать від конкретного набору даних та мети використання результатів. Кластерний аналіз як такий є не автоматизованим завданням, а ітераційним процесом виявлення знань або інтерактивної багатокритеріальної оптимізації, який містить спроби та невдачі. Часто доводиться змінювати процес опрацювання даних та параметри моделі поки не буде отримано з результат з заданими властивостями.

Окрім терміну кластеризація існує багато термінів з аналогічним значенням, серед яких автоматична класифікація, числова таксономія та типологічний аналіз. Тонкі розбіжності часто полягають у використанні результатів: для добування даних, отримані групи є предметом інтересу, при автоматичній класифікації, навпаки, більш важливий степінь розбіжності.

Кластерний аналіз походить з антропології, де він був започаткований Драйвером і Крьобером у 1932 році. В психологію він був введений Зубіним у 1938 році і Робертом Тріоном у 1939 році. Став відомий завдяки використанню Кеттелем для класифікації теорії ознак в психології особистості, починаючи з 1943 року.

Кластеризація даних є процесом розподілу елементів даних на класи або групи так, що елементи в одному класі є якомога близькими, а елементи різних класів є настільки різнорідними, наскільки це можливо. Залежно від характеру даних та мети кластеризації можуть використовуватися різні міри подібності для розміщення елементів в класах, причому міра подібності визначає самі кластери. Приклади мір, які можуть бути використані для кластеризації, включають відстань, зв’язок та інтенсивність.

У жорсткій кластеризації, дані розділені на окремі кластери, де кожен елемент даних належить одному кластеру. В нечіткій кластеризації, елементи даних можуть належати до більш ніж одного тематичного напряму, і з кожним елементом множини пов’язана функція належності до кожного кластеру. Вона вказує на силу зв’язку між цим елементом даних і конкретною групою. Нечітка

кластеризація є процесом присвоєння цих мір належності та їх використання для визначення складу кожного з кластерів.

Кластерний аналіз – це багатовимірна статистична процедура, яка виконує збір даних, що містять інформацію про вибірку об’єктів і потім упорядковує об’єкти в порівняно однорідні групи-кластери (Q-кластеризація, або Q-техніка, власне кластерний аналіз).

Основна мета кластерного аналізу – знаходження груп схожих об’єктів у вибірці. Спектр застосувань кластерного аналізу дуже широкий: його використовують в археології, антропології, медицині, психології, хімії, біології, державному управлінні, філології, маркетингу, соціології та інших дисциплінах.

Однак універсальність застосування привела до появи великої кількості несумісних термінів, методів і підходів, що ускладнюють однозначне використання і несуперечливу інтерпретацію кластерного аналізу.

1.2 Методи нечіткої кластеризації

Нечіткий кластерний аналіз використовується при побудові нейро-нечітких систем для визначення нечітких множин, якщо вони невідомі апріорі. Нечіткі множини знаходяться як проекції кластерів на кожну розмірність. Можливо поєднувати апріорні знання з кластерним аналізом, використовуючи його для уточнення параметрів функції приналежності. Недоліком такого методу визначення нечітких множин є складність їхньої інтерпретації.

Більшість методів нечіткої кластеризації спрямовані на мінімізацію суми [12-16]:

$$Goal(a,\mu,c)=\sum_{k=1}^{N}\sum_{q=1}^{m}\left(\mu_q^{\beta}d^2\left(a,c_q\right)\right),$$

при виконанні умов

$$\sum_{q=1}^{N}\mu_q(k)=1,0<\sum_{q=1}^{N}\mu_q(k)<N,$$

де $\mu_q(k)$ – рівень нечіткої належності спостереження $a(k)$ до q-го кластера;

c_q – прототип центроїд q-го кластеру;

$\beta>1$ – параметр фаззіфікації, що задає «розмитість» меж кластерів;

$d(a(k),c_q)$ – відстань між $a(k)$ та c_q у прийнятій метриці.

Вибір відстані між об’єктами є вузловим моментом дослідження, від нього залежить остаточний варіант розбиття об’єктів на класи при даному алгоритмі розбиття.

Найбільш простий шлях обчислення відстаней між об’єктами у багатовимірному просторі полягає у обчисленні евклідових відстаней. Якщо є дво- чи тривимірний простір, то цей захід є реальною геометричною відстанню між об’єктами в просторі.

$$d^2(a_k,c_q)=\|a_k-c_q\|_2^2.$$

Відстань міських кварталів (манхеттенська відстань). Ця відстань є просто сумою модулів різниць за координатами. У більшості випадків ця міра відстані призводить до таких же результатів, як і для звичайної відстані Евкліда. Однак зазначимо, що для цього заходу вплив окремих великих різниць (викидів) зменшується (оскільки вони не зводяться у квадрат). Манхеттенська відстань обчислюється за такою формулою:

$$d(a_k,c_q)=\|a_k-c_q\|_1=\sum_{i=1}^{n}|a_{ki}-c_{qi}|.$$

Вибір відстані між об’єктами є фокусом дослідження, від якого залежить кінцевий варіант розбиття об’єктів на класи для заданого алгоритму розбиття. Найпростішим способом обчислення відстані між об’єктами в багатовимірному просторі є обчислення евклідових відстаней, але евклідова метрика (і її квадрат) обчислюється з джерела, а не зі стандартизованих даних. У цьому випадку пропонується використовувати часткову відстань, описану формулою нижче:

$$d_p^2\left(a_k, c_q\right) = \frac{n}{\delta_{k\Sigma}} \sum_{i=1}^{n} \left(a_{ki} - c_{qi}\right)^2 \delta_{ki},$$

де c_{qi}-і компонент q-го прототипу (центроїда) відповідного кластера ($q = 1,2,...,m$),

$$\delta_{ki} = \begin{cases} 0 \,|\, a_{ki} \in A_G, \\ 1 \,|\, a_{ki} \in A_F, \end{cases}$$

$$\delta_{k\Sigma} = \sum_{i=1}^{n} \delta_{ki} \,,$$

де

$$A_F = \{a_k \in A \,|\, a_k \ - \text{ вектор, що містить всі складові}\};$$

$$A_P = \{a_{ki}, 1 \le i \le n, 1 \le k \le N \,|\, \text{ все значення } a_k, \text{що містяться в } A\};$$

$$A_G = \{a_{ki} = ?, 1 \le i \le n, 1 \le k \le N \,|\, \text{ всі значення } a_k, \text{що відсутні в } A\}.$$

Легко побачити, що часткова відстань [7] стає звичайною евклідовою метрикою. У протилежному випадку відстань між прототипом оцінюється на основі наявних компонентів, як показано на рисунку 1.3.

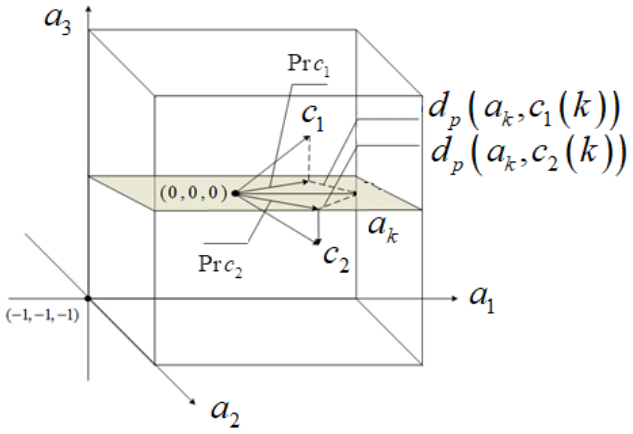


Рисунок 1.3 – Стратегія часткової відстані

Тут у тривимірному векторі a_k не вистачає однієї компоненти a_{k3}, тому відстань вимірюється на площині a_1, a_2, а замість прототипів і використовуємо їх проекції на цю площину $\Pr c_1$ і $\Pr c_2$.

Процес змагання, який лежить в основі навчання карти Кохонена [3], організований на основі оцінки часткових відстаней, тобто коли спотворений (неповний) вектор спостережень a_{k+1} надходить, спочатку оцінюється відстань між цим вектором і центроїдами $c_1(k), c_2(k), ..., c_m(k)$, а потім починається пошук нейрона-переможця $c_q(k)$ такий, що

$$d_p^2\left(a_{k+1}, c_q(k)\right) = \arg\min_q \left\{ d_p^2\left(a_{k+1}, c_1(k)\right), ..., d_p^2\left(a_{k+1}, c_m(k)\right) \right\}.$$

Далі відсутній компонент замінюється відповідним компонентом центроїда нейрона-переможця:

$$a_{k+1,i} = c_{qi}(k).$$

Більш складні методи кластеризації шукають кластери як гіпер еліпсоїди різного розміру [12, 17]. Такі методи називають частковими, вони не можуть вірно опрацьовувати шуми та викиди і віднаходити кластери з неопуклими поверхнями. Для проведення кластерного аналізу за допомогою часткового методу необхідно задати його параметри: діапазон значень змінних, кількість кластерів для кожної із змінних (або їх ширину), функцію приналежності, що описує кластери та інші параметри в залежності від обраного методу кластеризації.

За допомогою ієрархічних методів можна віднайти кластери, об'єднуючи менші кластери та розподіляючи більші [18, 19]. Таким чином знаходиться дерево кластерів, на різних рівнях якого можна отримати різне розподілення на кластери.

Щільнісні методи та сіткові методи дозволяють розподіляти на кластери різного розміру довільно розподілені екземпляри [20-24]. Вони також добре впізнають шуми та викиди, але потребують ретельного вибору параметрів, необхідних для реалізації методу.

Множину ознак у об'єктів кластеризації слід вибирати так, щоб всі значення ознак були виміряні в шкалі відношень або шкалі інтервалів. У цьому випадку результати нечіткої кластеризації мають змістовну інтерпретацію, адекватну проблемі знаходження нечітких кластерів.

Інтервальна шкала. В процесі виміру ознаки об'єкту ставиться у відповідність, як правило, деяке дійсне число, рівне значенню цієї ознаки. Допустимим перетворенням в шкалах інтервалів є довільна лінійна зростаюча функція між двома множинами значень ознак. Характерною властивістю цієї шкали є відсутність абсолютного нуля. Приклад ознаки, вимірюваної в інтервальній шкалі: температура в шкалах Цельсія і Фаренгейта.

Шкала відношень. В процесі виміру ознаки об'єкту ставиться у відповідність також деяке дійсне число, рівне значенню цієї ознаки. Допустимим перетворенням в шкалах відношень є довільна лінійна зростаюча функція, що проходить через нуль. Характерною властивістю цієї шкали є

наявність абсолютного нуля. Приклади ознак, вимірюваних в шкалі відношень: відстань в метрах і футах, маса в кілограмах і фунтах, швидкість в км/год і вузлах.

1.3 Метод FCM

Алгоритм нечіткої кластеризації називають FCM – алгоритмом (Fuzzy Classifier Means, Fuzzy *C*-Means).

Метою FCM – алгоритму кластеризації є автоматична класифікація множини об'єктів, які задаються векторами ознак в просторі ознак. Іншими словами, такий алгоритм визначає кластери і, відповідно, класифікує об'єкти. Кластери представляються нечіткими множинами і, крім того, межі між кластерами також є нечіткими.

FCM – алгоритм кластеризації припускає, що об'єкти належать всім кластерам з певною функцією належності. Ступінь належності визначається відстанню від об'єкту до відповідних кластерних центрів. Даний алгоритм ітеративно обчислює центри кластерів і нові ступені належності об'єктів.

Для заданої множини $A=\{a(1),a(2),...,a(k),...a(N)\}\subset P^n$, де $a(k)\in R^n$ – вектор-спостереження, k – або номер цього спостереження в масиві даних A, і кластерів c_q, що виділяються, передбачається, що будь-який належить будь-якому із ступенем належності, $\mu_q(k)\in[0,1]$, де q номер кластеру.

Мета алгоритму полягає в мінімізації суми всіх зважених відстаней

$$\sum_{k=1}^{N}\sum_{q=1}^{m}\mu_q^{\beta}(k)d^2\left(a(k),c_q\right)\to\min.$$

1.4 Метод Густафсона-Кесселя

Цей метод має значно більшу обчислювальну трудомісткість у порівнянні з методом FCM.

Результати нечіткого кластер-аналізу можна використовувати для синтезу нечітких правил. Кожен кластер буде являти собою деяке нечітке правило, що узагальнює підмножину екземплярів навчаючої вибірки, найбільш тісно розташованих у просторі ознак [25-28].

Функції належності термів у посилках правила отримують проектуванням ступенів належності відповідного кластера на вхідні змінні. Потім отримані множини ступенів належностей апроксимують придатними параметричними функціями належності.

Як висновок правила бази знань вибирають координату центра кластера. Висновки правил бази знань Мамдані знаходять також як і функції належності термів вхідних змінних. Висновки правил бази знань Т. Сугено знаходять за методом найменших квадратів. При кластеризації з використанням норми Махалонобіса як висновки правил типу Сугено можуть бути обрані рівняння довгих осей гіпер еліпсоїдів.

1.5 Метод ймовірнісної нечіткої кластеризації

Метод ймовірнісної нечіткої кластеризації пов’язаний з мінімізацією цільової функції [18, 29]

$$E\left(U_q(k), w_q\right) = \sum_{k=1}^{N} \sum_{q=1}^{m} U_q^{\beta}(k) D^2\left(x(k), w_q\right), \qquad (1.1)$$

за наявності обмежень

$$\sum_{q=1}^{N} U_q(k) = 1, 0 < \sum_{q=1}^{N} U_q(k) < N, \tag{1.2}$$

де $U_q(k)$ – рівень нечіткої належності спостереження $x(k)$ до q-го кластера $Cl_q (1 \le q \le m)$;

w_q – прототип центроїд q-го кластеру, що має уточнюватися в процесі послідовної рекурентної кластеризації;

$\beta > 1$ – параметр фаззіфікації, що задає «розмитість» границь кластерів;

$D\left(x(k), w_q\right)$ – відстань між $x(k)$ та w_q у прийнятій метриці, найчастіше метриці Ітакури-Сайто, окремим випадком якої є відстань П. Махаланобіса.

Вирішення задачі нелінійного програмування (1.1), (1.2) за допомогою алгоритма Ерроу-Гурвіца-Узави веде до процедури рекурентної кластеризації

$$\begin{cases} U_q(k+1) = \dfrac{\left(D^2\left(x(k+1), w_q(k)\right)\right)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}\left(D^2\left(x(k+1), w_l(k)\right)\right)^{\frac{1}{1-\beta}}}, \\ \\ w_q(k+1) = w(k) + \eta(k+1) * \\ *U_q^{\beta}(k+1)\left(x(k+1) - w_q(k)\right), \end{cases} \tag{1.3}$$

де $\eta(k)$ – параметр кроку навчання.

При значенні фаззіфікатора $\beta = 2$ приходимо до рекурентної версії нечітких C-середніх у вигляді

$$\begin{cases} U_q(k+1) = \dfrac{\left(D^2\left(x(k+1), w_q(k)\right)\right)^{-1}}{\sum\limits_{l=1}^{m}\left(D^2\left(x(k+1), w_l(k)\right)\right)^{-1}}, \\ w_q(k+1) = w(k) + \eta(k+1)U_q^2(k+1)* \\ *\left(x(k+1) - w_q(k)\right), \end{cases} \tag{1.4}$$

при цьому цікаво помітити, що другі співвідношення (1.3), (1.4) є за суттю правилом самонавчання Т. Кохонена [10] «Переможець отримує більше» з функцією сусідства $U_q^{\beta}(k+1)$ або $U_q^2(k+1)$ на кожному кроці налаштування.

Шляхом нескладних перетворень можна переписати перше співвідношення (1.3) у вигляді

$$U_q(k+1) = \frac{1}{1 + \dfrac{D^2\left(x(k+1), w_q(k)\right)^{\beta-1}}{\sum\limits_{\substack{l=1 \\ l\neq q}}^{m} D^2\left(x(k+1), w_l(k)\right)^{\frac{1}{1-\beta}}}}, \tag{1.5}$$

або для $\beta = 2$

$$\begin{cases} U_q(k+1) = \dfrac{1}{1 + \dfrac{D^2\left(x(k+1), w_q(k)\right)}{\sigma_q^2(k+1)}}, \\ \sigma_q^2(k+1) = \left(\sum\limits_{\substack{l=1 \\ l\neq q}}^{m}\left(D^2\left(x(k+1), w_l(k)\right)^{-1}\right)\right)^{-1}, \end{cases} \tag{1.6}$$

що за суттю є функцією щільності розподілу Коші з параметром ширини $\sigma^2(k+1)$, тобто відповідає умовам, що висуваються до функцій сусідства у процедурах Т. Кохонена.

1.6 Можливісні методи нечіткої кластеризації

Можливісні алгоритми нечіткої кластеризації пов’язані з мінімізацією цільової функції [7, 8, 18]

$$E\left(U_q(k), w_q, \mu_q\right) = \sum_{k=1}^{N}\sum_{q=1}^{m} U_q^{\beta}(k) D^2\left(x(k), w_q\right) + \\ + \sum_{q=1}^{m} \mu_q \sum_{k=1}^{N}\left(1 - U_q(k)\right)^{\beta}, \tag{1.7}$$

де параметр μ_q визначає відстань між спостереженням та центроїдом w_q, на якій рівень належності $U_q(k)$ набуває значення 0,5.

Онлайн версія алгоритму Р. Крішнапурама-Дж. М. Келлера [18] має вигляд

$$\begin{cases} U_q(k+1) = \left(1 + \left(\dfrac{D^2\left(x(k+1), w_q(k)\right)}{\mu_q(k)}\right)^{\frac{1}{\beta-1}}\right)^{-1}, \\ w_q(k+1) = w_q(k) + \eta(k+1) U_q^{\beta}(k+1) * \\ *\left(x(k+1) - w_q(k)\right), \\ \mu_q(k+1) = \dfrac{\sum\limits_{p=1}^{k+1} U_q^{\beta}(p) D^2\left(x(p), w_q(k+1)\right)}{\sum\limits_{p=1}^{k+1} U_q^{\beta}(p)}, \end{cases} \tag{1.8}$$

або при $\beta = 2$

$$\begin{cases} U_q(k+1) = \left(1 + \dfrac{D^2\left(x(k+1), w_q(k)\right)}{\mu_q(k)}\right)^{-1}, \\ w_q(k+1) = w_q(k) + \eta(k+1)U_q^2(k+1) * \\ *\left(x(k+1) - w_q(k)\right), \\ \mu_q(k+1) = \dfrac{\sum\limits_{p=1}^{k+1} U_q^2(p) D^2\left(x(p), w_q(k+1)\right)}{\sum\limits_{p=1}^{k+1} U_q^2(p)}. \end{cases} \quad (1.9)$$

І знов таки тут у першому співвідношенні (1.9) виникає функція Коші з параметром ширини $\mu_q(k)$, що визначається третім співвідношенням (1.9).

1.7 Постановка задачі

В якості альтернативи імовірнісним та можливісним методам нечіткої кластеризації пропонується розробити метод достовірної нечіткої кластеризації з рекурентною модифікацією, який базується на підході правдоподібності та мірі подібності для нечіткої кластеризації для того, щоб скоротити та пришвидшити процес кластеризації даних в онлайн режимі коли дані надходять на обробку в онлайн режимі.

Об’єктом роботи є тестові набори даних Іриси Фішера та Вина із Uci репозиторію для методів кластеризації та класифікації даних.

Метою роботи є розробка та моделювання методу нечіткої кластеризації даних, що базуються на теорії достовірності та дозволяють кластеризувати данні в онлайн режимі за умов перетинних кластерів.

Завданнями роботи, відповідно до мети, є:

– аналіз та опис предметної області;

– постановка завдання;

– розробка та моделювання методу нечіткої кластеризації, що є достатньо простим у чисельній реалізації та призначений для вирішення задач кластеризації в рамках Data Stream Mining та Big Data Mining;

– підбір програмних засобів для реалізації та моделювання методу кластеризації даних;

– програмна реалізація методу;

– порівняльний аналіз роботи запропонованого методу з відомими.

2 РЕКУРЕНТНА ДОВІРЧА НЕЧІТКА КЛАСТЕРИЗАЦІЯ ДАНИХ НА ОСНОВІ МІРИ ПОДІБНОСТІ

Розглянуті алгоритми кластеризації формують класи, що мають форму гіперсфер, що не завжди відповідає реальним умовам, коли ці кластери можуть мати довільну форму. Більш адекватними та зручними є кластери гіпереліпсоїальної форми з довільною орієнтацією осей у просторі ознак.

В якості альтернативи імовірнісним і можливим процедурам в [7] були введені алгоритми достовірної нечіткої кластеризації, які використовують як свою основу теорію достовірності [9-11] і значною мірою позбавлені недоліків відомих методів.

Вихідною інформацією для вирішення задачі кластеризації є масив багатовимірних векторів спостережень $X = \{x(1), x(2), ..., x(k), ...x(N)\} \subset R^n$, де $x(k) \in R^n$ – вектор-спостереження, k – або номер цього спостереження у масиві даних X, або поточний дискретний час у задачах Data Stream Mining.

Якщо дані надходять на опрацювання послідовно у онлайн режимі, ці дані повинні бути розбиті на m перетинних класів, при цьому для кожного $x(k)$ повинен бути також розрахований рівень нечіткої належності до кожного з кластерів $U_q(k), q = 1, 2, ...m$. Передбачається також, що дані, що надходять на опрацювання, передоброблені так, що $-1 \le x_i(k) \le 1$, де $x_i(k), i = 1, 2, ..., n$ – i-та компонента вектора спостережень $x(k)$.

Переважна більшість відомих алгоритмів нечіткої кластеризації передбачає, що вихідний масив даних X містить N спостережень і не змінюється в процесі аналізу. В той же час існує досить широкий клас задач Data Stream Mining, де дані надходять на обробку послідовно і їх обсяг апріорі є невідомим та Big Data Mining коли цей обсяг є настільки великим, що просто не дозволяє опрацьовувати ці дані у пакетному режимі. У таких ситуаціях на перший план виходять рекурентні алгоритми нечіткої кластеризації, за

допомогою яких ці дані аналізуються послідовно вектор за вектором в міру їх надходження в систему.

Тому є доцільним проаналізувати відомі рекурентні методи нечіткої кластеризації та запропонувати нові, що відрізняються більш широкими функціональними можливостями у порівнянні з вже існуючими алгоритмами [20-26].

2.1 Теорія довіри

Концепція нечіткої множини була запропонована Заде у 1965 році за допомогою функції належності. Для того щоб виміряти нечіткі залежності Заде у 1978 році запропонував концепцію міри можливості. З тих пір цю теорію вивчали багато дослідників, таких як Нахміас у 1978 році, Кауфман і Гупта у 1985 році, Циммерманн у 1985 році, Дюбуа і Праде у 1988 році, Клір і Юань у 1995 році, Де Куман у 1997 році і Лю у 2002 році. Хоча можливісна міра більш розповсюджена, але в ній відсутні властивості двоїсності. Самодуальна міра є абсолютно необхідною як в теорії, так і на практиці [30-34].

2.1.1 Міра довіри та простір довіри

Нехай Θ – непуста множина і нехай $\mathrm{P}(\Theta)$ набір потужностей Θ (тобто всіх підмножин Θ). Кожен елемент $\mathrm{P}(\Theta)$ називається подією. Для того щоб представити аксіоматичне визначення довіри, необхідно присвоїти кожній події A число $Cr\{A\}$ яке вказує на достовірність того, що A відбувається. Для того, щоб переконатись, що число $Cr\{A\}$ має певні математичні властивості, які інтуїтивно очікується з певною достовірністю, застосовуючи п’ять аксіом:

– $Cr\{\Theta\}=1$;

– Cr збільшується, тобто, $Cr\{A\} \le Cr\{B\}$ коли завгодно $A \subset B$;

– Cr є двоїстою, тобто, $Cr\{A\} + Cr\{A^c\} = 1$ для будь якого $A \in \mathrm{P}(\Theta)$;

– $Cr\{U_i A_i\} \wedge 0,5 = \sup_i Cr\{A_i\}$ для будь-якого $\{A_i\}$ з $Cr\{A_i\} \le 0,5$;

– нехай Θ будуть непусті множини на яких Cr_k задовольняє перші чотири аксіоми, $k = 1,2,...,n,$ відповідно, і нехай $\Theta = \Theta_1 \times \Theta_2 \times ... \times \Theta_n$.

Міра довіри до Θ є позитивною тоді і лише тоді, коли існує не більше двох елементів, які приймають ненульові значення. Це означає, що міра достовірності ідентична мірі ймовірністі, якщо в універсальній сукупності є фактично два елементи. Допущення, що Θ це непуста множина, а $Cr\{\theta\}$ є невід'ємною функцією Θ, яка задовольняє умові розширення довіри

$$\sup_{\theta \in \Theta} Cr\{\theta\} \ge 0,5,$$

$$Cr\{\theta^*\} + \sup_{\theta = \theta^*} Cr\{\theta\} = 1 \text{ якщо } Cr\{\theta^*\} \ge 0,5.$$

Тоді $Cr\{\theta\}$ має унікальне розширення до міри достовірності $\mathrm{P}(\Theta)$ наступним чином

$$Cr\{A\} = \begin{cases} \sup\limits_{\theta \in A} Cr\{\theta\}, \text{ якщо } \sup\limits_{\theta \in A} Cr\{\theta\} < 0,5 \\ 1 - \sup\limits_{\theta \in A^c} Cr\{\theta\}, \text{ якщо } \sup\limits_{\theta \in A} Cr\{\theta\} \ge 0,5. \end{cases}$$

2.1.2 Нечітка змінна

Традиційно нечітка змінна визначається функцією належності. Визначимо її як функцію на просторі довіри так само, як визначається випадкова величина та функція міри на просторі ймовірностей [35, 36].

Нехай $f:\Re^n \to \Re$ і нехай $\xi_1,\xi_2,...,\xi_n$ нечіткі змінні на просторі довіри $\left(\Theta,\mathrm{P}\left(\Theta\right),Cr\right)$.

Тоді $\xi = f\left(\xi_1,\xi_2,...,\xi_n\right)$ нечітка змінна, що визначається як $\xi\left(\theta\right) = f\left(\xi_1\left(\theta\right),\xi_2\left(\theta\right),...,\xi_n\left(\theta\right)\right)$ для будь-якого $\theta \in \Theta$.

Якщо нечіткі змінні визначені на різних просторах довіри, тоді $\xi = f\left(\xi_1,\xi_2,...,\xi_n\right)$є нечіткою змінною, визначеною на просторі довіри $\left(\Theta,\mathrm{P}\left(\Theta\right),Cr\right)$ як $\xi\left(\theta_1,\theta_2,...,\theta_n\right) = \xi = f\left(\xi_1\left(\theta_1\right),\xi_2\left(\theta_2\right),...,\xi_n\left(\theta_n\right)\right)$ для будь-яких $\left(\theta_1,\theta_2,...,\theta_n\right) \in \Theta$.

2.1.3 Функція належності

Нехай ξ буде нечіткою змінною визначеною простором довіри $\left(\Theta,\mathrm{P}\left(\Theta\right),Cr\right)$. Тоді функцію належності можна записати наступним чином:

$$\mu\left(x\right) = \left(2Cr\left\{\xi = x\right\}\right) \wedge 1,\ x \in \Re.$$

На практиці нечітка змінна може бути задана функцією належності. У цьому випадку необхідно обчислити значення достовірності нечіткої події.

Нехай ξ нечітка змінна з функцією належності μ. Тоді для будь якої множини B дійсних чисел маємо

$$Cr\left\{\xi \in B\right\} = \frac{1}{2}\left(\sup_{x\in B}\mu\left(x\right) + 1 - \sup_{x\in B^c}\mu\left(x\right)\right).$$

2.1.4 Розподіл довіри

Розподіл достовірності $\Phi:\Re\to[0,1]$ нечіткої змінної ξ визначається як

$$\Phi(x)=Cr\{\theta\in\Theta|\xi(\theta)\le x\}.$$

Тобто $\Phi(x)$ це достовірність того, що нечітка змінна ξ приймає значення менше або рівне x. Якщо нечітка змінна ξ задана функцією належності μ то розподіл її довіри визначається

$$\Phi(x)=\frac{1}{2}\left(\sup_{y\le x}\mu(y)+1-\sup_{y>x}\mu(y)\right),\ \forall x\in\Re.$$

Функція $\Phi:\Re\to[0,1]$ є розподілом довіри і тоді і лише тоді, коли це зростаюча функція

$$\lim_{x\to-\infty}\Phi(x)\le 0,5\le\lim_{x\to\infty}\Phi(x),$$

$$\lim_{y\downarrow x}\Phi(y)=\Phi(x)\ \text{якщо}\ \lim_{y\downarrow x}\Phi(y)>0,5\ \text{або}\ \Phi(x)\ge 0,5.$$

Функція щільності довіри $\Phi:\Re\to[0,1]$ нечіткої змінної ξ яка була запропонована Лю у 2004 році є такою що

$$\int_{-\infty}^{+\infty}\phi(y)dy=1,\ \Phi(x)=\int_{-\infty}^{x}\phi(y)dy,\ \forall x\in\Re,$$

де Φ розподіл достовірності нечіткої змінної ξ.

На відмінну від випадкової події, $Cr\{a\le\xi\le b\}\ne\int_a^b\phi(y)dy.$

Пен, Мок та Це у 2005 році представили нову концепцію нечіткого домінування, засновану на розподілі довіри, та надали деякі основні властивості нечіткого домінування.

2.2 Довірчі методи нечіткої кластеризації даних

Довірчі алгоритми нечіткої кластеризації пов'язані з мінімізацією цільової функції [8-11, 21, 37]

$$E\left(Cr_q(k), w_q\right) = \sum_{k=1}^{N}\sum_{q=1}^{m} Cr_q^{\beta}(k) D^2\left(x(k), w_q\right) \tag{2.1}$$

за наявності обмежень

$$\begin{cases} 0 \le Cr_q(k) \le 1 \forall q, k, \\ \sup Cr_q(k) \ge 0,5 \forall k, \\ Cr_q(k) + \sup Cr_l(k) = 1 \end{cases} \tag{2.2}$$

для всіх q, k, для яких $Cr_q(k) \ge 0$,

де $Cr_q(k)$ – рівень довіри того, що спостереження $x(k)$ належить кластеру Cl_q.

В процедурах довірчої кластеризації рівень нечіткої належності визначається функцією належності [9-11, 38]

$$U_q(k) = \varphi_q\left(D\left(x(k), w_q\right)\right), \tag{2.3}$$

де $\varphi_q\left(D\left(x(k), w_q\right)\right)$ – монотонно зменшується на інтервалі $[0, \infty]$ та $\varphi_q(0) = 1, \varphi_q(\infty) \to 0$.

Нескладно помітити, що (2.3) є не що інше, як міра подібності заснована на відстані [10, 11]. В якості такої міри у [7] було запропоновано використовувати функцію

$$U_q(k) = \frac{1}{1 + D^2\left(x(k), w_q\right)}, \tag{2.4}$$

що X знов таки є функцією Коші з одиничним параметром ширини при цьому ніяк не враховується характер розподілу даних у вхідному масиві X.

Тому на наш погляд, більш прийнятним є вибір замість (2.4) співвідношень (1.5), (1.6), що прив'язані саме до характеру даних як в цілому у масиві X, так і у кластерах $Cl_q, q = 1,2...m$.

Цікаво помітити, що функція Коші постійно виникає в задачах нечіткої кластеризації, як вже відзначених ймовірнісної, можливісної, довірчої, так, робастної кластеризації стійкої до аномальних викидів у вихідних даних.

Таким чином, якщо пакетний алгоритм довірчої нечіткої кластеризації має вигляд

$$\begin{cases} U_q(k) = \dfrac{1}{1 + D^2\left(x(k), w_q\right)}, \\ U_q^*(k) = \dfrac{U_q(k)}{\sup U_l(k)}, \\ Cr_q(k) = \dfrac{1}{2}\left(U_q^*(k) + 1 - \sup U_l^*(k)\right), \\ w_q = \dfrac{\sum\limits_{k=1}^{N} Cr_q^{\beta}(k) x(k)}{\sum\limits_{k=1}^{N} Cr_q^{\beta}(k)}, \end{cases} \tag{2.5}$$

а його рекурентна версія може бути записана у формі

$$
\begin{cases}
\sigma_q^2(k+1) = \sum\limits_{\substack{l=1 \\ l\neq q}}^{m} \left(D^2\left(x(k+1), w_l(k)\right)^{\frac{1}{1-\beta}} \right)^{-1}, \\
U_q(k+1) = \dfrac{1}{1+\dfrac{\left(D^2\left(x(k+1), w_q(k)\right)\right)^{\beta-1}}{\sigma_q^2(k+1)}}, \\
U_q^*(k+1) = \dfrac{U_q(k+1)}{\sup U_l(k+1)}, \\
Cr_q(k+1) = \dfrac{1}{2}\left(U_q^*(k+1) + 1 - \sup U_l^*(k+1)\right), \\
w_q(k+1) = w_q(k) + \eta(k+1) Cr_q^{\beta}(k+1) * \\
*\left(x(k+1) - w_q(k)\right)
\end{cases}
\tag{2.6}
$$

або для найбільш розповсюдженого фаззіфікатора $\beta = 2$

$$
\begin{cases}
\sigma_q^2(k+1) = \left(\sum\limits_{\substack{l=1 \\ l\neq q}}^{m} \left\| x(k+1) - w_l(k) \right\|^{-2} \right)^{-1}, \\
U_q(k+1) = \left(1 + \dfrac{\left\| x(k+1) - w_q(k) \right\|^2}{\sigma_q^2(k+1)} \right)^{-1}, \\
U_q^*(k+1) = \dfrac{U_q(k+1)}{\sup U_l(k+1)}, \\
Cr_q(k+1) = \dfrac{1}{2}\left(U_q^*(k+1) + 1 - \sup U_l^*(k+1)\right), \\
w_q(k+1) = w_q(k) + \eta(k+1) Cr_q^{2}(k+1) * \\
*\left(x(k+1) - w_q(k)\right).
\end{cases}
\tag{2.7}
$$

З обчислювальної точки зору рекурентний алгоритм довірчої нечіткої кластеризації не є складнішим у порівнянні з онлайн версіями ймовірнісних, можливісних та робасних процедур.

Найбільш поширений підхід у рамках ймовірнісної нечіткої кластеризації пов'язаний із мінімізацією цільової функції [1, 37, 38].

2.3 Довірчі методи нечіткої кластеризації викривлених даних

У ситуаціях коли масив вихідних даних $X = \{x(1), x(2)..., x(k),..., x(N)\}$ містить пропуски (відсутні спостереження) розглянутий вище підхід не може бути використаний і потребує суттєвої модифікації. Так було в [12] було запропоновано модифікація FCM – процедури з урахуванням стратегії часткових відстане. В рамках цієї стратегії в розгляд вводиться три підмасиви даних:

$$X_F = \{x(k) \in X\},$$

де вектор $x(k)$ містить всі компоненти;

$$X_p = \{x_i(k), 1 \le i \le n, 1 \le k \le N\},$$

де $x_i(k)$ – значення векторів спостережень, що містяться в X;

$$X_G = \{x_i(k) = none, 1 \le i \le n, 1 \le k \le N\},$$

де $x_i(k)$ – компоненти векторів спостережень, що відсутні в X.

Використовуючи часткову відстань, можемо переписати функцію цілі замість (2.1) у наступному вигляді:

$$E\left(U_q(k), w_q\right) = \sum_{k=1}^{N}\sum_{q=1}^{m} U_q^{\beta}(k) \frac{n}{\delta_{\Sigma}(k)} \sum_{i=1}^{n}\left(x_i(k) - w_{qi}\right)^2 \delta_i(k), \tag{2.8}$$

де

$$\delta_i(k) = \begin{cases} 0 \ if \ \ x_i(k) \in X_G, \\ 1 \ if \ \ x_i(k) \in X_F, \end{cases}$$

$$\delta_{\Sigma}(k) = \sum_{i=1}^{n} \delta_i(k).$$

Використовуючи метод невизначених множників Лагранжа, отримуємо:

$$\begin{cases} U_q(k) = \dfrac{\left(D_p^2\left(x(k), w_q\right)\right)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}\left(D_p^2\left(x(k), w_l\right)\right)^{\frac{1}{1-\beta}}}, \\ w_{qi} = \dfrac{\sum\limits_{k=1}^{N}\left(U_q(k)\right)^{\beta} \delta_i(k) x_i(k)}{\sum\limits_{k=1}^{N}\left(U_q(k)\right)^{\beta} \delta_i(k)}. \end{cases} \tag{2.9}$$

У рекурентній онлайн формі (2.9) може бути записана у вигляді

$$\begin{cases} U_q(k+1) = \dfrac{\left(D_p^2\left(x(k+1), w_q(k)\right)\right)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}\left(D_p^2\left(x(k+1), w_l(k)\right)\right)^{\frac{1}{1-\beta}}}, \\ w_{qi}(k+1) = w_{qi}(k) + \eta(k+1) U_q^{\beta}(k+1)\left(x_i(k+1) - w_{qi}(k)\right)\delta_i(k). \end{cases} \tag{2.10}$$

Аналогічно з використанням стратегії часткових відстаней може бути записана пакетна процедура достовірної нечіткої кластеризації

$$\begin{cases} U_q(k) = \left(1 + D_p^2\left(x(k), w_q\right)\right)^{-1}, \\ U_q^*(k) = U_q(k)\left(\sup U_l(k)\right)^{-1}, \\ Cr_q(k) = \dfrac{1}{2}\left(U_q^*(k) + 1 - \sup\limits_{l \neq q} U_l^*(k)\right), \\ w_{qi} = \dfrac{\sum\limits_{k=1}^{N}\left(Cr_q(k)\right)^{\beta}\delta_i(k)x_i(k)}{\sum\limits_{k=1}^{N}\left(Cr_q(k)\right)^{\beta}\delta_i(k)}, \end{cases} \tag{2.11}$$

та її онлайн версія:

$$\begin{cases} \sigma_q^2(k+1) = \dfrac{1}{\sum\limits_{\substack{l=1 \\ l \neq q}}^{m} D_p^2\left(x(k+1), w_l(k)\right)}, \\ U_q(k+1) = \left(1 + \dfrac{D_p^2\left(x(k+1), w_q(k)\right)}{\sigma_q^2(k+1)}\right)^{-1}, \\ U_q^*(k+1) = \dfrac{U_q(k+1)}{\sup U_l(k+1)}, \\ Cr_q(k+1) = \dfrac{1}{2}\left(U_q^*(k+1) + 1 - \sup\limits_{l \neq q} U_l^*(k+1)\right), \\ w_{qi}(k+1) = w_{qi}(k) + \eta(k+1)Cr_q^{\beta}(k+1)\left(\tilde{x}_i(k+1) - w_{qi}(k)\right)\delta_i(k). \end{cases} \tag{2.12}$$

Можна помітити, що метод (2.12) є узагальненням процедури (2.7) у разі обробки даних, не спотворених пропусками.

3 ОПИС ПРОГРАМНОЇ РЕАЛІЗАЦІЇ

3.1 Системи комп’ютерної математики

Системи комп’ютерної математики (СКМ) – це програмні засоби нового покоління, призначені для виконання чисельних та аналітичних розрахунків будь-якого рівня складності, спрямованих на розв’язання різноманітних задач, які допускають коректне формулювання за допомогою термінів математики. При цьому, як правило, у системах комп’ютерної математики реалізовано високий ступінь візуалізації як проміжних, так і кінцевих розрахунків. Це потужні програмні середовища, які може ефективно застосовувати будь-який користувач для створення власних інформаційних продуктів високого рівня, не будучи при цьому професійним програмістом та математиком. Характерною рисою СКМ є їх гнучкість – користувачеві дається можливість активно втручатися в хід обчислень, при необхідності спрямовуючи розв’язання задачі в потрібне йому русло. Такого не можна сказати про велику кількість пакетів існуючих прикладних програм.

За класифікацією, запропонованою В.П. Д’яконовим, усі сучасні СКМ можна поділити на 7 основних класів:

– системи для чисельних розрахунків;

– табличні процесори;

– матричні системи;

– системи для статистичних розрахунків;

– системи для символьних розрахунків (системи комп’ютерної алгебри);

– системи для спеціальних розрахунків;

– універсальні системи.

На сьогоднішній день до універсальних СКМ можна віднести відомі програмні продукти, розроблені відомими західними фірмами, такими як MathSoft, MathWorks, Waterloo Maple, Wolfram. Найбільшу популярність

мають системи MathCAD, Maple, Mathematica, MATLAB. Саме вони все частіше використовуються для розв’язання навчальних, інженерних, науково-дослідних задач у різних галузях природничих наук.

Незважаючи на тещо кожна з СКМ відрізняється від інших своїми можливостями, загальну її структуру можна подати схемою.

Центральне місце належить ядру системи. Ядро – це коди сукупності еталонно відкомпільованих процедур та функцій, які забезпечують працездатність системи та проведення обчислень. Інтерфейс дає можливість користувачу спілкуватись із системою, звертаючись до ядра з конкретними задачами та одержувати візуалізований на екрані дисплея результат обчислень. Інтерфейс більшості сучасних СКМ базується на засобах найбільш популярних операційних систем Windows /95/98/ NT / XP /…. Об’єм ядра іноді обмежують з метою забезпечення швидкої його роботи. При необхідності до ядра можуть підключатися бібліотеки додаткових процедур та функцій або так звані пакети розширення. Саме такий підхід застосовано розробниками системи Maple: у звичайному режимі роботи використовується стандартна бібліотека команд, яка забезпечує проведення найпоширеніших математичних операцій. У разі потреби, однак, можна підключити інші бібліотеки, що реалізують, наприклад, обчислення з застосуванням спеціального математичного апарата.

Важливу роль у СКМ відіграє довідкова система. Як правило, вона має глибоко продуману розгалужену структуру й містить великий об’єм інформації щодо функціональних можливостей системи, прийомів роботи з нею та конкретними прикладами розв’язання типових задач.

Усі перелічені системи пройшли у своєму розвитку декілька етапів. Вони постійно вдосконалюються розробниками, і на сьогоднішній день існують декілька версій кожної з систем. Зокрема, версія системи MathCAD 1.0, випущена в кінці 80-х років, працювала під керуванням MSDOS, і для її успішного використання було достатньо процесора типу Intel 80286. Ядро системи разом з допоміжними файлами та колекцією прикладів цілком

«вміщувалось» на дискету ємністю 720 Кбайт, отже, з системою можна було працювати навіть на вітчизняному комп'ютері «Іскра 1030», який мав дуже обмежений (640 Кбайт) об'єм оперативної пам'яті. На певному етапі досконаліть систем комп'ютерної математики залежала від рівня розвитку апаратних ресурсів персональних комп'ютерів. Так, версія MathCAD 3.0 вже мала засоби створення тривимірної графіки, але ефективність роботи з системою при цьому знизилась через підвищення вимог до оперативної пам'яті. У міру розвитку ринку персональних комп'ютерів вдосконалювалась і система MathCAD. По-справжньому користувач зміг оцінити переваги цієї системи з появою процесорів типу Pentium та з виходом версії MathCAD 5.0, яка працювала підкеруванням системи Windows 3.11. Особливістю цієї версії бувпотужний та дуже зручний для користувача інтерфейс, загальна структура якогозбереглась і в останніх версіях пакета MathCAD. У наступних версіях (MathCAD 6.0–8.0, MathCAD 2000 – 2003) розширювались, насамперед, обчислювальні можливості ядра системи, додавались нові вбудовані процедури та функції, засоби програмування, засоби для проведення аналітичних (нечисельних) розрахунків, удосконалювався інтерфейс та довідкова система. Зараз найбільш розповсюдженими є версії Math CAD 2001, Math CAD 2001i та Math CAD 11, ядро яких можна вважати цілком оптимізованим для ефективного розв'язання будь-яких обчислювальних задач. Досить важливо, що розробники намагаються вдосконалювати систему таким чином, щоб зберегти «спадкоємність» версій: у більшості випадків більш нові версії «розуміють» файли, створені користувачем за допомогою більш ранніх версій.

Значна роль у розвитку систем комп'ютерної математики належить мережі INTERNET. На сьогоднішній день створено спеціальні сайти, на яких можна одержати як загальну інформацію щодо функціонування конкретної системи, так і щодо застосування її для розв'язання задач з тієї чи іншої галузі знань (www.mathsoft.com, www.mapleapps.com). Характерно, що всі сучасні СКМ дозволяють працювати в режимі «співробітництва» (Collaboration): при

наявності Internet-браузера користувач має змогу звернутися за допомогою в розв’язанні своєї задачі до інших користувачів. Як правило, таке звернення супроводжується кваліфікованою відповіддю з боку фахівців, що використовують системи комп’ютерної математики для власних потреб. Серед подібних сайтів співробітництва треба відзначити англомовний http://collab.mathsoft.com (сайт для прихильників системи MathCAD).

3.2 MATLAB

«Matrix Laboratory» – термін, що відноситься до пакету прикладних програм для вирішення завдань технічних обчислень, а також до використовуваної в цьому пакеті мови програмування. MATLAB працює на більшості сучасних операційних систем, включаючи Linux, Mac OS і Microsoft Windows. MATLAB, як мова програмування, була розроблена в кінці 1970-х років з метою полегшення процесів програмування для студентів (мова розділу Simulink отримала назву візуального проектування). Нова мова була з великим інтересом зустрінута вченими, що працюють в області прикладної математики.

Вдосконалений варіант MATLAB на мові C з’явився в 1984 р. Спочатку MATLAB призначався для проектування систем управління, але швидко завоював популярність в багатьох інших наукових і інженерних областях. Він також широко використовувався і в освіті, зокрема, для викладання лінійної алгебри і чисельних методів. Мова MATLAB є високорівневою інтерпретуємою мовою програмування, що включає засновані на матрицях структури даних, широкий спектр функцій, інтегроване середовище розробки, об'єктно-орієнтовані можливості і інтерфейси до програм, написаних на інших мовах програмування. Основною особливістю мови MATLAB є його широкі можливості по роботі з матрицями, які творці мови виразили в гаслі «думай векторно» (англ. Think vectorized). MATLAB надає користувачеві велику

кількість (декілька сотень) функцій для аналізу даних, що покривають практично всі області математики. MATLAB надає зручні засоби для розробки алгоритмів, включаючи високорівневі з використанням концепцій об’єктно-орієнтованого програмування. У ньому є всі необхідні засоби інтегрованого середовища розробки, включаючи налагоджувач і профайлер. Функції для роботи з цілими типами даних полегшують створення алгоритмів для мікро контролерів і інших застосувань, де це необхідно. У складі пакету MATLAB є велика кількість функцій для побудови графіків, зокрема тривимірних, візуального аналізу даних і створення анімованих роликів. Вбудоване середовище розробки дозволяє створювати графічні інтерфейси користувача з різними елементами управління, такими як кнопки, поля введення і іншими. За допомогою компоненту MATLAB Compiler ці графічні інтерфейси можуть бути перетворені в самостійні застосування, для запуску яких на інших комп'ютерах необхідна бути встановлена бібліотека MATLAB Component Runtime.

Пакет MATLAB містить функції, які дозволяють йому діставати доступ до інших додатків середовища Windows так само, як і цим застосуванням діставати доступ до даних MATLAB, за допомогою технології динамічного обміну даними (DDE). Інтерфейс для послідовного порту пакету MATLAB забезпечує прямий доступ до периферійних пристроїв, таким як модеми, принтери і наукове устаткування, що підключається до комп’ютера через послідовний порт (COM-порт). Інтерфейс працює шляхом створення об’єкту спеціального класу для послідовного порту. Наявні методи цього класу дозволяють читати і записувати дані в послідовний порт, використовувати події і обробники подій, а також записувати інформацію на диск комп’ютера в режимі реального часу. Це буває необхідно при проведенні експериментів, симуляції систем реального часу і для інших застосувань. Для MATLAB є можливість створювати спеціальні набори інструментів (англ. toolbox), що розширюють його функціональність. Наборами інструментів є колекції функцій, написаних на мові MATLAB для вирішення певного класу завдань.

Компанія Mathworks поставляє набори інструментів, які використовуються в багатьох областях.

3.3 Експериментальні дослідження

Щоб перевірити розроблені методи, а також зробити порівняльний з іншими більш відомими підходами, дослідження було проведено з використанням добре відомих тестових наборів даних репозиторію UCI, таких як Вина та Іриси Фішера. Опис цих наборів даних наведено в таблиці 3.1.

Репозиторій UCI (UCI Machine Learning Repository) – найбільший репозиторій реальних та модельних завдань машинного навчання. Містить реальні дані щодо прикладних завдань у галузі біології, медицини, фізики, техніки, соціології та інше.

Усі дані про набори даних зберігаються у вигляді файлів як ftp-сховище. Опис кожного набору даних також зберігається як файл, що уповільнює швидкість пошуку інформації.

Переваги репозиторію: добре відсортовані дані, повнотекстовий пошук.

Недоліки репозиторію: тільки текстовий формат даних, незручний у використанні та для перетворень. Відсутня система персоналізації, користувач може додати вибірку лише за жорстким шаблоном, що потребує додаткових витрат на адміністрування системи.

Таблиця 3.1 – Опис набору даних

Вибірка	**Кількість спостережень**	**Кількість атрибутів**	**Кількість кластерів**	**Ресурс**
Wine	178	13	3	Forina et al.(1988)
Iris	150	4	3	Fisher (1936)

Кожен із наборів даних має власний номер атрибута, номер даних, номер кластера та джерело даних.

3.3.1 Вибірка даних Іриси Фішера

Іриси Фішера (англ. Iris flower data set) – це багатовимірний набір даних для задачі класифікації, на прикладі якого англійський статистик та біолог Рональд Фішер в 1936 році продемонстрував роботу розробленого ним методу дискримінантного аналізу. Іноді його також називають ірисами Андерсона – через те, що дані були зібрані американським ботаніком Едгаром Андерсоном. Цей набір даних став класичним і часто використовується в літературі для ілюстрації роботи різних статистичних алгоритмів.

Іриси Фішера складаються з даних про 150 вимірювань ірисів з трьох видів – Iris setosa, Iris virginica і Iris versicolor, по 50 вимірювань на вид. Для кожного екземпляра вимірювалися чотири характеристики (в сантиметрах):

– довжина зовнішньої частки оцвітини (англ. sepal length);
– ширина зовнішньої частки оцвітини (англ. sepal width);
– довжина внутрішньої частки оцвітини (англ. petal length);
– ширина внутрішньої частки оцвітини (англ. petal width).

За даними вимірів будують правила класифікації, що дозволяє визначити вид рослини за даними вимірювань.

Один з класів (Iris setosa) набору даних є лінійно-відокремленим від двох інших. На рисунку 3.1 продемонстровані Іриси Фішера.

Набір даних ірисів Фішера можна загрузити з пакету UCI репозиторію.

Рисунок 3.2 демонструє частину навчальної вибірки Ірисів Фішера.

Рисунок 3.1 – Іриси Фішера

Sepal length	Sepal width	Petal length	Petal width	Species
5.2	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.3	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

Рисунок 3.2 – Частина вибірки ірисів Фішера

3.3.2 Вибірка даних Вина

Wine Quality Dataset – це набори даних, які доступні в наборах даних розпізнавання машинного навчання UC-Irvine. Цей набір даних є результатом хімічного аналізу різних вин, вирощених у Португалії.

Є два набори даних, пов’язані зі зразками червоного та білого вина vinho verde на півночі Португалії. Основною метою є створення моделі якості вина на основі даних фізико-хімічних досліджень. На рисунку 3.3 продемонстрована частина вибірки Вина.

Цікава річ щодо цього набору даних якості вина полягає в тому, що ми можемо практикувати регресійне та класифікаційне моделювання.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	winecolor
818	6.2	0.30	0.17	2.8	0.040	24.0	125.0	0.99390	3.01	0.46	9.0	5	white
3032	6.7	0.14	0.46	1.6	0.036	15.0	92.0	0.99264	3.37	0.49	10.9	5	white
2722	7.6	0.40	0.27	5.2	0.030	32.0	101.0	0.99172	3.22	0.62	12.3	7	white
5798	8.9	0.31	0.36	2.6	0.056	10.0	39.0	0.99562	3.40	0.69	11.8	5	red
757	6.8	0.22	0.37	15.2	0.051	68.0	178.0	0.99935	3.40	0.85	9.3	6	white
5662	9.1	0.68	0.11	2.8	0.093	11.0	44.0	0.99888	3.31	0.55	9.5	6	red
346	5.6	0.34	0.10	1.3	0.031	20.0	68.0	0.99060	3.36	0.51	11.2	7	white
3526	8.9	0.27	0.28	0.8	0.024	29.0	128.0	0.98984	3.01	0.35	12.4	6	white
4188	5.3	0.33	0.30	1.2	0.048	25.0	119.0	0.99045	3.32	0.62	11.3	6	white
2821	6.6	0.40	0.46	6.2	0.056	42.0	241.0	0.99680	3.50	0.60	9.9	5	white

Рисунок 3.3 – Частина вибірки Вина

Набір даних, містить 178 спостережень та 13 атрибутів (змінних). Тип вина в одному з трьох класів: 1 (59 сортів), 2 (71 сорт) і 3 (48 сортів): алкоголь, яблучна кислота, лужність золи, магній, феноли (загальні феноли), флавоноїди, нефлавоноїди (нефлавоноїдні феноли), проантоціани, колір (інтенсивність кольору), відтінок, розведення (D280/OD315 розбавлених вин), пролін.

3.4 Програмна реалізація

Для моделювання методу достовірної нечіткої кластеризації даних, використовувались обидві вибірки для подальшого порівняльного аналізу якості кластеризації даних, в залежності від виду вибірок.

На рисунку 3.4 продемонстрована програмна реалізація достовірної кластеризації даних: завантаження вибірки даних Іриси Фішера в Матлаб, видалення рядка-належності до кластера, для того, щоб відкластеризувати дані.

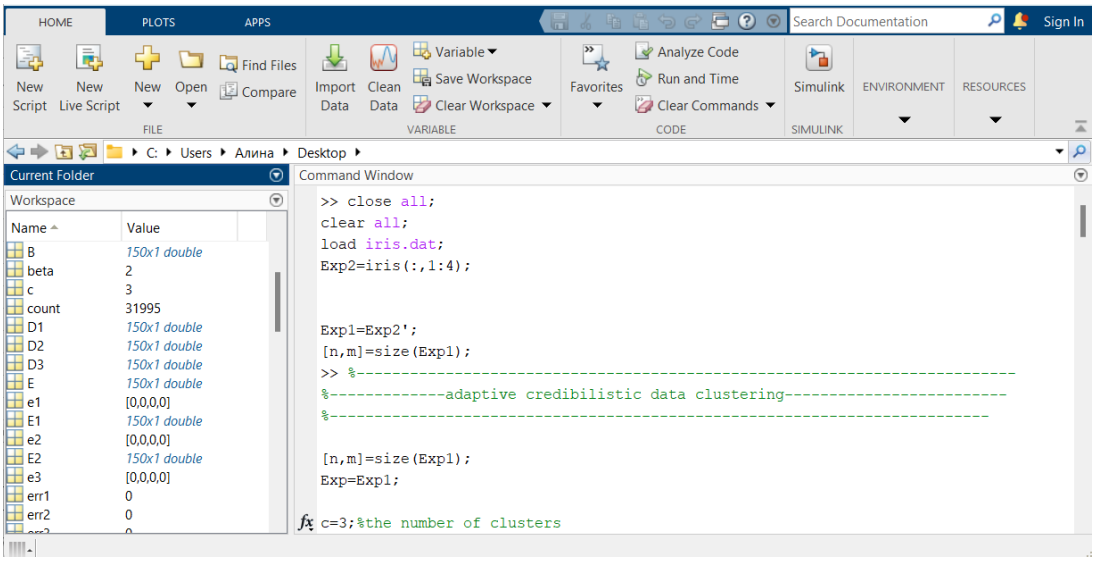
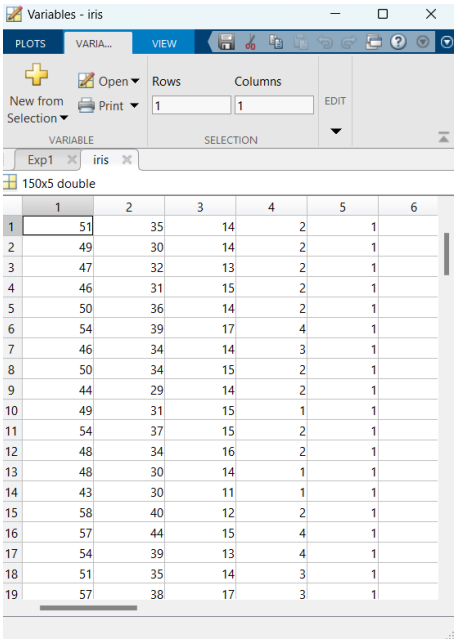


Рисунок 3.4 – Програмна реалізація методу достовірної нечіткої кластеризації даних

Наступним кроком є нормалізація вибірки даних в гіперкуб, для того, щоб в подальшому можна було перевірити правильність розрахунків за допомогою рівнів належностей спеціального типу.

Вихідні дані, що завантажені в систему, продемонстровано на рисунку 3.5, а нормовані в гіперкуб – рисунок 3.6.

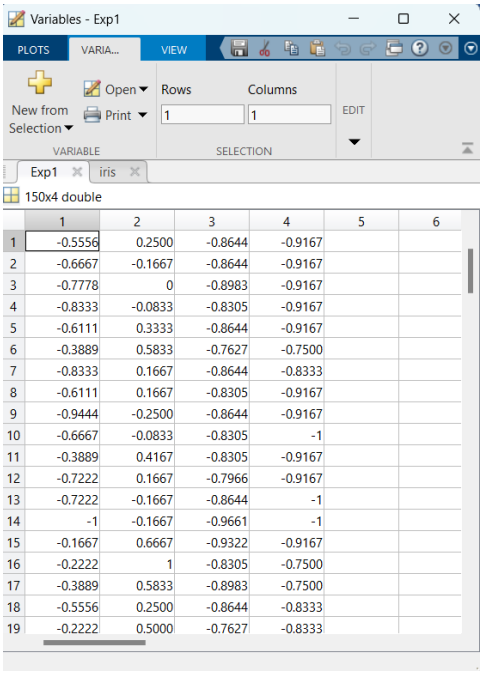


	1	2	3	4	5	6
1	51	35	14	2	1	
2	49	30	14	2	1	
3	47	32	13	2	1	
4	46	31	15	2	1	
5	50	36	14	2	1	
6	54	39	17	4	1	
7	46	34	14	3	1	
8	50	34	15	2	1	
9	44	29	14	2	1	
10	49	31	15	1	1	
11	54	37	15	2	1	
12	48	34	16	2	1	
13	48	30	14	1	1	
14	43	30	11	1	1	
15	58	40	12	2	1	
16	57	44	15	4	1	
17	54	39	13	4	1	
18	51	35	14	3	1	
19	57	38	17	3	1	

Рисунок 3.5 – Вихідні дані Іриси Фішера, що завантажені в систему

Як писалося раніше, перевірка правильності розрахунків, проходила за допомогою рівнів належності спеціального типу.

На рисунку 3.7 продемонстровано перевірку правильності роботи розробленого методу.



	1	2	3	4	5	6
1	-0.5556	0.2500	-0.8644	-0.9167		
2	-0.6667	-0.1667	-0.8644	-0.9167		
3	-0.7778	0	-0.8983	-0.9167		
4	-0.8333	-0.0833	-0.8305	-0.9167		
5	-0.6111	0.3333	-0.8644	-0.9167		
6	-0.3889	0.5833	-0.7627	-0.7500		
7	-0.8333	0.1667	-0.8644	-0.8333		
8	-0.6111	0.1667	-0.8305	-0.9167		
9	-0.9444	-0.2500	-0.8644	-0.9167		
10	-0.6667	-0.0833	-0.8305	-1		
11	-0.3889	0.4167	-0.8305	-0.9167		
12	-0.7222	0.1667	-0.7966	-0.9167		
13	-0.7222	-0.1667	-0.8644	-1		
14	-1	-0.1667	-0.9661	-1		
15	-0.1667	0.6667	-0.9322	-0.9167		
16	-0.2222	1	-0.8305	-0.7500		
17	-0.3889	0.5833	-0.8983	-0.7500		
18	-0.5556	0.2500	-0.8644	-0.8333		
19	-0.2222	0.5000	-0.7627	-0.8333		

Рисунок 3.6 – Вихідні дані Іриси Фішера, нормовані в гіперкуб, з видаленням стовпця, що надає інформацію про приналежність до кластера

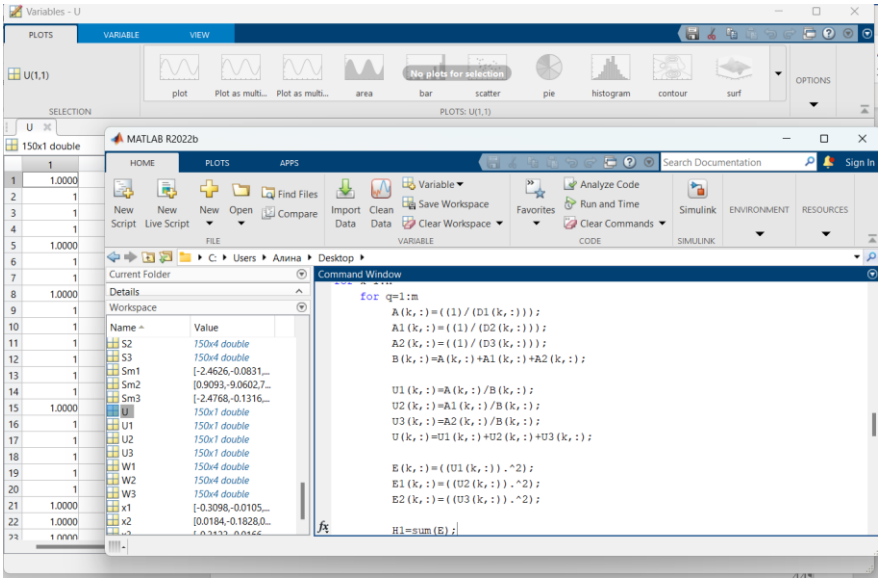


Рисунок 3.7 – Рівні належності, розраховані в достовірному методі кластеризації, що підтверджують правильність розрахунків

3.5 Методи оцінки якості кластеризації

Метод оцінки якості кластеризації – інструментарій для кількісної оцінки результатів кластеризації.

Прийнято виділити дві групи методів оцінки якості кластеризації:

– зовнішні (англ. External) методи, засновані на порівнянні результатів кластеризації з апріорно відомим розділенням на класи;

– внутрішні (англ. Internal) методи показують якість кластеризації лише за інформацією в даних.

Для оцінки якості кластеризації даних використовували індекс Силуету (SI), індекс Калінскі-Харабаса (CHI) та індекс Девіса-Болдуіна (DBI).

Індекс силуету – даний коефіцієнт не передбачає знання істинних міток об’єктів, і дозволяє оцінити якість кластеризації, використовуючи тільки саму (нерозмічену) вибірку і результат кластеризації.

Спочатку силует визначається окремо для кожного об’єкта. Позначимо через *n*-середня відстань від даного об’єкта до об’єктів з того ж кластера, через *m*-середня відстань від даного об’єкта до об’єктів з найближчого кластера (відмінного від того, в якому лежить сам об’єкт). Тоді силуетом даного об’єкта називається величина:

$$SI = \frac{m - n}{\max(m, n)}.$$

Силуетом вибірки називається середня величина силуету об’єктів даної вибірки. Таким чином, силует показує, наскільки середня відстань до об’єктів свого кластера відрізняється від середньої відстані до об’єктів інших кластерів. Дана величина лежить в діапазоні [- 1, 1] . Значення, близькі до -1, відповідають поганим (розрізненим) кластеризації, значення, близькі до нуля, кажуть про те, що кластери перетинаються і накладаються один на одного, значення, близькі до 1, відповідають «щільним» чітко виділеним кластерам.

Таким чином, чим більше силует, тим чіткіше виділені кластери, і вони є компактними, щільно згруповані хмари точок. За допомогою силуету можна вибирати оптимальне число кластерів *k* (якщо воно заздалегідь невідомо) – вибирається число кластерів, максимізуючи значення силуету. На відміну від попередніх метрик, силует залежить від форми кластерів, і досягає великих значень на більш опуклих кластерах, одержуваних за допомогою алгоритмів, заснованих на відновленні щільності розподілу.

Індекс Калінські-Харабаса. Компактність заснована на відстані від точок кластера до їх центроїдів, а розділеність – на відстані від центроїд кластерів до глобального центроїда:

$$CH(C) = \frac{N-K}{K-1} * \frac{\sum_{c_k \in C} |c_k| * \|\overline{c}_k - \overline{X}\|}{\sum_{c_k \in C} \sum_{x_i \in C} |c_k| * \|x_i - \overline{c}_k\|}.$$

Індекс Девіса-Болдуіна. Це, можливо, одна з найбільш використовуваних мір оцінок якості кластеризації. Вона вичисляє компактність як відстань від об’єктів кластера до їх центроїдів, а окремість – як відстань між центроїдами.

$$DB(C) = \frac{1}{K} \sum_{c_k \in C} \max_{c_l \in C \setminus c_k} \left\{ \frac{S(c_k) + S(c_l)}{\|\overline{c}_k - \overline{c}_l\|} \right\}.$$

3.6 Аналіз отриманих результатів в експериментальних дослідженнях

Результати кластеризації наборів даних показані в таблицях 3.2 та 3.3. Як показано в таблицях 3.2 та 3.3, метод нечіткої кластеризації на основі достовірного підходу демонструє хороші результати.

Таким чином, силуетний індекс показує, наскільки середня відстань до об’єктів кластера відрізняється від середньої відстані до об’єктів інших

кластерів. Це значення знаходиться в діапазоні [-1, 1]. Значення, близькі до -1, відповідають «поганим» (розрізненим) типам кластеризації. Значення, близькі до нуля, вказують на те, що кластери перетинаються і перекриваються. Значення, близькі до 1, відповідають «щільним» чітко виділеним кластерам. Таким чином, чим більший силует, тим чіткіші скупчення, і вони являють собою компактні, щільно згруповані хмари точок. Як видно з індексу силуету, метод відновлення даних працює досить добре.

Чим вище значення індексу Калінскі-Харабаса, тим краще рішення.

В індексі Девіса-Болдуіна значення, близькі до нуля, вказують на найкращу секцію, тобто, як бачимо, з майже всіма відсутніми даними розподіл є «хорошим», тому метод працював добре.

Таблиця 3.2 – Якість кластеризації

Методи кластеризації	**SI**	**CHI**	**DBI**
Адаптивна достовірна нечітка кластеризація даних	0,2325	922,01	1,25
Адаптивна достовірна нечітка кластеризація даних з пропусками	0,3335	965,42	1,05
FCM	0,2354	986,39	1,23
K-means	0,3676	1419,28	1,09

Алгоритм достовірної нечіткої кластеризації працює не тільки з повними даними, але й з даними, які містять відсутні значення. Для проведення експериментальних досліджень ми штучно ввели 10 відсутніх значень у набір даних Іриси Фішера. Рисунок 3.8 демонструє кластеризацію на основі методу достовірної нечіткої кластеризації даних (CFC) Іриси Фішера із 10 відсутніми значеннями.

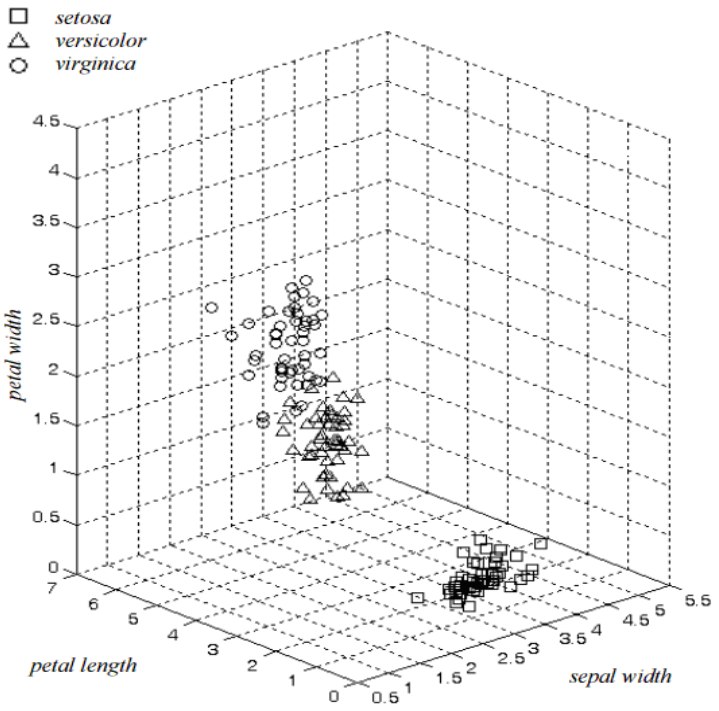


Рисунок 3.8 – Достовірна нечітка кластеризація даних Ірисів Фішера із 10 відсутніми значеннями

В таблиці 3.3 наведений порівняльний аналіз кластеризації різних даних із архіву Uci репозиторію.

Таблиця 3.3 – Порівняння точності кластеризації запропонованого методу із FCM на основі 100 спостережень для різних наборів даних

Вибірка	**Метод кластеризації**	**Загальна точність**		
		Найвища	**Середня**	**Дисперсія**
Вина	FCM	68,54	68,54	0
	CFC	67,98	67,98	0
Іриси Фішера	FCM	89,33	89,33	0
	CFC	91,33	90,06	0,04

Порівняльний аналіз даних було виконано з даними, запропонованими раніше методами кластеризації, які містять відсутні значення, такими як класичні алгоритми FCM і *K*-means.

ВИСНОВКИ

Проведені експерименти підтвердили дієвість запропонованих методів достовірної нечіткої кластеризації даних і дозволяють рекомендувати його для використання на практиці для вирішення задач автоматичної кластеризації даних різної природи. Запропонований метод призначений для використання в гібридних системах обчислювального інтелекту і, насамперед, у проблемах навчання штучних нейронних мереж, нейро-нечітких систем, а також у проблемах кластеризації та класифікації.

Отримані результати мають важливе практичне значення для створення систем відновлення та кластеризації даних, які надходять на обробку послідовно, в реальному часі. У ході дипломних досліджень отримано такі результати:

– запропонований метод достовірної нечіткої кластеризації даних, який дозволяє вирішувати задачу кластеризації у таблиці «об’єкт-властивість», що містять апріорі невідома кількість даних, а також забезпечує високу швидкодію і простоту чисельної реалізації;

– проведено імітаційне моделювання метода, виконана експериментальна оцінка похибок та якості кластеризації даних в онлайн режимі.

Результати роботи апробовано у вигляді тез доповідей під час 27 Міжнародного молодіжного форуму «РАДІОЕЛЕКТРОНІКА І МОЛОДЬ У XX СТОЛІТТІ» [39].

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Aggarwal, C. C. (2015). *Data mining: the textbook* (Vol. 1). New York: springer.

2. Bezdek, J. C. (2013). *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media.

3. Kohonen, T. (1991). Self-organizing maps: Ophmization approaches. In *Artificial neural networks* (pp. 981-990). North-Holland.

4. Krishnapuram, R., & Keller, J. M. (1993). A possibilistic approach to clustering. *IEEE transactions on fuzzy systems*, *1*(2), 98-110.

5. Xu, R., Xu, J., & Wunsch, D. C. (2012). A comparison study of validity indices on swarm-intelligence-based clustering. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *42*(4), 1243-1256.

6. Bodyanskiy, Y., Shafronenko, A., & Volkova, V. (2012). Adaptive clustering of incomplete data using neuro-fuzzy Kohonen network. *Artificial Intelligence Methods and Techniques for Business and Engineering Applications"–Rzeszow-Sofia: ITHEA*, 287-296.

7. Shafronenko, A., Dolotov, A., Bodyanskiy, Y., & Setlak, G. (2018, August). Fuzzy clustering of distorted observations based on optimal expansion using partial distances. In *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)* (pp. 327-330). IEEE.

8. Shafronenko, A., Bodyanskiy, Y., Pliss, I., & Irina, K. (2021, September). Online Credibilistic Fuzzy Clustering Method Based on Cauchy Density Distribution Function. In *2021 11th International Conference on Advanced Computer Information Technologies (ACIT)* (pp. 704-707). IEEE.

9. Shafronenko, A., Bodyanskiy, Y. V., Klymova, I., & Holovin, O. (2020, May). Online credibilistic fuzzy clustering of data using membership functions of special type. In *CMIS* (pp. 744-753).

10. Zhou, J., Wang, Q., Hung, C. C., & Yi, X. (2015). Credibilistic clustering: the model and algorithms. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *23*(04), 545-564.

11. Zhou, J., Wang, Q., Hung, C. C., & Yang, F. (2017). Credibilistic clustering algorithms via alternating cluster estimation. *Journal of Intelligent Manufacturing*, *28*, 727-738.

12. Bodyanskiy, Y., Shafronenko, A., & Mashtalir, S. (2020). Online robust fuzzy clustering of data with omissions using similarity measure of special type. In *Lecture Notes in Computational Intelligence and Decision Making: Proceedings of the XV International Scientific Conference "Intellectual Systems of Decision Making and Problems of Computational Intelligence"(ISDMCI'2019), Ukraine, May 21–25, 2019 15* (pp. 637-646). Springer International Publishing.

13. Бодянський, Є. В., Шафроненко, А. Ю., & Климова, І. М. (2021). Метод адаптивної достовірної нечіткої кластеризації даних на основі еволюційного алгоритму. *Збірник наукових праць Харківського національного університету Повітряних Сил*, (2 (68)), 80-83.

14. Bodyanskiy, Y., Shafronenko, A., & Volkova, V. (2012). Adaptive clustering of incomplete data using neuro-fuzzy Kohonen network. *Artificial Intelligence Methods and Techniques for Business and Engineering Applications"–Rzeszow-Sofia: ITHEA*, 287-296.

15. HOPPNER, F. K., KRUSE, R., & RUNKER, R. (1999). T.(1999) Fuzzy Cluster Analysis.

16. Gan, G., Ma, C., & Wu, J. (2020). *Data clustering: theory, algorithms, and applications*. Society for Industrial and Applied Mathematics.

17. Pereira, O. J., de Almeida Pacheco, L., Barreto, S. S., Emanuel, W., de Oliveira Fontes, C. H., & Cavalcante, C. A. M. T. (2012). Pattern Recognition using Multivariate Time Series for Fault Detection in a Thermoeletric Unit. In *Computer Aided Chemical Engineering* (Vol. 31, pp. 315-319). Elsevier.

18. Krishnapuram, R., & Keller, J. M. (1993). A possibilistic approach to clustering. *IEEE transactions on fuzzy systems*, *1*(2), 98-110.

19. Rastrigin, L. A. (1967). *Random Search in Optimization Problems for Multiparameter Systems*. FOREIGN TECHNOLOGY DIV WRIGHT-PATTERSON AFB OHIO.

20. Bodyanskiy, Y. V., Pliss, I. P., & Shafronenko, A. Y. (2022). КЛАСТЕРИЗАЦІЯ МАСИВІВ ДАНИХ НА ОСНОВІ КОМБІНОВАНОЇ ОПТИМІЗАЦІЇ ФУНКЦІЙ ЩІЛЬНОСТІ РОЗПОДІЛУ ТА ЕВОЛЮЦІЙНОГО МЕТОДУ КОТЯЧИХ ЗГРАЙ. *Radio Electronics, Computer Science, Control*, (4), 61-61.

21. Bodyanskiy, Y. V., Pliss, I. P., Shafronenko, A. Y., & Kalynychenko, O. V. (2022). НЕЧІТКА ДОВІРЧА КЛАСТЕРИЗАЦІЯ ДАНИХ НА ОСНОВІ АНАЛІЗУ ЩІЛЬНОСТІ РОЗПОДІЛУ ДАНИХ ТА ЇХ ПІКІВ. Radio Electronics, Computer Science, Control, (3), 58-68.

22. Maksymova, S., Matarneh, R., Lyashenko, V., & Belova, N. (2017). Voice control for an industrial robot as a combination of various robotic assembly process models.

23. Bodyanskiy, Y. V., Pliss, I. P., & Shafronenko, A. Y. (2022). ШВИДКА НЕЧІТКА ПРАВДОПОДІБНА КЛАСТЕРИЗАЦІЯ НА ОСНОВІ АНАЛІЗУ ПІКІВ ЩІЛЬНОСТІ РОЗПОДІЛУ ДАНИХ. *Radio Electronics, Computer Science, Control*, (1), 76-76.

24. Bodyanskiy, Y. V., Pliss, I. P., Shafronenko, A. Y., & Kalynychenko, O. V. (2022). НЕЧІТКА ДОВІРЧА КЛАСТЕРИЗАЦІЯ ДАНИХ НА ОСНОВІ АНАЛІЗУ ЩІЛЬНОСТІ РОЗПОДІЛУ ДАНИХ ТА ЇХ ПІКІВ. *Radio Electronics, Computer Science, Control*, (3), 58-58.

25. Bodyanskiy, Y., Shafronenko, A., & Pliss, I. (2022). Clusterization of vector and matrix data arrays using the combined evolutionary method of fish schools. *System research and information technologies*, (4), 79-87.

26. Bodyanskiy, Y., Shafronenko, A., Klymova, I., & Polyvoda, V. (2022). Robust Recurrent Credibilistic Modification of the Gustafson-Kessel Algorithm. In *Lecture Notes in Computational Intelligence and Decision Making: 2021 International Scientific Conference" Intellectual Systems of Decision-making and*

Problems of Computational Intelligence", Proceedings (pp. 613-623). Springer International Publishing.

27. Bodyanskiy, Y. V., Shafronenko, A., & Klymova, I. (2021, April). Adaptive Recovery of Distorted Data Based on Credibilistic Fuzzy Clustering Approach. In *COLINS* (pp. 6-15).

28. Shafronenko, A., Bodyanskiy, Y., Pliss, I., & Irina, K. (2021, September). Online Credibilistic Fuzzy Clustering Method Based on Cauchy Density Distribution Function. In *2021 11th International Conference on Advanced Computer Information Technologies (ACIT)* (pp. 704-707). IEEE.

29. Dubnitskiy, V., Kobylin, A., Kobylin, O., Kushneruk, Y., & Sheviakov, I. (2022). ОБЧИСЛЕННЯ ЗНАЧЕНЬ ФУНКЦІЙ КОМПЛЕКСНОЇ ЗМІННОЇ З ІНТЕРВАЛЬНИМ АРГУМЕНТОМ, ВИЗНАЧЕНИМ В ГІПЕРБОЛІЧНІЙ ФОРМІ. Advanced Information Systems, 6(3), 83-91.

30. Matarneh, R., Sotnik, S., Belova, N., & Lyashenko, V. (2018). Automated modeling of shaft leading elements in the rear axle gear. SPC.

31. Shafronenko, A. Y., & Rudenko, D. A. (2020). ONLINE RECURRENT METHOD OF CREDIBILISTIC FUZZY CLUSTERING. *ВВК 91*, 37.

32. Bodyanskiy, Y. V., Shafronenko, A. Y., Rudenko, D. A., & Klymova, I. N. (2020). Online Recurrent Method Of Credibilistic Fuzzy Clustering.

33. Bodyanskiy, Y., Shafronenko, A., & Pliss, I. (2021). Правдоподібна нечітка кластеризація даних на основі еволюційного методу божевільних котів. *System research and information technologies*, (3), 110-119.

34. Bodyanskiy, Y. V., Shafronenko, A. Y., & Klymova, I. N. (2021). ONLINE FUZZY CLUSTERING OF INCOMPLETE DATA USING CREDIBILISTIC APPROACH AND SIMILARITY MEASURE OF SPECIAL TYPE. *Radio Electronics, Computer Science, Control*, (1), 97-104.

35. Mohammad, A., Sotnik, S., Belova, N., & Lyashenko, V. (2018). Informational and Structural-Parametric Models of Inductions Micromotors.

36. Bodyanskiy, Y. V., Shafronenko, A. Y., & Klymova, I. N. (2021). ONLINE FUZZY CLUSTERING OF INCOMPLETE DATA USING

CREDIBILISTIC APPROACH AND SIMILARITY MEASURE OF SPECIAL TYPE. Radio Electronics, Computer Science, Control, (1), 97-104.

37. Bodyanskiy, Y., & Chala, O. (2023). Evolving Stacking Neuro-Fuzzy Probabilistic Networks and Their Combined Learning in Online Pattern Recognition Tasks. In Artificial Intelligence in Control and Decision-making Systems: Dedicated to Professor Janusz Kacprzyk (pp. 95-123). Cham: Springer Nature Switzerland.

38. Kobylin, O. A., Vyskrebentseva, S. O., & Petrova, R. V. (2019). Обробка даних, що містять пропуски в задачах кластеризації. Системи управління, навігації та зв'язку. Збірник наукових праць, 5(57), 45-50.

39. Фалько, М.К. (2023). ОГЛЯД МОЖЛИВИХ ПІДХОДІВ ДО ВИРІШЕННЯ ЗАДАЧ НЕЧІТКОЇ КЛАСТЕРИЗАЦІЇ. 27-й Міжнародний молодіжний форум «Радіоелектроніка і молодь у XXI столітті». Зб. матеріалів форуму. Т. 7, 8. Харків: ХНУРЕ. 2023, 86-87.