МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Харківський національний університет радіоелектроніки
Факультет Центр післядипломної освіти
(повна назва)

Кафедра Програмної інженерії
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Дослідження моделей еволюції кластерів в задачах розпізнавання образів
(тема)

Виконав:
Студент 2 курсу, групи ІПЗздм-19-1
Швець К.В.
(прізвище, ініціали)

Спеціальність 121 Інженерія програмного забезпечення
(код і повна назва спеціальності)

Тип програми освітньо-наукова

Керівник проф. Шубін І.Ю.
(посада, прізвище)

Допускається до захисту
Зав. кафедри _____ З.В. Дудар
(підпис) (прізвище, ініціали)

2021

Харківський національний університет радіоелектроніки

Факультет Центр післядипломної освіти
(повна назва)

Кафедра Програмної інженерії
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 121 Інженерія програмного забезпечення
(код і повна назва спеціальності)

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

Освітня програма Інженерія програмного забезпечення
(повна назва)

ЗАТВЕРДЖУЮ:
Зав.кафедри _____
(підпис)
« ____ » _____ 2021 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студента Швець Катерини Валеріївни
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження моделей еволюції кластерів в задачах розпізнавання образів

затверджена наказом університету від 26.03.2021 № 34 Стз

2. Термін подання роботи до екзаменаційної комісії 16 05 2021р.
3. Вихідні дані до роботи *проаналізувати існуючі алгоритми, що використовуються для вимог підтримки прийняття рішень, мови розробки програмного забезпечення*
4. Перелік питань, що потрібно опрацювати в роботі *мета роботи, аналіз проблемної галузі і постановка задачі, опис запропонованих варіантів оптимізації, використовувані методи та алгоритми, опис розробленої програмної системи, опис застосованих програмних рішень, аналіз можливих застосувань*
5. Перелік графічного матеріалу із зазначенням креслеників, схем, слайдів, ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри)

Мета завдання, обґрунтування доцільності розробки, постановка задачі, базові моделі, методи й алгоритми, структурно-логічна схема взаємодії даних, інтерфейс програмної системи, результати дослідної експлуатації програмної

системи, висновки

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
спецчастина	проф. Шубін І.Ю.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1.	Аналіз предметної галузі	26 березня 2021 р.	виконано
2.	Огляд існуючих методів	31 березня 2021 р.	виконано
3.	Розробка алгоритмів, проектування та розробка ПЗ	15 квітня 2021 р.	виконано
4.	Підготовка пояснювальної записки	28 квітня 2021 р.	виконано
5.	Спецчастина	30 квітня 2021 р.	виконано
6.	Підготовка презентації та доповіді	05 травня 2021 р.	виконано
7.	Попередній захист	10 травня 2021 р.	виконано
8.	Нормоконтроль, рецензування	12 травня 2021 р.	виконано
9.	Занесення роботи в електронний	14 травня 2021 р.	виконано
10.	Допуск до захисту в зав. кафедри	16 травня 2021 р.	виконано

Дата видачі завдання ____ _____ 2021р.

Студент _____
(підпис)

Керівник роботи _____ проф. Шубін І.Ю.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ /ABSTRACT

Пояснювальна записка до кваліфікаційної роботи магістра: 95 с, 5 табл., 46 рис., 6 дод., 34 джерела

КЛАСТЕРНИЙ АНАЛІЗ, РОЗПІЗНАВАННЯ ОБРАЗІВ, ЕВОЛЮЦІЙНІ ПЕРЕТВОРЕННЯ КЛАСТЕРНИХ ПОСЛІДОВНОСТЕЙ, ШТУЧНІ НЕЙРОННІ МЕРЕЖІ.

Метою роботи є розробка алгоритму і програмного забезпечення, орієнтованих на рішення завдань розпізнавання образів з довільною розмірністю простору ознак, що класифікують, при наявності динамічних змін

Предмет досліджень – еволюційні перетворення кластерних послідовностей, фактори ознак, що класифікують, які не приналежні простору кластера.

Результат – математичне моделювання і апробація обчислювальних алгоритмів штучної нейронної мережі і кластерного аналізу, що дозволяють розпізнавати образи.

CLUSTER ANALYSIS, PATTERN RECOGNITION, EVOLUTIONARY TRANSFORMATIONS OF CLUSTER SEQUENCES, ARTIFICIAL NEURAL NETWORKS.

The aim of the work is to develop an algorithm and software focused on solving image recognition problems with an arbitrary dimension of the space of classifying features, in the presence of dynamic changes

The subject of research is the evolutionary transformations of cluster sequences, classifying factors that do not belong to the cluster space.

The result is mathematical modelling and approbation of computational algorithms of artificial neural network and cluster analysis that allow to recognize images.

Я, Швець Катерина Валеріївна, студентка гр. ІПЗздм-19-1, здобувачка вищої освіти на другому (магістерському) рівні кафедри «Програмна інженерія», заявляю: моя кваліфікаційна робота на тему «Дослідження методів аналізу безпеки семантичних баз даних», що буде представлена в екзаменаційну комісію для публічного захисту, виконана самостійно, в ній не містяться елементи плагіату і вона може бути опублікована в електронному архіві відкритого доступу ElAr KhNURE. Всі запозичення з друкованих та електронних джерел мають відповідні посилання.

Я ознайомлена з діючим положенням «Про протидію академічному плагіату в ХНУРЕ», згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування дисциплінарних заходів.

ЗМІСТ

ВСТУП

Розпізнавання образів є на сьогоднішній день однієї з найбільш пріоритетних і актуальних проблем, що коштують перед людським співтовариством. Даний феномен пов'язаний із завданнями, що виникають у процесі проектування й розробки надійних охоронних систем; з питаннями теоретичної й прикладної робототехніки; із проблемами ефективного управління складними автоматизованими комплексами пошуку інформації й обробки інформаційних потоків і т.д.

Принцип, що лежить в основі всіх відомих процедур розпізнавання, досить простий: множину ознак, що характеризують досліджуваний об'єкт, співставляється з набором ознак, що втримуються в заздалегідь сформованій базі.

За результатами порівняння виноситься судження про можливість віднесення об'єкта до якої-небудь із існуючих категорій або вказується категорія, до якої об'єкт найбільш близький по властивостях у рамках прийнятої метрики. Якщо ж більшість характеристик не має аналогів у базі порівняння, об'єкт позначається як не ідентифікований і стає ядром нової категорії, що раніше не існувала. Також якщо є підстави вважати, що результати спостережень сумнівні або не достовірні, об'єкт не підлягає класифікації.

Незважаючи на простоту ідеї розпізнавання, процес її реалізації може виявитися досить трудомістким. Насамперед при значному обсязі бази порівнянь і великій кількості кваліфікуючих ознак у досліджуваного об'єкта процедура розпізнавання, виконана у вигляді попарного порівняння кожної ознаки з усіма елементами бази, украй не продуктивна через значну трудомісткість, по суті зводиться до повного перебору всіх можливих пар.

Ще однією серйозною перешкодою на шляху рішення завдання розпізнавання образів є відсутність загальноприйнятих формалізованих правил розбивання на категорії об'єктів і динамічна нестабільність сформованих категорій. Структура категорій або класів, визначається, насамперед, цілями й

завданнями досліджень, які можуть бути різні. Так само виявляються різними, а часом непорівнянними досвід і інтуїція дослідників, які визначають склад класів і рівень їх підготовки в предметній області. Крім того зміна, із плином часу, наповнення бази порівнянь і поява нових ознак в об'єктів, що підлягають розпізнаванню буде безсумнівно змінювати конфігурацію категорій, можливо аж до їхнього повного переформатування.

Таким чином, завдання розпізнавання образів, являє собою складне динамічне завдання, ряд етапів рішення якої на сучасному рівні наукових представлень не може бути строго обґрунтований. Зусилля вчених, що займаються дослідженням цього завдання, в основному зосереджені на розробці алгоритмів, що дозволяють прискорити процес пошуку рішення за рахунок систематизації процедури порівнянь.

Дуже важливим показником, що характеризує еволюційні перетворення кластерних послідовностей, є фактор ознак, що класифікують, які не приналежні простору також потужність множини елементів, що входять до складу кластера. Ця величина з плином часу може і повинна змінюватися, оскільки кількість елементів, що належать кластеру, може як зростати так і спадати. Зростання буде спостерігатися або при захопленні об'єктів з інших кластерів, або внаслідок поповнення навчальної вибірки в цілому. Спад же можливий через перехід частини елементів в інші кластери, а також в разі природної або примусової ліквідації будь-яких об'єктів. Якщо потужність множини елементів членів кластерної послідовності зберігається на постійному рівні, або має тенденцію до зростання, це є свідченням позитивної еволюційної динаміки, а кластер має ясні перспективи подальшого існування. Зворотна тенденція говорить про деградацію кластера аж до можливості його зникнення, тому прогноз його еволюційних змін невизначеності і даремний.

Коло галузей людської діяльності, де зустрічаються завдання, що пов'язані із проблемою розпізнавання образів, надзвичайно великий і продовжує неухильно розширюватися. Системи медичної діагностики, утворювальні Smart-системи, різного роду системи охорони й сигналізації, системи пошуку й обробки

інформації [1] , – от далеко не повний список тих сфер, де затребуваність цих завдань не викликає сумнівів.

Явище інформаційної глобалізації, характерне для сучасного миру, визначається насамперед надзвичайно більшими обсягами інформації й значним числом ознак, що класифікують. Це актуалізувало розробки нових математичних моделей і високопродуктивних алгоритмів для опису й обробки інформаційних потоків.

У цей час існує ряд технік, використовуваних у завданнях розпізнавання образів, проте, підвищення ефективності розпізнавання із застосуванням більш досконалих алгоритмів класифікації на основі нових математичних моделей є важливим і актуальним.

Одним з найбільш актуальних напрямків в області аналізу й обробки даних є побудова математичних і програмних додатків для розпізнавання образів у потоках даних. У цей час існує таке поняття як Big Data.

Big Data – це дані, які не можуть бути швидко оброблені й аналізовані за допомогою простих програм, таких як Excel. Усе явище й процеси, що відбуваються у світі є джерелами більших даних, такі як зміна клімату, аеронавігація, транспортний трафік і комунікаційні трафіки. Обсяг цифрових даних у сучасному світі збільшився й постійно зростає [1].

Незважаючи на наявність сучасних технологій і програм, людей у цей час контролює тільки 1% усіх світових цифрових даних, а інші дані поза можливостями людського контролю. При цьому поряд з підвищенням рівня технічних засобів усе більшу роль відіграють методи обробки даних, що поліпшують сприйняття, аналіз, розпізнавання образів для прийняття рішень і керування поведінкою технічних систем.

Однією із ключових проблем, що виникають при обробці інформаційних потоків, є проблема класифікації, тобто віднесення кожного об'єкта, виявленого інформаційною системою, до відповідного до класу по наявності деяких характерних ознак, яке виникає при рішенні багатьох практичних завдань: програмне забезпечення для аналізу даних користувачів, щоб скористатися цими

даними в різних і корисних областях, наприклад, антинаркотична сфера, анти екстремізм, анти тероризм. Деякі організації аналізують дані користувачів для добутку нових продуктів, модифікацій і поліпшення їх характеристик або прогнозування стилів і поведінки користувачів. Наприклад, аналіз даних студентів допомагає впровадити поліпшення в сфері освіти, охорони здоров'я й ін.

Створення й моделювання адаптивних нейронних мереж розглядається як найбільш затребуваний напрямок у вирішенні багатьох проблем штучного інтелекту й у системах інтелектуального аналізу даних. Синонімами терміна «інтелектуальний аналіз даних» є видобуток даних (Data mining), виявлення знань (knowledge discovery) .

Інтелектуальний аналіз даних пов'язаний з пошуком схованих нетривіальних і корисних закономірностей, що дозволяють одержати нові знання про досліджувані дані. Особливий інтерес до методів аналізу даних виник у зв'язку з розвитком засобів збору й зберігання даних, що дозволяють накопичувати більші обсяги інформації.

Перед фахівцями з різних галузей науки встало питання про обробку цифрових даних, що збираються, перетворення їх у знання й використання для розвитку різних галузей. Відомі статистичні методи показують лише частину потреб по обробці даних при невеликих обсягах даних, і для їх використання необхідно мати чітке уявлення про шукані закономірності. У такій ситуації методи інтелектуального аналізу даних здобувають особливу актуальність. Їхня основна особливість полягає у встановленні наявності й характеру схованих закономірностей у даних.

Серед методів інтелектуального аналізу даних особливе місце займають класифікація й кластерізація. Класифікація, при відомій заздалегідь угрупованні даних на підмножини (класи), установлює закономірність, по якій дані групуються саме таким чином, для рішення таких завдань використовується штучні нейронні мережі. Тому в роботі були розроблені нові алгоритми для побудови й навчання нейронних мереж. Штучна нейронна мережа (ШНМ) являє

собою математичну модель паралельних обчислень, що містить взаємодіючі між собою штучні нейрони. Перевагою нейронних мереж перед традиційними алгоритмами є можливість їх навчання.

Навчання ШНМ може вестися із учителем або без учителя. У першому випадку мережі пред'являються значення як вхідних, так і бажаних вихідних сигналів, і вона по деякому внутрішньому алгоритму підбудовує ваги своїх синоптичних зв'язків.

У другому випадку виходи ШНМ формуються самостійно, а ваги змінюються по алгоритму, що враховує тільки вхідні й похідні від них сигнали. Існує множина алгоритмів навчання, які поділяються на два класи: детерміністські й стохастичні. У першому підстроюванні ваг являє собою послідовність дій, у другому – воно проводиться на основі дій, що підкоряються деякому випадковому процесу.

У цей час одним з актуальних питань інформаційних технологій в області штучного інтелекту є проблема розпізнавання й класифікації образів із застосуванням принципово нових методів і алгоритмів, на основі розроблених нових математичних моделях. Налагоджені нейронні мережі можна застосовувати для рішення всіляких завдань, від відновлення пропусків у даних до аналізу й пошуку закономірностей.

У роботі розглядається математичне моделювання й проводиться апробація обчислювальних алгоритмів штучної нейронної мережі й кластерного аналізу, що дозволяють розпізнавати образи.

1 АНАЛІЗ СТАНУ РОЗВ'ЯЗАННЯ ПРОБЛЕМИ ТА ОБҐРУНТУВАННЯ ЦІЛЕЙ ДОСЛІДЖЕННЯ

1.1 Аналіз теорії розпізнавання образів

Перші спроби автоматизації процесу розпізнавання образів ставляться до початку 1950-х років. У цей час комп'ютери вже стали широко використовуваним засобом для обробки інформації. Мета цих досліджень полягає в розробці систем розпізнавання друкованих знаків – букв і цифр [3]. Одна з перших моделей запам'ятовування й організації даних, реалізованих по принципу людського мозку, була запропонована Розенблатом в 1950 році на основі персептрона. До цього часу в 40-ві роки вже штучні нейронні мережі виділилися як окремий науковий предмет. Дослідники в цьому напрямку створили апаратні й програмні моделі біологічного нейрона і його зв'язків [4].

Метою цих моделей є відтворення функцій людського мозку[6]. Пізніше такими моделями стали називати персептрони. Перші спроби дослідників привели до грубих апроксимацій, перш ніж вони досягли більш глибокого розуміння нерівної системи людину. Вражаючі результати перших спроб на персептронах стимулювали подальші дослідження, що привело до створення більш досконалих мереж.

Штучні нейронні мережі були вперше вивчені систематично в 1943 року Маклаком і Питтсом [5]. На рис. 1.1 показана проста нейронна модель, яку вони використовували у своїй роботі в більшій частині. На вхід цієї схеми надходять двійковий сигнал x, який множиться на вагу w, і потім підсумовується. Значення на виході персептрона залежить від значення суми, якщо значення цієї суми більше значення заданого переділу, то вихід дорівнює одиниці або нулю. Такі системи і їм подібні називаються персептронами [6].

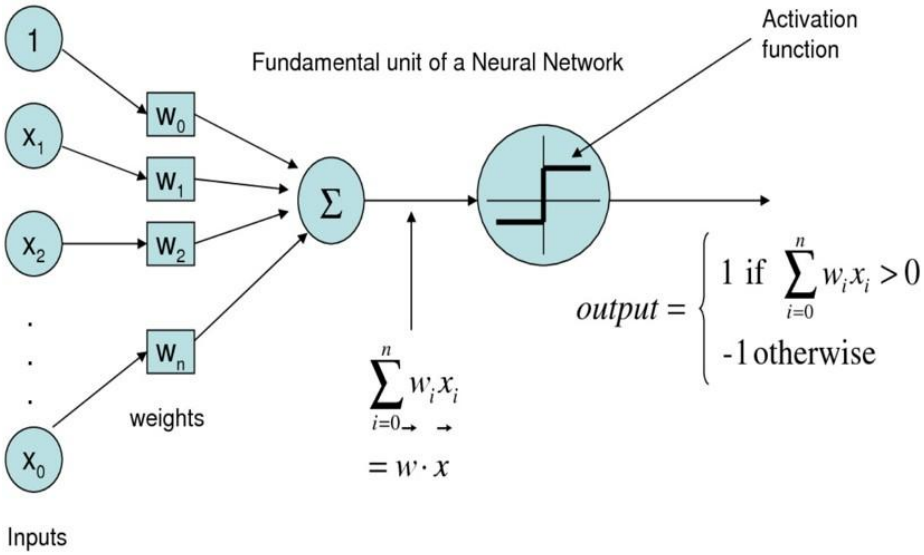

Рисунок 1.1 – Схема персептрона нейронної мережі

В 60-ті роки персептрони широко поширилися й одержали великий інтерес і оптимізмом. У цей час Розенблат довів теорему про навчання персептрона [7]. Також було запропоновано кілька переконливих представлень про системи персептронів, що викликав інтерес у дослідників в усьому світі до вивчення можливостей цих систем [8]. На початку досліджень виявилося, що персептрони не можуть навчатися рішенню ряд простих завдань, що приводило до розчарування. Ці проблеми були вирішені Мінським, коли він аналізував проблему й з'ясував обмеження одношарових персептронів, а також їх здатності у виконанні завдань і навчанні [9]. Однак, у цей час багатошарові персептрони і їх методи навчання не існували, а дослідники перейшли в нові області, у яких методи одношарових персептронов стали не ефективними. Поява нових методів навчання багатошарових персептронів викликало новий інтерес у дослідників.

Дослідження методів створення систем розпізнавання образів в 60-х роках зростало залежно від ступеня використання обчислювальних машин і потреби в збільшення швидкості й ефективності взаємодії між людиною й машиною [6].

До аналітичних методів був запропонований додатковий синтаксичний похід для використання результатів теорії машинних мов при рішенні конкретних типів завдань і розпізнаванні візуальних образів [10].

Інформаційний вибух в 700х роках став одним з головних проблем, що викликають інтерес до розвитку теорій і методів побудови систем автоматичного

розпізнавання образів [10].

Психології припустили, що до 2000 року можливість людського мозку засвоювати підвищений потік інформацій стане болісним.

Завдання розпізнавання образів мають більшу практичну значимість. Терміни «розпізнавання образів» і «класифікації образів» є частково взаємозамінними [11].

У деяких галузях науки дані терміни розглядаються як різні, кожний з яких має свої сфери застосування й ці терміни інтерпретуються залежно від специфіки завдань.

1.2 Аналіз змісту категорії «розпізнавання образів»

Розпізнавання образів – це один з видів машинного навчання, який фокусується на розпізнаванні образа й закономірностей в інформаційних даних [12].

Шаблон може бути визначений як об'єкт, якому може бути привласнена функція – образ відбитка пальця, візерунок рукописного введення, подоба людських обличчя, образ голосових сигналів або шаблон послідовності ДНК [12]. Розпізнавання образів – це певна техніка, яка класифікується як вид машинного навчання, яка наділяє машину можливістю розуміти навколишнє середовище й інші сфери нашому життя. Розпізнавання образів також використовується для прийняття більш точних і підходящих рішень для кожної категорії образа [11].

Існує три типи процедури навчання, використовувані для розпізнавання образів, тобто навчання із учителем, яке припускає надання навчальної вибірки. Другий тип навчання – це навчання без учителя, яке полягає в знаходженні відповідних образів в інформаційних даних без надання навчальної вибірки для виявлення образа. Третій тип навчання – напівконтрольоване навчання, яке

використовує комбінацію мічених і немічених даних для класифікації об'єктів.

Крім того, існують різні області застосування розпізнавання образів: аналіз зображення [13], класифікація при пошуку в Інтернеті, витяг мультимедійних баз даних, розпізнавання мови [14], обробка тексту природньою мовою [15], біометричне розпізнавання [16], медичне розпізнавання [1], автоматичного розпізнавання цілей у військовій сфері [17], використовуючи оптичне, інфрачервоне зображення й дистанційне зондування і т.д. Важливість і цінність реальних проблем, які можуть бути вирішені за допомогою розпізнавання образів, підтверджується визначенням різних методів розпізнавання й класифікації об'єктів, запропонованих у сучасній літературі (статистичне розпізнавання образів, штучна нейронна мережа й машина опорних векторів і т.д.).

Певна інформація (наприклад, результати вимірів) доступна для кожного елементу набору, і при цьому в неї є особливість, яку має тільки частина даного елемента. Якщо володіння цією особливістю елемента відсутнє, виникає припущення, що в доступній інформації виникає проблема по виявленню елементів, що володіють цією особливістю. Цю проблему можна розв'язати, побудувавши модель на основі механічних, фізичних, хімічних або інших наукових даних, які могли б пояснити взаємозв'язок між доступним джерелом інформації й розглянутою функцією. Але в багатьох випадках складність системи робить застосування такої моделі практично неможливої, тому доцільно застосовувати методи розпізнавання образів.

1.3 Області застосування методів розпізнавання образів

Виявлення зон, підданих землетрусам [18]. Проблема полягає в тому, щоб визначити в регіоні області, де можливі сильні землетруси (з магнітудою $M \geq M_0$, де M_0 – заданий поріг). Об'єктами є обрані геоморфологічні структури (перетинання лінєаментів, морфоструктурні вузли, і т.д.) регіону. Можливість

сильного землетрусу поблизу об'єкта – це розглянуте завдання. Доступна інформація – це топографічні, геологічні, геоморфологічні й геофізичні дані об'єктів. Проблема розпізнавання образів полягає в тому, щоб розділити обрані структури на два класи: структури, у яких можуть виникати землетруси з магнітудою $M \geq M'$; структури, у яких можуть виникнути тільки землетруси з магнітудою $M < M'$.

Перерозподіл шарів багатих вуглеводнями [19]. Розглядаються шари, що зустрічаються в шпарі. Проблема полягає в тому, щоб визначити, які шари містять нафту або воду. Об'єктами є шари, заповнені нафтовими шарами, визначення яких є розглянутим завданням. Доступна інформація – це геолого-геофізичні дані, обмірювані для шарів. Проблема розпізнавання образів полягає в тому, щоб розділити шари на два класи: шари, що містять нафту; шари, що не містять нафту.

Медична діагностика [20] – розглядається конкретне захворювання. Проблема полягає в тому, щоб діагностувати хвороба з використанням результатів медичних тестів. Об'єкти розглядаються людьми. Хвороба – це завдання на розгляді. Доступна інформація – це дані, отримані за допомогою медичних тестів. Проблема як розпізнавання образів полягає в тому, щоб розділити досліджених людей на два класи:

– люди, у яких є хвороба;

– люди, у яких немає хвороби.

Класифікація й кластерний аналіз: множина W поділяється на групи (кластери) (див. рис.1.2) на основі деякого підходу в m-мірному простір w_1, w_2, ... , w_n.

Позначено $p(\boldsymbol{w}\ \boldsymbol{v})$ відстань між двома m-мірними векторами $\boldsymbol{w} = (w_1, w_-, \ldots, w_0)$ і $\boldsymbol{v} = (v_1, v_2, \ldots, v_m)$.

Определіліти класифікацію й оцінити в той же час її особливу функцію. Найкраща класифікація дає екстремум цієї функції.

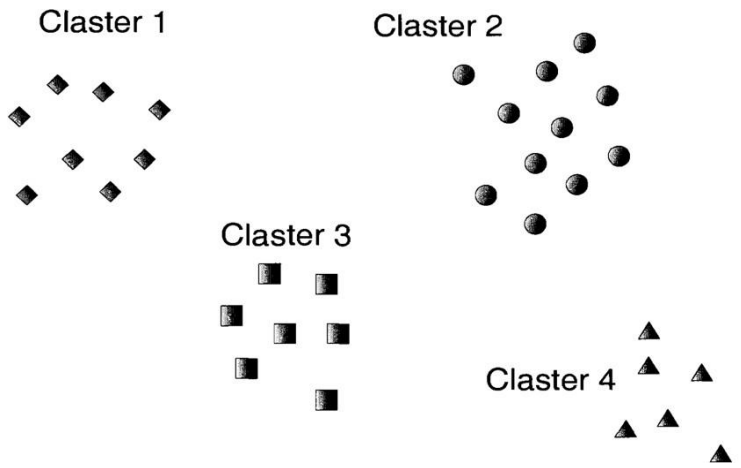

Рисунок 1.2 – Кластерізація об'єктів у двовимірному просторі

Приклади функцій [23]. Нехай W − кінцева множина. Можуть бути використані наступні функції:

$$J_1 = \frac{(k-1)\sum_{k=1}^{K}\rho_k}{2\sum_{k=1}^{K-1}\sum_{j=k+1}^{K}\rho_{kj}} \to min$$

$$J_2 = \frac{1}{K}\left(\sum_{k=1}^{K}\rho_k - \frac{2}{K-1}\sum_{k=1}^{K-1}\sum_{j=k+1}^{K}\rho_{kj}\right) \to min,$$

$$\rho_k = \frac{2}{m_k(m_k-1)}\sum_{i=1}^{m_k-1}\sum_{s=i+1}^{m_k}\rho_{(\mathbf{w}^i,\mathbf{w}^s)},$$

$$\rho_{kj} = \frac{1}{m_k m_j}\sum_{i=1}^{m_k}\sum_{s=1}^{m_j}\rho_{(\mathbf{w}^i,\mathbf{v}^s)},$$

де K – число груп;

m_k ,m_j числа об'єктів у групі з номером k і в групі з номером j відповідно;

$w^1, w^2, \ldots, w^{mk}$ є об'єктами групи з номером k ,

$v^1, v^2, \ldots, v^{mj}$ є об'єктами групи з номером j.

Після визначення груп можна сформулювати наступну проблему: знайти загальну особливість об'єктів, що належать до одній і тій же групі.

1.4 Класифікація процедури навчання

Якщо заздалегідь відомо про деяких об'єктах, до яких конкретно груп (класам) вони відносяться, то ця інформація може використовуватися для визначення класифікації для інших об'єктів.

У дослідженні [23] множина W розділена на два класи, наприклад, D і N

Дані апріорні приклади об'єктів кожного класу. Їх називають тренуванням множини W:

$$W' \subset W,$$

$$W' = D \cup N'.$$

де D − навчальна вибірка (апріорні приклади) об'єктів, що належать класу D,

N − навчальна вибірка (апріорні приклади), що належать класу N.

Навчаюча вибірка W використовується для визначення апріорно невідомого розподілу об'єктів множини W між класами D і N.

Результат розпізнавання образів: правило визнання, воно дозволяє розпізнавати до якого класу належить об'єкт знаючи вектор $\boldsymbol{w}$ об'єкт, що характеризує цей? фактичний поділ об'єктів на окремі класи відповідно до цього правила (див. рисунок 1.3): $W = D \cup N$ якщо є об'єкти з невизначеною класифікацією, тоді $W = (D \cup N) \cup U$

Аналіз отриманого правила визнання дасть інформацію для розуміння зв'язків між характеристикою, яка відрізняється класами D і N, з одного боку, і описом об'єктів (компонентів векторів **w**)с іншої.

В області розпізнавання образів використовуються багато підходів, одним з яких є застосування стандартних моделей класифікаторів, що навчаються із учителем.

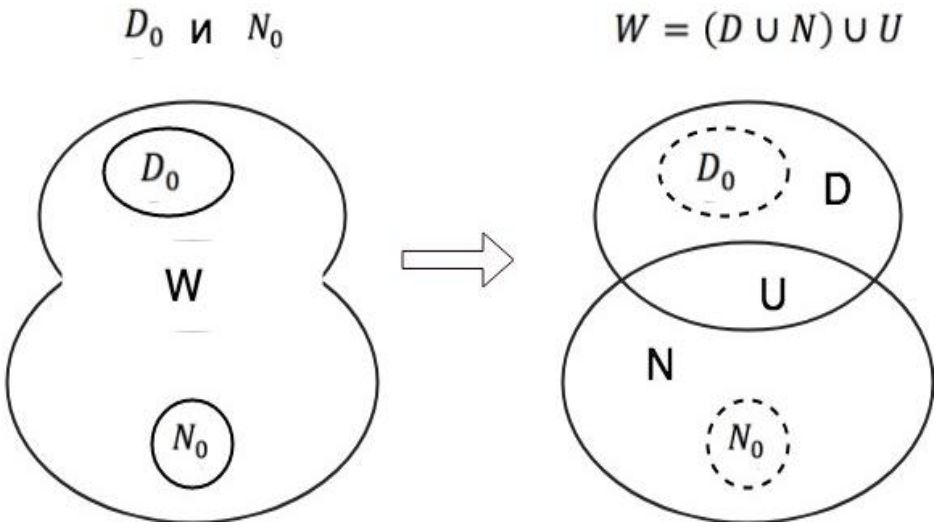

Рисунок 1.3 – Класифікація процесу з навчанням

У таких моделях для їхнього навчання використовуються відзначені вибірки даних, які складаються із двох масивів (масив об'єктів і відповідним масив оцінок), які визначають класи, до яких ставляться об'єкти.

Під час навчання масив даних розділяється на дві частини неоднакових по розміру. Потім за допомогою моделі певного правила конкретного алгоритму навчання, параметри моделей накладаються за допомогою навчальної вибірки таким чином, щоб у якості вхідних даних об'єктів модель привласнювала на виході оцінку класу, до якого даний об'єкт належить.

Даний підхід складається з множини моделей. У середовищі цих моделей найпоширенішими й широко використовуваними є модель зворотного поширення помилок, багатошарова штучна нейронна мережа, векторний метод, дерево рішень і колекції моделей, які є сукупністю перерахованих моделей [24].

У багатошарових персептронах навчання здійснюється за допомогою методу зворотного поширення помилок. Даний метод у широкому масштабі використовується для розпізнавання різних класів картин, таких як рукописні символи, почерк, особи людей, дані візуальних датчиків роботизованих систем [22].

Модель багатошарового персептрона складається з безлічі штучних нейронів, обчислювальної одиниці моделей, об'єднаних у вигляді шарів в

ієрархічній послідовності.

Штучний нейрон відбиває структуру біологічного нейрона (нервової клітини), що полягає з одного або декількох входів, одного з декількох виходів і функції активації [25]. Крім того входи нейрона мають асоціативний коефіцієнт (вага).

Нейрон поводяться в такий спосіб: нехай у нейрона є m+1 входів, вхідні значення яких x_0, $x_1,\ldots,x_m$, ці значення множаться на значення вектора ваг $w_0,w_1,\ldots w_m$, на перший вхідний елемент подаються фіксовані значення зсуву $x_0=1$.

У цьому випадку виходам нейрона привласнюються вихідне значення функції активації від визначальної суми вихідних значень нейрона:

$$y = f(\sum_{i=0}^{m} w_i x_i)$$

Функція активації може не мати властивості нелінійності, нормалізації вихідних даних і іншими властивостями. З функцій, що виступають у якості функції активації відомі функції сигмоїд.

У багатошарових штучних нейронах вхідні значення одного шару є вихідними значеннями попереднього шару.

В цьому випадку нейрони вхідного шару багатошарової мережі посередньо ухвалюють вихідні значення даних, що підлягають розпізнаваню, наприклад, у випадку розпізнавання картинок, дані являють собою значення інтенсивності їх складових пікселів.

Вихідний шар змінюється залежно від виду завдань. Але в стандартних архітектурах число вихідних нейронів відповідає рівній кількості класів розпізнавання, а вихідні значенні кожного нейрона одержують значення в інтервалі від 0 до 1, що представляють собою ймовірність приналежності вхідної картинки певному класу.

Дослідники відзначають, що багатошарові нейрони можуть виражати будь-яку математичну функцію за допомогою випадкового набору нейронів [32,33]. Оскільки існує складність у формуванні аналітичного правила класифікації образів за критерієм розпізнавання.

Здатність навчальної вибірки позначена тим, що являє собою нейронні мережі й подобні їм моделі активними для розпізнавання природніх образів навколишнього середовища, які відрізняються нечіткою структурою й безліччю варіацій у переділах певних класів.

Навчання нейронних мереж методом зворотного поширення помилки полягає в наступному: нехай є функція розпізнавання $g: X \rightarrow Y$, а аргументом: $x_n \in X$. Аргументом є вектор ознак образа, а значеннями функції є сукупність класів $y \in Y$. Навчальна вибірка є підмножиною значень даної функції $D = \{(x_0, y_0), (x_1, y_1), \ldots, (x_m, y_m)\}$. У цьому випадку завдання навчання полягає в знаходженні такої функції $h: X \rightarrow Y$, яка представляє функцію g на всій її області визначення, у тому числі значеннях, що не входять в (D). Шукана функція являє собою додаток теорії оптимізації [26].

На виході мережі одержано вихідне значення $h(x)$, яке одержано шляхом послідовного навчання нейронів усіх шарів, $g(x)$ – значення шуканої функції для того самого образа. Далі надходить процедура виконання етапу зворотної похибки, який полягає в обчисленні приватної похідної для кожного нейрона мережі стосовно його ваг:

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial o_j} \frac{\partial o_j}{\partial net_j} \frac{\partial net_j}{\partial w_{ij}}.$$

Дані на кожному кроці навчання, ваги нейронів кожного шару збільшуються значеннями часток похідних відповідно до методу градієнтного спуска. Зміни алгоритму навчання включає особливі додаткові заходи її організації в цілях захисту від повторення навчання, а також застосування різних оптимізаторів.

Багатошарові мережі використовуються для розпізнавання образів деяких різних класів, таких як язикові знаки, почерк, рукопис. У цей час багато програм використовують безпосереднє навчання із учителем для розпізнавання образів. У нейронних мережах застосовуються методом опорних векторів, що є більш ефективним рішенням із точки зору обсягу ресурсів [27].

Недоліки алгоритму зворотного поширення помилки. Даний алгоритм здійснюється по методу градієнтного спуска щодо площини помилок. Це виражається в наступному: у певній крапці площини існує напрямок швидкого спуска, потім здійснюється стрибок униз, на якусь відстань, пропорційно параметру швидкості навчання й складності нахилу, при цьому йде прагнення до збереження колишнього напрямку руху. Для всіх навчальних спостережень, узятих у довільному порядку окремо обчислюються крок спуска, що приводить до одержання гарної апроксимації множини у загальній площині помилок. Метод назад поширення досить простий і застосовується для рішення багатьох завдань, але при цьому має безліч серйозних недоліків. Процес навчання алгоритму по даному методу тривалий. У деяких випадках для навчання мережі потрібно кілька днів, а іноді кілька тижнів, що приводить до неможливості проведення навчання [27]. Це відбувається по наступних причинах.

Збій мереж – значення даних у результаті корекції можуть стати дуже більшими величинами в процесі навчання мережі, у результаті цього на виході всіх або більшості нейронів будуть видаватися більші значення, де функції активації буде не ефективна.

Процес навчання може практично стати замороженим, тому що помилка, що посилає назад під час навчання пропорційна похідною функцією активації. Подібні кроки розглядаються як експериментальні для запобігання збія мережі або для відновлення мереж після його походження.

Алгоритм зворотного поширення помилки прагне до мінімізації, використовуючи різновид градієнтного спуска, тобто здійснює перегони вниз по поверхні помилки, постійно регулюється в напрямки до мінімуму. Поверхня помилки комплексної мережі сильно стискується й складається з різнорідних

відрізків (ярів, долін, пагорбів, складок). Коли поруч існує більш глибокий мінімум, то мережа може перебуває в локальному мінімумі. У такий спосіб мережа може вибратися із крапки локального мінімуму, коли весь його напрямку ведуть нагору. Можна уникнути цієї пастки за допомогою статичних методів [28].

Алгоритм зворотного поширення помилки має розбіжності. Це полягає в нескінченній малій корекції ваг. На практиці видне, що даний доказ нездійсненний, тому що це веде до довгому, а навіть до нескінченного часу навчання. З досвіду пропонується визначити розмір кроку. Постійна нестійкість і збій можуть виникати при більших розмірах кроку. При малих розмірах кроку сходження занадто повільне.

Навчання повинне бути стійким, тому що нема рації в навчанні, якщо нова навчальна безліч забувається. У доказі сходження дана умова виконується, але при виконанні корекції необхідно надати мережі всі вектори навчального безлічі. На всій безлічі обчислюються необхідні зміни ваг, але це вимагає додаткової пам'яті. Ваги прагнуть до мінімальної помилки після трохи навчальних циклу. Коли мережа перебуває в зовнішньому середовищі, що безупинно міняє, те даний метод буде пошукам через не повторення того самого вектора. Коли мережа безцільно блукає або багато осцилюється, процес навчання колись не закінчиться. Алгоритм зворотного поширення помилки в цьому випадку не відбиває біологічну систему.

1.5 Класифікація за допомогою вирішальних функцій

У завданні розпізнавання образів, завдання опису й утвору класів є однієї з основних завдань. Наприклад, є якась (кінцева) множина класів $A = \{\omega_1,\ldots, \omega_m\}$. Кожний існуючий образ x характеризується деяким набором ознак у просторі ознак – вектором x . Увесь простір розпізнаваних ознак X розбивається на m + 1

попарно неспільних множин (повну групу множин) $X_0, X_1, \ldots, X_m$: $X_i \cap X_j = \varnothing$ $\forall\ i \neq j$, таким чином, що $x \in \omega_i$, якщо $x \in X_i$ [28].

Якщо $x \in X_0$, то образ x не належить не одному класу тобто потрапив в область «невизначеності» і в такому випадку не можливо класифікувати його. Визначено, що множина X_i є множиною переваги класу ω_i у просторі ознак X (див. рис. 1.5). Таким чином, границями класів розпізнаваних образів будуть границі областей X_i (i = 0, 1,…, m).

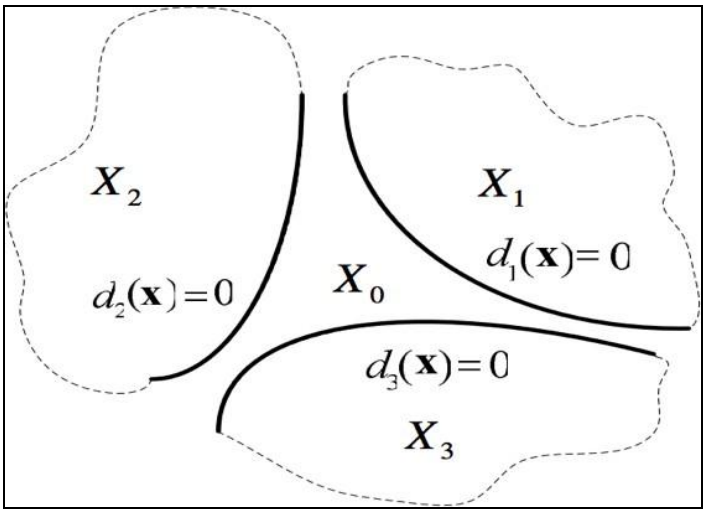

Рисунок 1.5 – Простір ознак

Автоматичне знаходження границь класів – одне з головних завдань теорії розпізнавання образів. Границі класів розпізнаваних образів дозволено визначати по-різному, приміром, з підтримкою думки вирішальної функції. Припущення, якщо простір ознак є n-мірним метричним простором R^n.

У цьому випадку передбачається, що існує m +1 функція $d_j(x)$, $x \in R^n$ (вирішальні або дискримінантні функції) такі, що $X_j = \{ x \in R^n : d_j(x) > 0\}$. Поверхня $S_j = \{x \in R^n : d_j(x) = 0\}$ називається поділяючою [11]. Можна вважати, що образ x належить класу ω_i, якщо виконуються нерівності $dj(x) < 0, \forall i \neq j$ та $d_i(x) > 0$.

1.6 Алгоритм Хо-Кашьяпа

Процедури алгоритму Хо-Кашьяпа [42]. Крок 1. На початку обирається довільний вектор y^0 с N позитивними координатами, потім обчислюється $w^0 = V^0y^0$ і покладається $k = 0$.

Крок 2. Перевіряється умова останова $\boldsymbol{Vw}^k > 0$. Якщо умова виконується, то зупиняється й завершує алгоритм роботу, якщо не виконується тобто а якщо ні, то необхідно переходити до пункту 3.

Крок 3. Обчислюються вектори $y^{(k+1)}$ та $w^{(k+1)}$ за формулами:

$$\mathbf{y}^{(k+1)} = \mathbf{y}^k + h_k\left(\mathbf{Vw}^{(k)} - \mathbf{y}^{(k)}\right)^+, k = \overline{1,N} \text{ и } \quad \mathbf{w}^{(k+1)} = \mathbf{Vy}^{(k+1)}, k = \overline{1,N},$$

якщо нарощується k, то необхідно переходити до пункту 2.

НСД-алгоритм є цікавим, тому що має одну властивість:
якщо на деякому k-му кроці алгоритму виявиться, що всі помилки, $(\mathbf{Vw}^{(k)} - \mathbf{y}^{(k)})^+ = 0$, але, в цьому випадку $\mathbf{Vw}^{(k)} - y^{(k)} \neq 0$ це означає, що класи точно не є лінійно роздільними.

У противному випадку тобто якщо класи є лінійно роздільними, то алгоритм обчислення вагового вектора w по формулах

$$\mathbf{y}^{(k+1)} = \mathbf{y}^k + h_k\left(\mathbf{Vw}^{(k)} - \mathbf{y}^{(k)}\right)^+, k = \overline{1,N} \ ,$$

$$\mathbf{w}^{(k+1)} = \mathbf{Vy}^{(k+1)}, k = \overline{1,N} \ .$$

сходиться. Саме такий підхід до знаходження вирішальної функції називається алгоритмом найменшої середньоквадратичної помилки (НСК-алгоритмом) або алгоритмом Хо-Кашьяпа (Ho Y.C., Kashayp R.L.).

У деяких випадках коли координати векторів, що класифікують ознак є випадковими величинами, те й помилка неправильної класифікації стане випадковою подією. Тоді завдання побудови вирішальної функції зводиться до знаходження такої функції, яка мінімізувала б можливість неправильної класифікації.

Баєсівський підхід [29] заснований на статичному характері спостережень. Його підставою є припущення про те, що на просторі образів існує імовірнісний захід, який або можна оцінити, або вона відома. У цьому випадку метою є розробка класифікаторів, які будуть правильно оцінювати ймовірності й відносити образи до класу, що походить, за значенням імовірності. Це означає, що завдання полягає у виділенні найбільш відповідного класу по ймовірності [30].

Нехай завдана множина M класів $\omega_1, \omega_2, \ldots, \omega_m$, а також $P(\omega_i \mid x)$, $i = 1,\ldots, M$ ймовірність того, що об'єкт, що не був ідентифікований та представлений вектором ознак x належить класу ω_i . $P(\omega_i \mid x)$ завдається в результаті експерименту після отримання вектору ознак і називається апостеріорною ймовірністю.

Апріорна ймовірність $P(\omega \mid x)$ отримується таким чином, якщо відомі $P(\omega)$ та $P(x \mid \omega_i)$.

$$P(AB) = P(A|B)P(B), P(AB) = P(B|A)P(A)$$
$$P(A|B)P(B) = P(B|A)P(A)$$
$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Таким чином, задача співставлення по апостеріорної ймовірності приводиться до обчислення значень $P(\omega_1)$, $P(\omega_2)$, $P(x \mid \omega_1)$, $P(x \mid \omega_2)$.

1.7 Постановка задач дослідження

Якщо є достатньо даних для визначення ймовірності приналежності об'єкта кожному із класів $P(\omega_i)$, $i = 1,2$.

Такі ймовірності називаються апріорними ймовірностями класів. А також якщо відомі функції розподілення вектора ознак для кожного класу $P(x \mid \omega_i)$, $i = 1,2$. Вони називаються функціями правдоподібності x по відношенню до ω_i. Класична постановка задачі кластерного аналізу передбачає продовження процесу

побудови нових кластерів аж до повного вичерпання елементів вихідної навчальної вибірки. На практиці, це не завжди можливо, оскільки у вихідній базі об'єктів можуть знайтись такі, образи яких не можуть бути віднесені до жодного з сформованих кластерів.

Підхід Баєса заснований на гіпотезі об існуванні деякого розподілу ймовірностей для кожного параметра. Недоліком цього методу є у випадку $x \notin \omega_i$ неможливо класифікувати об’єкт. Як правило, це об'єкти, частина ознак яких лежать в безпосередній близькості від точок верхніх або нижніх граней множини X за відповідними координатами. Ці об'єкти залишаються внекластерними, але, при певних обставинах можуть стати центрами нових кластерних утворень.

Особливо слід звернути увагу на неприпустимість наявності загальних елементів у різних кластерів. Якщо ж в ході кластерізації виявиться факт перетину хоча б пари кластерів, то це є незаперечним свідченням слабо вираженої тенденції угруповання в навчальній вибірці і неефективності апарату кластерного аналізу стосовно до даної множини.

Метою роботи є розробка алгоритму і програмного забезпечення, орієнтованих на рішення завдань розпізнавання образів з довільною розмірністю простору ознак, що класифікують, при наявності динамічних змін.

Для досягнення мети необхідно розв'язати наступні завдання:

– розробити алгоритм із можливостями саморозвитку й самоорганізації, який би міг ефективно використовуватися для вирішення завдань розпізнавання незалежно від розмірності простору ознак;

– розробити й апробувати алгоритми кластерізації, що діють на заданій множині ознак, що характеризують, певну предметну область;

– створити спеціалізований програмний комплекс для практичної реалізації цих алгоритмів;

– провести програмний експеримент ефективності розроблених методів.

2 ОПИС ПРОВЕДЕНИХ ТЕОРЕТИЧНИХ ДОСЛІДЖЕНЬ

2.1 Аналіз методів розпізнавання образів

Коло областей людської діяльності, де зустрічаються завдання, пов'язані із проблемою розпізнавання образів, надзвичайно великий і продовжує неухильно розширюватися. Системи медичної діагностики, утворювальні Smart-системи, різного роду системи охорони й сигналізації, система пошуку й обробки інформації – от далеко не повний список тих сфер, де актуальність цих завдань не викликає сумнівів. У самому загальному виді завдання розпізнавання образів може бути сформульована досить просто: по деякому набору ознак установити приналежність об'єкта одному з відомих класів, або обґрунтувати неможливість його класифікації. Однак, відмінності предметних областей, для яких ставиться та або інше завдання, роблять їх суттєво різними й диктують різні вимоги, пропоновані до використовуваних алгоритмів.

Головними з них є дві відмінності. Перше, безпосередньо пов'язано з особливостями розв'язуваного завдання, розмірність факторного простору ознак, що класифікують. Друге, обумовлене характеристиками проектованої системи, швидкість розпізнавання. Складні, багатостадійні процедури, застосовувані для завдань, де розмірність простору ознак значний, будуть явно надлишкові для завдань із невеликим числом ознак, що класифікують. Висока швидкодія обов'язкова для виявлення осередку загоряння в системах пожежної сигналізації буде зайвим в утворювальних системах.

Наведені умови дозволяють стверджувати, що не слід розглядати доцільність застосування алгоритмів, що мають певний набір якостей, для завдань розпізнавання образів взагалі, без врахування їх приналежності до однієї із чотирьох категорій, які визначаються з однієї сторони розмірністю простору ознак, що класифікують завдання, а з іншої вимогами до швидкодії проектованої системи. Схема категорування завдань розпізнавання образів представлена в таблиці 2.1.

Таблиця 2.1 – Схема категорій

Розмірність простору ознак, що класифікують			
мала		більша	
Категорія 1 Швидкодія низька	Категорія 2 Швидкодія висока	Категорія 3 Швидкодія низька	Категорія 4 Швидкодія висока

Аналіз методів кластерного розпізнавання показав шість основних якостей, яки повинні мати ефективні алгоритми розпізнавання.

– простота реалізації;

– чутливість до шумів;

– здатність до саморозвитку;

– здатність до самоорганізації;

– інтенсивність потоку даних;

– можливість роботи в режимі реального часу.

Аналіз даних показує, що для всіх категорій завдань пріоритетними визнано такі якості алгоритмів як здатність до саморозвитку й самоорганізації. Це означає, що системи розпізнавання образів високого рівня повинні бути самонавчальні, мати можливості масштабування, обробляти динамічні послідовності. Тобто, процедура розпізнавання образів є а систематизований ланцюжок дій, що змінюється під впливом зовнішнього середовища, і формує формалізоване представлення про неї. Простота реалізації більш затребувана для алгоритмів, що задіяні на завданнях з невеликою розмірністю простору ознак, а можливість роботи в режимі реального часу більш значима для систем з високими стандартами швидкодії. Приблизно однаковими для всіх категорій виявилися вимоги до завадостійкості.

Наступним етапом є ранжирування методів, застосовуваних у завданнях розпізнавання образів. Показник швидкодії є, скоріше, вимогою до системи, яка алгоритм реалізує, чому характеристикою алгоритму або властивістю вхідного потоку даних, то зазвичай обмеження розбивки завдань тільки на дві категорії: по розмірності простору кваліфікуючих ознак.

Алгоритми розпізнавання охоплюють весь спектр ідей, якими на сьогоднішній день мають фахівці в цій області.

– метод опорних векторів;
– метод «найближчого сусіда»;
– метод Баєса;
– метод галузей і границь;
– метод потенційних функцій;
– алгоритм внутрігрупових середніх;
– нейронні мережі.
– алгоритм хвильової кластерізації.

Саме цей набір методів, використовуваних у завданнях розпізнавання образів, обрано з точки зору їх застосовності для обробки масивів, де число кваліфікуючих ознак мале, і масивів, де кількість таких ознак велика.

Перевірка статистичної значимості значень коефіцієнтів рангової кореляції проводиться за критерієм χ^2 для числа ступенів свободи $f = n - 1$, розрахункове значення якого набуде табличне значення цього критерію при рівні значимості 5% $\chi^2_{0,05} = 14.1$.

Обчислене значення коефіцієнта рангової кореляції $K = 0.854$, що, у сукупності з високим значенням критерію $\chi^2_{\text{роз}} = 47.8$, показує статистичну вірогідність отриманих результатів. Це, у свою чергу, дозволяє використовувати експертні оцінки для формування обґрунтованих суджень і висновків.

При аналізі матеріалів таблиці 2.2 представлено відразу кілька методів рівною мірою прийнятними (або неприйнятними) для рішення певного класу завдань. Лідерами є алгоритми методу хвильової кластерізації й нейронних мереж. Вони цілком ефективні для завдань розпізнавання з невеликою розмірністю простору ознак.

Ще однією важливою їхньою перевагою є здатність до самонавчання і самоорганізованість, що дозволяє досить успішно використовувати ці методи в динамічних системах розпізнавання образів. Далі з невеликим відставанням йдуть

методи найближчого сусіда й Баєса.

Таблиця 2.2 – Оцінки методів для завдань із малою розмірністю простору ознак, що класифікують

Оцінювані методи	Експерти								$\sum_{i=1}^{m} r_{ij}$
	1	2	3	4	5	6	7	8	
Метод опорних векторів	7	5	8	6	7	5,5	7	8	53,5
Метод найближчого сусіда	2	4	1,5	4	4	2,5	3	2	23
Метод Баєса	4,5	2,5	3,5	1,5	2	2,5	4	4	24,5
Метод галузей і границь	4,5	6	5	4	5	5,5	5,5	6	41,5
Метод потенційних функцій	8	7	6	7,5	7	7,5	8	6	57
Метод внутрігрупових середніх	6	8	7	7,5	7	7,5	5,5	6	54,5
Нейронні мережі	2	2,5	1,5	4	2	2,5	1,5	2	18
Метод хвильовий кластерізації	2	1	3,5	1,5	2	2,5	1,5	2	16
Зв'язані ранги	T_1=30	T_2=6	T_3= 12	T_4= 36	T_5= 48	T_6= 72	T_7= 12	T_8= 48	$\sum_{j=1}^{8} T_j = 264$

Ці алгоритми не представляють складності в реалізації, однак властивостями самонавченості й самоорганизуємості не мають. Явними аутсайдерами списку виявилися метод опорних векторів, метод потенційних функцій і метод внутрігрупових середніх. Оцінки, проставлені експертами, указують на те, що в силу ряду причин застосування цих методів для завдань із малої розмірності простору ознак, що класифікують, недоцільно.

В табл. 2.3 наведено аналіз тих же методів стосовно до завдань із великою розмірністю простору ознак, що класифікують.

Обчислене по даним цієї таблиці значення коефіцієнта рангової кореляції також виявилося досить високим, хоча й меншим ніж для завдань із малою розмірністю простору ознак, а саме $K = 0.702$.

Розрахункове значення критерію, перевищує табличне $\chi^2_{\text{роз}} = 39.3$. Ти методи, які для рішення завдань великої розмірності малопридатні, це саме ті методи, які для завдань із малою розмірністю простору ознак мали найвищі ранги:

метод найближчого сусіда, нейронні мережі, метод хвильовий кластеризації. Спроби застосувати ці методи до завдань великої розмірності натрапляли на значні труднощі різної природи й успіху не мали.

Таблиця 2.3 – Оцінки завдань із великою розмірністю простору ознак

Оцінювані методи	Експерти								$\sum_{i=1}^{m} r_{ij}$
	1	2	3	4	5	6	7	8	
Метод опорних векторів	5	6	7	6,5	1,5	2	1,5	4	33,5
Метод найближчого сусіда	7	5	7	6,5	8	7	6	8	54,5
Метод Баєса	1,5	4	2	3	1,5	4,5	1,5	2	20
Метод галузей і границь	1,5	3	4	1	4	4,5	3,5	5	26,5
Метод потенційних функцій	4	1	2	3	5	2	3,5	2	22,5
Метод внутрігрупових середніх	3	2	2	3	3	2	5	2	22
Нейронні мережі	7	7	5	6,5	6,5	8	7	6,5	53,5
Метод хвильовий кластеризации	7	8	7	6,5	6,5	7	7	6,5	55,5
Зв'язані ранги	T_1=30	T_2=0	T_3= 48	T_4= 84	T_5= 12	T_6= 30	T_7= 36	T_8= 30	$\sum_{j=1}^{8} T_j = 270$

Сумарні ранги інші методів значно вище й досить близькі друг до друга. З дуже незначним відривом поставлений метод Баєса, що виглядає цілком природньо, враховуючи імовірнісний характер задіяних у ньому процедур. Однозначних рекомендацій з вибору методу рішення завдань розпізнавання образів з більшим набором класів і ознак не існує. Вибір же найбільш підходящого алгоритму представляє самостійну проблему й залежить не тільки від факторів розмірності, але й від конкретних особливостей і змістовного змісту завдання.

У цілому за результатами аналізу можуть бути зроблені наступні висновки:

– найважливішими якостями алгоритмів розпізнавання образів, у баченні

експертного співтовариства, є здатність до саморозвитку й самоорганізації, що має на увазі можливість їх використання в багато-стадійних динамічних процедурах;

– застосовність тих або інших методів для рішення завдань розпізнавання в самій значній мері залежить від розмірності простору ознак, що класифікують;

– актуальною і затребуваною є проблема розробки алгоритму з можливостями саморозвитку й самоорганізації, який би міг ефективно використовуватися для рішення завдань розпізнавання будь-якої розмірності.

2.2 Розробка математичної моделі динамічної кластерізації

У кваліфікаційній роботі для рішення завдання розпізнавання пропонується алгоритм на базі апарата кластерного аналізу [96]. На відміну від відомих [93-95], цей алгоритм дозволяє враховувати динамічні зміни бази порівнянь і потоку розпізнаваних образів і використовувати їх у ході подальших досліджень.

Кластером (cluster) у традиційнім розумінні, яке прийнято в справжній статті, називається сукупність об'єктів (образів) $\{x_i\}$, що задовольняють вимозі $\|x_i - x_j\| < d$, де $\|.\|$ символ має сенс заходу близькості між об'єктами; d – заздалегідь визначене граничне значення відповідно до обраного заходу. У данім дослідженні, як і в більшості подібних робіт, у якості заходу близькості використана евклидова метрика [11]: $\|x_i - x_j\| = \sqrt{\sum_{k=1}^{n}(x_{ik}^2 - x_{ij}^2)}$, де n –обсяг простору ознак, що характеризують розпізнаваний об'єкт.

Застосування кластерного аналізу як інструмента для рішення завдання розпізнавання образів буде тим більше успішним, чим вище тенденція математичних образів досліджуваних об'єктів, як точка n-мірного простору, до угруповання близько деяких центрів. Разом з тим, як відзначалося вище, було б нереалістичним очікувати збереження незмінним положення цих центрів із часом. Звідси випливає, що кластери в алгоритмах розпізнавання образів повинні

розглядатися як динамічні структури.

Нехай є деяка множина об'єктів X з відомими властивостями, образи яких можуть бути задані точками в просторі R^n. Надалі ці множини слід назвати навчальною вибіркою. Розіб'ємо її на m непересічних підмножинткластерів $X_1, X_2, \dots, X_m$, так, щоб $X_1 \cup X_2 \cup \dots \cup X_m = X$, а $X_i \cap X_j = \emptyset$. Для $\forall i \neq j$ операція, що представляє собою перший етап кластерного аналізу, є вкрай невизначеною й слабко формалізованою. Геометрія, розміри кластера, чисельний склад елементів, що входять у нього, та критерії подібності між ними – усі ці характеристики визначаються змістом і особливостями конкретного завдання, і їх вибір практично цілком залежить від чисто суб'єктивних факторів: професійної кваліфікації й інтуїції дослідника.

Згідно аналізу поточного стану розробок в роботі пропонується сферична форма кластерів. Складність математичного опису сфери, як геометричного тіла, мало залежить від розмірності простору. І, оскільки число кваліфікуючих ознак у завданні розпізнавання образів може бути досить великим, опис кластерів у формі сфероїдів забезпечить простоту представлення й інтерпретації результатів.

Для побудови кластерних сфероїдів необхідно всі ознаки, що класифікують, привести до єдиних безрозмірних одиниць. Із цією метою, шляхом аналізу апріорної інформації й матеріалів навчальної вибірки, установлюються точна верхня x_{jsup} точна нижня x_{jinf} грані по кожному з n ознак у проведенім дослідженні й виконується перехід до безрозмірних змінним $y_j, j = \overline{1,n}$ за формулою:

$$y_j = M \frac{x_i - x_{jinf}}{x_{jsup} - x_{jinf}}, \qquad (2.1)$$

де M − масштабуючий множник, обраний з міркувань зручності представлення даних.

Також, x_{jsup} і x_{jinf} не обов'язково повинні збігатися з найбільшим і найменшим значенням ознак, що класифікують, елементів вихідної навчальної вибірки.

На множині елементів $\{y_j\}$ модифікованої навчальної вибірці Y виділено

підмножину образів, що утворюють достатньо тісне угруповання, і будується перший кластерний сфероїд $B^1(r)$ із центром у точці e_o^1, яка розташована в безпосередній близькості від геометричного центру угруповання. Точку e_o^1 названо початковою базовою точкою; верхній індекс указує номер кластера, а нижній індекс номер точки.

Радіус сфероїда r визначається згідне цілям дослідження й особливостям угруповання елементів навчальної вибірки. Формалізовані рекомендації з його вибору дотепер не розроблені. Таким чином, до складу кластера $B^1(r, e_o^1)$ входять елементи модифікованої навчальної вибірки Y, що задовольняють умові $\|e_o^1 - y\|_i \le r$.

Стан отриманого кластера можна уточнити, оскільки геометричний центр області, займаний досліджуваною підмножиною навчальної вибірки, як правило, не збігається із центром ваги угруповання. Із цією метою обчислюється центр ваги сукупності елементів, що перебувають в границях сфероїда $B^1(r, e_o^1)$ за формулою:

$$e_o^1 = \frac{1}{|B^1(r, e_o^1)|} \sum_{y \in B^1(r, e_o^1)} y_i \qquad (2.2)$$

де $|B^1(r, e_o^1)|$ – потужність підмножини елементів у складі сфероїда $B^1(r, e_o^1)$.

Точка e_1^1 буде першою базовою точкою і геометричним центром сфероида $B^1(r, e_1^1)$. В результаті цих действійв $B^1(r, e_o^1)$, з'являються елементи, яких не було $B^1(r, e_1^1)$, а частина елементів, що входили до складу $B^1(r, e_o^1)$, буде втрачена.

В сфероїді $B^1(r, e_1^1)$ геометричний центр не буде збігатися з центром тяжкості, тому в тому ж порядку знаходяться координати другої базової точки e_2^1 є географічним центром сфероида $B^1(r, e_2^1)$ і т.д. Послідовність точок $\{e_h^1\}$ сходиться, як обмежена послідовність, певна на компакті, а значить, за кінцеве число кроків досягається виконання умови $\|e_h^1 - e_h^1\| < \varepsilon$, де ε – будь-який

заданий наперед позитивне число, що визначає бажану точність ітераційної процедури. Саме з цієї причини положення початкової базової точки e_0^1 принципового значення не має.

В ході практичної реалізації описаного алгоритму може з'ясуватися, що радіус сферичної оболонки кластера обраний невдало і частина елементів угруповання до складу кластера $B^1(r, e_h^1)$ не увійшло і це явно суперечить змістовному змістом завдання. В цьому випадку слід збільшити розміри сфероида $B^1(r, e_h^1)$ так, щоб найвіддаленіший від базової точки e_h^1 об'єкт, з тих, що повинні стати елементами кластера, потрапив на кордон нової розширеної сфероида. Позначено множину об'єктів модифікованої навчальної вибірки, які необхідно додатково впровадити в кластер B^1 через $\{\hat{y}_j\}$, і знайдено величину:

$$R = \max_{\{\hat{y}_j\}} \left\| e_h^1 - \hat{y}_j \right\| , \qquad (2.3)$$

яка і визначатиме радіус шуканого кластерного сфероїда $B^1(r, e_h^1)$ з центром в базовій точці e_h^1. Приріст радіуса при цьому складе $\Delta r = R - r$.

Після зміни радіуса і додавання до його складу нових елементів, центр ваги кластера зміститься щодо геометричного центру точки e_h^1. Цей зсув легко усувається за допомогою описаної вище процедури ітерації. Це явище матиме місце і тоді, коли формування кластера на базі вихідної навчальної вибірки завершиться, і система почне працювати в режимі розпізнавання образів, обробляючи потік об'єктів, що надходять на її вхід.

2.3 Алгоритм динамічної кластерізації

Якщо систему передбачається використовувати як динамічну , таку що сама навчається структуру, необхідно здійснювати корекцію положення і розмірів

кластерних сфероїдів постійно протягом всього періоду експлуатації. Зауважимо, однак, що в міру наповнення кластерів, зміщення центру ваги під час вступу нових об'єктів стає все менш і менш значним. І в тих випадках, коли воно не виходить за межі ε- околиці, коригуючий вплив перестає бути обов'язковим.

Крок 1 – на початку задається множина об'єктів навчальної вибірки A = {a_1, a_2, …, a_N }, також початкові кордону кластерів $Y = \{\{y_{1inf}, y_{1sub}\}, \ldots, \{y_{minf}, y_{msub}\}\}$, де y_{iinf} – верхня межа і y_{isub} – нижня межа.

Необхідно також задати число ε. Число *m* визначає кількість кластерів, заданий користувачем для угруповання розпізнаваних об'єктів. В результаті виконання алгоритму число *m* може збільшиться в залежності від появи об'єктів, що не входять в межі кластерів. Процес утворення нових кластерів для цих об'єктів відбувається на другому етапі роботи алгоритму.

Крок 2 – створюється цикл для утворення необхідних кластерів. За замовчуванням задається значення $j = 0$ і $j \le m$, j-го кластера. Встановлюються початкові кордону j-го кластера $\{y_{1inf}, y_{1sub}\}$. Перевіряється які точки об'єктів входять в межах першого кластера.

Крок 3: після визначення вхідних точок об'єктів в межах кластера розраховується радіус і будується *k*-й сфероїд для j-го кластера за формулою:

$$r^j{}_k = \max_{\{\hat{y}_i\}} \left\| e^j{}_0 - \hat{y}_j \right\| .$$

Крок 4 – розраховується центр ваги для *k*-го сфероїда *j*-го кластера за формулою:

$$e^j{}_k = \frac{1}{|B^{\,j}{}_k(r^j{}_k, e^j{}_{k-1})|} \sum_{y \in B^{\,j}{}_k(r^j{}_k, e^j{}_{k-1})} y_i .$$

Крок 5: перевіряється виконання умови $\| e^j_k - e^j_{k-1} \| < \varepsilon$ і, якщо воно не виконано, то число *k* збільшується на одиницю тобто переходить в наступний сфероїд (крок 3), в іншому випадку даний *k*-й сфероїд є остаточним. Далі, число *j* збільшується на одиницю і перевіряється виконання умови $j>1$ тобто процес утворення кластера переходить в наступний крок 5.

На рисунку 2.1 представлені кроки виконання алгоритму:

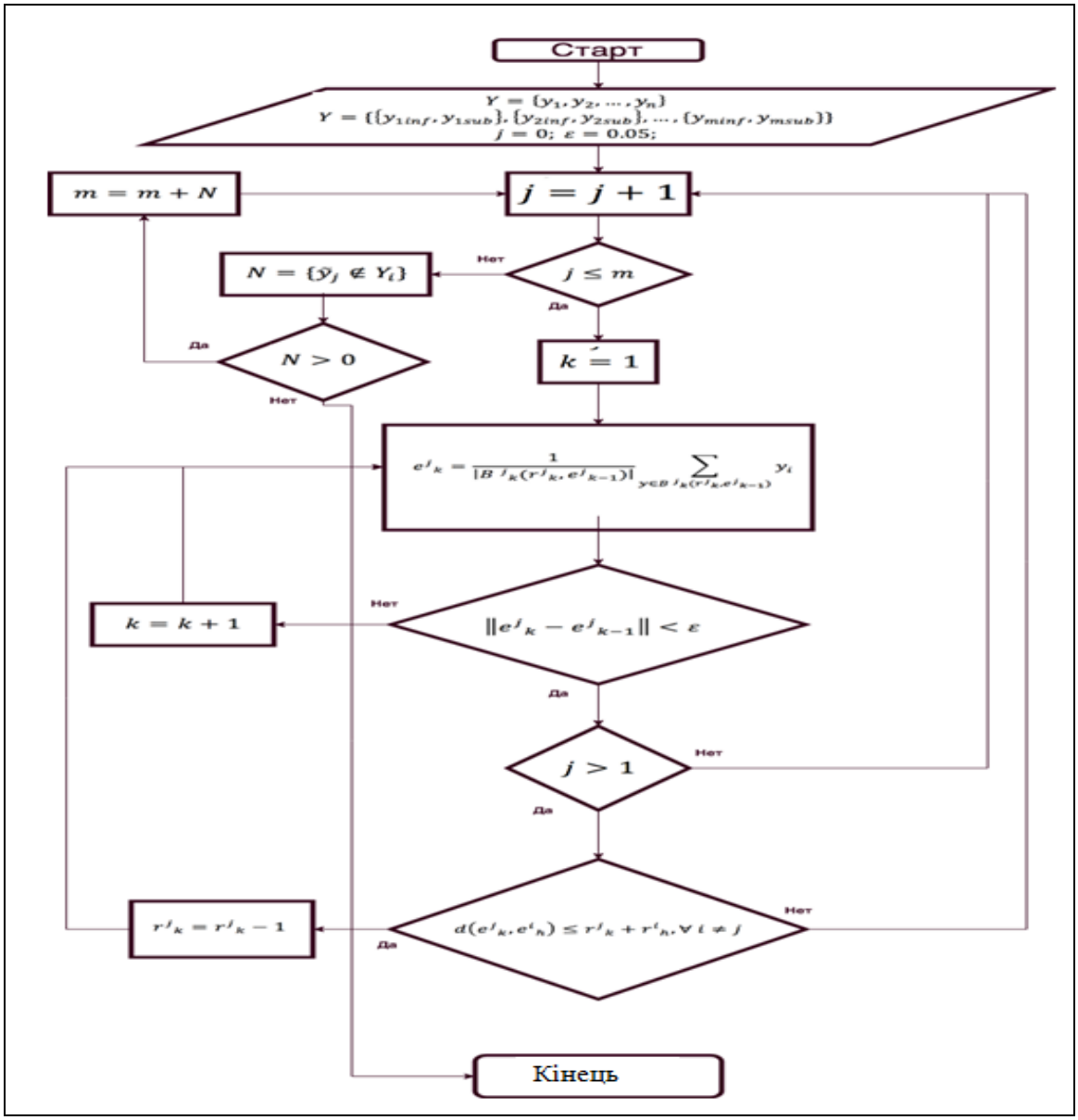

Рисунок 2.1 – Логічна схема алгоритму динамічної кластеризації

Крок 6: перевіряється виконання умови $d(e^j_k\ ,\ e^j_h) \le r^j_k + r^i_{h,}\ \forall$ i ≠□, і, якщо воно виконано, то *k*-й кластерний сфероїд має перетин з іншими раніше утвореними кластерними сфероїдами.

В такому випадку необхідно зменшити радіус *k*-го кластерного сфероїда. Після зменшення переходити до кроку 3. Якщо умови не виконуються тобто немає перетину з іншими кластерними сфероїді, то переходити до кроку 2.

Крок 7. При виконанні умови $j > m$ і $N \leq 0$, алгоритм закінчує роботу. По завершенні вказується число N – кількість об'єктів, що не входять в утворені кластерні сфероїди.

2.4 Аналіз алгоритму динамічної кластерізації

Після розробці методів і алгоритми класифікації об'єктів, для аналізу їх ефективності необхідна експериментальна перевірка на різних класах імітованих і реальних даних.

Вирішення цього завдання вимагає розробки алгоритмічного і програмного забезпечення. Для цього був розроблений комплекс прикладних програм, що дозволяє класифікувати об'єкти в різних предметних областях, а також вирішувати допоміжні завдання: пошук об'єкта за зразком, угруповання об'єктів та інші дії з об'єктами.

На рисунку 2.2 множини елементів навчальної вибірки, що наведені до безрозмірних змінних, представлено точками площини в координатах $y - V$ з масштабуючим множником $\mathrm{M} = 100.\varepsilon = 0.5$, де ε – завдане позитивне число, що визначає бажану точність ітераційної процедури.

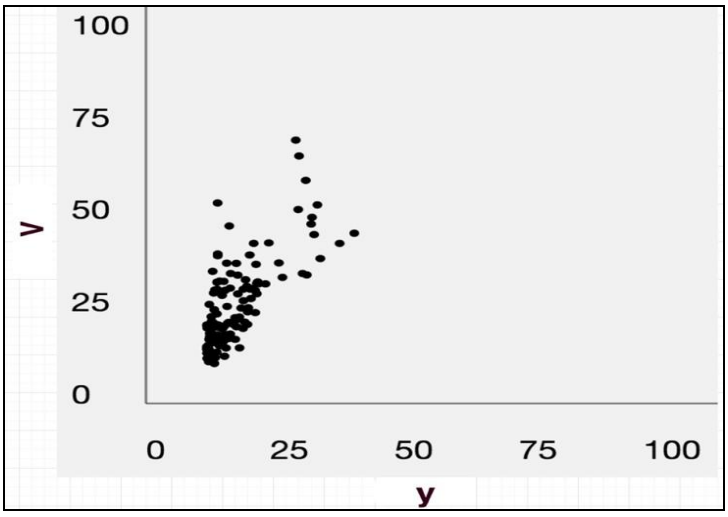

Рисунок 2.2 – Множина елементів навчальної вибірки в безрозмірних координатах

Для початку процедури кластеризації було прийнято рішення сформувати три вихідних кластера, кордони, яких і положення їх геометричних центрів. Запропонований поділ служить для того, щоб зафіксувати початкове положення для запуску викладеного вище алгоритму динамічної кластеризації. Процес еволюції кластерних сфероїдів, як результат дії цього алгоритму, що для визначення остаточного першого кластера виявилося досить двох ітерацій, при цьому радіус кластерного сфероїда зберіг своє первісне значення. Це пояснюється тим, що угруповання об'єктів вибірки в області першого кластера виражена найбільш чітко.

Процес перетворення другого і третього кластерів зажадав 5 і 6 ітерацій відповідно і супроводжувався більш значними змінами, оскільки вони торкнулися не тільки положення центрів кластерних сфероїдів, але і величини їх радіусів. Особливо помітно це проявляється для другого кластера, де радіус, в порівнянні з вихідним, змінився більш ніж в 3 рази, що було викликано необхідністю усунення можливості перетину кластерів. Помітно також, що зміна положення центрів кластерних утворень стійко направлено в сторону менших значень характеризують ознак, тобто в область більш вираженою щільності угруповання.

Проміжні етапи і кінцевий результат розрахунків, виконаних відповідно до розробленого алгоритму, докладно проілюстровано на рисунку 2.3.

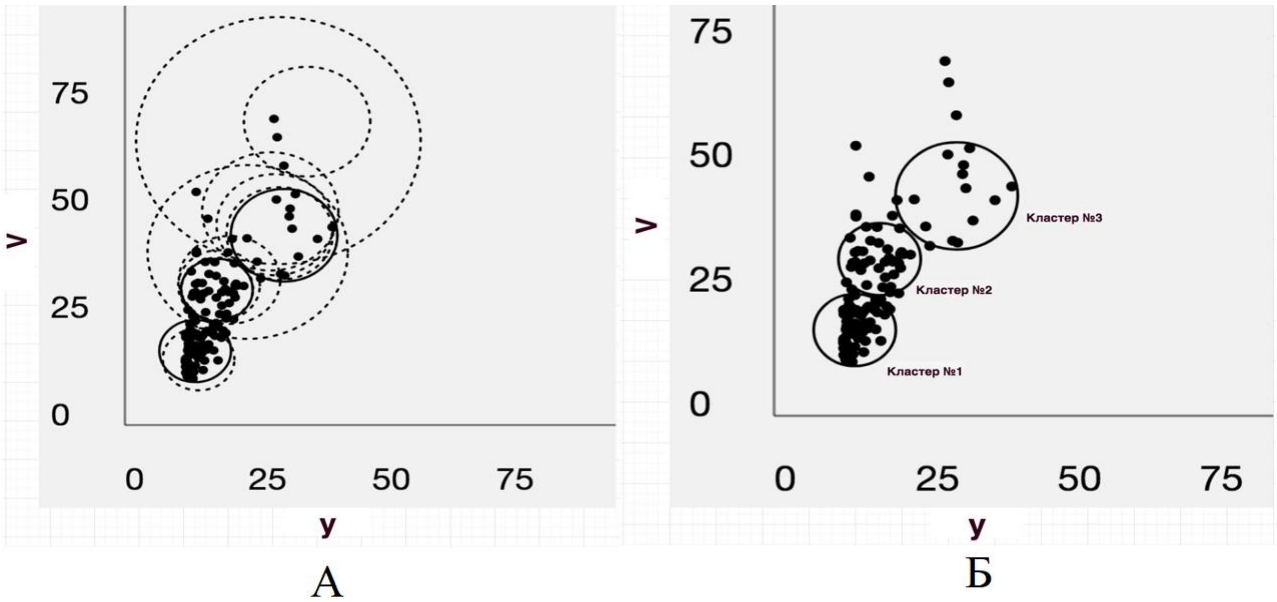

Рисунок 2.3 – Етапи і кінцевий результат розрахунків

У першій частині А рисунка представлена уся послідовність перетворень кластерних сфероїдів. Пунктирними колами зображені кордони проміжних кластерів, а суцільними – їхня кінцева позиція, яку можна вважати жорстко зафіксованої для обраного показника точності обчислень ε і заданого обсягу навчальної вибірки.

Матеріали таблиці 2.4 і рисунок 2.3 дають підставу виявити важливу властивість алгоритму динамічної кластеризації, а саме його самоорганізуємість – розміри радіуса другого кластера суттєво зменшилися, а положення центру змістилося в бік менших значень в порівнянні з вихідним. Це свідчить про те, що по-перше розміри кластерних сфероїдів не слід встановлювати, орієнтуючись на «круглі» числа, а по-друге в рамках навчальної вибірки є даними, що характеризують ознаки.

Таблиця 2.4 – Етапи перетворення кластерів

№	Центр кластера	Радіус *r*
1		
1	[220698, 102.5]	7.2
2	[227734, 108.1]	7.2
2		
1	[567468, 334.4]	10.2
2	[461127, 316.1]	8.2
3	[442760, 305.5]	7.2
4	[439437, 301.9]	8.2
3		
1	[1059716, 788.7]	29.7
2	[978274, 557.8]	13.7
3	[1024047, 514.5]	11.7
4	[1073611, 485.8]	10.7
5	[1117677, 480.0]	11.7
6	[1117677, 480.0]	10.7

З розгляду правої частини рисунку 2.3 видно, що кілька об'єктів сформованих кластерів та можливість їх впровадження в один зі створених кластерів шляхом збільшення радіусу – відсутня.

Для цих елементів навчальної вибірки алгоритм запускається повторно, результатом цього стає поява трьох нових кластерних утворень (див. рис. 2.4), число елементів в яких значно менше, ніж в перших трьох кластерах. Це не знижує цінність отриманих результатів і не дискредитує розроблений алгоритм, оскільки в міру поповнення навчальної вибірки і природних змін входять до неї об'єктів, структура кластерного спільноти повинна і буде змінюватися, що і складає головний змістовний сенс ідеї динамічної кластерізації.

Одним з аспектів кластеризації є те, наскільки легко вони перетворюються в діаграми та графіки.

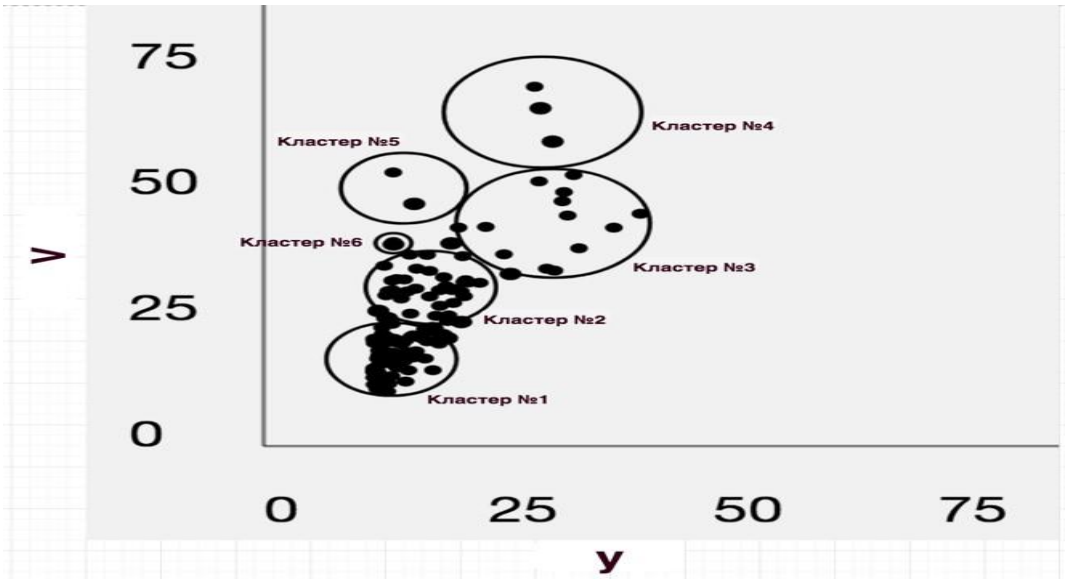

Рисунок 2.4 – Етапи і повторний результат розрахунків

У розглянутому прикладі розраховується кількість елементів в кожному кластері. Це розраховується шляхом підсумовування міст в кожному кластері. До складу кластера $B^1(r, e_h^1)$ входять елементи модифікованої навчальної вибірки Y, що задовольняють умові $\|e_0^1\|_i \leq r$.

Алгоритм розпізнавання, що сам розвивається, дозволяє в просторі класифікованих ознак формувати кластери і здійснювати коригування їх положення і розмірів згідно зі станом навчальної вибірки.

2.5 Моделювання процесів еволюції кластерних утворень

При цьому, однак, давався опис лише результату дії змін навчальної вибірки, і ніяк не досліджувався вплив їх основної рушійної сили – часу. Тобто за допомогою яких засобів може бути змодельована і відстежена зміна кластерних утворень в часі, тобто процес їх еволюції. Нехай в фазовому просторі класифікують ознак сформовані кілька послідовностей кластерів, заданих координатами центрів своїх кластерних сфероїдів.

Кожному кластеру поставлений у відповідність певний момент часу t_i, , на який цілком встановлено положення його центру відповідно до алгоритму, докладно описаному в попередньому розділі. На рисунку 2.7 наведено приклад, де таких послідовностей три, а положення кластерних центрів позначено символами x, Δ, ▌.

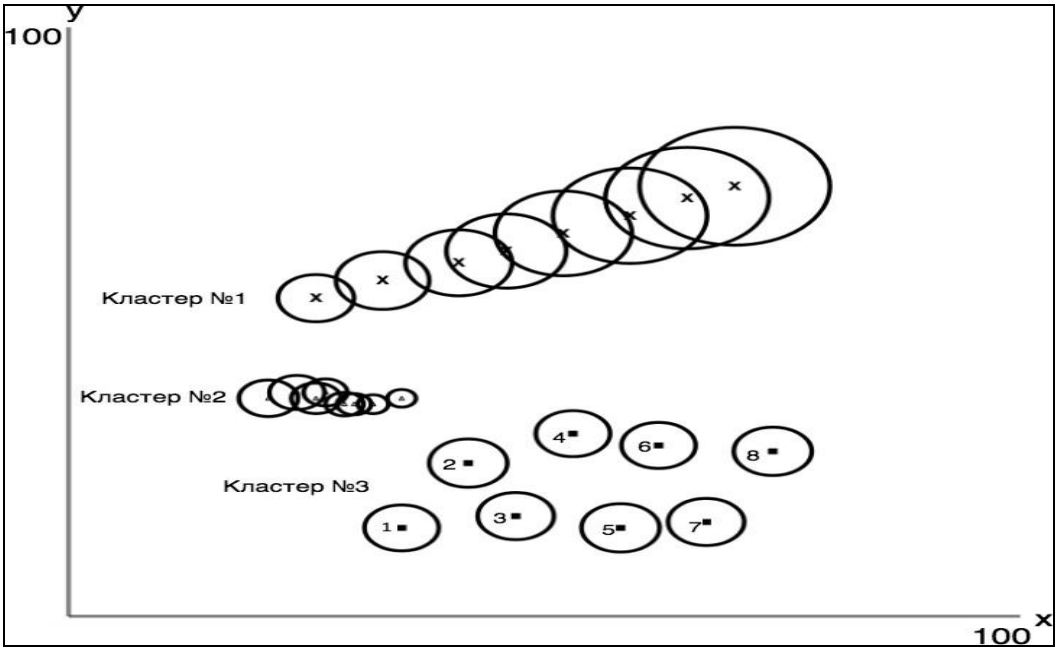

Рисунок 2.7 – Приклад розташування кластерних послідовностей

Таким чином, фактично, описані вище послідовності є ряди динаміки, рівні яких задаються не окремими числами, а числовими множинами, чиї потужності рівні розмірності простору ознак, що класифікують. Якщо ця розмірність дорівнює m, то для побудови моделі поведінки кожної кластерної послідовності необхідно отримати m рівнянь виду

$$x_j = p_j(t) \quad j = 1, \dots, m,$$

де x_j – елементи множини ознак.

Чим більшим є число m, тим більша складність обчислення при ідентифікації параметрів моделі і складнощів з інтерпретацією результатів моделювання. Отже, еволюційні зміни кожної з виявлених кластерних послідовностей повністю характеризуються m часовими рядами за кількістю класифікують ознак досліджуваної навчальної вибірки.

Значення рівнів часових рядів формуються як результат сукупної дії трьох складових: трендової T, циклічної S і випадкової E. Тому аналіз часових рядів рекомендується починати з оцінки незалежного вкладу кожної з них. Це здійснюється шляхом побудови діаграм кореляції або таблиць коефіцієнтів автокореляції. Якщо буде виявлено, що переважний вплив на формування рівнів ряду має трендова складова, то відповідний класифікують ознака визнається значимо впливає на хід еволюційних змін кластерної послідовності. Після цього встановлюється тип тренда і будується регресійна модель, яка після належної статистичної перевірки може бути використана для прогнозування подальших етапів еволюції. При наявності переважаючих трендових складових у двох і більше ознак слід перевірити гіпотезу ко-інтегрованості, тобто встановити, чи є виявлені тренди взаємообумовленими чи ні.

Для вирішення цього завдання можуть бути використані, наприклад, метод відхилення від тренда і ряд інших. Якщо гіпотеза ко-інтегрованості знайшла своє підтвердження, то сукупний вплив трендів відповідних ознак має бути обов'язково враховане при прогнозуванні еволюційного поведінки кластерної послідовності. В іншому випадку збереження трендових тенденцій в моменти часу, такі за які спостерігаються, не гарантоване і, отже, використання їх в прогнозуванні неправомірно.

У разі переважання, або помітної присутності в часовому ряду будь-якої ознаки з циклічною складовою, її вплив має бути елімінований. У згладженому таким чином часовому ряді оцінюється трендова складова, але він невиразний на

тлі циклічних змін, або обґрунтовується її незначимість на хід еволюційного процесу. Якщо виявиться, що вплив тренда істотний, його внесок враховується відповідно до описаної вище процедури, інакше його слід розглядати як шум – випадкову складову з імовірно нормальним законом розподілу. При цьому дія, що виникає характерною ознакою на еволюцію визнається нікчемним.

Значення ознак, що визначають положення центрів кластерів, в рівновіддалені один від одного моменти часу. передбачаються відомими так само як кількість елементів, наявних в складі кластерів.

3 ОПИС АЛГОРИТМІВ НАВЧАННЯ Й КЛАСИФІКАЦІЇ ОБ'ЄКТІВ

3.1 Структура і склад комплексу алгоритмів

Функціональний склад комплексу програм представлений на рис. 3.1, де наведено основні модулі програм і їх зв'язки.

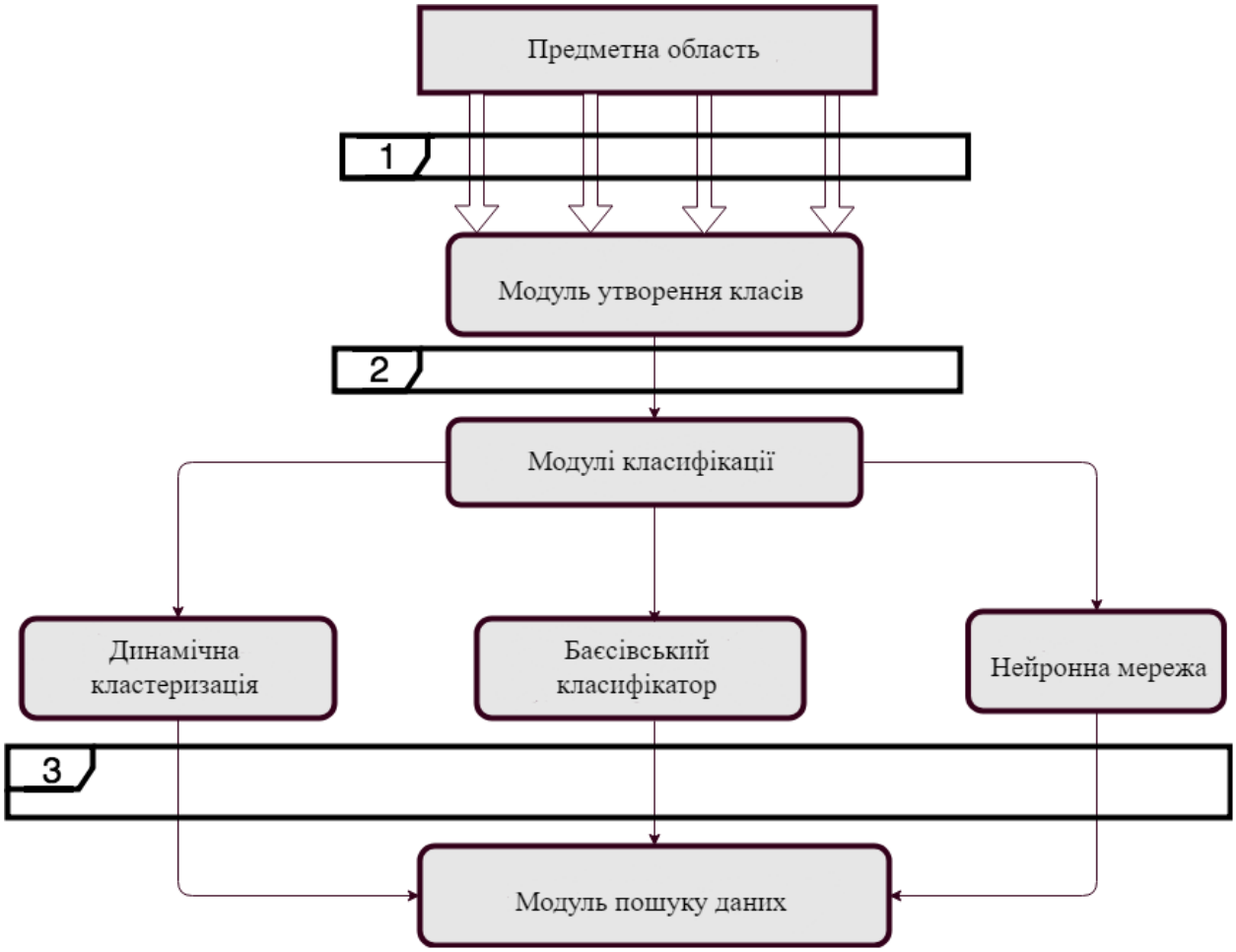

Рисунок 3.1 – Функціональний склад комплексу алгоритмів

Комплекс програм являє собою сукупність взаємозалежних модулів, які в процесі рішення завдань розпізнавання образів можуть працювати і як єдине ціле й окремо.

Пункт 1 – на вхід системи надходить потік об'єктів, що належать навчальній вибірці з деякої предметної області, які необхідно класифікувати. У роботі комплексу може бути виділено три пункти по числу блоків, наявних у його складі. У процесі класифікації навчальна вибірка може поповнюватися новими протягом усього часу роботи класифікаторів.

Пункт 2: результатом роботи модуля формування класів буде множина класів для розпізнавання M = { ω_1, ω_2, … , ω_m}, m − кількість класів. Якщо на початку процедури класифікації множина класів невизначена, вибирається класифікатор моделі динамічної кластеризації, при цьому навчання несе назву «навчання без учителя» [14]. Якщо опис множини відомо, слід вибрати класифікатор моделі нейронної мережі, тому що в нейронних мережах необхідно заздалегідь задати бажані значення на виході. Такий процес «навчання називається навчання з учителем».

Пункт 3 – у результаті класифікації формуються організовані класи або кластери, кожний клас має свою множина об'єктів ω_i = { a_1, a_2,…, a_N,} причому $\omega_i \cap \omega_j = \varnothing$, $\forall i \neq j$.

Модуль класифікації. Процедура класифікації об'єктів відіграє найважливішу роль у завданнях розпізнавання образів. У даній роботі модуль класифікації містить у собі три самостійні блоки, кожний з яких орієнтований на рішення певного типу завдань. Це блок динамічної кластеризації, що представляє авторську розробку, байесовский класифікатор і нейронні мережі. Нижче дана коротка характеристика кожного з них.

Блок динамічної кластеризації використовується як апарат кластерного аналізу стосовно до рішення завдання розпізнавання образів у припущенні, що образи об'єктів можуть бути інтерпретовані як вектори простору кінцевої розмірності. На відміну від інших відомих робіт, присвячених аналогічної тематиці, тут кластери розглядаються як динамічні утвори, положення й розміри яких змінюються в міру поповнення навчальної вибірки новими об'єктами. Розроблений простий і легко реалізований алгоритм, що дозволяє виконувати розрахунок необхідних змін.

Блок Баєсівського класифікатора. Значення умовних ймовірностей, одержувані по формулі Баєса, дають можливість установити раціональну черговість класифікації, указуючи варіанти можливої правильної класифікації один по одному убування їх рангів. Цей метод добре проявляє себе при роботі з навчальними вибірками великого обсягу.

Блок класифікації за допомогою нейронної мережі призначений для завдань, де характеристики класів відомі. У цьому випадку створюється нейронна мережа із зазначеними нейронами на вході й виході, а далі використовується розроблений алгоритму для навчання мережі використовується алгоритм методу зворотного поширення помилок.

3.2 Алгоритм створення нових класів

Поділ розглянутої множини об'єктів на класи може бути заданий такими способами:

Перерахування – кожний клас задається шляхом прямої вказівки його членів. Такий підхід використовується в тому випадку, якщо доступна повна апріорна інформація про всі можливі об'єкти розпізнавання. Пропоновані системі образи рівняються із заданими описами представників класів і ставляться до того класу, якого належать найбільш подібні з ними зразки. Такий підхід називають методом порівняння з еталоном. Він, приміром, застосуємо при розпізнаванні машинопечатних символів певного шрифту. Його недоліком є слабка стійкість до шумів і викривленням у розпізнаваних образах.

Завдання загальних властивостей: клас задається вказівкою деяких ознак, властивих усім його членам. Розпізнаваний об'єкт у такому випадку не рівняється прямо із групою еталонних об'єктів. У його первинному описі виділяються значення певного набору ознак, які потім рівняються із заданими ознаками класів. Такий підхід називається зіставленням за ознаками. Він економніше методу порівняння з еталоном у питанні кількості пам'яті, необхідної для зберігання описів класів. Крім того, він допускає деяку варіативність розпізнаваних образів. Однак, головною складністю є визначення повного набору ознак, що точно відрізняють членів одного класу від членів усіх інших.

Кластеризація – у випадку, коли об'єкти описуються векторами ознак або вимірів, клас можна розглядати як кластер. Розпізнавання здійснюється на основі розрахунку відстані опису об'єкта до кожного з наявних кластерів. Якщо кластери досить рознесені в просторі, при розпізнаванні добре працює метод оцінки відстаней від розглянутого об'єкта до кожного із кластерів. Складність розпізнавання зростає, якщо кластери перекриваються. Звичайно це є наслідком недостатності вихідної інформації й може бути дозволене збільшенням кількості вимірів об'єктів. Для завдання вихідних кластерів доцільно використовувати процедуру навчання.

На рис. 3.1 блок А, показано, що X_1 – область переваги першого класу ω_1, а X_2 – область переваги другого класу ω_2. $X_1 \cap X_2 = X$ − загальні ознаки. $(X_1 \cup X_2) \backslash (X_1 \cap X_2)$ − уніфіковані ознаки. І так далі – при вступі нових послуг у систему створюються нові класи.

На рис. 3.2. представлена процедура виконання даного алгоритму.

У блоці 1 А це навчальна вибірка об'єктів, у якій визначаються правила вибору ознак і кількість ознак даної предметної області. Множина *X* є сукупністю ознак предметної області, що складається з m ознак (у наведеному нижче прикладі *m =1,..., 18*), *x* – вектор ознак розпізнаваного образа, що утворюється з n ознак і зазвичай n < m.

У блоці 2 рівняється m з n, якщо m = n, та множина залишається такою ж й завдання (розпізнаваний образ) ставиться до загального класу.

У блоці 3 якщо n < m, те для порівняння визначається початкова ознака множини предметної області i=1 і початкова ознака розпізнаваного образа j=1, і починається порівняння початкової ознаки розпізнаваного образа, з усіма ознаками множини предметної області починаючи з першого. Якщо збігу з першою ознакою предметної області ні, то порівняння проводиться із другою ознакою предметної області *i+1*, і т. д. аж до того, поки не найдеться співпадаючий елемент у безлічі ознак або і станеться рівним *m* (*i = m*).

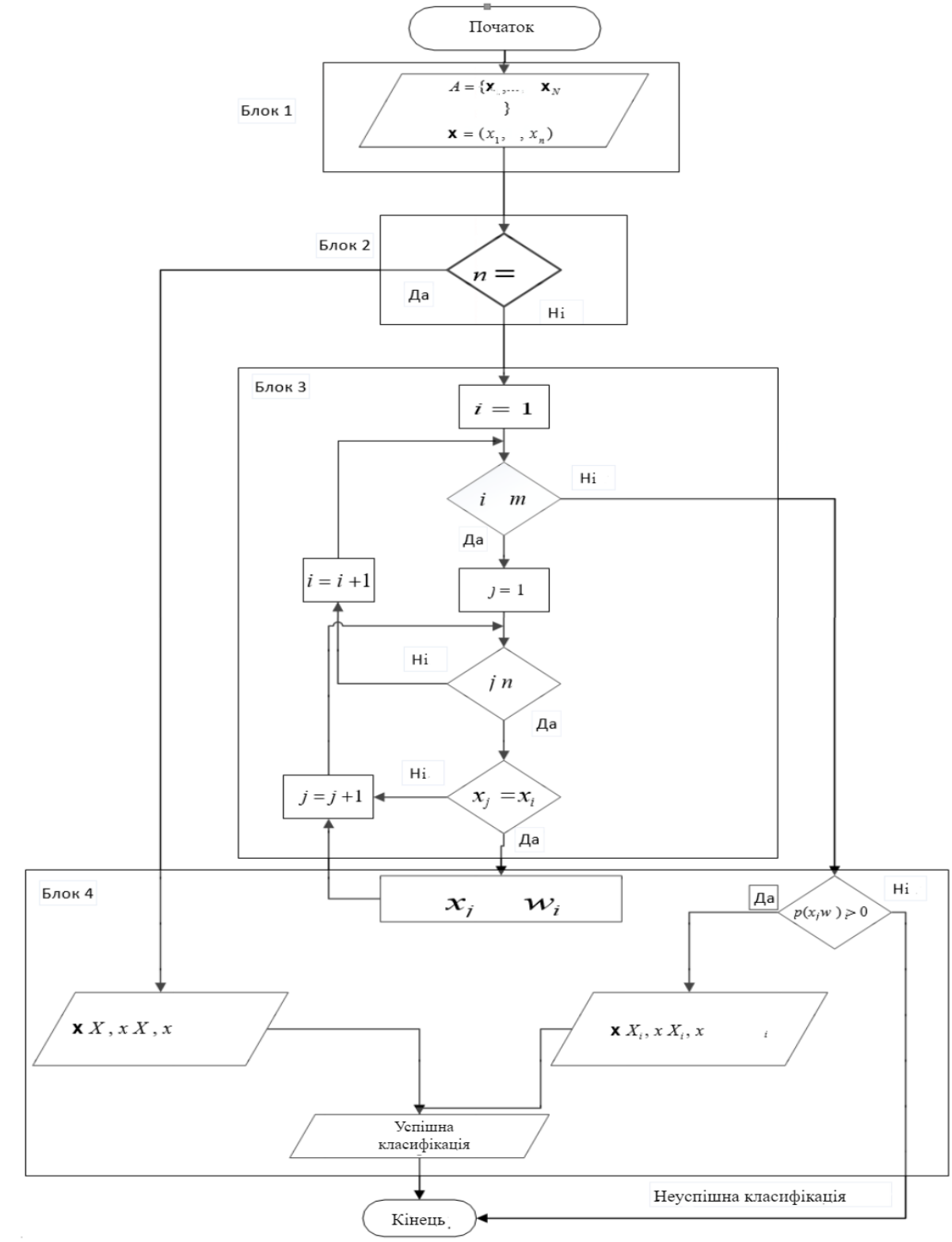

Рисунок 3.2 –Схема алгоритму класифікації предметної області

Потім на порівняння береться наступна ознака розпізнаваного образа j+1, і рівняється з усіма ознаками предметної області і т. д. аж до того моменту, коли j = n, див. рис. 3.1.

У блоці 4 після порівняння всіх ознак і умови того, що ознаки

розпізнаваного образа збігаються з ознаками в даній предметній області множини предметної області розбивається й з'являється новий клас ω_i та $x_j \in \omega_i$. Поява нового класу залежить від значення ймовірності збігу ознак розпізнаваного образа з ознаками множини предметної області $p(x_j \in \omega_i)$.

3.3 Модель класифікації об'єктів на основі теорії Баєса

Розроблено алгоритм для класифікації об'єктів, що використовує оцінку двомірних просторових характеристик об'єкта. Дана модель може використовуватися для формування класів і класифікації об'єктів різних типів даних на основі теорія Баєса (при цьому в данім дослідженні докладно розглядається приклади класифікації). У даній главі представлене аналітичне формулювання моделі і її компонентів.

Основу запропонованої моделі становлять два рівні абстракцій: вивчення предметної області (формування навчальної вибірки й утвору класів) і рівень класифікації екземпляра нових об'єктів.

Розроблена модель інтегрована в комплексній системі, за допомогою якої можна застосувати експериментальні дослідження й використовувати її для подальшої обробки даних. Розроблена система класифікації об'єктів, функціонально взаємозалежна сукупність методів і технічних засобів, що здійснює процес синтезу й аналізу розпізнаваних образів.

Запропонована модель підходить для різних предметних областей не залежно від їхнього розміру й типу даних, у тому числі охоплює якісні й кількісні характеристики. Розглядається предметна область із кінцевою безліччю різних об'єктів. Будь-який об'єкт із предметної області однозначно характеризується деякою безліччю ознак. При цьому, два різні об'єкти можуть мати загальні ознаки. Уся предметна область, у цілому, характеризується набором ознак, що входять у неї об'єктів. У даній главі розглядається два взаємозалежні завдання.

У завданні 1 потрібно розбити всю предметну область на класи об'єктів, тобто представити її як об'єднання класів об'єктів. Для рішення цього завдання розроблена математична модель і відповідний алгоритм процедури розбивки, який дозволяє, виходячи з розмірності простору ознак об'єкта і їх змісту, виділити окремі класи об'єктів у предметній області. Після встановлення всіх класів у предметній області розглядається завдання 2, де розглядається побудову математичної моделі й алгоритму розпізнавання екземпляра об'єкта з погляду приналежності його до якого- те класу об'єктів предметної області. Це завдання зважується на основі Баєсівського підходу, тобто з використанням формул повної ймовірності й Баєса. Для роз'яснення процедур класифікації й розпізнавання в роботі приводиться приклад використання запропонованих алгоритмів.

Як приклад докладно розглядається область комунікаційних завдань.

Запропоновано: U множина образів комунікаційних завдань із загальними ознаками X =$\{ x_1, \ldots, x_ö\}$, *m* – кількість ознак предметної області (на приклад, m =1,…, 18).

Образ – це опис об'єкта або процесу, що дозволяє виділяти його з навколишнього середовища й групувати з іншими об'єктами або процесами для прийняття необхідних рішень [122] (у розглянутій предметній області образ «послуга» або «завдання»). Цю множину необхідно розділити на підмножини ознак, відповідні до класів запитів завдань даної предметної області. За допомогою списку правил, система становить словник ознак [123]. Цю множину буде позначено через *X*. Поки не зроблено перший запит у систему, існує тільки один загальний клас ω область переваги якого є *X*. У множині образів комунікаційних завдань предметної області *U* буде цікавити деякі підмножини – класи залежно від типів комунікаційних послуг. Ті категорії об'єктів, які необхідно виділити або розділити всю множину образів у процесі розпізнавання, звичайно називають класами.

Множина класів $\Omega = \{\omega_1,\ldots, \omega_m\}$ є кінцевою (у даному завданні розпізнавання m =1,…, 4, і дорівнює числу класів послуг: ω_1 – клас послуг інтернету, ω_2 – клас послуг тарифів, ω_3 – клас послуг оплати, ω_4 – клас послуг підтримок), і класи утворюють повну групу підмножин з U (поділ множини

образів U), тобто $U^{m}\ \omega_i = U$, та $\omega_i \cap \omega_j = \varnothing$ для усіх $i \neq j$. Класифікувати об'єкт $x \in U$ по класах ωі, значить знайти таку індикаторну функцію g: U → Y, де Y={y_i}, а y_i – ознаки об'єкта, яка ставить у відповідність образу $x \in U$ мітку $y_i \in Y$ того класу ω_i, якому він належить тобто g(x) = y_i, якщо $x \in \omega_i$).

Система автоматично створює класи за допомогою вирішальних функцій при розпізнаванні вступників на систему невідомих образів. У системі кількість класів збільшується з поділом на підмножини класів залежно від, послуг що з'являються, і ці класи можуть перетинатися між собою володіючи загальними ознаками.

Нехай у системі з'явилася неідентифікована послуга *x*, для якої по своїх ознаках, що описують, x = (x_1, …, x_n), n ≤ m, де *n* − кількість ознак завдання й необхідно створити новий клас із множини ознак предметної області X={X_1, X_2}. Ця послуга описана деякими ознаками в даній предметній області.

Ці ознаки рівняються з ознаками предметної області на предмет збігу. Якщо всі ці ознаки повністю збігаються або з великою ймовірністю збігаються з ознаками предметної області $p(x_j \in \omega_i) > p(x_j \in \omega_r) \quad \forall\ i \neq r$, те новий клас не створюється.

Якщо більшість ознак збігаються, але не всі ознаки, то в предметній області вводяться нові ознаки й створюється новий клас.

У випадку, коли з'являється нове завдання й у предметній області вже існують деякі класи $\omega = U^{m}\ \omega_i$, тj дана послуга перевіряється на приналежність якомусь із існуючих класів, якщо дані ознаки збігаються з ознаками деякого класу, те дане завдання ставиться до даного класу. Але якщо ознаки, що описують дане завдання не збігаються або ймовірність збігу $p(x_j \in \omega_i) = 0$, та перевіряється на збіг із загальними ознаками предметної області, щоб створити новий клас у даній предметній області.

На рис. 3.3 блок С – це множина ознак X_i, $i = 1, \ldots, 4$.

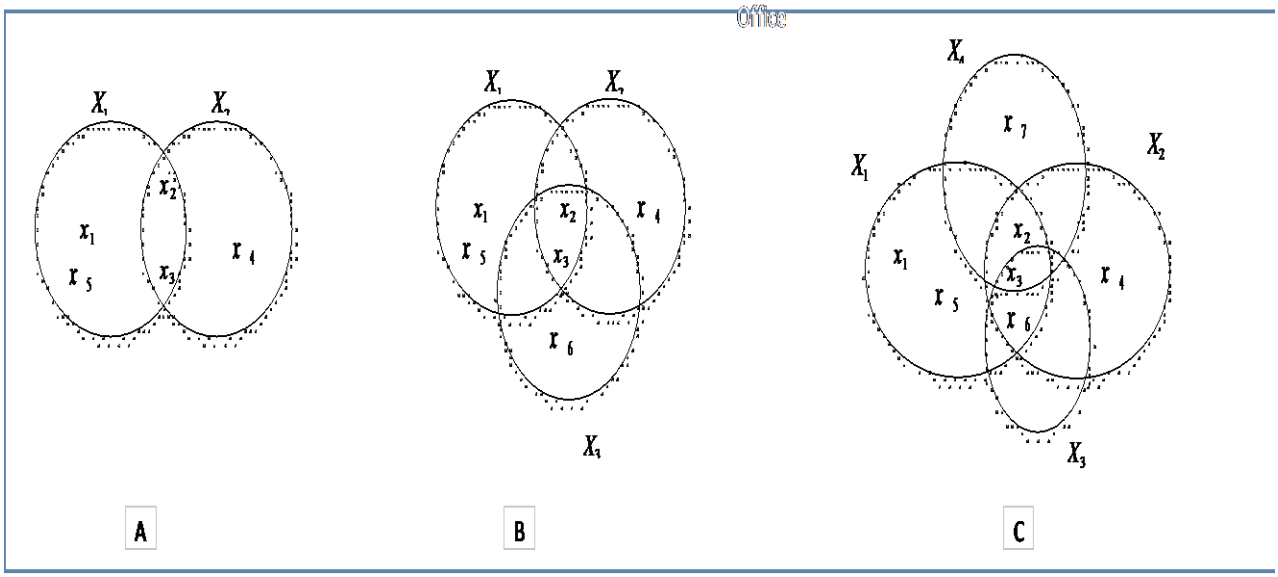

Рисунок 3.3 – Поділ множини ознак на підмножини класів

Тоді

$$X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7\}$$
$$X_1 = \{x_1, x_2, x_3, x_5, x_6\}$$
$$X_2 = \{x_2, x_3, x_4, x_6\}$$
$$X_3 = \{x_3, x_6\}$$
$$X_4 = \{x_2, x_3, x_7\}$$

Предикати:

$$P(x_1 \in \omega_i): i = \overline{1,4} => 1; 0; 0; 0$$
$$P(x_2 \in \omega_i): i = \overline{1,4} => 1; 1; 0; 1$$
$$P(x_3 \in \omega_i): i = \overline{1,4} => 1; 1; 1; 1$$
$$P(x_4 \in \omega_i): i = \overline{1,4} => 0; 1; 0; 0$$
$$P(x_5 \in \omega_i): i = \overline{1,4} => 1; 0; 0; 0$$
$$P(x_6 \in \omega_i): i = \overline{1,4} => 1; 1; 1; 0$$
$$P(x_7 \in \omega_i): i = \overline{1,4} => 0; 0; 0; 1$$

Тоді вирішувальна функція буде мати вигляд

$$d(x) = d\left(\frac{x_1}{7}, \frac{x_2}{7}, \dots, \frac{x_n}{7}\right)$$
$$d(x) = \frac{x_1}{7}, + \frac{x_2}{7} + \cdots + \frac{x_n}{7};$$

$$d_1(x) = \left(\frac{1}{7}x_1 + \frac{1}{7}x_2 + \frac{1}{7}x_3 + \frac{1}{7}x_5 + \frac{1}{7}x_6\right) = \frac{5}{7};$$

$$(x_1, x_2, x_3, x_4, x_5, x_6, x_7) \in \omega_2$$

$$d_2(x) = \left(\frac{1}{7}x_2 + \frac{1}{7}x_3 + \frac{1}{7}x_4 + \frac{1}{7}x_6\right) = \frac{4}{7};$$

$$(x_1, x_2, x_3, x_4, x_5, x_6, x_7) \in \omega_3$$

$$d_3(x) = \left(\frac{1}{7}x_3 + \frac{1}{7}x_6\right) = \frac{2}{7};$$

$$(x_1, x_2, x_3, x_4, x_5, x_6, x_7) \in \omega_4$$

$$d_4(x) = \left(\frac{1}{7}x_2 + \frac{1}{7}x_3 + \frac{1}{7}x_7\right) = \frac{3}{7};$$

тобто якщо *di(x) > dj(x)*, то $x \in \omega_i$.

Повна вірогідність того, що настане подія A(P(A)) обчисляється за формулою:

$$P(A) = \sum_{i=1}^{N} P(H_i)P(A|H_i).$$

Графік таких ймовірностей надано на рис. 3.4

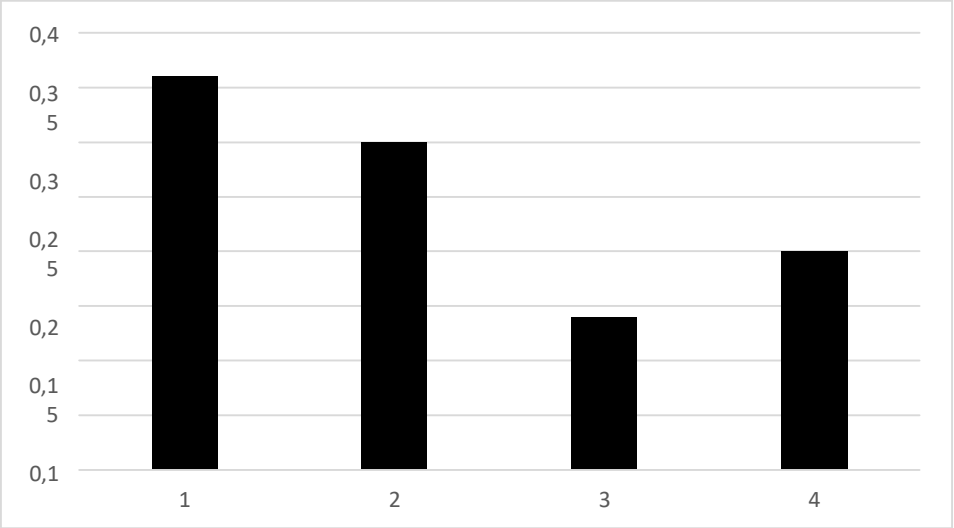

Рисунок 3.4 – Графік ймовірностей приналежності об'єкта *i*-класу ω

Значення ймовірностей, одержувані по формулі Баєса, надають черговість для правильної класифікації, вказуючи варіанти правильної класифікації один по одному убування їх значень, тобто перший можливий варіант буде той варіант, для якого значення ймовірності більше, для того що б скоротити повний перебір варіантів класифікації як на рис. 3.5.

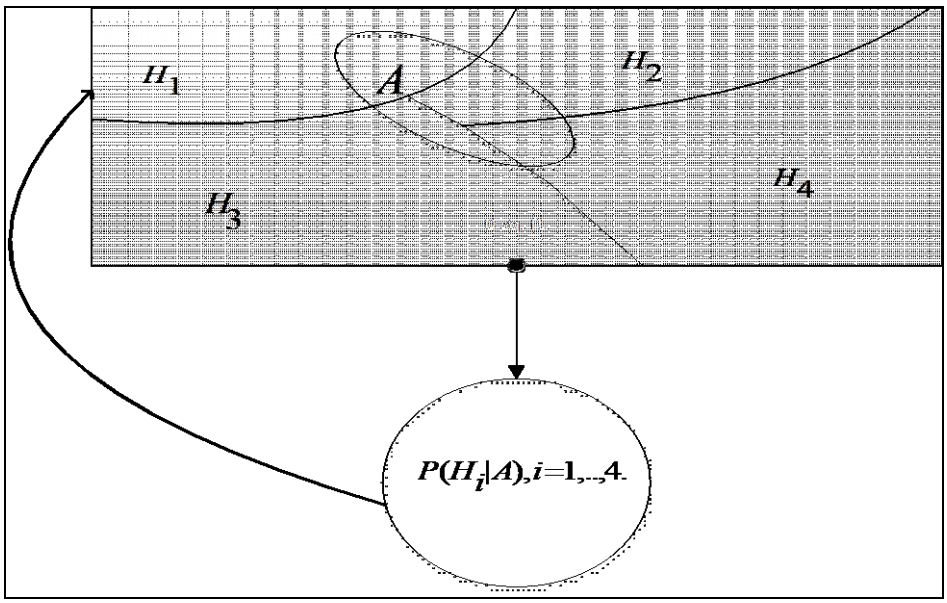

Рисунок 3.5 – Позначення події *A* при виконанні події H_i, *i* =1,..,4

3.4 Модуль побудови й навчання нейронної мережі

В якості прикладу, що дозволяє перевірити теоретичні розрахунки, розглянуті проблеми розпізнавання деяких об'єктів по конкретних завданням сфери комунікаційних послуг, які надходять із мережі на вхід класифікатора. Кожний об'єкт *x* характеризується множиною ознак, тобто для кожного об'єкта будується інформаційний вектор $x = (x_1, \ldots, x_n)$, де через *n* позначена кількість ознак. Цей вектор ознак по логіці предикатів перетвориться у вектор з булевими значеннями 0 або 1 залежно від приналежності його до будь-якого класу ω_i, i = 1,…, N, де *N* − кількість класів у предметній області.

Екземпляр об'єкта може мати загальні ознаки в багатьох класах, тобто згідно з ознаками, може частково відноситися до того або іншого класу, при цьому з'являється необхідність оцінити ймовірність приналежності об'єкта до певного класу. З використанням формули Баєса завдання може бути вирішене і це дозволяє значно скоротити перебір класів при навчанні самої системи (класифікатора). Таким чином розглядається завдання навчання системи безпомилкової класифікації об'єктів за допомогою перцептрона.

Для представленої моделі класифікації об'єктів була розроблена сукупність алгоритмів, що включає в себе:

– алгоритм побудови штучних нейронних мереж, за допомогою якої виконується автоматичне створення нових нейронних мереж залежно від вихідних параметрів.

– алгоритм навчання штучних нейронних мереж, яким завершується побудова нейронних мереж.

Кількість нейронів у вхідному шарі визначається залежно від розмірності матриці X^1 наприклад, якщо кількість ознак рівно 10, будується 10 нейронів у вхідному шарі, і кожний нейрон буде мати певні значення ознак. У моделі створення нейронної мережі необхідно задати N − кількість нейронів *i* у вхідному шарі де $10000 \geq N \geq 1$. Залежно від заданого числа будуються нейрони у вхідному шарі. Максимальне число нейронів у вхідному шарі не більше 10000, тому якщо ні, то нейронна мережа буде дуже великою і процес її навчання буде досить складно й довго відбуватися в рамках технічних обладнань, доступних на цей час.

Матриця Y^1 − вихідних сигналів мереж, тобто навчання, отримані на виході мережі для *i*-ї навчальної вибірки.

$$Y^1 = \begin{bmatrix} y^1_{11} & y^1_{12} \\ y^1_{21} & y^1_{22} \\ \cdots & \cdots \\ y^1_{N1} & y^1_{N2} \end{bmatrix}$$

Для кожного a_i *i*-го об'єкта в навчальної вибірки ставиться у відповідність певний клас ω_i, до якого об'єкт повинен належати, тобто $a_i \in \omega_i$. Як тільки закінчується процес побудови нейронної мережі, в алгоритмі автоматично задається W^1 − матриця випадкових значень початкових ваг нейронної мережі.

$$W^1 = \begin{bmatrix} w^1_{11} & w^1_{12} \\ w^1_{21} & w^1_{22} \\ \cdots & \cdots \\ w^1_{N1} & w^1_{N2} \end{bmatrix}$$

Кількість ітерацій обмежується залежно від значення E середньоквадратичної помилки. Якщо E = 0, або E ≤ ∂ де ∂ − припустима максимальна середньоквадратична помилка. Параметр 6. □ − припустима максимальна середньоквадратична помилка.

Приклад створення нейронної мережі, яка буде навчатися даними в наступній таблиці матриці значень істинності, називану «ексклюзивним» або «XOR» (або 1, або 0, але не обидва):

Щоб розв'язати проблему, потрібно ввести новий шар у нейронну мережу – це шар, який називається «схованим шаром», та дозволяє мережі створювати й підтримувати внутрішні представлення введених даних на вхід мережі. Мережа з одним схованим шаром, яка буде відображати таблицю істинності XOR (див. рис. 3.6).

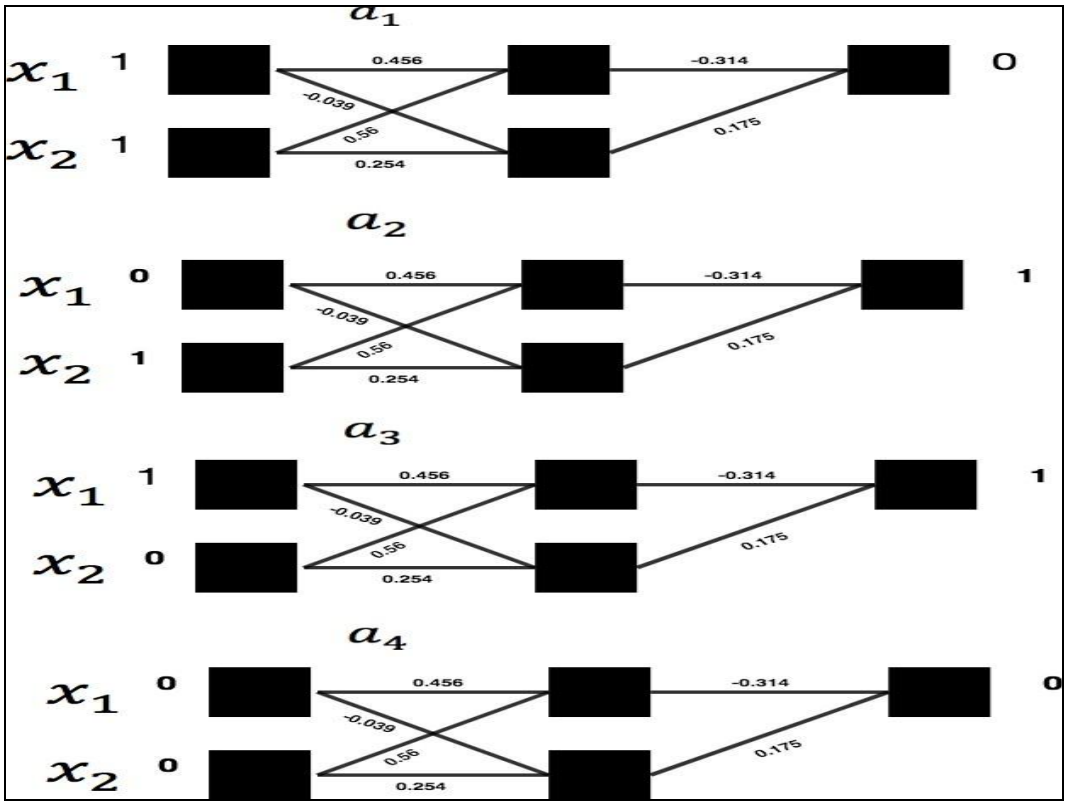

Рисунок 3.6 – Графічні результати побудови нейронної мережі XOR зі значенням даних між нейронами

Вектор отриманих значень даних у результаті навчання створеної мережі буде Y = (0.456, −0.039, 0.56, 0.254, −0.314, 0.175).

Розрахунок показує що за 138 ітерації – досягається E = 0.00003832.

Це задовольняє умові $E \leq \partial$, $\partial = 0.0001$. Де ∂ − припустима максимальна середньоквадратична помилка при навчанні нейронної мережі та завдається користувачем при вводі параметрів.

Ці значення використовуються для правильної класифікації нових об'єктів що подаються на вхід мережі.

4 ОПИС ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

4.1 Розробка програмного забезпечення

Розроблено програмне забезпечення для класифікації комунікаційних завдань. У комунікаційну систему приходять різні запити від різних користувачів через інтернет або стільникові мобільні телефони. Класифікація завдань по їхніх класах і вибори методів для виконання таких завдань є дуже важливим питанням для забезпечення швидкої й високоефективної роботи. Для цього потрібно створити програмне забезпечення самонавчальної системи, що дозволяє розпізнавати завдання, що надходять у комунікаційну систему й методи їх виконання.

Комплекс прикладних програм написаний мовами Java і Javascript з використанням об'єктно-орієнтованого підходу. Програми класифікації об'єктів виконані в середовищі Netbeans. У розроблене програмне забезпечення впроваджено й застосовано вищеописані математичні моделі й алгоритми за допомогою мови програмування високого рівня Java. Також були використані інші мови програмування й бібліотеки, такі як Javascript, Ajax, Jquery. Комплекс програм дозволяє розглядати ефективність розроблених алгоритмів, а також перевірити й оцінювати їхні результати.

Однією з основних переваг створених математичних моделей і алгоритмів є їхня ефективність в аналізі даних великої розмірності. Ефективність розроблених алгоритмів підтверджується тим, що обсяг цифрових даних досить великий. Для прикладу розглядається рисунок 4.1, який показує кількість надходжень і завдань, що класифікуються у систему комунікаційних завдань упродовж місяця, де кожен стовпець на рис. 4.1 є одним кластером. За допомогою розробленого програмного забезпечення є можливість вибрати певний час для одержання необхідного звіту про кількість об'єктів у кластерах системи.

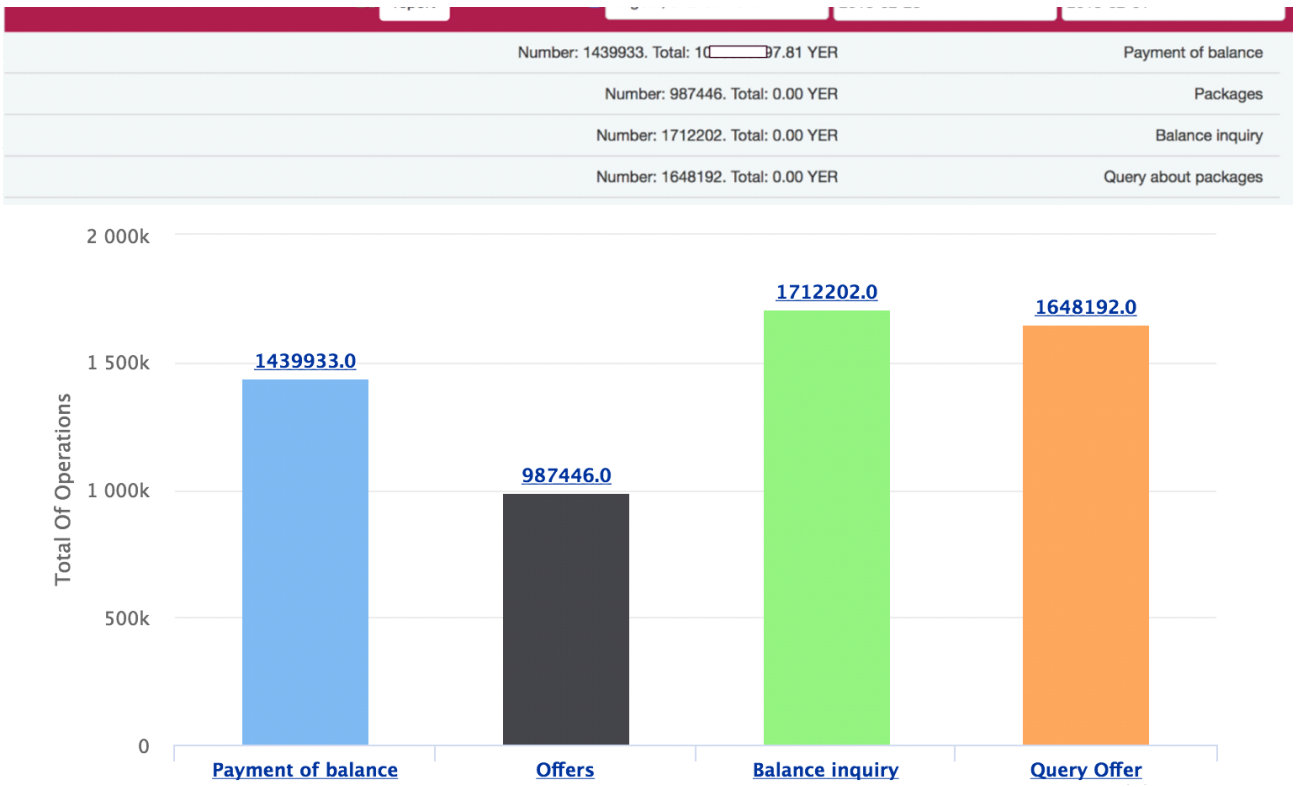

Рисунок 4.1 – Діаграма кількості завдань у 4-х комунікаційних кластерах

Як відзначено у прикладі, існує 4 класи комунікаційних послуг (Paymentofbalance, Offers, Balanceinquire, Queryoffer), тобто множина класів $M = \{ \omega_1 , \omega_2 , \omega_3 , \omega_4 \}$. Кожен клас має свої ознаки. Деякі з них є загальними, тобто класи можуть перетинатися один з одним .

У комплексі програм було розроблено інтерфейс, через який експерт може налаштувати систему. Досить важливо в конфігурації системи задати основні характеристики (ключові слова) предметної області. На основі даних характеристик система буде навчатися, тобто створювати навчальні вибірки (розпізнавані образи) і утворювати необхідні класи.

Зазвичай процес визначення характеристик даної предметної області відбувається автоматично в ході роботи системи за допомогою розпізнаваних вхідних образів. Необхідно задати кілька множин характеристик об'єктів предметної області для початку процесу навчання системи (див. рис. 4.2). Усі характеристики додаються й зберігаються в базі даних з можливістю їх редагування й видалення з бази при необхідності.

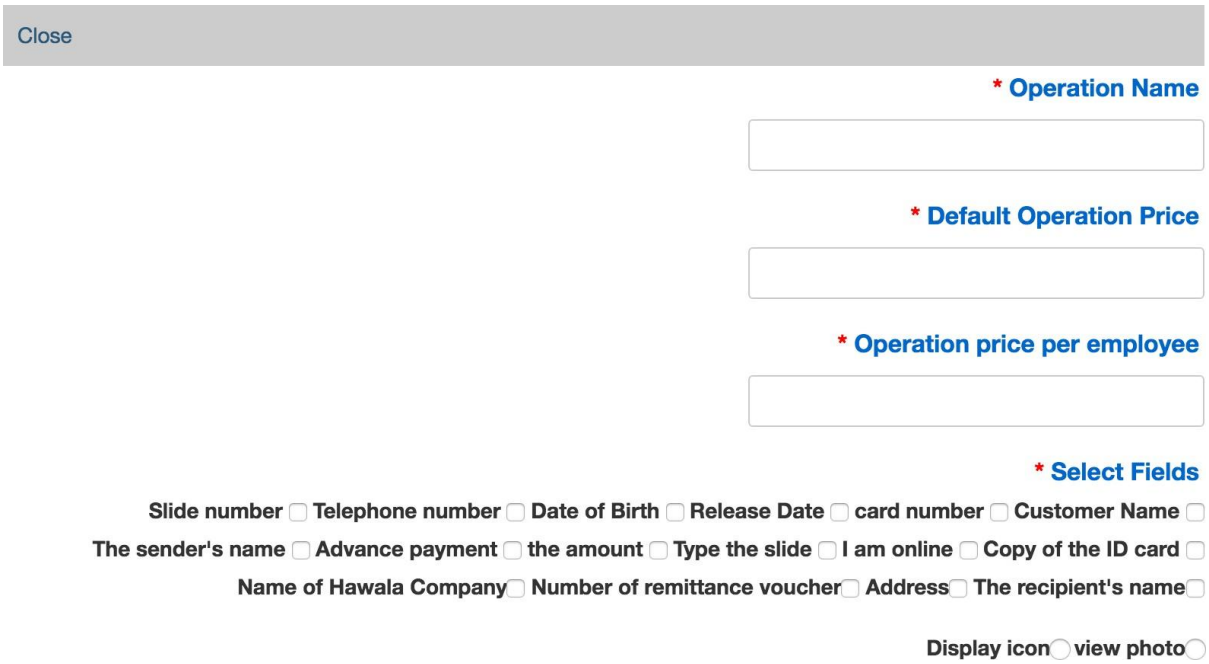

Рисунок 4.2 – Інтерфейс створення ознак, що описують предметну область системи

Після визначення ознак об'єкти, що розпізнані, надходять на вхід системи. Кожен об'єкт розглядається окремо й виділяється своїми ознаками. Якщо у вхідному об'єкті з'явилися нові характеристики, то вони виділяються й зберігаються в базі даних.

Number: 19

More	The default operation price	Operation name	id
More	0	MTN Packages	
More	0	Receipt of a transfer and its addition to insurance	
More	0	Send a transfer	
More	0	Wholesale Instant Shipping	
More	0	Send a photo notification of insurance transfer	
More	0	Activate a bouquet of balance	
More	0	(Package with extra balance (package price + balance	
More	0	Immediate charging all networks	
More	0	Fixed telephone payment	
More	0	Paying Internet ADSL	
More	0	Raise your billing numbers	
More	2100	Transfer from prepay to billing	
More	1600	New number programming bills	

Рисунок 4.3 – Список утворених класів.

Новий клас зберігається в базі даних під найменуванням категорії завдань, що входять у нього. Для експерта найменування завжди доступне для редагування, а також для інших дій. Кожен клас (категорія) зі списку класів має свої власні ознаки в тому числі й загальні ознаки, які можуть бути в списку ознак інших класів. Відзначимо, що процес класифікації в системі відбувається як «без учителя», так і «з учителем» залежно від установлених налаштувань системи, тому при класифікації з учителем існує можливість корегувати ознаки в кожному кластері. Експерт одержує можливість корегування ознак, що описують створені класи, після надання доступу адміністратором системи.

При надходженні нових об'єктів у класифікатори системи, кожен об'єкт ставиться до того або іншому класу з існуючих утворених класів. Якщо об'єкт жодному класу не належить, то він буде представлятися як окремий кластер.

Кожен об'єкт у базі даних представляється як множина ознак, властивостей, що його характеризують. Інформація про об'єкт зберігається у форматі Json Data. Кожен атрибут має своє значення яке може бути кількісним або якісним (див. рис. 4.4). Для перегляду був розроблений інтерфейс, за допомогою якого можна переглядати всю інформацію про об'єкти.

Export to Excel

No. 4139613

Preparation	Case	Type	Date	Customer	Customer Number	the amount	phone number	Display	Operation type	operation number
	Success	unknown	2018-04-06 23:58:33.0	Abdul Malik al-Sharabi / The Old System	181	0.00	770723952	New / Package 200 MB - Prepaid SIM card	Packages	6142174
	Success	pre paid	2018-04-06 23:58:30.0	Abdul Malik al-Sharabi / The Old System	181	820.00	770723952		Payment of balance	6142173
	Success	unknown	2018-04-06 23:58:26.0	Abdul Malik al-Sharabi / The Old System	181	0.00	770723952	Remove / Package 200 MB - Prepayment Slice	Packages	6142172
	Success	unknown	2018-04-06 23:58:24.0	Abdul Malik al-Sharabi / The Old System	181	0.00	770723952		Query about packages	6142171
	Success	pre paid	2018-04-06 23:58:14.0	Abdul Malik al-Sharabi / The Old System	181	0.00	770723952		Balance inquiry	6142170

Рисунок 4.4 – Список класифікованих об'єктів у базі даних

У системі об'єкти можуть мати декілька ознак, їх кількість не обмежена й відрізняється кількістю ознак у кожному об'єкті. За допомогою розробленого інтерфейсу користувач може розглядати всі значення ознак даного об'єкта

натисканням кнопки «Перегляд» (див. рис. 4.5). Кожна ознака визначається ключовим словом і має своє значення.

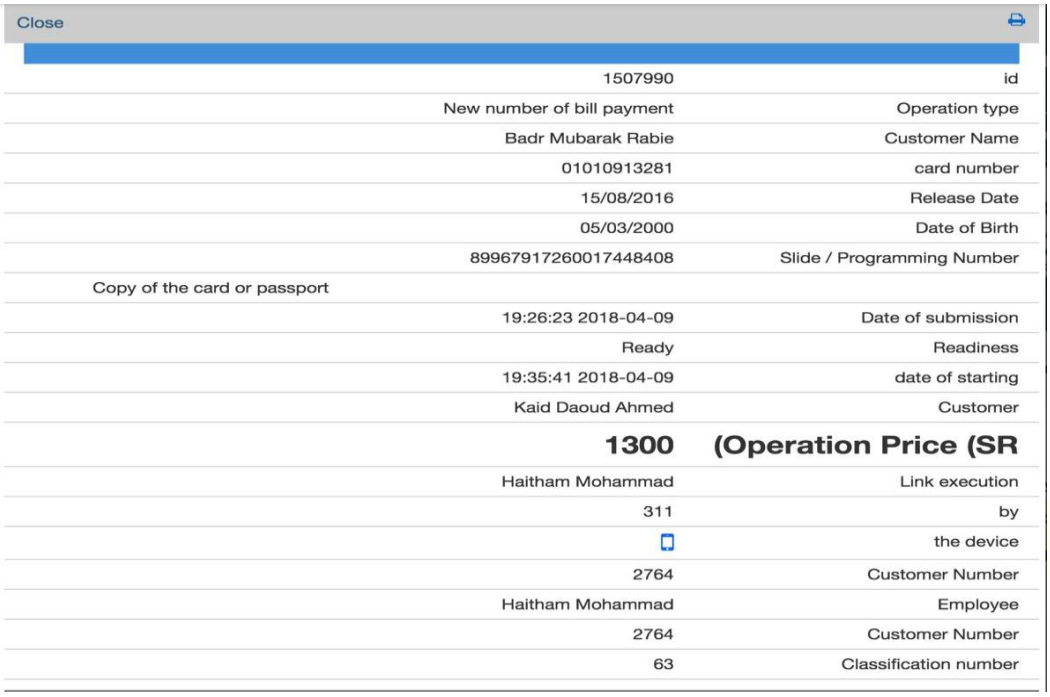

Рисунок 4.5 – Значення ознак, що характеризують об'єкт.

Однією з головних цілей класифікації об'єктів у розробленому комплексі програм є швидкодія, для полегшення й прискорення процесу обробки даних у подальшій роботі системи, тому в комплексі програм було створено модуль пошуку об'єктів за деякими значеннями. Модуль пошуку даних є одним з головних завдань аналізу даних великої розмірності, тому звичайно, в таких модулях використовують методи розпізнавання образів, для того щоб класифікувати об'єкти по певних категоріях. Чим більше ефективність методу розпізнавання образів, тем ефективніше модуль пошуку даних.

У модулі пошуку даних необов'язково задавати всі значення атрибутів. Для виконання процесу пошуку необхідно задати значення мінімум одного атрибута або, як максимум, значення всіх атрибутів. Час для виконання пошуку й вибору підходящих об'єктів залежно від заданих параметрів пошуку малий, при досить великій кількості об'єктів у системі.

Швидкість виконання пошуку також залежить від властивостей обчислювальної машини (потужність процесору, оперативна пам'ять і інші).

Для ілюстрації процесу у візуальному вираженні була розроблена імітаційна модель класифікації об'єктів під час надходження об'єктів на вхід системи

У лівій частині рисунка 4.6 показані об'єкти до проходження через класифікатор. Для прикладу обрано 4 класи X = {X_1, X_2, X_3, X_4}, тому що класів у системі може бути більше ніж 4. У правій частині рисунка 4.6 показані об'єкти після класифікації, тобто після віднесення кожного об'єкта певному класу, якому об'єкт належить.

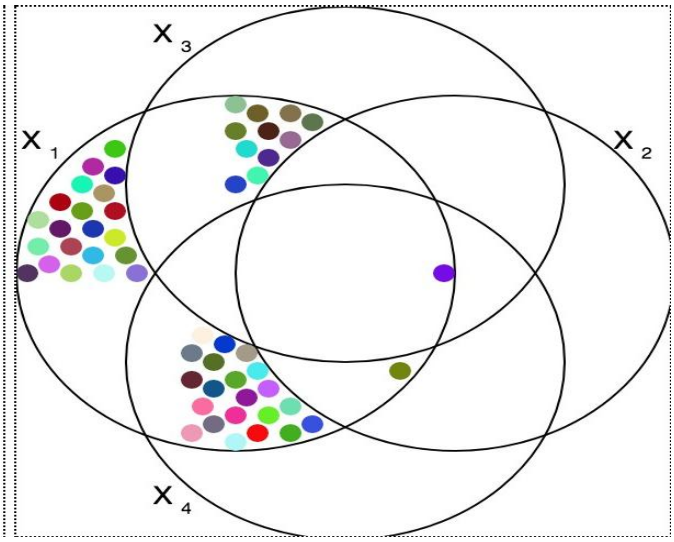

Рисунок 4.6 – Імітаційна модель класифікації об'єктів

Помітно, що існує перетинання між класами, куди попадають об'єкти, у яких є загальні ознаки. Потрапляння об'єкта в ту або іншу область повністю визначається тими ознаками, які він має.

Колір об'єктів на рисунку 4.6 є їхньою назвою, що відрізняє їх один від одного. Конвергенція кольорів не означає, що об'єкти, у яких є схожі кольори сходяться один з одним за ознаками. У даному модулі для визначення приналежності об'єктів до певного класу використаний класифікатор на основі методу Баєса.

На основі отриманої ймовірності в результаті процесу визначення й порівняння ознак об'єктів з ознаками класами вибирається клас із максимальною ймовірністю.

4.2 Опис програмного забезпечення кластерізації

По заданих ключових критеріях, створюються первинні кластери, і будується кластерне дерево. При надходженні нового об'єкта □, алгоритм розпізнає об'єкт за його ознакою, і видає вектор ознак об'єкта: значення ознак необов'язково повинні бути кількісними.

Алгоритм виділяє некількісні ознаки об'єкта, і по їх значеннях загальний кластер розбивається на підмножини кластерів. Кластери нумеруються за зростанням: кожному кластеру надається свій номер C_n $n = \overline{1, N}$, як ідентифікатор. Кожен кластер у кластерному дереві буде мати свої незалежні лінійні функції з іншими підкластерами, а також окремий шлях до своїх об'єктів.

У моделі використовується формат типу даних JSON, у вигляді ключа й значення кожної ознаки в об'єкті:

```
{«name»:»xxx», «type»:»xxx», «size»:»xxx»,»time»:»xxx»},
```

де ххх – значення ознаки, яке може бути кількісне або якісне.

Також було розроблено програмний інтерфейс для застосування алгоритму динамічної кластеризації в завданнях розпізнавання образів.

Пункт 1, вводиться назва предметної області, у якій застосовується алгоритм кластеризації (у даному прикладі <<Список студентів>>).

Пункт 2, назва осі координат x, у даному прикладі буква *y*, яка позначає кількість елементів кластерів.

Пункт 3, назва осі координат *y*, в даному прикладі буква V, яка позначає групи.

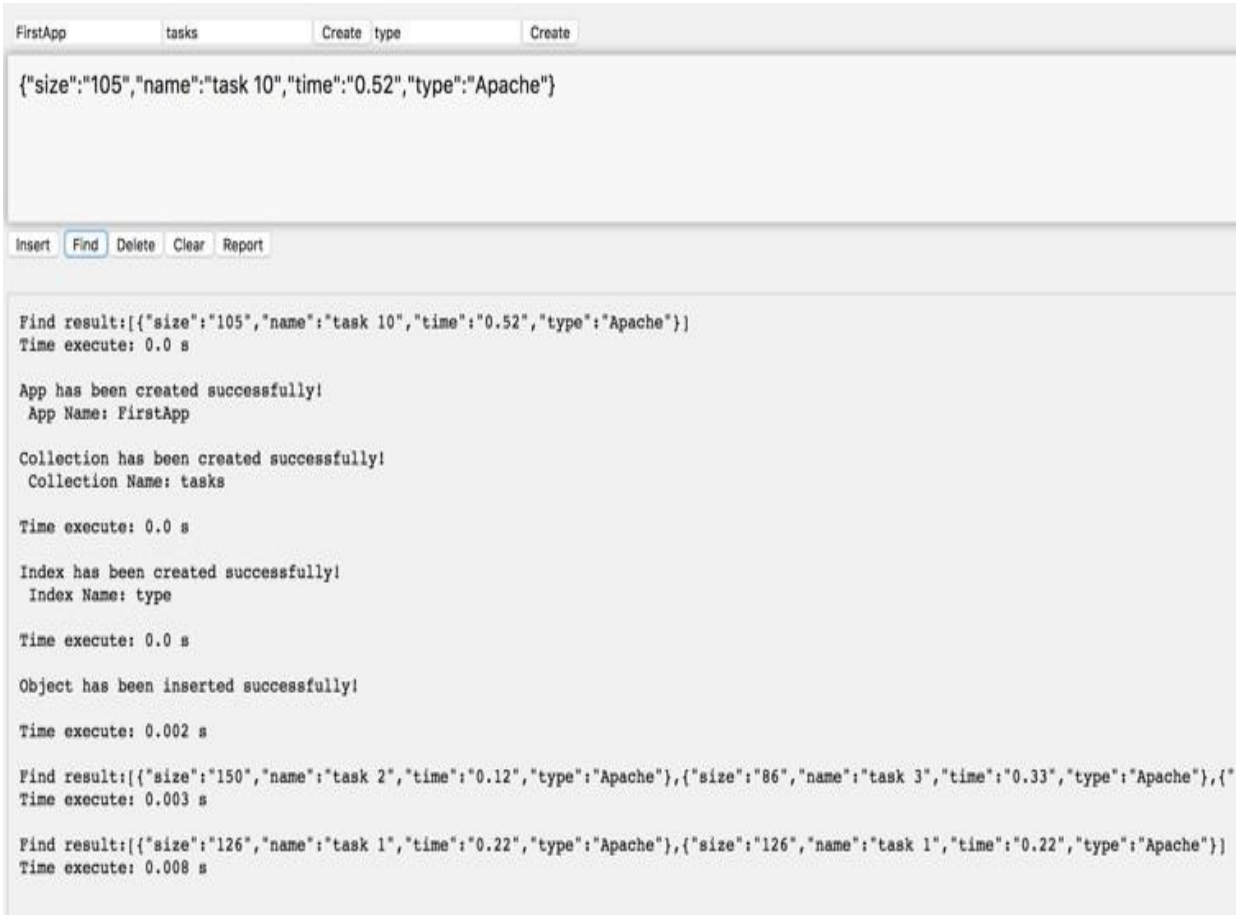

Рисунок 4.8 –API моделі класифікації об'єктів

Пункт 4, набір даних навчальної вибірки. Дані навчальної вибірки вводяться у форматі типу даних JSON. Кожен об'єкт представляється набором ознак з їхніми значеннями.

Пункт 5, вводиться початкові границі кластерів. Можна не вводити усі границі кластерів навчальної вибірки, тому що якщо не усі границі кластерів задані, то алгоритм автоматично будує нові кластери й визначає границі для нових створених кластерів самостійно. Якщо значення границь кластерів не задані, алгоритм автоматично задає початкові границі залежно від розміру навчальної вибірки.

5 ОПИС МОЖЛИВОСТІ ВИКОРИСТАННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

«Кластер» у рамках розроблених алгоритмів являє собою сукупність одного або декількох вузлів (серверів), які разом зв'язуються для зберігання даних і забезпечення інтегрованих функції індексування й пошуку по всіх вузлах, тобто $C = \{c_1, \ldots, c_m\}$, $_m$ – кількість вузлів у кластері, кластер позначати великою буквою *C*. Кластер ідентифікується унікальним іменем, яке за замовчуванням є «`FirstApp`». Це ім'я важливе, тому що вузол може бути тільки частиною кластера, тому що вузол налаштований на об'єднання кластерів по імені.

«Вузол» – це один окремий сервер, який є частиною створеного кластера, зберігає дані й бере участь у параметрах індексування й пошуку кластера. Вузол буде позначатися буквою *c*. Точно так само, як і кластер, вузол ідентифікується ім’ям, яке за замовчуванням є довільним універсальним унікальним ідентифікатором, який призначається вузлу при його створенні. Користувач може призначити вузлу будь-яке ім'я, але якщо воно не призначене, то система дасть найменування за замовчуванням «`FirstNode`». Ім'я важливе для подальшої роботи з даними з метою адміністрування й керування. Вузол може бути налаштований на приєднання певного кластера по імені кластера. За замовчуванням кожний вузол налаштований на об'єднання кластера з іменем «`FirstApp`». В одному кластері може бути стільки вузлів, скільки користувач забажає, без обмеження.

«Індекс» являє собою набір документів, які мають схожі характеристики. Наприклад, перший індекс для даних клієнта, другий індекс для каталогу продуктів і третій індекс для даних замовлення. Індекс ідентифікується по імені, і це ім'я використовується для посилання на індекс при виконанні операцій обробки даних індексування, пошуку, відновлення й видалення документів у ньому. Позначено індекс буквою *e*.

«Документ» являє собою базову одиницю інформації, яка може бути проіндексована. Наприклад, можна мати документ для одного клієнта, інший

документ для одного продукту, а третій – для одного замовлення. Цей документ виражається в JSON (JavaScriptObjectNotation), який є розповсюдженим форматом обміну інтернет-даними. Будемо позначати документ буквою *x*, де c = {x_1, …, x_n}, n = 1,…N. Документи це об'єкти, які необхідно класифікувати по класах схожих характеристик.

Для використання розробленої моделі в інших системах необхідно встановити Java на ПК, що використовується. Програмний код являє собою компілятор мовою Java і інтегрується з іншими системами за допомогою додавання *.jar файл у програмі користувача.

Після встановлення модель надає повний і потужний API, який можна використовувати для взаємодії зі створеними кластерами. Серед дій, які можна виконати за допомогою API, доцільно виділити наступні:

– визначення стану, статусу й статистики кластерів, вузлів і індексів;

– адміністрування кластерів, вузлів, індексів і даних;

– виконання обробки даних (створювати, читати, обновляти й видаляти) і операції пошуку в створених індексах.

Команда для створення нового кластеру з іменем «University».

```
POST /network/cluster1
```

Команда для створення нового індексу з іменем «Students».

```
POST /network/cluster1/Students
```

Додаються нові документи до створеного індексу «Students». Індексується простий студентський документ в індекс Students з ідентифікатором 1 у такий спосіб:

```
POST /network/cluster1/Students/1
{«name»:»John», «course»:2, «age»:22, «faculty»: «Informaticsandappliedmathematics»}
```

Для знаходження документа за допомогою розробленого алгоритму для пошуку, наприклад, витяг документу, який щойно про індексували (див. рис. 5.1):

```
за ідентифікатором
GET /network/cluster1/Students/1
Результат:
{«name»:«John», «course»:2, «age»:22, «faculty»:
«Informaticsandappliedmathematics»}.
за іншими критеріями:
GET /network/cluster1/Students/name= John
Результат:
{«name»: «John», «course»:2, «age»:22, «faculty»:
«Informaticsandappliedmathematics»}.
```

Рисунок 5.1 – Фрагмент програмного коду

За допомогою кнопки «Report» можна отримати звіт про кількість об'єктів у базі (див. рис. 5.2).

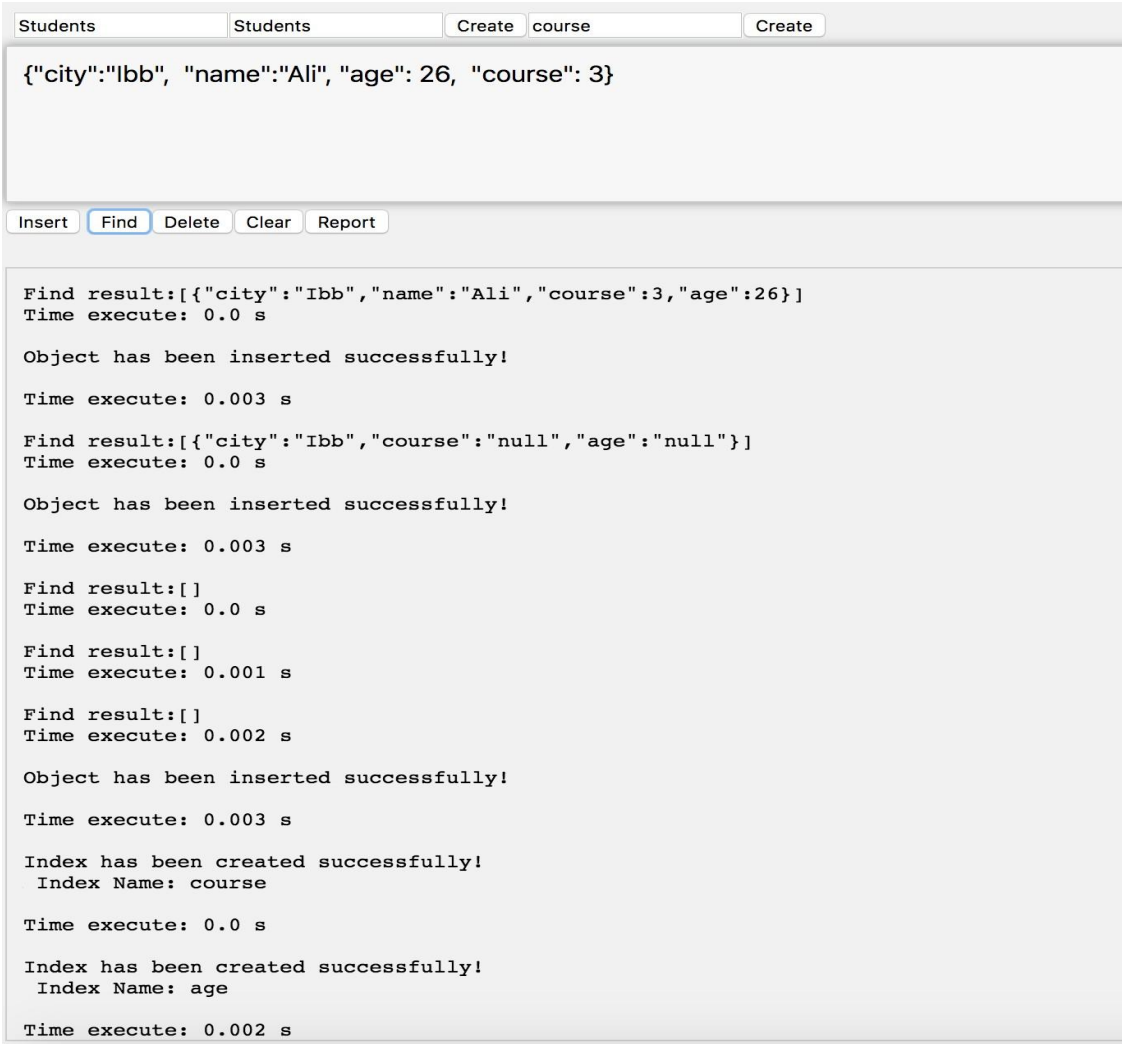

Рисунок 5.2 – Інтерфейс класифікації об'єктів

Для реалізації розроблених математичних моделей і алгоритмів було розроблено програмне забезпечення систем класифікацій. У першому лівому полі інтерфейсу, розробленого програмного забезпечення, користувач уводить назву

будь-якого проекту, залежно від конкретної предметної області. Після введення назви необхідно натиснути на кнопку «Create».

Після створення проекту створюються атрибути ознак, тобто ключові слова. Кнопка «Insert» призначена для введення даних у класифікатори системи, які зберігаються в базі після проходження процесу класифікації.

Також інтерфейс представляє можливість пошуку й видалення даних за допомогою кнопок «Find» і «Delete».

ВИСНОВКИ

Кластери слід розглядати як динамічні утвори, що змінюються під впливом потоку розпізнаваних образів. Розроблений алгоритм кластерізації працездатний у просторі більших даних.

Запропоновано математичні моделі для рішення завдань розпізнавання образів за допомогою, теорії кластерної системи, що само організується.

За допомогою апарата аналізу часових рядів потрібно дослідити процес еволюційних змін кластерних утворів.

На основі імовірнісного підходу побудовані модель і алгоритм обчислювальної процедури, що дозволяє ідентифікувати приналежність об'єкта до певного класу предметної області.

Розроблений і апробований алгоритм динамічної кластерізації, що діє на заданій множині ознак, що характеризують, певної предметної області.

Розроблений спеціалізований програмний комплекс для практичної реалізації алгоритмів.

Розроблено математичне й програмне забезпечення на основі запропонованої моделі динамічної кластерізації для класифікації запитів-завдань комунікаційних послуг, придатне для практичного використання в режимі он-лайн.

Розглядається хід еволюційних змін трьох кластерних сфероїдів, заданих в двовимірному просторі ознак.

Проведений порівняльний аналіз методів, що використовуються для рішення завдань розпізнавання образів, за результатами якої визначені області їх ефективного використання. У цілому за результатами експертизи зроблені наступні висновки:

Найважливішими якостями алгоритмів розпізнавання образів, у баченні експертного співтовариства, є здатність до саморозвитку й самоорганізації, що

має на увазі можливість їх використання в багато-стадійних динамічних процедурах.

Застосування тих або інших методів для вирішення завдань розпізнавання в значній мірі залежить від розмірності простору ознак, що класифікують.

Актуальної й затребуваної є проблема розробки алгоритму з можливостями саморозвитку й самоорганізації, який би міг ефективно використовуватися для рішення завдань розпізнавання будь-якої розмірності.

Показано, що кластери є динамічними утвореннями, що змінюються під впливом потоку розпізнаваних образів. Це явище було показано коли формування кластера на базі вихідної навчальної вибірки завершується й система починає працювати в режимі розпізнавання образів, обробляючи потік об'єктів, що надходять до неї.

Результати наведених прикладів указують на наявність в запропонованому алгоритмі властивостей самоорганізації й еволюційних змін у породжених ним кластерних утворень.

Розроблено алгоритм обчислювальної процедури, які дозволяють класифікувати об'єкти в деякій предметній області, по їхнім властивостям (ознакам), за умови, що частина ознак різних об'єктів збігаються. Це алгоритм дозволяв ранжувати класи по ймовірності приналежності їм класифікуємого об'єкта й указати кращу черговість перевірки об'єктів на предмет приналежності їх тому чи іншому класу. Даний алгоритм особливо ефективний при обробці великих потоків запитів, оскільки, як правило, знімає необхідність повного перебору й скорочує тривалість процедури розпізнавання.

Для практичної реалізації алгоритмів розроблений програмний комплекс. У розроблених системах наочно виявилася ефективність розроблених математичних моделей і алгоритмів, як засобу обробки великого обсягу даних у багатомірному просторі ознак.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Методы интеллектуального анализа данных / IBM URL: https://www.ibm.com/developerworks/ru/library/ba-data-mining-techniques

2. Methods of multidimensional classification in problems of linguistic localization / Shubin I., Kozyriev A. // Proceedings of the III International Conference "Innovative Technologies in Science and Education". November 14, 2019 in Amsterdam, The Netherlands, 2019. pp 398-402

3. Chetverikov G., Puzik O., Vechirska I. Multiple-valued structures of intellectual systems //Proceedings of the with Internations Computer Sciences and Information Technologies (CSIT). 2016, 7589907. -pp. 204-207

4. Решения для интеллектуального анализа данных [Электронный ресурс] / Microsoft URL: https://msdn.microsoft.com/ru-RU/library/ms174861(v=sql.120).aspx –

5. Чубукова И.А. Data Mining: БИНОМ.:/ И.А. Чубукова. – М.: Лаборатория знаний, 2008 – 382 с.

6. Прикладная статистика: Классификация и снижение размерности: / С.А. Айвазян, В.М. Бухштабер, И.С. Юнюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1989. – 383 с.

7. Кречетов, Н.В. Продукты для интеллектуального анализа данных: / Н.В. Кречетов. – М : Рынок программных средств, 2007. – 237 с.

8. Android Interfaces and Architecture Android Developers –: URL: https://source.android.com/devices/

9. Data Mining – интеллектуальный анализ данных / Site of Information Technologies: URL: http://www.inftech.webservis.ru/it/database/datamining/ar2.html

10. Киселев М. В. Средства добычи знаний в бизнесе и финансах:/ М. В. Киселев, Е.Г Соломатин. – М.: Открытые системы, 2012 – 285 с.

11. Гиг Дж., Прикладная общая теория систем:/ Дж. Гиг. – М.: Мир, 2001 – 336 с.

12. Дюк В.А. Обработка данных на ПК в примерах:/ В.А. Дюк. – СПб: Питер, 1997. – 285 с.

13. Тельнов Ю.Ф. Интеллектуальные информационные системы в экономике:/ Ю.Ф. Тельнов. – М.: СИНТЕГ, 2011 – 306 с.

14. Васильев В.П. Информационно– аналитические системы/ В.П. Васильев. – М.: МЭСИ, 2007– 452 с.

15. Методы и модели анализа данных: OLAP и Data Mining/ А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод. – СПб.:БХВ– Петербург, 2004 – 336 с.

16. Введение в анализ ассоциативных правил / BaseGroup Labs URL: http://www.basegroup.ru/library/analysis/association_rules/intro/

17. Эккель Б. Философия JAVA, Библиотека программиста:/ Б. Эккель. – СПб.: Питер, 2013 г. – 638 с.

18. Ульман Д., Уиндом Д. Введение в системы баз данных:/ Д.Ульман, Д. Уиндом. – К.: КПІ, 2018 – 328 с.

19. Гаврилова Т.А., Хорошевский В.Ф. Базы данных интеллектуальных систем:/ Т.А. Гаврилова, В.Ф. Хорошевский. – СПб.: Питер, 2011 – 384 с.

12. J. Nielsen Top 10 Mistakes in Web Design, 2021, http://www.nngroup.com/articles/top-10-mistakes-web-design/.

21. Slywotzky, K. Weber, Demand: Cracking the Code of What People Really Desire. – New York: Business Plus, 2020.

22. G.G. Chetverikov, I.D. Vechirska, S.S.Tanyanskiy, “The methods of algebra finite predicates in the intellectual system of complex calculations of telecommunication companies,” International Conference Proceedings Crimean Microwave and Telecommunication Technology (CriMiCo), 6959425, 2014, pp.

23. Shubin, I., Snisar, S., Zhyrnov, V., Slavhorodskyi, V.// Practical Application of Formal Representation of Information for Intelligent Radar Systems 2018 International Scientific-Practical Conference on Problems of Infocommunications Science and Technology, PIC S and T 2018 - Proceedings, 2019, pp. 433-436, 8632103

24. L. Ardissono, P. Torasso. Dynamic user modeling in a Web store shell. In

ECAI, 2021. (pp. 621-625).

25. R. Brancato. Web interfaces and audience analysis. 2021, http://www.slideshare.net/RochelleBrancato/web-interfaces-and-audience-analysis

26. Resource Description Framework (RDF) Schema Specification 1.0, W3C Candidate Recommendation // World Wide Web Consortium, - March 2020. http://www.w3.org/TR/2020/Cr-rdf-schema-20000327/.

27. OWL Web Ontology Language // World Wide Web Consortium, - September 2004. http://www.w3.org/2014/OWL/.

28. OWL Web Ontology Language Reference, W3C Recommendation // World Wide Web Consortium, - February 2014. http://www.w3c.org /TR/owl-ref/.

29. DAML+OIL Reference Description // World Wide Web Consortium, - March 2020. http://www.w3.org/TR/daml+oil-reference

30. SWRL: A Semantic Web Rule Language Combining OWL and Ruleml // World Wide Web Consortium, - November 2003. http://www.ruleml.org.

31. Integration of information Systems: Bridging Heterogeneous Databases. / Editing by Amar Gupta, Sloan School of Management, Massachusetts Institute of Technology.-1996.

32. G. Wiederhold. Mediators in the architecture of future information systems// IEEE Computer, -March 2016. - p. 38-49.

33. Gruber T. Towards principles for the design of ontologies used for knowledge sharing// International Journal of Human-Computer Studies, 43: 907-928,- 2014.

34. N. Guarino. Some ontological principles for designing upper level lexical resources. In Proceedings of the First International Conference on Language Resources and Evaluation, Granada, 2008. pp. 3–15