Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)
Кафедра Системотехніки
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Розробка методу класифікації за допомогою дерева рішень на базі теорії приблизних множин
(тема)

Виконав:
студент 2 курсу, групи ІТПм-22-2
Чернишов Д.В.
(прізвище, ініціали)

Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма Інформаційні технології проектування
(повна назва спеціалізації)
Керівник проф. Ситніков Д.Е.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ Гребеннік І.В.
(підпис) (прізвище, ініціали)

2024 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Системотехніки
(повна назва)

Рівень вищої освіти другий (магістерський)

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

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

Освітня програма Інформаційні технології проектування
(повна назва)

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

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Чернишову Дмитру Владиславовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Розробка методу класифікації за допомогою дерева рішень на базі теорії приблизних множин
затверджена наказом університету від 06 жовтня 20 23 р. № 1157Ст
2. Термін подання студентом роботи до екзаменаційної комісії 08 січня 20 24 р.
3. Вихідні дані до роботи науково-технічні публікації та дані різних печатних та інтернет-джерел щодо тематики кваліфікаційної роботи

4. Перелік питань, що потрібно опрацювати в роботі
1. Аналіз сучасних тенденцій в теорії приблизних множин
2. Дослідження принципів роботи дерев рішень
3. Розробка моделі дерева рішень
4. Інтеграція дерев рішень з теорією приблизних множин
5. Експериментальна перевірка розробленої гібридної моделі
6. Аналіз результатів експерименту та формулювання висновків

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри)_ _____
_____
_____
_____
_____
_____
_____
_____
_____
_____

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

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	06.10.2023	виконано
2	Аналіз предметної області та постановка задачі	06.10.2023-01.12.2023	виконано
3	Дослідження метода розв'язання задачі	01.12.2023-15.12.2023	виконано
4	Побудова системи	15.12.2023-25.12.2023	виконано
5	Підготовка пояснювальної записки та презентації	15.12.2023-02.01.2024	виконано
6	Попередній захист	01.01.2024-05.01.2024	виконано
7	Захист перед Екзаменаційною Комісією	09.01.2024	

Дата видачі завдання 16 жовтня 20 23 р.

Студент _____
(підпис)

Керівник роботи ___ _____ проф.. каф. СТ, Ситніков Д. Е.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка охоплює: 56 сторінки, 12 зображень, 39 джерел.

ДЕРЕВА РІШЕНЬ, ТЕОРІЯ ГРУБИХ МНОЖИН, КЛАСИФІКАЦІЯ ДАНИХ, ВИБІР ОЗНАК, ВИЛУЧЕННЯ ПРАВИЛ, МАШИННЕ НАВЧАННЯ, ОБРОБКА ШУМУ, ДОБУВАННЯ ДАНИХ.

Об’єктом дослідження є оптимізація моделей дерев рішень в застосуваннях у інтернет технологіях.

Предметом дослідження є вдосконалена модель дерев рішень, інтегрована з теорією приблизних множин (ТПМ), для більш ефективного вибору ознак шумами.

Метою роботи є синергія інтерпретаційних переваг дерев рішень і ефективності обробки шумів у ТПМ. Головний акцент робиться на вдосконаленні процесу вибору ознак і видобування правил для підвищення точності та надійності класифікації моделі, особливо в шумних середовищах даних.

Методи дослідження охоплюють детальне вивчення теоретичних основ дерев рішень і ТПМ, ретельний аналіз їх комбінованого застосування для поліпшеної класифікації та реалізацію цієї інтегрованої моделі на наборі даних.

В результаті дослідження була розроблена вдосконалена модель дерев рішень, включаючи ТПМ для поліпшеного вибору ознак та стійкості до шумів. Ця модель продемонструвала вищу точність в завданнях класифікації і представляє собою значний крок вперед у застосуванні інтелектуальних технік обробки даних в галузі машинного навчання та інтернет технологій.

ABSTRACT

The explanatory note contains: 56 pages, 12 images, 39 sources.

DECISION TREES, ROUGH SET THEORY, DATA CLASSIFICATION, FEATURE SELECTION, RULE EXTRACTION, MACHINE LEARNING, NOISE HANDLING, DATA MINING.

The object of the research is the optimization of decision tree models in internet technologies applications.

The subject of the research is an improved decision tree model integrated with rough set theory (RST) for more effective feature selection amidst noise.

The purpose of the work is to synergize the interpretational advantages of decision trees with the noise-handling efficiency of RST. The main focus is on refining the feature selection process and rule extraction to enhance the model's accuracy and reliability in classification, especially in noisy data environments.

The research methods encompass a detailed study of the theoretical foundations of decision trees and RST, a thorough analysis of their combined application for improved classification, and the implementation of this integrated model on a dataset.

As a result of the research, an enhanced decision tree model was developed, incorporating RST for improved feature selection and noise resistance. This model demonstrated higher accuracy in classification tasks and represents a significant advancement in the application of intelligent data processing techniques in the field of machine learning and internet technologies.

ЗМІСТ

Перелік умовних позначень, символів, одиниць, скорочень і термінів . 6
Вступ 7
1 Теоретична основа. 8
1.1 Задача класифікації 8
1.1.1 Огляд задачі класифікації в аналізі даних 8
1.1.2 Історія еволюції методів класифікації 10
1.1.3 Важливість та Сфери Застосування Класифікації . . 11
1.2 Основи методів дерев рішень 13
1.2.1 Базові концепти у методах дерев рішень 13
1.2.2 Архітектура та конструкція дерев рішень 16
1.2.3 Ключові алгоритми у конструкції дерев рішень . . . 18
1.2.4 Переваги та обмеження дерев рішень 19
1.3 Розуміння теорії приблизних множин 21
1.3.1 Вступ до Теорії приблизних множин 21
1.3.2 Ключові принципи та концепції в теорії приблизних множин 23
1.3.3 Теорія приблизних множин в аналізі даних 24
1.4 Аналіз предметної області валідаційних датасетів 26
1.5 Постановка проблеми 27
2 Розробка методу 29
2.1 Інтеграція теорії приблизних множин та дерев рішень 29
2.1.1 Методологічні підходи до інтеграції дерев рішень і теорії приблизних множин 29
2.1.2 Алгебраічний підхід до до інтеграції дерев рішень і теорії приблизних множин 30
2.1.3 Інтеграція метрики важливості у дерева рішень . . . 31
2.1.4 Деталі імплементації дерева рішень 32
2.1.5 Імплементація гібрідної моделі 32

2.2 Попередня обробка даних 36
2.3 Оцінка та порівняння отриманих моделей 37
2.3.1 Оцінка точності у задачі 37
2.3.2 Стратегія валідації 38
3 Аналіз отриманих результатів 40
3.1 Порівняння точності 40
3.2 Візуальне порівняння методів 41
3.3 Аналіз кореляції між отриманими методами 42
3.4 Вплив шуму на гібрідний метод 44
3.5 Подальші дослідження 47
3.5.1 Дослідження різних типів даних 48
3.5.2 Вдосконалені техніки відбору ознак 48
3.5.3 Оптимізація параметрів моделі 48
3.5.4 Порівняльні дослідження з іншими моделями 49
3.5.5 Застосування в реальних сценаріях 49
3.5.6 Дослідження масштабованості 49
3.5.7 Вдосконалення реалізації теорії приблизних множин 49
3.5.8 Моделі, які зрозумілі для користувачів 50
Висновки 51
Перелік джерел посилання 52

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ДР – дерево рішень

ТПМ – теорія приблиних множин

МН – машинне навчання

ВО – важливість ознаки

Survival forecasting – прогназування виживання

ВСТУП

Ландшафт прийняття рішень на основі даних швидко розвивається завдяки появі складних моделей класифікації, які є ключовими у витягуванні цінних знань з великих та складних наборів даних. Ця магістерська робота зосереджена на вивченні взаємодії між деревами рішень та теорією приблизних множин (ТПМ) у завданнях класифікації, області, яка набула значного поширення як у академічних, так і у промислових дослідженнях.

Дерева рішень, відомі своєю простотою та інтерпретованістю, широко використовуються у різноманітних застосуваннях, від сегментації клієнтів до діагностики захворювань. Вони працюють, розбиваючи набір даних на менші підмножини, одночасно розробляючи відповідне дерево рішень поступово. Сила дерев рішень полягає у їхній здатності моделювати складні процеси прийняття рішень, використовуючи деревоподібну структуру рішень та їх можливі наслідки.

Теорія приблизних множин, запроваджена Здзіславом Павлаком у 1980-х роках, пропонує математичний підхід до нечіткості та невизначеності у даних. ТПМ довела свою ефективність у роботі з неточною та неповною інформацією, роблячи її цінним інструментом для аналізу даних, особливо в сфері вибору ознак та індукції правил. Інтеграція ТПМ з деревами рішень представляє унікальну можливість підвищити надійність та точність моделей класифікації, особливо в наборах даних, що містять неоднозначності та невизначеності.

Ця магістерська робота має на меті дослідити та розробити комплексний фреймворк, який використовує сильні сторони як дерев рішень, так і теорії приблизних множин. Ціллю є створення моделі класифікації, яка не тільки забезпечує високу точність, але й зберігає інтерпретованість та надійність у випадках шумних та неповних даних. Роблячи можливим використання такого методу у різних сферах, таких як фінанси або кібербезпека.

1 ТЕОРЕТИЧНА ОСНОВА

1.1 Задача класифікації

1.1.1 Огляд задачі класифікації в аналізі даних

Класифікація в аналізі даних - це метод ідентифікації, до якої з набору категорій належить нове спостереження на основі навчального набору даних, що містить спостереження, членство в категоріях яких відоме. Це форма навчання з учителем, що означає, що навчання моделі відбувається за допомогою попередньо визначеного набору навчальних прикладів. Цей підхід є важливим у різних галузях, включаючи машинне навчання, статистику та розпізнавання образів.

Однією з ранніх застосувань класифікації в аналізі даних є робота Рональда Фішера з лінійним дискримінантом в 1936 році. Метод спочатку використовувався для відокремлення двох або трьох видів квіток ірису (рис. 1.1) на основі довжини і ширини їхніх чашолистиків та пелюсток [1]. Ця рання робота створила основу для подальших розвитків методів класифікації.

З роками методи класифікації значно еволюціонували, великою мірою завдяки досягненням у обчислювальній потужності та доступності даних. Сучасні методи класифікації різноманітні, включаючи дерева рішень, нейронні мережі, методи опорних векторів (SVM), а також ансамблеві методи, такі як випадкові ліси. Ці методи відрізняються за складністю, інтерпретованістю і придатністю для різних видів даних та завдань [2].

У практичних застосуваннях класифікація використовується для вирішення різноманітних завдань, від фільтрації електронної пошти (спам чи не спам) до медичної діагностики (хворі чи здорові). Універсальність методів класифікації робить їх невід’ємною частиною в епоху великих обсягів даних, де потрібно ефективно аналізувати та категоризувати велику кіль-

кість інформації [3].

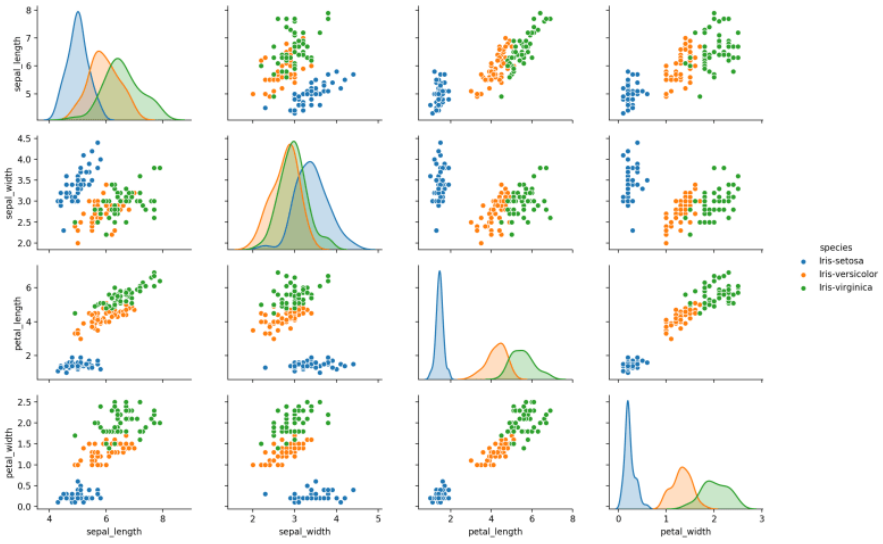

Рисунок 1.1 – Іріси Фішера

Одним з ключових аспектів класифікації є її залежність від якості та кількості даних. Ефективність моделі класифікації великою мірою залежить від репрезентативності та різноманітності навчальних даних. Неточності у навчальних даних можуть призвести до моделей, які є спрямованими або неефективними в реальних сценаріях [4].

Ще однією розглянутою проблемою є баланс між складністю та інтерпретованістю моделі. При більш складних моделях, таких як глибокі нейронні мережі, може бути вища точність, але вони часто розглядаються як ”чорні скриньки”через відсутність інтерпретованості. З іншого боку, простіші моделі, наприклад, дерева рішень, легше інтерпретувати, але вони можуть не так ефективно виявляти складні патерни в даних [5].

Слід відзначити, що класифікація в аналізі даних є динамічною та важливою галуззю, яка постійно розвивається завдяки технологічним досягненням. Її застосування охоплюють багато галузей та надають важливі інсайти, сприяючи процесам прийняття рішень. Оскільки обсяг та складність даних продовжують зростати, розробка більш надійних, ефективних та ін-

терпретованих методів класифікації залишається ключовим напрямком дослідження.

1.1.2 Історія еволюції методів класифікації

Історія еволюції методів класифікації в аналізі даних є подорожжю через різні наукові дисципліни та технологічні досягнення. Класифікація, в її основі, полягає в категоризації чи позначенні даних на основі їх атрибутів, концепт, який був невід'ємною частиною людських знань та розуміння протягом століть.

Одним із найраніших формальних підходів до класифікації в статистиці був розроблений на початку 20-го століття. У цей період відбувся народження лінійного дискримінантного аналізу, запропонованого Рональдом Фішером у 1936 році. Метод Фішера спрямовувався на лінійне виділення різних класів на основі їхніх ознак, що було революційним у таких галузях, як біологія для класифікації видів [1].

Середина 20-го століття призначена появою комп'ютерів, що значно прискорило розвиток технік класифікації. Одним із перших методів класифікації, які використовували комп'ютерну допомогу, був алгоритм k - найближчих сусідів (k-NN), запропонований у 1950-х роках. Цей алгоритм класифікує дані на основі найближчих прикладів з навчального набору в просторі ознак, акцентуючи увагу на важливості метрик відстаней у класифікації [6].

У 1970-х і 1980-х роках сфера машинного навчання почала набирати обертів, і разом з цим виникли дерева рішень, популяризовані алгоритмами, такими як ID3 і C4.5, розробленими Россом Квінланом. Дерева рішень дозволили створювати більш інтуїтивне і візуальне представлення класифікації, що дозволило аналізувати більш складні структури даних [7].

1990-і роки відзначилися появою методу опорних векторів (SVM), запропонованими Вапніком і Кортесом. SVM були значним проривом і нада-

ли надійну продуктивність у високорозмірних просторах. Вони працюють, знаходячи оптимальну гіперплощину, яка максимізує відстань між різними класами, проявляючи особливу ефективність у завданнях класифікації тексту та зображень [8].

21-й століття відзначається бумом великих даних та зростаючою складністю завдань класифікації. Ансамблеві методи, такі як випадкові ліси, стали популярними завдяки здатності поєднувати кілька класифікаторів для покращення точності та надійності. Інтеграція технік глибокого навчання, переважно через нейронні мережі, також стала важливим кроком у роботі з обсягом та складністю великих наборів даних, особливо у галузях, таких як розпізнавання зображень та мови [9].

Сьогодні техніки класифікації є синтезом статистики, інформатики та штучного інтелекту, постійно еволюціонуючи для врахування зростаючої складності та обсягу даних у різних галузях.

1.1.3 Важливість та Сфери Застосування Класифікації

Класифікація є важливим елементом аналізу даних, відіграючи ключову роль у інтерпретації та розумінні великих обсягів даних, що генеруються у сучасному світі. Її важливість полягає в здатності організовувати дані у значущі категорії, що спрощує аналіз та прийняття рішень у різних сферах.

Однією з найважливіших сфер застосування класифікації є медицина. У медичній діагностиці алгоритми класифікації допомагають ідентифікувати категорії хвороб на основі симптомів та результатів тестів. Ця діяльність є важливою для своєчасної та точної діагностики, яка є необхідною для ефективного лікування. Наприклад, моделі машинного навчання все більше використовуються для виявлення та класифікації раку на основі даних із зображень, покращуючи точність та швидкість діагностики порівняно з традиційними методами [10].

У фінансах методи класифікації використовуються для оцінки кредитних ризиків, категоризації інвестиційних можливостей та виявлення фінансових шахраївств. Фінансові установи використовують алгоритми для класифікації заявників на кредити в різні ризикові категорії, що впливає на прийняття рішень про надання позики. Так само при виявленні шахраївства класифікація допомагає в ідентифікації паттернів операцій, які ймовірно є обманними, тим самим захищаючи фінансові активи [11].

Ще однією значущою областю застосування є сфера соціальних медіа та маркетингу. Алгоритми класифікації аналізують дані та відгуки споживачів для категоризації їхніх вподобань та поведінки. Ця інформація є надзвичайною для цільового маркетингу та стратегій розробки продуктів, оскільки вона дозволяє підприємствам налаштовувати свої пропозиції для конкретних сегментів споживачів [12].

У галузі екологічних наук класифікація використовується для аналізу використання і покриття землі, що є важливим для моніторингу та планування навколишнього середовища. Шляхом класифікації супутникових та аерофотознімків у різні типи покриття землі, дослідники та політики можуть відслідковувати зміни в екосистемах, розширення міст і наслідки змін клімату [13].

Важливість класифікації також поширюється на зростаючу галузь автономних автомобілів. Алгоритми класифікації є фундаментальними для того, щоб ці автомобілі могли інтерпретувати дані сенсорів, розрізняючи різні об’єкти, такі як автомобілі, пішоходи та дорожні знаки, що є важливим для безпечного руху [14].

Широке застосування технік класифікації підкреслює їхню універсальність та важливу роль у сучасному аналізі даних. Шляхом перетворення сирого даних у категоризовану та керовану інформацію методи класифікації допомагають в галузях від охорони здоров’я до управління навколишнім середовищем, сприяючи розвитку та ефективності в кожній галузі.

1.2 Основи методів дерев рішень

1.2.1 Базові концепти у методах дерев рішень

Дерева рішень є фундаментальною частиною машинного навчання та аналізу даних, надаючи модель для прийняття рішень та класифікації, яка є як зрозумілою, так і універсальною. Дерево рішень - це структура, схожа на блок-схему, в якій кожен внутрішній вузол представляє тест на атрибут, кожна гілка представляє результат тесту, а кожний листовий вузол представляє мітку класу або рішення, прийняте після обчислення всіх атрибутів. Шляхи від кореня до листя представляють правила класифікації [15, 7, 3].

В сутності дерево рішень розбиває дані на підмножини на основі значень атрибутів (рис. 1.2). Цей процес відомий як ”розгалуження”. Розгалуження продовжується, поки алгоритм не визначить, що додаткові розгалуження не призведуть до кращих передбачень, або досягне попереднього критерію зупинки. Краса дерев рішень полягає в їх простоті та легкості інтерпретації, що робить їх доступним інструментом для вчених-дослідників та фахівців у галузі аналізу даних.

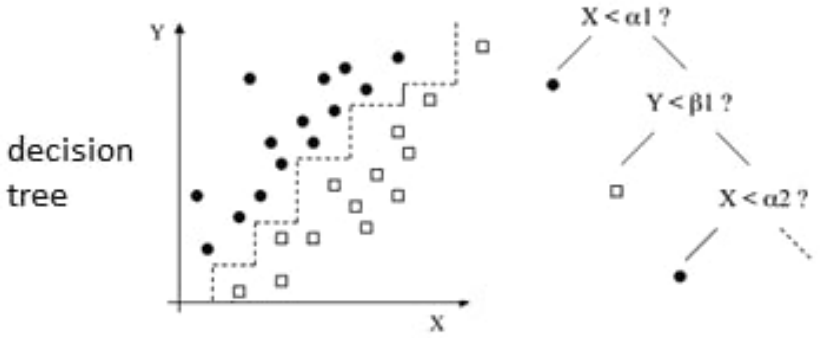

Рисунок 1.2 – Розбиття деревом рішення

Будівництво дерева рішень починається з визначення атрибуту, який

найкраще розділяє вибірку. Для вимірювання найкращого розгалуження використовуються різні метрики, включаючи Gini, приріст інформації та відношення приросту. Gini вимірює частоту помилкової класифікації [15] будь-якого елемента набору даних при випадковому визначенні класу 1.1.

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \tag{1.1}$$

де p_i - це ймовірність того, що елемент буде класифікований в клас i у наборі даних D.

З іншого боку, приріст інформації базується на концепції ентропії з теорії інформації. Він вимірює очікуваний зменшення ентропії або випадковості після розгалуження набору даних за певним атрибутом. Вищий приріст інформації вказує на більше значуще зменшення ентропії, що вказує на краще розгалуження [7]. Ентропія набору даних визначається наступним чином:

$$Entropy(D) = -\sum_{i=1}^{m} p_i \log_2 p_i \tag{1.2}$$

де p_i знову ж таки є ймовірність того, що елемент буде класифікований в клас i.

Подібність між індексом Джині та ентропією полягає в їх призначенні та поведінці (рис. 1.3). Обидві ці метрики є вимірами чистоти вузла в дереві рішень і мінімізуються в процесі побудови дерева. Вони надають критерій для оцінки того, наскільки добре певний розподіл даних розділяє дані на однорідні класи. На практиці обидві ці метрики часто дають схожі результати щодо структури дерева рішень та точності класифікації.

Однак є відмінності в їх математичних властивостях та обчислювальній ефективності. Ентропія, завдяки своєй логарифмічній функції, зазвичай є більш обчислювально інтенсивною порівняно з індексом Джині. Ло-

гарифмічна функція в ентропії може реагувати по-різному на зміни у ймовірностях класів, що іноді призводить до трохи відмінних розподілів порівняно з індексом Джині, особливо в випадках з багатьма мітками класів.

У підсумку, хоча індекс Джині та ентропія різняться у своїх математичних формулах, вони схожі за своєю роллю в алгоритмах дерева рішень. Обидві вони вимірюють організованість вузла та ведуть алгоритм до знаходження найкращих розподілів для розділення даних на однорідні класи, тим самим відіграючи важливу роль у ефективності дерева рішень як моделі класифікації.

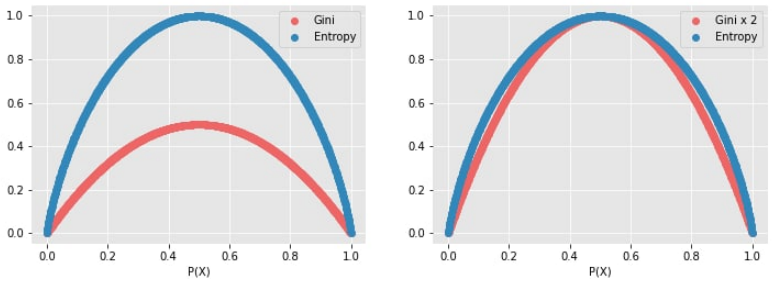

Рисунок 1.3 – Порівняння ентропії з індексом Джинні

Алгоритм дерева рішень перебирає кожну ознаку для визначення розгалуження, яке призводить до максимального приросту інформації або мінімальної Gini. Після визначення найкращого атрибуту для розгалуження кореня, набір даних розділяється, і процес повторюється рекурсивно на кожному розділеному піднаборі. Це рекурсивне розділення є важливою складовою будівництва дерева рішень і сприяє його ефективності в задачах класифікації.

Важливим аспектом будівництва дерева рішень є уникнення перенавчання (рис. 1.4), поширеної проблеми, коли модель стає надто складною і починає захоплювати шум в даних, а не відображати фактичні базові закономірності. Використовуються методи, такі як обрізка (вилучення частини дерева, які надають мало потужності для класифікації екземплярів), щоб

забезпечити, що дерево залишається узагальненням.

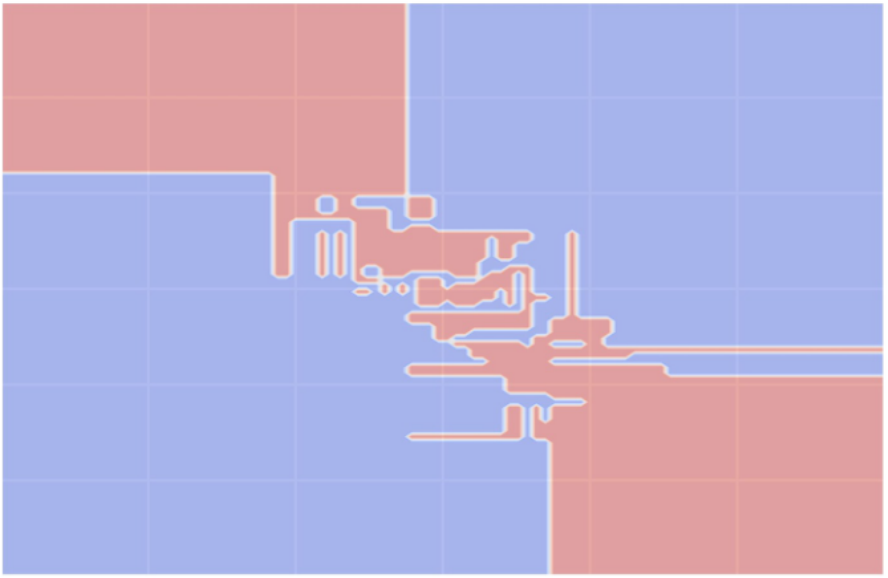

Рисунок 1.4 – Приклад перенавчання у випадку дерева рішення

Заключно, дерева рішень є потужним інструментом для класифікації та прийняття рішень. Їх проста, правило-заснована структура дозволяє легко інтерпретувати та аналізувати їх, роблячи їх популярним вибором в різних застосуваннях, від сегментації клієнтів до медичної діагностики. Як фундаментальна техніка у галузі машинного навчання, розуміння основних концепцій дерев рішень є важливим для будь-якого аналітика даних або дослідника.

1.2.2 Архітектура та конструкція дерев рішень

Архітектура та конструкція дерев рішень включають систематичні процеси для створення моделі, яка точно класифікує або передбачає результати на основі вхідних змінних. Архітектура дерева рішень є ієрархічною за природою і складається з вузлів, гілок та листя, кожний з яких відіграє важливу роль у процесі прийняття рішень [4] .

Зазвичай дерево рішень починається з кореневого вузла, який містить весь набір даних. Цей вузол поділяється на підвузли на основі конкретних критеріїв, які часто визначаються за допомогою показників, таких як Gini чи приріст інформації, як обговорювалося раніше. Кожен внутрішній вузол в дереві представляє тест на атрибут, що призводить до розгалуження набору даних на менші підмножини. Вузли, на яких більше не відбувається подальшого розгалуження, називаються листовими вузлами або кінцевими вузлами. Ці листя представляють рішення або класифікаційний результат.

Вибір атрибуту для розгалуження на кожному вузлі є ключовим кроком у конструкції дерева. Цей процес керується статистичними показниками. Наприклад, показник Gini оцінює однорідність міток на вузлі 1.1

Підхід на основі приросту інформації використовує концепцію ентропії 1.2, що представляє собою розлад або випадковість інформації. Зменшення ентропії після поділу набору даних D за атрибутом A є приростом інформації 1.3.

$$Gain(D, A) = Entropy(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} Entropy \tag{1.3}$$

де $Values(A)$ - це різні значення атрибуту A, а D_v - це підмножина D для якої атрибут A має значення v.

Конструкція дерева рішень включає рекурсивне поділ набору даних до досягнення критерію зупинки. Цей критерій може бути максимальною глибиною дерева, мінімальною кількістю вибірок у листі або прийнятним рівнем чистоти в листових вузлах. Під час цього рекурсивного процесу алгоритм вибирає найкращий атрибут для розгалуження даних на кожному вузлі на основі вибраного статистичного показника.

Для запобігання перенавчанню, дерева рішень часто піддають стрижці. Стрижка дерева включає в себе обмеження дерева, щоб видалити розгалуження, які мають мало або жодної цінності для класифікації екземпля-

рів. Існують два основні методи стрижка: стрижка заздалегіть, яка завершує конструкцію дерева рано, і післястрижка, яка видаляє гілки з повністю вирощеного дерева. Загальна стратегія для післстрижки - це стрижка з помилкових дерев, де гілки заміщуються листовим вузлом, якщо заміна не значно не знижує точність класифікації на валідаційному наборі.

1.2.3 Ключові алгоритми у конструкції дерев рішень

Конструкція дерев рішень є процесом, центральним для багатьох завдань машинного навчання, який характеризується підходом до поділу даних на гілки для створення структури дерева. Ефективність та ефективність дерева рішень в значній мірі залежать від алгоритмів, які використовуються для його створення. З багатьох алгоритмів, розроблених протягом років, виділяються три, які вирізняються своєю широкою поширеністю та фундаментальними ролями в галузі: ID3, C4.5 і CART.

Розроблений Росом Квінланом в 1980-х роках, алгоритм ID3 [7] є одним із найстаріших і найвпливовіших алгоритмів для конструкції дерев рішень. ID3 використовує жадний пошук зверху вниз через простір можливих гілок без повернення назад. Як критерій вибору атрибутів, які найкраще розділять вибірки на цільові класи, використовується приріст інформації. Приріст інформації для атрибута A обчислюється на основі ентропії - міри безладу або невизначеності інформації 1.2.

де S - це множина всіх вибірок, $Values(A)$ - це різні значення атрибута A, а S_v - вибірки в S, для яких атрибут A має значення v.

C4.5, розширення ID3, також був розроблений Квінланом на початку 1990-х років [17]. Він поліпшує ID3, працюючи як з неперервними, так і з дискретними атрибутами, вирішуючи проблему відсутніх значень та впроваджуючи форму післяобрізки для боротьби з перенавчанням. C4.5 використовує концепцію відношення приросту, модифікацію приросту інформації, яка включає нормалізаційний фактор, вирішуючи проблему спрямо-

ваності ID3 на атрибути з більшою кількістю рівнів. Відношення приросту для атрибута A визначається як:

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInfo(S, A)} \tag{1.4}$$

де $SplitInfo(S, A)$ - це потенційна інформація, створена поділом навчального набору даних S на підмножини відповідно до значень атрибута A. p_i - це ймовірність того, що об'єкт в S належить класу i.

1.2.4 Переваги та обмеження дерев рішень

Дерева рішень, популярний інструмент у арсеналі аналітиків даних та практиків машинного навчання, мають комбінацію переваг і обмежень, що робить їх підходящими для деяких застосувань, але менш ідеальними для інших.

Однією з основних переваг дерев рішень є їхня зрозумілість та легкість інтерпретації. На відміну від багатьох моделей "чорних ящиків дерева рішень надають чітке візуальне представлення процесу прийняття рішень, дозволяючи як технічним, так і нетехнічним зацікавленим особам зрозуміти логіку за прогнозами. Ця прозорість особливо цінна в галузях, де розуміння обгрунтування рішення є так само важливим, як саме рішення, наприклад, в сфері охорони здоров'я та фінансів.

Ще однією значущою перевагою є їхня гнучкість у роботі як з числовими, так і з категоріальними даними. Дерева рішень можуть природно обробляти і розбивати дані різних типів без необхідності виконувати обширну підготовку даних, як це часто потрібно для інших алгоритмів машинного навчання, таких як one-hot encoding або нормалізація.

Крім того, дерева рішень відносно стійкі до викидів і можуть ефективно обробляти відсутні дані, особливо в алгоритмах, таких як C4.5, які включають механізми для вирішення проблем з неповними даними. Ця стійкість

робить їх практичним вибором в реальних сценаріях, де дані часто є неповними.

Незважаючи на ці переваги, дерева рішень мають свої обмеження. Одним з найбільш помітних є їхня схильність перенавчання на навчальних даних. Дерева рішень можуть створювати надто складні структури, які погано узагальнюються на нові дані, особливо якщо дерево глибоке. Це перенавчання в основному пов’язане з їхньою ієрархічною природою, де кожне рішення базується на підмножині даних, що може призводити до потенційної великої дисперсії у прогнозах.

Іншим обмеженням є їхня нестабільність. Невеликі зміни в даних можуть призвести до значно різних структур дерева. Ця проблема особливо виразна у деревах рішень через їхній ієрархічний, жадібний підхід, де кожний розгалужений вузол приймається без урахування подальших розгалужень. Цю нестабільність можна пом’якшити за допомогою методів ансамблю, таких як випадковий ліс, але вона залишається проблемою при використанні окремих дерев рішень.

Дерева рішень також мають складнощі в моделюванні певних відносин, таких як ті, які включають проблеми XOR, паритет або мультиплексор. У таких сценаріях кордон між класами не легко визначити ієрархією рішень, що вимагає надзвичайно складної структури дерева.

На завершення, дерева рішень спрямовані на вибір атрибутів із більшою кількістю рівнів. Цю спрямованість можна виправити за допомогою алгоритмів, таких як C4.5, які використовують відношення приросту замість чистого приросту інформації, але вона є вродженою обмеженістю в простіших алгоритмах дерев рішень.

У підсумку, дерева рішень пропонують переваги в зрозумілості, гнучкості та стійкості до викидів і відсутніх даних, але вони також схильні до перенавчання, нестабільні та мають складності в моделюванні складних границь класів. Розуміння цих переваг і обмежень є важливим для аналітиків даних і практиків машинного навчання під час вибору відповідної

моделі для своєї конкретної задачі.

1.3 Розуміння теорії приблизних множин

1.3.1 Вступ до Теорії приблизних множин

Теорія приблизних множин, вперше представлена Здзіславом Павлаком на початку 1980-х років [18], є математичним підходом до роботи з вагомістю та невизначеністю в аналізі даних. На відміну від класичної теорії множин, яка вимагає чіткої класифікації, теорія приблизних множин дозволяє аналізувати ситуації, де є нездатність точно визначити приналежність до множини через неповну або недостатньо точну інформацію.

У своїй сутності теорія приблизних множин базується на апроксимації множин. У цьому контексті будь-який данний підмножина об’єктів, яка характеризується відносинами нерозрізненості, апроксимується двома множинами - нижньою апроксимацією і верхньою апроксимацією 1.5. Нижня апроксимація множини складається з усіх елементів, які точно належать множині на основі наявної інформації, тоді як верхня апроксимація містить елементи, які, можливо, належать множині.

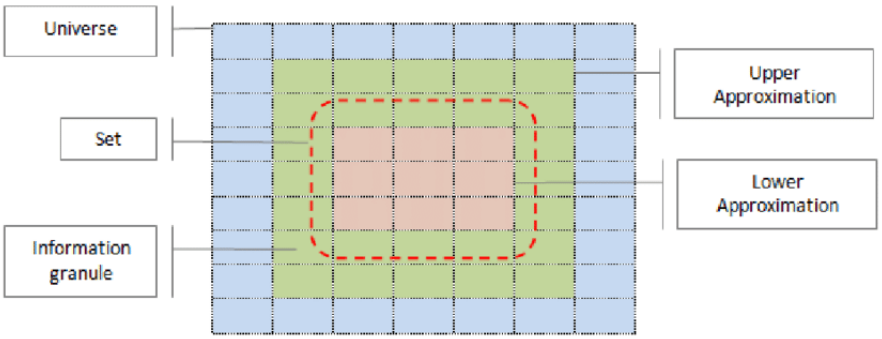

Рисунок 1.5 – Основні елементи ТПМ

Математично, для підмножини X у всесвіту U та відношенні еквіва-

лентності R, нижня апроксимація і верхня апроксимація визначаються так:

$$\underline{apr}_R(X) = \{x \in U \mid [x]_R \subseteq X\} \tag{1.5}$$

$$\overline{apr}_R(X) = \{x \in U \mid [x]_R \cap X \neq \emptyset\} \tag{1.6}$$

де $[x]_R$ позначає клас еквівалентності x за відношенням R. Пара цих апроксимацій утворює приблизну множину.

Основною перевагою теорії приблизних множин є її здатність обробляти невизначеність без потреби в будь-якій попередній або додатковій інформації про дані, такі як ймовірнісні розподіли чи функції належності, які є важливими в ймовірнісних та нечітких підходах до множин [19]. Це робить теорію приблизних множин особливо корисною в експлораторному аналізі даних та виявленні знань, де точна природа даних не повністю відома.

Ще одним важливим аспектом теорії приблизних множин є акцент на зменшенні кількості атрибутів у наборі даних зі збереженням його потужності для класифікації. Це зменшення досягається шляхом визначення та видалення атрибутів, які є зайвими чи несуттєвими для задачі класифікації.

Теорія приблизних множин також сприяє видобутку правил прийняття рішень із даних. Ці правила походять з нижньої та верхньої апроксимацій і можуть бути використані для класифікації та прогнозування. Можливість генерації явних правил із даних є цінною в галузях, де пояснюваність та інтерпретованість рішень моделі є важливими.

Теорія приблизних множин пропонує унікальний та ефективний підхід до роботи з невизначеністю та нечіткістю, які є властивими багатьом реальним наборам даних. Її акцент на апроксимаціях, зменшенні атрибутів та генерації правил робить її потужним інструментом для аналізу да-

них, особливо в областях, де точність компрометується через неповну або нечітку інформацію.

1.3.2 Ключові принципи та концепції в теорії приблизних множин

Теорія приблизних множин(ТПМ), математичний фреймворк для роботи з невизначеністю та нечіткістю в даних, ґрунтується на кількох ключових принципах та концепціях. Ці основні елементи дозволяють теорії ефективно управляти неточною чи неповною інформацією, роблячи її цінним інструментом у аналізі даних.

Концепція недискретності є центральною для ТПМ. Вона базується на ідеї, що об'єкти вивчення можуть бути нерозрізнені один від одного на основі доступних атрибутів. Відношення еквівалентності, зазвичай позначене як R, визначається на всесвіті дискусії U і розбиває його на класи еквівалентності. Ці класи групують разом елементи, які є нерозрізними один від одного щодо R. Відношення недискретності є фундаментальною концепцією, яка лежить в основі підходу теорії до обробки невизначеності.

Приблизні множини характеризуються своїми нижньою та верхньою апроксимаціями. Точність апроксимації множини X в ТПМ - це вимірювання того, наскільки добре X можна приблизити за допомогою доступної інформації. Вона визначається як відношення між кількістю елементів нижньої апроксимації 1.5 до кількості елементів верхньої апроксимації 1.6:

$$\alpha_R(X) = \frac{|\underline{apr}_R(X)|}{|\overline{apr}_R(X)|} \tag{1.7}$$

Це відношення, яке знаходиться в межах від 0 до 1, вказує на ступінь невизначеності при визначенні членства елементів у множині X. Відношення 1 вказує на чітку множину, тоді як відношення менше 1 вказує на наближену множину.

ТПМ надає фреймворк для аналізу залежності атрибутів. Атрибут є залежним від іншого, якщо класифікація множини в універсумі може бути змінена цим атрибутом. Ця концепція є важливою для розуміння значущості кожного атрибута в наборі даних, що призводить до скорочення атрибутів та спрощення процесу аналізу даних.

Редукт - це мінімальний набір атрибутів, який може бути використаний для визначення тієї ж самої класифікації, що і весь набір атрибутів. Перетин усіх редуктів формує ядро, яке є набором всіх незамінних атрибутів. Визначення редуктів та ядра є важливим в Теорії Наближених Множин для спрощення даних без втрати важливої інформації.

ТПМ, з акцентом на відношеннях еквівалентності, апроксимаціях та залежності атрибутів, надає міцний математичний фреймворк для роботи з невизначеністю чи неповною інформацією. Її здатність виділяти важливі атрибути та формулювати наближення робить її потужним інструментом в областях від видобутку даних до когнітивної науки.

1.3.3 Теорія приблизних множин в аналізі даних

Теорія приблизних множин (ТПМ) виросла в значущий інструмент аналізу даних, особливо у сфері вилучення корисної інформації з неточних або неповних наборів даних. Її унікальний підхід до роботи з невизначеністю та нечіткістю відрізняє її від інших методологій аналізу даних.

У традиційних методах аналізу даних вирішення питань невизначеності часто вимагає додаткової інформації, такої як розподіли ймовірностей чи ступені належності. Однак ТПМ підходить до цього інакше. Працюючи з нижніми та верхніми апроксимаціями, вона може впоратися з даними, де межі категорій не чітко визначені. Ця особливість особливо корисна в реальних сценаріях, де дані не завжди точні або повні.

Однією з найбільш помітних застосувань ТПМ в аналізі даних є передобробка даних. ТПМ може значно скоротити складність даних, виявля-

ючи та видаляючи зайві або неінформативні ознаки. Цей процес, відомий як скорочення атрибутів або вибір ознак, підвищує ефективність аналізу даних, особливо в наборах даних високої вимірковості. Зберігаючи лише найінформативніші атрибути, ТПМ спрощує моделі, покращує їх інтерпретовність та часто підвищує їх продуктивність.

ТПМ виявляється особливо ефективною у видобутку правил при аналізі даних [23, 24]. Ці правила прості та часто мають форму ”якщо-то що робить їх легкими для інтерпретації та застосування. У багатьох галузях, таких як фінанси та охорона здоров’я, де інтерпретовність так само важлива, як і точність, ці правила надають безцінні відомості. Наприклад, у медичній діагностиці правила, видобуті через ТПМ, можуть допомогти зрозуміти взаємозв’язки між симптомами та захворюваннями, полегшуючи процеси прийняття рішень.

В галузі відкриття знань ТПМ застосовується для виявлення шаблонів та взаємозв’язків в даних, які не завжди очевидні. Вона особливо корисна в областях, де дані є невизначеними чи неповними, таких як аналіз ринку чи дослідження вибору споживачів. Можливість ТПМ наближати множини на основі наявних знань підтримує створення систем підтримки рішень, які можуть працювати ефективно, навіть із неповними даними.

ТПМ порівнюють з іншими теоріями, такими як теорія нечітких множин [21] та традиційні статистичні методи. У той час як Теорія нечітких множин розглядає нечіткість шляхом призначення ступенів належності, ТПМ розглядає її через призму наближення, без необхідності додаткової інформації. Ця відмінність робить ТПМ більш застосовною в ситуаціях, де призначення ступенів належності є неможливим або занадто суб’єктивним.

Незважаючи на переваги, використання ТПМ в аналізі даних ставить перед собою деякі виклики. Один з найважливіших викликів - це обчислювальна складність, особливо в великих наборах даних з багатьма атрибутами. Крім того, вибір відношення еквівалентності, яке є ключовим для

формування приблизних множин, може суттєво вплинути на результати, вимагаючи обережного розгляду та експертизи в галузі.

Теорія приблизних множин пропонує унікальний підхід до аналізу даних, особливо в роботі з невизначеною чи неповною інформацією. Її застосування в виборі ознак, видобутку правил [22] та виявленні шаблонів робить її цінним інструментом у різних галузях. Оскільки дані продовжують зростати за складністю та обсягом, роль ТПМ в аналізі даних може стати все більш важливою.

1.4 Аналіз предметної області валідаційних датасетів

У інтеграції дерев рішень з теорією приблизних множин, особливо коли використовується важливість ознак як критерій для розбиття дерева, важливу роль відіграють природа та підготовка набору даних. У цьому розділі наведено типи даних, які використовуються, та необхідні кроки передпроцесінгу для ефективного застосування цього інтегрованого підходу.

Набір даних ”Титанік”[34], відомий в галузі машинного навчання та науки про дані, створює ідеальні умови для тестування покращеної моделі дерева рішень. Цей набір даних включає інформацію про пасажирів знаменитого корабля ”Титанік включаючи такі характеристики, як вік, стать, клас, вартість квитка та статус виживання. Складність та різноманітність цих даних роблять їх ідеальними для оцінки продуктивності моделі у роботі з реальними даними, які є невизначеними та категоріальними.

У сфері кібербезпеки, зокрема для класифікації шкідливого програмного забезпечення, набір даних ”Microsoft Malware Classification Challenge (BIG 2015)”[35] виступає як важливий бенчмарк. Він включає велику кількість об’єктів, кожен з яких представляє різні типи шкідливого програмного забезпечення і характеризується різними ознаками, такими як байткод, асемблерний код та інші атрибути файлів. Складність цього набору даних та його критичне значення в застосуванні роблять його ідеальним для оцін-

ки продуктивності розширеної моделі дерева рішень в умовах високих ризиків і технічної складності.

Набір даних Home Credit Default Risk [36] надає повну основу для тестування покращеної моделі дерева рішень в контексті оцінки кредитного ризику. Цей набір даних особливо важливий для фінансових установ, які прагнуть точно передбачити можливість повернення позикодавцями кредитів. Він містить різноманітні характеристики, включаючи кредитну історію, деталі позики та демографічні дані позичальника, що робить його корисним для оцінки ефективності моделі в фінансовому середовищі.

1.5 Постановка проблеми

Інтеграція теорії приблизних множин (ТПМ) з деревами рішень (ДР) спрямована на вирішення значущої проблеми в галузі аналізу даних: покращення точності класифікації при обробці властивої невизначеності та складності реальних даних. Ця інтеграція особливо актуальна в сценаріях, де дані є неповними, нечіткими або шумними, що є загальними характеристиками у різних галузях, таких як охорона здоров’я, фінанси та аналіз соціальних медіа.

Традиційні методи класифікації, включаючи стандартні дерева рішень, часто мають труднощі з наборами даних, що містять невизначені, нечіткі або неповні відомості. Жорсткість цих методів у роботі з даними може призвести до помилкової класифікації або перенавчання, особливо при роботі з складними або високо-вимірними наборами даних. Крім того, інтерпретованість цих моделей може бути пожертвована в погоні за більшою точністю, особливо в методах, які використовують складні перетворення або покладаються на ”чорні ящики”алгоритми.

ТПМ пропонує рамки для роботи з невизначеністю в даних без потреби в додатковій інформації чи припущеннях. Її підхід до класифікації даних через нижні та верхні наближення надає механізм для більш ефе-

ктивної обробки випадків на межі категорій. Однак ТПМ сама по собі може бути недостатньою для складних завдань класифікації, оскільки вона в основному фокусується на передпроцесингу та скороченні даних, а не на прогностичному моделюванні.

Запропонована інтеграція ДР та ТПМ має на меті використовувати переваги обох підходів. Дерева рішень привносять структурований ієрархічний підхід до класифікації, пропонуючи інтерпретованість та легкість впровадження. У поєднанні з можливістю теорії приблизних множин вирішувати невизначеність та скорочувати простір атрибутів, гіпотетично інтегрована модель може досягти більшої точності та надійності, ніж кожен з методів окремо.

Метою цього дослідження є розробка гібридної моделі класифікації, яка поєднує надійність ТПМ у роботі з невизначеними та неповними даними з структурованим процесом прийняття рішень ДР. Модель спрямована на покращення точності класифікації в наборах даних, де традиційні методи зазнають труднощі через питання якості даних. Дослідження буде досліджувати методологічні підходи до цієї інтеграції, вдосконалювати алгоритми для підвищення точності дерев рішень за допомогою ТПМ та оцінювати ефективність об’єднаного методу в різних сферах застосування.

Гіпотеза полягає в тому, що інтеграція ТПМ та ДР призведе до моделі класифікації, яка перевершує традиційні моделі дерев рішень, особливо в наборах даних, що відзначаються невизначеністю та неповністю. Очікуваним результатом є набір методологічних вказівок та алгоритмів, які можуть бути застосовані в різних галузях та призвести до більш точних, надійних та інтерпретованих моделей класифікації.

2 РОЗРОБКА МЕТОДУ

2.1 Інтеграція теорії приблизних множин та дерев рішень

2.1.1 Методологічні підходи до інтеграції дерев рішень і теорії приблизних множин

Інтеграція Дерев Рішень (ДР) та Теорії Приблизних Множин (ТПМ) в аналізі даних представляє методологічний прогрес, який використовує переваги обох підходів. Ця інтеграція спрямована на підвищення точності класифікації, покращення інтерпретованості та управління невизначеністю в даних.

Поточні дослідження підкреслюють важливість врахування невизначеності даних в деревах рішень [25], область, в якій теорія приблизних множин може внести свій внесок. Крім того, у літературі обговорюється [26, 27], як алгебраїчні методи теорії приблизних множин можуть надати систематичний спосіб для їх інтеграції, визнаючи складність цього завдання та потребу в інноваційних методах, які можуть ефективно об’єднати ці області.

Останні розробки [30, 31] в алгоритмах дерев рішень вказують на можливість такої інтеграції, тоді як вдосконалення в теорії приблизних множин, зокрема в алгебраїчних методах [32, 33], вказують на потенційні покращення в класифікації за допомогою дерев рішень.

З алгоритмічної точки зору ця інтеграція передбачає розробку гібридних моделей, які поєднують в собі принципи як ТПМ, так і алгоритмів дерев рішень. Такі моделі вимагають ретельного розгляду того, як взаємодіють обидва методи, особливо щодо потоку даних та порядку операцій. Наприклад, можна спочатку застосувати ТПМ для зменшення кількості атрибутів, а потім використовувати алгоритм дерева рішень на скороченому наборі даних, або, навпаки, використовувати ТПМ для післяобробки

результатів дерева рішень для роботи з невизначеністю в класифікаціях.

Інтеграція Дерев Рішень і Теорії Приблизних Множин є методологічним прогресом в аналізі даних, що поєднує переваги обох методів для розв’язання складних завдань класифікації. Ця інтеграція вимагає ретельного розгляду того, як найкраще комбінувати можливості апроксимації та поділу обох підходів, що призводить до більш точних, інтерпретованих та надійних моделей.

2.1.2 Алгебраїчний підхід до до інтеграції дерев рішень і теорії приблизних множин

Інтеграція Дерев Рішень (ДР) з Теорією Приблизних Множин (ТПМ) включає об’єднання методологій, де алгебраїчний підхід до ТПМ відіграє важливу роль. Цей алгебраїчний метод [26, 27] був досліджений у контексті генерації логічних правил, результати цього дослідженання кажуть про можливість ефективної інтегровані моделі на базі ДР для покращення моделей класифікації.

Алгебраїчний підхід в ТПМ [26, 27], надає нову перспективу для роботи з нечіткістю та невизначеністю в даних. Визначаючи приблизні множини за допомогою набору предикатів та операцій над предикатами, які називаються мовою приблизних множин, цей підхід дозволяє здійснювати більш витончені класифікації. Інтеграція цих алгебраїчних приблизних множин ТПМ з ДР може значно підвищити точність дерев, особливо в наборах даних, де традиційні бінарні розділи не можуть належним чином відобразити складність даних.

Впровадження алгебраїчного підходу ТПМ разом з ДР включає використання бінарних кодів для швидкого обчислення приблизних множин та ефективної генерації логічних правил на основі цих приблизних множин

Основний виклик в інтеграції алгебраїчного підходу ТПМ з ДР полягає в складності обробки булевих операцій та забезпеченні ефективного

обчислення точних верхніх (1.6) та нижніх 1.5 приблизних множин. Однак переваги, які надає цей підхід, такі як здатність генерувати точні логічні правила та обробляти різноманітні приблизні множини, роблять подолання цих викликів виправданими для досягнення більш точних та надійних моделей класифікації.

На базі цього алгебраїчного підходу, був розроблен метод [28] визначення важливості ознаки 2.1, що беруть участь в описі приблизної множини. Ця оцінка особливо корисна при виборі найважливіших ознак для завдання класифікації.

$$V(P_i) = 100\% * \frac{\Delta BN}{M(X)} \tag{2.1}$$

У цій формулі $M(X)$ - це кількість істинних значень у булевому векторі X, а ΔBN позначає зміну булевого представлення при розгляді ознаки P_i.

Ключовим етапом в оцінці важливості ознак є встановлення порогу, який називається ‘minDeterioration‘ і визначається аналітиком. Цей поріг допомагає класифікувати, чи є ознака важливою для завдання класифікації. Якщо $V(P_i) \geq$ minDeterioration, то ознака P_i вважається важливою. З іншого боку, якщо $V(P_i) <$ minDeterioration, то ознака вважається несуттєвою.

Оцінка важливості ознак в комбінованій моделі ДР та ТПМ є важливим етапом в удосконаленні моделі класифікації. Ця оцінка допомагає не лише зрозуміти внесок кожної ознаки, але й спрямовує на оптимізацію моделі для підвищення її продуктивності та ефективності.

2.1.3 Інтеграція метрики важливості у дерева рішень

Інтеграція важливості ознак 2.1, отриманої з алгебраїчного підходу Теорії Приблизних Множин (ТПМ), як критерію розгалуження у побудові

Дерева Рішень (ДР), пропонує нову методологію в галузі аналізу даних та машинного навчання. Цей підхід відрізняється від традиційних критеріїв розгалуження, таких як приріст інформації або індекса Джині, і зосереджується на силі класифікації окремих ознак.

У комбінованій моделі ДР та ТПМ важливість ознаки, обчисленої за формулою 2.1, може бути використана для визначення розгалужень у дереві рішень [29]. Ця важливість є мірою внеску певної ознаки в завдання класифікації, виміряної в термінах змін, які вона призводить до класифікації при її врахуваSubtopic 5нні. 2.3.1 Оцінка точності у задачі

2.1.4 Деталі імплементації дерева рішень

У звичайному дереві рішень розділення на кожному вузлі визначається на підставі метрик, таких як приріст інформації , який обчислює зменшення ентропії, або індексу Джині, яка вимірює однорідність міток.

2.1.5 Імплементація гібрідної моделі

Інтеграція Теорії Приближених Множин (ТПМ) з Деревами Рішень (ДР) представляє собою новий гібридний підхід у класифікації та аналізі даних. Цей метод поєднує надійність ТПМ у роботі з невизначеністю даних із структурованою можливістю прийняття рішень ДР. Впровадження цього гібридного методу передбачає декілька ключових етапів, кожен із яких використовує переваги обох методологій.

Потім застосовується ТПМ для обчислення нижніх та верхніх приближень даних. Ці приближення допомагають виявити найважливіші ознаки або атрибути та розуміти граничні області в наборі даних. Важливість кожної ознаки оцінюється за допомогою метрик на основі ТПМ, визначаючи їх внесок у завдання класифікації. Цей етап має велике значення для ідентифікації ключових ознак, які будуть використовуватися для розділу в

Input: Dataset D with features F and target T, Maximum depth d_{max}, Minimum samples per leaf s_{min}

Output: Decision tree model

```
function BuildTree(D, F, T, current_depth)
    if current_depth = d_max or |D| < s_min then
        return most frequent label in T
    end if
    BestSplit ← FindBestSplit(D, F, T)
    Create a node N representing BestSplit
    for each split S in BestSplit do
        D_left, D_right ← split D based on S
        N_child ← BuildTree(D_left, F, T, current_depth + 1)
        Attach N_child to N
    end for
    return N
end function
return BuildTree(D, F, T, 0)
```

Алгоритм 1 – Алгоритм будування дерева рішень

дереві рішень.

Завдяки метрикам важливості ознак, отриманим за допомогою ТПМ, будується дерево рішень. У відміну від традиційних ДР, які використовують нечистоту Джині або приріст інформації для розділу вузла, це гібридне дерево використовує важливість ознак, отриману за допомогою ТПМ. Кожен вузол в дереві розбивається на основі найважливішої ознаки, визначеної за допомогою ТПМ, забезпечуючи, що структура дерева включає найбільш інформативні шляхи прийняття рішень.

Input: Dataset D, Set of attributes A, Target attribute T

Output: Size of the boundary region

```
function CalculateBoundaryRegionSize(D, A, T)
    L ← LowerApproximation(D, A, T)
    U ← UpperApproximation(D, A, T)
    BoundaryRegion ← U  L
    return |BoundaryRegion|
end function
function LowerApproximation(D, A, T)
    Initialize LowerApproximationSet as an empty set
    for each subset S in D based on attributes A do
        if all elements in S have the same value of T then
            Add S to LowerApproximationSet
        end if
    end for
    return LowerApproximationSet
end function
function UpperApproximation(D, A, T)
    Initialize UpperApproximationSet as an empty set
    for each subset S in D based on attributes A do
        if there exists at least one element in S with the same value of T then
            Add S to UpperApproximationSet
        end if
    end for
    return UpperApproximationSet
end function
return CalculateBoundaryRegionSize(D, A, T)
```

Алгоритм 2 – Знаходження граничної області

2.2 Попередня обробка даних

У сфері моделювання з бінарними ознаками, особливо для дерев рішень і теорії приблизних множин, попередня обробка різних типів даних в бінарний формат має важливе значення. Це включає перетворення неперервних змінних, дискретних категорій і інших форм даних у бінарне представлення. Неперервні змінні, такі як вік або дохід, які мають широкий діапазон значень, спочатку групуються або дискретизуються на категорії. Ці категорії після дискретизації перетворюються у бінарний формат за допомогою кодування одним з багатьма (OHE). Кожна категорія в OHE представлена бінарним вектором, де ’1’ розміщено в позиції, що відповідає категорії, і ’0’ призначено в інших місцях.

Для дискретних категоріальних даних, таких як стать чи країна, можливе пряме застосування OHE. Кожна унікальна категорія в цих змінних перетворюється на окрему бінарну ознаку. У випадку ординальних даних, де існує природний порядок (наприклад, рейтинги), рішення про збереження цього порядку впливає на підхід до бінаризації. Висока кардинальність категоріальних даних, яка включає велику кількість унікальних значень, вимагає ретельного вивчення. OHE в таких сценаріях може призвести до просторів високої розмірності, тому потребується використання технік бінування на основі частоти або інших методів зменшення розмірності.

Робота з відсутніми даними також є важливою. Варіанти включають в себе імпутацію, використання моделей, стійких до відсутніх значень, або призначення унікальної бінарної ознаки, яка позначає відсутність даних. Після бінаризації важливо переконатися, що масштаб цих бінарних ознак є сумісним із залишковими ознаками в моделі, щоб уникнути вигнутості важливості ознак. Ця трансформація часто призводить до високорозмірних даних, особливо після обширного застосування OHE, що може потребувати методів зменшення розмірності. Крім того, розрідженість даних, як загальний результат OHE, слід керувати за допомогою форматів даних, які

ефективно обробляють розріджені матриці. Через цю систематичну конвертацію в бінарний формат, моделі стають більш адаптивними і ефективними для різних наборів даних.

2.3 Оцінка та порівняння отриманих моделей

2.3.1 Оцінка точності у задачі

Крива характеристики робочої характеристики приймача (ROC) - це графік 2.1, який ілюструє діагностичну здатність бінарної системи класифікації при зміні порогу відокремлення. Площа під кривою (AUC) ROC є мірою здатності моделі розрізняти класи. В суті, вона надає одну міру, яка узагальнює продуктивність моделі у всіх можливих порогах класифікації.

Показник AUC особливо корисний в сценаріях з несбалансованими наборами даних, оскільки він не ґрунтується на конкретному порозі для класифікації. Модель з AUC, близьким до 1, вказує на велику роздільну здатність між класами, тоді як AUC, близький до 0,5, підказує відсутність роздільної сили, подібно до випадкового вибору.

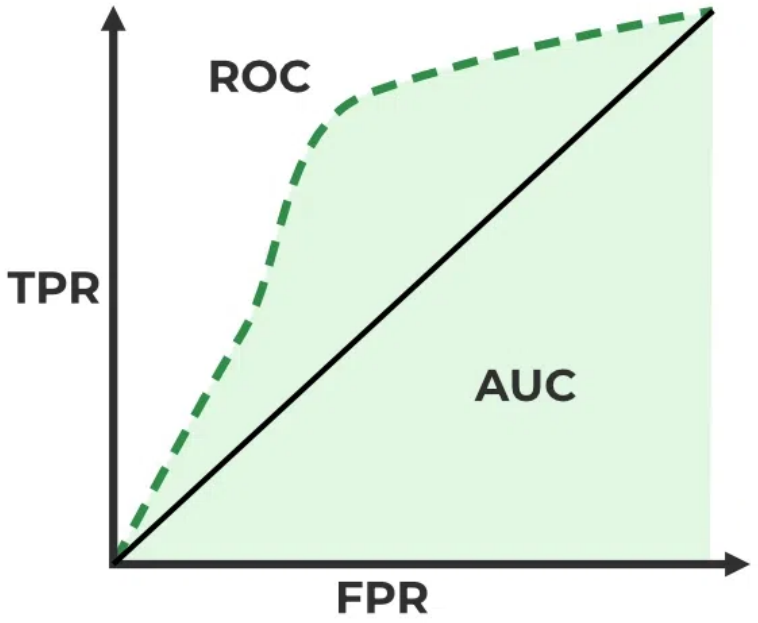

Рисунок 2.1 – Крива ROC

2.3.2 Стратегія валідації

У машинному навчанні, особливо в задачах класифікації, крос-валідація є стандартною технікою для оцінки того, як результати статистичного аналізу будуть узагальнюватися до незалежного набору даних. Стратифікована k-факторна крос-валідація, де k=10, є варіацією k-факторної крос-валідації, яка використовується переважно в задачах бінарної або багатокласової класифікації. Ця техніка розбиває дані на k підвибірок, забезпечуючи, що кожна підвибірка є відмінним представником всього набору даних. Особливо в стратифікованій крос-валідації, кожна підвибірка зберігає відсоток зразків для кожного класу, як відзначено у повному наборі даних, що є важливим для роботи з несбалансованими наборами даних.

Процес передбачає розділення даних на 10 різних піднаборів (підвибірок). Для кожної ітерації 9 підвибірок використовуються для навчання моделі, а решта одна - для тестування. Цей процес повторюється 10 разів, при цьому кожна підвибірка служить тестовим набором один раз. Резуль-

тати з 10 підвибірок потім усереднюються, щоб отримати одне значення оцінки. Цей підхід особливо ефективний у наданні більш точної міри прогностичної продуктивності моделі.

3 АНАЛІЗ ОТРИМАНИХ РЕЗУЛЬТАТІВ

3.1 Порівняння точності

Це дослідження має на меті не лише порівняти точність моделей дерев рішень, побудованих за допомогою цих критеріїв, а також розглянути структурні відмінності у моделях, які вони формують. Аналізуючи моделі з різних наборів даних, дослідження спрямоване на виявлення переваг, обмежень і застосовності кожного критерію в різноманітних сценаріях реального світу. Цей комплексний аналіз вносить свій внесок у активний діалог щодо оптимізації моделей дерев рішень, надаючи уявлення про потенційні переваги інтеграції ТПМ в традиційні методології дерев рішень.

Таблиця 3.1 – Порівняння точності

Метод	**Titanic**	**Malware**	**Credit Risk**
Приріст інформації	0.828	0.5658	0.5645
Зміна граничної області	0.841	0.5747	0.5787

Це покращення в наборі даних ”Titanic”можна пояснити тим, що критерій зміни граничної області ефективніше захоплює внутрішню структуру даних. Цей критерій враховує не тільки приріст інформації від розбиття за ознакою, але також оцінює розмір граничної області, яка утворюється в результаті розбиття. Ця двійкова увага допомагає вибирати ознаки, які більш чітко відокремлюють дані на класи.

3.2 Візуальне порівняння методів

Також важливим етапом аналізу результатів є проведена візуальна оцінка двох різних моделей дерев рішень для набору даних ”Titanic”. Ці візуалізації демонструють вплив інтеграції теорії приблизних множин на моделі дерев рішень. Була побудована візуалізація для обох критеріїв - приросту інформації (рис. 3.1) та зміни граничної області (рис. 3.2), що надає уявлення про відмінності у пріоритетах ознак за кожним з критеріїв.

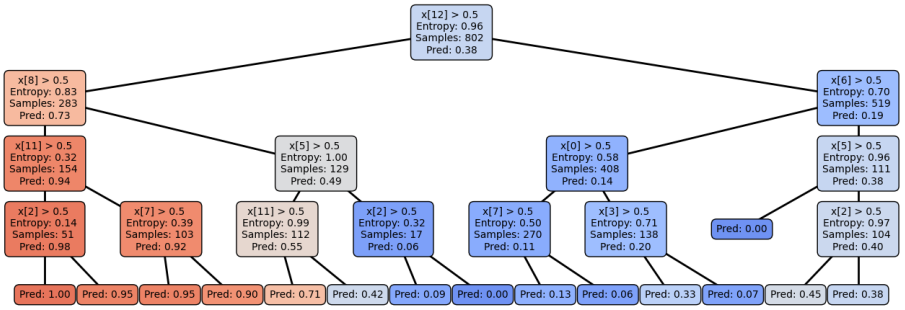

Рисунок 3.1 – Дерево рішень на базі ентропії

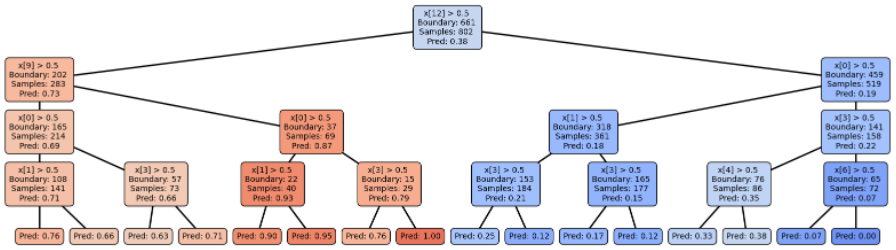

Рисунок 3.2 – Дерево рішень на базі зміни граничної області

Спільний візуальний порівняльний аналіз дерев рішень, побудованих з використанням критерію приросту інформації і критерію зміни граничної області, надав глибокі уявлення про їх відмінні методології. Стало зрозуміло, як кожен критерій впливає на пріоритетизацію ознак, що призводить до відмінних структур і глибини дерев. Дерева відрізнялися за складністю,

що підтверджувалося їх відповідним розгалуженням і глибиною, причому критерій зміни граничної області часто призводив до структури, яка, можливо, краще відображувала тонкі відмінності в даних.

Відстеження шляхів від кореня до листя обох дерев дозволило спостерігати вплив різних ознак на кінцевий результат класифікації. Цей аспект був особливо відкриттям для розуміння того, як кожна модель обробляє та класифікує дані за різних критеріїв. Крім того, візуалізації слугували важливим інструментом для оцінки інтерпретованості моделей. Хоча критерій зміни граничної області може пропонувати вищу точність, він може призвести до моделей, які менш інтуїтивні порівняно з тими, що базуються на прирості інформації.

Цей порівняльний візуальний аналіз підкреслив практичні наслідки застосування різних критеріїв для побудови дерев рішень. Він висвітлив не тільки нюансові відмінності в обробці та класифікації моделей, але й важливість таких аналізів у виборі та розробці моделей. Розуміння цих відмінностей має велике значення, особливо в галузях, де баланс між інтерпретованістю моделі та точністю є важливою складовою.

3.3 Аналіз кореляції між отриманими методами

У цьмоу розділі було проведено глибоке вивчення взаємозв’язку між змінами граничної області та ентропією при побудові дерев рішень. Цей аналіз був ключовим при оцінці взаємодії між новим критерієм зміни граничної області, отриманим за допомогою теорії приблизних множин (ТПМ), та традиційною мірою приросту інформації, яка використовується в деревах рішень.

Було створено три докладні графіки кореляції 3.3, один для кожного набору даних: ”Titanic ”Класифікація вредоносного програмного забезпечення”та ”Кредитний ризик за замовчуванням”. Ці графіки відображали зміни граничної області в залежності від значень ентропії для кожної озна-

ки в межах наборів даних. Основною метою було визначити, чи впливали зміни в граничній області, що свідчать про невизначеність або неоднозначність в класифікації даних, систематично впливають на ентропію, міру чистоти або безладу, яка використовується в традиційних алгоритмах дерев рішень.

Результати цього кореляційного аналізу були цікавими. Вони показали слабку кореляцію між змінами граничної області та ентропією в усіх трьох наборах даних. Це свідчить про те, що, хоча зміни в граничній області впливають на ентропію до певної міри, ефект не є однаковим або лінійним для різних типів ознак і наборів даних.

Така слабка кореляція вказує на те, що критерій зміни граничної області додає унікальний аспект до процесу прийняття рішень при побудові дерев, який не в повній мірі враховується традиційними мірами, що базуються на ентропії. Критерій зміни граничної області, здається, розкриває аспекти складності та неоднозначності даних, які ентропія сама по собі не повністю охоплює. Цей нюансований взаємозв’язок підкреслює складність інтеграції ТПМ з моделями дерев рішень, підкреслюючи потенціал цих моделей для виявлення більш складних закономірностей і взаємозв’язків в даних, які традиційні методи можуть упустити.

Кореляційний аналіз, отже, надає не лише емпіричні уявлення про взаємодію теорії приблизних множин і алгоритмів дерев рішень, але також відкриває нові можливості для подальших досліджень. Глибше розуміння цього взаємозв’язку може призвести до розробки більш вдосконалених моделей дерев рішень, які використовують переваги як ТПМ, так і традиційних методів класифікації. Ці моделі можуть потенційно забезпечити покращену точність, особливо в наборах даних, що характеризуються високим рівнем невизначеності або складності.

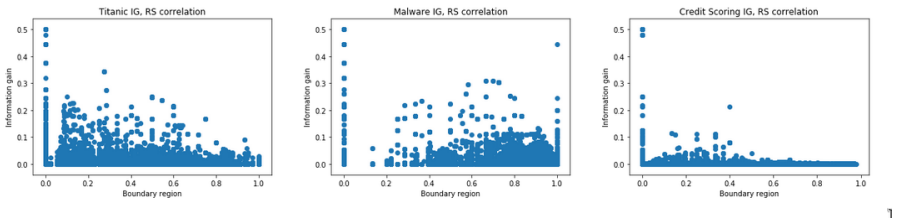

Рисунок 3.3 – Кореляція зміни граничної області та приросту інформації

3.4 Вплив шуму на гібрідний метод

У цьому розділі досліджуються вплив шуму [37, 38, 39] на механізми теорії приблизних множин. ТПМ фундаментально працює з поняттям нерозрізненості, яке об’єднує об’єкти, які не можна відрізнити за наявними ознаками. Шум, за своєю природою, змінює ці відносини нерозрізненості, впливаючи на нижні та верхні наближення, які є важливими для ТПМ.

Основу цього аналізу становить математична основа ТПМ, з фокусом на впливі шуму на матрицю розрізненості та формування редуктів - це мінімальні підмножини атрибутів, які зберігають потужність класифікації повного набору. Присутність шуму в ознаках може приховувати фактичні відносини між атрибутами та класами рішень, що може призвести до вибору незначущих або вводячих в оману атрибутів в редуктах.

У експериментальній фазі було використано бінаризовану версію набору даних ”Titanic з фокусом на 13 вибраних бінарних ознаках. Експеримент включав поступове додавання шумових ознак до основного набору даних і спостереження наступних змін у відносній граничній області. Для кращого візуального розуміння, відносна гранична область була представлена замість класичної граничної області, яка квантує об’єкти за (3.1). Отриманий графік (рис. 3.4) відображає вплив додавання N шумових ознак на граничну область. Крім того, був проведений аналіз у гіршому сценарії, використовуючи бінарний характер наших ознак. Як показано в (3.2), нерозрізненість може поділитися навпіл з додаванням кожної бінарної ознаки.

$$\text{Relative Boundary} = \frac{\text{Boundary Size}}{\text{Total Rows}} \tag{3.1}$$

$$\text{WorstBoundaryDecay}(x) = \frac{\text{WorstBoundaryDecay}(x-1)}{2} \tag{3.2}$$

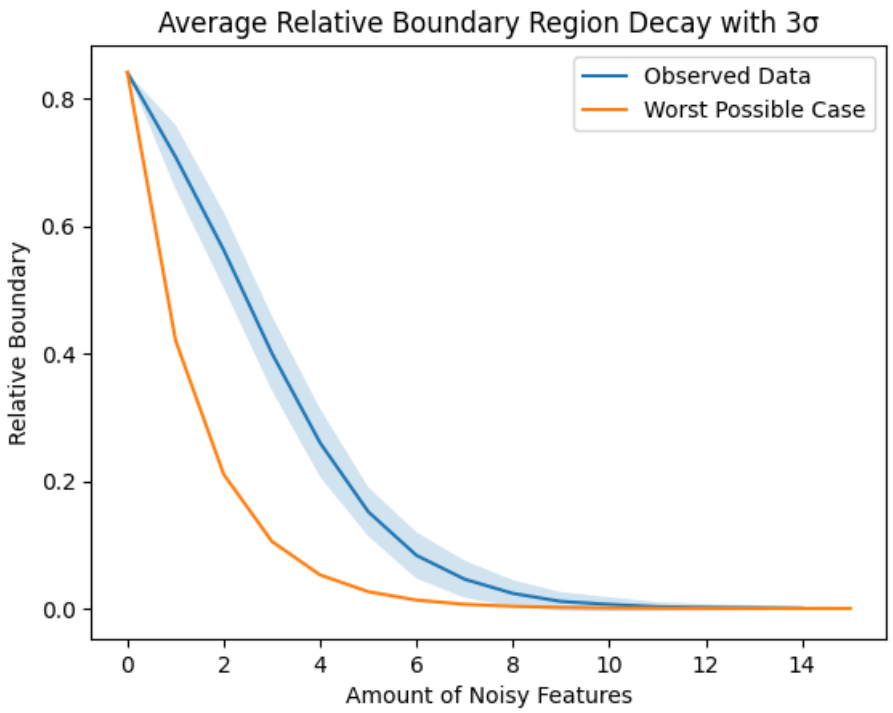

Рисунок 3.4 – Зменшення граничної області

Спостережені результати показують, що Середня Гранична Область відображає послідовний, монотонно спадаючий тренд, як і очікувалося. У іншому сценарії був проведений аналіз випадково створеного набору даних (рис. 3.5) з 512 вибірками. По-перше, розмір Граничної Області був значним, але почав відзначати помітне зменшення після певної межі. У класичній ТПМ, метрика залежності вимірює ступінь залежності класифікації об’єктів в одному наборі від атрибутів в іншому наборі. Шум може

порушити ці відносини залежності, що може призвести до неправильних припущень про значущість певних атрибутів.

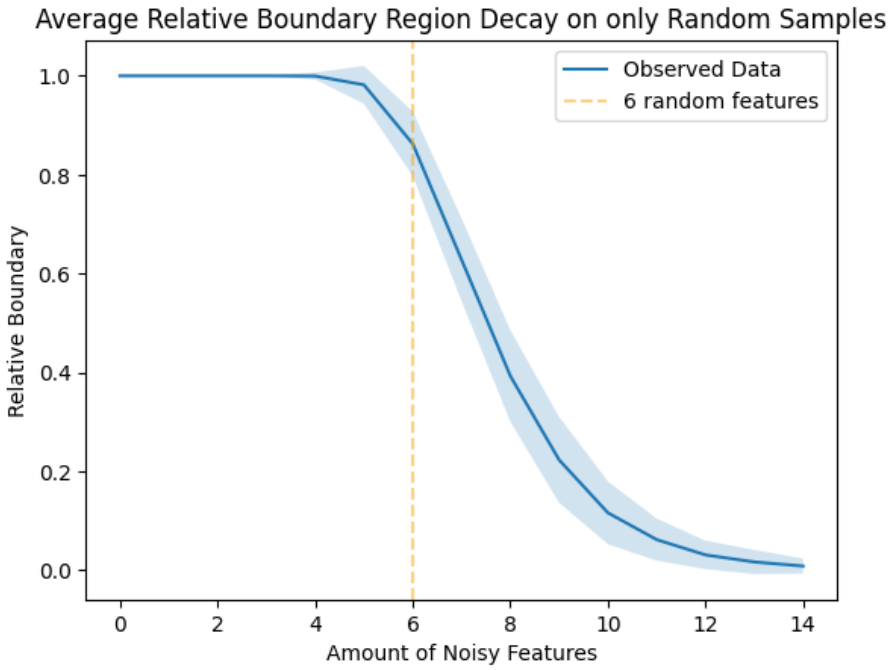

Рисунок 3.5 – Гранична область на повністю випадковому наборі даних

Рисунок 3.4 ілюструє Зменшення Середньої Відносної Граничної Області на 3σ, тоді як рисунок 3.5 показує Середню Відносну Граничну Область у Випадковому Наборі Даних.

Ця закономірність вказує на те, що гранична область у нашому випадково створеному наборі даних зменшується в межах від $(\log_2 M) - 3$ до $(\log_2 M) + 3$ в

ипадкових ознак, де M представляє розмір набору даних. Ця гіпотеза була підтверджена після проведення експериментів на наборах даних різного розміру.

Рисунок 3.6 демонструє Кількість Ознак, які Призводять до Нульової Граничної Області відносно Розміру Набору Даних.

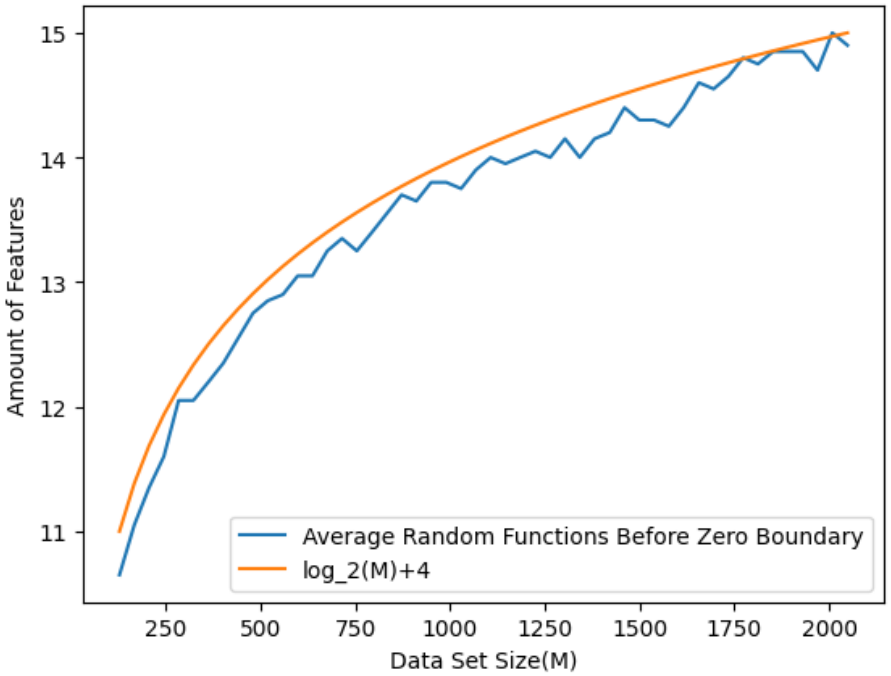

Рисунок 3.6 – Кількість ознак, що призводять до нульової граничної області відносно розміру набору даних

Теоретичний аналіз спрямований на вимірювання впливу шуму на ТПМ і надання уявлень про стійкість різних методів вибору ознак на основі ТПМ в умовах шумового середовища. Метою є отримання більш деталізованого розуміння сильних та слабких сторін класичної теорії приблизних множин у роботі з шумовими даними, що, в свою чергу, визначатиме майбутні застосування та дослідження в галузі управління шумом.

3.5 Подальші дослідження

Інтеграція теорії приблизних множин (ТПМ) з деревами рішень (ДР) представляє собою новий підхід у класифікації даних, який демонструє обіцяні результати у покращенні точності моделі та роботи з складними даними. Проте початкові дослідження, хоча і обнадійливі, відкривають декілька напрямків для подальших досліджень. Ці області не лише обіцяють поглибити наше розуміння цього гібридного підходу, але також мають по-

тенціал значно розширити галузь машинного навчання та аналізу даних.

3.5.1 Дослідження різних типів даних

Початкові дослідження були спрямовані на конкретні набори даних, кожен з унікальними характеристиками. Майбутні дослідження повинні спрямовуватися на застосування гібридної моделі ТПМ-ДР для більшого спектру наборів даних, включаючи ті, що мають вищі розмірності, різні рівні шуму та різноманітні області застосування. Це надасть більш всебічне розуміння можливостей та обмежень моделі.

3.5.2 Вдосконалені техніки відбору ознак

Хоча поточна модель використовує важливість ознак, отриману з ТПМ та міри ентропії, існує можливість дослідження більш вдосконалених методів відбору ознак. Техніки, які враховують взаємодію ознак та нелінійні зв’язки, можуть бути особливо корисними для покращення передбачувальної сили моделі.

3.5.3 Оптимізація параметрів моделі

Параметри, використовувані при побудові дерева рішень, такі як глибина дерева та мінімальна кількість вибірок у листі, мають значний вплив на продуктивність моделі. Дослідження оптимізаційних технік для цих параметрів, можливо, за допомогою алгоритмів машинного навчання, таких як генетичні алгоритми чи пошук по сітці, може привести до створення моделей, які є одночасно більш точними і ефективними.

3.5.4 Порівняльні дослідження з іншими моделями

Порівняння гібридної моделі ТПМ-ДР з іншими моделями класифікації, такими як Випадковий ліс, Метод опорних векторів чи Штучні нейронні мережі, може надати цінних висновків. Такі порівняльні дослідження допоможуть в позиціонуванні моделі ТПМ-ДР в загальному контексті алгоритмів машинного навчання.

3.5.5 Застосування в реальних сценаріях

Застосування моделі в реальних сценаріях, таких як діагностика у сфері охорони здоров’я, фінансовий прогноз або кібербезпека, дозволить перевірити її практичну застосовність. Це не лише підтвердить модель в реальних умовах, але також може призвести до значних практичних внесків у різних галузях.

3.5.6 Дослідження масштабованості

Питання масштабовості моделі при роботі з обсяжними наборами даних має важливе значення. Майбутні дослідження повинні вивчати, як модель працює з точки зору обчислювальної ефективності та використання пам’яті при обробці великих обсягів даних.

3.5.7 Вдосконалення реалізації теорії приблизних множин

Подальші дослідження щодо аспектів впровадження ТПМ у гібридну модель можуть розкрити більш ефективні або точні способи обчислення приблизних множин ТПМ, що призведе до покращення продуктивності моделі.

3.5.8 Моделі, які зрозумілі для користувачів

Одержання більш зрозумілих користувачам моделей важливо в багатьох сферах застосування. Дослідження можливостей зробити модель ТПМ-ДР більш зрозумілою для користувачів було б корисним. Це може включати розробку візуалізаційних інструментів чи методів для пояснення рішень моделі зрозумілою мовою.

ВИСНОВКИ

Ця раобота досліджувала інтеграцію дерев рішень та теорії приблизних множин (ТПМ) в задачі класифікації з фокусом на покращенні продуктивності та інтерпретованості моделей Дерев Рішень в умовах шумних та складних даних. Дослідження продемонструвало, як здатності ТПМ до відбору ознак та видобування правил можуть значно оптимізувати моделі дерев рішень, що призводить до більш точних та надійних результатів класифікації.

Підхід інтеграції вирішував ключові виклики, такі як перенавчання та чутливість до шуму в деревах рішень, підтверджуючи, що синергія між деревами рішень та ТПМ може створити моделі, які є не лише стійкими, але й мають високий рівень інтерпретованості. Це особливо важливо в прийнятті рішень на основі даних, де одночасно важлива як точність, так і розуміння рішень моделі.

Крім того, дослідження надало висновки щодо потенційних застосувань цього інтегрованого підходу в різних галузях, виокремлюючи його важливість у сучасному аналізі даних та процесах прийняття рішень. Знахідки відкривають шлях для майбутніх досліджень та розвитку в цій області, особливо в покращенні технік класифікації даних для роботи з ростучою складністю та обсягом даних у різних галузях.

У підсумку, це дослідження вносить значні внески у галузь науки про дані та машинного навчання, пропонуючи новаторський підхід, який поєднує переваги дерев рішень та ТПМ для вирішення деяких найважливіших завдань у класифікації даних.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Fisher R. A. ”The use of multiple measurements in taxonomic problems.”Annals of Eugenics, vol. 7, no. 2, 1936, pp. 179-188.

2. Hastie T., Tibshirani R., and Friedman J. ”The Elements of Statistical Learning: Data Mining, Inference, and Prediction.”Springer Series in Statistics, 2009.

3. James G., Witten D., Hastie T., and Tibshirani R. ”An Introduction to Statistical Learning.”Springer, 2013.

4. Witten I. H., Frank E., Hall M. A., and Pal C. J. ”Data Mining: Practical Machine Learning Tools and Techniques.”Morgan Kaufmann, 2016.

5. Molnar, C. (2020). Interpretable Machine Learning. Available at: https://christophm.github.io/interpretable-ml-book/.

6. Cover T. and Hart P. ”Nearest neighbor pattern classification.”IEEE Transactions on Information Theory, vol. 13, no. 1, 1967, pp. 21-27.

7. Quinlan J. R. ”Induction of decision trees.”Machine Learning, vol. 1, no. 1, 1986, pp. 81-106.

8. Cortes C. and Vapnik V. ”Support-vector networks.”Machine Learning, vol. 20, no. 3, 1995, pp. 273-297.

9. Goodfellow I., Bengio Y., and Courville A. ”Deep Learning.”MIT Press, 2016.

10. Kourou K., Exarchos T. P., Exarchos K. P., Karamouzis M. V., and Fotiadis D. I. ”Machine learning applications in cancer prognosis and prediction.”Computational and Structural Biotechnology Journal, vol. 13, 2015, pp. 8-17.

11. Ngai E. W., Hu Y., Wong Y. H., Chen Y., and Sun X. ”The application of data mining techniques in financial fraud detection: A

classification framework and an academic review of literature.”Decision Support Systems, vol. 50, no. 3, 2011, pp. 559-569.

12. Fan S. and Gordon M. D. ”The power of social media analytics.”Communications of the ACM, vol. 57, no. 6, 2014, pp. 74-81.

13. Lu D. and Weng Q. ”A survey of image classification methods and techniques for improving classification performance.”International Journal of Remote Sensing, vol. 28, no. 5, 2007, pp. 823-870.

14. Janai J., Güney F., Behl A., and Geiger A. ”Computer Vision for Autonomous Vehicles: Problems, Datasets and State-of-the-Art.”arXiv preprint arXiv:1704.05519, 2017.

15. Breiman L., Friedman J. H., Olshen R. A., and Stone C. J. ”Classification and Regression Trees.”Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

16. Rokach L. and Maimon O. ”Data Mining with Decision Trees: Theory and Applications.”World Scientific Pub Co Inc, 2005.

17. Quinlan J. R. ”C4.5: Programs for Machine Learning.”Morgan Kaufmann, 1993.

18. Pawlak Z. ”Rough sets.”International Journal of Computer & Information Sciences, vol. 11, no. 5, 1982, pp. 341-356.

19. Pawlak Z. ”Rough Sets: Theoretical Aspects of Reasoning about Data.”Kluwer Academic Publishers, 1991.

20. Skowron A. and Rauszer C. ”The discernibility matrices and functions in information systems.”In Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory, pp. 331-362, Springer, 1992.

21. Komorowski J., Pawlak Z., Polkowski L., and Skowron A. ”Rough sets: A tutorial.”In Rough-Fuzzy Hybridization: A New Trend in Decision-Making, pp. 3-98, Springer, Singapore, 1999.

22. Grzymala-Busse J. W. ”A new version of the rule induction system LERS.”Fundamenta Informaticae, vol. 31, no. 1, 1997, pp. 27-39.

23. Wang Z., Zhang X., and Deng J. ”The uncertainty measures for covering rough set models.”Soft Computing, Springer Science and Business Media LLC, vol. 24, no. 16, 2020, pp. 11909–11929, DOI: 10.1007/s00500-020-05098-x.

24. Geetha M. A., Acharjya D. P., and Iyengar N. Ch. S. N. ”Algebraic properties and measures of uncertainty in rough set on two universal sets based on multi-granulation.”Proceedings of the 6th ACM India Computing Convention, ACM, 2013, DOI: 10.1145/2522548.2523168.

25. Qian Y., Xu H., Liang J., Liu B., and Wang J. ”Fusing Monotonic Decision Trees.”IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 10, 2015, pp. 2717–2728, DOI: 10.1109/TKDE.2015.2429133.

26. Sitnikov D., Titova O., Romanenko O., and Ryabov O. ”A method for finding minimal sets of features adequately describing discrete information objects.”Data Mining X, WIT Press, 2009, DOI: 10.2495/data090141.

27. Sitnikov D. and Ryabov O. ”An Algebraic Approach To Defining Rough Set Approximations And Generating Logic Rules.”Data Mining V, WIT Press, 2004, DOI: 10.2495/data040171.

28. D. Sitnikov, O. Titova, O. Romanenko, and O. Ryabov, ”An approach to finding reduced sets of information features describing discrete objects based on rough sets theory,”Data Mining IX, WIT Press, 2008, DOI: 10.2495/data080011.

29. D. Chernyshov and D. Sytnikov, “Binary classification based on a combination of rough set theory and decision trees,” Innovative Technologies and Scientific Solutions for Industries, no. 4(26), Kharkiv National University of Radioelectronics, 2023, pp. 87–94, DOI: 10.30837/itssi.2023.26.087.

30. Blockeel H., Devos L., Frénay B., Nanfack G., and Nijssen S. ”Decision trees: from efficient prediction to responsible AI.”Frontiers in Artificial Intelligence, Frontiers Media SA, vol. 6, 2023, DOI: 10.3389/frai.2023.1124553.

31. Hu X., Rudin C., and Seltzer M. ”Optimal Sparse Decision Trees.”arXiv, 2019, DOI: 10.48550/ARXIV.1904.12847.

32. Chiaselotti G., Gentile T., and Infusino F. ”Decision systems in rough set theory: A set operatorial perspective.”Journal of Algebra and Its Applications, World Scientific Pub Co Pte Lt, vol. 18, no. 01, 2019, pp. 1950004, DOI: 10.1142/s021949881950004x.

33. Xu J., Qu K., Meng X., Sun Y., and Hou Q. ”Feature selection based on multiview entropy measures in multiperspective rough set.”International Journal of Intelligent Systems, Hindawi Limited, vol. 37, no. 10, 2022, pp. 7200–7234, DOI: 10.1002/int.22878.

34. Cukierski, W. (2012). Titanic - Machine Learning from Disaster. Kaggle. Available at: https://kaggle.com/competitions/titanic.

35. Ronen R., Radu M., Feuerstein C., Yom-Tov E., and Ahmadi M. ”Microsoft Malware Classification Challenge.”arXiv, 2018, DOI: 10.48550/ARXIV.1802.10135.

36. Montoya, A., Odintsov, K., and Kotek, M. (2018). Home Credit Default Risk. Kaggle. Available at: https://kaggle.com/competitions/home-credit-default-risk.

37. Чернишов, Д., Ситніков, Д. (2023). ДОСЛІДЖЕННЯ ВПЛИВУ ШУМУ В ТЕОРІЇ ПРИБЛИЗНИХ МНОЖИН. Матеріали конференцій МЦНД, 22.12.2023; Одеса, Україна, с. 302–304. Вилучено із https://archive.mcnd.org.ua/index.php/conference-proceeding/article/view/953.

38. Chernyshov, D., Sytnikov, D., (2023). Exploring the Effects of

Noise in Rough Set Theory. In Proceedings of the XIV International Scientific and Practical Conference, International Science Group, 19.12.2023. DOI: 10.46299/isg.2023.2.14.

39. Chernyshov, D., Sytnikov, D. (2023). Noise Impact in Classical Rough Set Theory: Challenges and Insights. In Abstracts of the 10th International Conference ”Information Technology and Implementation (Satellite) – 2023”. ISBN 978-960-801-907-4.