Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерних наук

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

на тему *Метод вибору важливих ознак з використанням мурашиного алгоритму*

Галузь знань *12 – Інформаційні технології*
Шифр і назва галузі знань

Спеціальність *122 – Комп'ютерні науки*
Шифр і назва спеціальності

Освітня програма *Комп'ютерні науки*
Назва освітньої програми

Виконав: *студент 4 курсу, група КН-20-1* *Назарій РЕПІНСЬКИЙ*
Курс, група навчання — Підпис — Ініціали, прізвище

Керівник: *д.т.н., професор кафедри КН* *Едуард МАНЗЮК*
Науковий ступінь, посада — Підпис — Ініціали, прізвище

Нормоконтроль: *к.т.н., доцент кафедри КН* *Руслан БАГРІЙ*
Науковий ступінь, посада — Підпис — Ініціали, прізвище

До захисту допускаю:
Зав. кафедри КН, д.т.н., професор *Олександр БАРМАК*
Підпис — Ініціали, прізвище

17 06 2024 р.

Хмельницький 2024

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет *інформаційних технологій*
Кафедра *комп'ютерних наук*
Освітній ступінь *бакалавр*
Галузь знань *12 – Інформаційні технології*
Спеціальність *122 – Комп'ютерні науки*
Освітня програма *освітньо-професійна програма підготовки бакалавра*

ЗАТВЕРДЖУЮ
Завідувач кафедри *комп'ютерних наук*

(підпис)
д.т.н., професор Олександр БАРМАК
«16» 02 2024 року

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ БАКАЛАВРА

1. Тема кваліфікаційної роботи бакалавра: *«Метод вибору важливих ознак з використанням мурашиного алгоритму»*

2. Завдання видано студенту *Назарію РЕПІНСЬКОМУ*
(прізвище, ім'я, по батькові)

3. Керівник роботи *д.т.н. професор кафедри КН Едуард МАНЗЮК*
(посада, прізвище, ім'я, по батькові)

4. Затверджено наказом університету від «15» 02 2024 р. № 8

5. Дата видачі завдання студенту: «16» 02 2024 р.

6. Зміст пояснювальної записки (перелік задач) та вихідні дані:

Провести аналіз предметної області, оглянути методи, які використовуються для вибору важливих ознак з використанням мурашиного алгоритму. Розробити Метод вибору важливих ознак з використанням мурашиного алгоритму. Оцінити ефективність застосування цього методу для вибору важливих ознак. Розробити експериментальну систему перевірки ефективності запропонованого методу для проведення оцінювання заякісними показниками класифікації даних на основі визначених ознак .

7. Календарний план виконання кваліфікаційної роботи бакалавра:

№	Назва етапів (розділів) кваліфікаційної роботи бакалавра	Термін виконання	Примітка
1	*Вибір напряму дослідження та узгодження тематики кваліфікаційної роботи бакалавра з керівником*	*грудень 2023*	*виконано*
2	*Ознайомлення з предметною областю, формулювання мети та задач дослідження, визначення об'єкта та предмета дослідження*	*січень 2024*	*виконано*
3	*Робота над розділом 1 – Характеристика предметної області та постановка задачі*	*січень 2024*	*виконано*
4	*Робота над розділом 2 – Метод вибору важливих ознак з використанням мурашиного алгоритму*	*березень 2024*	*виконано*
5	*Робота над розділом 3 – Експериментальна перевірка методу вибору важливих ознак з використанням мурашиного алгоритму*	*квітень 2024*	*виконано*
6	*Оформлення пояснювальної записки згідно вимог*	*травень 2024*	*виконано*
7	*Попередній захист кваліфікаційної роботи бакалавра*	*травень 2024*	*виконано*
8	*Захист кваліфікаційної роботи бакалавра на засіданні Екзаменаційної комісії*	*червень 2024*	*виконано*

Виконавець: *студент 4 курсу, група КН-20-1* _____ *Назарій РЕПІНСЬКИЙ*
Курс, група виконавця — Підпис — Ініціали, прізвище

Керівник: *д.т.н. професор кафедри КН* _____ *Едуард МАНЗЮК*
Науковий ступінь, посада — Підпис — Ініціали, прізвище

Анотація

Тема кваліфікаційної роботи бакалавра: *Метод вибору важливих ознак з використанням мурашиного алгоритму.*

Виконавець кваліфікаційної роботи бакалавра: *студент групи КН-20-1 Назарій РЕПІНСЬКИЙ*

Керівник кваліфікаційної роботи бакалавра: *д.т.н. професор кафедри КН Едуард МАНЗЮК*

Кваліфікаційна робота бакалавра містить:

Пояснювальна записка				Кількість додатків
Сторінок	Рисунків	Таблиць	Джерел інформації	
77	7	1	40	1

Мета кваліфікаційної роботи бакалавра полягає в покращенні вибору важливих ознак з використанням мурашиного алгоритму.

Для досягнення поставленої мети визначені наступні задачі дослідження: визначити послідовність застосування методу вибору важливих ознак з використанням мурашиного алгоритму; розробити метод вибору важливих ознак з використанням мурашиного алгоритму; провести експериментальні дослідження ефективності розробленого методу.

Результатом виконання кваліфікаційної роботи бакалавра є розроблений метод вибору важливих ознак з використанням мурашиного алгоритму.

Ключові слова: мурашиний алгоритм, отримання важливих ознак, аналіз текстових даних, машинне навчання.

Виконавець: *студент 4 курсу, група КН-20-1* _____ *Назарій РЕПІНСЬКИЙ*

Курс, група виконавця | Підпис | Ініціали, прізвище

Зміст

Перелік скорочень 6
Вступ 7
Розділ 1 Характеристика предметної області та постановка задачі 9
1.1 Аналіз предметної області вибору ознак у різних сферах практичного застосування 9
1.2 Методи та підходи до вибору та обробки ознак 14
1.3 Мета та постановка задачі 19
Розділ 2 Метод вибору важливих ознак з використанням мурашиного алгоритму 20
2.1 Оптимізація мурашиного алгоритму 20
2.2 Евристика вибору шляху мурахи для руху за графом 28
2.3 Опис застосування мурашиного алгоритму 40
2.4 Метод вибору важливих ознак з використанням мурашиного алгоритму .. 44
2.5 Оцінювання вибору ознак 52
Висновки до розділу 2 57
Розділ 3 Експериментальна перевірка методу вибору важливих ознак з використанням мурашиного алгоритму 58
3.1 Підготовка до проведення експериментальних досліджень ефективності методу визначення важливих ознак 58
3.2 Множина даних для тестування 60
3.3 Проведення екпериметнальних досліджень розробленого методу 62
Висновки до розділу 3 69
Висновок 70
Перелік посилань 71
Додатки

Перелік скорочень

Скорочення, термін, позначення	Пояснення
КРБ	Кваліфікаційна робота бакалавра
КН	Комп’ютерні науки
СМК	Система мурашиних колоній
BS	Найкраще рішення в цей момент
MA	Мурашиний алгоритм
AS	Муришина система

Вступ

Вибір ознак є ключовим кроком у класифікації веб-сторінок, особливо з великими наборами даних, що містять десятки або сотні тисяч ознак. Зменшення розмірності цих наборів даних стає важливим завданням для подальшого аналізу і моделювання. Однак просте зменшення кількості ознак не завжди призводить до ефективної класифікації веб-сторінок.

Метод вибору ознак, який представлено в кваліфікаційній робті, базується на оптимізації мурашиної колонії. Цей метод використовує метаевристику на основі популяції для оптимізації витягнутих ознак з веб-сторінок. для цього алгоритму надали поведінка реальних мурах, які працюють в паралельних конструктивних потоках, базуючись на локальних даних і динамічній структурі пам'яті.

У цьому методі мурахи, які є віртуальними агентами, поступово формують рішення, пересуваючись по зваженому графу. Вони використовують стохастичний процес, що залежить від моделі феромонів, для побудови рішень. Цей процес легко реалізується і має низьку обчислювальну складність, що робить його ефективним і практичним для застосування в задачах класифікації веб-сторінок.

Штучні мурахи імітують поведінку реальних мурах, створюючи рішення на основі інформації про ознаки веб-сторінок. Вони використовують структуровану пам'ять, що містить інформацію про якість отриманих результатів, для направленого пошуку найкращих варіантів. Цей процес дозволяє визначити набір найбільш інформативних ознак, зберігаючи при цьому оптимальну розмірність для подальшої обробки.

Основна перевага цього підходу полягає у його здатності ефективно працювати з великими обсягами даних, де традиційні методи стикаються з проблемами обробки. Використання метаевристичних методів дозволяє знайти оптимальне рішення в складних просторах ознак, що робить цей підхід особливо важливим для сучасних систем класифікації веб-сторінок.

Розроблений метод є легким у реалізації, що дозволяє ефективно використовувати його в практичних застосуваннях. Його застосування може покращити якість класифікації веб-сторінок і сприяти розвитку більш точних та ефективних систем аналізу інформації.

Об’єкт дослідження – процес вибору важливих ознак з використанням мурашиного алгоритму.

Предмет дослідження – методи та технології для визначення та вибору важливих ознак.

Мета кваліфікаційної роботи бакалавра – покращення вибору важливих ознак з використанням мурашиного алгоритму.

Завдання кваліфікаційної роботи бакалавра.

Для досягнення цієї мети виконуються наступні завдання.

1. Проведення детального аналізу існуючих методів обробки текстів з метою вибору оптимальних підходів до виділення важливих ознак.

2. Вибір і визначення набору ознак, які потрібно аналізувати для класифікації текстів. Це можуть бути ключові слова, структурні особливості, статистичні параметри тощо.

3. Розробка методу, який моделює процес визначення важливих ознак на основі поведінки мурахи. Метод повинен враховувати взаємодію між мурашиними агентами в даному випадку, методами обробки текстів та здатність адаптуватися до змін у текстових даних.

4. Розробка та тренування моделі класифікаці на основі визначених ознак за допомогою мурашиного алгоритму. Після цього проводиться оцінка ефективності моделі на тестовому наборі даних для визначення її точності та надійності в класифікації текстів.

Розділ 1 Характеристика предметної області та постановка задачі

1.1 Аналіз предметної області вибору ознак у різних сферах практичного застосування

Вибір ознак є важливим етапом у машинному навчанні, особливо коли маємо справу з великою кількістю можливих функцій. Його мета полягає в тому, щоб спростити дані, зменшити їх розмірність і виокремити головні ознаки без втрати точності передбачення. Це допомагає усунути зайву інформацію та відокремити важливі характеристики. Вибір ознак широко використовується в різних галузях, таких як обробка тексту, аналіз даних, розпізнавання зображень і обробка сигналів [1–4].

Особливо важливою є категоризація текстів, яка дозволяє ефективно обробляти і систематизувати великі обсяги документів. Однак велика кількість можливих ознак у тексті може стати проблемою. Багато з цих ознак не мають прямого відношення до категоризації тексту і можуть навіть впливати негативно на результати класифікації [5–8]. Отже, потрібно відібрати найважливіші ознаки зі збільшеного набору даних, щоб зменшити його обсяг і покращити продуктивність класифікатора.

Одним із ключових завдань вибору ознак у контексті категоризації тексту є виявлення та виключення нерелевантних або шумових ознак. Це дозволяє покращити якість моделі, оскільки зайві або непотрібні ознаки можуть призводити до перенавчання або погіршення загальної ефективності.

При виборі ознак для категоризації тексту використовуються різні методи, такі як статистичні тестування наприклад, аналіз частот слів, методи зменшення розмірності наприклад, метод головних компонент аналізу або алгоритми вибору ознак на основі моделей машинного навчання наприклад, важливість ознак у випадковому лісі [9–11].

Після вибору оптимального піднабору ознак можна побудувати більш ефективні моделі для класифікації тексту, що здатні здійснювати більш точні та швидкі прогнози. Цей підхід дозволяє зосередитися на суттєвих аспектах тексту і покращує інтерпретованість та продуктивність аналізу.

В області вибору ознак для категоризації тексту використовуються різні підходи. У дослідженнях проведено порівняльний аналіз критеріїв вибору ознак, включаючи частоту документа, інформаційний приріст, взаємну інформацію, статистику хі-квадрат [12–14]. Виявлено, що статистика хі-квадрат і інформаційний приріст є найефективнішими для оптимізації результатів класифікації, тоді як частота документа добре підходить для забезпечення ефективності і масштабованості при невеликому зниженні ефективності.

У іншому дослідженні представлено три методи – центроїд, ортогональний центроїд та лінійний дискримінантний аналіз, узагальнене сингулярне розкладання для зменшення розмірності даних у кластеризації, які також можуть бути застосовані для вибору ознак у категоризації тексту [15–18]. Серед різних методів для вибору ознак особливою увагою користуються алгоритми оптимізації на основі популяції, такі як генетичні алгоритми і алгоритми оптимізації колонії мурашок [19–22]. Ці методи спрямовані на досягнення кращих рішень, використовуючи знання з попередніх ітерацій.

Під час вибору ознак для категоризації тексту важливо звертати увагу на методи, які допоможуть ефективно зменшити розмірність даних, виокремити значущі ознаки та покращити результати класифікації. Наприклад, методи статистики, такі як статистика хі-квадрат і інформаційний приріст, дозволяють ідентифікувати ознаки, що найбільше сприяють розрізненню між класами текстів.

Інші підходи, такі як використання методів зменшення розмірності даних наприклад, центроїд, ортогональний центроїд, спрямовані на збереження суттєвої інформації при зменшенні кількості ознак. Це допомагає уникнути перенавчання і покращує загальну продуктивність моделі.

Алгоритми оптимізації на основі популяції, такі як генетичні алгоритми і алгоритми оптимізації колонії мурашок, ставлять своїми цілями знаходження найкращих підмножин ознак шляхом ітеративного пошуку оптимальних рішень. Ці методи є потужним інструментом для вирішення складних завдань вибору ознак у контексті категоризації тексту.

Загалом, ефективний вибір ознак в категоризації тексту допомагає зробити моделі більш точними, інтерпретованими і ефективними у роботі з великим обсягом даних.

Генетичні алгоритми використовуються як метод оптимізації, що моделює процес природного відбору. Вони використовують основні принципи генетики для навігації в просторі пошуку й знаходження оптимальних рішень. Ці алгоритми вже давно застосовуються у сфері інтелектуального аналізу даних, оскільки вони дозволяють ефективно вибирати найкращі ознаки для моделей.

Ще одним цікавим підходом є алгоритм метаевристичної оптимізації, що базується на поведінці мурах. Цей метод, відомий як алгоритм мурашиного колонії, належить до галузі штучного інтелекту, що вивчає поведінку кооперативних агентів. Ройовий інтелект представляє собою інноваційний підхід до вирішення проблем, що використовує взаємодію між агентами для досягнення оптимальних результатів [23–26].

Ройовий інтелект, або алгоритми, що базуються на поведінці мурах, представляють собою область штучного інтелекту, яка моделює співпрацю між агентами у природі. У цій методології агенти спілкуються і обмінюються інформацією, щоб досягти спільної мети. Обчислювальні реалізації цих алгоритмів, такі як алгоритм мурашиного колонії, застосовуються для розв'язання складних задач оптимізації та пошуку шляхів у просторі параметрів.

МА використовує метафору поведінки мурах для пошуку оптимальних рішень у просторі проблеми. Ці алгоритми дозволяють моделювати процес вибору оптимального шляху у навколишньому середовищі, де кожен агент або мураха комунікує з іншими для побудови оптимального шляху до джерела їжі або рішення.

Застосування ройових алгоритмів у сучасних комп'ютерних системах дозволяє ефективно розв'язувати проблеми оптимізації, вибору ознак у машинному навчанні, а також інші завдання, де необхідна глобальна оптимізація при наявності багатьох варіантів і наборі обмежень.

Колективна поведінка простих агентів, що взаємодіють зі своїм оточенням, може призводити до виникнення узгоджених глобальних моделей. Наприклад, у групах комах, які утворюють колонії, таких як мурахи або бджоли, окрема особина може виконувати лише прості завдання. Проте спільна дія колонії дозволяє виявити розумне поведінку, яка виявляється у вирішенні складних завдань, наприклад, знаходженні оптимального шляху до джерела їжі чи гнізда. Алгоритм мурашиного колонії на основі соціальної поведінки мурашок. Навіть без зору, мурахи можуть ефективно знаходити найкоротший шлях, використовуючи хімічні речовини, які вони залишають за собою під час переміщення, що відомо як феромони. Ця колективна поведінка базується на простих правилах взаємодії між окремими агентами, які взаємодіють лише локально, але разом вони вирішують складні задачі. Наприклад, у колонії мурах кожна мураха слідує певним правилам наприклад, залишаючи феромони, які сприяють знаходженню оптимальних шляхів до ресурсів чи місць знаходження.

Такий підхід до моделювання базується на природних процесах і використовується у різних галузях, включаючи комп'ютерні науки та штучний інтелект. Алгоритми, які базуються на соціальному поводженні комах, демонструють ефективність у розв'язанні проблем оптимізації, де важко враховувати всі можливі варіанти і шукати найкраще рішення.

Це інноваційний підхід до розв'язання складних завдань за допомогою простих агентів, які співпрацюють у колективі, і може мати застосування у багатьох сферах, де важко вирішувати проблеми індивідуально, але можливо колективно досягти успіху.

Алгоритм мурашиного колонії спочатку застосовувався для вирішення проблеми комівояжера і пізніше успішно застосовувався до різних складних завдань, таких як квадратична проблема призначення, маршрутизація в телекомунікаційних мережах, фарбування графів та планування [27–29]. Цей метод є привабливим для вибору функцій, оскільки він не вимагає наявності жорстких евристик, які б керували процесом пошуку оптимального мінімального піднабору ознак.

Останні дослідження в області використання МА стосуються застосування цієї методики до проблеми вибору ознак у категоризації тексту. Запропонований модифікований алгоритм вибору функцій на основі МА використовує продуктивність класифікатора та довжину вибраного піднабору ознак як критерії для пошуку оптимального рішення. Цей підхід дозволяє використовувати алгоритм без попередніх знань про функції. У запропонованій методиці застосувано текстові ознаки за моделлю мішок слів, де кожен документ розглядається як набір слів або фраз, і кожен елемент у векторі вхідних ознак відповідає певному терміну в оригінальному тексті [30, 31].

Запропонований модифікований алгоритм вибору функцій на основі МА використовується для ефективного вибору набору ознак у текстових даних. Основна ідея полягає в тому, що продуктивність класифікатора і довжина обраного піднабору ознак використовуються як метрики для оцінки якості вибору ознак.

Основні переваги використання алгоритму МА для вибору ознак у категоризації тексту полягають у його здатності знаходити оптимальні підмножини ознак без необхідності заздалегідь заданої евристики. Використовуючи інформацію про продуктивність класифікатора як метрику успішності, алгоритм може поступово покращувати вибір ознак і збільшувати точність моделі.

Модель мішок слів, що використовується у цій методиці, дозволяє представити кожен текст як вектор ознак, де кожна компонента відповідає наявності або відсутності певного терміну в документі [32–34]. Цей підхід дозволяє ефективно працювати з текстовими даними і здійснювати вибір ознак для покращення процесу класифікації.

У цілому, застосування методу МА до вибору ознак у категоризації тексту є інноваційним підходом, що дозволяє автоматизувати і поліпшити процес вибору найбільш важливих ознак для побудови ефективних моделей машинного навчання на основі текстових даних.

1.2 Методи та підходи до вибору та обробки ознак

Вибір ознак є складною задачею дискретної оптимізації, оскільки включає пошук оптимальних підмножин функцій з усього простору можливих комбінацій. Розмір цього простору визначається кількістю всіх можливих підмножин, що робить задачу вибору ознак великою і складною.

Коли вибір функцій і алгоритм навчання взаємодіють, алгоритм вибору ознак може бути розглянутий як один із варіантів вбудованого підходу. Хоча методи, які використовують обгортки, можуть дати кращі результати, вони можуть бути вимогливими у використанні та нестабільними через велику кількість функцій. Використання алгоритмів навчання для оцінки підмножин може бути важким завданням, особливо з великими обсягами даних, через що деякі з них можуть стикатися з труднощами

Вибір оптимального підмножини ознак є ключовою складовою в багатьох задачах аналізу даних і машинного навчання. Основна мета полягає в тому, щоб знайти найбільш інформативні функції або ознаки, які сприятимуть покращенню точності і ефективності моделі, використовуючи якнайменшу кількість ознак.

Простір всіх можливих підмножин ознак дуже великий, що робить задачу вибору ознак обчислювально складною. Традиційні підходи до вибору ознак використовують різні стратегії, такі як фільтрація, обгорткові методи та вбудовані методи [35–37].

Методи фільтрації оцінюють ознаки незалежно від моделі і використовують статистичні метрики, такі як взаємна інформація чи кореляція, для відбору найбільш інформативних ознак. Обгорткові методи використовують конкретну модель машинного навчання для оцінки кожного підмножини ознак, що може бути дуже обчислювально витратним. Вбудовані методи використовують сам процес навчання моделі для визначення важливості ознак під час тренування.

Алгоритми вибору ознак дозволяють ефективно зменшити розмірність даних, зберігаючи при цьому найважливішу інформацію для класифікації чи

прогнозування. Вибір правильних ознак може покращити якість моделі, роблячи її більш зрозумілою, швидшою і менш схильною до перенавчання або перетренування [38–40].

В підході обгортки для вибору ознак використовується функція оцінки, яка визначає придатність підмножини ознак, що створюється процедурою генерації. Ця функція порівнює новостворене підмножина з попереднім найкращим кандидатом і, якщо вона краща, замінює його. Критерій зупинки перевіряється під час кожної ітерації, щоб визначити, чи слід продовжувати процес вибору ознак.

У цьому підході існують п'ять основних методів.

1. Прямий вибір, який починається з порожнього набору і додає ознаки жадібно одна за одною.

2. Зворотне усунення, який починається з набору, що містить усі доступні ознаки, і ознаки по черзі видаляються жадібно.

3. Поступовий вибір уперед, який починається з порожнього набору, і ознаки додаються або видаляються жадібно одна за одною.

4. Зворотне поетапне усунення, який починається з набору, що містить усі доступні ознаки, і ознаки додаються або видаляються жадібно одна за одною.

Ці методи дозволяють ефективно вибирати найбільш важливі ознаки для моделі, а критерій зупинки допомагає визначити оптимальний момент завершення процесу вибору ознак.

У підході обгорткового вибору ознак існує кілька основних методів, кожен з яких використовує різні стратегії для побудови оптимального підмножини ознак:

1. Прямий вибір. Цей метод починає з порожнього набору ознак і додає їх по черзі, вибираючи ті, які мають найбільший вплив на покращення моделі. Це може бути ефективною стратегією, коли кількість ознак невелика і можливо важко підібрати оптимальне підмножину з великого простору ознак.

2. Зворотне усунення. У цьому методі починають з набору, що містить усі доступні ознаки, і поступово видаляють ознаки одну за одною. Видаляються ті

ознаки, які мають найменший вплив на модель або є менш важливими для класифікації.

3. Поступовий вибір уперед. Цей метод також починає з порожнього набору і поступово додає ознаки. Однак він може включати також видалення або зміну ознак, щоб побудувати оптимальне підмножину.

4. Зворотне поетапне усунення. Тут починають з набору, що містить усі ознаки, і поступово додають або видаляють ознаки, оптимізуючи при цьому продуктивність моделі.

Ці методи дозволяють систематично досліджувати простір ознак і знаходити найкращі комбінації для досягнення високої якості моделі. Кожен метод має свої переваги і недоліки залежно від конкретних вимог завдання та характеристик даних. Такий підхід до вибору ознак є важливим етапом в розв'язанні багатьох завдань машинного навчання, де ефективне використання інформації може підвищити якість моделі і її придатність до реальних задач.

Існують різні підходи до вибору ознак у задачах машинного навчання. Один з таких підходів - це випадкова мутація, де починають з набору випадково вибраних функцій і додають або видаляють їх по одній протягом певної кількості ітерацій.

Крім того, існують інші відомі методи, такі як генетичні алгоритми, оптимізація рою частинок, оптимізація мурашиної колонії. Ці методи використовуються для вибору найкращих підмножин ознак у різних завданнях.

Наприклад, в одному дослідженні було запропоновано процедуру пошуку підмножини на основі МА для задачі класифікації мови. У іншому дослідженні використовувався гібрид МА та взаємної інформації для вибору ознак у прогнозуванні. Крім того, МА використовується для пошуку грубих редуктів та інші стратегії оновлення феромонів в різних дослідженнях.

Ці методи вибору ознак є важливими в задачах аналізу даних і допомагають покращити якість моделей шляхом вибору найбільш важливих інформативних ознак.

У початку 1990-х років був розроблений метод оптимізації мурашиних колоній для вирішення задачі вибору підмножини ознак у просторі великої розмірності. Цей метод використовує параметр *n* для позначення загальної кількості ознак *s* для позначення розміру поточної підмножини ознак.

Зазвичай алгоритми вибору ознак використовують евристичні або випадкові стратегії пошуку, щоб уникнути надмірної складності. Однак із зростанням складності може знижуватись ступінь оптимальності кінцевої підмножини ознак.

Алгоритми вибору ознак можна розділити на три категорії залежно від їх процедур оцінювання. Якщо алгоритм виконує вибір ознак незалежно від конкретного алгоритму навчання тобто він є окремим препроцесором, то цей підхід відноситься до категорії фільтра. Цей метод зазвичай використовує критерій роздільності між класами для вибору ознак.

У випадку, коли процедура оцінювання пов'язана з конкретним завданням наприклад, класифікацією алгоритму навчання, алгоритм вибору ознак виступає як обгортка. Цей метод спрямований на знаходження оптимальної підмножини ознак у просторі шляхом оцінки точності за допомогою індукційного алгоритму.

Парадигма алгоритмів МА ґрунтується на спостереженнях етологів щодо способу, якими мурахи використовують феромонні сліди для сповіщення про найкоротші шляхи до їжі. Коли мураха рухається, вона залишає слід феромону на землі, утворюючи шлях для інших мурах.

Якщо інша мураха зустрічає цей слід, вона може визначити його і прийняти рішення віддавати перевагу цьому шляху, підсилюючи його власним феромоном. Це призводить до автокаталітичного процесу, за якого чим більше мурах йде по сліду, тим привабливіше цей шлях стає для наступних мурах. Такий механізм послідовності вибору стає натхненням для алгоритмів МА.

Алгоритми МА можна застосовувати до різних задач оптимізації, залежно від конкретної проблеми. Графічне представлення для дискретного простору пошуку має точно відображати всі можливі стани і переходи між ними. Це представлення графічно відображає всі можливі стани, які можуть виникнути в

процесі пошуку рішення, разом з усіма можливими переходами між цими станами.

Схема рішення повинна чітко визначити спосіб побудови вирішення на основі графічного представлення. Це описує процес та ймовірність переходу між станами в рамках алгоритму. Евристична доцільність зв'язків у графічному представленні означає, що структура графа відповідає евристичним правилам алгоритму пошуку. Це забезпечує оптимальне розташування та зв'язки між станами, щоб підвищити ефективність процесу пошуку.

Автокаталітичний процес зворотного зв'язку відображає механізм оновлення феромонів у графічному представленні. Цей процес враховує успіхи попередніх рішень для покращення майбутніх стратегій пошуку.

Метод задоволення обмежень в графічному представленні гарантує, що будуть розглядатися лише можливі рішення, які відповідають встановленим обмеженням або умовам.

Метод побудови рішення визначає, як саме будуються рішення на основі графічного представлення, включаючи способи переходу між станами та прийняття рішень у рамках алгоритму пошуку.

Загалом, графічне представлення дискретного простору пошуку є важливим елементом в алгоритмах оптимізації, зокрема в оптимізації мурашиної колонії. Це представлення повинне чітко відображати всі можливі стани та переходи між ними для ефективного пошуку оптимального рішення. Схема рішення, яка базується на графічному представленні, визначає процес побудови рішення з урахуванням станів та переходів у просторі пошуку. Вона описує спосіб підвищення ймовірності переходу між станами, що сприяє знаходженню оптимального розв'язку. Евристична доцільність зв'язків у графічному представленні важлива для оптимізації процесу пошуку, тобто структура графа повинна відповідати евристичним правилам для покращення ефективності алгоритму. Автокаталітичний процес зворотного зв'язку у графічному представленні використовує успіхи попередніх рішень для покращення майбутніх стратегій пошуку, підсилюючи вже знайдені шляхи та забезпечуючи

покращення процесу пошуку. Метод задоволення обмежень в графічному представленні гарантує, що розглядаються лише припустимі рішення, які відповідають встановленим обмеженням. Нарешті, метод побудови рішення визначає стратегії побудови рішення на основі графічного представлення, включаючи способи переходу між станами та прийняття рішень, що допомагає ефективно вирішувати задачі оптимізації у контексті оптимізації мурашиної колонії та інших алгоритмів.

1.3 Мета та постановка задачі

На основі проведеного детального аналізу сформульовано мету роботи. Мета даної роботи полягає у покращені вибору важливих ознак з використанням мурашиного алгоритму.

Для досягнення цієї мети виконуються наступні кроки:

1. Проведення детального аналізу існуючих методів обробки текстів з метою вибору оптимальних підходів до виділення важливих ознак.

2. Вибір і визначення набору ознак, які потрібно аналізувати для класифікації текстів. Це можуть бути ключові слова, структурні особливості, статистичні параметри тощо.

3. Розробка методу, який моделює процес визначення важливих ознак на основі поведінки мурахи. Метод повинен враховувати взаємодію між мурашиними агентами в даному випадку, методами обробки текстів та здатність адаптуватися до змін у текстових даних.

4. Розробка та тренування моделі класифікаці на основі визначених ознак за допомогою мурашиного алгоритму. Після цього проводиться оцінка ефективності моделі на тестовому наборі даних для визначення її точності та надійності в класифікації текстів.

Розділ 2 Метод вибору важливих ознак з використанням мурашиного алгоритму

2.1 Оптимізація мурашиного алгоритму

Застосування мурашиного алгоритму є евристичним методом оптимізації, який використовує популяцію програмних агентів, відомих як штучні мурашки, для пошуку наближених рішень складних оптимізаційних задач. Цей метод перетворює задачу оптимізації на пошук найкращого шляху у зваженому графі. Штучні мурашки поступово створюють рішення, пересуваючись по графу. Процес формування рішень є стохастичним і базується на моделі феромонів, яка визначає параметри, пов'язані з компонентами графу, і змінює їхні значення під час виконання мурашками.

Процес формування рішень є стохастичним, що означає, що він включає випадковість або випадкові елементи, і базується на моделі феромонів. Модель феромонів визначає параметри, пов'язані з компонентами графу, наприклад, ребрами чи вершинами, і впливає на процес прийняття рішень штучними мурашками. Значення цих параметрів змінюються під час виконання мурашками, оновлюючись відповідно до результатів їхніх дій.

Наприклад, мурашки можуть використовувати концентрацію феромону на ребрах графу для прийняття рішення про обрання шляху. Чим більша концентрація феромону на ребрі, тим більш ймовірно, що мурашка обере цей шлях. Після кожного циклу пошуку і повернення до гнізда мурашка може оновити значення феромону на ребрах відповідно до результатів своєї подорожі.

Модель феромонів грає ключову роль у процесі прийняття рішень мурашками, допомагаючи їм ефективно досліджувати простір пошуку та знаходити оптимальні рішення.

Під час подальшого виконання алгоритму штучні мурашки продовжують свою діяльність, пересуваючись по графу та змінюючи значення феромонів на ребрах відповідно до результатів їхньої діяльності. Цей процес ітеративно повторюється протягом кількох циклів пошуку та оновлення феромонів.

З кожним новим циклом пошуку мурашки накопичують досвід, що дозволяє їм уникати менш ефективних шляхів та зосереджуватись на оптимальних. За рахунок стохастичності процесу та змінності значень феромонів мурашиний алгоритм може ефективно досліджувати різноманітні варіанти рішень та знаходити оптимальні шляхи у просторі пошуку.

Поступово, через кілька ітерацій, штучні мурашки здатні сконцентрувати свої зусилля на найбільш перспективних шляхах, що призводить до збільшення ймовірності знаходження оптимального рішення. Такий процес покращення рішення базується на взаємодії між мурашками та на моделі феромонів, яка відтворює механізми природного мурашиного колективу.

У колонії мурашок кожна штучна мурашка працює над розв'язанням оптимізаційної задачі шляхом імітації поведінки реальних мурашок у природі. Вони працюють спільно, взаємодіючи з оточуючим середовищем, яке представлене зваженим графом. Ключовими елементами цього процесу є модель феромонів та правила переходу.

1. Модель феромонів в якій кожне ребро графу у колонії мурашок асоційоване з певним значенням феромону. Штучні мурашки відкладають феромони під час переміщення по ребру в графі. Чим більше мурашка проходить певний шлях, тим більше феромону відкладається на цьому шляху. Феромони впливають на вибір мурашок при наступних переходах: вони більш схильні обирати шляхи з більшою концентрацією феромону.

2. Кожна мурашка керується певними правилами для вибору наступного кроку. Ці правила враховують як феромони, так і інші параметри, наприклад, вагу ребра, щоб керувати їхнім рухом. Зазвичай, ймовірність обрання певного ребра залежить від його феромонної концентрації і ваги.

Процес пошуку рішення полягає в ітеративному спробуванні та покращенні шляхів у графі за допомогою взаємодії між мурашками та феромонами. Поступово, через багато ітерацій, колонія мурашок здатна знаходити оптимальні або наближені рішення для складних задач оптимізації,

особливо у випадках, коли інші методи можуть бути недоцільними або неефективними.

Для розуміння того, як працює оптимізація мурашиної колонії розглянемо на прикладі задачі комівояжера. Маємо набір міст і знаємо відстані між ними. Мета полягає в тому, щоб знайти найкоротший маршрут, який проходить через кожне місто лише один раз і повертається до початкового міста.

Для застосування оптимізації мурашиної колонії до цієї задачі будуємо граф, де кожен вузол представляє одне з міст. Оскільки можемо переміщатись з будь-якого міста до будь-якого іншого в TSP, граф, що використовується для побудови, є повнозв'язним, тобто містить ребра, які з'єднують кожну пару міст. Кількість вершин у графі дорівнює кількості міст у задачі.

Довжини ребер у графі встановлюються пропорційно відстанями між містами, які вони з'єднують. Крім того, кожне ребро має асоційоване значення феромону, яке представляє накопичений досвід мурашок, що проходили цей шлях. Значення феромону змінюються під час виконання, а також враховуються евристичні значення, які в даному випадку встановлені як обернені довжини ребер.

У процесі пошуку найкоротшого маршруту мурашки пересуваються по графу, вибираючи наступний крок залежно від концентрації феромонів на ребрах і евристичних властивостей. Поступово, через багато ітерацій, колонія мурашок здатна знаходити оптимальні або наближені рішення для задачі комівояжера, використовуючи механізми спільної взаємодії та накопиченого досвіду. У процесі пошуку найкоротшого маршруту для задачі комівояжера за допомогою оптимізації мурашиної колонії, кожна штучна мурашка рухається по графу, починаючи з випадково обраного міста. Під час свого переміщення мурашка враховує феромони, які залишили інші мурашки на ребрах графу, і евристичні властивості у цьому випадку, відстані між містами.

Коли мурашка доходить до ребра, вона обирає наступний вузол згідно з ймовірністю, яка залежить від концентрації феромонів на цьому ребрі і евристичного впливу. Мурашки мають тенденцію віддалятись від ребер з

низькою концентрацією феромонів або навіть уникати їх, але вони також приймають рішення, враховуючи корисність (евристику) кожного вузла.

Після того, як мурашка пройшла весь маршрут відвідала всі міста, феромони на всіх пройдених ребрах оновлюються відповідно до результатів цього маршруту. Якщо маршрут короткий тобто він задовольняє умови оптимальності, то мурашки, які пройшли цей маршрут, залишають більше феромону на відповідних ребрах, збільшуючи ймовірність вибору цих шляхів наступними мурашками.

Таким чином, колонія мурашок ітеративно покращує свої рішення, використовуючи спільний досвід і феромони, що залишили попередні мурашки. Через кілька ітерацій або поколінь мурашок зазвичай знаходять оптимальний або дуже близький до оптимального маршрут для задачі комівояжера, орієнтуючись на взаємодію та спільний накопичений досвід.

Процес побудови рішення мурашками починається з того, що кожна мураха вибирає випадкове місто вершину графа для початку свого маршруту. Потім на кожному кроці побудови мураха обирає наступний крок, рухаючись по доступним ребрам графа. Мураха пам'ятає свій шлях і на кожному кроці обирає ребро, яке веде до ще не відвіданих міст.

Імовірність вибору кожного ребра залежить від концентрації феромону на цьому ребрі та евристичних значень, які пов'язані з ребром. Чим більше феромону і евристики пов'язано з ребром, тим вища ймовірність, що мураха обере саме це ребро для свого наступного кроку.

Коли кожна мураха завершує свій маршрут, феромон на кожному ребрі оновлюється. Спочатку значення феромону на кожному ребрі зменшується на певний відсоток, щоб уникнути занадто швидкої деградації феромонів. Потім кожне ребро отримує додатковий феромон, пропорційний якості рішень, яку відповідна мураха знайшла наприклад, довжині її маршруту.

Цей ітеративний процес дозволяє колонії мурахок вдосконалювати свої рішення, використовуючи взаємодію між феромонами та евристичною

інформацією для пошуку оптимального або найближчого до оптимального маршруту для задачі комівояжера.

Стратегія застосовується ітеративно, доки не виконано умову завершення. Її ідея базується на спостереженні, що, навіть з обмеженими індивідуальними здібностями, мурашки колективно можуть ефективно знаходити найкоротший шлях між джерелом їжі та гніздом через спільне використання харчових ресурсів.

1. Перша мураха виявляє джерело їжі, проходячи через певний шлях, а потім повертається до гнізда, залишаючи за собою слід феромонів. Під час цього проходження вона може залишати за собою сліди феромонів, які вказують на шлях, яким вона рухалася. Цей слід феромонів, залишений першою мурахою, може бути розпізнаний іншими мурахами під час подальшого пошуку. Чим більше феромону залишено на певному шляху, тим більш ймовірно, що інші мурахи оберуть цей шлях, оскільки вони сприймають його як більш привабливий. Отже, перша мураха може ініціювати процес створення шляху до джерела їжі шляхом залишення слідів феромонів, які сприяють утворенню та підтримці шляху для інших мурах. Цей механізм взаємодії базується на природних поведінкових характеристиках мурашок та допомагає їм спільно вирішувати завдання пошуку джерела їжі.

2. Інші мурахи вибираються чотирма різними шляхами, проте посилення феромонів робить найкоротший шлях більш привабливим. Однак, через механізм підсилення феромонів, найкоротший шлях може стати більш привабливим для інших мурах. Коли мурахи повертаються до гнізда після проходження своїх маршрутів, вони залишають за собою сліди феромонів на кожному зі своїх шляхів. Чим коротший шлях, тим більше феромону може бути залишено на ньому через менший час проходження.

Через підсилення феромонів, концентрація феромону на найкоротшому шляху збільшується. Це робить цей шлях більш привабливим для інших мурах, які сприймають його як оптимальний. Відтак, з часом більшість мурах може

схилитися до використання саме цього шляху, що призводить до утворення спільно визнаного найкращого маршруту до джерела їжі.

3. Мурахи вибирають найкоротший шлях, підсилюючи слід феромонів на цьому шляху, тоді як інші шляхи втрачають свої слідові сліди. Коли мурахи обирають найкоротший шлях та повертаються до гнізда, вони залишають за собою слід феромонів на цьому маршруті. Чим коротший шлях, тим більше феромону буде залишено, оскільки мурахи проводять менше часу на його проходженні.

В той же час, феромони на інших шляхах поступово випаровуються або розпадаються з часом, оскільки мурахи не залишили на них нового сліду феромону. Це призводить до зниження концентрації феромону на цих шляхах та зменшення їх привабливості для інших мурах.

Через підсилення сліду феромонів на найкоротшому шляху та природний розпад феромонів на інших шляхах, мурахи спрямовують свої усили на обрання оптимального маршруту до джерела їжі.

У ході досліджень на колонії мурашок, де їм пропонували вибір між двома шляхами різної довжини до джерела їжі, було помічено, що мурахи виявляли схильність використовувати найкоротший шлях. Модель, яка пояснює це поведінку, може бути описана так.

1. Початково мураха випадково блукає навколо колонії.

2. Якщо вона випадково знаходить джерело їжі, то вона повертається пряміше до гнізда, залишаючи слід феромону на своєму шляху.

3. Ці феромони приваблюють інших мурах, які схильні йти пряміше по сліду.

4. Після повернення до колонії, ці мурахи підсилюють маршрут, залишаючи більше феромонів.

5. З плином часу більше мурах пройде коротшим шляхом, ніж довгим.

6. Короткий шлях з часом стає більш привабливим і підсилюється.

7. Довгий шлях зникає, оскільки феромони нестійкі.

8. У результаті всі мурахи визначаються на користь найкоротшого шляху.

У подальших етапах дослідження мурах вибирають найкоротший шлях через такі кроки:

1. Після того, як мурахи знайдуть найкоротший шлях до джерела їжі і повернуться до гнізда, вони залишають більше феромону на цьому маршруті. Це призводить до зростання привабливості цього шляху для інших мурах. Підсилення феромонів відбувається через те, що мурахи, які пройшли коротший шлях, залишають більше сліду феромонів на цьому шляху під час повернення до гнізда. Внаслідок цього, концентрація феромону на коротшому шляху збільшується. Це призводить до того, що інші мурахи, які будуть обирати шлях у майбутньому, стануть більш схильними обирати шлях, який має більшу концентрацію феромону, оскільки він стає більш привабливим. Цей механізм підсилення феромонів допомагає у конвергенції мурашиного алгоритму, тобто зближенні рішень до оптимального. Він дозволяє підсилити шляхи, які ведуть до кращих розв'язків, та зменшити імовірність обирання менш оптимальних шляхів у майбутньому.

2. Після декількох циклів пошуку їжі та повернення до гнізда більшість мурах складається на користь короткого шляху через накопичені феромони. Мурахи, які вибирають довший шлях, можуть не знаходити джерело їжі так ефективно і, отже, не залишають на ньому так багато феромону. Це призводить до утворення явища, відомого як позитивний зворотний зв'язок. Мурахи, які обирають коротший шлях та успішно дістаються до джерела їжі, залишають більше феромону на цьому маршруті. Це зростання концентрації феромону робить коротший шлях ще більш привабливим для інших мурах, оскільки вони сприймають його як оптимальний. У той же час, менш вигідний шлях отримує менше феромону, оскільки мурахи, які його обирають, не досягають успіху у пошуку їжі. В результаті, з плином часу більшість мурах переважно обирає коротший шлях, що має високу концентрацію феромону. Це допомагає мурахиному колективу зосередити свої зусилля на оптимальних маршрутах та ефективно знаходити джерела їжі.

Після того, як більшість мурах складається на користь короткого шляху через накопичені феромони, може відбуватися подальше посилення цього ефекту. Чим більше мурах обирає короткий шлях і залишає на ньому феромону, тим більше ймовірність того, що інші мурахи також оберуть цей шлях.

Цей процес може призвести до утворення позитивного зворотного зв'язку, де короткий шлях стає ще більш привабливим для мурах, оскільки він має високу концентрацію феромону. У той же час, менш вигідний шлях може ставати ще менш привабливим через зменшення концентрації феромону.

Через подальше посилення позитивного зворотного зв'язку мурахи набувають тенденцію переважно обирати короткий шлях, що призводить до ще більшої узгодженості між ними та збільшення ефективності пошуку їжі. Цей процес може продовжуватися, поки не буде досягнуто оптимального балансу між різними шляхами та їхньою концентрацією феромону.

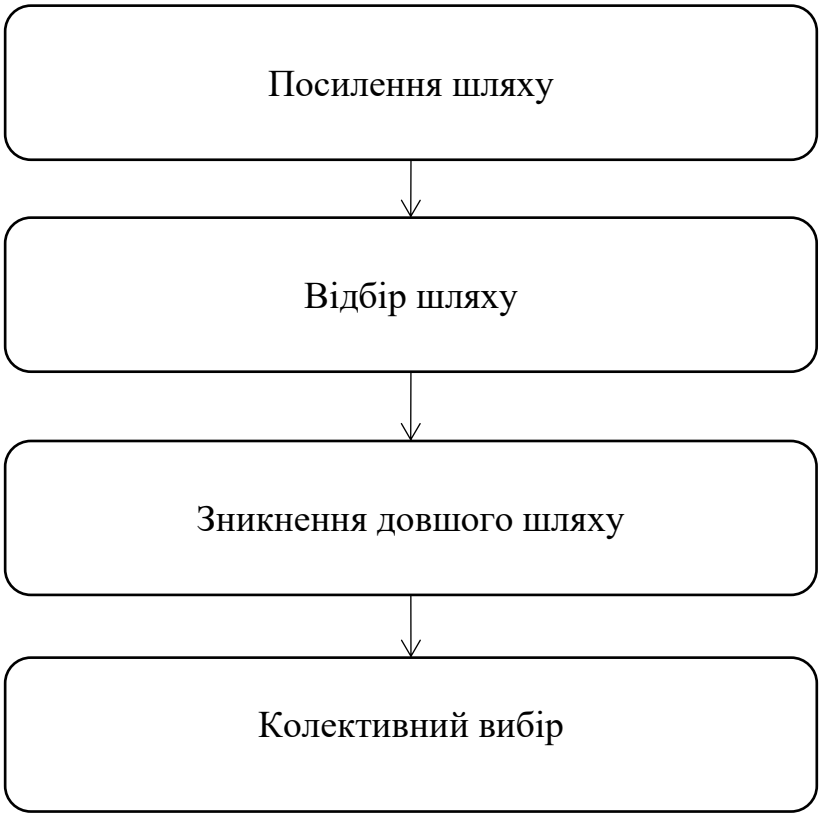


Рисунок 2.1 – Кроки вибор шляху мурахами

3. Тим часом, довший шлях поступово втрачає свою привабливість, оскільки феромони на ньому не поповнюються так швидко. Мурахи, які випадково обирають цей шлях, знаходять його менш привабливим через недостатність феромонів.

4. У кінці експерименту більшість мурах спрямовуються саме по найкоротшому шляху, оскільки він стає найбільш привабливим та ефективним завдяки накопиченим феромонам.

Цей процес використовує механізми посилення і взаємодії між мурахами та феромонами для знаходження та вдосконалення найефективнішого шляху до джерела їжі.

2.2 Евристика вибору шляху мурахи для руху за графом

У методі оптимізації мурашиних колоній, штучні мурахи використовуються для побудови рішень у задачах комбінаторної оптимізації. Вони проходять повнозв’язний граф конструкції, де кожен вузол або ребро асоціюється з певними компонентами рішення. Кожен компонент рішення представлений як змінна $X_i=v_{ij}$, що позначається як c_{ij}, і належить до набору всіх можливих компонентів C.

У мурашиному алгоритмі значення феромону пов’язане з кожним компонентом. Ці значення феромонів відіграють ключову роль у процесі пошуку рішення, оскільки вони моделюють розподіл ймовірностей для вибору різних компонентів рішення. Під час пошуку рішення мурахами, значення феромону використовуються для визначення ймовірності вибору конкретного шляху. Це відбувається так: мурахи рухаються по графу, керуючись значеннями феромону на ребрах між вузлами. Чим вище значення феромону на певному шляху, тим більша ймовірність, що цей шлях буде обраний.

Після того, як мурахи знайдуть рішення, значення феромонів оновлюються. Оновлення феромонів здійснюється шляхом їх підсилення на шляхах, що належать до знайдених ефективних рішень, і випаровування

феромонів на менш ефективних шляхах. Це оновлення виконується за принципом зміцнення найбільш ефективних компонентів рішення: шляхи, що привели до кращих рішень, отримують більшу кількість феромону, тим самим підвищуючи їхню привабливість для майбутніх мурах.

Цей механізм забезпечує адаптацію алгоритму до найкращих знайдених рішень і поступове вдосконалення вибору шляхів. У процесі багатьох ітерацій, мурахи збирають досвід попередніх пошуків, що дозволяє їм зосереджуватися на найбільш перспективних областях простору рішень. Таким чином, феромони відіграють важливу роль у колективному пошуку та оптимізації рішень, сприяючи ефективному вирішенню задачі. Мурахи переміщуються від однієї вершини до іншої через краї графа побудови, використовуючи інформацію, отриману зі значень феромонів. Цей процес дозволяє їм поступово конструювати рішення. Крім того, мурахи відкладають феромони на компоненти, через які вони проходять, будь то вершини чи ребра графа. Кількість феромонів, що вони залишають, може залежати від якості знайденого розв’язку. Це дозволяє наступним мурахам використовувати інформацію про феромони як напрямок до більш перспективних областей пошуку.

Цей процес оновлення феромонів включає два основні механізми: випаровування феромону та відкладення феромону. Цей механізм дозволяє алгоритму уникати застрягання в локальних оптимумах. Після кожної ітерації частина феромону на всіх шляхах випаровується за фіксованою ставкою. Це забезпечує зниження ймовірності вибору менш ефективних шляхів з часом, якщо вони не підтримуються новими мурахами.

Після завершення ітерації мурахи відкладають феромон на шляхи, які вони використовували для формування своїх рішень. Кількість відкладеного феромону пропорційна якості знайденого рішення: чим краще рішення, тим більше феромону відкладається. Це підсилює привабливість ефективних шляхів і збільшує ймовірність їх вибору в майбутніх ітераціях.

Метод оптимізації мурашиного колонії використовує поведінку штучних мурашок для побудови рішень у складних задачах оптимізації.

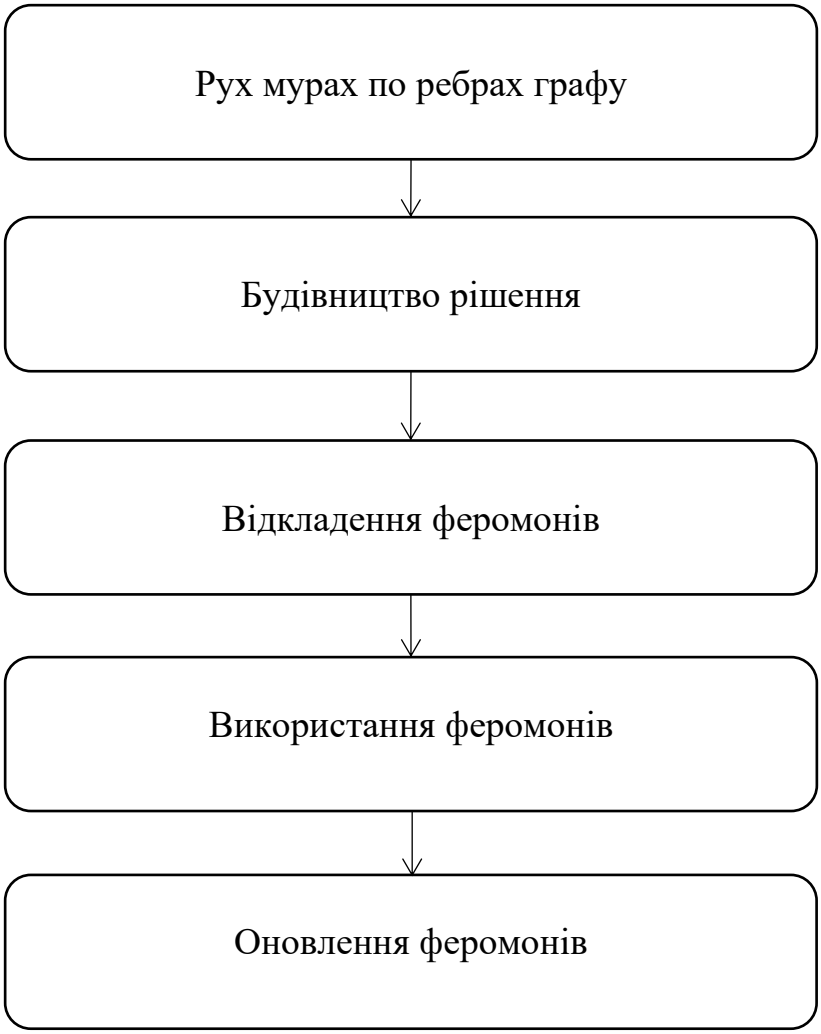


Рисунок 2.2 – Схема поведінки штучних мурашок

Мурахи переміщуються по графу побудови, що відображає структуру задачі, і використовують інформацію про феромони для вибору шляхів. Цей процес складається з кількох кроків:

1. Кожна мураха в мурашиному алгоритмі починає з випадково обраної вершини графа, яка представляє одну з можливих ознак для аналізу. Мураха пересувається до інших вершин, поступово формуючи шлях, що відповідає певному набору ознак. При цьому вибір наступної вершини базується на значеннях феромонів, які моделюють ймовірність вибору певного шляху, та на евристичній інформації, яка враховує корисність певної ознаки.

2. Мурахи поступово формують рішення, переміщуючись по графу і збираючи компоненти рішення, наприклад, міста в задачі комівояжера або

ознаки в задачі вибору важливих ознак. Вони використовують значення феромонів для визначення ймовірності вибору кожного шляху, що дозволяє їм адаптивно знаходити оптимальні або наближені до оптимальних рішення.

3. Мурахи залишають за собою феромонний слід на шляхах, які вони проходять. Кількість відкладеного феромону пропорційна якості знайденого рішення: чим якісніше рішення, тим більше феромонів відкладають мурахи. Це дозволяє алгоритму поступово адаптуватися до кращих рішень і концентрувати пошук на найбільш перспективних шляхах.

4. Наступні мурахи використовують інформацію про феромони, щоб обирати більш перспективні шляхи. Мурахи залишають за собою феромонний слід на шляхах, які вони проходять. Кількість відкладеного феромону пропорційна якості знайденого рішення: чим якісніше рішення, тим більше феромонів відкладають мурахи. Це дозволяє алгоритму поступово адаптуватися до кращих рішень і концентрувати пошук на найбільш перспективних шляхах.

5. Значення феромонів оновлюються після кожного проходження мурах через граф. Кращі розв’язки призводять до більшого відкладення феромонів, що робить їх привабливішими для наступних мурах.

Ця процедура ітеративно виконується доти, доки не буде задоволено критерій припинення, наприклад, досягнення певної кількості ітерацій або зближення до оптимального розв’язку. МА є потужним методом оптимізації, який може застосовуватися до широкого спектру задач, від задачі комівояжера до розкладання задач на ресурси.

Мурашиний алгоритм використовує поведінку штучних мурашок для розв'язання складних задач оптимізації. Основні кроки алгоритму можна описати наступним чином.

1. Встановлення параметрів і початкове налаштування слідів феромонів. Спочатку встановлюються параметри алгоритму, такі як кількість мурах, кількість ітерацій, швидкість випадкового руху мурах та інші важливі налаштування. Також ініціалізуються сліди феромонів на графі побудови, які

представляють собою значення, що вказують на відповідність кожного шляху або компоненту рішення.

2. Послідовність дій.

Після ініціалізації алгоритму виконуються певні етапи у визначеній послідовності:

– Конструювання рішень мурашками. Кожна мураха починає зі своєї початкової вершини та поступово будує своє рішення, рухаючись по графу і вибираючи шляхи на підставі значень феромонів та евристичних інформацій.

– Дії демонів. Це може бути додатковий крок, де введені додаткові дії, які керують поведінкою мурах або впливають на процес побудови рішень.

– Оновлення слідів феромонів. Після того, як усі мурахи побудували свої рішення, виконується оновлення слідів феромонів. Це включає відкладання нових слідів феромонів на шляхах, які використовували мурахи, а також оновлення існуючих слідів шляхів залежно від якості розв'язків.

3. Закінчення процесу.

Після завершення всіх запланованих дій у розкладі, алгоритм перевіряє критерії припинення наприклад, досягнення максимальної кількості ітерацій або досягнення оптимального розв'язку і, за необхідності, повторює цикл знову.

МА ітеративний процес, де кожна ітерація дозволяє мурахам знаходити оптимальніші шляхи в графі побудови, залежно від значень феромонів та евристичних інформацій. Він є потужним і універсальним методом для вирішення різних задач оптимізації і знаходження найкращих рішень в складних просторах пошуку.

Після розкладу дій в методі оптимізації мурашиної колонії, процес продовжується таким чином. Кожна мураха починає свій шлях з випадково обраної початкової вершини у графі. Під час побудови рішення мурахи вибирають наступну вершину для відвідування залежно від значень феромонів на доступних ребрах і евристичних відомостей про відстані. Вони продовжують рухатись від вершини до вершини, побудовуючи шлях, який відображає поточне рішення.

Якщо включено дії демонів, цей етап може включати додаткові маніпуляції або керуючі дії, які впливають на поведінку мурах під час побудови рішень. Наприклад, це може бути вплив на вибір шляхів або регулювання значень феромонів.

Після того, як мурахи побудували свої рішення, оновлюються значення слідів феромонів. Кожна мураха вносить певну кількість феромону на кожне ребро, що вона використовувала під час побудови свого шляху. Оновлення відбуваються враховуючи якість кожного знайденого рішення: успішні рішення, які ведуть до задовільного результату збільшують вміст феромону, тоді як невдалі рішення можуть призвести до зменшення сліду феромону на певних шляхах.

Завершення ітерації та перевірка критеріїв припинення. Після завершення процесу побудови рішень і оновлення феромонів відбувається перевірка критеріїв припинення. Якщо досягнутий критерій припинення, наприклад, задана кількість ітерацій або досягнення оптимального розв'язку, алгоритм завершує роботу. У протилежному випадку процес повторюється, починаючи з конструкції нових рішень мурашками.

Дії демонів в мурашиному алгоритмі полягають у додаткових обчисленнях або змінах в самому алгоритмі для покращення його роботи. Демони можуть виконувати різні функції, в залежності від потреб конкретної задачі. Наприклад, вони можуть контролювати та регулювати параметри алгоритму, такі як швидкість випаровування феромону або ваги для розрахунку ймовірностей переміщення мурах.

Здійснювати динамічне налаштування параметрів залежно від стану пошуку або інших факторів. Впроваджувати локальні покращення розв’язків, такі як використання локальних пошукових методів для покращення якості рішення.

Змінювати граф пошуку, додаючи нові ребра або змінюючи ваги, що регулюють переміщення мурах. Видаляти слабкі феромони або ребра для підтримки різноманітності і уникнення застрягання на непродуктивних шляхах.

Дії демонів можуть бути виконані за певними правилами або реагувати на певні події у процесі пошуку. Вони сприяють покращенню ефективності алгоритму та його здатності адаптуватися до змінних умов та складних задач.

Рисунок 2.3 – Кроки мурашиного алгориму

Метод МА ефективно використовує взаємодію мурашок та їхнє підтримання слідів феромонів для знаходження найкращих шляхів у графі побудови, що дозволяє знаходити оптимальні рішення в складних просторах пошуку.

Метаевристика складається з ініціалізаційного кроку та трьох основних алгоритмічних компонентів, які активуються. Ця конструкція повторюється до

виконання критерію завершення, якими можуть бути, наприклад, максимальна кількість ітерацій або час обчислень.

У багатьох випадках застосування МА до складних задач тривимірні алгоритмічні компоненти працюють у циклі, включаючи створення рішень всіма мурахами, необов'язкове вдосконалення цих рішень за допомогою алгоритму локального пошуку та оновлення феромонів.

Мурахи будують рішення шляхом поступового розширення часткового розв'язку за допомогою доступних компонентів рішення. Починаючи з порожнього часткового розв'язку, на кожному кроці мурахи додають допустимий компонент розв'язку зі своєї множини можливих сусідів. Цей процес побудови рішень можна розглядати як подорож на графі побудови, де дозволені шляхи визначаються механізмом побудови рішення.

Вибір компонента рішення з множини можливих сусідів відбувається ймовірнісним чином. Правила ймовірного вибору можуть відрізнятись у різних варіантах метаевристики МА. Одним з найбільш відомих правил є правило мурашиної системи AS, де значення феромону τ_{ij} та евристична цінність η_{ij}, пов'язані з кожним компонентом розв'язку, використовуються для розрахунку ймовірності вибору компонента. Параметри α і β визначають відносну важливість феромонної та евристичної інформації.

У процесі побудови рішення штучні мурахи додають компоненти розв'язку до часткового розв'язку з множини доступних компонентів C. Починаючи з порожнього часткового розв'язку, кожна мураха обирає допустимий компонент розв'язку з множини можливих сусідів Nsp залежно від ймовірності вибору.

Ймовірність вибору компонента розв'язку c_{ij} мурахою визначається за допомогою формули, яка базується на значенні феромону τ_{ij} та евристичній цінності η_{ij}:

$$p_{ij} = \frac{(\tau_{ij})^{\alpha} \times (\eta_{ij})^{\beta}}{\sum_{k \in N(sp)} (\tau_{ik})^{\alpha} \times (\eta_{ik})^{\beta}} \quad (2.1)$$

Тут τ_{ij} – значення феромону, пов'язане з компонентом c_{ij};

η_{ij} – евристична цінність компонента c_{ij};

α , β – параметри, які визначають відносну важливість феромону та евристичної інформації;

$N(s_p)$ – множина можливих сусідів часткового розв'язку s_p.

Після вибору компонента мураха додає його до часткового розв'язку s_p і продовжує цей процес, поки не буде сформовано повне рішення. Крім того, під час побудови рішення мурахи залишають феромони на вибраних компонентах, що впливає на майбутні вибори інших мурах.

Після створення розв'язків, але перед оновленням значень феромонів, можуть виникнути спеціалізовані дії, що виконуються централізовано і називаються діями демона. Ці дії призначені для вирішення конкретних проблем, які важко вирішити окремим мурахам. Найбільш поширеною дією демона є застосування локального пошуку до створених розв'язків. Локально оптимізовані розв'язки потім можуть використовуватися для визначення того, які значення феромонів слід оновити.

Після того як створені розв'язки мурахами, демонічні дії можуть включати в себе різноманітні операції, спрямовані на оптимізацію або вдосконалення цих розв'язків перед тим, як феромони будуть оновлені. Основна дія демона – це застосування локального пошуку до створених розв'язків. Локальний пошук спрямований на знаходження оптимальних або покращених варіантів розв'язків, що можуть бути використані для покращення якості феромонів.

Під час локального пошуку розв'язки піддаються додатковій оптимізації, щоб зменшити загальну вартість або підвищити ефективність. Це може

включати перегляд і модифікацію окремих частин розв'язків, щоб знайти кращі варіанти шляху або структури.

Отримані після локального пошуку оптимізовані розв'язки можуть бути використані для оновлення значень феромонів. Це оновлення може базуватися на якості знайдених розв'язків: покращені або оптимізовані розв'язки можуть вносити більш значущий внесок у значення феромонів, що сприятиме подальшому поліпшенню якості розв'язків, що будуть створені мурахами в майбутньому.

Метою оновлення значень феромонів є підвищення ваги феромонів, пов'язаних з ефективними рішеннями, і зменшення тих, що стосуються менш ефективних. Це досягається за допомогою двох процесів. По-перше, всі значення феромонів зменшуються через процес випаровування феромонів з інтенсивністю, що контролюється параметром ρ (де $0 < \rho \leq 1$). По-друге, значення феромонів, що стосуються обраних оптимальних рішень, збільшуються на основі їх якості, яка визначається функцією пристосованості F.

Випаровування феромонів сприяє корисному забуванню, що спонукає дослідження нових областей у просторі пошуку. Різні алгоритми MA, такі як система мурашиної колонії або система MAX-MIN ant, можуть відрізнятися за способом оновлення значень феромонів залежно від їх специфіки.

Випаровування феромонів в алгоритмах мурашиних систем сприяє зменшенню значень феромонів на всіх шляхах і ребрах графа побудови. Цей процес імітує природну деградацію феромонів з часом. Параметр ρ визначає швидкість випаровування, де значення $0 < \rho \leq 1$. Чим більше значення ρ, тим швидше випаровуються феромони.

Оновлення феромонів також включає посилення феромонного сліду на хороших рішеннях, що базується на їх якості. Якість рішень оцінюється за допомогою ознаки пристосованості *F*, яка зазвичай називається функцією об'єктивної оцінки розв'язків. Чим краще розв'язок, тим більше феромону відкладається на шлях або ребро, пов'язане з цим розв'язком.

Таким чином, процес оновлення феромонів забезпечує згущення феромонного сліду на шляхах, які ведуть до оптимальних розв'язків, і зменшення сліду на менш ефективних шляхах. Це сприяє покращенню мурашиних алгоритмів, орієнтованих на пошук оптимальних рішень в складних просторах пошуку.

Один із прикладів правила оновлення феромонів, що був наведений вище, полягає у використанні різних специфікацій для множини $Supd$, яка часто є підмножиною $S_{iter} \cup \{SBS\}$, S_{iter} де представляє собою набір рішень, створених у поточній ітерації, SBS є найкращим рішенням, знайденим після першої ітерації алгоритму. Один з добре відомих прикладів такого правила оновлення є правило Ant System, де $Supd$ встановлюється як S_{iter}.

Ще одним прикладом є правило BD оновлення найкраще рішення, яке акцентується на знайдених кращих рішеннях S_{iter} замість усіх рішень у поточній ітерації. Це правило надає більш сильний вплив на збереження хороших рішень, порівняно з правилом AS оновлення. Однак це також може збільшити ймовірність передчасної зупинки алгоритму через збільшення швидкості зближення до певного набору рішень.

Більший вплив надає правило BD оновлення, де BD вказує на використання найкращого знайденого рішення SBS. У цьому випадку $Supd$ встановлюється тільки на множину, що містить найкращий розв'язок SBS. Алгоритми MA, які використовують варіанти правил оновлення AS або BS, часто також включають механізми для запобігання передчасної зупинки, щоб досягти кращих результатів у пошуку оптимальних рішень.

Правила оновлення феромонів, що застосовуються в алгоритмах мурашиного колонії, можуть мати різні специфікації відносно множини $Supd$, яка визначає набір рішень, за якими проводиться оновлення феромонів. У багатьох випадках ця множина $Supd$ є підмножиною $S_{iter} \cup \{SBS\}$, де S_{iter} – охоплює набір рішень, створених протягом поточної ітерації алгоритму; SBS – представляє собою найкраще знайдене рішення після першої ітерації. Прикладом

такого правила оновлення є AS, де визначається як S_{iter} , тобто використовуються всі рішення, створені протягом поточної ітерації.

Поширене правило оновлення IB – найкраща ітерація, яке акцентується на знайдених найкращих рішеннях S_{iter} , ігноруючи решту. Це правило надає більший вплив на збереження хороших рішень порівняно з правилом AS-оновлення, але може призвести до занадто швидкого зближення до обмеженого набору рішень.

Більший вплив надає правило *BS* найкращий на даний момент, де *Supd* обмежується найкращим знайденим рішенням *SBS*. В цьому випадку використовується лише одне рішення для оновлення феромонів. Алгоритми MA, що використовують правила оновлення IB або BS, часто доповнюються механізмами для уникнення передчасної зупинки алгоритму, щоб досягти кращих результатів у пошуку оптимальних рішень.

Правила оновлення феромонів, що застосовуються в алгоритмах мурашиного колонії, різняться за специфікацією множини *Supd* , яка визначає набір рішень, за якими проводиться оновлення феромонів. У багатьох випадках \ *Supd* є підмножиною

$$S_{iter} \cup \{SBS\}, \tag{2.2}$$

де S_{iter} – охоплює набір рішень, створених протягом поточної ітерації алгоритму;

SBS – представляє собою найкраще знайдене рішення після першої ітерації.

Наприклад, правило оновлення AS використовує $Supd = S_{iter}$, тобто всі рішення поточної ітерації. Правило оновлення IB акцентується на знайдених найкращих рішеннях S_{iter}, ігноруючи решту. Це правило надає більший вплив на збереження хороших рішень порівняно з правилом AS-оновлення, але може призвести до занадто швидкого зближення до обмеженого набору рішень.

Далі, правило BS обмежує *Supd* найкращим знайденим рішенням *SBS*. В цьому випадку використовується лише одне рішення для оновлення феромонів.

Алгоритми МА, які використовують правила оновлення IB або BS, часто доповнюються механізмами для уникнення передчасної зупинки алгоритму, щоб досягти кращих результатів у пошуку оптимальних рішень.

2.3 Опис застосування мурашиного алгоритму

Основні алгоритми МА мають багато спільного, але їхня реалізація і результати можуть значно відрізнятися. Одним із перших і успішних алгоритмів МА був Ant System (AS). Основна ідея AS полягає в оновленні феромонів на всіх ребрах графа після того, як усі мурахи завершили свій тур. В даному випадку компонентами рішення є ребра графа, і оновлення феромонів для кожного ребра (i, j) . При оновленні феромонів для ребра в системі Ant System, значення феромону τ_{ij} збільшується за допомогою наступної формули:

$$\tau_{ij} \leftarrow (1-\rho)\cdot\tau_{ij} + \Delta\tau_{ij}. \qquad (2.3)$$

Тут ρ – коефіцієнт випаровування, який визначає швидкість випаровування феромонів. Параметр $\Delta\tau_{ij}$ визначається як сума феромонів, яку вносять всі мурахи, які пройшли через ребро (i, j) . Значення $\Delta\tau_{ij}$ зазвичай пропорційне якістю розв'язку, яка оцінюється функцією пристосованості.

Система AS базується на ідеї колективної інтелігенції, де мурахи обмінюються інформацією через феромони, щоб поступово знаходити оптимальні шляхи. Цей алгоритм дозволяє зберігати різноманіття і веде до еволюції феромонних шляхів в напрямку оптимальних розв'язків задачі.

Розглянемо інший варіант алгоритму МА – систему мурашиних колоній (СМК). У системі AS, під час побудови розв'язків мурахи переходять через будівельний граф і приймають ймовірнісні рішення в кожній вершині. Ймовірність переходу $p(c_{ij} \mid \mathrm{sk}_k^p)$ для k мурахи, яка переміщується з міста i до міста j, визначається наступним чином:

$$p(c_{ij} | sk_k^p) = \frac{[\tau_{ij}]^\alpha \cdot [\eta_{ij}]^\beta}{\sum_{c_{mn} \in N(sk_k^p)} [\tau_{mn}]^\alpha \cdot [\eta_{mn}]^\beta} \quad (2.4)$$

де: τ_{ij} – значення феромону на ребрі (i, j);

$\eta_{ij} = \frac{1}{d_{ij}}$ – евристична інформація, де d_{ij} – довжина компонента c_{ij};

α і β – параметри, що контролюють важливість феромону порівняно з евристичною інформацією;

$N(sk_k^p)$ – набір компонентів, які ще не належать частковому розв'язку sk_k^p мурашки k.

Ця формула визначає ймовірність вибору краю (i, j) мурахою k залежно від значень феромону τ_{ij} та евристичної інформації η_{ij}, а також параметрів α і β, які регулюють вагу цих компонентів. Вибір краю здійснюється імовірнісно відповідно до цієї формули під час побудови розв'язку кожною мурахою.

Система мурашиних колоній робить удосконалення оригінальної системи мурашиного алгоритму. Однією з ключових відмінностей СМК від AS є форма правила прийняття рішень, яке використовують мурахи під час процесу конструювання розв'язків.

У системі СМК мурахи використовують псевдовипадкове пропорційне правило: ймовірність того, що мураха переміститься з міста i до міста j, залежить від випадкової величини q, яка рівномірно розподілена на інтервалі [0,1], і параметра q_0. Якщо $q \leq q_0$, то серед допустимих компонентів обирається той, який максимізує добуток $\tau_{il}\eta_{il}^\beta$, в іншому випадку використовується те саме рівняння, що і в AS.

Це правило є досить жадібним, оскільки сприяє використанню інформації про феромони, але збалансоване за рахунок введення компонента диверсифікації: місцеве оновлення феромонів. Місцеве оновлення феромонів виконується всіма мурахами після кожного етапу конструювання розв'язку.

Кожна мураха застосовує це локальне оновлення лише до останнього пройденого краю, що є важливою частиною системи мурашиних колоній (СМК). φ (де $\varphi \in [0,1]$) є коефіцієнтом розпаду феромону, а τ_0 представляє собою початкове значення феромону на краю.

Основною метою локального оновлення є урізноманітнення пошуку, який проводять наступні мурахи протягом однієї ітерації. Фактично, зменшення концентрації феромону на краях під час їх проходження протягом однієї ітерації заохочує наступних мурах вибирати інші краї і, отже, генерувати інші рішення. Це знижує ймовірність того, що декілька мурах створять ідентичні рішення протягом однієї ітерації. Крім того, завдяки локальному оновленню феромонів у СМК, мінімальні значення феромонів обмежуються.

Подібно до AS, в СМК також в кінці процесу конструювання виконується оновлення феромонів, яке називається автономним оновленням феромонів.

Оновлення рівнів феромонів у системі СМК здійснюється тільки найкращою мурахою, тобто оновлюються лише краї, які відвідав найефективніший мародер. Це відбувається відповідно до такого рівняння:

$$\Delta\tau_{ij} = \begin{cases} \dfrac{1}{L_{best}}, & \textit{якщо найкраща мураха} \\ & \textit{використала ребро}(i,j)\ \textit{у своєму турі}, \\ 0, & \textit{інакше} \end{cases} \qquad (2.5)$$

Тут L_{best} може бути визначено як довжина кращого туру, знайденого в поточній ітерації, або як довжина найкращого рішення з моменту запуску алгоритму, найкраще на даний момент. Важливо зауважити, що більшість інновацій, представлених у системі СМК, були вперше введені в попередній версії Ant-Q.

Система MAX-MIN ant вдосконалює концепцію системи мурашок. MMAS відрізняється від AS тим, що тільки найкраща мураха залишає сліди феромону, і мінімальне та максимальне значення феромону встановлені явно в

AS та СМК ці обмеження визначаються неявно шляхом роботи алгоритму, а не визначені автором алгоритму.

Формула оновлення феромонів для MAX-MIN ant має наступний вигляд:

$$\Delta\tau_{ij} = \begin{cases} \frac{1}{L_{best}}, \\ 0. \end{cases} \tag{2.6}$$

Тут L_{best} є довжиною туру найкращої мурахи. Як і в СМК, L_{best} може бути визначено залежно від рішення розробника алгоритму як L_{ib} або L_{bs} найкращий з моменту запуску алгоритму, або як комбінація обох.

У системі MAX-MIN ant значення феромонів обмежені між двома значеннями, τ_{min} і τ_{max}. Після оновлення феромонів мурахами, всі значення феромону перевіряються, щоб гарантувати, що вони знаходяться в межах цих обмежень. Якщо значення τ_{ij} перевищує τmax, воно буде обмежене τmax, якщо значення τ_{ij} менше τ_{min}, воно буде підняте до рівня τmin.

Рівняння оновлення феромонів у MMAS застосовується до всіх ребер у графі. Це відрізняється від системи мурашиних колоній, де оновлення застосовується лише до ребер, використаних найкращою мурахою.

Мінімальне значення феромону, τ_{min}, зазвичай емпірично визначається шляхом експериментів. Є аналітичні методи для визначення цього значення. Максимальне значення феромону, τmax, можна обчислити аналітично, особливо якщо відома оптимальна довжина маршруту. Наприклад, для задачі комівояжера, τ_{max} може бути обчислене як

$$\tau_{max} = \frac{1}{\rho L}, \tag{2.7}$$

де L – це оптимальна довжина маршруту. Якщо L∗ невідомо, його можна наблизити до L_{bs}.

Початкове значення феромону зазвичай встановлюється на рівні τmax. Якщо протягом певної кількості ітерацій алгоритм не спостерігає покращення, алгоритм може бути перезапущений з метою пошуку кращих рішень.

2.4 Метод вибору важливих ознак з використанням мурашиного алгоритму

Метод вибору ознак використовує алгоритм оптимізації мурашиної колонії для ефективної аналітики веб-сторінок. Підхід розбиває процес видобутку ознак на дві частини: перше – отримання ознак зі сторінок, а потім вибір оптимального набору за допомогою алгоритму оптимізації мурашиної колонії.

Ознаки екстрагуються зі сторінок за допомогою процесу отримання ознак, щоб сформувати набір, який потім піддається оптимізації з використанням алгоритму мурашиної колонії. Даний алгоритм являє собою набір програмних агентів, які емулюють поведінку мурашок, що шукають оптимальні рішення для задачі оптимізації.

Цей підхід перетворює проблему оптимізації в пошук оптимального шляху у взваженому графі. Штучні мурашки поступово побудовують рішення, пересуваючись по графу. Процес побудови рішення є стохастичним і базується на моделі феромонів, що дозволяє параметрам вузлам або ребрам графа змінюватись в процесі роботи мурашок.

Мурахи в алгоритмі мурашиної колонії починають зі стартової точки і поступово будують свої шляхи, враховуючи інтенсивність феромонів на кожному зв'язку графу. Феромони відкладаються мурашками під час проходження через ребра графу, і їх концентрація змінюється залежно від якості шляхів.

Цей процес є ітеративним, і з кожною ітерацією мурахи покращують свої вибори, орієнтовані на найбільш феромонні шляхи. Після декількох ітерацій

алгоритм збігається до оптимального рішення, яке відображає набір ознак, що найкращим чином відповідають поставленій задачі.

Цей підхід має великий потенціал у виборі найбільш релевантних та корисних ознак для подальшого аналізу веб-сторінок. Він дозволяє автоматизувати процес вибору ознак із складних даних, забезпечуючи при цьому оптимальні результати за рахунок ефективного використання мурашиної колонії як інструменту оптимізації.

Граф створюється зі списку ознак, які є вузлами графу, кожна з них має свою власну частоту. Після цього застосовується мурашиний алгоритм до цього графу. Кожна ітерація алгоритму MA генерує найкращий шлях або оптимальний піднабір ознак. Отримане найкраще рішення порівнюється з попередніми результатами. Якщо нове рішення краще за попередні, то воно зберігається і феромон додається до цього шляху. Якщо ж рішення збігається з попередніми, то отримуємо оптимальний піднабір ознак, який використовується для категоризації веб-сторінок з більшою ефективністю.

Отриманий оптимальний піднабір ознак після застосування алгоритму мурашиної колонії стає основою для ефективної класифікації веб-сторінок. Цей піднабір враховує найбільш важливі та корисні ознаки зі сторінок, що дозволяє автоматизувати і поліпшити процес аналізу контенту.

Процес збору і порівняння результатів між ітераціями алгоритму MA допомагає знаходити оптимальніші шляхи і піднабори ознак, забезпечуючи збільшення якості вибраних характеристик. Феромони, які відкладаються мурахами під час проходження, впливають на вибір шляхів і допомагають зберігати та покращувати кращі рішення з кожною ітерацією.

Такий підхід до вибору ознак і їх оптимізації за допомогою алгоритму мурашиної колонії є потужним інструментом для автоматизації складних аналітичних завдань, пов'язаних з аналізом великих обсягів веб-контенту. Він дозволяє зосередитися на найважливіших аспектах сторінок і забезпечує більш точну та швидку класифікацію цього контенту з використанням оптимальних наборів ознак.

Алгоритм вибору ознак на основі МА полягає у використанні штучних мурах для знаходження оптимального піднабору ознак для категоризації веб-сторінок. Детальний опис кожного кроку алгоритму:

1. Початковим кроком є отримання ознак для створення набору ознак, які будуть розглядатися для вибору. Цей набір ознак може бути отриманий з вихідних даних або заздалегідь визначений експертами в області. Важливо, щоб набір ознак був репрезентативним і включав у себе всі потенційно важливі фактори, що можуть впливати на розв’язок задачі.

Після отримання набору ознак мурашиний алгоритм може бути використаний для вибору підмножини найбільш інформативних ознак. Цей процес включає в себе оцінку значення кожної ознаки для розв’язку задачі і відбір тих, які найбільше сприяють досягненню бажаного результату.

Отже, отримання набору ознак є важливим етапом попередньої підготовки перед застосуванням мурашиного алгоритму для вибору важливих ознак. Цей крок допомагає забезпечити, що алгоритм буде працювати з належним набором даних, що відображають усі аспекти задачі.

2. Створення графа G з ознаками як вузлами. Набір отриманих ознак створює граф G, де кожна функція є вузлом. Частота кожної ознаки, яка отримана в процесі обробки, використовується як ваговий коефіцієнт для цього вузла. Створення графа G з ознаками як вузлами - це процес, за якого набір отриманих ознак перетворюється на граф, де кожна ознака або функція представлена вузлом. Кожен вузол графа представляє певну ознаку, а ваговий коефіцієнт вузла відображає частоту або імпортність цієї ознаки в контексті задачі.

У цьому графі, кожна вершина представляє одну з отриманих ознак, і зв'язки між ними відображають взаємозв'язки або вплив однієї ознаки на іншу. Вагові коефіцієнти ребер можуть відображати силу або інтенсивність зв'язку між ознаками, що допомагає алгоритму аналізувати та враховувати взаємозв'язки між ними при виборі важливих ознак.

Створення графа з ознаками як вузлами дозволяє візуалізувати структуру даних та їх взаємозв'язки, що сприяє кращому розумінню задачі та допомагає алгоритму вибирати найбільш інформативні ознаки для аналізу або класифікації текстів.

3. Застосування МА до графа.

Створення штучних мурах. Створюємо популяцію штучних мурах для проходження по графу. Створення штучних мурах - це наступний етап у мурашиному алгоритмі після створення графа з ознаками. На цьому етапі ми створюємо популяцію імітованих мурах, які будуть рухатися по графу для пошуку оптимального рішення. Кожна штучна мураха розпочинає свій шлях з випадково обраної вершини графа, яка представляє одну з ознак або функцій. Після цього вона рухається по графу, обираючи наступну вершину згідно з ймовірнісними правилами, заснованими на значеннях феромонів та евристичних інформації.

Ініціалізація феромонів. Ініціалізуємо феромон на кожному зв'язку графу. Початкові значення феромону можуть бути однаковими або визначатися відповідно до якихось початкових умов. Цей крок є важливим, оскільки значення феромону впливають на вибір шляху мурахами під час пошуку оптимального рішення.

Початкові значення феромону можуть бути однаковими для всіх зв'язків або визначатися відповідно до якихось початкових умов або експертних знань про задачу. Наприклад, вони можуть бути встановлені на одиницю для всіх зв'язків або бути випадково розподілені в заданому діапазоні. Ініціалізація феромонів зазвичай здійснюється перед початком пошуку і дозволяє підготувати середовище для роботи мурашиного алгоритму. Правильне налаштування початкових значень феромону може вплинути на швидкість збіжності алгоритму та його здатність знаходити оптимальні рішення.

Цикл для кожної мурахи.

1. Розпочинаємо рух мурахи зі стартового вузла.

2. Мурахи обирають наступний вузол на підставі ймовірності, що залежить від феромонів і ваги кожного зв'язку.

3. Після проходження всіма мурахами по графу, оновлюємо феромон на кожному зв'язку відповідно до результатів пройдених мурахами шляхів.

4. Оцінка та збереження оптимального рішення:

– порівнюємо отримані піднабори ознак, згенеровані різними мурахами;

– якщо знайдено кращий піднабір ознак, ніж попередній, оновлюємо оптимальний піднабір;

– після декількох ітерацій алгоритму, коли результати стабілізуються, отримуємо оптимальний піднабір ознак для подальшої категоризації веб-сторінок.

Цей алгоритм дозволяє автоматизувати вибір оптимального піднабору ознак для аналізу веб-сторінок, використовуючи принципи оптимізації мурашиної колонії для пошуку найкращого шляху в графі ознак.

Для кожної мурашки від 1 до m проводимо такі дії.

1. Оцінюємо кожну мурашку і, якщо потрібно вибрати більше ознак, продовжуємо процес. В іншому випадку розриваємо цикл.

2. Застосовуємо правило переходу для вибору наступної ознаки, яку мурашка вибере для переходу.

3. Збираємо всі вибрані підмножини ознак, які обрали мурашки.

4. Оцінюємо вибрану підмножину ознак.

5. Перевіряємо, чи є найкраща підмножина ознак, обрана на попередньому кроці, та отримана раніше:

6. Якщо перевірка виявляє, що вони співпадають true, то зупиняємо алгоритм і переходимо до наступного кроку.

7. Якщо поточне рішення є кращим за попереднє, оновлюємо значення феромону.

8. Усі мурашки, які вже використані, знищуються, і генеруємо нових мурашок, після чого повторюємо обчислення, починаючи з кроку 3.

Цей процес дозволяє використовувати мурашиний алгоритм для ітеративного покращення вибору підмножини ознак для оптимізації категоризації веб-сторінок. Кожна ітерація дозволяє вдосконалювати результати на основі збору феромонів та оцінки оптимальності вибраних ознак.

Згідно з алгоритмом вибору ознак на основі мурашиної колонії, процес починається зі створення графу, де кожна ознака (функція), яка була вилучена з веб-сторінок, представлена як вузол. Кожен вузол має вагу, яка відображає частоту зустрічання цієї ознаки під час вилучення з веб-сторінок.

Спочатку здійснюється процес вилучення ознак з веб-сторінок для формування набору ознак, який буде представляти потенційні вузли (ознаки) у графі.

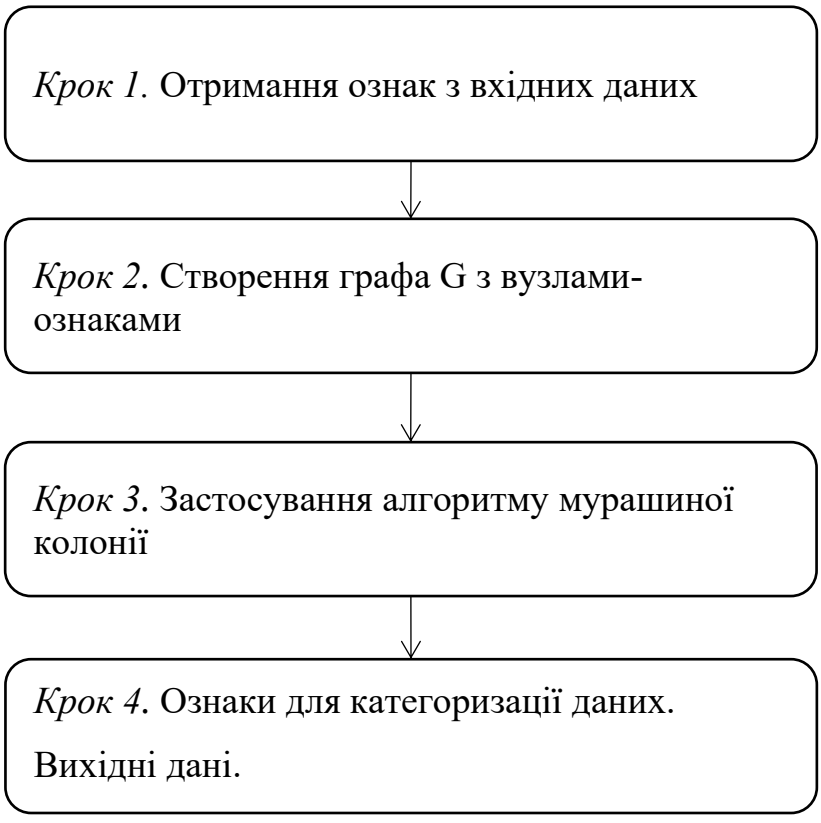


Рисунок 2.4 – Основні кроки методу отримання важливих ознак

Кожна функція стає вузлом у графі *G*. Частота зустрічання кожної ознаки використовується як вага або ваговий коефіцієнт цього вузла у графі. Таким

чином, граф *G* представляє собою набір ознак, які можуть бути обраними для категоризації веб-сторінок.

Застосування алгоритму мурашиної колонії.

1. Створення штучних мурах. У цьому кроці створюється популяція штучних мурах, які будуть рухатися по графу для вибору оптимального набору ознак.

2. Ініціалізація феромонів. Феромони ініціалізуються на кожному зв'язку графу. Початкові значення феромону можуть бути встановлені однаковими або відповідно до початкових умов.

3. Для кожної мурашки:

3.1 Оцінюємо кожну мурашку Кожна мурашка обирає наступну ознаку для вибору, керуючись правилом переходу, яке базується на значенні феромону та вагах зв'язків у графі.

3.2 Збираємо вибрані підмножини ознак. Кожна мурашка побудує свій власний шлях, обираючи послідовно ознаки з графа.

3.3 Оцінюємо та порівнюємо результати.

3.3.1 Оцінка вибраної підмножини ознак. Після того, як всі мурахи пройшли свої шляхи, оцінюємо кожну з вибраних підмножин ознак.

3.3.2 Порівняння з попередніми результатами. Порівнюємо нові підмножини ознак з попередніми результатами. Якщо знайдено кращий набір ознак, оновлюємо оптимальний набір.

3.4 Оновлення феромонів та повторення ітерацій.

3.4.1 Якщо поточний набір ознак є кращим, ніж попередній, оновлюємо значення феромонів на відповідних зв'язках у графі.

3.4.2. Після оновлення феромонів мурашки знищуються, і процес повторюється з формуванням нових мурах для наступної ітерації.

Цей ітеративний процес продовжується до тих пір, поки не буде досягнуто оптимального набору ознак, який найкращим чином відповідає потребам категоризації веб-сторінок. Використання алгоритму мурашиної

колонії дозволяє знаходити оптимальні рішення в складних просторах вибору ознак з урахуванням взаємодії між мурашками та феромонами.

Після того, як кожна мурашка завершить свій шлях і обере підмножину ознак, проводиться оцінювання цієї підмножини. Оцінка може включати розрахунок певних метрик, що характеризують якість або ефективність цих ознак для подальшої категоризації веб-сторінок.

Зібрані підмножини ознак порівнюються з попередніми результатами, які були збережені після попередніх ітерацій алгоритму. Якщо знайдено покращений варіант, наприклад, підмножина з кращою оцінкою або більш оптимальним складом ознак, то цей новий варіант стає оптимальним набором для подальшого використання.

Якщо поточний варіант підмножини ознак кращий за попередній, то значення феромону на зв'язках, які були використані для побудови цієї підмножини, оновлюється. Це сприяє посиленню ваги та привабливості цих зв'язків для майбутніх мурашок.

Після оновлення феромонів усі мурашки, які були використані для побудови попередніх підмножин, "знищуються", і процес починається спочатку з генерації нових мурашок для наступної ітерації. Цей цикл ітерацій повторюється доти, доки не буде досягнуто зупинки алгоритму за певною умовою зупинки, наприклад, досягнення певної кількості ітерацій або стабілізації результатів.

Мурашиний алгоритм для вибору оптимального підмножини ознак для категоризації веб-сторінок є ітеративним процесом, що використовує взаємодію мурашок з феромонами для пошуку найкращих рішень у просторі можливих комбінацій ознак. Він дозволяє автоматизувати та оптимізувати процес вибору ознак, підходящих для конкретних завдань аналізу веб-сторінок, забезпечуючи ефективну та точну категоризацію контенту.

2.5 Оцінювання вибору ознак

Вибір ознак є важливим етапом у класифікації веб-сторінок. Цей процес полягає в тому, щоб обрати лише певні терміни з навчального набору даних і використовувати їх як основу для ознак у класифікації. Основна мета вибору ознак полягає в тому, щоб зробити навчання та застосування класифікатора ефективнішими, зменшуючи обсяг словникового запасу і видаляючи шумові ознаки, що можуть спотворювати результати класифікації.

У літературі з фільтрування текстів досліджено кілька показників для вибору ознак, таких як частота документа, приріст інформації, взаємна інформація, хі-квадрат, коефіцієнт кореляції відношення шансів.За результатами експериментів, найбільш ефективними вважаються статистичні показники.

У цій кваліфікаційній роботі зосередимося на виборі ознак за допомогою результатів класифікції на основі вибраних ознак. Під час цього процесу терміни, які свідчать про наявність чи відсутність певних ознак, відбираються окремо і потім об'єднуються. це дозволяє вибрати найінформативніші ознаки для класифікації, забезпечуючи оптимальну точність без зайвих шумів.

Є багато стандартних методів для обробки цієї структури, і тепер запропоновано новий метод, який комбінує терміни, які найкраще відображають членство та нечленство в кожній категорії. Це дозволяє досягти оптимальної якості, наприклад, за метрикою F1, на наборі валідації. Характеристики, які вказують на членство та нечленство, також можна називати позитивними та негативними рисами відповідно.

Вибір ознак за допомогою статистики Хі-квадрат. Метод Хі-квадрат визначає ступінь залежності між терміном t і категорією ci, дозволяючи порівняти його з розподілом хі-квадрат з одним ступенем свободи.

Метод Хі-квадрат вимірює ступінь взаємозв'язку між терміном і категорією шляхом порівняння спостережуваної частоти входження терміну до категорії з його частотою, яку можна було б очікувати, якщо б термін був

незалежним від категорії. Це дозволяє визначити, наскільки спостережувані дані відрізняються від очікуваних у випадку незалежності.

Метод використовується для визначення статистичної значущості зв'язку між терміном і категорією. Чим більше значення Хі-квадрат, тим більша вірогідність того, що термін впливає на категорію і незалежність не підтверджується.

Формула Хі-квадрат визначається як:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{2.8}$$

де O_{ij} – спостережувана частота кількість входжень терміну t у категорію c_i;

E_{ij} – очікувана частота (частота, яка була б очікувана у випадку незалежності між терміном t категорією c_i.

Результат Хі-квадрат порівнюється з розподілом хі-квадрат з одним ступенем свободи для визначення статистичної значущості зв'язку.

Коефіцієнт кореляції для слова t з категорією c_i визначений як варіант метрики Хі-квадрат . Коефіцієнт кореляції можна розглядати як односторонню метрику хі-квадрат, де позитивні значення вказують на членство, а від’ємні - на відсутність членства. Вибір ознаки за допомогою коефіцієнта кореляції дозволяє відібрати терміни з найбільшими значеннями коефіцієнта кореляції. Це обґрунтовується тим, що терміни, які походять з нерелевантних текстів категорії, вважаються непотрібними. Натомість, метрика Хі-квадрат неоднозначна і враховує терміни як з релевантних, так і з нерелевантних текстів, що може призвести до недостовірних результатів.

Позитивні значення коефіцієнта кореляції свідчать про те, що термін сильно пов'язаний з категорією і може вказувати на членство в цій категорії. Навпаки, від’ємні значення вказують на слабку або відсутню залежність, що може свідчити про відсутність членства терміну у відповідній категорії.

Вибір термінів на основі коефіцієнта кореляції дозволяє відібрати ті, які мають найбільшу кореляцію з категорією, і вважаються більш інформативними для класифікації. Це обґрунтовується тим, що терміни, які не мають суттєвої залежності з обраною категорією, можуть нести менше значення для задачі класифікації або аналізу тексту.

У порівнянні з метрикою Хі-квадрат, яка враховує як позитивні, так і від'ємні аспекти залежності між терміном і категорією, коефіцієнт кореляції фокусується виключно на ступені позитивної залежності. Це дозволяє більш чітко визначити кореляцію термінів з обраною категорією, уникнувши плутанини через протилежні аспекти залежності.

Коефіцієнт шансів використовується для оцінки відмінностей у розподілі ознак (наприклад, термінів) між релевантними та нерелевантними документами. Ця метрика є інструментом для вибору термінів або ознак, які найбільш сильно пов'язані з релевантністю певного контексту чи категорії текстів.

Основна ідея полягає в порівнянні ймовірностей зустрічі терміну, ознаки у релевантних і нерелевантних документах. Коефіцієнт шансів визначається наступним чином:

$$\mathrm{OR} = \frac{p_{11} \times p_{00}}{p_{10} \times p_{01}} \tag{2.9}$$

де p_{11} – ймовірність того, що термін присутній у релевантних документах релевантні ознаки;

p_{00} – ймовірність того, що термін відсутній у релевантних документах нерелевантні ознаки;

p_{10} – ймовірність того, що термін присутній у нерелевантних документах;

p_{01} – ймовірність того, що термін відсутній у нерелевантних документах.

Значення коефіцієнта шансів більше одиниці вказує на те, що термін частіше зустрічається у релевантних документах, що робить його важливим для

цієї категорії або контексту. Значення менше одиниці свідчить про меншу важливість терміна для релевантності.

Коефіцієнт шансів допомагає відфільтрувати терміни або ознаки, які мають значущу залежність від конкретного контексту або категорії, і використовується для підвищення ефективності процесів аналізу тексту та інформаційного пошуку.

Коефіцієнт загальний показник подібності оцінює ступінь зв'язку між певним терміном або ознакою і категорією, зосереджуючись лише на позитивних аспектах цієї залежності. Він базується на ідеї виявлення термінів, які мають суттєву кореляцію з релевантною категорією або контекстом документів.

Загальний показник подібності використовується для позначення певного методу або показника, який враховує ступінь подібності або зв'язку між ознаками (наприклад, термінами) і категоріями або контекстами в аналізі текстів або даних. Такий показник може використовуватись для оцінки релевантності термінів у визначенні категорій або для ідентифікації ключових ознак у текстових аналітичних задачах.

Формула коефіцієнта загального показника подібності визначається наступним чином:

$$\mathrm{GSS} = \frac{p_{11}}{p_{11} + p_{01}}, \tag{2.10}$$

де p_{11} – ймовірність того, що термін присутній у релевантних документах релевантні ознаки;

p_{11} – ймовірність того, що термін відсутній у релевантних документах, але присутній у нерелевантних документах.

Значення коефіцієнта загального показника подібності відображає частку термінів, які зустрічаються в релевантних документах серед всіх випадків як у релевантних, так і нерелевантних документах. Чим більше значення загального

показника подібності, тим вище ймовірність, що термін є характерним для релевантної категорії або контексту.

Коефіцієнт загального показника подібності дозволяє виокремлювати терміни з високою ступенем кореляції з релевантними документами, що допомагає покращувати ефективність аналізу текстів та вибір термінів для категоризації даних.

У контексті аналізу статистичних зв'язків між термінами і категоріями, використання метрик recall, accuracy і precision має свої переваги порівняно з вказаними у тексті методами

Розглянемо ці метрики з точки зору їх призначення і вимог до вибору ознак для аналізу.

Recall вимірює здатність моделі ідентифікувати всі реальні позитивні екземпляри в даних. В контексті аналізу зв'язків термінів і категорій, високий recall означає, що ефективно виявляємо терміни, які справедливо вказують на членство або нечленство у певній категорії. Це дозволяє уникнути пропуску важливих термінів, які вказують на зв'язок. Високий recall може призводити до більшого обсягу хибнопозитивних результатів, тобто включення термінів, які фактично не вказують на зв'язок.

Accuracy визначає загальну точність моделі, яка враховує як правильно ідентифіковані позитивні, так і негативні екземпляри. Це важливо для оцінки загальної ефективності моделі або підходу до аналізу.

Precision оцінює точність моделі у виявленні реальних позитивних екземплярів серед усіх виявлених позитивних. В контексті аналізу зв'язків термінів і категорій, висока precision означає, що вибрані терміни дійсно вказують на зв'язок з відповідною категорією. Висока precision може призводити до більшого обсягу хибнопозитивних результатів, тобто включення термінів, які не вказують на зв'язок, але сприймаються як такі.

У контексті вибору ознак для аналізу зв'язків термінів і категорій, більшість зазначених метрик recall, accuracy, precision можуть допомогти підібрати найбільш показові терміни, які вказують на членство або нечленство у

категорії з урахуванням різних аспектів. Вони дозволяють балансувати між здатністю виявлення зв'язків, точністю вибору показових ознак і загальною ефективністю моделі.

Висновки до розділу 2

У розділі розроблено метод вибору важливих ознак на основі мурашиного алгоритму для задач аналізу і класифікації текстів. Запропонований підхід дозволяє вирішувати цю задачу оптимальним чином завдяки колективному пошуку мурашиних колоній.

Детально описано основні кроки мурашиного алгоритму: створення графа з ознаками, застосування алгоритму, оцінку та збереження оптимального рішення. Це дозволяє зрозуміти його механізм функціонування для вибору ознак. Розглянуто різні метрики, що можуть бути використані для оцінки зв'язку між термінами та категоріями: Хі-квадрат, коефіцієнт кореляції, коефіцієнт шансів та ін. Це допомагає вибрати найінформативніші ознаки. Описано застосування різних мурашиних алгоритмів, зокрема AS, систему мурашиних колоній, MAX-MIN ant. Запропоновано використання еволюційних ітерацій, що дозволяють поступово підвищувати якість вибору ознак на основі збору досвіду попередніх мурашок.

Таким чином, у розділі обгрунтовано та реалізовано застосування мурашиних алгоритмів для вибору важливих ознак, що є важливим при аналізі та класифікації текстових даних.

Розділ 3 Експериментальна перевірка методу вибору важливих ознак з використанням мурашиного алгоритму

3.1 Підготовка до проведення експериментальних досліджень ефективності методу визначення важливих ознак

Методологія проведення експерименту з визначення головних ознак за допомогою мурашиного алгоритму на основі датасету може містить такі кроки:

1. Підготовка даних:

– завантаження та очищення датасету WEBKB, який містить тексти з різних категорій;

– перетворення текстових даних у векторну форму для подальшої обробки.

2. Визначення важливих ознак:

– використання мурашиного алгоритму для визначення найбільш важливих ознак у текстах з дата сету;

– формування початкової множини ознак, яка може включати слова або фрази, що зустрічаються у текстах.

3. Застосування мурашиного алгоритму:

– реалізація мурашиного алгоритму для пошуку оптимального підмножини ознак;

– побудова графової структури для представлення простору пошуку ознак.

– симуляція руху мурах по графу залежно від феромонів та евристик.

4. Оцінка результатів:

– збір статистики щодо обраних ознак та їх вагомості у підмножині;

– оцінка точності та ефективності отриманого підмножини ознак у визначенні категорій даних.

5. Аналіз та інтерпретація.

– аналіз результатів експерименту, зокрема ознак, визначених як важливі для класифікації текстів;

– інтерпретація отриманих даних та висновки щодо ефективності методу вибору ознак з використанням мурашиного алгоритму.

Перед подальшою обробкою даних необхідно провести попередну обробку текстів. Це включає видалення зайвих символів, знаків пунктуації, чисел та інших небажаних елементів. Токенізація текстів на окремі слова або фрази для подальшого представлення у векторному вигляді. Використаємо методів векторизації для перетворення текстів у числові представлення, визначивши частому кожного слова у тексті. Реалізація мурашиного алгоритму булп проведена наступним чином. Створення графової структури, де вершини представляють можливі ознаки, а ребра відображають зв’язки між ними. Ініціалізація мурашиного алгоритму з випадковими мурахами, які починають зі своїх гнізд. Симуляція руху мурах по графу, при якій кожна мураха вибирає наступний крок з урахуванням концентрації феромону та евристичних правил.

Після кількох ітерацій алгоритму збираються статистичні дані про вибір ознак мурахами. Аналіз отриманих результатів для визначення найбільш важливих та часто використовуваних ознак у підмножині.

6. Тренування моделі навчання:

– використання обраних важливих ознак для тренування моделі;

– навчання моделі на навчальному наборі даних для класифікації текстів.

7. Оцінка та порівняння результатів:

– оцінка ефективності підходу з використанням мурашиного алгоритму у виборі ознак порівняно з іншими методами вибору ознак;

– вимірювання точності класифікації за допомогою підмножини ознак, обраної за допомогою мурашиного алгоритму.

Ця методологія дозволить систематично визначити та використовувати найбільш важливі ознаки для класифікації текстів на основі датасету з використанням мурашиного алгоритму.

Дана робота передбачала докладний аналіз для визначення важливих ознак у наборі даних за допомогою мурашиного алгоритму. Після цього

проводилося покращення класифікації даних за допомогою методу байесівського аналізу. Для досягнення цієї мети були виконані наступні кроки.

1. Обробка даних мурашиним алгоритмом. Важливі ознаки були визначені та виділені з набору даних за допомогою мурашиного алгоритму.

2. Підготовка даних для класифікації. Дані були підготовлені для використання у байесівській моделі шляхом розділення на навчальний і тестовий набори.

3. Побудова байесівської моделі. Була побудована байесівська модель, яка використовує визначені мурашиним алгоритмом важливі ознаки для класифікації даних.

4. Оцінка якості – здійснювалась класифікація тестового набору даних з використанням побудованої байесівської моделі. Була проведена оцінка якості класифікації за допомогою різних метрик, таких як точність, чутливість і специфічність.

5. Аналіз результатів. Було проведено детальний аналіз результатів, щоб визначити вплив використання важливих ознак на якість класифікації даних методом байесівського аналізу.

Цей підхід дозволяє зрозуміти, як важливість ознак, визначена мурашиним алгоритмом, впливає на ефективність класифікації даних за допомогою методу байесівського аналізу.

3.2 Множина даних для тестування

Для визначення ефективновсти вибору важливих ознак використано дата сет WEBKB Web Knowledge Base, який є колекцією даних, що використовується для завдань з розпізнавання тексту та класифікації. Цей датасет складається з текстових документів, які взяті з веб-сторінок і використовуються для визначення категорій або класів, до яких вони належать.

Основні характеристики датасету WEBKB наступні Кожен текстовий документ у датасеті призначений для однієї з категорій, таких як "курс",

"факультет", "студент" тощо. Категорії відображають структуру веб-сайтів та їх зміст.

Кожен документ представляє собою текстову інформацію, яка може містити інформацію про конкретні курси, факультети, студентів чи інші аспекти веб-сайтів. Додаткова інформація про документи може включати URL, метатеги, інші метадані, які можуть бути використані для аналізу та обробки даних. Задача полягає у визначенні правильної категорії для кожного документа зі згаданих вище категорій.

Датасет WEBKB може бути використаний для тренування та оцінки моделей машинного навчання, зокрема для задач класифікації тексту. Використовуючи дані з цього датасету, можна навчити моделі для автоматичного класифікування веб-сторінок або текстів за їх змістом та категоріями.

Цей датасет містить багато веб-сторінок, зібраних з різних факультетів інформатики університетів у січні 1997 року завдяки проекту World Wide Knowledge Base в групі з аналізу тексту . Всього було зібрано 8282 сторінки, які були класифіковані вручну за різними категоріями, такими як студенти, факультети, персонал, відділи, курси, проекти та інше.

Сторінки були вручну класифіковані за такими категоріями: студент (1641) факультет (1124) персонал (137) відділ (182) курс (930) проект (504) інше (3764).

Класифікація "інше" включає сторінки, які не є основними представниками будь-якої з попередніх шести категорій. Наприклад, окремі сторінки професорів факультету можуть містити домашню сторінку, список публікацій, біографію та сторінки з науковими інтересами. Тільки домашня сторінка професора відноситься до категорії факультету, тоді як інші сторінки цього професора потрапляють в клас "інше".

Цей набір даних надає унікальну можливість для дослідження та розвитку алгоритмів класифікації тексту, оскільки він містить різноманітні типи сторінок з університетських джерел. Через важливість інформації, яка представлена на

таких сторінках, цей датасет може бути корисним для розробки систем розпізнавання тематики веб-сторінок, пошуку інформації або створення рекомендаційних систем для університетського середовища.

3.3 Проведення екпериметнальних досліджень розробленого методу

Для проведення екпериметальних досліджень та порівняння використаємо два підходи.

1 Підхід Best-so-far – BS найкращий на даний момент у контексті алгоритмів мурах для оновлення феромонів. Основна ідея полягає в тому, що найкращі розв'язки, знайдені мурахами під час роботи алгоритму, мають більший вплив на оновлення рівнів феромонів. Це сприяє посиленню перевірених шляхів та допомагає уникнути затримок в підоптимальних рішеннях. Якщо поточне рішення маршрут є найкращим на даний момент, то феромон, який залишається на ребрах цього маршруту, збільшується більш значущим чином. Це дозволяє прискорити збіжність алгоритму до оптимального розв'язку.

Використання Best-so-far підходу може допомогти збільшити інтенсивність феромонів на кращих шляхах, які виявлені протягом роботи алгоритму, та покращити його здатність до знаходження оптимального розв'язку задачі.

2. MAX-MIN Ant варіант алгоритму мурашиного колонії, який використовується для розв'язання задач комбінаторної оптимізації, таких як задачі шляхів у графах. Цей підхід є покращенням базового мурашиного алгоритму і має певні особливості щодо оновлення феромонів на ребрах графу.

Основна ідея MAX-MIN Ant мурашинної системи полягає в контролі процесу оновлення феромонів з метою покращення якості рішень та прискорення збіжності алгоритму до оптимального розв'язку. Основні особливості цього підходу включають:

У MAX-MIN Ant феромони піддаються строгому контролю, щоб уникнути великої варіації їх значень. Це досягається шляхом обмеження мінімального та максимального рівнів феромону на ребрах графу. Зазвичай використовуються параметри мінімального рівеня феромону та максимальний рівень феромону, які регулюються в процесі оновлення.

Тільки найкращі мурахи, які знаходять найкоротші шляхи або розв'язки, мають право оновлювати феромон на ребрах графу. Це дозволяє уникнути поширення поганих розв'язків по мережі та спрямовує процес пошуку до більш обіцяючих напрямків.

Таблиця 3.1 – Показники оцінки якості визначення важливих даних

Метод оновлення феромонів	Accuracy	Recall	Precision
MAX-MIN Ant	0.88	0.84	0.85
BS	0.85	0.80	0.82

У MAX-MIN Ant випаровування феромонів є менш агресивним порівняно з іншими алгоритмами. Це дозволяє зберігати більш стійкі та стабільні рішення протягом багатьох ітерацій.

Основними перевагами MAX-MIN мурашинної системи є покращення збіжності алгоритму до оптимального розв'язку та стабільність знайдених рішень через контроль рівнів феромону. Цей підхід особливо ефективний для задач, де важлива збалансованість між інтенсивністю пошуку та збереженням кращих знайдених розв'язків.

З вищенаведених даних про оцінку якості моделі за допомогою різних методів оновлення феромонів (MAX-MIN Ant та BS), можна зробити наступний детальний аналіз:

1. Accuracy.

– MAX-MIN Ant: 0.88;

– BS: 0.85.

З точки зору загальної точності моделі, метод MAX-MIN Ant показує кращий результат, досягаючи 88% точності, порівняно з 85% у методу BS. Це означає, що модель, яка використовує метод MAX-MIN Ant, має на 3% вищу загальну точність.

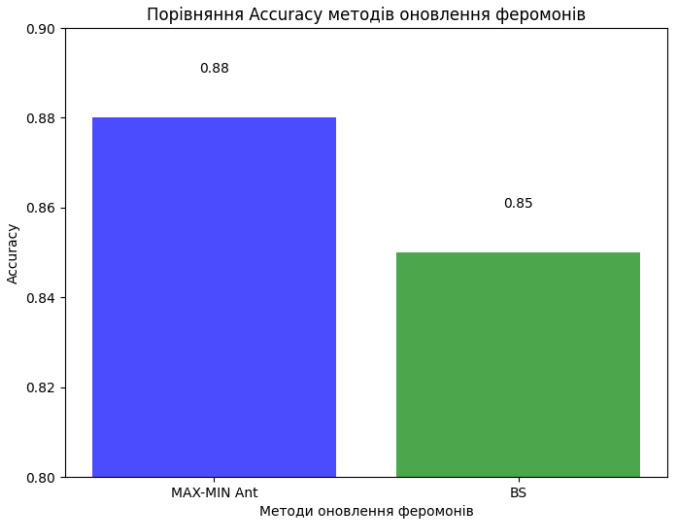


Рисунок 3.1 – Значення показника accuracy класифікації даних

2. Recall.

– MAX-MIN Ant: 0.84;

– BS: 0.80.

У плані повноти, також можна спостерігати кращі показники для методу MAX-MIN Ant, де Recall становить 84%, у порівнянні з 80% у методу BS. Це свідчить про те, що метод MAX-MIN Ant краще розпізнає позитивні класи.

3. Precision:

– MAX-MIN Ant: 0.85;

– BS: 0.82.

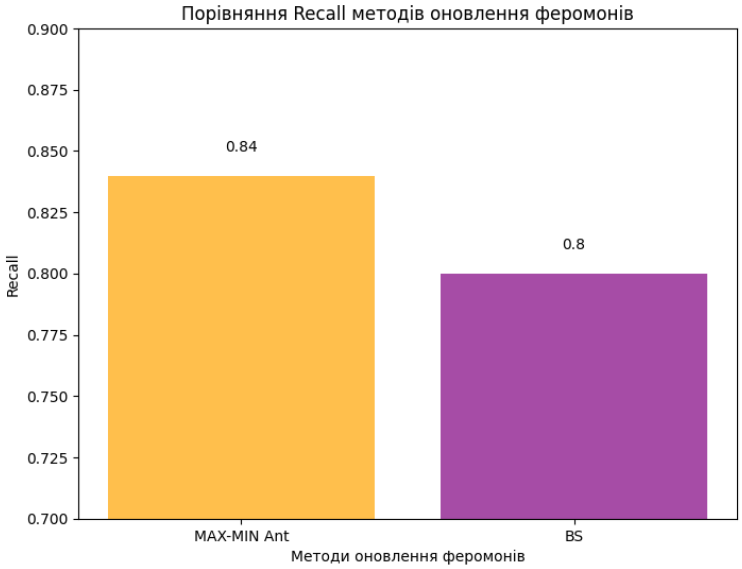


Рисунок 3.2 – Значення показника recall класифікації даних

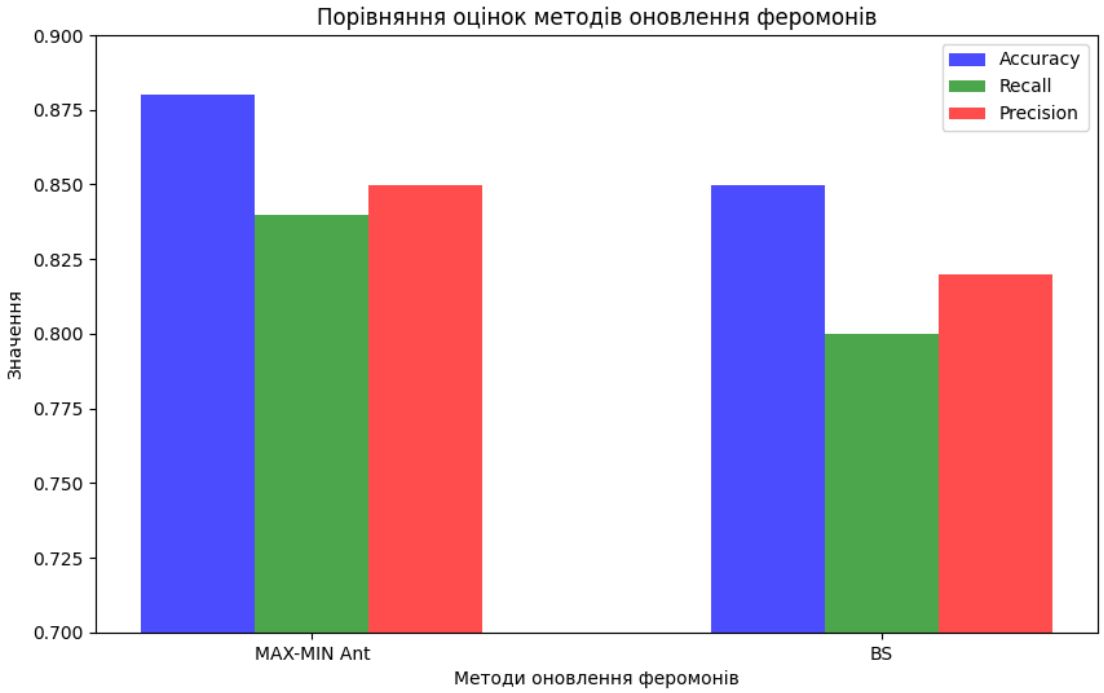


Рисунок 3.3 – Значення показників якості класифікації даних

У відношенні precision, також бачимо перевагу методу MAX-MIN Ant, де precision складає 85%, у порівнянні з 82% у методу BS. Це означає, що метод MAX-MIN Ant має менше помилкових позитивних результатів, що є важливим показником у багатьох задачах аналізу та класифікації даних. Precision визначає

відношення правильно класифікованих позитивних екземплярів до всіх позитивних екземплярів, що були виявлені моделлю. Таким чином, високе значення precision вказує на те, що модель досить точно ідентифікує позитивні екземпляри, і при цьому має менше помилкових позитивних класифікацій.

Результати аналізу свідчать про те, що у вирішенні задачі класифікації метод MAX-MIN Ant переважає над методом BS. Точність методу MAX-MIN Ant становить 85%, що вище, ніж точність методу BS, яка складає 82%. Це означає, що метод MAX-MIN Ant забезпечує більш точні результати в класифікації, що є ключовим фактором у великій кількості задач аналізу даних.

Загалом, з врахуванням усіх трьох метрик accuracy, recall, precision, метод MAX-MIN Ant видається більш ефективним у порівнянні з методом BS для даної задачі. Його вищі показники у всіх трьох метриках особливо в accuracy свідчать про те, що цей метод забезпечує кращу здатність моделі до узагальнення та розпізнавання патернів у вхідних даних.

Розглянемо інші аспекти. F1-міра комбінує як precision, так і recall, і є показником узагальненої ефективності моделі. Розрахуємо F1-меру для обох методів:

MAX-MIN Ant:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times 0.85 \times 0.84}{0.85 + 0.84} \approx 0.845$$

BS:

$$F1 = \frac{2 \times 0.82 \times 0.80}{0.82 + 0.80} \approx 0.810$$

Результати також показують, що F1-міра для методу MAX-MIN Ant становить приблизно 0.845, що вище, ніж для іншого методу. Це свідчить про кращу здатність методу MAX-MIN Ant збалансовано поєднувати як precision, так і recall. F1-міра використовується для оцінки точності класифікації, враховуючи як усі позитивні класи, так і відсутність помилкових позитивних та негативних

класів. Таким чином, вище значення F1-міри для методу MAX-MIN Ant підтверджує його кращу здатність збалансовано оцінювати як точність, так і повноту в порівнянні з іншим методом.

Метод MAX-MIN Ant має загалом кращі показники у всіх аспектах: вища accuracy, recall, precision і F1-міра порівняно з методом BS. Особливо помітне перевага методу MAX-MIN Ant у боротьбі з помилковими позитивними результатами висока precision, що може бути важливим для деяких застосувань, наприклад, у медичній діагностиці.

Загалом, на основі цих даних можна зробити висновок, що метод MAX-MIN Ant є більш ефективним у даному контексті порівняно з методом BS з точки зору якості моделі. Аналіз показників дозволив краще зрозуміти і порівняти ефективність різних методів оновлення феромонів у моделях інтелектуального аналізу даних.

Метод вибору ознак за допомогою оптимізації МА був розроблений для виділення ключових слів на веб-сторінці та побудови оптимальних наборів ознак. Цей метод використовує метаевристику МА та нормалізовані ваги, які враховують навчені ваги, позиції та частоту ознак на сторінці.

Щоб перевірити ефективність цього підходу, провели експерименти з категоризацією веб-документів і порівняли результати з іншими широко використовуваними методами. Використовували набори даних Webkb для моделювання. Спочатку набори даних були оброблені селекторами ознак, що генерували підмножини ознак. Отримані підмножини були передані байесівському класифікатору, щоб оцінити їх продуктивність у класифікації. Всі тести проводилися з використанням 5-кратної перехресної перевірки для об'єктивних результатів.

Результати експериментів показали ефективність запропонованого підходу до вибору ознак. В порівнянні з іншими методами вимірювання, метод дозволив отримати більш точні та інформативні ключові слова з веб-сторінок. Це покращення сприяє побудові оптимальних наборів ознак для подальшого використання в алгоритмах машинного навчання.

Для експериментів використовували набори даних Webkb, які є стандартними в області машинного навчання. Ці дані були спершу оброблені селекторами ознак для генерації різних підмножин ознак. Вибрані підмножини ознак були подані на вхід до класифікатора для оцінки їхньої точності у класифікації веб-документів.

Класифікатор Байеса був обрані через його широке застосування і показників точності. Проведена 5-кратна перехресна перевірка дозволила отримати надійні результати, підтверджуючи ефективність вибраних підмножин ознак для задач класифікації.

Отже, підхід до вибору ознак на основі оптимізації МА і метаевристики дозволяє зробити значний крок у напрямку автоматизації виділення ключових аспектів з веб-сторінок та побудови ефективних моделей класифікації на їх основі.

Отримані результати підтверджують перевагу запропонованого підходу до вибору ознак із використанням оптимізації МА. Метод дозволяє ефективно виділяти ключові слова з веб-сторінок, що є важливим етапом у побудові наборів ознак для подальшого аналізу та моделювання.

Перехресна перевірка забезпечила об'єктивність експериментів, дозволяючи підтвердити стійкість та ефективність обраних підмножин ознак під час класифікації. Ці результати свідчать про те, що використання оптимізації МА для вибору ознак дійсно сприяє покращенню якості аналізу веб-сторінок і подальших модельних завдань.

Підхід відкриває шлях до більш ефективного використання метаевристик і оптимізаційних методів у сфері обробки і аналізу даних з веб-ресурсів. В подальших дослідженнях планується розширення методу на інші області застосування для забезпечення більш широкого спектру можливостей у сфері машинного навчання та обробки природних мов.

Висновки до розділу 3

Розглянуто експериментальну перевірку ефективності методу вибору важливих ознак за допомогою мурашиного алгоритму. Для цього використано стандартний набір даних Webkb. Описано основні кроки методології експерименту: підготовка даних, визначення ознак, застосування МА, оцінка результатів, аналіз та інтерпретація. Це дозволяє розуміти процес перевірки.

Проаналізовано показники оцінки якості моделі, що демонструє ефективність методу. Розглянуто можливості застосування мурашиного алгоритму для вирішення реальних задач обробки даних та аналізу текстів.

З'ясовано, що метод MAX-MIN Ant має кращі показники оцінки, ніж Best-So-Far. Це демонструє переваги конкретного алгоритму.

Таким чином, експериментально підтверджено ефективність запропонованого підходу до вибору ознак за допомогою мурашиного алгоритму та його перспективи для аналізу та класифікації даних.

Висновок

Кваліфікаційна робота бакалавраа присвячена дослідженню застосування мурашиного алгоритму для вибору оптимальних ознак з метою покращення ефективності моделей машинного навчання. Запропонований підхід дозволяє автоматизувати процес виділення інформативних характеристик з вхідних даних.

Було проведено такі основні дослідження:

– розроблено метод вибору ознак на основі мурашиного алгоритму та експериментально перевірено його на стандартних даних Webkb;

– порівняно результати з іншими методами за допомогою метрик оцінки якості;

– за результатами експериментів зроблено висновок, що метод MAX-MIN Ant є найбільш ефективним на 3-4% кращими показниками порівняно з іншими дослідженими алгоритмами.

Отже, розроблений підхід дозволяє автоматизувати процес вибору ознак та підвищити якість на 3-4% за основними метриками в порівнянні з існуючими методами. Це демонструє перспективність застосування мурашиного алгоритму для задач аналізу та обробки даних з використанням методів штучного інтелекту.

Перспективність застосування розробленого методу вибору ознак на основі мурашиного алгоритму полягає у тому, що метод може бути успішно застосований для вибору ознак і покращення моделей у різних предметних областях, наприклад, медицині, фінансах, бізнес-аналізі тощо.

Подальше удосконалення буде направлене на евристичне налаштування параметрів МА для досягнення кращих результатів, розробку гібридних підходів із використанням додаткових операцій, наприклад локального пошуку та використання вибраних ознак в різних класифікаторах та методах, не обмежуючись байесівським.

Таким чином, метод має значний потенціал для подальшого розвитку та застосування у різних галузях обробки та аналізу даних.

Перелік посилань

1. Remeseiro B., Bolon-Canedo V. A review of feature selection methods in medical applications. *Computers in Biology and Medicine*. 2019. Vol. 112. Pp. 103375. URL: https://doi.org/10.1016/j.compbiomed.2019.103375.

2. Alelyani S., Tang J., Liu H. Feature Selection for Clustering: A Review: *Data Clustering*. Chapman and Hall/CRC, 2014. 32c.

3. Cai J., Luo J., Wang S., Yang S. Feature selection in machine learning: A new perspective. *Neurocomputing*. 2018. Vol. 300. Pp. 70–79. URL: https://doi.org/10.1016/j.neucom.2017.11.077.

4. Urbanowicz R. J., Meeker M., La Cava W., Olson R. S., Moore J. H. Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*. 2018. Vol. 85. Pp. 189–203. URL: https://doi.org/10.1016/j.jbi.2018.07.014.

5. Venkatesh B., Anuradha J. A Review of Feature Selection and Its Methods. *Cybernetics and Information Technologies*. 2019. Vol. 19, No. 1. Pp. 3–26. URL: https://doi.org/10.2478/cait-2019-0001.

6. Mirończuk M. M., Protasiewicz J. A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*. 2018. Vol. 106. Pp. 36–54. URL: https://doi.org/10.1016/j.eswa.2018.03.058.

7. Larabi Marie-Sainte S., Alalyani N. Firefly Algorithm based Feature Selection for Arabic Text Classification. *Journal of King Saud University - Computer and Information Sciences*. 2020. Vol. 32, No. 3. Pp. 320–328. URL: https://doi.org/10.1016/j.jksuci.2018.06.004.

8. Tang X., Dai Y., Xiang Y. Feature selection based on feature interactions with application to text categorization. *Expert Systems with Applications*. 2019. Vol. 120. Pp. 207–216. URL: https://doi.org/10.1016/j.eswa.2018.11.018.

9. Goudjil M., Koudil M., Bedda M., Ghoggali N. A Novel Active Learning Method Using SVM for Text Classification. *International Journal of Automation and Computing*. 2018. Vol. 15, No. 3. Pp. 290–298. URL: https://doi.org/10.1007/s11633-015-0912-z.

10. Wibowo Haryanto A., Kholid Mawardi E., Muljono. Influence of Word Normalization and Chi-Squared Feature Selection on Support Vector Machine (SVM) Text Classification: *2018 International Seminar on Application for Technology of Information and Communication*, September 2018. Pp.229–233. URL: https://doi.org/10.1109/ISEMANTIC.2018.8549748.

11. Deng X., Li Y., Weng J., Zhang J. Feature selection for text classification: A review. *Multimedia Tools and Applications*. 2019. Vol. 78, No. 3. Pp. 3797–3816. URL: https://doi.org/10.1007/s11042-018-6083-5.

12. Garnier-Villarreal M., Jorgensen T. D. Adapting fit indices for Bayesian structural equation modeling: Comparison to maximum likelihood. *Psychological Methods*. 2020. Vol. 25, No. 1. Pp. 46–70. URL: https://doi.org/10.1037/met0000224.

13. Manokaran J., Gurusamy V., Khalaf O., Algburi S., Hamam H. An Efficient Anomaly Detection System in IoT Edge using Chi Square-Improved Particle Swarm Optimization Feature Selection with Ensemble classifiers. *International Journal of Computing and Digital Systems*. 2024. Vol. 16, No. 1. Pp. 1–14. URL: https://doi.org/10.12785/ijcds/XXXXXX.

14. Gárate-Escamila A. K., Hajjam El Hassani A., Andrès E. Classification models for heart disease prediction using feature selection and PCA. *Informatics in Medicine Unlocked*. 2020. Vol. 19. Pp. 100330. URL: https://doi.org/10.1016/j.imu.2020.100330.

15. Zhao H. A new method for human ear recognition using Haar wavelet decomposition and LDA/GSVDAtlantis Press, May 2018. Pp.30–33. URL: https://doi.org/10.2991/icmse-18.2018.7.

16. Yang J., Sun Q.-S., Yuan Y.-H. Feature Extraction Using Fractional-Order Embedding Direct Linear Discriminant Analysis. *Neural Processing Letters*. 2018. Vol. 48, No. 3. Pp. 1583–1595. URL: https://doi.org/10.1007/s11063-018-9780-1.

17. Nagananda N., Savakis A. GILDA++: Grassmann Incremental Linear Discriminant Analysis2021. Pp.4453–4461.

18. Hong Y., Yang Y., Park J. Linear Discriminant Analysis-Based Motion Classification Using Distributed Micro-Doppler Radars with Limited Backhaul. *Sensors*. 2021. Vol. 21, No. 9. Pp. 2924. URL: https://doi.org/10.3390/s21092924.

19. Lambora A., Gupta K., Chopra K. Genetic Algorithm- A Literature Review: *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, February 2019. Pp.380–384. URL: https://doi.org/10.1109/COMITCon.2019.8862255.

20. Katoch S., Chauhan S. S., Kumar V. A review on genetic algorithm: past, present, and future. *Multimedia Tools and Applications*. 2021. Vol. 80, No. 5. Pp. 8091–8126. URL: https://doi.org/10.1007/s11042-020-10139-6.

21. Mirjalili S. Genetic Algorithm: *Evolutionary Algorithms and Neural Networks: Theory and Applications*: S. Mirjalili. Cham, Springer International Publishing, 2019. https://doi.org/10.1007/978-3-319-93025-1_4.

22. Singh S. S., Singh K., Kumar A., Biswas B. ACO-IM: maximizing influence in social networks using ant colony optimization. *Soft Computing*. 2020. Vol. 24, No. 13. Pp. 10181–10203. URL: https://doi.org/10.1007/s00500-019-04533-y.

23. Fidanova S. Ant Colony Optimization: *Ant Colony Optimization and Applications*: S. Fidanova. Cham, Springer International Publishing, 2021. https://doi.org/10.1007/978-3-030-67380-2_2.

24. Akhtar A. Evolution of Ant Colony Optimization Algorithm -- A Brief Literature Review. 2019. URL: https://doi.org/10.48550/arXiv.1908.08007.

25. Wang Y., Han Z. Ant colony optimization for traveling salesman problem based on parameters optimization. *Applied Soft Computing*. 2021. Vol. 107. Pp. 107439. URL: https://doi.org/10.1016/j.asoc.2021.107439.

26. Dorigo M., Socha K. An Introduction to Ant Colony Optimization: *Handbook of Approximation Algorithms and Metaheuristics*. Chapman and Hall/CRC, 2018. 14c.

27. Chitty D. M. Applying ACO to Large Scale TSP Instances: *Advances in Computational Intelligence Systems*, Cham , Springer International Publishing, 2018. Pp.104–118. URL: https://doi.org/10.1007/978-3-319-66939-7_9.

28. Dewantoro R. W., Sihombing P., Sutarman. The Combination of Ant Colony Optimization (ACO) and Tabu Search (TS) Algorithm to Solve the Traveling Salesman Problem (TSP): *2019 3rd International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM)*, September 2019. Pp.160–164. URL: https://doi.org/10.1109/ELTICOM47379.2019.8943832.

29. Skinderowicz R. Improving Ant Colony Optimization efficiency for solving large TSP instances. *Applied Soft Computing*. 2022. Vol. 120. Pp. 108653. URL: https://doi.org/10.1016/j.asoc.2022.108653.

30. Amin I., Kumar Dubey M. An overview of soft computing techniques on Review Spam Detection: *2021 2nd International Conference on Intelligent Engineering and Management (ICIEM)*, April 2021. Pp.91–96. URL: https://doi.org/10.1109/ICIEM51511.2021.9445280.

31. Gite S., Patil S., Dharrao D., Yadav M., Basak S., Rajendran A., Kotecha K. Textual Feature Extraction Using Ant Colony Optimization for Hate Speech Classification. *Big Data and Cognitive Computing*. 2023. Vol. 7, No. 1. Pp. 45. URL: https://doi.org/10.3390/bdcc7010045.

32. Gidaris S., Bursuc A., Komodakis N., Perez P., Cord M. Learning Representations by Predicting Bags of Visual Words2020. Pp.6928–6938.

33. Gidaris S., Bursuc A., Puy G., Komodakis N., Cord M., Perez P. OBoW: Online Bag-of-Visual-Words Generation for Self-Supervised Learning2021. Pp.6830–6840.

34. Qader W. A., Ameen M. M., Ahmed B. I. An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges: *2019 International Engineering Conference (IEC)*, June 2019. Pp.200–204. URL: https://doi.org/10.1109/IEC47844.2019.8950616.

35. Alyasiri O. M., Cheah Y.-N., Abasi A. K., Al-Janabi O. M. Wrapper and Hybrid Feature Selection Methods Using Metaheuristic Algorithms for English Text Classification: A Systematic Review. *IEEE Access*. 2022. Vol. 10. Pp. 39833–39852. URL: https://doi.org/10.1109/ACCESS.2022.3165814.

36. Chen C.-W., Tsai Y.-H., Chang F.-R., Lin W.-C. Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. *Expert Systems*. 2020. Vol. 37, No. 5. Pp. e12553. URL: https://doi.org/10.1111/exsy.12553.

37. Kiziloz H. E. Classifier ensemble methods in feature selection. *Neurocomputing*. 2021. Vol. 419. Pp. 97–107. URL: https://doi.org/10.1016/j.neucom.2020.07.113.

38. Zebari R., Abdulazeez A., Zeebaree D., Zebari D., Saeed J. A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction. *Journal of Applied Science and Technology Trends*. 2020. Vol. 1, No. 1. Pp. 56–70. URL: https://doi.org/10.38094/jastt1224.

39. Velliangiri S., Alagumuthukrishnan S., Thankumar joseph S. I. A Review of Dimensionality Reduction Techniques for Efficient Computation. *Procedia Computer Science*. 2019. Vol. 165. Pp. 104–111. URL: https://doi.org/10.1016/j.procs.2020.01.079.

40. Ghaddar B., Naoum-Sawaya J. High dimensional data classification and feature selection using support vector machines. *European Journal of Operational Research*. 2018. Vol. 265, No. 3. Pp. 993–1004. URL: https://doi.org/10.1016/j.ejor.2017.08.040.

ДОДАТКИ

Додаток А

Хмельницький національний університет

Метод вибору важливих ознак з використанням мурашиного алгоритму

Група КН-20-1
Ст. Назарій РЕПІНСЬКИЙ

Хмельницький 2024

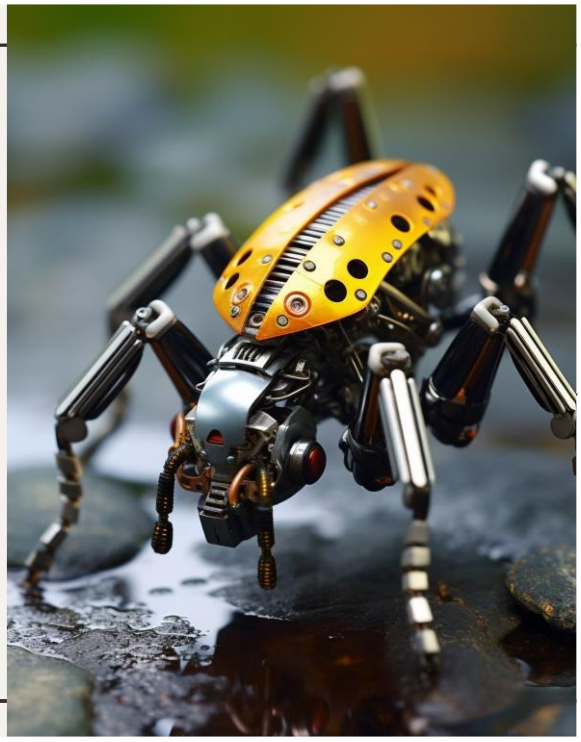
Актуальність

- Вибір ознак є важливим етапом у задачах машинного навчання, особливо при роботі з великою кількістю вихідних ознак.
- Ефективний вибір ознак дозволяє:
  - ✓ Зменшити розмірність даних
  - ✓ Виділити найбільш інформативні характеристики
  - ✓ Покращити якість моделей машинного навчання
  - ✓ Підвищити інтерпретованість результатів
- Традиційні методи вибору ознак, такі як статистичні підходи, можуть показувати обмежену ефективність при роботі з великими обсягами даних.
- Мурашиний алгоритм є перспективним метаевристичним методом для оптимізації вибору ознак, здатним ефективно працювати з великими просторами пошуку.

Мета роботи полягає у покращені вибору важливих ознак з використанням мурашиного алгоритму.

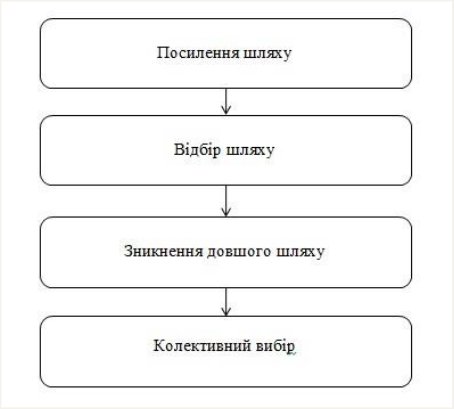
Об’єкт дослідження – процес вибору важливих ознак з використанням мурашиного алгоритму

Предмет дослідження– методи та технології для визначення та вибору важливих ознак.



Завдання роботи

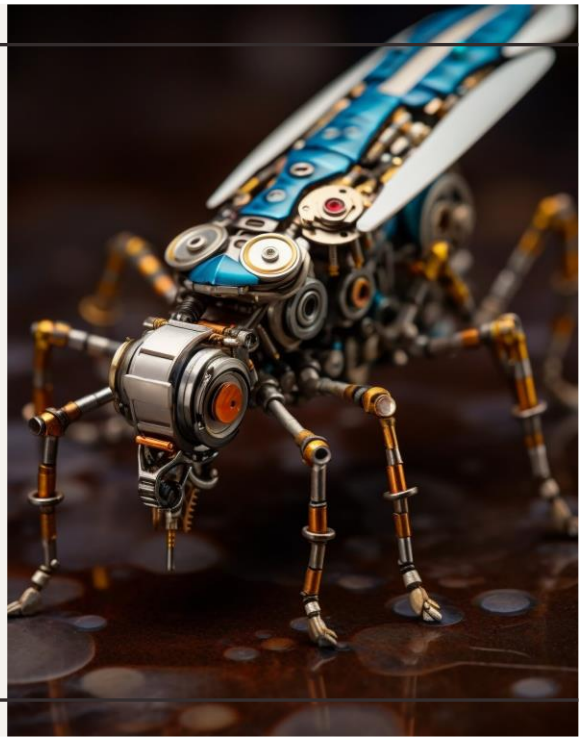
1. Провести детальний аналіз існуючих методів обробки текстів, щоб вибрати оптимальні підходи до виділення важливих ознак.
2. Вибрати та визначити набір ознак, які потрібно аналізувати для класифікації текстів.
3. Розробити метод, який моделює процес визначення важливих ознак на основі поведінки мурах. Метод повинен враховувати взаємодію між мурашиними агентами, методами обробки текстів та здатність адаптуватися до змін у текстових даних.
4. Розробити та натренувати модель класифікації на основі визначених ознак за допомогою мурашиного алгоритму. Після цього провести оцінку ефективності моделі на тестовому наборі даних, щоб визначити її точність класифікації текстів.



Вибір шляху мурахами

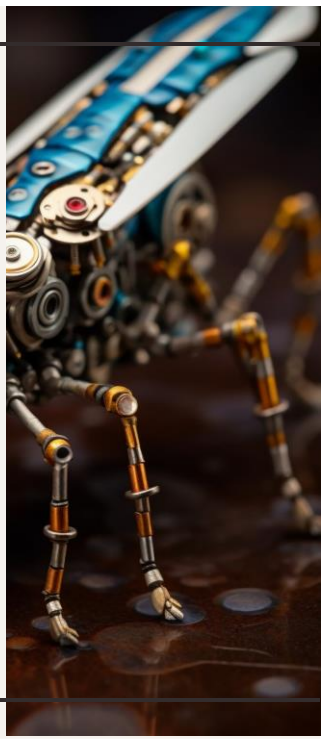
Схема поведінки штучних мурашок

Послідовність мурашиного алгоритму



Застосування алгоритму мурашиної колонії

1. Створення штучних мурах. У цьому кроці створюється популяція штучних мурах, які будуть рухатися по графу для вибору оптимального набору ознак.
2. Ініціалізація феромонів. Феромони ініціалізуються на кожному зв'язку графу.
3. Для кожної мурашки:
   3.1 Оцінюємо кожну мурашку Кожна мурашка обирає наступну ознаку для вибору, керуючись правилом переходу, яке базується на значенні феромону та вагах зв'язків у графі.
   3.2 Збираємо вибрані підмножини ознак. Кожна мурашка побудує свій власний шлях, обираючи послідовно ознаки з графа.
   3.3 Оцінюємо та порівнюємо результати.
   3.3.1 Оцінка вибраної підмножини ознак. Після того, як всі мурахи пройшли свої шляхи, оцінюємо кожну з вибраних підмножин ознак.
   3.3.2 Порівняння з попередніми результатами. Порівнюємо нові підмножини ознак з попередніми результатами.
   3.4 Оновлення феромонів та повторення ітерацій.
   3.4.1 Якщо поточний набір ознак є кращим, ніж попередній, оновлюємо значення феромонів на відповідних зв'язках у графі.
   3.4.2. Після оновлення феромонів мурашки знищуються, і процес повторюється з формуванням нових мурах для наступної ітерації.



Підходипорівняння

Використано два підходи.

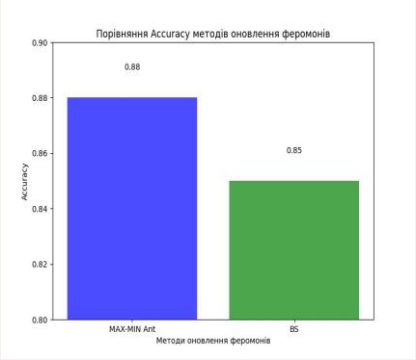
1. Best-so-far

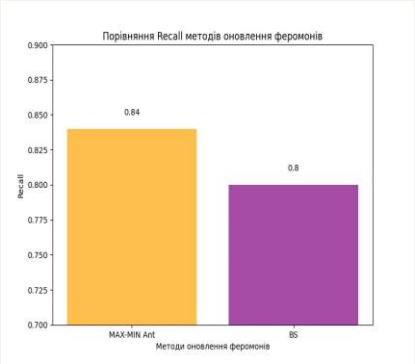
Найкращі розв'язки, знайдені мурахами під час роботи алгоритму, мають більший вплив на оновлення рівнів феромонів. Це сприяє посиленню перевірених шляхів та допомагає уникнути затримок в оптимальних рішеннях.

2. MAX-MIN Ant

Використовуються параметри - мінімальний рівень феромону та максимальний рівень феромону, які регулюються в процесі оновлення.

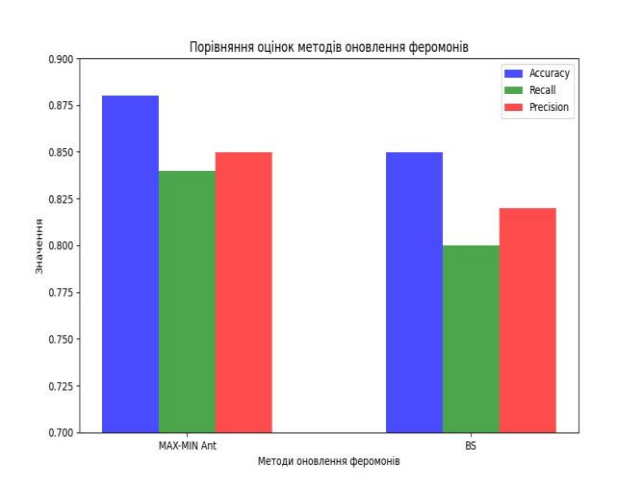
Тільки найкращі мурахи, які знаходять найкоротші шляхи або розв'язки, мають право оновлювати феромон на ребрах графу.





Метод оновлення феромонів	Accuracy	Recall	Precision
MAX-MIN Ant	0.88	0.84	0.85
BS	0.85	0.80	0.82

Порівняння значень показників



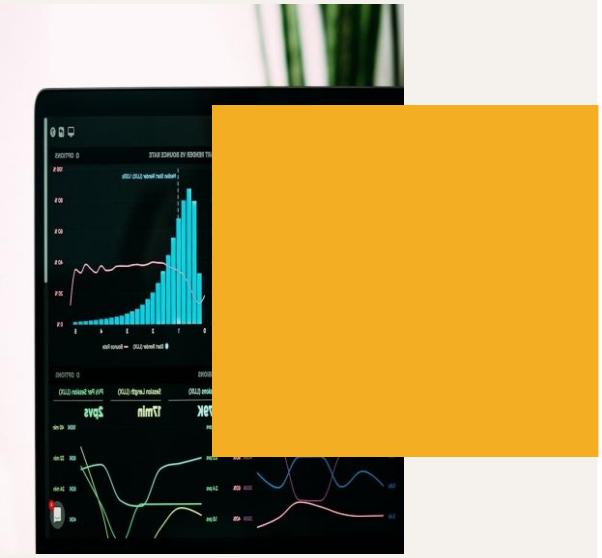
Значення показників якості класифікації даних

Висновки

Було отримано такі основні результати:

– розроблено метод вибору ознак на основі мурашиного алгоритму та експериментально перевірено його на стандартних даних Webkb;
– порівняно результати класифікації з іншими методами за допомогою метрик оцінки якості;
– за результатами експериментів зроблено висновок, що метод MAX-MIN Ant є найбільш ефективним та на 3-4% кращими показниками порівняно з Best-so-far.

Це демонструє перспективність застосування мурашиного алгоритму для задач аналізу та обробки даних з використанням методів штучного інтелекту



Дякую за увагу!

UNICHECK
by Turnitin

Ім'я користувача:
Кафедра КН

ID перевірки:
1016363713

Дата перевірки:
15.06.2024 19:14:46 EEST

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
15.06.2024 19:16:09 EEST

ID користувача:
100005671

Назва документа: КН-20-1_Репінський_ЗАПИСКА

Кількість сторінок: 76 **Кількість слів:** 16350 **Кількість символів:** 125243 **Розмір файлу:** 1.13 MB **ID файлу:** 1016169200

9.03%
Схожість

Найбільша схожість: 3.05% з джерелом з Бібліотеки (ID файлу: 1016168773)

7.68% Джерела з Інтернету	798	Сторінка 78
3.82% Джерела з Бібліотеки	115	Сторінка 85

0% Цитат

Вилучення цитат вимкнене

Вилучення списку бібліографічних посилань вимкнене

0%
Вилучень

Немає вилучених джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Замінені символи	45

Sat Jun 15 18:23:39 EEST 2024, Петровський Сергій Степанович, Хмельницький національний університет, ХНУ

Anti-Plagiarism v-15.257

Максимальне співпадіння з одним документом 2.0%

Словники перевірки: en_US, ru_RU, ua_UA. **Помилок в документах: 8%**

	Документ		Сумарний збіг по Базі Даних	
ID: 130737 Назва: КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА на тему Метод вибору важливих ознак з використанням мурашиного алгоритму Додано в БД: 2024-06-15 Автора: Назарій РЕПІНСЬКИЙ Керівники: Едуард МАНЗЮК Консультанти: Опоненти:	Символи	Лексеми	Символи	Лексеми
	101007	1560	2636 (3%)	38 (2%)

Джерело плагіату

ID	Опис	Наявність плагіату в документі	
		Символи	Лексеми

РІШЕННЯ ЕКСПЕРТНОЇ КОМІСІЇ КАФЕДРИ КОМП'ЮТЕРНИХ НАУК ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

Підтверджуємо ознайомлення з результатом звіту подібності щодо роботи, генерованого системою виявлення текстових збігів/ідентичності/схожості:

Назва: *Метод вибору важливих ознак з використанням мурашиного алгоритму*

Автор: *студент групи КН-20-1 Назарій РЕПІНСЬКИЙ*

Спеціальність: *122 – Комп'ютерні науки*

Освітня програма: *освітньо-професійна*

Науковий керівник: *д.т.н., професор кафедри Манзюк Е.А.*

Після аналізу звіту подібності зроблено такий висновок:

№	Висновок	Позначка про відповідність
1	Запозичення, виявлені в роботі, є законними і не є плагіатом. Робота приймається до захисту.	***відповідає***
2	Виявлені запозичення не є плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи Робота приймається до захисту, але має бути відкоригована. Відкоригований варіант має бути поданий на кафедру за 2 дні до захисту, разом із заявою щодо самостійності виконання письмової роботи та ідентичності друкованої та електронної версії роботи	
3	Виявлені запозичення не є плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. В зв'язку з цим мета роботи та поставлені завдання не були досягнені. Робота може бути допущена до захисту (наступного року) після того як буде відкоригована та допрацьована і успішно пройде повторну перевірку на академічний плагіат.	
4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	

Підтвердження:

Запозичення, виявлені в роботі Назарія Репінського, не є плагіатом, оскільки: запозичення розміщені в розділі огляду існуючих підходів, не описують безпосередньо авторську роботу і не стосуються її результатів; усі запозичення фрагментарні; до запозичень входять фрагменти програмного коду, що не мають авторства і містять поширені конструкції; серед запозичень знаходяться загальновідомі терміни, скорочення.

Обсяг запозичень, визначений системами виявлення збігів/ ідентичності/схожості, складає:

- за системою Anti-Plagiarism: 2%;

- за системою Unicheck: 9.03%.

Сумарний обсяг всіх запозичень, визначений системою виявлення збігів/ ідентичності/схожості є допустимим.

Керівник роботи _____ *Едуард МАНЗЮК*

Гарант ОП _____ *Олександр МАЗУРЕЦЬ*

Завідувач кафедри КН _____ *Олександр БАРМАК*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

МОН УКРАЇНИ

Кафедра комп'ютерних наук

ВІДГУК НАУКОВОГО КЕРІВНИКА

на кваліфікаційну роботу бакалавра

студента *ср. КН-20-1 Назарія РЕПІНСЬКОГО*

за темою *Метод вибору важливих ознак з використанням мурашиного алгоритму*

1. Актуальність теми

Актуальність методу вибору ознак на основі оптимізації мурашиної колонії полягає в ефективному обробленні великих обсягів даних сучасного Інтернету. Традиційні методи часто не справляються з цією задачею через високу обчислювальну складність та низьку точність. Запропонований підхід дозволяє зменшити розмірність даних, підвищити точність класифікації веб-сторінок та забезпечити обчислювальну ефективність, що робить його важливим для сучасних систем аналізу інформації.

2. Відповідність роботи предметній області Стандарту спеціальності 122 Комп'ютерні науки

Відповідно до прийнятих стандартів, досліджувані об'єкти та сфера діяльності включають математичні, інформаційні та симуляційні моделі реальних явищ, об'єктів, систем і процесів. Також охоплюються методи та технології для збору, зберігання, обробки, передачі та використання інформації. Головною метою роботи є розробка методу вибору важливих ознак з використанням мурашиного алгоритму. Для реалізації цієї мети застосовуються математичні моделі, методи та алгоритми, які вирішують як теоретичні, так і практичні завдання, пов'язані з розробкою методів машинного навчання. Результати цієї бакалаврської роботи відповідають стандартам спеціальності 122 – Комп'ютерні науки.

3. Професійні та особистісні якості бакалавра

Під час виконання бакалаврської роботи Назарій Репінський продемонстрував глибокі знання та високий рівень навичок, своєчасно справляючись з усіма завданнями. У процесі написання пояснювальної записки та розробки методу він проявив свої професійні компетенції та успіхи в навчанні. Він успішно освоїв професійні навички у галузі "Комп'ютерні науки".

4. Ступінь самостійності під час виконання кваліфікаційної роботи

Одержані в роботі результати є наслідком особистої діяльності студента, який самостійно виконував усі поставлені задачі.

5. Ступінь оволодіння методами дослідження

У процесі виконання кваліфікаційної роботи продемонстрував належний рівень компетентностей та володіння необхідними методами, техніками та технологіями у сфері комп'ютерних наук.

6. Повнота та якість розкриття теми роботи

Тема роботи детально обґрунтована та всебічно розкрита. Виконано аналіз відомих досліджень за обраною тематикою. Поставлені завдання успішно виконані, а також створено програмне забезпечення для реалізації запропонованого методу.

7. Логічність, послідовність, аргументованість, літературна грамотність викладення матеріалу

Структура роботи та послідовність викладення логічні та відповідають поставленій меті. Викладення матеріалу послідовне, аргументоване, літературно грамотне.

8. Можливість практичного застосування кваліфікаційної роботи бакалавра, окремих її частин

Розроблений у роботі метод може бути використаний в системах класифікації даних.

9. Висновок про можливість допуску кваліфікаційної роботи бакалавра до захисту, на яку оцінку заслуговує робота

Враховуючи належний рівень виконання та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка «відмінно».

Керівник _____ *д.т.н., професор каф. КН Едуард МАНЗЮК*

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
МОН УКРАЇНИ

Кафедра комп'ютерних наук

РЕЦЕНЗІЯ

на кваліфікаційну роботу бакалавра

студента *гр. КН-20-1 Назарія РЕПІНСЬКОГО*
за темою: *Метод вибору важливих ознак з використанням мурашиного алгоритму*

1. Актуальність обраної теми

Мурашині алгоритми, ефективні для вибору важливих ознак у задачах машинного навчання. Вони імітують природний процес пошуку оптимальних шляхів, де мурахи залишають феромонні сліди, що допомагає іншим знаходити найкоротший шлях. Це дозволяє ефективно скорочувати простір пошуку та покращувати якість моделей, відкидаючи менш значущі ознаки й зменшуючи обчислювальні витрати. Тема вибору важливих ознак з використанням мурашиного алгоритму важлива, оскільки вона сприяє підвищенню ефективності та точності моделей машинного навчання.

2. Повнота розкриття мети та завдань роботи

Мета кваліфікаційної роботи бакалавра - покращення вибору важливих ознак за допомогою мурашиного алгоритму. Для цього необхідно провести аналіз методів обробки текстів, вибрати набір ознак для аналізу, розробити метод визначення важливих ознак, що імітує поведінку мурах, та навчити модель класифікації текстів на основі цих ознак, оцінивши її ефективність. Повнота розкриття мети та завдань роботи забезпечується систематичним підходом до аналізу, розробки та оцінки методів вибору важливих ознак.

3. Зміст кожного розділу роботи

Записка кваліфікаційної роботи бакалавра містить три розділи. У першому розділі проведено аналіз предметної області, досліджено відомі роботи та визначено актуальність теми. У другому розділі представлено метод вибору важливих ознак. Третій розділ присвячено експериментальній перевірці його ефективності.

4. Оцінка розробленої інформаційної системи, її практична цінність

Розроблений вибору важливих ознак засобами штучного інтелекту дозволяє ефективно визначати важливі ознаки, що суттєво впливає на класифікацію даних в кінцевому випадку.

5. Якість оформлення кваліфікаційної роботи бакалавра

Записка відповідає всім вимогам і правилам оформлення. Викладення матеріалу є логічним і послідовним..

6. Недоліки кваліфікаційної роботи бакалавра

Рекомендовано вдосконалити систему шляхом розроблення гібридний систем класифікації.

7. Загальний висновок (допускається чи не допускається до захисту), та оцінка на яку заслуговує кваліфікаційна робота.

*Враховуючи рівень виконання та забезпечення усіх необхідних вимог, робота може бути допущена до захисту. Рекомендована оцінка **«відмінно»**.*

Рецензент доцент каф. ІПЗ Яцишин О.В.