Міністерство освіти і науки України

Харківський національний університет радіоелектроніки

Факультет комп’ютерної інженерії та управління
(повна назва)

Кафедра електронних обчислювальних машин
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА

Пояснювальна записка

Рівень вищої освіти другий (магістерський)

Методи виявлення аномалій у масивах багатовимірних даних

(тема)

Виконав:

студент II курсу, групи СПм-22-6

Усатенко М.В.
(прізвище, ініціали)

Спеціальність

123 «Комп’ютерна інженерія»
(код і повна назва спеціальності)

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

Освітня програма

Системне програмування
(повна назва освітньої програми)

Керівник: зав. каф. Коваленко А.А.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри ЕОМ _____ Коваленко А.А.
(підпис) (прізвище, ініціали)

2024 р.

Харківський національний університет радіоелектроніки

Факультет комп'ютерної інженерії та управління

Кафедра електронних обчислювальних машин

Рівень вищої освіти другий (магістерський)

Спеціальність 123 «Комп'ютерна інженерія»
(код і повна назва)

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

Освітня програма Системне програмування
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
"____" _____ 20__ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

студенту Усатенку Максиму Віталійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Методи виявлення аномалій у масивах багатовимірних даних

затверджена наказом по університету від " 01 " квітня 2024 р. № 257 Ст

2. Термін подання студентом роботи до екзаменаційної комісії 15 червня 2024 р.

3. Вхідні дані до роботи 1) масиви багатовимірних даних;
2) методи для порівнянн: PCA, випадкова проєкція, DOBIN;
3) базовий алгоритм HDoutliers.

4. Перелік питань, що потрібно опрацювати у роботі
1) аналіз проблеми виявлення аномалій у багатовимірних даних;
2) огляд існуючих технологій виявлення аномалій;
3) розробка удосконаленого алгоритму;
4) проведення експериментальних досліджень;
5) висновки.

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) Слайд-презентація – 12 слайдів

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

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Аналіз проблеми виявлення аномалій у	02.04.24-10.04.24	
	багатовимірних даних		
2	Огляд існуючих технологій виявлення	11.04.24-22.04.24	
	аномалій		
3	Розробка удосконаленого алгоритму	23.04.24-06.05.24	
4	Проведення експериментів	07.05.24-23.05.24	
5	Оформлення матеріалів кваліфікаційної роботи	24.05.24-03.06.24	
6	Подання кваліфікаційної роботи керівникові та	04.06.24-07.06.24	
	її попередній захист		
7	Подання кваліфікаційної роботи на рецензу-	08.06.24-12.06.24	
	вання		

Дата видачі завдання 01 квітня 2024 р.

Студент _____
(підпис)

Керівник роботи _____ зав. каф. Коваленко А.А.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка кваліфікаційної роботи: 72 с., 12 рис., 3 табл., 1 дод., 41 джерело.

БАГАТОВИМІРНІ ДАНІ, БАГАТОФАКТОРНІ ДАНІ, ВИЯВЛЕННЯ АНОМАЛІЙ, ВИЯВЛЕННЯ ВИКИДІВ, КЛАСТЕР, ПОТОКОВІ ДАНІ, ПОШУК НАЙБЛИЖЧОГО СУСІДА.

Метою кваліфікаційної роботи є розробка методу виявлення аномалій у масивах багатовимірних даних.

В роботі розглянуті методи виявлення аномалій, зокрема в багатовимірних і багатофакторних даних. Обговорено останні дослідження у відповідній галузі.

Алгоритм HDoutliers – це потужний алгоритм для виявлення аномалій у даних великої розмірності. Однак він страждає від кількох обмежень, які значно перешкоджають його здатності виявляти аномалії в певних ситуаціях. У цьому дослідженні запропоновано вдосконалений алгоритм, який усуває ці обмеження.

ABSTRACT

Master’s thesis: 72 pages, 12 figures, 3 tables, 1 appendix, 41 sources.

ANOMALY DETECTION, CLUSTER, MULTIDIMENSIONAL DATA, MULTIVARIATE DATA, NEAREST NEIGHBOR SEARCH, OUTLIER DETECTION, STREAMING DATA.

The purpose of the qualification work is to develop a method for detecting anomalies in arrays of multidimensional data.

Methods of detecting anomalies, in particular in multidimensional and multivariate data, are considered in the work. The latest research in the relevant field is discussed.

The HDoutliers algorithm is a powerful algorithm for detecting anomalies in high-dimensional data. However, it suffers from several limitations that greatly hinder its ability to detect anomalies in certain situations. This study proposes an improved algorithm that overcomes these limitations.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

DAE – глибокий автокодувальник (англ., Deep Autoencoder)

DBN – глибока довірча мережа (англ., Deep Belief Network)

DOBIN – базис викидів на основі відстані з використанням сусідів (англ., Distance-Based Outlier Basis using Neighbours)

FCR – коефіцієнт помилкової класифікації (англ., Faulty Classification Rates)

KNN – K найближчих сусідів (англ., K nearest neighbor)

ODR – коефіцієнт виявлення викидів (англ., Outlier Detection Rates)

PCA – аналіз головних компонент (англ., principal component analysis)

ROBEM – надійна максимізація очікувань (англ., Robust Expectation Maximization)

ROC – характеристика оператора приймача (англ., Receiver Operator Characteristics)

ВСТУП

У сучасному світі щохвилини, щосекунди генерується велика кількість даних завдяки розвитку технологій. Усе це призводить до ери великих даних, коли дані швидко зростають, а останні події також сприяють величезному об’єму даних.

Виявлення аномалій привернуло велику увагу через його важливість у багатьох сферах, включаючи вторгнення в мережу, шахрайство з кредитними картками, управління енергією, фінанси, статистику, контроль процесів, обробку сигналів, а також машинне навчання. Виявлення аномалій сприяє ранньому виявленню нерелевантних моделей або незвичайних подій. Таке є корисним для попередньої обробки даних і очищення для пошуку підозрілих даних. Виявлення аномалій залишається предметом інтенсивних досліджень, і, визначаючи аномалії, дослідники можуть отримати життєво важливі знання, які допоможуть краще зрозуміти дані. Крім того, добре мати фундаментальне розуміння аномалій, які можуть призвести до кращого аналізу та, водночас, уникнути будь-якого незначного впливу на якість даних. Прикладом найпоширеніших програм для виявлення аномалій є виявлення шахрайства з кредитними картками та обробка заявок на кредит.

Проблему виявлення аномалії, як правило, непросто вирішити, і це важко, особливо в багатовимірних даних великої розмірності. Було виявлено, що різні сфери, такі як біомедицина, Інтернет, освіта, медицина, бізнес і соціальні медіа, застосовують багатовимірні дані. Раніше виявлення аномалій проводилося за допомогою статистичних методів. Передові технології, величезний попит на використання машинного навчання через умови великого обсягу даних, з якими традиційні методи не можуть впоратися.

Однак, якщо статистичні методи та методи машинного навчання сліпо використовуються для даних, що містять аномалії, ці методи можуть негативно вплинути на отримані результати, наприклад, неправильна

специфікація моделі, упереджена оцінка параметрів і, зрештою, оманливі результати. Протягом багатьох років було визнано, що існує проблема, пов’язана з пошуком аномалій, зокрема з їх усуненням. Дуже важливо бути спостережливим, щоб зрозуміти, чому аномалії потрібно виявити та що вони представляють. Більше того, сучасні дані часто мають велику розмірність, і традиційне виявлення аномалій може зіткнутися з труднощами при обробці даних великої розмірності. Таким чином, для виявлення аномалій було розроблено численні методи машинного навчання, такі як методи на основі відстані, кластеризації, щільності та класифікації.

Крім того, більшість із цих методів також вирішують багатовимірну проблему високої розмірності при виявленні аномалії, яка створює серйозну проблему, що призводить до величезної обчислювальної складності, отримання недійсних результатів і, одночасно, ускладнює завдання.

Алгоритм HDoutliers є потужним неконтрольованим алгоритмом для виявлення аномалій у даних великої розмірності з міцною теоретичною основою. Однак він страждає від деяких обмежень, які значно перешкоджають його продуктивності за певних обставин. В роботі пропонується модифікація алгоритму, яка усуває ці обмеження. Аномалія визначається як спостереження, яке помітно відрізняється від більшості з великим розривом у відстані. Для розрахунку аномального порогу використовується підхід, заснований на теорії екстремальних значень. Використовуючи різноманітні синтетичні та реальні набори даних, демонструється широка застосовність і корисність алгоритму. Також демонструється, як цей алгоритм може допомогти у виявленні аномалій, присутніх в інших структурах даних за допомогою розробки функцій. Аналізуються ситуації, коли алгоритм HDoutliers є ненайкращим як за точністю, так і за часом обчислення. Ця структура реалізована в пакеті R з відкритим кодом.

1 ПРОБЛЕМА ВИЯВЛЕННЯ АНОМАЛІЙ У БАГАТОВИМІРНИХ ДАНИХ

1.1 Загальні відомості

Багатофакторні дані включають дві або більше змінних чи ознак [1, 2]. Це досить складно, тому що багатофакторні дані вимагають розуміння зв’язків між багатьма змінними, і зазвичай людський мозок перевантажений простою масою даних. Крім того, для обробки багатофакторних даних потрібно більше математики, ніж однофакторних, для отримання висновку.

Велика розмірність означає наявність кількох змінних, ознак, атрибутів у наборі даних, що перевищує кількість спостережень [3]. Велика розмірність або збільшення розмірності може призвести до розрідженості даних, у результаті чого дані мають багато суперечливих властивостей, які є більш розпорошеними та більш ізольованими, що створює серйозну проблему для аналізу даних. Ця проблема широко відома як «прокляття розмірності» [4].

Оскільки аномалії є незвичайними за визначенням і можуть суттєво відрізнятися одна від одної, вони викликають інші проблеми та виклики, ніж класифікація, яка регулярно контролюється [5, 6]. Незважаючи на це, в різних областях були ефективно реалізовані алгоритми виявлення аномалій. З іншого боку, вже розроблено багато різних методів, оскільки кожна область застосування має своє визначення ненормальності та обмежень застосування [7, 8]. Перша половина завдання полягає у виявленні аномалій, друга – в інтерпретації аномалій, які були виявлені. Часто набори даних реального світу матимуть умови, коли деякі точки поводяться інакше, ніж решта наборів даних. Дуже важливо вміти виявляти аномалії, які можуть зіпсувати результат аналізу або також можуть містити цінну інформацію. Дослідження [9] підкреслюють, що помилки можуть спричиняти аномалії, але також можуть виникати в незвичних обставинах. Також мається на увазі, що потрібно якось дослідити та зрозуміти їх з різних точок зору, а не просто

видаляти.

Оскільки світ дедалі більше керується даними, а в той же час розповсюджуються нові технології, зібрані дані поступово стають величезними за розміром і вимірністю. Більшість традиційних методів виявлення аномалій не в змозі впоратися з великовимірними даними. Проведені дослідження [10] також підтвердили твердження, що аналіз даних великої розмірності став складним процесом. Зі збільшенням розмірності дані стають більш розрідженими, що спричиняє труднощі у виявленні та аналізі аномалій. Залучаючи багатовимірні дані, важливо зробити правильну інтерпретацію. Причина полягає в тому, що це допомагає користувачам оцінити аномальну вибірку для кожної додаткової інформації з метою повного розуміння даних.

Останні методи виявлення аномалій були розроблені для наборів даних низької розмірності і стикаються з труднощами, коли розміри збільшуються. Водночас із збільшенням розмірності даних існуючі методи потребують великих обчислювальних витрат [11]. Крім того, пряме застосування може дати недійсні результати. Протягом багатьох років було розроблено численні алгоритми для числових даних великої розмірності. Незважаючи на те, що були розроблені різні методики, важливо знати, що традиційне виявлення аномалії є менш значущим зі зростанням розмірності [12]. Крім того, деякі дослідники [13] стверджують, що можлива стратегія обробки даних великої розмірності полягає в застосуванні методів зменшення розмірності для покращення виявлення аномалій. Одним із сучасних рішень у виявленні аномалій є використання алгоритмів машинного навчання.

Також слід зазначити, що дані великої розмірності негативно впливають на продуктивність алгоритмів машинного навчання. Незважаючи на те, що алгоритми машинного навчання здатні передбачати завдання, їхня продуктивність часто обмежена, а іноді дає погані результати щодо якості представлення даних, особливо з даними великої розмірності та умовою з великою кількістю функцій. Крім того, більшість даних реального світу

мають більше однієї функції, змінних і ознак, широко визнаних як багатофакторні дані. Виявлення аномалій у таких даних стає все більш важливим, особливо в дослідженнях. Наприклад, це стосується деяких галузей, таких як планування охорони здоров'я, фабричні системи та транспортні системи. При роботі з багатофакторними даними точка даних може не узгоджуватися з шаблоном основних даних. Таким чином, аномалію можна не помітити під час перевірки. Такий вид аномалії можна визначити за допомогою статистичного інструменту.

Навпаки, в статистиці розроблені різні алгоритми для виявлення аномалій, але більшість методів застосовуються лише до однофакторних випадків [14]. Процес визначення аномалії є більш складним у багатофакторних наборах даних порівняно з однофакторними наборами даних. Незважаючи на те, що багато досліджень сфокусовані на методах виявлення аномалій, вони зосереджені лише на однофакторних наборах даних, і лише деякі розглядали багатофакторні набори даних. Це призводить до збільшення труднощів у виявленні аномалій.

Було зроблено багато спроб осмислити питання виявлення аномалій. Однак це нелегко, коли аномалії є в межах багатовимірних та багатофакторних даних. Дослідження [15] свідчать, що виявлення аномалій у багатовимірних і багатофакторних даних стає надзвичайно складним. Дослідницька робота [11] припускає, що в багатофакторних даних великої розмірності необхідно враховувати відстань спостереження від центроїда, а також форму даних. Крім того, необхідно знати про кількість ознак, які необхідно враховувати (однофакторні чи багатофакторні); інакше усунення аномалій правильних даних може призвести до значної втрати інформації. У ще одні дослідницькій роботі [16] зазначено, що кожну змінну в багатофакторному наборі даних слід аналізувати разом, щоб врахувати їх зв'язок. Це інша природа аномалій у багатофакторних умовах, яка залежить від співвідношення між змінними. Крім того, його неможливо легко виявити за допомогою таких методів візуалізації, як гістограми, прямокутні діаграми

або діаграми розсіювання. Обмеження методів візуалізації корисно лише для просторів до 3D і не є корисним для просторів, розмірність яких перевищує 3D [11].

1.2 Аспекти проблеми виявлення аномалій

Проблема виявлення аномалій має багато різних аспектів, і на методи виявлення може сильно впливати те, як визначати аномалії, тип вхідних даних і очікуваний результат. Ці відмінності призводять до різноманітних формулювань проблем, які необхідно вирішувати за допомогою різних аналітичних методів. Хоча зараз існує кілька корисних обчислювальних методів, розробка нових методів виявлення аномалій продовжує залишатися активною, привабливою міждисциплінарною областю досліджень через різні аналітичні проблеми в різних областях застосування, таких як моніторинг навколишнього середовища, відстеження об’єктів, епідеміологічні спалахи, мережева безпека, виявлення шахрайств. Постійно зростаючі обчислювальні ресурси та передові технології збору даних, які наголошують на великомасштабних даних у реальному часі, є ще однією причиною такого зростання, оскільки вони створюють нові аналітичні проблеми з їх зростаючим розміром, швидкістю та складністю, які вимагають ефективних аналітичних та обчислювальних методів.

Виявлення аномалій має дві основні цілі, які є суперечливими: одна знижує цінність аномалій і намагається їх усунути, тоді як інша вимагає приділяти особливу увагу аномаліям і проводити аналіз першопричин. Наявність аномалій у даних можна вважати недоліками даних або помилками вимірювання, які можуть призвести до упередженого оцінювання параметрів, неправильної специфікації моделі та оманливих результатів, якщо сліпо застосовувати класичні методи аналізу [17]. У таких ситуаціях основна увага полягає в тому, щоб знайти можливості для видалення аномальних точок і, таким чином, покращити як якість даних, так і результати подальшого

аналізу даних. Навпаки, у багатьох інших застосуваннях аномалії самі по собі є основними носіями важливої та часто критичної інформації, такої як екстремальні погодні умови (наприклад, лісова пожежа, цунамі, повінь, землетрус, виверження вулкана та забруднення води), несправності (наприклад, відстеження польотів і відстеження кабелю живлення), шахрайство, які можуть завдати значної шкоди життю або цінним активам, якщо їх швидко не виявити та не виправити.

Багатовимірні набори даних існують у багатьох областях досліджень. Деякі алгоритми виявлення аномалій також використовують розробку ознак як техніку зменшення розмірності і таким чином перетворюють інші структури даних, такі як колекція часових рядів за допомогою функцій часових рядів [18], колекція діаграм розсіювання з використанням скагностики [19] і геномні мікроматриці та хімічні композиції в біології [20] у багатовимірні дані до процесу виявлення з метою легкого контролю. За сценарієм даних великої розмірності всі атрибути можуть бути одного типу даних або суміші різних типів даних, наприклад категоріальних або числових, що безпосередньо впливає на реалізацію та область застосування алгоритму. Багато дослідницької уваги було приділено виявленню аномалій для числових даних. Доступні обмежені методи, які обробляють як числові, так і категоріальні дані за допомогою аналізу відповідності [21].

Багатовимірні аномалії можуть виникати у всіх атрибутах або підмножині атрибутів. Якщо всі аномалії у високовимірному просторі даних були аномаліями в нижчому вимірі, тоді виявлення аномалії можна виконати за допомогою паралельних зображень по осі або шляхом включення додаткового кроку вибору змінної для процесу виявлення. Однак на практиці певні багатовимірні екземпляри сприймаються лише як аномалії, якщо розглядати їх як багатовимірні проблеми та кореляційну структуру всіх розглянутих атрибутів. В іншому випадку вони, як правило, не помічаються, якщо атрибути розглядаються окремо [21].

Проблема виявлення аномалій широко вивчалася протягом останніх

десятиліть у багатьох областях застосування. Загалом було проведено кілька досліджень методів виявлення аномалій для певних доменів даних, таких як високовимірні дані, мережеві дані, часові дані, машинне навчання та статистичні домени, виявлення новизни, виявлення вторгнень та невизначені дані. Деякі алгоритми є специфічними для програми та використовують переваги основної структури даних або інших предметно-специфічних знань [18]. Також доступні більш загальні алгоритми без предметно-специфічних знань з власними перевагами та обмеженнями.

Серед багатьох можливостей алгоритм HDoutliers [21] – це потужний неконтрольований алгоритм із міцною теоретичною основою для виявлення аномалій у даних великої розмірності. Хоча цей алгоритм має багато переваг, кілька характеристик перешкоджають його продуктивності. Зокрема, за певних обставин він має тенденцію до збільшення частоти хибних негативних результатів (тобто детектор ігнорує точки, які здаються справжніми аномаліями), оскільки він використовує лише відстані до найближчих сусідів для розпізнавання аномалії. Крім того, для роботи з великими наборами даних із численними спостереженнями він використовує алгоритм Leader, який утворює кілька кластерів точок за один прохід через набір даних за допомогою кулі фіксованого радіуса. Використовуючи цей метод кластеризації, він намагається отримати можливість ідентифікувати аномальні кластери точок. Однак за наявності дуже близьких сусідніх аномальних кластерів це має тенденцію до збільшення частоти помилкових негативів. Крім того, цей додатковий етап кластеризації має серйозний негативний вплив на обчислювальну ефективність алгоритму при роботі з великими наборами даних.

Це дослідження має три фундаментальні внески. По-перше, пропонується алгоритм, що представляє «аномальний пошук і трасування», який усуває обмеження алгоритму HDoutliers. Алгоритм зосереджений саме на швидкому, точному обчисленні аномального результату за допомогою простих, але ефективних методів для покращення продуктивності. По-друге,

представляється пакет R для реалізації пропонованого алгоритму і пов’язаних з ним функцій. По-третє, демонструється широка застосовність і корисність пропонованого алгоритму з використанням різних наборів даних.

Отже, покращений алгоритм повинен надати такі переваги:

- можливість застосовувати як до одновимірних, так і до багатовимірних даних;
- неконтрольований за своєю природою і не вимагає навчальних наборів даних для процесу побудови моделі;
- аномальний поріг є пороговим значенням, керованим даними, і має дійсну імовірнісну інтерпретацію, оскільки базується на теорії екстремальних значень;
- використовуючи k-відстань до найближчих сусідів для обчислення аномального показника, він отримує можливість мати справу з проблемою маскування;
- завдяки використанню швидких механізмів пошуку найближчих сусідів він може забезпечити підтримку наборів даних, які передаються у великих кількостях майже в реальному часі;
- може працювати з даними, які можуть мати мультимодальний розподіл для типових екземплярів даних;
- виробляє оцінку (щоб вказати, наскільки аномальні екземпляри) і двійкову класифікацію (щоб зменшити простір пошуку під час візуального аналізу та аналізу першопричини) для кожного екземпляра даних як результат;
- може виявляти як викиди, так і інлієри.

1.3 Типи аномалій у даних великої розмірності

Проблеми виявлення аномалій у високовимірних даних полягають у трьох аспектах (рисунок 1.1), що передбачає виявлення:

- глобальних аномалій;

- локальних аномалій;
- мікрокластерів або кластерів аномалій.

На рисунку 1.1. аномалії представлені червоними трикутниками, а чорні точки відповідають типовій поведінці.

Більшість існуючих методів виявлення аномалій для даних великої розмірності можуть легко розпізнати глобальні аномалії, оскільки вони дуже відрізняються від щільної області щодо своїх атрибутів. Навпаки, локальна аномалія є аномалією лише тоді, коли вона відрізняється від свого місцевого оточення та порівнюється з ним. В одному з досліджень [22] вводиться набір алгоритмів, заснованих на визначенні щільності або відстані аномалії, який головним чином зосереджується на локальних аномаліях у високовимірних даних. Мікрокластери або кластери аномалій можуть спричинити проблеми з маскуванням. Цій проблемі приділено дуже мало уваги порівняно з двома іншими категоріями. Алгоритм HDoutliers певною мірою вирішує цю проблему шляхом групування екземплярів, які дуже близькі за високовимірного простору, а потім вибір репрезентативного члена з кожного кластера перед обчисленням відстаней до найближчих сусідів для вибраних екземплярів. У цій роботі розглядаються всі три типи аномалій.

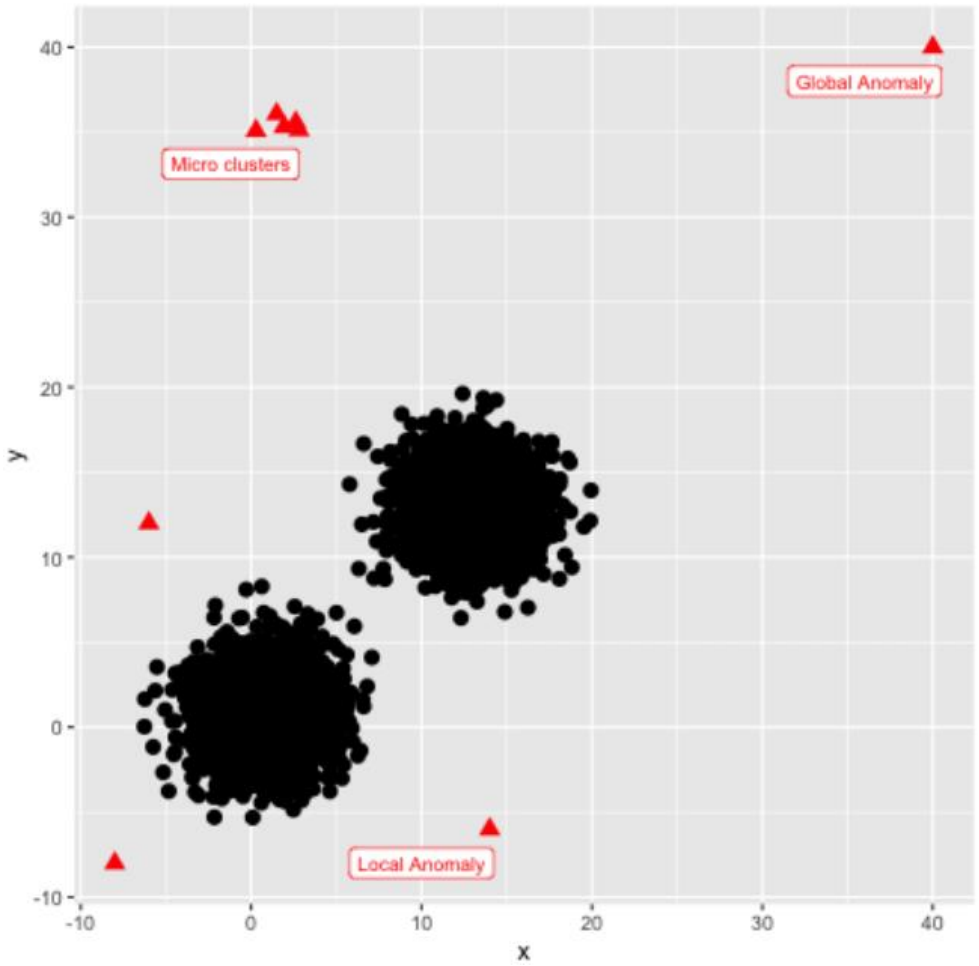


Рисунок 1.1 – Різні типи аномалій у багатовимірних даних

1.4 Визначення аномалій у багатовимірних даних

Аномалії часто згадуються в літературі під кількома альтернативними термінами, такими як викиди, новизна, недоліки, відхилення, неузгоджені спостереження, екстремальні значення/випадки, точки змін, рідкісні події, вторгнення, неправильне використання, винятки, аберації, несподіванки, особливості, дивні значення і забруднюючі речовини в різних сферах застосування. З них два терміни аномалії та викиди використовуються загально та взаємозамінно в літературі, що описує дослідження, пов’язані з темою.

Термін inlier також стосується теми, але рідко зустрічається в літературі з виявлення аномалій. Внутрішні точки – це ті точки, які

з’являються між типовими кластерами, не приєднуючись до жодного з них, але все ще знаходяться в діапазоні, визначеному типовими кластерами. Навпаки, відповідне поняття «викид» зазвичай використовується для позначення екземпляра даних, який з’являється поза межами простору ближче до хвоста розподілу, визначеного типовими екземплярами даних. Деякі класичні методи, пов’язані з цією темою, не в змозі виявити інлієри і зосереджуються лише на викидах [23]. Однак виявлення інлієрів є не менш важливим, оскільки вони можуть призвести до помилок інтерполяції. В цій роботі розглядаються як внутрішні, так і викидні показники. Щоб уникнути будь-якої плутанини, буде використано термін «аномалія».

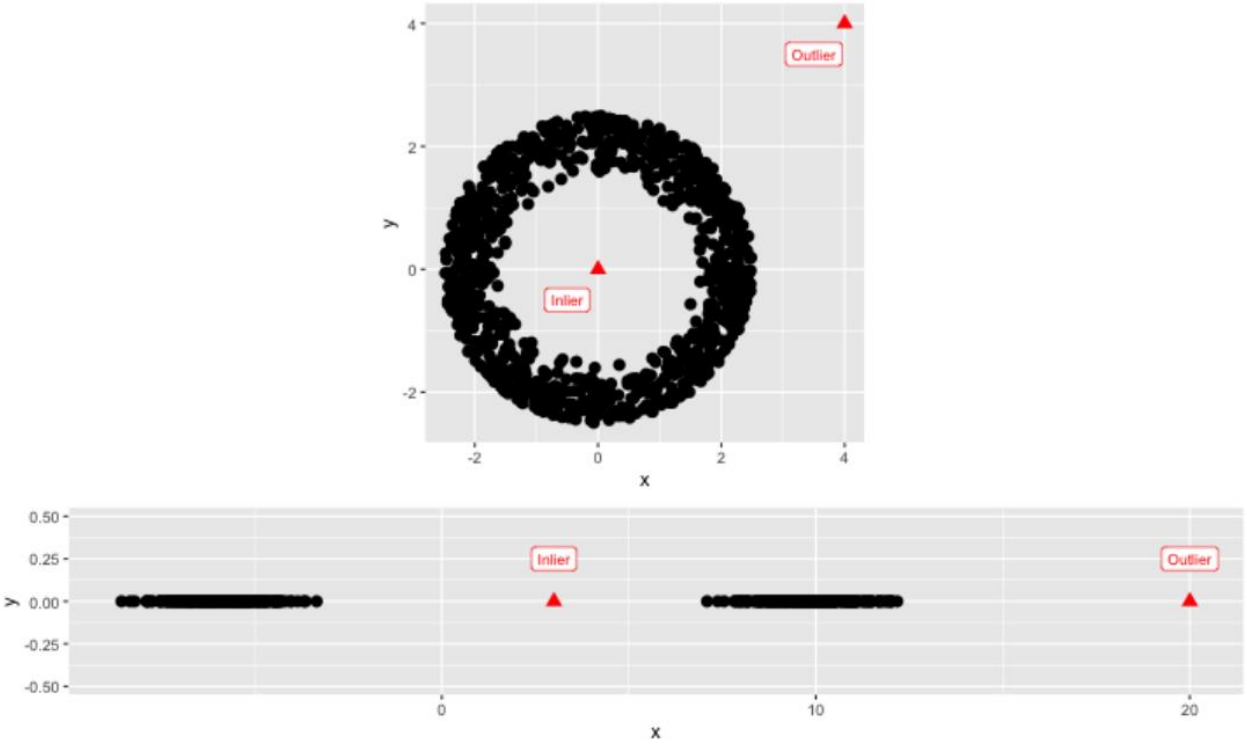


Рисунок 1.2 – Внутрішні та зовнішні точки

Через складний характер проблеми важко знайти уніфіковане визначення аномалії; визначення часто залежить від фокусу дослідження та структури вхідних даних, доступних системі [24]. Однак є деякі визначення, які є достатньо загальними, щоб впоратися з наборами даних у різних

областях застосування. Аномалію визначають як спостереження, яке помітно відрізняється від інших членів набору даних. Однак це відхилення можна визначити в термінах відстані або щільності. В ряді запропонованих методів виявлення аномалій [21, 25] аномалії визначаються в термінах відстані. В інших методах [26, 27] аномалії визначають відносно щільності або ймовірності появи спостережень. Також можна знайти низку алгоритмів виявлення аномалій на основі відстані та щільності.

У даній роботі аномалію визначено як спостереження, яке помітно відрізняється від більшості з великим розривом у відстані за припущенням, що існує велика відстань між типовими даними та аномаліями порівняно з відстанню між типовими даними.

2 ОГЛЯД ТЕХНОЛОГІЙ ВИЯВЛЕННЯ АНОМАЛІЙ У БАГАТОВИМІРНИХ ДАНИХ

Як було зазначено в розділі 1, дві функції, які найбільше впливають на проблеми виявлення аномалій, – це «багатовимірність» і «багатофакторність». Проблема багатовимірних і багатофакторних даних не тільки ускладнює розпізнавання аномалій, але й створює нові перешкоди, такі як обчислювальна вартість, нерелевантні результати, якщо виконуються прямі програми, несумісні точки з первинними даними, а також розрідженість даних [3]. Методи, які вирішують проблему виявлення аномалій у багатовимірних і багатовимірних даних, підсумовані в таблиці 2.1. Кожен метод має переваги та недоліки, коли потрібно виправляти різні проблеми залежно від характеру даних.

Таблиця 2.1 – Порівняння продуктивності алгоритмів виявлення аномалій

Алгоритм	Переваги	Недоліки
PCA	широко використовується завдяки простоті та ефективності	з великою розмірністю оцінка зазвичай складна; наявність аномалії може вплинути на продуктивність PCA
Випадкова проєкція	може бути використана будь-яка комбінація розмірів і вимірів вибірки	немає чітких рекомендацій щодо кількості бажаних проекцій
DOBIN	дозволяє виявити аномалію за допомогою меншої кількості компонентів	Чутливий
STRAY	застосовується як для одновимірних, так і для високовимірних даних, а процес	необхідно провести оптимізацію за найкращим значенням K

	побудови моделі не потребує використання навчальних наборів даних	
ROBEM	для виявлення аномалії використовується критичне значення; таким чином, це призводить до успішної продуктивності щодо виявлення аномалій	найповільніший алгоритм
DAE-KNN	знижує обчислювальні витрати та покращує ефективність виявлення порівняно з одним детектором аномалій	побудова DAE займає багато часу, якщо набір даних великий
OCP	немає необхідності оцінювати коваріацію, ідеально підходить для даних великої розмірності	час обчислення вище

Загалом, стратегії вирішення проблем виявлення аномалій можна розділити на кілька категорій: зменшення розмірів, машинне навчання та гібридні.

2.1 Підхід зменшення розмірності

Процес пошуку маловимірних функцій у високовимірних даних для усунення бар’єрів високовимірних даних. Такий процес допомагає зменшити кількість вхідних змінних у наборі даних. Простими словами, перетворення багатовимірного подання даних у низьковимірне, зберігаючи якомога більше корисних вихідних значень даних. Для зменшення розмірності можна використовувати кілька застосовних методів, наприклад аналіз головних компонент (PCA), вибір ознак, генетичний алгоритм, лінійний

дискримінантний аналіз і машинне навчання. Після цього [28] дані перетворюються за допомогою розмірного зменшення для того, щоб зробити сприятливе подання даних для точної генерації продуктивності алгоритму машинного навчання в інших областях дослідження. Здатність підходів зменшення розмірності трансформувати її є дуже вагомою через наявність таких складних даних, що робить методи широко використовуваними для аналізу та візуалізації даних великої розмірності [10].

Аналіз головних компонент. Це найстаріший і найпопулярніший підхід. Він також відомий як один із підходів, здатних вирішити проблему великої розмірності [28]. Це означає, що методи PCA часто використовуються для подолання «прокляття розмірності». PCA має на меті витягти всі відповідні значення з набору даних і об’єднати їх у нові ортогональні змінні, відомі як головні компоненти [29]. Це будуть лінійні комбінації корельованих змінних з меншою кількістю компонентів, ніж вихідні.

Перші головні компоненти представляють велику кількість початкових відхилень даних після других головних компонент, які містять другу велику дисперсію. Метод передбачає, що перші головні компоненти зберігають велику варіацію на початку та зменшують розмірність від p до k. Крім того, вони можуть бути обчислені як лінійна зважена комбінація ознак.

Випадкова проєкція. Більшість методів виявлення аномалій у багатофакторних і багатовимірних даних вимагають інформації з коваріаційної матриці. Однак із збільшенням розмірності даних складнішою стає оцінка матриці. Дослідження [30] запропонувало виявлення аномалій за допомогою випадкової проєкції, щоб уникнути необхідності оцінювати матрицю. Запропонований метод використовує проєкції як техніку зменшення розмірності. Певним чином, йому не потрібно оцінювати середнє значення та стандартне відхилення.

Базис викидів на основі відстані з використанням сусідів (англ., Distance-Based Outlier Basis using Neighbours, DOBIN). Метод DOBIN [13] діє

як стратегія попередньої обробки, яку можна застосувати будь-яким методом виявлення аномалії. Звичайним є використання PCA для виявлення аномалій, коли дані мають велику розмірність. Однак, за результатами аналізу, DOBIN є кращим за PCA. Більше того, DOBIN має два способи використання: по-перше, це спрощення шляхом розгляду лише меншої кількості компонентів для виявлення аномалій. По-друге, ще одне використання DOBIN – це допомогти виявити аномалії у формі візуалізації. Основна конструкція для DOBIN полягає в максимізації K відстаней до найближчих сусідів (KNN).

Основні кроки DOBIN:

- визначення простору Y для певного набору даних;
- формування основи;
- перетворення вхідного простору Z.

2.2 Підхід машинного навчання

Раніше виявлення аномалій проводилося за допомогою статистичних методів. Після появлення великих даних машинне навчання є широко використовуваною технікою через величезну кількість даних, з якою традиційні методи не можуть працювати. Надвисока розмірність даних може спричинити проблеми з моделями машинного навчання, такими як точна категоризація, ідентифікація шаблонів і подання. Приклади підходів машинного навчання включають лінійну регресію, автокодер-декодер і підходи на основі кластеризації.

Алгоритм «заблудження» (англ., Stray, узятий зі слів Search і Trace AnomalY) запропонований для подолання обмежень та розширення можливостей іншого методу виявлення аномалій HDoutliers. Блукаючий алгоритм – це підхід, заснований на відстані, який використовує евклідові відстані для пошуку k-найближчого сусіда. Для кожного окремого спостереження потрібно обчислити k-найближчі сусідні відстані KNN, де i=1, 2...,k. Після цього обчислюються послідовні різниці відстаней. Потім

береться k-відстань найближчого сусіда з найбільшим розривом.

2.3 Гібридний підхід

Поєднання підходу машинного навчання з іншими методами, такими як статистичні чи інші застосовні методи, називається гібридним підходом. На ранніх стадіях виявлення аномалії може бути виконаний простий аналіз даних, наприклад описова статистика. Це допомагає ідентифікувати аномальні спостереження, отримати розуміння даних, що зрештою може призвести до модифікацій, включаючи комбінацію інших методів.

Метод DAE (англ., Deep Autoencoder) з ансамблем KNN. DAE створюється за допомогою Deep Belief Network (DBN), похідної від RBM. З іншого боку, RBM – це неорієнтована графова модель, що складається з видимих одиниць v і прихованих одиниць h, які представляють спостереження та особливості. DAE намагається зіставити високовимірні дані в низьковимірний простір ознак. Остаточне рішення буде щодо ненормального зразка, якщо він вказує 1, і нормального зразка, якщо він вказує -1.

Метод OCP (англ., One Class Peeling). Підхід OCP – це гнучка структура для виявлення аномалій у багатовимірних даних, яка об’єднує статистичні методи та методи машинного навчання. У стратегію включено методи щільності ядра та статистичної відстані. Крім того, це не передбачає обчислення коваріаційної матриці. Потім метод OCP включає вимірювання відстані між кожним спостереженням і центром і надійно прогнозує центр. Формулювання дається шляхом визначення центру багатофакторних даних за допомогою ітераційного методу очищення на основі меж, отриманих із SVDD. Для оцінки стійкості часто використовується кінцева точка руйнування заміни вибірки (англ., Finite Sample Replacement Breakdown Point, FSRBP).

Ключові етапи методу OCP:

- визначення граничного значення, h;
- обчислення стійкої оцінки за допомогою SVDD із ядерною функцією Гауса;
- обчислення відстані ядра між кожним вектором спостереження та оцінка надійності у центрі даних;
- масштабування відстані;
- позначення спостережень більшими, ніж з потенційною аномалією.

Надійна максимізація очікувань (англ., Robust Expectation Maximization, ROBEM). Для пошуку аномалій було розроблено багато методів машинного навчання та статистичних методів. Одним із способів ідентифікації аномалій є кластеризація. Метод кластеризації є переконливим у сфері машинного навчання. Було запропоновано [29] новий алгоритм кластеризації шляхом поєднання алгоритму кластеризації EM, а також надійного аналізу головних компонент (ROBPCA). Крім того, запропонований метод складається з двох етапів: 1) аномалії виявляються за допомогою алгоритму ROBPCA та 2) доступний набір даних кластеризується за допомогою алгоритму EM-кластеризації. Після етапу 1 буде використано алгоритм ROBPCA для обчислення балів головних компонентів і ортогональних відстаней.

Основні етапи реалізації методу ROBEM:

- відбувається виявлення аномалії за допомогою алгоритму ROBPCA; аномалії визначаються як спостереження, які перевищують критичні значення як для оцінки, так і для ортогональних відстаней (розрахованих за допомогою ROBPCA), і надсилаються до кластера аномалій. Для порівняння, очищені дані містять усі спостереження, що залишилися;
- кластеризація на етапі, коли спостереження в очищених даних кластеризувались за допомогою алгоритму EM.

2.4 Огляд виявлення аномалій

Слід детально розглянути структуру потоку виявлення аномалій у багатовимірних і багатофакторних даних. Структура включає наступні етапи, як показано на рисунку 2.1.

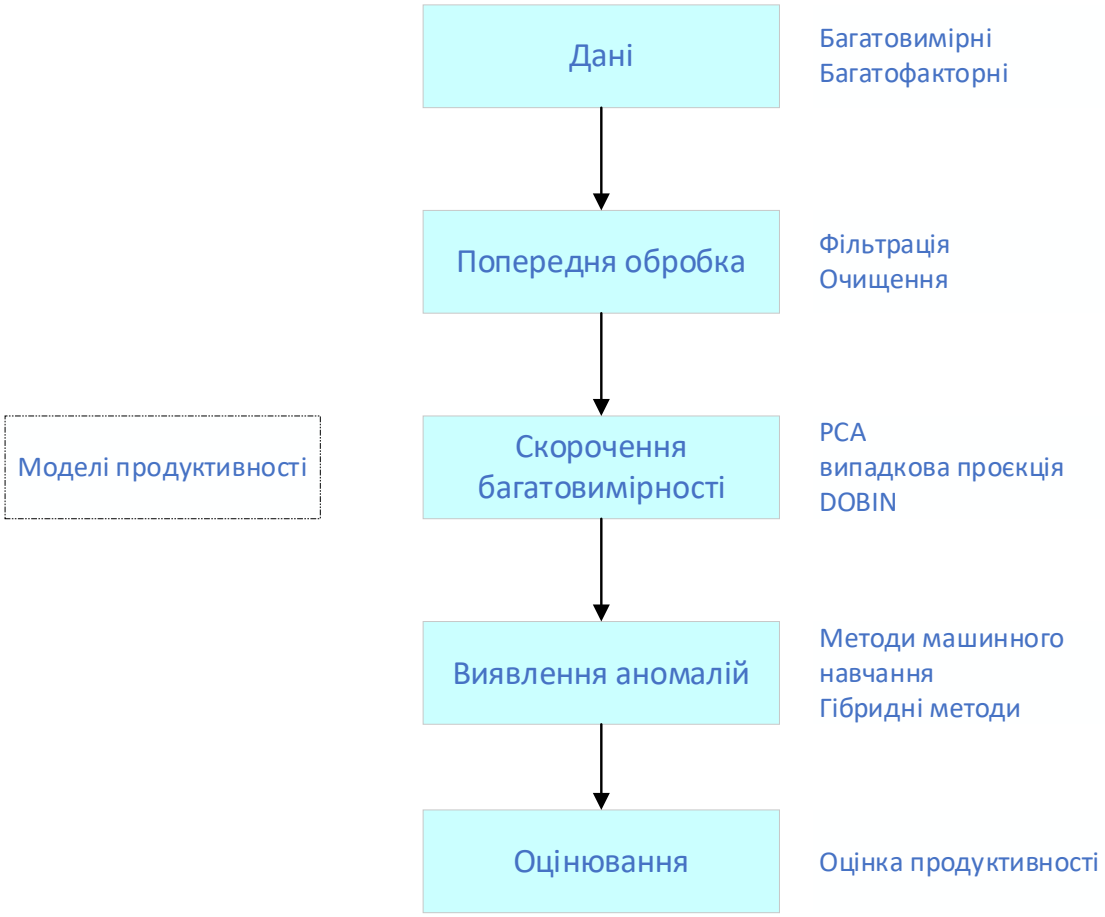


Рисунок 2.1 – Фреймворк виявлення аномалій [31]

Дані. Етап підготовки даних, на якому вибираються відповідні набори даних для виявлення аномалій. У цьому випадку розглядаються як багатофакторні, так і багатовимірні.

Попередня обробка даних. На цьому етапі багатовимірні та багатофакторні набори даних були очищені та відфільтровані, щоб переконатися, що не було невизначеностей, і далі розділені на навчальні та тестові набори даних.

Зменшення розмірів. Процес пошуку низьковимірних характеристик

високовимірних даних. Допомога в усуненні перешкод високовимірних даних, оскільки більшість існуючих методів не можуть добре працювати в умовах великої розмірності.

Виявлення аномалії. Мета виявлення аномалій – дослідити наявність аномалій у даних. Форми виведення будуть у формі балів і міток. Технічно бали сортуються, і вибирається поріг для позначення аномалій. Між тим, мітки приймаються за допомогою двійкового рішення про те, чи є алгоритм аномалією чи ні.

Оцінювання. Модель інтегрується на завершальному етапі. Цей етап є критичним кроком, оскільки перевіряє надійність і можливість узагальнення моделі. Здебільшого продуктивність вимірюватиметься площею під характеристиками оператора приймача (англ., Receiver Operator Characteristics, AUC), коефіцієнтами виявлення викидів (англ., Outlier Detection Rates, ODR), коефіцієнтами помилкової класифікації (англ., Faulty Classification Rates, FCR), а також кривою ROC, особливо для завдань класифікації.

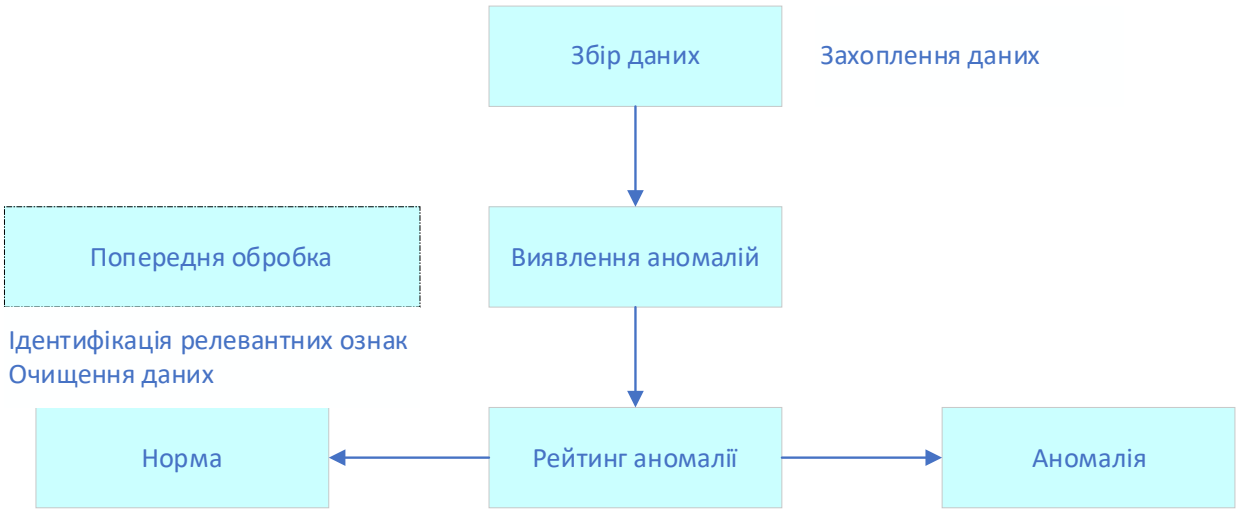


Рисунок 2.2 – Фаза виявлення загальної аномалії

Фаза виявлення загальної аномалії припиняється до моменту визначення аномальних і нормальних даних [32]. Немає додаткових пояснень

щодо того, чи класифікуються точки як аномалії, які мають бути видалені, і великі нормальні спостереження (екстремальні), як показано на рисунку 2.2. Існує кілька попередніх підходів до виявлення аномалій, які були розглянуті вище. Однак суттєва відмінність між деякими з цих підходів і випадком, який аналізується, полягає в тому, що немає подальшого пояснення різниці між аномальними та екстремальними спостереженнями.

Різні дослідники провели багато експериментальних тестів для вимірювання ефективності виявлення аномалій у багатовимірних і багатофакторних даних [33]. Для відповідного порівняння продуктивності було вибрано різні показники ефективності. Комплексна порівняльна оцінка різних методів, заснованих на виявленні аномалій, представлена в таблиці 2.2. Для детального огляду слід аналізувати чотири характеристики: швидкість навчання, корисність, ефективність і вимоги до ресурсів. Швидкість навчання вказує на ступінь ефективності запропонованого методу в навчанні. Тим часом, корисність описує сферу застосування техніки, незалежно від того, чи вона застосовна лише до багатофакторної, багатовимірної чи обох. Далі ефективність відноситься до продуктивності запропонованого методу на відміну від традиційного, а вимога до ресурсів відноситься до обчислювальних вимог запропонованого методу.

Таблиця 2.2 – Порівняльний аналіз обраного способу

	Швидкість навчання	Корисність	Ефективність	Вимоги до ресурсів с
PCA	висока	багатовимірні дані	–	низькі
Випадкова проєкція	середня	обидва типи	більш стабільний, коли розмір змінюється	високі

DOBIN	висока	багатовимірні дані	кращий інструмент зменшення розмірів порівняно з PCA.COV4	–
STRAY	висока	обидва типи	перевершує HDoutliers з точки зору точності та постійного обчислення	низькі
ROBEM	середня	обидва типи	більш успішний порівняно з існуючим	високі
DAE-KNN	висока	багатовимірні дані	точніші порівняно з автономними алгоритмами	високі
OCP	середня	багатофакторні дані	до 88% точніше при правильній класифікації	високі

PCA, DOBIN, STRAY і DAE-KNN мають високу швидкість навчання, що показує ідеальний результат і було доведено порівняно з методами випадкової проекції, ROBEM і OCP. Крім того, більшість методів, які застосовуються для обох умов, є багатофакторними та багатовимірними, оскільки ці дві умови пов’язані одна з одною та є взаємозамінними. Якщо методи неефективні, вони займають занадто багато часу для виявлення аномалій. Згідно з повідомленими дослідженнями, більшість методів продемонстрували чудову здатність долати прокляття розмірності та багатоваріантних характеристик у виявленні аномалій. Нарешті, більшість методів також займають дуже багато часу. Однак ми вважаємо, що кожен метод має свої переваги незалежно від складності проблеми в часі.

2.5 Висновок за розділом

Розділ зосереджений на огляді та обговоренні останніх досліджень, пов’язаних із методами виявлення аномалій у багатовимірних і багатофакторних даних. Крім того, він також надає переваги та недоліки кожного методу відповідно, щоб можна було розробити більш надійний метод. Було виявлено дві проблеми в алгоритмах виявлення аномалій. По-перше, вибір відповідного методу зменшення на основі даних має важливе значення для деяких підходів до зменшення розмірності, тому що іноді важлива інформація може бути втрачена під час процесу зменшення розмірності. Наприклад, є деякі недоліки PCA за наявності шуму. Проте PCA та його модифіковані варіанти, такі як надійний PCA та розріджений PCA, все ще широко використовуються в багатьох додатках завдяки своїй простоті та ефективності.

Тим часом, для випадкової проекції та DOBIN ці методи діють як інструменти зменшення розмірів у попередній обробці даних, щоб допомогти будь-якому алгоритму виявлення аномалій знайти аномалії. Розробка методів пов’язана з відсутністю інтерпретації традиційних методів зменшення розмірів. З іншого боку, метод OCP поєднує статистичне та машинне навчання, зосереджуючись на виявленні аномалії в багатоваріантних умовах. Останніми будуть алгоритм DAE-KNN, ROBEM і STRAY, підхід машинного навчання, який застосовується для виявлення аномалій у багатовимірних і багатофакторних умовах. Ці методи створено не лише для виявлення аномалій, але й для покращення можливостей існуючих методів, за рахунок їх додавання. Наприклад, для DAE-KNN, поєднуючи автокодер і K-найближчих сусідів, ROBEM на основі алгоритму кластеризації ROBPCA та EM, і, нарешті, алгоритм STRAY спрямований на подальше покращення можливостей HDoutliers. По-друге, більшість методів дуже добре справляються з виявленням аномалій, але немає належного тесту, щоб дізнатися, чи є виявлені аномалії справжніми аномаліями чи просто

великими нормальними значеннями. Після цього не слід якось їх видаляти, а, можливо, належним чином дослідити, оскільки аномалії не обов’язково є помилками.

Після дослідження для порівняння різних методів проблем виявлення аномалій у багатовимірних і багатовимірних даних дослідники перейшли до наступного кроку. Наступним кроком є сформулювання більш надійного алгоритму виявлення аномалій, який може добре працювати в багатовимірних і багатовимірних даних і може належним чином розрізняти аномалії, які можуть погано вплинути на дані, або аномалії, які містять цінну інформацію. Потім потрібно оцінити запропонований алгоритм виявлення аномалії з існуючими, щоб порівняти ефективність і точність. Реалізація запропонованих алгоритмів виявлення аномалій може допомогти в прийнятті рішень, покращити продуктивність і вирішити різні складні проблеми.

3 УДОСКОНАЛЕННЯ АЛГОРИТМУ HDOUTLIERS

3.1 Обмеження алгоритму HDoutliers

Хоча алгоритм HDoutliers [21] має багато переваг, деякі характеристики обмежують його можливості. Далі це пояснюється детально.

3.1.1 Розрізнення аномалій через відстань до найближчого сусіда

Алгоритм HDoutliers використовує алгоритм Leader [34] для формування невеликих кластерів точок перед обчисленням відстані до найближчого сусіда. В алгоритмі Leader кожен кластер є кулею у просторі даних великої розмірності. В алгоритмі HDoutliers радіус цієї кулі вибирається таким чином, щоб він був значно нижчим від очікуваного значення відстаней між n(n - 1)/2 парами точок, розподілених випадковим чином у d-вимірному одиничному гіперкубі.

Після формування кластерів за допомогою алгоритму Leader алгоритм HDoutliers вибирає репрезентативні члени з кожного кластера. Потім він обчислює відстані найближчих сусідів для кожного з цих репрезентативних членів. Далі ці відстані використовуються для ідентифікації аномалій на основі припущення, що аномалії створюють великі розриви між типовими даними та аномаліями порівняно з розривами між самими типовими даними. Таким чином, згідно з цим припущенням вважається, що будь-який аномальний кластер з’явиться далеко від кластерів типових точок даних. У результаті відстань найближчого сусіда для цього аномального кластера буде значно вищою, ніж у кластерів типових даних, і таким чином ідентифікуватиме його як аномальний кластер. Потім усі точки даних, що містяться в аномальному кластері, позначаються як аномальні точки в заданому наборі даних.

Однак ще одним припущенням для належної роботи цього методу є те, що будь-які аномальні кластери, присутні в наборі даних, ізольовані. Наприклад, можна розглянути ситуацію, в якій два аномальних скупчення знаходяться дуже близько одне до одного, але далеко від решти типових скупчень. Два кластери стануть найближчими сусідами один до одного, і вони спільно проєктуватимуть їх, будучи аномальними, надаючи дуже малі відстані найближчих сусідів для обох кластерів, які сумісні з відстанями найближчих сусідів решти типових кластерів. Рисунок 4.2 (в, г) далі пояснюють цей аргумент. У цих двох прикладах алгоритм HDoutliers (з кроком кластеризації) оголошує точки аномаліями, лише якщо вони ізольовані, і не вдається виявити аномальні кластери, які мають кілька сусідніх кластерів. Незважаючи на те, що алгоритм HDoutliers включає етап кластеризації з метою ідентифікації аномальних кластерів точок, через дуже малий розмір кулі, яка використовується для створення кластерів (прикладів) у d-вимірному просторі, він не може привести всі точки до одного кластера і натомість створює кілька аномальних кластерів, які знаходяться дуже близько один до одного. Потім ці аномальні кластери стають найближчими сусідами один з одним і мають дуже малі найближчі відстані для репрезентативного члена кожного кластера. Оскільки виявлення аномалій повністю залежить від цих відстаней до найближчих сусідів, і оскільки аномальні кластери не виявляють жодних значних відхилень від типових кластерів щодо відстаней до найближчих сусідів, алгоритм тепер не може виявити ці точки як аномалії і тим самим збільшує кількість хибно-негативних результатів.

3.1.2 Проблеми через кластеризацію алгоритму Leader

Після формування кластерів точок даних алгоритм HDoutliers повністю ігнорує щільність точок даних. Коли він формує кластери точок даних за допомогою алгоритму Leader, він вибирає репрезентативний член із кожного

кластера та виконує подальший аналіз лише з використанням цих репрезентативних членів. Рисунок 4.2 (д) надає приклад, пов’язаний з цим питанням. Цей набір даних є бімодальним набором даних з аномальною точкою, розташованою між двома типовими класами.

Весь набір даних містить 2001 точку даних. Точки даних, зібрані в крайньому лівому верхньому куті, представляють один типовий клас із 1000 точок даних. Другий типовий клас точок даних зібраний у крайньому правому нижньому куті з ще 1000 точками даних. Оскільки цей другий клас точок даних є ущільненим за змістом, 1000 точок даних тепер загорнуті в одну кулю під час формування кластерів за допомогою алгоритму Leader.

В алгоритмі HDoutliers наступним кроком є вибір одного члена з кожного з цих кластерів. Коли він вибирає репрезентативний член із цієї кулі, що містить 1000 точок даних, він ігнорує решту 999 точок даних при виявленні аномалій. Цей крок вводить алгоритм в оману, а інші кроки алгоритму розглядають цей репрезентативний елемент як ізольовану точку даних, хоча він оточений 999 сусідніми точками даних у вихідному наборі даних. Таким чином, усі точки даних у всьому цьому класі оголошуються алгоритмом як аномалії, хоча він містить половину набору даних. Було висунуто пропозицію [35] розглядати джиттер не як ідеальне рішення, а як альтернативу для пом’якшення цієї проблеми. Також стверджується, що проблема, як правило, не виникає у просторах даних великої розмірності, де така деталізація менш імовірна. Однак тоді це породжує проблему сусідніх аномальних кластерів (як показано на рисунку 4.2 (в, г)), які окремо здаються типовими або викликають обмежену підозру (через наявність інших сусідніх аномальних кластерів), однак їх співпоширення є дуже аномальним.

Рисунок 4.2 (е) демонструє іншу ситуацію, в якій хибно-негативні результати збільшуються через етап кластеризації. Цей двовимірний набір даних містить 1001 точку даних. Точки даних, зібрані в крайньому лівому верхньому куті, представляють типовий клас, що охоплює 1000 точок даних, а ізольована точка даних у крайньому правому нижньому куті представляє

аномалію. Оскільки цей типовий клас з 1000 точок даних дуже компактний, він створює лише 14 кластерів за допомогою алгоритму Leader. Загалом набір даних утворює 15 кластерів із кластером, створеним ізольованою точкою, розташованою в крайньому правому нижньому куті. Незважаючи на те, що вихідний набір даних містить 1001 точку даних, алгоритм розглядає лише 15 точок даних (репрезентативний член кожного кластера) для розрахунку аномального порогу. Тепер це число недостатньо велике, щоб дати стабільну оцінку аномального порогу. Через це незнання щільності вхідного набору даних тепер не вдається виявити очевидну аномальну точку в крайньому лівому нижньому куті.

3.1.3 Проблема з обчисленням порогу

Для реалізації алгоритму HDoutliers доступний пакет R [36]. Згідно з реалізацією пакета R, поточна версія алгоритму HDoutliers використовує наступного потенційного кандидата на аномалії при обчисленні аномального порогу в кожній ітерації алгоритму пошуку знизу вгору. Такий підхід за певних обставин призводить до підвищення рівня помилкового виявлення. В запропонованому в даній роботі алгоритму можна уникнути цього обмеження.

3.2 Запропонований покращений алгоритм

У цьому розділі розглядається запропонований покращений алгоритм для виявлення аномалій у даних великої розмірності. Запропонований алгоритм має на меті подолати обмеження алгоритму HDoutliers і таким чином розширити його можливості.

3.2.1 Вхідні дані алгоритму

Вхідними даними для алгоритму є набір екземплярів даних, де кожен екземпляр даних може бути реалізацією лише одного атрибута або набору атрибутів (функції, вимірювання та розміри). Це дослідження обмежено обговоренням кількісних даних; отже, входом може бути вектор, матриця або кадр даних d (>1) числових змінних, де кожен стовпець відповідає атрибуту, а кожен рядок відповідає спостереженню цих атрибутів. Тоді фокус зосереджений на виявленні аномальних екземплярів (рядків) у наборі даних.

3.2.2 Нормалізація стовпців

Оскільки пропонований алгоритм базується на визначенні відстані аномалії, відстані найближчих сусідів між екземплярами даних у просторі даних великої розмірності є ключовою інформацією для алгоритму виявлення аномалій. Однак змінні з великою дисперсією можуть справляти непропорційний вплив на обчислення евклідової відстані. Щоб зробити змінні еквівалентної ваги, стовпці даних спочатку нормалізуються таким чином, щоб дані були обмежені одиничним гіперкубом. Цю нормалізацію зазвичай називають мінімаксною (min-max) нормалізацією, яка передбачає лінійне перетворення вхідних даних у результат в діапазоні від 0 до 1. Цей тип перетворення не змінює розподіл і не стискає точки разом, маскуючи аномалії.

3.2.3 Пошук найближчого сусіда

Після нормалізації стовпців набору даних обчислюється k-відстань найближчого сусіда з максимальним розривом для кожного екземпляра. Використовуючи цей показник, вдалося подолати вищезазначені обмеження алгоритму HDoutliers.

Для кожного окремого спостереження алгоритм спочатку обчислює k-відстань до найближчих сусідів, d_i, KNN, де i = 1,2, ...,k . Потім він обчислює

послідовні різниці між відстанями Δ_i, KNN. Далі він вибирає k-відстань найближчого сусіда з максимальним розривом Δ_i, max. Рисунок 3.1 ілюструє, як ці дії допомагають вдосконаленому алгоритму виявляти аномальні точки або аномальні кластери точок; аномалії представлені червоними трикутниками, а чорні точки відповідають типовій поведінці.

На рисунку 3.1(а) набір даних містить лише одну аномалію в (15; 16,5). Для цього набору даних відстань найближчого сусіда може відрізнити аномальну точку від решти типових точок, оскільки відстань найближчого сусіда для аномальної точки значно більша (14,8), ніж для решти типових точок. Рисунок 3.1(б) показує зміну k-відстаней до найближчих сусідів аномалії в (15; 16,5). Для цього набору даних k-відстань найближчого сусіда з максимальним розривом виникає, коли k = 1. Другий набір даних на рисунку 3.1(в), має три аномалії (мікрокластер) навколо (15; 16,5). Якщо обчислювати лише відстані найближчих сусідів для кожного спостереження, тоді три аномалії не можна відрізнити від типових точок, оскільки їхні значення дуже малі (0,7) порівняно з значеннями більшості типових точок із відстанями найближчих сусідів приблизно (0,0015 до 2,5).

Однак три аномалії можна відрізнити від своїх типових точок відносно k-найближчих сусідніх відстаней із максимальним розривом (рисунок 3.1(г)). Для трьох аномалій на рисунку 3.1(г), відстань третього найближчого сусіда має максимальний розрив (рисунок 3.1(д)), і тепер три точки легко розрізнити як аномалії відносно k-відстані найближчих сусідів з максимальним розривом. Таким чином, використовуючи k-відстань найближчих сусідів з максимальним проміжком, алгоритм розбіжності отримує здатність виявляти як аномальні синглетони, так і мікрокластери. Завдяки цьому підходу можна зменшити частоту хибних виявлень і, таким чином, усунути обмеження алгоритму HDoutliers, одночасно отримавши можливість виявляти мікрокластери. Це також дуже проста, але розумна інвестиція порівняно з часом, який витрачає алгоритм Lesder на формування невеликих кластерів для виявлення мікрокластерів (особливо для наборів

даних великих розмірів) у алгоритмі HDoutliers. Крім того, для кожної точки відповідні k-відстані найближчого сусіда з максимальним розривом діють як аномальна оцінка, що вказує на ступінь аномалії.

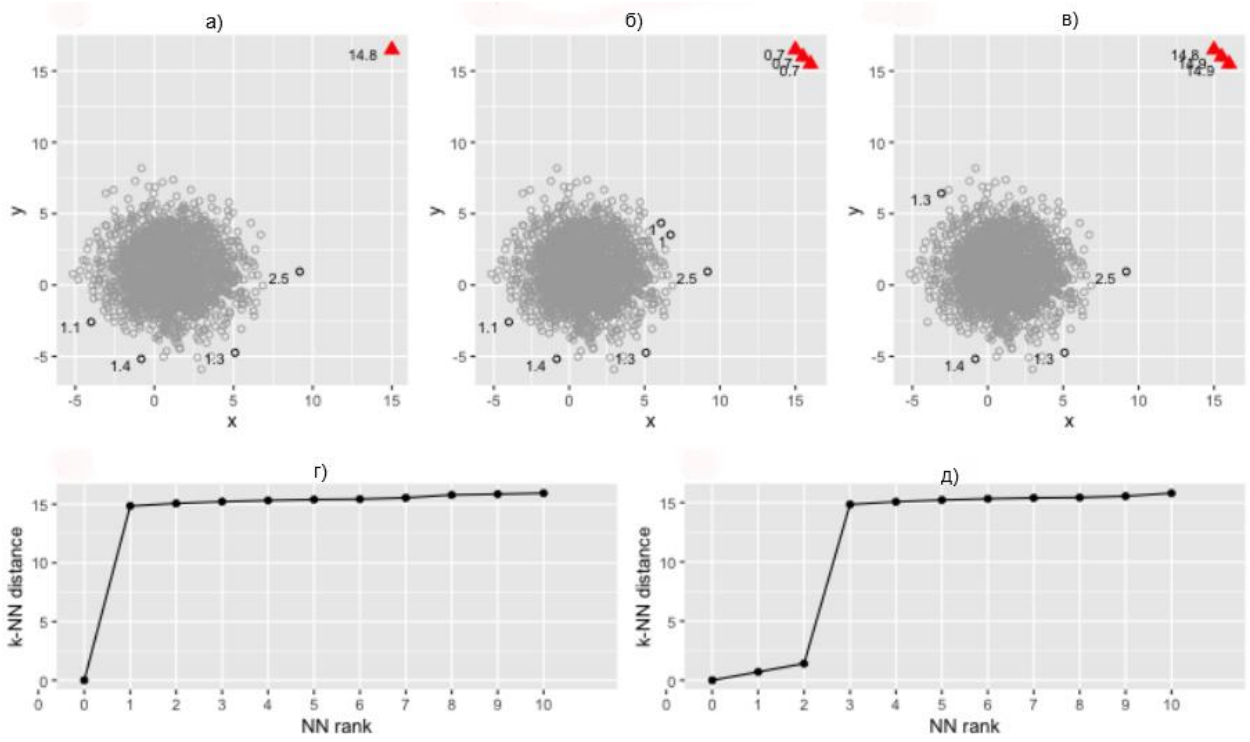


Рисунок 3.1 – Різниця між відстанню найближчого сусіда та k-відстанню найближчого сусіда з максимальним розривом

У цій роботі розглядаються як точні, так і наближені методи пошуку k-найближчих сусідів. Грубий пошук передбачає проходження всіх можливих пар точок для виявлення k-найближчих сусідів для кожного екземпляра даних, і тому досліджуються точні k-найближчі сусіди. І навпаки, k-вимірні дерева (дерева k-d) використовують структури просторових даних, які розділяють простір, щоб забезпечити ефективний доступ до визначеної точки запиту [37]. Таким чином, це передбачає пошук приблизно k-найближчих сусідів навколо вказаної точки запиту.

У поточному алгоритмі параметр k, який визначає розмір околиці, вводиться як параметр, визначений користувачем, який можна вибрати

відповідно до програми. Один із способів інтерпретації ролі k у алгоритмі збій – розглядати його як мінімально можливий розмір для типового кластера в даному наборі даних. Якщо розмір аномального кластера менший за k, він буде виявлений як мікрокластер за допомогою пропонованого алгоритму. Вибір k має різні ефекти для різних вимірів і розмірів даних. Можна встановити k рівним 1, якщо в наборі даних немає мікрокластерів, і таким чином зосередитися на локальних і глобальних аномальних точках. Високі значення k рекомендуються для наборів даних із високими розмірами через «прокляття розмірності».

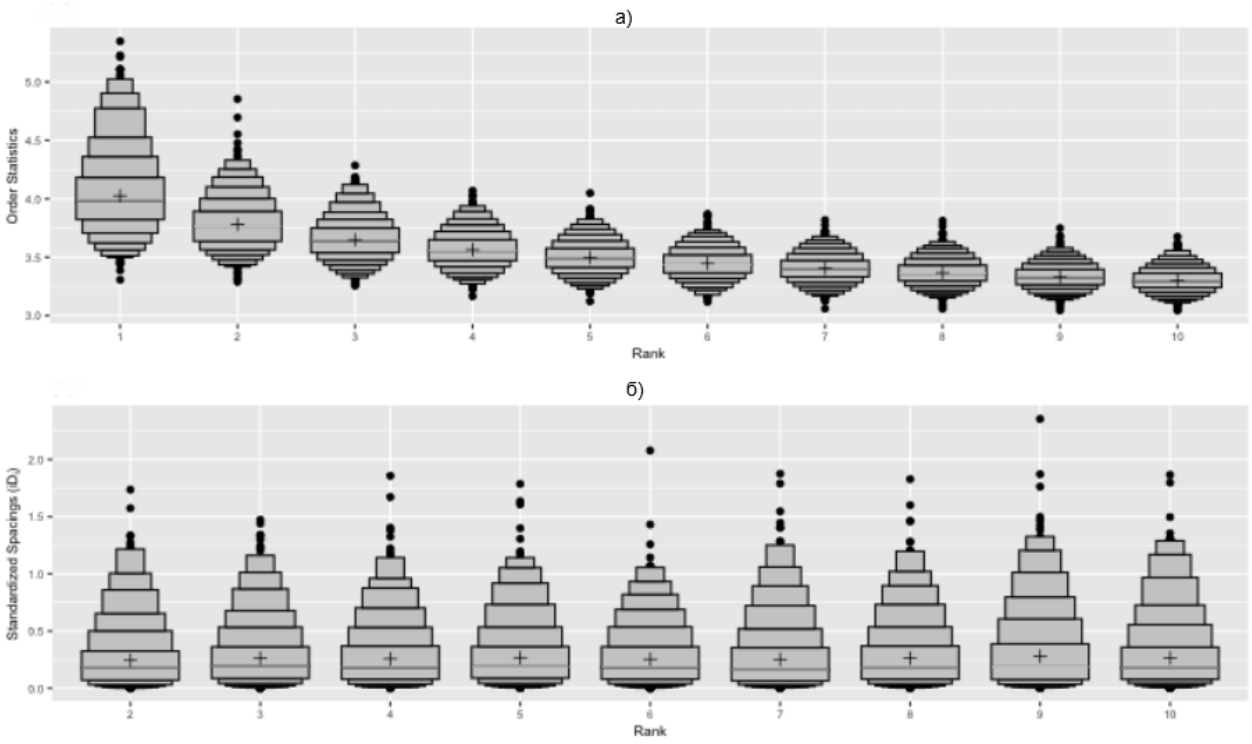
3.2.4 Розрахунок порогу

Аномальні бали призначають кожній точці ступінь аномалії. Однак для певних застосувань також важливо класифікувати типові та аномальні точки для подальшого аналізу першопричини. В ідеалі слід віддати перевагу універсальному порогу, щоб однозначно відрізняти аномальні точки від типових точок. Алгоритм HDoutliers визначає аномальний поріг на основі теорії екстремальних значень, розділу теорії ймовірностей, який стосується поведінки статистики екстремального порядку в даній вибірці.

Розрахунок аномального порогу є застосуванням теореми про відстань Вайсмана (теорема 1), який застосовний до розподілу даних, охоплених максимальною областю тяжіння розподілу Гамбеля. Ця вимога задовольняється широким діапазоном розподілів, починаючи від розподілів із легкими хвостами до помірно важких хвостів, які спадають до нуля швидше, ніж будь-яка степенева функція [38]. Приклади включають експоненціальний, гамма-, нормальний і лог-нормальний розподіли з експоненціально спадаючими хвостами.

Нехай $X_1, X_2, ..., X_n$ – вибірка з функції розподілу F і нехай $X_{1:n} \geq X_{2:n} \geq ... \geq X_{n:n}$ – статистика порядку. Доступні дані $X_{1:n}, ..., X_{k:n}$ для деякого фіксованого k.

Теорема 1 (теорема про відстань). Нехай $D_{i,n} = X_{i:n} - X_{i+1:n}, (i = 1,...,k)$ інтервал між послідовними порядковими статистиками. Якщо F знаходиться в області максимального притягання розподілу Гамбеля, відстані $D_{i,n}$ асимптотично незалежні та експоненціально розподілені із середнім, пропорційним i^{-1}.



(а) Розподіл статистики у порядку спадання $X_{i:n}$

(б) розподіл стандартизованих інтервалів $iD_{i,n}$ для $i \in \{1,...,10\}$ для 1000 вибірок, кожна з яких містить 20000 випадкових чисел зі стандартного нормального розподілу

Рисунок 3.2 – Функції розподілу

Проілюструємо цю теорему за допомогою рисунку 3.2, який показує розподіл статистичних даних у порядку спадання $(X_{i:n})$ і стандартизованих інтервалів $(iD_{i,n})$ для $i \in \{1,...,10\}$ для 1000 вибірок, кожна з яких містить 20000 випадкових чисел зі стандартного нормального розподілу. Рисунок

3.2(а) показує розподіл $X_{i:n}$ із середніми $X_{i:n}$, зображеними чорними хрестами. Проміжки між послідовними чорними хрестиками дають проміжки між статистичними даними вищого порядку $\left(D_{i,n}\right)$. Слід зауважити, що нормальний розподіл знаходиться в області максимального тяжіння розподілу Гамбеля і що цей приклад не містить викидів. Наслідок теореми 1 полягає в тому, що стандартизовані інтервали $\left(iD_{i,n}\right)$ для $\left(i=1,...,K\right)$ є приблизно iid [25]. Рисунок 3.2(б) показує розподіл стандартизованих інтервалів $\left(iD_{i,n}\right)$ для $\left(i=1,2,...,10\right)$ для 1000 зразків розміром 20000. Кожна діаграма літерних значень має приблизно форму експоненціального розподілу.

Далі обчислюється аномальний поріг з підмножини точок, що охоплює 50 відсотків із них із найменшими аномальними балами, припускаючи, що ця підмножина містить аномальні бали, які відповідають типовим точкам даних, а решта підмножини містить бали, які відповідають можливим кандидатам на аномалії. Дотримуючись теореми про відстань Вайсмана, підбирається експоненціальний розподіл до верхнього хвоста викидних балів першої підмножини, а потім обчислюються верхні 1-α точки підігнаної кумулятивної функції розподілу, таким чином визначаючи аномальний поріг для наступного аномального рахунку. Потім із підмножини, що залишилася, вибирається точка з найменшою аномальною оцінкою. Якщо ця аномальна оцінка перевищує граничну точку, всі точки в решті підмножини позначаються як аномалії, і пошук аномалій припиняється. В іншому випадку точка оголошується як типова точку та додається до підмножини типових точок. Потім оновлюється гранична точка, включаючи останнє додавання. Цей алгоритм пошуку продовжується, доки не буде знайдено аномальний показник, який перевищує останню граничну точку. Цей алгоритм відомий як алгоритм «пошуку знизу вгору». Розрахунок порогу виконується за припущенням, що розподіл k-найближчих сусідів із максимальним розривом знаходиться в максимальній області тяжіння розподілу Гамбеля, який

охоплює широкий діапазон розподілів.

3.2.5 Вихід

Аномалії вимірюються за двома шкалами: (1) бінарна класифікація та (2) оцінка викидів. Відповідно до двійкової класифікації екземпляри даних класифікуються як типові або аномальні за допомогою керованого даними порогового значення аномальності на основі теорії екстремальних значень. Цей тип класифікації важливий, якщо наступні етапи процесу аналізу даних автоматизовані. Розсіяний алгоритм також призначає аномальний бал кожному екземпляру даних, щоб вказати ступінь відмінності кожного вимірювання. Ці аномальні бали дозволяють користувачеві ранжувати та вибирати найбільш серйозні або відповідні аномальні точки для аналізу першопричини та негайного вжиття запобіжних заходів. Алгоритм HDoutliers, яка забезпечує лише двійкову класифікацію, не дозволяє користувачеві безпосередньо зробити такий вибір, щоб спрямувати свою увагу на більш значні аномальні випадки. І навпаки, різні методи, запропоновані в літературі, дають аномальні бали, але аномальний поріг визначається користувачем і залежить від програми. Вихідні дані, створені пропонованим алгоритмом, – це комплексне рішення, яке містить необхідні вимірювання аномалій для подальших дій.

4 ЕКСПЕРИМЕНТИ

Алгоритм HDoutliers є потужним алгоритмом у сучасних найсучасніших методах виявлення аномалій у даних великої розмірності. Основна увага алгоритму збитків полягає в усуненні деяких обмежень алгоритму HDoutliers, які перешкоджають його продуктивності за певних обставин. Тут ми виконуємо експериментальну оцінку точності та обчислювальної ефективності нашого паразитного алгоритму порівняно з алгоритмом HDoutliers. Хоча ці приклади досить обмежені в кількості та здебільшого обмежені двовимірними наборами даних, їх слід розглядати лише як просту ілюстрацію ключових особливостей алгоритму збій, який перевершує алгоритм HDoutliers.

4.1 Оцінка продуктивності

Перший експеримент (рисунок 4.1) був розроблений для перевірки впливу розмірності, розміру даних і методу пошуку k-найближчого сусіда на час роботи різних версій двох алгоритмів: пропонованого і HDoutliers.

Алгоритм HDoutliers має дві версії. Перша версія обчислює відстань до найближчого сусіда для кожного екземпляра даних і не передбачає жодного етапу кластеризації перед обчисленням відстані до найближчого сусіда. Ця версія алгоритму (версія 1 HDoutliers, далі) рекомендована для малих вибірок ($n<10000$).

Друга версія використовує алгоритм Leader для формування кількох кластерів точок, а потім вибирає репрезентативного члена з кожного кластера. Далі обчислюються відстані найближчих сусідів лише для вибраних репрезентативних членів. Порівняно з версією 1 алгоритму HDoutliers (без кроку кластеризації) (рисунок 4.1(a)), версія 2 із кроком кластеризації є надзвичайно повільною для вищих розмірів (>10), а час

виконання збільшується швидше зі збільшенням розміру вибірки. На рисунку 4.1(б)) показана лише частина вимірювань повного експерименту другої версії алгоритму HDoutliers для чіткого порівняння різних версій двох алгоритмів (пропонований і HDoutliers). Рисунок 4.1(є) представляє повну версію рисунка 4.1(б); чорна рамка в 4.1(є) покриває область зображення 4.1(б). Додатковий етап кластеризації у другій версії алгоритму HDoutliers, який є важливим для виявлення мікрокластерів, займає надзвичайно багато часу, особливо з великими вибірками з більшими розмірами.

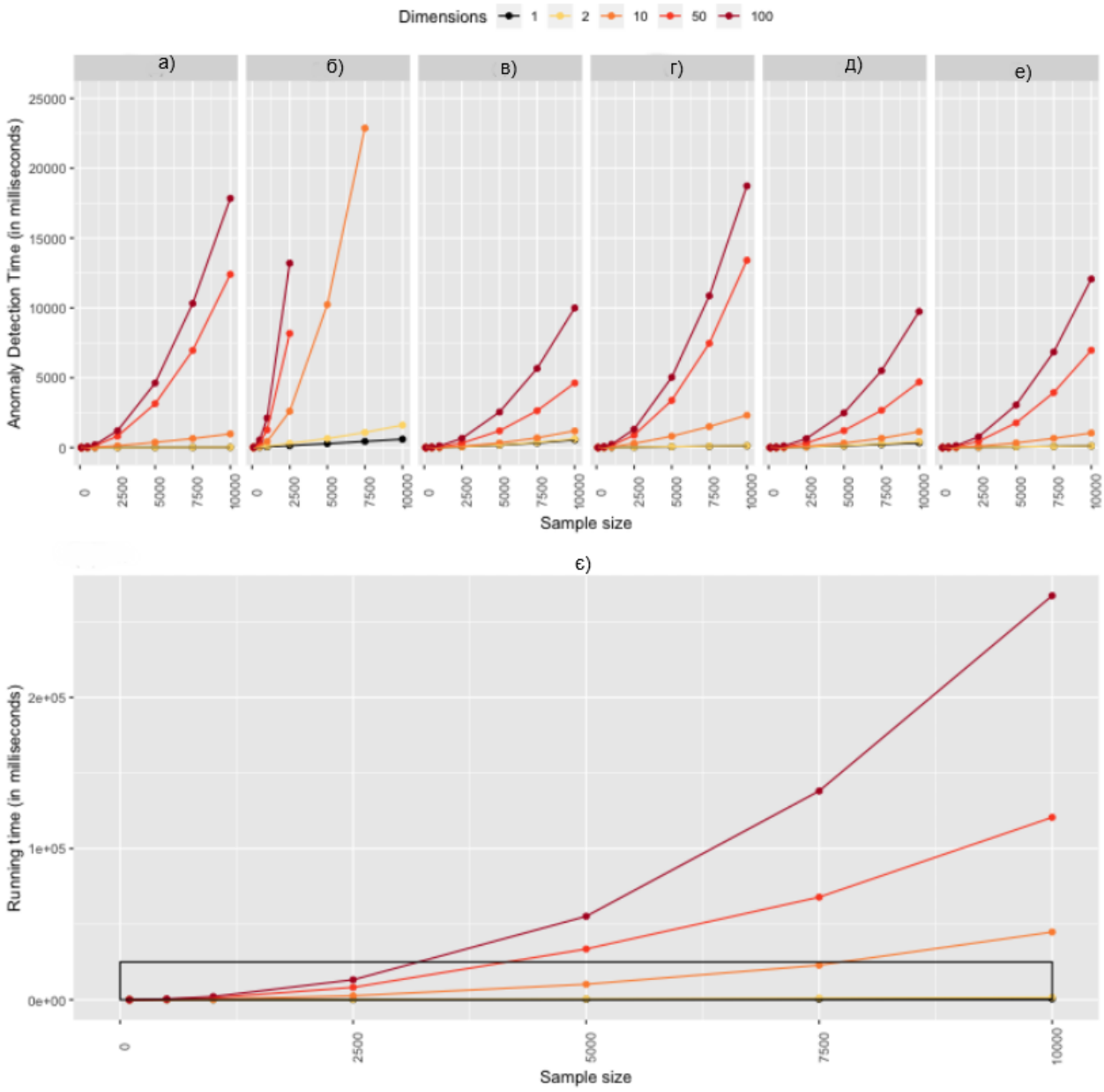


Рисунок 4.1 – Продуктивність масштабованості

Рисунки 4.1(в-е) відповідають пропонованому алгоритму. У цьому експерименті, для визначення впливу методів пошуку k-найближчих сусідів, розглянуті як точні (brute force), так і наближені (kd-treess) алгоритми пошуку найближчих сусідів.

Для програмного середовища R доступні багато реалізацій алгоритмів пошуку k-найближчих сусідів. В цій роботі для порівняльного аналізу розглянуті пакети FNN [39] (рисунок 4.1(в, г)) і «nabor» [37] (рисунок 4.1(д, е)) системи R. Набір пакетів R містить швидку бібліотеку k-найближчого сусіда, написану на шаблоні C++. Було помічено, що пошук k-(>1) найближчих сусідів (рис 4.1(а), у цьому прикладі k встановлено на 10) замість лише одного (k=1) найближчого сусіда (рисунок 4.1(г)) лише трохи збільшує час виконання, оскільки збільшується кількість екземплярів.

Результати на обох рисунках 4.1(а) і 4.1(г) базуються на приблизних відстанях найближчих сусідів з використанням алгоритму пошуку найближчих сусідів kd-treess. помічено, що реалізація kd-treess у пакеті «nabor» (рисунок 4.1(е)) набагато швидше, ніж реалізація пакету FFN (рисунок 4.1 (г)). Здається парадоксом, але зі збільшенням розмірності час роботи пропонованого алгоритму з kd-trees (рисунок 4.1 (г, е)) зростає набагато швидше, ніж алгоритм brute force, який передбачає пошук усіх можливих пар точок для виявлення k-найближчих сусідів для кожного екземпляра даних (рисунок 4.1 (в, д)).

Інші дослідження також свідчать про подібний результат для багатьох алгоритмів, заснованих на kd-trees в різних варіантах. Це може бути пов’язано з розпаралелюванням і моделями доступу до пам’яті двох механізмів пошуку. Алгоритм brute force легко розпаралелювати, оскільки він передбачає незалежний пошук усіх можливих кандидатів для кожного екземпляра даних. На відміну від цього, алгоритм пошуку kd-trees природно послідовний, і тому його важко реалізувати в паралельних системах із помітним прискоренням.

4.2 Оцінка якості

Далі було оцінено частоту хибно-позитивних результатів (типові точки, неправильно визначені як аномалії) помилкового алгоритму, запустивши його багато разів на випадкових даних. Значення, представлені в таблиці 4.1, ґрунтуються на 1000 ітераціях і є середніми.

Таблиця 4.1 – Показники продуктивності – показники хибно-позитивних результатів

Метод	Розмір	100	500	1000	2500	5000	7500	10000
HDoutliers WoC	1	0,017	0,011	0,008	0,007	0,005	0,005	0,004
	10	0,002	0,002	0,002	0,002	0,002	0,002	0,002
	100	0,001	0,001	0,001	0,001	0,001	0,001	0,001
HDoutliers WC	1	0,036	0,024	0,024	0,019	0,017	0,014	0,013
	10	0,006	0,006	0,006	0,005	0,005	0,005	0,005
	100	0,003	0,003	0,003	0,003	0,003	0,003	0,003
Пропонований – brute force	1	0,006	0,003	0,002	0,002	0,002	0,001	0,001
	10	0,001	0,001	0,001	0,001	0,001	0,001	0,000
	100	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Пропонований - FNN kd-trees	1	0,006	0,003	0,002	0,002	0,002	0,001	0,001
	10	0,001	0,001	0,001	0,001	0,001	0,001	0,000
	100	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Пропонований – «nabor» brute	1	0,006	0,003	0,002	0,002	0,002	0,001	0,001
	10	0,001	0,001	0,001	0,001	0,001	0,001	0,000
	100	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Пропонований - «nabor» kd-trees	1	0,006	0,003	0,002	0,002	0,002	0,001	0,001
	10	0,001	0,001	0,001	0,001	0,001	0,001	0,000
	100	0,000	0,000	0,000	0,000	0,000	0,000	0,000

Були застосовані різні версії двох алгоритмів (пропонований і HDoutliers) до наборів даних, де кожен стовпець генерується випадковим чином із стандартизованого нормального розподілу. Усі набори даних не містять аномалій HDoutliers WoC: алгоритм HDoutliers без етапу кластеризації; HDoutliers WC: алгоритм HDoutliers із кроком кластеризації.

У кожному тесті критичне значення α було встановлено рівним 0,05. Порівняно з алгоритмом HDoutliers низький рівень хибно-позитивних результатів був досягнутий для паразитного алгоритму для всіх вимірів і розмірів вибірки. На відміну від алгоритму HDoutliers, у пропонованому спостерігався набагато менший рівень помилкового виявлення навіть для невеликих наборів даних із меншими розмірами. Жодної різниці не спостерігалося в різних версіях алгоритму розбіжності з різними механізмами пошуку найближчих сусідів та їх різними реалізаціями.

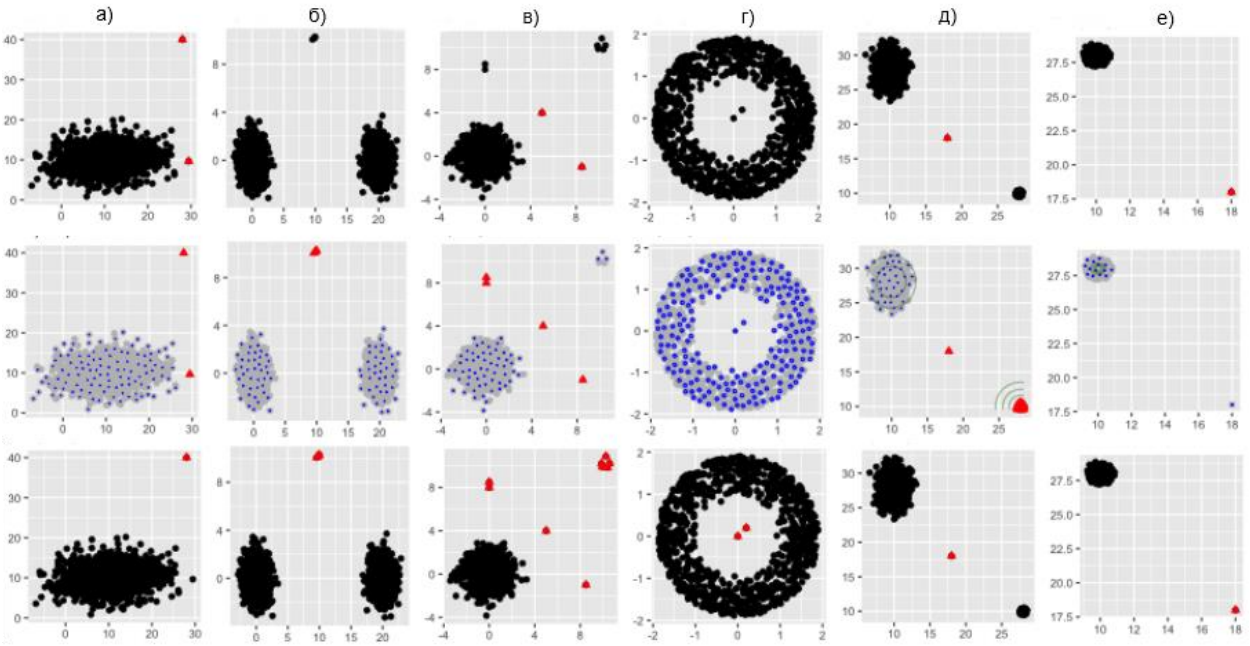


Рисунок 4.2 – Якість алгоритму

Рисунок 4.2 демонструє, як пропонований алгоритм перевершує дві версії алгоритму HDoutliers за різних обставин. Верхній ряд показує результати алгоритму HDoutliers без етапу кластеризації. Середній ряд

показує результати алгоритму HDoutliers з кроком кластеризації; синім кольором позначені репрезентативні члени, вибрані з кожного кластера, сформованого алгоритмом Leader. Нижній ряд показує результати вдосконаленого алгоритму з пошуком k-найближчого сусіда з використанням brute force. Виявлені аномалії позначені червоними трикутниками. Цей обмежений набір прикладів було вибрано з метою висвітлення деяких ключових особливостей пропонованого алгоритму.

Особливість 1. Всі три алгоритми змогли правильно зафіксувати аномальну точку в крайньому правому верхньому куті рисунка 4.2(а). Однак дві версії алгоритму HDoutliers мають тенденцію генерувати деякі помилкові спрацьовування, особливо з малими розмірами.

Особливість 2. Рисунок 4.2(б) демонструє здатність мати справу з мультимодальними типовими класами. Два кластери внизу графіка представляють два типові класи. Виявити мікрокластер у центрі зверху, який містить три екземпляри аномальних даних, змогла лише друга версія алгоритму HDoutliers, що використовує етап кластеризації. Однак формування невеликих кластерів до обчислення відстані не завжди допомагає виявити мікрокластери.

Особливість 3. Рисунок 4.2(в) показує ситуацію, коли навіть друга версія алгоритму HDoutliers не може виявити мікрокластери. Алгоритм Leader в HDoutliers використовує дуже маленьку кульку фіксованого радіуса для формування кластерів, і тому тепер йому не вдається захопити п’ять точок в один кластер і натомість генерує три маленькі кластери, які знаходяться дуже близько один до одного. Обидві версії алгоритму HDoutliers тепер не можуть виявити мікрокластер у крайньому правому верхньому куті, оскільки набір даних порушує одну з основних вимог щодо ізоляції аномальних точок або аномальних кластерів. У пропонованому алгоритмі значення k було встановлено рівним 10. Можна інтерпретувати значення k як максимально допустимий розмір для мікрокластеру. Тобто, щоб невеликий кластер був мікрокластером, кількість точок даних у цьому

кластері має бути меншою за k. В іншому випадку кластер вважається типовим кластером.

Особливість 4. Рисунок 4.2 (г) демонструє здатність виявляти інлієри. Алгоритм HDoutliers також має здатність виявляти інлієри лише тоді, коли є ізольовані інлієри, вільні від аномальних сусідів. Обидві версії алгоритму HDoutliers не в змозі виявити два внутрішніх значення, оскільки вони знаходяться дуже близько одне до одного і, таким чином, спільно проєктують їх як аномальні.

Особливість 5. Як пояснено у 3.1, рисунок 4.2 (д) показує, що етап кластеризації другої версії алгоритму HDoutliers може ввести в оману процес виявлення і, таким чином, збільшити рівень хибних спрацьовувань. Щільні області набору даних позначені кривими щільності. Видно два типові кластери, один у крайньому лівому верхньому куті, а інший у крайньому правому нижньому куті. Проміжний елемент також присутній між двома типовими класами. Після формування кластера за допомогою алгоритму Leader з кожного кластера вибирається лише один репрезентативний член для розрахунку відстані найближчого сусіда. Вибраний учасник тепер ізольований і отримує дуже високу аномальну оцінку, очолюючи весь типовий кластер у крайньому правому нижньому куті з 1000 балами, які будуть визначені як аномальні. На відміну від цього, пропонований алгоритм позбавлений цих проблем, оскільки він не передбачає жодного етапу кластеризації перед обчисленням відстані найближчого сусіда.

Особливість 6. Як пояснено у 3.2, рисунок 4.2 (е) показує, що етап кластеризації може збільшити рівень хибно-негативних результатів. Цей набір даних містить один типовий клас, який щільно стиснутий за змістом (крайній лівий верхній кут) і очевидну аномалію в крайньому правому нижньому куті. Оскільки типовий клас є щільним кластером, лише кілька точок даних вибираються з типового класу для обчислення найближчого сусіда. У цьому прикладі етап кластеризації суттєво зменшує вибірку вхідного набору даних, що призводить до величезної втрати інформації в

представленні даних. Сині точки на рисунку 4.2 (е) представляють вибрані члени з кожного кластера для розрахунків найближчих сусідів. Тепер зменшеного розміру вибірки недостатньо для правильного розрахунку аномального порогу на основі теорії екстремальних значень.

4.3 Практичне використання

Було застосувано пропонований алгоритм до набору даних, отриманих від автоматизованої системи підрахунку пішоходів із 43 датчиками в місті Мельбурн, Австралія [40], з метою визначити незвичну поведінку пішоходів у муніципалітеті. Виявлення такої незвичайної критичної поведінки пішоходів у різних місцях міста в різний час доби є важливим, оскільки це є прямим показником економічних умов міста, пов’язаної діяльності та безпеки і зручності пішоходів. Це також направляє та інформує про прийняття рішень і планування. Практичний приклад ілюструє, як пропонований алгоритм можна використовувати для роботи з іншими структурами даних, такими як часові дані та потокові дані, використовуючи розробку функцій.

4.4 Обробка часових даних

Для наочності період навчання обмежений одним місяцем з 1 по 31 грудня 2018 року. Рисунок 4.7 показує кількість пішоходів у 43 місцях міста Мельбурн у різний час доби. Кожна діаграма розсіювання відповідає негативно викривленому розподілу. Загалом, будні мають двомодальний розподіл, а вихідні – унімодальний. Тепер мета полягає в тому, щоб виявити дні з незвичною поведінкою. Оскільки це включає велику колекцію діаграм розсіювання, ручний моніторинг займає багато часу, а незвичайну поведінку важко визначити візуальним оглядом.

Виявлення аномальних діаграм розсіювання з великої колекції діаграм

розсіювання потребує певної попередньої обробки. Зокрема, щоб застосувати пропонований алгоритм, потрібно перетворити цей вхідний набір даних із великою колекцією діаграм розсіювання у набір даних великої розмірності. Простіший підхід полягає у використанні функцій, які описують різні форми та шаблони діаграм розсіювання. Обчислювальні функції, які описують значущі форми та шаблони на заданій діаграмі розсіювання, є простими за допомогою scagnostics (діагностики діаграми розсіювання), розробленої [19].

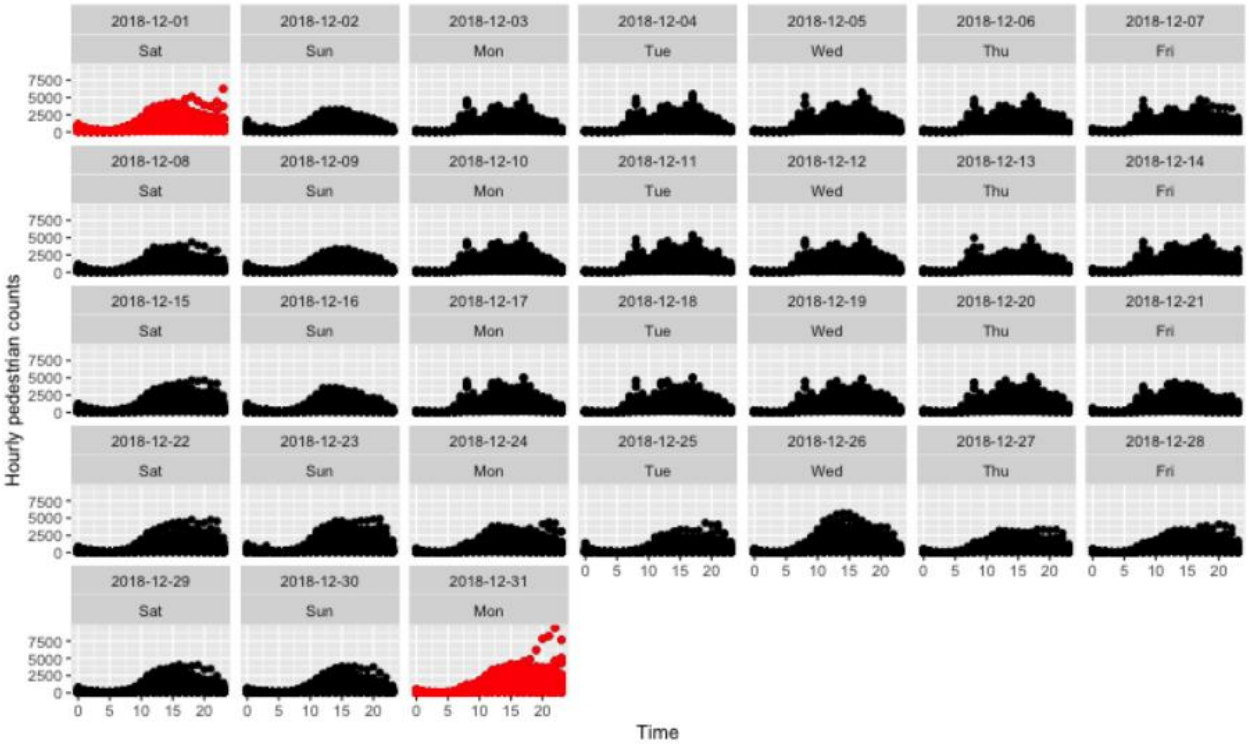


Рисунок 4.3 – Діаграма розсіювання щогодини підрахунку пішоходів у 43 місцях у місті Мельбурн, Австралія, з 1 по 31 грудня 2018 року

Для поточного дослідження вибрано п’ять ознак: віддалені, опуклі, тонкі, тягучі та монотонні [19]. Спеціально відібрано ці функції для конкретного випадку використання. Витягнувши ці п’ять ознак із кожної діаграми розсіювання, перетворюємо оригінальну колекцію діаграм розсіювання у багатовимірний набір даних із п’ятьма вимірами та 31 екземпляром даних. рисунок 4.3 забезпечує відображення на основі ознак

оригінальної колекції діаграм розсіювання. Кожна точка в цьому багатовимірному просторі даних відповідає одній діаграмі розсіювання (або дню) у вхідній колекції діаграм розсіювання. У цьому багатовимірному просторі алгоритм виявляє дві аномальні точки, які позначені червоним кольором на рисунку 4.4. На рисунку 4.3 відповідні діаграми розсіювання (або дні) позначені червоним кольором. На кожному графіку рисунка 4.4 аномалії, визначені алгоритмом, представлені червоним кольором.

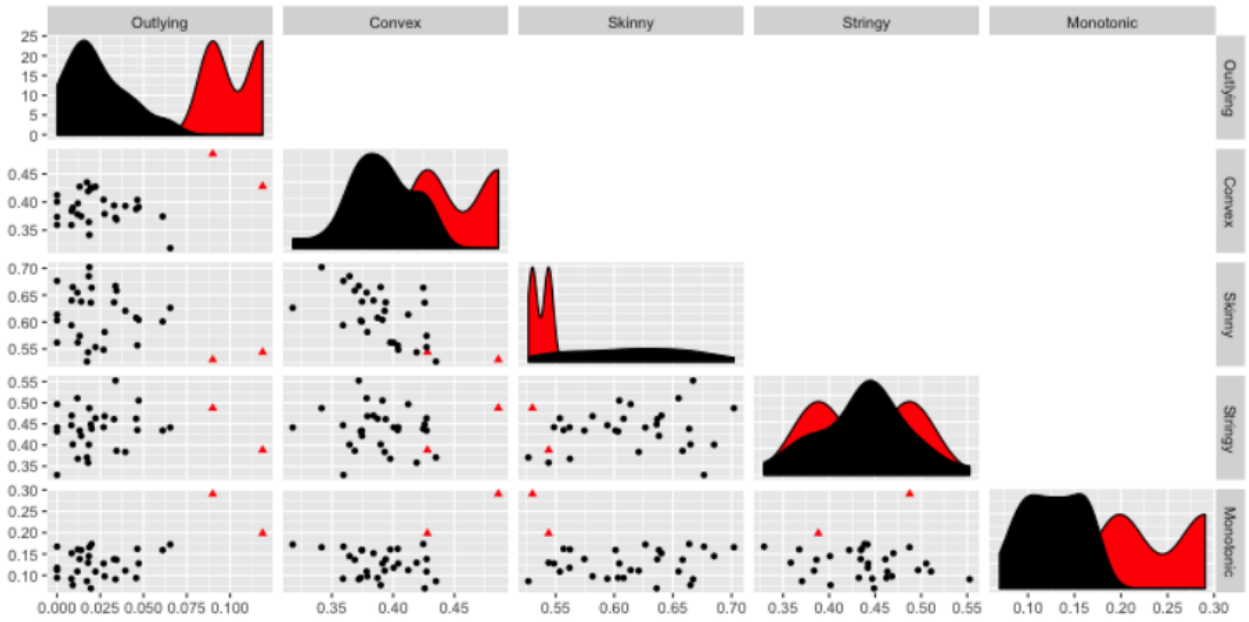


Рисунок 4.4 – Представлення колекції діаграм розсіювання на основі функцій за допомогою Scagnostics

Візуальний огляд також підтверджує аномальну поведінку цих двох діаграм розсіювання. Обидва дні, 1 грудня 2018 року та 31 грудня 2018 року, демонструють незвичайне підвищення пізніше того ж дня. Один вибраний день, 31 грудня 2018 року, є очевидною аномалією, оскільки це новорічна ніч, а пов’язаний із цим феєрверк у Саутбенку в місті Мельбурн приваблює багато тисяч відвідувачів. Подальші розслідування щодо 1 грудня 2018 року показали, що на Мельбурнському майданчику для крикету відбувся музичний концерт о 20:00, і незвичайне підвищення пізніше цього дня могло

бути пов’язане з учасниками концерту.

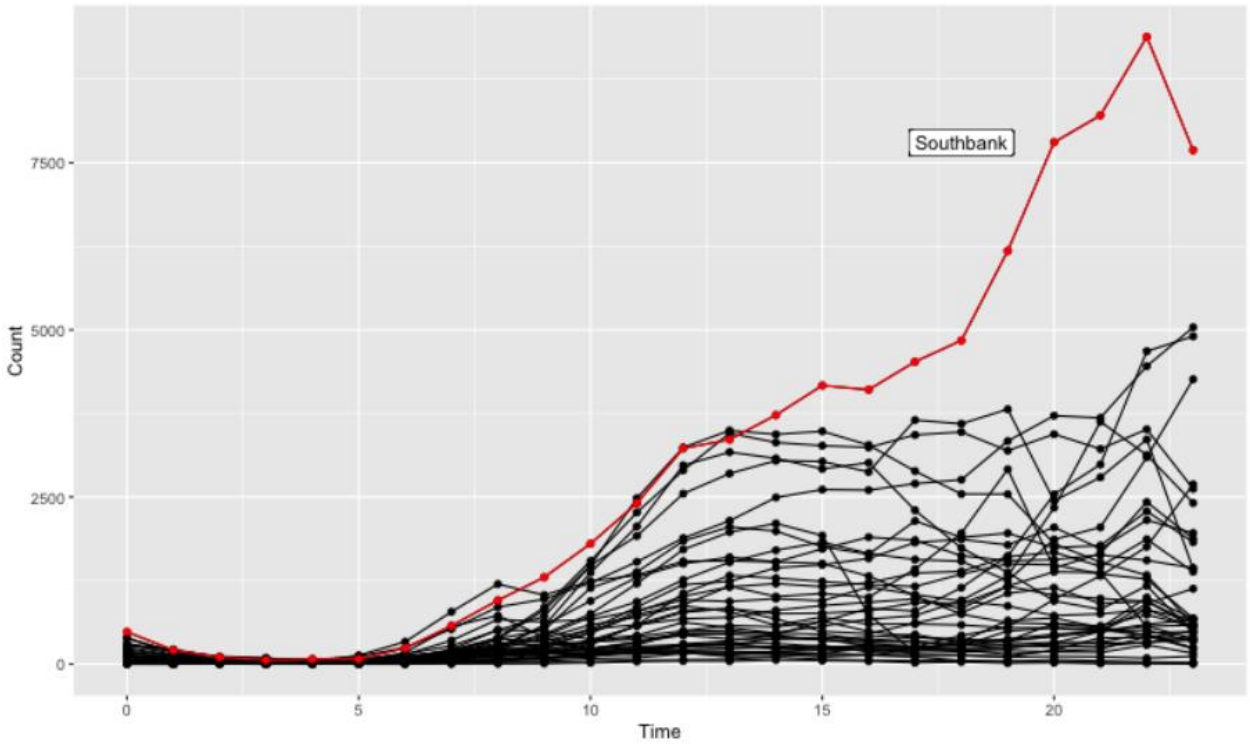


Рисунок 4.5 – Графік багатовимірного часового ряду щогодини підрахунку пішоходів, виміряного за допомогою 43 різних датчиків у місті Мельбурн 31 грудня 2018 року

Після виявлення аномальних діаграм розсіювання або днів з аномальною поведінкою пішоходів для кожного дня проводиться подальше дослідження, щоб виявити місця з аномальною поведінкою протягом вибраного дня. Якщо зосередитися на одному дні, можна отримати колекцію з 43 часових рядів із погодинною кількістю пішоходів, згенерованих 43 датчиками, розташованими в різних географічних точках міста (рисунок 4.5). Для цього аналізу виділяємо сім характеристик часових рядів і перетворюємо оригінальну колекцію часових рядів у багатовимірний простір даних із семи вимірами та 43 екземплярами даних (рисунок 4.6). Тепер кожна точка в цьому багатовимірному просторі відповідає одному часовому ряду (або датчику) на рисунку 4.5. Пропонований алгоритм оголошує одну точку як

аномальну точку в цьому багатовимірному просторі даних. Ця точка відповідає положенням датчиків у Саутбанку в Мельбурні, де новорічні феєрверки щорічно приваблюють мільйони глядачів. Аномальний часовий ряд, виявлений алгоритмом за допомогою функцій часового ряду, позначено червоним кольором.

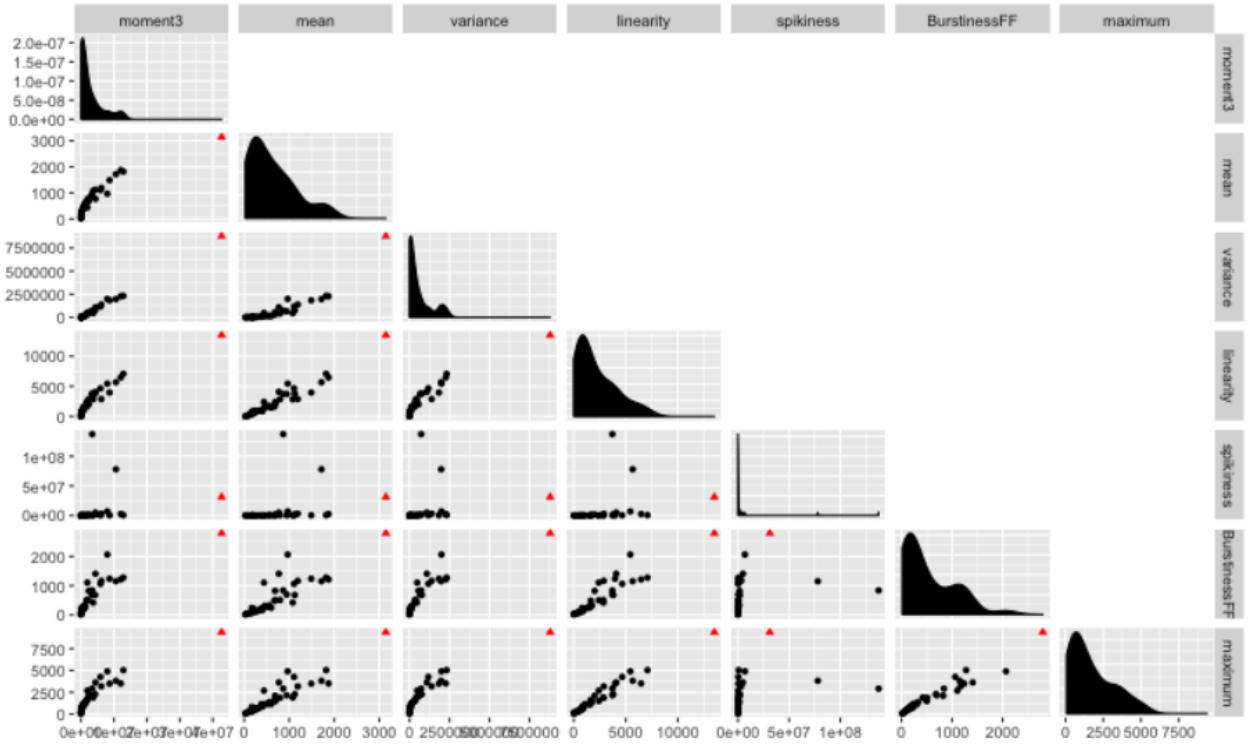


Рисунок 4.6 – Представлення колекції часових рядів на основі ознак на 31 грудня 2018 року

Ці типи висновків відіграють вирішальну роль для прийняття рішень щодо міського планування та управління; визначити можливості покращення пішохідних та транспортних заходів у місті; розуміти вплив великих подій та інших екстремальних умов на діяльність пішоходів і, таким чином, допомагати приймати рішення щодо вимог до безпеки та ресурсів; а також для планування та реагування на надзвичайні ситуації тощо.

4.5 Обробка потокових даних

Завдяки неконтрольованому характеру пропонованого алгоритму його можна легко розширити для потокових даних. Ковзне вікно фіксованої довжини можна використовувати для обробки потокових даних. Тоді набори даних у кожному вікні можна розглядати як пакетний набір даних, і алгоритм може бути застосований до кожного вікна для виявлення аномалій у наборах даних, визначених відповідним вікном.

Алгоритм також можна використовувати для ідентифікації аномальних часових рядів у великій колекції потокових часових даних. Нехай $W[t, t+\omega]$ представляє ковзне вікно, що містить n окремих часових рядів довжиною ω. Спочатку витягуємо m ознак з кожного часового ряду в цьому вікні. Цей крок створює матрицю ознак n×m, де кожен рядок тепер відповідає часовому ряду в оригінальній колекції часових рядів. Після того, як буде перетворено оригінальну колекцію часових рядів у багатовимірний набір даних, можна застосувати пропонований алгоритм для ідентифікації аномальних точок у цьому m-вимірному просторі даних. Потім відповідні часові ряди оголошуються як аномальні ряди у великій колекції часових рядів у відповідному ковзному вікні.

ВИСНОВКИ

В роботі розглянуті методи виявлення аномалій, зокрема в багатовимірних і багатофакторних даних. Обговорено останні дослідження з керуванням проблемами, пов’язаними з багатовимірними та багатофакторними даними.

Алгоритм HDoutliers – це потужний алгоритм для виявлення аномалій у даних великої розмірності. Однак він страждає від кількох обмежень, які значно перешкоджають його здатності виявляти аномалії в певних ситуаціях. У цьому дослідженні запропоновано вдосконалений алгоритм, який усуває ці обмеження. Аномалія визначається тут як спостереження, яке помітно відрізняється від більшості з великим розривом у відстані. Продемонстровано, як пропонований алгоритм може допомогти у виявленні аномалій, присутніх в інших структурах даних за допомогою розробки функцій. На додаток до мітки, алгоритм також призначає аномальний бал кожному екземпляру даних для того, щоб вказати ступінь відмінності кожного вимірювання.

Незважаючи на те, що алгоритм HDoutliers є потужним, в роботі надано кілька класів контрприкладів, де структурні властивості даних не дозволяли HDoutliers виявити певні типи викидів. На цих контрприкладах було продемонстровано, що пропонований алгоритм перевершує HDoutliers як з точки зору точності, так і часу обчислення. Звичайною практикою є оцінка міцності алгоритму за допомогою наборів тестових завдань з різними складними властивостями. Однак слід визнати, що ці контрприклади недостатньо різноманітні та складні, щоб дозволити прокоментувати унікальні сильні та слабкі сторони цих двох алгоритмів, ані узагальнити висновки та вважати, щоб пропонований алгоритм завжди буде кращим.

Це дослідження слід розглядати як спробу змоделювати подальше дослідження алгоритму HDoutliers та його наступників з кінцевою метою

досягти подальших покращень у всьому проблемному просторі, визначеному різними масивами даних великої розмірності. Отже, важливою відкритою дослідницькою проблемою є оцінка ефективності цих алгоритмів у найширшому можливому просторі проблем, визначеному різними наборами даних з різними властивостями. Постає задача – дослідити вплив інших класів проблем з різними структурними властивостями на продуктивність пропонованого алгоритму та де можуть бути його слабкі місця. Цей вид аналізу простору екземплярів забезпечить подальше розуміння покращеного дизайну алгоритму.

Проблеми виявлення аномалій зазвичай виникають у багатьох програмах у різних доменах програм. Таким чином, є надія, що різні люди з різним рівнем знань будуть використовувати пропонований алгоритм для багатьох різних цілей. Тому очікується, що в майбутніх дослідженнях будуть розроблені інструменти інтерактивної візуалізації даних, які дозволять досліджувати аномалії за допомогою комбінації графічних і чисельних методів.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. E. Uzabaci, I. Ercan, and O. Alpu, “Evaluation of outlier detection method performance in symmetric multivariate distributions,” Communications in Statistics-Simulation and Computation, vol. 49, no. 2, pp. 516-531, 2020.

2. R. A. Johnson, D. W. Wichern, and others, Applied multivariate statistical analysis, vol. 6. Pearson London, UK:, 2014.

3. S. Thudumu, P. Branch, J. Jin, and J. J. Singh, “A comprehensive survey of anomaly detection techniques for high dimensional big data,” J Big Data, vol. 7, no. 1, pp. 1-30, 2020.

4. H. Liu, X. Li, J. Li, and S. Zhang, “Efficient Outlier Detection for High-Dimensional Data,” IEEE Trans Syst Man Cybern Syst, vol. 48, no. 12, pp. 2451-2461, Dec. 2018, doi: 10.1109/TSMC.2017.2718220.

5. V. S. L’vov, A. Pomyalov, and I. Procaccia, “Outliers, extreme events, and multiscaling,” Phys Rev E, vol. 63, no. 5, p. 56118, 2001.

6. X. Xu, H. Liu, and M. Yao, “Recent progress of anomaly detection,” Complexity, 2019.

7. K. Malik, H. Sadawarti, and K. G S, “Comparative analysis of outlier detection techniques,” in IJCA, 2014, vol. 97, no. 8, pp. 12-21.

8. D. Ghosh and A. Vogt, “Outliers: An evaluation of methodologies,” in Joint statistical meetings, 2012, vol. 2012.

9. P. J. Rousseeuw and M. Hubert, “Anomaly detection by robust statistics,” Wiley Interdiscip Rev Data Min Knowl Discov, vol. 8, no. 2, p. e1236, 2018.

10. S. Ayesha, M. K. Hanif, and R. Talib, “Overview and comparative study of dimensionality reduction techniques for high dimensional data,” Information Fusion, vol. 59, pp. 44-58, 2020.

11. J. M. Kim and C. S. Park, “Elimination of multidimensional outliers for a compression chiller using a support vector data description,” Sci Technol Built

Environ, vol. 27, no. 5, pp. 578-591, 2021.

12. G. Horvath, E. Kovacs, R. Molontay, and S. Novaczki, “Copula-based anomaly scoring and localization for large-scale, high-dimensional continuous data,” ACM Transactions on Intelligent Systems and Technology (TIST), vol. 11, no. 3, pp. 1-26, 2020.

13. S. Kandanaarachchi and R. J. Hyndman, “Dimension reduction for outlier detection using DOBIN,” Journal of Computational and Graphical Statistics, vol. 30, no. 1, pp. 204-219, 2021.

14. S. Suboh and I. A. Aziz, “Anomaly Detection with Machine Learning in the Presence of Extreme Value-A Review Paper,” in 2020 IEEE Conference on Big Data and Analytics (ICBDA), 2020, pp. 66-72.

15. X. Chen, B. Zhang, T. Wang, A. Bonni, and G. Zhao, “Robust principal component analysis for accurate outlier sample detection in RNA-Seq data,” BMC Bioinformatics, vol. 21, no. 1, pp. 1-20, 2020.

16. R. Foorthuis, “On the nature and types of anomalies: a review of deviations in data,” Int J Data Sci Anal, vol. 12, no. 4, pp. 297-331, 2021.

17. Abuzaid, A, A Hussin & I Mohamed (2013). Detection of outliers in simple circular regression models using the mean circular error statistic. Journal of Statistical Computation and Simulation 83(2), 269-277.

18. Talagala, PD, RJ Hyndman, C Leigh, K Mengersen & K Smith-Miles (2019b). A feature-based framework for detecting technical outliers in water-quality data from in situ sensors. arXiv preprint arXiv:1902.06351.

19. Wilkinson, L, A Anand & R Grossman (2005). Graph-theoretic scagnostics. In: IEEE Symposium on Information Visualization, 2005. INFOVIS 2005. IEEE, pp.157-164.

20. Liu, S, D Maljovec, B Wang, PT Bremer & V Pascucci (2016). Visualizing high-dimensional data: Advances in the past decade. IEEE Transactions on Visualization and Computer Graphics 23(3), 1249-1268.

21. Wilkinson, L (2017). Visualizing big data outliers through distributed aggregation. IEEE Transac¬tions on Visualization and Computer Graphics 24(1),

256-266.

22. Madsen, JH (2018). DDoutlier: Distance and Density-Based Outlier Detection. R package version 0.1.0. https://CRAN.R-project.org/package=DDoutlier.

23. Jouan-Rimbaud, D, E Bouveresse, D Massart & O De Noord (1999). Detection of prediction outliers and inliers in multivariate calibration. Analytica Chimica Acta 388(3), 283-301.

24. Williams, KT (2016). “Local parametric density-based outlier detection and ensemble learning with applications to malware detection”. PhD thesis. The University of Texas at San Antonio.

25. Burridge, P & AMR Taylor (2006). Additive outlier detection via extreme-value theory. Journal of Time Series Analysis 27(5), 685-701.

26. Hyndman, RJ (1996). Computing and graphing highest density regions. The American Statistician 50(2), 120-126.

27. Clifton, DA, S Hugueny & L Tarassenko (2011). Novelty detection with multivariate extreme value statistics. Journal of Signal Processing Systems 65(3), 371-389.

28. O. O. Aremu, R. A. Cody, D. Hyland-Wood, and P. R. McAree, “A relative entropy based feature selection framework for asset data in predictive maintenance,” Comput Ind Eng, vol. 145, p. 106536, 2020.

29. Y. Oner and H. Bulut, “A robust EM clustering approach: ROBEM,” Communications in Statistics-Theory and Methods, vol. 50, no. 19, pp. 4587-4605, 2021.

30. P. Navarro-Esteban and J. A. Cuesta-Albertos, “High-dimensional outlier detection using random projections,” TEST, pp. 1-27, 2021.

31. S. Anitha and M. Metilda, “An efficient and robust cluster based outlying points detection in multivariate data sets,” Internation a l Journal of Engineering & Technology, vol. 7, no. 4, pp. 2881-2885, 2018.

32. M. A. Hayes and M. A. M. Capretz, “Contextual anomaly detection framework for big sensor data,” J Big Data, vol. 2, no. 1, p. 2, 2015.

33. T. Fujiwara, N. Sakamoto, J. Nonaka, K. Yamamoto, K.-L. Ma, and others, “A visual analytics framework for reviewing multivariate time¬series data with dimensionality reduction,” IEEE Trans Vis Comput Graph, vol. 27, no. 2, pp. 1601-1611, 2020.

34. Hartigan, JA & J Hartigan (1975). Clustering Algorithms. Vol. 209. Wiley New York.

35. Unwin, A (2019). Multivariate outliers and the O3 Plot. Journal of Computational and Graphical Statistics, 1-11.

36. Fraley, C (2018). HDoutliers: Leland Wilkinson’s Algorithm for Detecting Multidimensional Outliers. R package version 1.0. https://CRAN.R-project.org/package=HDoutliers.

37. Elseberg, J, S Magnenat, R Siegwart & A Nuchter (2012b). Comparison of nearest-neighbor- search strategies and implementations for efficient shape registration. Journal of Software Engineering for Robotics 3(1), 2-12.

38. Embrechts, P, C Kluppelberg & T Mikosch (2013). Modelling Extremal Events: for Insurance and Finance. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg.

39. Beygelzimer, A, S Kakadet, J Langford, S Arya, D Mount & S Li (2019). FNN: Fast Nearest Neighbor Search Algorithms and Applications. R package version 1.1.3. https : // CRAN . R - project.org/package=FNN.

40. City of Melbourne (2019). Pedestrian Volume in Melbourne. Last accessed 2019-07-23. http : //www.pedestrian.melbourne.vic.gov.au.

41. Білоконь А. С., Борисов С. О., Усатенко М.В., Федорченко В. М. Аналіз функціонування розподілених систем обробки та зберігання даних // «Системи управління навігації та зв’язку», – Випуск 7 (77), – Полтава, Національний університет “Полтавська політехніка імені Юрія Кондратюка”, – 2024. – С. 47-51.