ДОДАТОК А
ГРАФІЧНИЙ МАТЕРІАЛ КВАЛІФІКАЦІЙНОЇ РОБОТИ

Харківський національний університет радіоелектроніки
Кафедра ЕОМ

МЕТОДИ ВИЯВЛЕННЯ АНОМАЛІЙ У МАСИВАХ БАГАТОВИМІРНИХ ДАНИХ

КВАЛІФІКАЦІЙНА РОБОТА
ДРУГИЙ (МАГІСТЕРСЬКИЙ) РІВЕНЬ

Автор:
Усатенко М.В.,
студ. гр. СПм-22-6

Керівник:
Коваленко А.В.,
зав. каф. ЕОМ

МЕТА І ЗАДАЧІ РОБОТИ

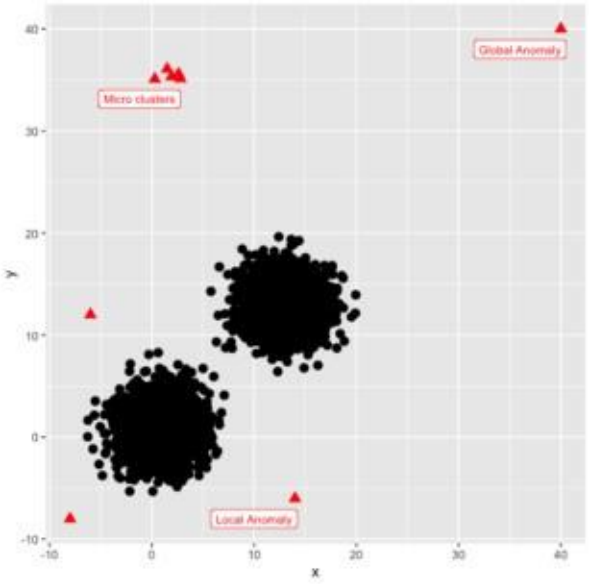
Мета кваліфікаційної роботи:

- дослідження методів виявлення аномалій у масивах багатовимірних даних.

Задачі:

- проаналізувати проблему виявлення аномалій в даних;
- розглянути існуючі методи виявлення аномалій в багатовимірних даних;
- запропонувати ефективний метод виявлення аномалій для масивів як одновимірних, так і багатовимірних, даних;
- провести експериментальне дослідження запропонованого методу.

РІЗНІ ТИПИ АНОМАЛІЙ У БАГАТОВИМІРНИХ ДАНИХ



3

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 побудови моделі не потребує використання навчальних наборів даних	необхідно провести оптимізацію за найкращим значенням K
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

4

МОДЕЛЬ ВИЯВЛЕННЯ АНОМАЛІЙ

5

ФАЗА ВИЯВЛЕННЯ ЗАГАЛЬНОЇ АНОМАЛІЇ

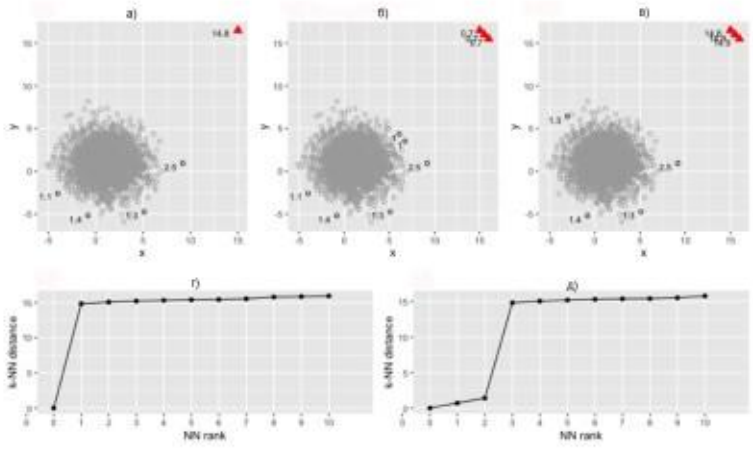
6

ПОСЛІДОВНІСТЬ ЕТАПІВ АЛГОРИТМУ

- Етап 1. Прийом вхідних даних.
- Етап 2. Нормалізація стовпців.
- Етап 3. Пошук найближчого сусіда.
- Етап 4. Розрахунок порогу.
- Етап 5. Видача результату.

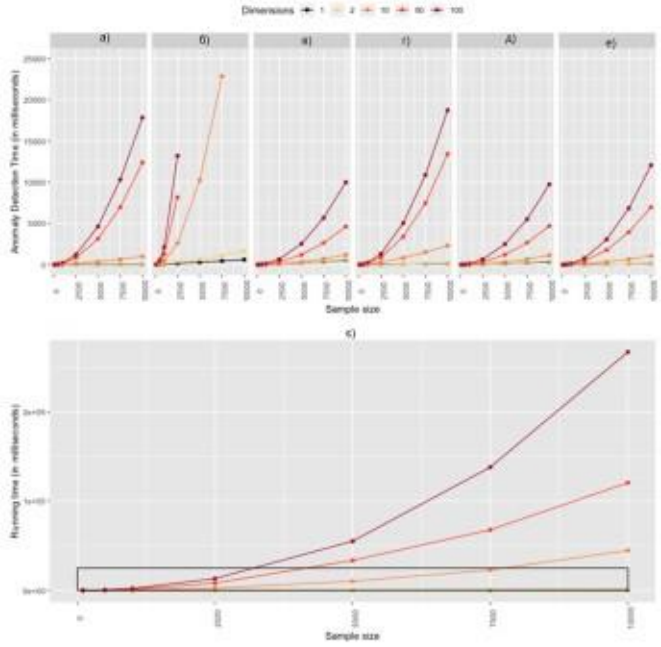
7

РІЗНИЦЯ МІЖ ВІДСТАНННЮ НАЙБЛИЖЧОГО СУСІДА ТА K-ВІДСТАНННЮ НАЙБЛИЖЧОГО СУСІДА З МАКСИМАЛЬНИМ РОЗРИВОМ



8

ПОРІВНЯННЯ ПРОДУКТИВНОСТІ



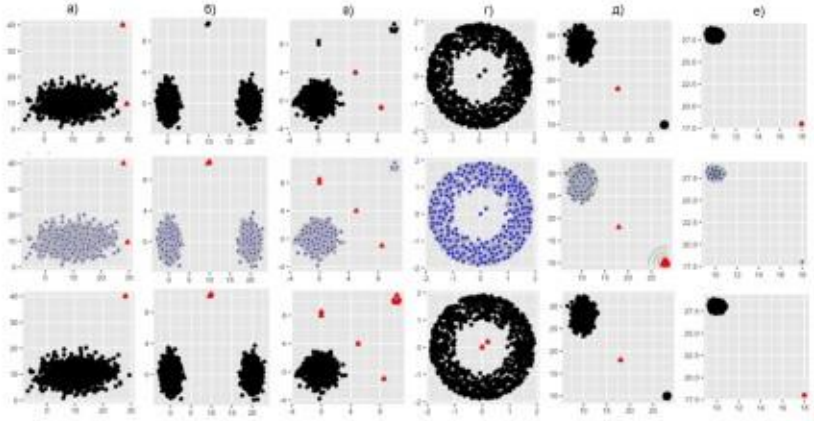
9

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

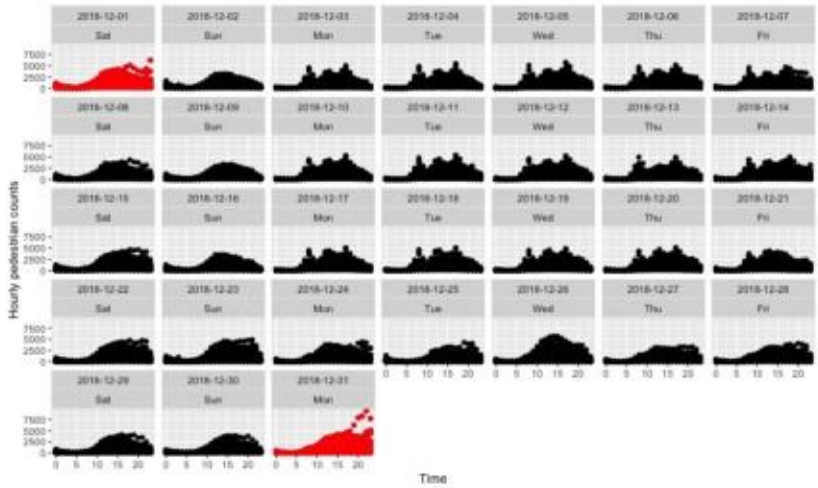
10

ЯКІСТЬ АЛГОРИТМУ



11

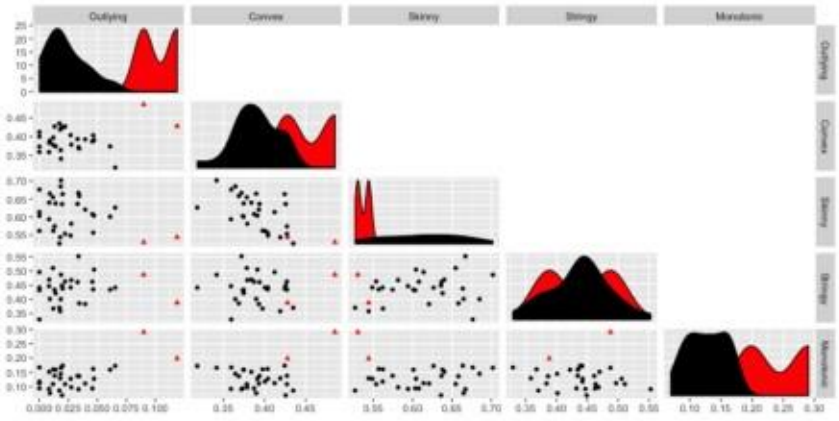
ЕКСПЕРИМЕНТАЛЬНА ПЕРЕВІРКА



Підрахунок пішоходів у 43 місцях у місті Мельбурн

12

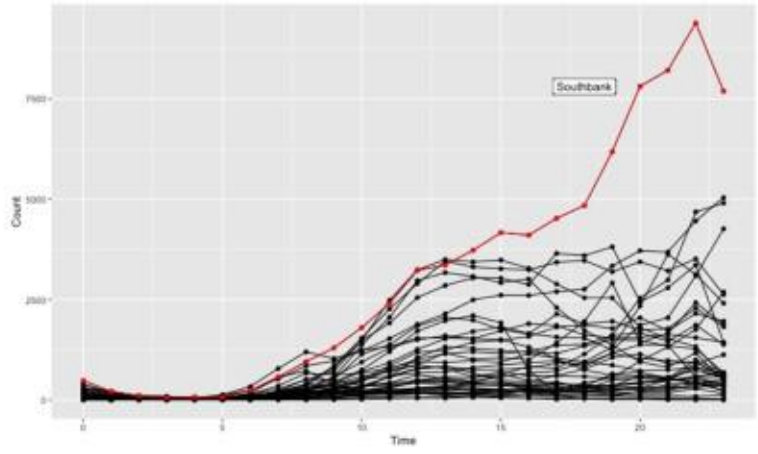
ЕКСПЕРИМЕНТАЛЬНА ПЕРЕВІРКА



Представлення колекції діаграм розсіювання за допомогою Scagnostics

13

ЕКСПЕРИМЕНТАЛЬНА ПЕРЕВІРКА



Графік багатовимірного часового ряду щогодинного підрахунку пішоходів

14

ВИСНОВКИ

- Розглянуті методи виявлення аномалій, зокрема в багатовимірних і багатофакторних даних. Проаналізовано останні дослідження з керуванням проблемами, пов'язаними з даними подібного виду.
- Важливою відкритою дослідницькою проблемою є оцінка ефективності цих алгоритмів у найширшому можливому просторі проблем, визначеному різними наборами даних з різними властивостями.
- 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 цій роботі запропоновано вдосконалений алгоритм, який усуває вказані обмеження.
- В роботі надано кілька класів прикладів, де структурні властивості даних не дозволяли HDoutliers виявити певні типи викидів, а 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 приклади 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.

15

ВИСНОВКИ

- 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.
- Цю роботу 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.
- 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.
- Оскільки проблеми виявлення аномалій зазвичай виникають у багатьох випадках, є надія, що пропонований алгоритм будуть використовувати для багатьох різних цілей різні користувачі з різним рівнем знань. Тому очікується, що в майбутніх дослідженнях будуть розроблені інструменти інтерактивної візуалізації даних, які дозволять досліджувати аномалії за допомогою комбінації графічних і чисельних методів.

16