ДОДАТОК А

Графічний матеріал кваліфікаційної роботи

Міністерство освіти та науки України
Харківський національний університет радіоелектроніки
Факультет комп'ютерної інженерії та управління
Кафедра ЕОМ

Магістерська кваліфікаційна робота

Методи виявлення аномального трафіку в IoT

Виконав:
ст.гр. СПм-22-3
Марченко Р.М.

Керівник:
зав.каф. Коваленко А.А.

Мета роботи та завдання

Об'єктом дослідження є процес виявлення аномального трафіку в IoT.
Предметом дослідження є методи виявлення аномального трафіку в IoT.
Метою кваліфікаційної роботи є підвищення точності виявлення аномального трафіку в IoT за рахунок підбору гіперпараметів для моделей машинного навчання.

Завдання:

- провести аналіз архітектури та протоколів в IoT;
- провести аналіз класифікації методів виявлення аномалій в IoT;
- обрати групу методів, що доцільно використовувати для виявлення аномального трафіку в IoT;
- реалізувати та провести навчання моделей машинного навчання для виявлення аномального трафіку в IoT;
- підібрати гіперпараметри моделей для покращенння точності виявлення аномалій;
- провести аналіз отриманих результатів.

Поняття IoT

Інтернет речей(IoT, Internet of Things) — це концепція об'єднання фізичних пристроїв, які мають вбудовану електроніку, програмне забезпечення, датчики та підключення до мережі для маніпуляції даними.

Застосування IoT:
- розумні будинки;
- розумні міста;
- промисловість;
- охорона здоров'я;
- логістика та транспорт.

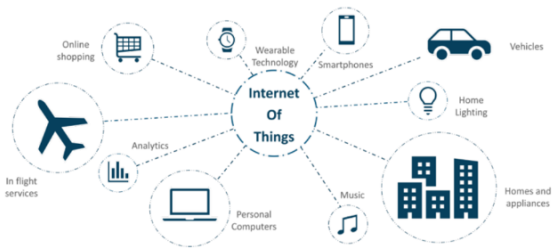


Рисунок 1. - Сфери використання IoT

3

Архітектура та протоколи IoT

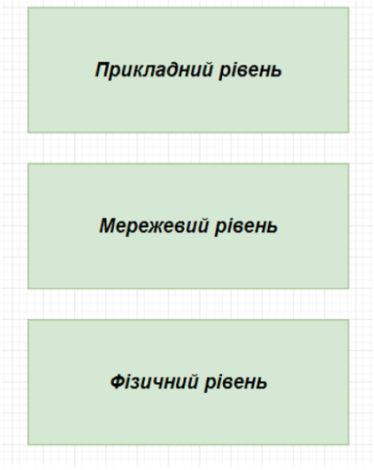


Рисунок 1 - Трирівнева архітектура

Рисунок 2 - П’ятирівнева архітектура

Рисунок 3 - Протоколи IoT

4

Класифікація методів виявлення аномального трафіку

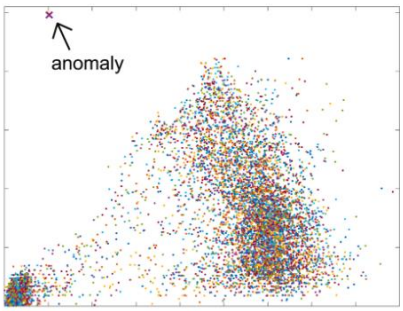


Рисунок 1 - Діаграма розсіювання з прикладом аномалії

Таблиця 1 – Основні переваги та недоліки методів виявлення аномалій

Категорія методу	Переваги	Недоліки
	За методом	
Геометричні методи	Добре підходять для даних з чітко визначеними структурами.	Можуть бути неефективними для даних із складними структурами або часово залежними даними.
Статистичні методи	Можуть моделювати різноманітні розподіли даних.	Вимагають чіткого розуміння розподілу даних, що моделюються, і можуть бути неефективними для даних зі складними структурами або змінними з часом.
Методи машинного навчання та глибокого навчання	Можуть виявляти складні аномалії та залежності між даними.	Вимагають великої кількості даних для тренування. Можуть бути складними для налаштування та оптимізації.
	За застосуванням	
Конструктивні застосування	Надають користь та виршують практичні завдання.	Вимагають розробки специфічних застосунків для кожного випадку.
Деструктивні застосування	Допомагають виявляти та запобігати шкідливим діям та атакам.	Зазвичай потребують додаткових заходів для захисту системи. Можуть призводити до фальсифікації або неправильного реагування.
Застосування для очищення даних	Допомагають видалити непотрібні дані та шум з даних.	Можуть втрачати корисну інформацію. Вимагають заздалегідь відомих шаблонів для очищення.
	За типом аномалії	
Пунктові аномалії	Відокремлюють аномалії, які виникають в окремих точках даних.	Можуть пропустити аномалії, які виникають лише в контексті.
Контекстуальні аномалії	Враховують контекст та поведінкові характеристики для виявлення аномалій.	Вимагають складніших аналітичних методів та більше обчислювальних ресурсів.
Колективні аномалії	Визначають аномалії на основі всього набору даних та структури взаємозв'язків між даними.	Можуть бути обчислювально витратними та вимагати великої кількості даних для навчання.
	За затримкою	
Online алгоритми	Здатні обробляти дані під час їх збору та аналізувати їх в реальному часі.	Можуть бути обмеженими за ресурсами та вимагати низької затримки.
Offline алгоритми	Мають доступ до всього набору даних і можуть використовувати більш складні обчислювальні методи.	Зазвичай вимагають більше обчислювальних ресурсів та можуть бути повільнішими в роботі.

5

Метод дерева прийняття рішень

- збір даних з IoT пристроїв;
- підготовка даних;
- маркування даних;
- побудова дерева прийняття рішень;
- навчання моделі дерева прийняття рішень;
- оцінка моделі дерева прийняття рішень;
- виявлення аномалій;
- аналіз результатів.

$$G(S,A) = E(S) - E(A), \quad (1)$$

де S – множина даних;
E – ентропія;
A – атрибут, за яким розділяється множина даних.

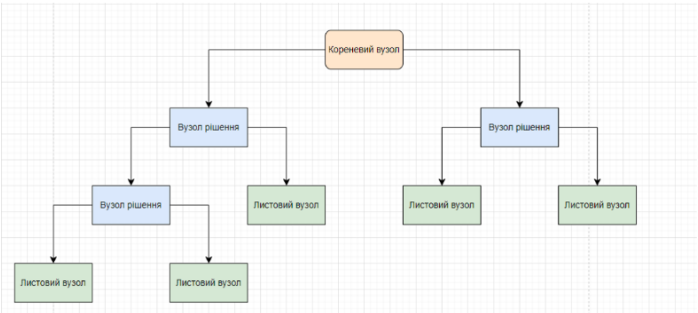


Рисунок 1 – Структура дерева прийняття рішень

6

Метод випадкового лісу

- збір даних;
- маркування даних;
- навчання моделі випадкового лісу;
- оцінка моделі випадкового лісу;
- виявлення аномалій;
- аналіз результатів.

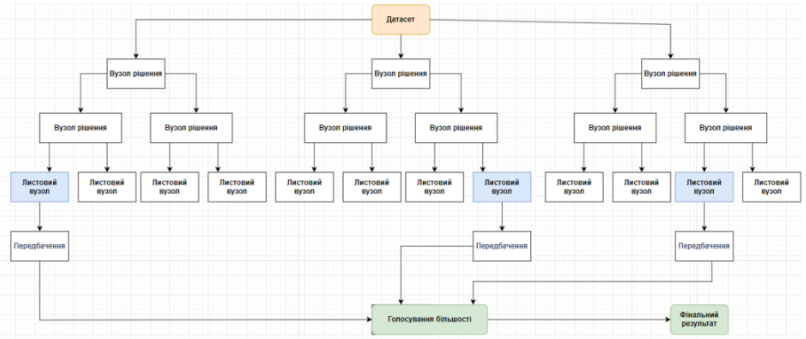


Рисунок 1. - Схема створення випадкового лісу

Функція передбачення для випадкового лісу визначається:

$$F(x) = \frac{1}{J}\sum_{j=1}^{j} f_j(x), \qquad 2)$$

де J – кількість дерев у лісі.

7

Метод опорних векторів

- збір даних;
- підготовка даних;
- маркування даних;
- вибір ядра;
- навчання моделі опорних векторів;
- оцінка моделі опорних;
- виявлення аномалій;
- аналіз результатів.

Гіперплощина визначається рівнянням

$$\vec{w} \times \vec{x} - b = 0, \qquad (3)$$

де w – вектор нормалі до гіперплощини;
x – вектор ознак;
b – скалярний зсув.

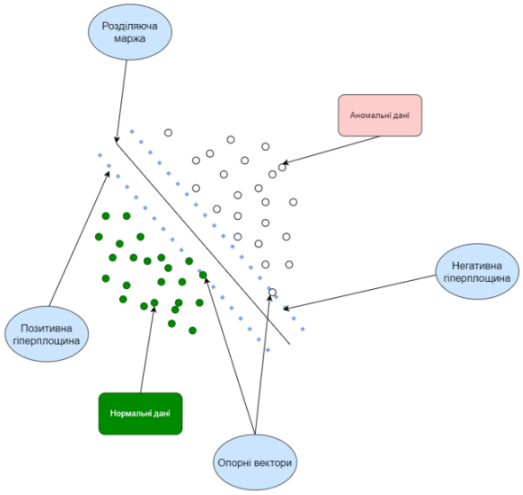


Рисунок 1 - Компоненти методу опорних векторів

8

Проведення попередньої обробки даних

Таблиця 1 - Короткий опис шкідливих сценаріїв з датасету

#	Name of Dataset	Duration (hrs)	#Packets	#ZeekFlows	Pcap Size	Name
1	CTU-IoT-Malware-Capture-34-1	24	233,000	23,146	121 MB	Mirai
2	CTU-IoT-Malware-Capture-43-1	1	82,000,000	67,321,810	6 GB	Mirai
3	CTU-IoT-Malware-Capture-44-1	2	1,309,000	238	1.7 GB	Mirai
4	CTU-IoT-Malware-Capture-49-1	8	18,000,000	5,410,562	1.3 GB	Mirai
5	CTU-IoT-Malware-Capture-52-1	24	64,000,000	19,781,379	4.6 GB	Mirai
6	CTU-IoT-Malware-Capture-20-1	24	50,000	3,210	3.9 MB	Torii
7	CTU-IoT-Malware-Capture-21-1	24	50,000	3,287	3.9 MB	Torii
8	CTU-IoT-Malware-Capture-42-1	8	24,000	4,427	2.8 MB	Trojan
9	CTU-IoT-Malware-Capture-60-1	24	271,000,000	3,581,029	21 GB	Gagfyt
10	CTU-IoT-Malware-Capture-17-1	24	109,000,000	54,659,864	7.8 GB	Kenjiro
11	CTU-IoT-Malware-Capture-36-1	24	13,000,000	13,645,107	992 MB	Okiru
12	CTU-IoT-Malware-Capture-33-1	24	54,000,000	54,454,592	3.9 GB	Kenjiro
13	CTU-IoT-Malware-Capture-8-1	24	23,000	10,404	2.1 MB	Hakai
14	CTU-IoT-Malware-Capture-35-1	24	46,000,000	10,447,796	3.6G	Mirai
15	CTU-IoT-Malware-Capture-48-1	24	13,000,000	3,394,347	1.2G	Mirai
16	CTU-IoT-Malware-Capture-39-1	7	73,000,000	73,568,982	5.3GB	IRCBot
17	CTU-IoT-Malware-Capture-7-1	24	11,000,000	11,454,723	897 MB	Linux,Mirai
18	CTU-IoT-Malware-Capture-9-1	24	6,437,000	6,378,294	472 MB	Linux.Hajime
19	CTU-IoT-Malware-Capture-3-1	36	496,000	156,104	56 MB	Muhstik
20	CTU-IoT-Malware-Capture-1-1	112	1,686,000	1,008,749	140 MB	Hide and Seek

Приклади значень з датасету:
- tf – мітка часу захоплення;
- uid – ідентифікатор захоплення;
- id_orig.h – IP-адреса джерела атаки;
- id_orig.p – порт джерела атаки;
- id_resp.h – IP пристрою IoT;
- id_resp.p – порт пристрою IoT;
- proto – протокол транспортного рівня з'єднання;
- duration – кількість часу обміну даними між пристроєм IoT і зловмисником;
- orig_bytes – кількість даних, надісланих на пристрій IoT;
- resp_bytes – кількість даних, надісланих пристроєм IoT;
- label – тип захоплення, доброякісне чи шкідливе.

9

Навчання та дослідження моделей (1)

	Макросередня точність	Макросередній відгук	Макросередня F1-міра
Модель випадкового лісу	99.98%	99.99%	99.98%
Модель опорних векторів	97.50%	97.48%	97.47%
Модель дерева рішення	99.97%	99.97%	99.97%

Рисунок 1 - Макросередні показники для кожної моделі

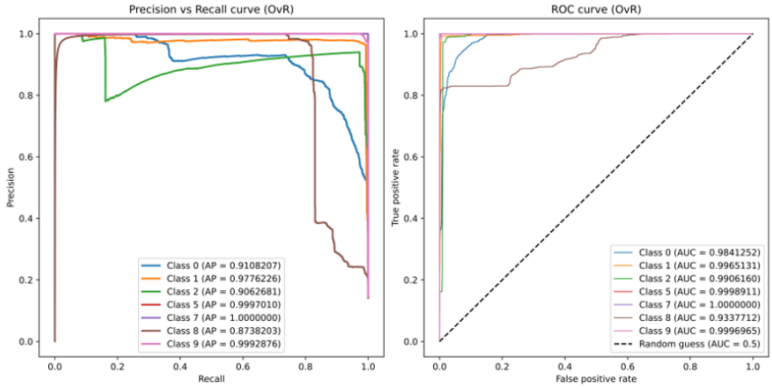


Рисунок 2 - Криві точності та відгуку для моделі опорних векторів

10

Навчання та дослідження моделей (2)

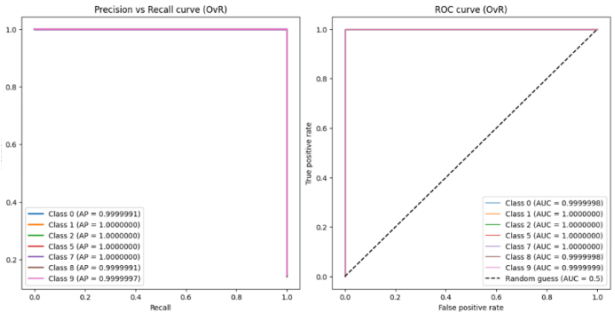


Рисунок 1 - Криві точності та відгуку для моделі випадкового лісу

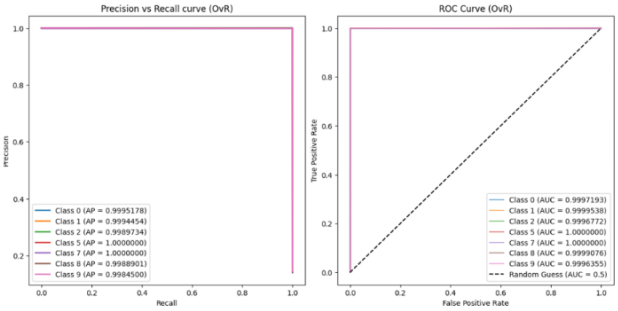


Рисунок 2 - Криві точності та відгуку для моделі дерева рішення

11

Навчання та дослідження моделей (3)

Рисунок 1 - Конфузійні матриці для реалізованих моделей

12

Висновки

У ході виконання кваліфікаційної роботи були досліджені методи виявлення аномального трафіку IoT, такі як: метод випадкового лісу, метод опорних векторів та метод дерева прийняття рішень.

У результаті проведеного дослідження було здійснено підбір оптимальних гіперпараметрів для реалізованих моделей для їх кращої продуктивність при виявленні аномального трафіку в IoT.

Для вибору оптимальних гіперпараметрів був використаний метод перехресної валідації, що дозволяє підібрати різні комбінації гіперпараметрів та вибирати ті, які забезпечують найкращу продуктивність моделі.

Для методу випадкового лісу було підібрано оптимальну кількість дерев та їх максимальну глибину, що дозволило збільшити точність виявлення аномалій за рахунок зменшення ймовірності перенавчання.

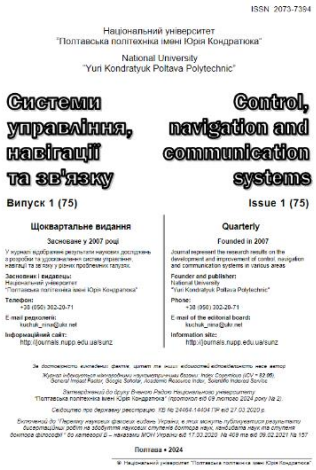
Для методу опорних векторів були підібрані значення параметра регуляризації C та тип ядра з урахуванням необхідності балансування між похибками першого та другого роду.

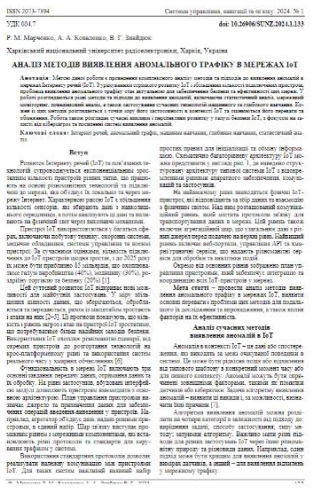
Для методу дерева прийняття рішень були підібрані глибина дерева та мінімальна кількість зразків для розщеплення з метою забезпечення стабільності моделі при обробці великої кількості даних IoT.

Результати дослідження показують, що метод випадкового лісу є найбільш перспективним методом для виявлення аномального трафіку в IoT.

13

Апробація результатів кваліфікаційної роботи





Посилання на публікацію: https://journals.nupp.edu.ua/sunz/issue/view/115/63

14