УДК 004.056.55:004.77:004.93

Мєшков В.І., аспірант спеціальності 122 Комп'ютерні науки
(Національний технічний університет «Дніпровська політехніка», м. Дніпро, Україна)

АНАЛІЗ КОРЕЛЯЦІЙНОЇ МАТРИЦІ ДЛЯ ПОКАЗНИКІВ НАБОРУ ДАНИХ CSE-CIC-IDS2017

В сучасних умовах швидкого розвитку інформаційних технологій та мережевих систем питання забезпечення кібербезпеки стає надзвичайно актуальним. Аналіз мережевих даних є одним з основних підходів до виявлення аномалій та попередження кібератак. Набір даних CSE-CIC-IDS2017 є одним з найпоширеніших стандартів для досліджень у цій сфері, оскільки містить великий обсяг мережевих показників, що дозволяє моделювати та виявляти різноманітні види атак.

У межах цього дослідження було проаналізовано різноманітні показники CSE-CIC-IDS2017, які надають детальну інформацію про мережевий трафік і дозволяють виявляти аномальні патерни. Основні категорії показників включають:

1. Сеансові та часові характеристики (тривалість потоку, середні та максимальні інтервали між пакетами) – дозволяють аналізувати тривалість та інтенсивність трафіку.
2. Розмір і швидкість передачі пакетів (кількість переданих пакетів і байтів, середній розмір та швидкість) – допомагають оцінити обсяги та варіативність переданих даних.
3. TCP-прапори (FIN, SYN, RST, PSH, ACK тощо) – використовуються для ідентифікації нестандартної поведінки, наприклад, сканування портів.
4. Розподіл за напрямками передачі (пакети і байти в обох напрямках) – дозволяє оцінити збалансованість трафіку.
5. Активність і бездіяльність – середній та максимальний час активного та неактивного станів сеансу, що допомагає ідентифікувати патерни поведінки пристроїв.
6. Інші технічні показники – початковий розмір вікна TCP, мінімальний розмір сегмента, та інші параметри, які доповнюють аналіз трафіку.

Перелік показників: Destination Port, Flow Duration, Total Fwd Packets, Total Backward Packets, Total Length of Fwd Packets, Total Length of Bwd Packets, Fwd Packet Length Max, Fwd Packet Length Min, Fwd Packet Length Mean, Fwd Packet Length Std, Bwd Packet Length Max, Bwd Packet Length Min, Bwd Packet Length Mean, Bwd Packet Length Std, Flow Bytes/s, Flow Packets/s, Flow IAT Mean, Flow IAT Std, Flow IAT Max, Flow IAT Min, Fwd IAT Total, Fwd IAT Mean, Fwd IAT Std, Fwd IAT Max, Fwd IAT Min, Bwd IAT Total, Bwd IAT Mean, Bwd IAT Std, Bwd IAT Max, Bwd IAT Min, Fwd PSH Flags, Bwd PSH Flags, Fwd URG Flags, Bwd URG Flags, Fwd Header Length, Bwd Header Length, Fwd Packets/s, Bwd Packets/s, Min Packet Length, Max Packet Length, Packet Length Mean, Packet Length Std, Packet Length Variance, FIN Flag Count, SYN Flag Count, RST Flag Count, PSH Flag Count, ACK Flag Count, URG Flag Count, CWE Flag Count, ECE Flag Count, Down/Up Ratio, Average Packet Size, Avg Fwd Segment Size, Avg Bwd Segment Size, Fwd Header Length.1, Fwd Avg Bytes/Bulk, Fwd Avg Packets/Bulk, Fwd Avg Bulk Rate, Bwd Avg Bytes/Bulk, Bwd Avg Packets/Bulk, Bwd Avg Bulk Rate, Subflow Fwd Packets, Subflow Fwd Bytes, Subflow Bwd Packets, Subflow Bwd Bytes, Init Win bytes forward, Init Win bytes backward, act_data_pkt_fwd, min_seg_size_forward, Active Mean, Active Std, Active Max, Active Min, Idle Mean, Idle Std, Idle Max, Idle Min, Attack Number [1,2].

Кожен з перелічених показників вносить свій вклад у формування загальної картини мережевого трафіку, що дозволяє детально аналізувати його поведінку та виявляти аномалії, пов'язані з потенційними загрозами. Використання кореляційних матриць та методів візуалізації сприяє виявленню закономірностей між показниками, що допомагає ідентифікувати типові патерни нормального та аномального трафіку, а також підвищує точність і ефективність моделей для автоматизованого виявлення аномалій.

На рисунку 1 представлена кореляційна матриця для аналізованих мережевих показників, яка демонструє рівень взаємозв'язків між ними.

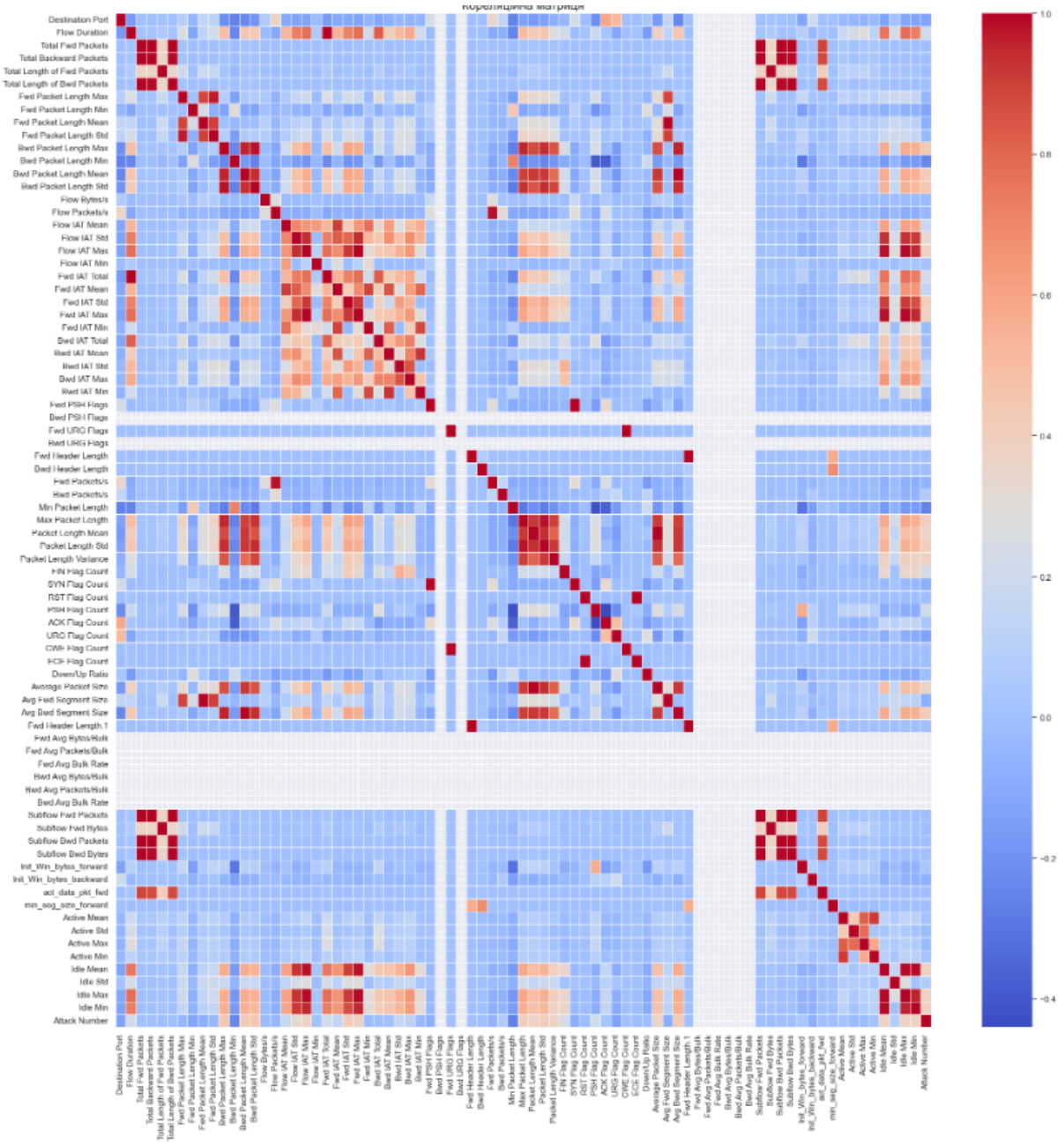

Рисунок 1 – Кореляційна матриці

Аналіз кореляційної матриці для показників набору даних CIC-IDS2017 дозволяє виявити основні взаємозв'язки між різними характеристиками мережевого трафіку, що може бути корисним для виявлення аномалій та вивчення поведінки мережі під час атак. Далі наведено ключові спостереження щодо показників, внесених у кореляційну матрицю:

1. Сеансові показники (Flow Duration, Flow Bytes/s, Flow Packets/s). Тривалість потоку (Flow Duration) має високу кореляцію з кількістю переданих байтів та пакетів, оскільки довший сеанс зазвичай супроводжується більшим обсягом даних. Такі залежності можуть вказувати на легітимні довготривалі з’єднання або на аномальну поведінку при надто великих значеннях. Швидкість передачі байтів та пакетів в одиницю часу (Flow Bytes/s, Flow Packets/s) часто корелює з параметрами, пов'язаними з розміром і частотою передачі пакетів, що може бути корисним для виявлення пікових навантажень або атак типу DoS.

2. Розміри пакетів (Fwd Packet Length Max, Min, Mean, Std; Bwd Packet Length Max, Min, Mean, Std). Показники довжини пакетів (максимальне, мінімальне, середнє значення та стандартне відхилення для пакетів вперед і назад) мають високу кореляцію між собою, що дозволяє оцінити стабільність їх розмірів. Стабільні показники

характерні для легітимного трафіку, тоді як різкі зміни можуть вказувати на аномалії, такі як спроби обману системи захисту.

3. Інтервали між пакетами (Flow IAT Mean, Std, Max, Min; Fwd IAT Mean, Std, Max, Min; Bwd IAT Mean, Std, Max, Min). Інтервали між пакетами, що передаються, особливо в одному напрямку (вперед чи назад), є важливими для оцінки інтенсивності передачі. Високі значення цих показників можуть вказувати на затримки або нестабільність з'єднання, що характерно для певних типів атак або аномальної поведінки. Ці показники часто мають кореляцію з такими показниками, як Flow Duration, що може свідчити про затяжні або спорадичні з'єднання.

4. TCP-прапори (FIN, SYN, RST, PSH, ACK, URG). Кількість різних TCP-прапорів дозволяє ідентифікувати особливі типи трафіку та поведінки, наприклад, спроби сканування портів чи ініціацію нових з'єднань. Показники, пов'язані з SYN та FIN прапорами, часто корелюють із загальною кількістю пакетів, оскільки вони сигналізують про відкриття та завершення з'єднань. Високі значення SYN, PSH або RST прапорів можуть свідчити про аномальні спроби встановлення з'єднань або атаки, такі як сканування портів або DoS-атаки.

5. Сегменти TCP та початковий розмір вікна (Init Win bytes forward/backward, min_seg_size_forward). Початковий розмір вікна TCP та мінімальний розмір сегмента допомагають у виявленні характеристик початкового етапу з'єднання, що є важливим для аналізу надійності з'єднання. Висока кореляція між цими показниками та загальним розміром пакетів може свідчити про певні особливості конфігурації мережевого обладнання чи наявність специфічного трафіку.

6. Активність і бездіяльність (Active Mean, Active Max, Idle Mean, Idle Max). Середній та максимальний час активності й бездіяльності в сесіях допомагає виявляти патерни використання мережі. Часті перерви можуть вказувати на нестабільне або аномальне з'єднання, характерне для певних атак. Наприклад, довга бездіяльність із подальшою активністю може свідчити про очікування зовнішніх тригерів для атаки.

7. Атаки (Attack Number). Показник визначає тип атаки, присутній у даних, що дозволяє аналізувати та порівнювати поведінкові патерни різних видів атак. Кореляція цього показника з іншими мережевими характеристиками допомагає виділити унікальні патерни, властиві конкретним типам атак, таким як DoS, Brute Force, DDoS тощо. Ці виявлені патерни можуть бути використані для навчання нейронної мережі, яка здатна розпізнавати та класифікувати аномалії в реальному часі. Нейронна мережа, навчена на основі таких кореляційних залежностей, зможе більш точно ідентифікувати потенційні загрози, відрізняючи нормальний трафік від підозрілих дій, що відповідають різним типам атак.

На основі проведеного аналізу показників набору даних CIC-IDS2017 та їхніх кореляцій можна зробити висновок, що детальне вивчення взаємозв'язків між мережевими характеристиками дозволяє ідентифікувати специфічні патерни, властиві різним типам атак. Виділення цих патернів є важливим кроком для створення моделей машинного навчання, зокрема нейронних мереж, здатних автоматично розпізнавати аномалії в трафіку. Виявлені залежності між показниками дозволяють підвищити точність і надійність таких моделей, оскільки вони зможуть класифікувати загрози в реальному часі та попереджати про потенційні кібератаки.

Список використаних джерел:

1. Prokopov V. Розробка системи виявлення кіберзагроз на основі аналізу даних з веб-ресурсів на мові програмування python / V. Prokopov, Ye. Meleshko, M. Yakymenko, V. Reznichenko, S. Shymko // Системи управління, навігації та зв'язку. Збірник наукових праць. – Полтава: ПНТУ, 2022. – Т. 2 (68). – С. 79-84. – doi: https://doi.org/10.26906/SUNZ.2022.2.079

2. Лущевський Б. Алгоритми машинного навчання для виявлення та прогнозування атак на мережеву інфраструктуру / Б. Лущевський // Кваліфікаційна робота – Тернопіль, 2023. – http://dspace.wunu.edu.ua/bitstream/316497/50177/1/ВКР_Лущевський Борис.pdf.