УДК 681.12

Зелений В.Є.
Арсенюк І.Р.

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ДЛЯ ВИЯВЛЕННЯ HTTP ЗАПИТІВ З АНОМАЛЬНОЮ ПОВЕДІНКОЮ

Вінницький Національний Технічний Університет

Анотація

Проаналізовано існуючі комерційні загальнодоступні рішення виявлення HTTP запитів з аномальною поведінкою, запропоновано новий формат програмного продукту для виконання задачі аналізу. Проаналізовано способи машинного навчання для виявлення аномалій. Виконано аналіз усіх даних з набору та запропоновано власну модель HTTP запиту для машинного навчання. Розглянуто кореляції між ознаками. Для імплементації обрано модель дерева рішень відштовхуючись від оцінок роботи моделей машинного навчання. Спроєктовано швидкісний мікросервіс, що є носієм інформаційної технології для виявлення HTTP запитів з аномальною поведінкою та усуває недоліки існуючих аналогів, не потребуючи втручання технічного спеціаліста, знань предметної області та тривалого навчання.

Ключові слова: *виявлення аномалій, машинне навчання, дерево рішень, HTTP запит.*

Abstract

Existing commercial publicly available solutions for detecting HTTP requests with anomalous behavior are analyzed, and a new software product format is proposed for performing the analysis task. Methods of machine learning to detect anomalies are analyzed. Analysis of all data from the set was performed and a custom HTTP request model was proposed for machine learning. Correlations between features are considered. For implementation, a decision tree model was chosen based on the performance evaluations of machine learning models. A high-speed microservice is designed, which is a carrier of information technology for detecting HTTP requests with anomalous behavior and eliminates the shortcomings of existing analogues, without requiring the intervention of a technical specialist, knowledge of the subject area and long-term training.

Keywords*: anomaly detection, machine learning, decision tree, HTTP request.*

Вступ

Сучасне життя напряму залежить від інтернету. Його значимість важко переоцінити, адже з ним пов’язане навчання, робота, комунікація, відпочинок. Інтернет дав можливість реалізувати безліч ідей та задумів, починаючи зі звичайного обміну даними на відстані. Він однозначно спростив людське життя та покращив ефективність будь-якої діяльності.

Будь-який веб-додаток, що відіграє серйозну роль, повинен бути захищеним. Що вагоміший вплив веб-додатку на ті чи інші процеси – тим вищі повинні бути вимоги до якості його захисту. Зазвичай, атаки на інтернет-ресурси спрямовані з метою унеможливлення доступу користувачів, але існують атаки, мета яких – зміна або викрадення чутливих даних.

Одним із найпоширеніших методів нападу є насичення атакованого комп'ютера або мережевого устаткування великою кількістю зовнішніх запитів (часто безглуздих або неправильно сформульованих) таким чином атаковане устаткування не може відповісти користувачам, або

відповідає настільки повільно, що стає фактично недоступним. Взагалі відмова сервісу здійснюється:

1) примусом атакованого устаткування до зупинки роботи програмного забезпечення/устаткування або до витрат наявних ресурсів, внаслідок чого устаткування не може продовжувати роботу;
2) заняттям комунікаційних каналів між користувачами і атакованим устаткуванням, внаслідок чого якість сполучення перестає відповідати вимогам.

Якщо атака відбувається одночасно з великої кількості IP-адрес, то її називають розподіленою (англ. Distributed Denial-of-Service — DDoS) [1]. Один із найпоширеніших способів реалізації такої атаки полягає у виконанні HTTP запитів.

Відомо, що для детектування комп'ютерних атак часто використовуютьаналіз ряду ознак комп'ютерного трафіку. Для підвищення ефективності такого детектування дуже важливо не лише забезпечити відбір з доступної множини ознак найінформативніших, а й визначити таке їх сполучення, яке дасть змогу найточніше, найповніше та найшвидше здійснювати детектування, вказуючи наявність та прогнозований вид атаки. Таким чином, важливість відбору ознак для задачі детектування комп'ютерних атак не викликає сумнівів [2].

Результати дослідження

Проведено аналіз предметної області, розглянуто та проаналізовано аналоги, виявлено їх недоліки та переваги. Відштовхуючись від недоліків, поставлено задачу для створюваної інформаційної технології, серед переваг якої буде відсутність необхідності втручання технічного спеціаліста, знань з предметної області та тривалого навчання.

Базуючись на висновках теореми про відсутність безкоштовних сніданків [3], задля дотримання високоточних результатів, досліджено та виконано практичну перевірку сучасних моделей машинного навчання, задача яких пов'язана з класифікацією об'єктів.

Використовуючи набір даних NF-UNSW-NB15 [4], проаналізовано кореляцію між змінними:

Рисунок 1 – Кореляція між змінними набору даних

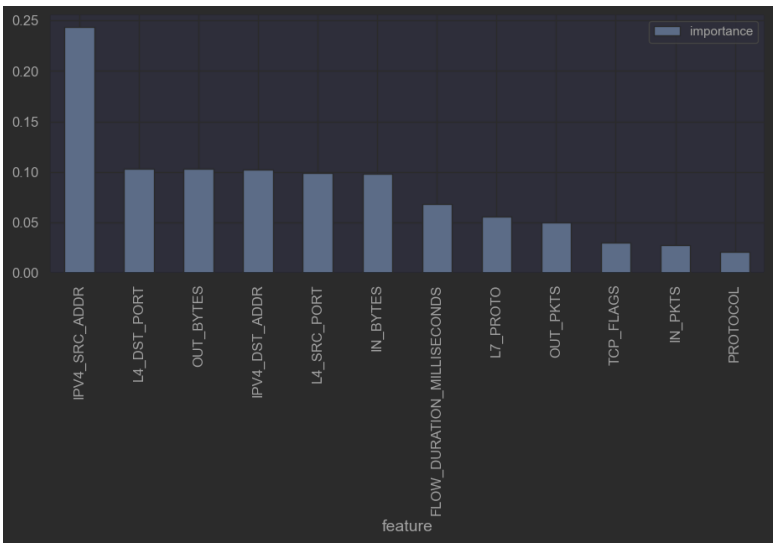

Рисунок 2 – Графік важливості ознак для обраної моделі

На основі незалежного тестування результатів моделей обрано дерево рішень в якості головної моделі машинного навчання, що також було обрано у роботі «Зменшення кількості інформативних ознак для задачі детектування комп'ютерних атак» [2]. Відштовхуючись від найвпливовіших ознак, виявлених на етапі обирання ознак, запропоновано та визначено модель HTTP запиту, що була застосована на етапах машинного навчання:

```
['L4_SRC_PORT',
 'L4_DST_PORT',
 'L7_PROTO',
 'IN_BYTES',
 'OUT_BYTES',
 'OUT_PKTS',
 'TCP_FLAGS',
 'FLOW_DURATION_MILLISECONDS',
 'IPV4_SRC_ADDR',
 'IPV4_DST_ADDR']
```

Рисунок 3 – Обрані ознаки для моделі HTTP запиту

Визначена модель при машинному навчанні дала наступні результати при ентропічній математичній моделі дерева рішень:

```
Decision Tree Classifier evaluation

Cross Validation Mean score:
 0.9660818350038441
Accuracy:
 0.9909441663646693
Confusion matrix:
```

Рисунок 4 – Результат дерева рішень

Ентропія - це кількість інформації, необхідна для точного опису деякої вибірки. Отже, якщо зразок однорідний, це означає, що всі елементи подібні, тоді ентропія дорівнює 0, інакше, якщо зразок поділено порівну, ентропія дорівнює максимум 1.

Отже, ліва чаша має найменшу ентропію, середня чаша має більшу ентропію, а права чаша має найвищу ентропію.

Математично [5]:

$$Entropy = \sum_{i=1}^{n} \mathrm{p_i} * \log(\mathrm{p_i}); \quad (1);$$

Для усунення недоліків існуючих аналогів-сервісів, розроблено мікросервіс на базі фреймворку FastAPI. Перевага такого мікросервісу в його швидкості та легковісності, простоті структури та лаконічності в загальному.

Загальна схема роботи інформаційної технології зображена на рисунку 5.

Рисунок 5 – Загальна схема алгоритму роботи інформаційної технології для виявлення HTTP запитів з аномальною поведінкою

Загальна схема алгоритму роботи модуля, що аналізує дані з HTTP запитами зображена на рисунку 6.

Рисунок 6 – Схема алгоритму роботи модуля, що аналізує дані з HTTP запитами

Висновки

Базуючись на висновках теореми про відсутність безкоштовних сніданків, задля дотримання високоточних результатів, досліджено та виконано практичну перевірку сучасних моделей машинного навчання, задача яких пов’язана з класифікацією об’єктів. На основі результатів моделей обрано дерево рішень в якості головної моделі машинного навчання. Відштовхуючись від найвпливовіших ознак, виявлених на етапі обирання ознак, визначено модель запиту, яка була застосована на етапах машинного навчання.

Інформаційну технологію імплементовано на основі мікросервісу, усунувши недоліки як точності навчання, так і необхідності втручання технічного спеціаліста.

Дані дослідження в подальшому можливо використати для створення нових моделей HTTP запитів з метою підвищення точності та якості ознак на етапі їх вибору.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. DoS-атака [Електронний ресурс] Режим доступу – www.wikiwand.com/uk/DoS-атака
2. Арсенюк І.Р., «Зменшення кількості інформативних ознак для задачі детектування комп'ютерних атак» [Електронний ресурс] Режим доступу – https://conferences.vntu.edu.ua/index.php/all-fitki/all-fitki-2018/paper/view/5097/4306
3. No Free Lunch in search and optimization [Електронний ресурс] Режим доступу – wikiwand.com/en/No_free_lunch_in_search_and_optimization
4. ML-Based NIDS Datasets [Електронний ресурс] Режим доступу – https://staff.itee.uq.edu.au/marius/NIDS_datasets/#RA5
5. Math behind Decision Tree Algorithm [Електронний ресурс] Режим доступу – https://ankitnitjsr13.medium.com/math-behind-decision-tree-algorithm-2aa398561d6d

Зелений Владислав Євгенович *– студент групи 2КН-21м, Факультет Інтелектуальних Інформаційних Технологій та Автоматизації, Вінницький Національний Технічний Університет, м. Вінниця, email: vladyslavzelenyi@gmail.com*

Арсенюк Ігор Ростиславович *– к. т. н., доцент, доцент кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця.*

Vladyslav Zelenyi *– Department of Intelligent Information Technologies and Automatization, Vinnytsia National Technical University, Vinnytsia, email: vladyslavzelenyi@gmail.com*

Ihor Arsenyuk *– Cand. Sc., Assistant Professor of the Chair of Computer Science, Vinnytsia National Technical University, Vinnytsia.*