УДК 004.77

О.М. Барановський, канд. тех. наук, доцент, А.В. Жилін, канд. тех. наук, доцент, Г.С. Голич
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

ЗАСТОСУВАННЯ МЕТОДУ МАШИННОГО НАВЧАННЯ ДЛЯ ВИРІШЕННЯ ЗАДАЧІ ДЕТЕКТУВАННЯ ТОЧКОВИХ АНОМАЛІЙ У МЕРЕЖЕВОМУ ТРАФІКУ ЗАСОБАМИ SIEM SPLUNK

O.M. Baranovskyi, Ph.D., Assoc. Prof., A.V. Zhylin, Ph.D., Assoc. Prof., H.S. Holych

APPLYING OF MACHINE LEARNING METHOD FOR DETECTING POINT ANOMALIES IN NETWORK TRAFFIC BY MEANS OF SIEM SPLUNK

Проблема детектування аномалій є поширеною в рамках виконання різних наукових і прикладних задач, зокрема і в аналізі мережевого трафіку. Актуальним залишається питання пошуку оптимального (по вимогливості до ресурсів, часу відпрацювання, якості отриманих результатів) способу детектування аномалій у мережевому трафіку через часту необхідність роботи з непомаркованими різнородними наборами даних, що призводить до значного зниження якості і слабкого практичного застосуваня окремих методів. Метод машинного навчання, за результатми окремних досліджень, є одним з найефективніших при вирішенні такого роду задач.

Метою роботи є дослідження інтегрованих алгоритмів машинного навчання в модулі ML Toolkit SIEM Splunk для визначення алгоритмів, що будуть доцільними у вирішенні задачі детектування відхилень (outliers) у мережевому трафіку.

Детектування точкових аномалій за використання ПЗ SIEM Splunk (SPL запитів) найчастіше здійснюється наступними методами:

- використання статичного порогового значення (threshold). У результаті виконання SPL запиту точковими аномаліями визначаються ті, що перевищують, або є нижчими за встановлене порогове значення.
- Використання середнього значення з статичним множником. У результаті виконання SPL запиту обраховується середнє значення по наявним числовим даним (з можливістю групування за додатковими полями) і обирається множник, від якого буде залежати верхнє порогове значення.
- Використання середнього значення з множником стандартного відхилення [1]. У результаті виконання SPL запиту, як і в попередньому методі, обраховується середнє значення і застосовується множник стандартного відхилення, який визначає фіксоване верхнє і нижнє порогове значення.
- Використання динамічного середнього значення з множником стандартного відхилення. У результаті виконання SPL запиту обраховується середнє значення для числових даних у групах і встановлюється крок, з яким буде обраховуватись середнє значення для наступної групи. Тому отримані верхнє і нижнє порогові значення будуть динамічними.

Також, для виявлення точкових аномалій засобами SIEM Splunk може використовуватись модуль MLTK (Machine Learning Toolkit), що представляє необхідний функціонал для виконання більш інтелектуальних обчислень, ніж тих, що запропоновані в попередніх методах. Було розглянуто стандартні алгоритми, інтегровані в модуль ПЗ Splunk MLTK, що стосуються категорії виявлення аномалій – DensityFunction, LOF (LocalOutlierFactor) та OneClassSVM [2].

Для порівняння вищеописаних методів детектування точкових аномалій було використано помаркований тестовий датасет traffic_1h.csv, який містить 17779 записів про

статистику отриманих мережевих даних з 3 сенсорів (sensor1, sensor2, sensor3) за протоколами (DNS, HTTP, HTTPS) протягом 1 години. Аномальною поведінкою для такого випадку вважаються відхилення від нормального розподілу у значеннях count (екстремуми), які залежать від конкретного сенсору і протоколу.

Тестування методів детектування точкових аномалій засобами SIEM Splunk продемонструвало більш якісні результати при застосуванні методу машинного навчання порівняно з іншими, незважаючи на більшу часову затримку в обрахунках.

Для порівняння результатів тестування інтегрованих алгоритмів модулю Splunk MLTK було застосовано підхід, що базується на побудові матриці похибок Confusion Matrix, де оцінка алгоритмів залежить від співвідношень елементів матриці. А саме:

- Precision (точність) – частка виявлених викидів, які є істинними аномаліями;
- Recall (повнота) – частка реальних викидів, які було виявлено системою;
- Accuracy (точність) – частка правильно класифікованих результатів до загального числа випадків (довжини вибірки);
- F1 – міра, що є середнім гармонічним між повнотою (precision) і точністю (recall).

Результати порівняння наведено в таблиці 1.

Таблиця 1

Результати тестування інтегрованих алгоритмів модулю Splunk MLTK

Метод детектування точкових аномалій	**Precision**	**Recall**	**Accuracy**	**F1**
Алгоритм DensityFunction	0.83	0.12	0.97	0.19
Алгоритм LOF	0.64	0.031	0.88	0.041
Алгоритм OneClassSVM	0.46	0.0054	0.56	0.011

Крім цього, візуалізація результатів відпрацювання алгоритмів в модулі Splunk MLTK є більш зручною для регулювання окремих параметрів, що спрощує процес порівняння результатів відпрацювання ML алгоритмів між собою для визначення оптимального способу вирішення поставленої задачі.

За використання модулю Splunk MLTK з мінімальною конфігурацією параметрів найбільш точного результату було досягнуто за використання алгоритму DensityFunction. Менш точними були результати з використанням алгоритму LOF (LocalOutlierFactor), де значній кількості середньостатистичних значень було надано мітку isOutlier=1 (визначено як викиди). Недоліком подальшого використання даного алгоритму може бути те, що зберігання моделей не підтримується і тому модель, відтренована на тестових даних не може бути застосована до нових даних. Найменш точні результати було отримано за використання алгоритму OneClassSVM. Більшість середньостатистичних значень отримала мітку isNormal=-1 (визначено як викиди), на що практично не впливала конфігурація параметрів. Загалом, алгоритм застосовується переважно для пошуку новизни в даних, тому ймовірно його використання було б більш доцільним для іншого набору даних або в ансамблі з іншим алгоритмом.

Як висновок, на основі відпрацювання тестового датасету доцільним вважається створення і подальше тренування моделей з використанням алгоритмів DensityFunction та LocalOutlierFactor.

Література:

1. Finding and removing outliers . Splunk Cloud Platform. Search Manual URL Режим доступу: https://docs.splunk.com/Documentation/SplunkCloud/8.2.2109/Search/Findingandremovingoutliers.

2. Algorithms in the Machine Learning Toolkit Splunk Machine Learning Toolkit. User Guide URL: https://docs.splunk.com/Documentation/MLApp/5.3.0/User/Algorithms.