КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 3 (11), 2021
ISSN 2663 - 4023
DOI 10.28925/2663-4023.2021.11.183194
УДК 004.056.5:004.3

Радівілова Тамара Анатоліївна
к.т.н., доцент, доцент кафедри інфокомунікаційної інженерії ім.В.В. Поповського
Харківський національний університет радіоелектроніки, Харків, Україна
ORCID ID: 0000-0001-5975-0269
tamara.radivilova@gmail.com

Кіріченко Людмила Олегівна
д.т.н., професор, професор кафедри прикладної математики
Харківський національний університет радіоелектроніки, Харків, Україна
ORCID ID: 0000-0002-2780-7993
lyudmila.kirichenko@nure.ua

Тавалбех Максим Хаджарович
аспірант кафедри інфокомунікаційної інженерії ім.В.В. Поповського
Харківський національний університет радіоелектроніки, Харків, Україна
ORCID ID: 0000-0002-9629-4183
tavalbeh@icloud.com

Ільков Андрій Анатолійович
старший помічник начальника навчального відділу
Харківський національний університет повітряних сил ім.І.Кожедуба, Харків, Україна
ORCID ID: 0000-0002-9709-2946
andreyilkov428@gmail.com

ВИЯВЛЕННЯ АНОМАЛІЙ В ТЕЛЕКОМУНІКАЦІЙНОМУ ТРАФІКУ СТАТИСТИЧНИМИ МЕТОДАМИ

Анотація. Виявлення аномалій є важливим завданням у багатьох сферах людського життя. Для виявлення аномалій використовується множина статистичних методів. У даній роботі для виявлення аномалій були обрані статистичні методи аналізу даних, такі як аналіз виживання, аналіз часових рядів (фрактальний), метод класифікації (дерева прийняття рішень), кластерний аналіз, ентропійний метод. Також наводиться опис вибраних методів. Для аналізу аномалій були взяті реалізації трафіків і атак з відкритого датасету. Для аналізу описаних методів було використано понад 3 млн. пакетів з набору даних. Датасет містив легітимний трафік (75%) і атаки (25%). Проведено імітаційне моделювання обраних статистичних методів на прикладі реалізацій мережного трафіку телекомунікаційних мереж різних протоколів. Для реалізації імітаційного моделювання були написані програми на мові програмування Pyton. Як аномалії були обрані DDoS-атаки, UDP-flood, TCP SYN, ARP-атаки і HTTP-flood. Був проведений порівняльний аналіз продуктивності обраних статистичних методів щодо виявлення аномалій (атак) за такими параметрами як ймовірність виявлення аномалій, ймовірність хибнопозитивного виявлення, час роботи кожного методу для виявлення аномалії. Результати експериментів показали працездатність кожного методу. Метод дерева рішень є найкращим за ймовірністю ідентифікації аномалій, меншій кількості хибнопозитивних спрацьовувань і часу виявлення аномалій. Метод ентропійного аналізу дещо повільніше і дає трохи більше помилкових спрацьовувань. Далі слідує метод кластерного аналізу, який дещо гірше виявляє аномалії. Тоді як метод фрактального аналізу показав меншу ймовірність виявлення аномалій, велику ймовірність помилкових спрацьовувань і більший час роботи. Найгіршим виявився метод аналізу виживання.

Ключові слова: виявлення аномалій; трафік; дерева рішень; фрактальний аналіз; кластерний аналіз; атаки; помилкові спрацьовування

ВСТУП

Аномальні відхилення в різній кількості присутні в багатьох явищах. Це і незвичайні коливання температури, серцевого ритму, викиди сонячної енергії і багато іншого. Також аномалії присутні в інформаційному телекомунікаційному просторі (пікові викиди трафіку, відмова обладнання, кібератаки). Дуже часто аномалії відбуваються через наявність кібератак. Кіберзлочинність – одна з найбільших сучасних проблем, з якими стикається людство. Кількість атак і збиток від них постійно ростуть і коливається в межах верхньої межі шкали. Атаки призводять до пошкодження і знищення даних, зниження продуктивності, крадіжці інтелектуальної власності, особистих або фінансових даних, грошових коштів.

Крім цього, існує висока ймовірність збоїв в роботі обладнання після атаки, також необхідний додатковий час для криміналістики, відновлення і видалення зламаних даних і систем [1]. Таким чином детектування аномалій є важливим завданням. У всіх областях уявлення про аномалії схожі: це дані, які сильно відрізняються від нормальних. В [2] виявлення аномалій - це процес пошуку об’єктів даних, поведінка яких сильно відрізняється від очікуваної. Однак необхідно відзначити, що в аналізі даних існує два напрямки пошуку аномалій: детектування викидів (англ. Outlier Detection) і детектування новизни (англ. Novelty Detection). На відміну від викиду, новий об’єкт в самій вибірці поки відсутній (він з’явиться через деякий час, і завдання полягає в тому, щоб виявити його при появі). Тобто, якщо йде спостереження за роботою системи, то після проникнення в неї вірусу, робота системи стає новизною [3]. Наприклад, якщо аналізується кількість службового трафіку і відкидаються аномально великі чи маленькі значення, то аналізується новизна і йде боротьба з викидами. А якщо кількість трафіку для кожного нового виміру порівнюється з минулими і відкидаються аномальні, то аналізуються викиди і йде боротьба з новизною. При аналізі аномалій зазвичай робиться кілька припущень про нормальність даних, а потім виділяються об’єкти, що порушують її. Таким чином, кластери – це групи схожих за характеристиками точок, а аномалії – об’єкти, що вибиваються із загального набору.

Дані, які виходять за межі кластера, можуть бути як шумом так і аномаліями. Шум (англ. noise) - це слабкий викид (він може розмивати кордони класа/кластера). Аномаліями ж є сильні викиди, які спотворюють кордони класа/кластера. Виявлення аномалій дозволяє виділити шум і нормальні сигнали, а також визначити потенційні фактори, які сприяли появі цих сигналів або аномалій [4]. Іншими словами, це дозволяє визначити, які статистичні флуктуації є суттєвими, а які – ні, першопричину істинної аномалії і надійні метричні прогнози.

Статистичні методи аналізу даних використовуються при виявленні аномалій для визначення того, як певна метрика змінилася відносно попередніх даних. Статистичними методами досліджується випадковість і детермінованість спостережень (частота і кількість з’єднань, переданих даних); теоретико-ймовірнісні моделі; основні характеристики спостережуваних випадкових подій; граничні теореми теорії ймовірностей і їх застосування; статистичні гіпотези і їх розрізнення; поліноміальні схеми і схеми розміщень як моделі випадкових процесів інформаційної безпеки; оцінка розладки процесу спостережень як метод виявлення реалізації інциденту інформаційної безпеки; багатовимірні моделі і кореляційні зв’язки та ін. Основними методами, які використовуються в інформаційній безпеці при аналізі подій і стану системи є статистичні методи, деякі з них описані нижче. [5-7].

Критеріями якості детектування аномалій були обрані ймовірність ідентифікації аномалій, ймовірність хибнопозитивних ідентифікацій аномалій (ідентифікація наявності аномалії при її реальній відсутності), час потрібний для виявлення аномалій [8].

Метою статті є аналіз статистичних методів виявлення аномалій в телекомунікаційному трафіку, що дає змогу обрати більш підходящий метод для кожної задачі.

СТАТИСТИЧНІ МЕТОДИ ДОСЛІДЖЕННЯ

Аналіз виживання – статистичний метод аналізу тривалості деякого процесу до моменту його припинення. Під процесом розуміють тривалість будь-якого явища в часі (наприклад, аналіз тривалості сесії з'єднання) [9, 10]. Аналіз виживання зазвичай включає в себе два основних етапи:

–оцінка терміну настання до аналізованого події (побудови таблиць дожиття методом Каплана-Мейра або ін. Методами);

–моделювання ризику настання аналізованого події (регресія Кокса).

Для аналізу роботи даного методу було проведено аналіз коректної роботи системи на основі легітимного трафіку і стандартного режиму роботи. При нестандартних режимах роботи та при атаках припинення роботи системних протоколів і системи в цілому ідентифікувалися як покращення роботи спостережуваного процесу. Ризиком настання події вважається ймовірність атаки за набраними статистичними даними.

Аналіз часового ряду – метою є побудова прогнозу значень часового ряду на майбутні періоди. Основними завданнями аналізу часового ряду є необхідність визначення, вплив яких компонент формує значення часового ряду, і побудова математичної моделі для кожної компоненти або їх сукупності. Метод фрактального аналізу, детально описаний в роботах [8, 11-13], застосовувався для ідентифікації аномалій. Аномалії (атаки) ідентифікувалися в залежності від змін мультифрактальних характеристик часового ряду.

Мультифрактальні об'єкти є статистично неоднорідними самоподібними об'єктами та мають більш складну скейлінгову поведінку. У цьому випадку скейлінговою характеристикою є нелінійна функція $h(q)$ – узагальнений показник Херста. Значення $h(q)$ при $q=2$ збігаються зі значеннями ступеня самоподібності H. Для монофрактальних процесів узагальнений показник Херста $h(q)=H$. Діапазон значень узагальненого показника Херста $\Delta h(q)=h(q1)-h(q2)$ визначає ступінь мультифрактальності: чим більше значення $\Delta h(q)$, тим більше виражені мультифрактальні властивості процесу. У разі монофрактальності $\Delta h(q)=0$. Одним з найбільш затребуваних на практиці методів аналізу самоподібних часових рядів є мультифрактальний детрендований флуктуаційній аналіз. Він дозволяє оцінювати узагальнений показник Херста $h(q)$ для нестаціонарних часових рядів.

Дерева рішень (англ. decision trees) – це статистичний метод, що дозволяє передбачати приналежність спостережень або об'єктів до того чи іншого класу категоріальної залежної змінної або середнє значення кількісної змінної в залежності від відповідних значень однієї або декількох незалежних змінних. Метод дерев рішень

можна застосувати для вирішення задач класифікації, що виникають в найрізноманітніших областях, і вважається одним з найефективніших. [14]

Метод дерев рішень для завдання класифікації або прогнозування полягає в тому, щоб здійснювати процес розподілу вихідних даних на групи, поки не будуть отримані однорідні (або майже однорідні) їхні підмножини. Сукупність правил, які дають таке розділення, дозволяє потім робити прогноз (цільова змінна), отриманий в результаті оцінки деяких вхідних ознак для нових даних (предикторів). [8, 15, 16]

Дерева рішень поділяються на дерева регресії і дерева класифікації. Дерева регресій працюють із кількісною цільовою змінною. Алгоритм навчання (або формування) дерева діє за принципом рекурсивного секціонування. Секціонування набору даних (тобто розбиття на непересічні підмножини) здійснюється на основі використання найбільш підходящого для цієї ознаки. У дереві створюється відповідний вузол прийняття рішень, і процес триває рекурсивно до тих пір, доки не виконається критерій зупинки.

При побудові дерев класифікації передбачається приналежність об'єкта до тієї чи іншої категорії цільової змінної (класу) в залежності від відповідних значень предикторів (ознак). Наприклад, класифікуються легітимний і атакований трафік в залежності від його властивостей. Ознаками можуть виступати самі значення вибірки. З набору ознак для побудови розбиття, потрібно вибрати такі, що дозволять отримати якомога більше однорідних (чистих) груп.

Кластерний аналіз (англ. cluster analysis) – сукупність багатовимірних статистичних методів класифікації об'єктів за ознаками, які їх характеризують, поділ сукупності об'єктів на однорідні групи, близькі по визначальним критеріям, виділення об'єктів певної групи. Кластер – це групи об'єктів, виділені в результаті кластерного аналізу на основі заданої міри схожості або відмінностей між об'єктами. Об'єкт – це конкретні предмети дослідження, які необхідно класифікувати. [17]

Існують різні алгоритми кластерного аналізу. Розглянемо DBSCAN метод основною задачею якого є можливість ідентифікувати нетипові об'єкти із набору даних. Основна ідея полягає в розподілі в кластерах схожих об'єктів по щільності. Для початку визначається радіус близькості і кількість об'єктів, які повинні бути розташовані в межах цього радіусу близькості. Щільно розташованими або досить схожими є об'єкти, які знаходяться на відстані рівній або менше заданому радіусу близькості. [4, 18, 19, 20]

Нехай буде набір даних, що містить n об'єктів; $\rho(x, y)$ визначена функція подібності, r – радіус близькості; m – мінімальна кількість об'єктів, яка повинна знаходитися в межах цього радіусу. Є функція M, яка визначає кількість об'єктів, що знаходяться в радіусі близькості

$$M(x_i) = \sum_{\substack{j=1 \\ j \neq i}}^{n} c(x_i, x_j),$$

де $c(x_i, x_j)$ – функція членства:

$$c(x_i, x_j) = \begin{cases} 1, \rho(x_i, x_j) \leq r, \\ 0, \rho(x_i, x_j) > r. \end{cases}$$

Об'єктами ядра є об'єкти, для яких умова $M(x_i) \geq m$ є істиною. Граничними об'єктами є об'єкти, для яких умова $M(x_i) < m$ є істинною і існує такий об'єкт ядра y_i,

ISSN 2663 - 4023

що умова $\rho(x_i, y_i) \leq r$ єдина. Об'єктами шума є об'єкти, для яких умова $M(x_i) < m$ є істиною і немає такого об'єкта ядра y_i, що умова $\rho(x_i, y_i) \leq r$ є істиною. У термінах методу DBSCAN шум – це набір нетипових об'єктів.

Основні кластери визначаються об'єктом ядра. Потім за цими кластерам розподіляються граничні об'єкти, а потім шумові об'єкти:

- Якщо в радіусі r немає іншого об'єкта шуму, то цей об'єкт розподіляється по окремим кластерам.
- Якщо в радіусі r є хоча б один об'єкт шуму, вони об'єднуються в загальну групу.
- Якщо в радіусі r є хоча б один об'єкт ядра або граничний об'єкт, то шумовий об'єкт додається в кластер, який містить цей об'єкт ядра або прикордонний об'єкт.

Основним недоліком такого способу розподілу є неможливість визначення нетипових об'єктів наприклад граничних об'єктів.

Аналіз ентропії використовується у виявленні аномалій для формування статистичного критерію з метою перевірки приналежності досліджуваного екземпляра аномальному класу. Суть методу максимуму ентропії полягає в побудові моделі, яка б максимізувала значення ентропії. Це відповідає тому припущенню, що при збільшенні числа унікальних записів вони рівномірно розподіляються між обраними класами множини, що призводить до збільшення ентропії.

З метою ефективності вибору функції ознак для обчислень часто вибираються найбільш важливі характеристики навчальних даних в моделі, і, в свою чергу, модель відображає емпіричний розподіл з найменшою кількістю функцій ознак і властивостей.

Ентропія множини ψ визначається наступним чином [21]:

$$H = -\sum_{\omega \in \psi} P_\omega \log_2 P_\omega .$$

де P_ω позначає ймовірність появи елемента ω в ψ множині.

Для виявлення аномалій в [21] спершу застосовується метод максимуму ентропії для створення нормальної моделі, в якій виділені класи даних мають найкращий рівномірним розподілом. Далі застосовується умовна ентропія для виявлення відмінностей між розподілом записів в поточних даних в порівнянні з розподілом, знайденим в результаті методу максимуму ентропії. З метою ефективності функції ознак для обчислень часто вибираються найбільш важливі характеристики навчальних даних в моделі, і, в свою чергу, модель описує емпіричний розподіл з найменшою кількістю функцій ознак і властивостей [22-24].

МЕТОДИКА ТА РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Наведені вище методи були протестовані за набором реальних даних трафіка. Для оцінки методів виявлення аномалій було використано мову програмування Pyton. Реалізації трафіка було агреговано за часом для аналізу за часовим рядом. Для аналізу іншими наведеними методами використовувались функції, що працюють з кожною дейтаграмою IP (UDP) для повторного складання потоку, видачі повідомлень, обчислення статистичних характеристик властивостей трафіка для кожного атрибута

ISSN 2663 - 4023

пакета (наприклад, довжина датаграми і розмір вікна TCP, IP-адреси джерела і призначення) та інших маніпуляцій з пакетами. Наприклад, якщо є 1000 послідовних пакетів і функція обчислює частоту зустрічі кожної унікальної IP-адреси джерела в цих 1000 пакетів, тоді матимемо модель розподілу адрес джерела. Подальші обчислення з цим розподілом дозволяють вимірювати випадковість або однорідність адрес, а також «коефіцієнт якості» розподілу порівняно з попередніми вимірами. Потім застосовується обраний метод для визначення відмінностей між розподілом класів пакетів у поточному трафіку порівняно з розподілом, обчисленим раніше.

Для аналізу наведених методів було розглянуто понад 3 млн. пакетів з набору даних [25]. Датасет містив легітимний трафік (75%) та атаки (25%). В якості аномальних даних були використані реалізації атак: DDoS, UDP-flood, потоків TCP SYN, ARP атаки та HTTP flood. В ході експериментів враховувались такі параметри пакетів трафіка як [22, 26-28]:

– duration – тривалість з'єднання в секундах;
– protocol_type – тип протоколу: TCP, UDP та інше;
– service – тип обслуговування: HTTP, FTP, TELNET і інше;
– flag – прапорець з'єднання: норма або помилка;
– scr_bytes – кількість байт даних від джерела до одержувача;
– dst_bytes – кількість даних від одержувача до джерела;
– wrong_fragments – кількість «неправильних» фрагментів;
– urgent – кількість термінових (urgent) пакетів;
– hot – кількість «гарячих» індикаторів;
– num_failed_logins – кількість помилкових спроб входу;
– logged_in – 1 – успішний вхід, 0 в іншому випадку;
– num_compromised – число скомпрометованих умов;
– root_shell – 1 – якщо отримана коренева оболонка, 0 – в іншому випадку;
– su_attempted – 1 – якщо була спроба виконати «su root», 0 – в іншому випадку;
– num_root – кількість доступів типу «root»;
– num_file_creations – кількість операцій створення файла;
– num_shells – кількість «підказок оболонки»;
– num_access_files – кількість отримань доступу до контролю над файлами;
– is_host_login – 1 – якщо логін належить до «host» списку;
– is_quest_login – 1 – якщо підключення типу «гість»;
– count – кількість підключень до хоста за останні дві секунди;
– srv_count – кількість підключень до сервісу за останні дві секунди;
– serror_rate – відсоток підключень з syn помилками.

В ході експерименту проводилася ідентифікація не тільки DDoS атак, але і UDP-flood, HTTP flood, TCP SYN, ARP-spoofing атаки. В таблиці 1 представлені значення ймовірності ідентифікації аномалій P_a, ймовірність хибнопозитивних ідентифікацій аномалій P_f та час, необхідний для виявлення аномалій різними методами. В ході експериментів виявлено, що запропоновані методи мають хибнопозитивні спрацьовування. Це може відбуватися через високу швидкість трафіка (генерація портів відбувається автоматично і ці номери портів рідко зустрічаються в навчальній вибірці), через обмеження інформації заголовка пакета (зашифровані дані) і т.ін.

ISSN 2663 - 4023

ТАБЛИЦЯ 1

ЙМОВІРНОСТІ ІДЕНТИФІКАЦІЇ АНОМАЛІЙ P_A, ЙМОВІРНІСТЬ ХИБНОПОЗИТИВНИХ ІДЕНТИФІКАЦІЙ АНОМАЛІЙ P_F ТА ЧАС, НЕОБХІДНИЙ ДЛЯ ВИЯВЛЕННЯ АНОМАЛІЙ РІЗНИМИ МЕТОДАМИ

	DDoS		UDP-flood		TCP SYN		ARP		HTTP flood		Час для виявлення аномалій, сек
	P_a	P_f	P_a	P_f	P_a	P_f	P_a	P_f	P_a	P_f	
Аналіз виживання	0.7	0.21	0.69	0.22	0.76	0.12	0.6	0.16	0.8	0.32	95
Аналіз часового ряда	0.9	0.1	0.72	0.25	0.63	0.27	0.77	0.31	0.79	0.12	39
Метод дерев рішень	0.96	0.05	0.95	0.05	0.94	0.08	0.96	0.11	0.94	0.09	29
DBSCAN метод	0.94	0.1	0.93	0.17	0.94	0.18	0.92	0.12	0.92	0.09	34
Ентропійний метод	0.96	0.06	0.94	0.09	0.94	0.07	0.95	0.1	0.93	0.07	34

Таким чином, розглянуті методи виявлення аномалій вторгнень показали високі значення ймовірності виявлення атак (близько 94%) та низькі значення помилковопозитивного індексу (близько 10%). Отже, кожен з приведених методів може використовуватися для виявлення різних типів атак.

Якщо відбувається аномалія, незалежно від її типу та того наскільки повільно збільшується трафік, вона може бути виявлена. Розглянуті методи потребують постійного обсягу пам’яті і складаються з навчальних виборок, статистичних підрахунків даних трафіка. Порівняльний аналіз ефективності застосування цих методів для виявлення аномалій (атак) за ймовірністю виявлення аномалій, ймовірністю хибнопозитивного виявлення, часу роботи кожного методу для виявлення аномалії. Результати експериментів показали, що метод дерева рішень є найкращим за всіма параметрами порівняння. Метод ентропійного аналізу дещо повільніше і дає трохи більше помилкових спрацьовувань. Далі слідує метод кластерного аналізу, який дещо гірше виявляє аномалії. А метод фрактального аналізу показав меншу ймовірність виявлення аномалій, велику ймовірність помилкових спрацьовувань і більший час роботи. Найгіршим був метод аналізу виживання.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

В роботі проведено аналіз статистичних методів детектування аномалій. Наведено короткий опис статистичних методів детектування аномалій, таких як аналіз виживання, аналіз часового ряду (фрактальний), метод класифікації (дерева рішень), кластерний аналіз (DBSCAN), метод аналізу ентропії. Також в роботі проведено імітаційне моделювання описаних методів. Виявлення аномалій проводилося на реалізаціях трафіку і атак, які були обрані з відкритого набору даних. Як аномалії були обрані реалізації DDoS атак, UDP-flood, TCP SYN, ARP атак та HTTP flood. В результаті експериментів було виявлено, що всі методи ідентифікують аномалії. Найкраще ідентифікує аномалії з ймовірністю 96% метод класифікації дерев рішень. Трохи гірше працює метод аналізу ентропії 96%, але хибнопозитивних спрацьовувань на 1% більше. Метод кластеризації DBSCAN показав 94% ідентифікації аномалій і майже на більше 2% хибнопозитивних спрацьовувань. Метод аналізу часового ряду показав 90% ймовірності ідентифікації, а метод виживання ідентифікує аномалії з 76% ймовірністю, але меншою кількістю хибнопозитивних значень.

У наступних роботах планується провести експерименти і зробити обчислення для ансамблів методів вибравши найбільш точні методи виявлення аномалій, найбільш

інформативні ознаки, оцінки аномальності і потім метод усереднення отриманих оцінок.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1 Radivilova, T., Kirichenko, L., Tawalbeh, M., Zinchenko, P., & Bulakh, V. (2020). THE LOAD BALANCING OF SELF-SIMILAR TRAFFIC IN NETWORK INTRUSION DETECTION SYSTEMS. *Cybersecurity: Education, Science, Technique*, *3*(7), 17–30. https://doi.org/10.28925/2663-4023.2020.7.1730

2 Han, J., Kamber, M., Pei, J. (2011). Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems*, *5*(4), 83-124. https://doi.org/10.1016/C2009-0-61819-5

3 Chandola, V., Banerjee, A., Kumar, V. (2009). Anomaly Detection: A Survey. *ACM Computing Survey*, *41*, 1–58.

4 Kirichenko, L., Radivilova, T., & Tkachenko, A. (2019). Comparative Analysis of Noisy Time Series Clustering. У *COLINS-2019: Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Systems, Volume I: Main Conference Kharkiv, Ukraine* (с. 184–196).

5 Madhuri, G. S. (2020). Usha Rani M. Statistical Approaches to Detect Anomalies. У Venkata Krishna P., Obaidat M. (eds) Emerging Research in Data Engineering Systems and Computer Communications. *Advances in Intelligent Systems and Computing*. https://doi.org/10.1007/978-981-15-0135-7_46.

6 Bendich, P., Chin, S. P., Clark, J., Desena, J., Harer, J., Munch, E., Newman, A., Porter, D., Rouse, D., Strawn, N., & Watkins, A. (2016). Topological and statistical behavior classifiers for tracking applications. *IEEE Transactions on Aerospace and Electronic Systems*, *52*(6), 2644–2661. https://doi.org/10.1109/taes.2016.160405

7 Goldstein, M., & Uchida, S. (2016). A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. *PLOS ONE*, *11*(4), Стаття e0152173. https://doi.org/10.1371/journal.pone.0152173

8 Kirichenko, L., Radivilova, T., & Bulakh, V. (2019). Machine Learning in Classification Time Series with Fractal Properties. *Data*, *4*(5), 1-13. https://doi.org/10.3390/data4010005

9 Han, M. L., Kwak, B. I., & Kim, H. K. (2018). Anomaly intrusion detection method for vehicular networks based on survival analysis. *Vehicular Communications*, *14*, 52–63. https://doi.org/10.1016/j.vehcom.2018.09.004

10 Pinto, J. D.(2015). Outlier Detection in Survival Analysis: Thesis to obtain the Master of Science Degree in Electrical and Computer Engineering.

11 Zhang, R., Zhou, M., Gong, X., He, X., Qian, W., Qin, S., & Zhou, A. (2014). Detecting anomaly in data streams by fractal model. *World Wide Web*, *18*(5), 1419–1441. https://doi.org/10.1007/s11280-014-0296-y

12 Gong, X., Qian, W., Qin, S., Zhou, A. (2003). Fractal Based Anomaly Detection over Data Streams. In: Ishikawa Y., Li J., Wang W., Zhang R., Zhang W. (eds) Web Technologies and Applications. *Lecture Notes in Computer Science*. https://doi.org/10.1007/978-3-642-37401-2_54

13 Radivilova, T., Kirichenko, L., Alghawli, A. S., Ilkov, A., Tawalbeh, M., Zinchenko, P. (2020). The complex method of intrusion detection based on anomaly detection and misuse detection. У *DESSERT: Proceedings of 2020 IEEE 11th International Conference on Dependable Systems, Services and Technologies* (с. 133-137). https://doi.org/10.1109/DESSERT50317.2020.9125051.

14 Kirichenko, L., Radivilova, T., & Bulakh, V. (б. д.). Binary classification of fractal time series by machine learning methods. У V. Lytvynenko, S. Babichev, W. Wójcik, O. Vynokurova, S. Vyshemyrskaya & S. Radetskaya (Ред.), *Lecture notes in computational intelligence and decision making* (с. 701–711). Advances in Intelligent Systems and Computing.

15 Reif, M., Goldstein, M., Stahl, A., Breuel, T. M. (2008). Anomaly detection by combining decision trees and parametric densities. *19th International Conference on Pattern Recognition*: Proceedings (с. 1-4).

16 Botana, I. L.-R., Eiras-Franco, C., & Alonso-Betanzos, A. (2020). Regression Tree Based Explanation for Anomaly Detection Algorithm. *Proceedings*, *54*(1), 7. https://doi.org/10.3390/proceedings2020054007

17 Кириченко, Л.О., Ткаченко, А. Е., Радивилова, Т. А. (2019). Кластеризация зашумленных временных рядов. *Системні технології. Регіональний міжвузівський збірник наукових праць*, *3*(122), 133-139.

18 Alam, M. (2020, 10 жовтня). *DBSCAN—a density-based unsupervised algorithm for fraud detection*. Medium. https://towardsdatascience.com/dbscan-a-density-based-unsupervised-algorithm-for-fraud-detection-887c0f1016e9

19 Sheridan, K., Puranik, T. G., Mangortey, E., Pinon-Fischer, O. J., Kirby, M., Mavris, D. N. (2020). An application of dbscan clustering for flight anomaly detection during the approach phase. *AIAA*: Proceedings of Scitech 2020 Forum, (с. 1851). https://doi.org/10.2514/6.2020-1851

20 Saeedi Emadi, H., & Mazinani, S. M. (2017). A Novel Anomaly Detection Algorithm Using DBSCAN and SVM in Wireless Sensor Networks. *Wireless Personal Communications*, *98*(2), 2025–2035. https://doi.org/10.1007/s11277-017-4961-1

21 Gu, Y., McCallum, A., Towsley, D. (2005). Detecting Anomalies in Network Traffic Using Maximum Entropy Estimation. *SIGCOMM*: Proceedings of the 5th ACM conference on Internet Measurement (с. 32–32).

22 Radivilova, T., Kirichenko, L., Alghawli, A. S. (2019). Entropy Analysis Method for Attacks Detection. *PIC S&T*: Proceedings of 2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology, (с. 443-446). https://doi.org/10.1109/PICST47496.2019.9061451

23 Callegari, C., Giordano, S., Pagano, M. (2017). Entropy-based network anomaly Detection. *ICNC: Proceedings of 2017 International Conference on Computing* (с. 334-340), Networking and Communications. https://doi.org/10.1109/ICCNC.2017.7876150.

24 Shukla, A. S., & Maurya, R. (2018). Entropy-Based Anomaly Detection in a Network. *Wireless Personal Communications*, *99*(4), 1487–1501. https://doi.org/10.1007/s11277-018-5288-2

25 *UGR'16 Dataset*. (б. д.). NESG - Home. https://nesg.ugr.es/nesg-ugr16/

26 Kalita, J. K., Bhuyan, M. H., & Bhattacharyya, D. K. (2017). *Network Traffic Anomaly Detection and Prevention: Concepts, Techniques, and Tools*. Springer.

27 Saad, A., Sisworahardjo, N. (2017). Data analytics-based anomaly detection in smart distribution network. *ICHVEPS*: *Proceedings of the 2017 International Conference on High Voltage Engineering and Power Systems*, IEEE.

28 Fernandes, G., Rodrigues, J. J. P. C., Carvalho, L. F., Al-Muhtadi, J. F., & Proença, M. L. (2018). A comprehensive survey on network anomaly detection. *Telecommunication Systems*, *70*(3), 447–489. https://doi.org/10.1007/s11235-018-0475-8

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 3 (11), 2021
ISSN 2663 - 4023

Tamara Radivilova
Ph.D, associated professor,
V.V. Popovskyy department of infocommunication engineering
Kharkiv National University of Radio Electronics, Kharkiv, Ukraine
ORCID ID: 0000-0001-5975-0269
tamara.radivilova@gmail.com

Lyudmyla Kirichenko
Dr., professor, professor of applied mathematics department
Kharkiv National University of Radio Electronics, Kharkiv, Ukraine
ORCID ID: 0000-0002-2780-7993
lyudmila.kirichenko@nure.ua

Maksym Tawalbeh
Post graduate student V.V. Popovskyy department of infocommunication engineering,
Kharkiv National University of Radio Electronics, Kharkiv, Ukraine
ORCID ID: 0000-0002-9629-4183
tavalbeh@icloud.com

Andrii Ilkov
senior assistant of the head of educational department
Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine
ORCID ID: 0000-0002-9709-2946
andreyilkov428@gmail.com

DETECTION OF ANOMALIES IN THE TELECOMMUNICATIONS TRAFFIC BY STATISTICAL METHODS

Abstract. Anomaly detection is an important task in many areas of human life. Many statistical methods are used to detect anomalies. In this paper, statistical methods of data analysis, such as survival analysis, time series analysis (fractal), classification method (decision trees), cluster analysis, entropy method were chosen to detect anomalies. A description of the selected methods is given. To analyze anomalies, the traffic and attack implementations from an open dataset were taken. More than 3 million packets from the dataset were used to analyze the described methods. The dataset contained legitimate traffic (75%) and attacks (25%). Simulation modeling of the selected statistical methods was performed on the example of network traffic implementations of telecommunication networks of different protocols. To implement the simulation, programs were written in the Pyton programming language. DDoS attacks, UDP-flood, TCP SYN, ARP attacks and HTTP-flood were chosen as anomalies. A comparative analysis of the performance of these methods to detect anomalies (attacks) on such parameters as the probability of anomaly detection, the probability of false positive detection, the running time of each method to detect the anomaly was carried out. Experimental results showed the performance of each method. The decision tree method is the best in terms of anomaly identification probability, fewer false positives, and anomaly detection time. The entropy analysis method is slightly slower and gives slightly more false positives. Next is the cluster analysis method, which is slightly worse at detecting anomalies. Then the fractal analysis method showed a lower probability of detecting anomalies, a higher probability of false positives and a longer running time. The worst was the survival analysis method.

Keywords: anomaly detection; traffic; decision trees; fractal analysis; cluster analysis; attacks; false positives.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1 Radivilova, T., Kirichenko, L., Tawalbeh, M., Zinchenko, P., & Bulakh, V. (2020). THE LOAD BALANCING OF SELF-SIMILAR TRAFFIC IN NETWORK INTRUSION DETECTION SYSTEMS. *Cybersecurity: Education, Science, Technique*, *3*(7), 17–30. https://doi.org/10.28925/2663-4023.2020.7.1730

2 Han, J., Kamber, M., Pei, J. (2011). Data mining concepts and techniques third edition. *The Morgan Kaufmann Series in Data Management Systems*, *5*(4), 83-124. https://doi.org/10.1016/C2009-0-61819-5

3 Chandola, V., Banerjee, A., Kumar, V. (2009). Anomaly Detection: A Survey. *ACM Computing Survey*, *41*, 1–58.

4 Kirichenko, L., Radivilova, T., & Tkachenko, A. (2019). Comparative Analysis of Noisy Time Series Clustering. У *COLINS-2019: Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Systems, Volume I: Main Conference Kharkiv, Ukraine* (p. 184–196).

5 Madhuri, G. S. (2020). Usha Rani M. Statistical Approaches to Detect Anomalies. У Venkata Krishna P., Obaidat M. (eds) Emerging Research in Data Engineering Systems and Computer Communications. *Advances in Intelligent Systems and Computing*. https://doi.org/10.1007/978-981-15-0135-7_46.

6 Bendich, P., Chin, S. P., Clark, J., Desena, J., Harer, J., Munch, E., Newman, A., Porter, D., Rouse, D., Strawn, N., & Watkins, A. (2016). Topological and statistical behavior classifiers for tracking applications. *IEEE Transactions on Aerospace and Electronic Systems*, *52*(6), 2644–2661. https://doi.org/10.1109/taes.2016.160405

7 Goldstein, M., & Uchida, S. (2016). A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. *PLOS ONE*, *11*(4), Стаття e0152173. https://doi.org/10.1371/journal.pone.0152173

8 Kirichenko, L., Radivilova, T., & Bulakh, V. (2019). Machine Learning in Classification Time Series with Fractal Properties. *Data*, *4*(5), 1-13. https://doi.org/10.3390/data4010005

9 Han, M. L., Kwak, B. I., & Kim, H. K. (2018). Anomaly intrusion detection method for vehicular networks based on survival analysis. *Vehicular Communications*, *14*, 52–63. https://doi.org/10.1016/j.vehcom.2018.09.004

10 Pinto, J. D.(2015). Outlier Detection in Survival Analysis: Thesis to obtain the Master of Science Degree in Electrical and Computer Engineering.

11 Zhang, R., Zhou, M., Gong, X., He, X., Qian, W., Qin, S., & Zhou, A. (2014). Detecting anomaly in data streams by fractal model. *World Wide Web*, *18*(5), 1419–1441. https://doi.org/10.1007/s11280-014-0296-y

12 Gong, X., Qian, W., Qin, S., Zhou, A. (2003). Fractal Based Anomaly Detection over Data Streams. In: Ishikawa Y., Li J., Wang W., Zhang R., Zhang W. (eds) Web Technologies and Applications. *Lecture Notes in Computer Science*. https://doi.org/10.1007/978-3-642-37401-2_54

13 Radivilova, T., Kirichenko, L., Alghawli, A. S., Ilkov, A., Tawalbeh, M., Zinchenko, P. (2020). The complex method of intrusion detection based on anomaly detection and misuse detection. У *DESSERT: Proceedings of 2020 IEEE 11th International Conference on Dependable Systems, Services and Technologies* (с. 133-137). https://doi.org/10.1109/DESSERT50317.2020.9125051.

14 Kirichenko, L., Radivilova, T., & Bulakh, V. Binary classification of fractal time series by machine learning methods. У V. Lytvynenko, S. Babichev, W. Wójcik, O. Vynokurova, S. Vyshemyrskaya & S. Radetskaya (Ред.), *Lecture notes in computational intelligence and decision making* (с. 701–711). Advances in Intelligent Systems and Computing.

15 Reif, M., Goldstein, M., Stahl, A., Breuel, T. M. (2008). Anomaly detection by combining decision trees and parametric densities. *19th International Conference on Pattern Recognition*: Proceedings (с. 1-4).

16 Botana, I. L.-R., Eiras-Franco, C., & Alonso-Betanzos, A. (2020). Regression Tree Based Explanation for Anomaly Detection Algorithm. *Proceedings*, *54*(1), 7. https://doi.org/10.3390/proceedings2020054007

17 Kirichenko, L.O., Tkachenko, A.E., Radivilova, T.A. (2019). Clustering of noisy time series. System technologies. Regional mizhvuzivskiy zbirnik naukovikh prats, 3 (122), 133-139.

18 Alam, M. (2020). *DBSCAN—a density-based unsupervised algorithm for fraud detection*. Medium. https://towardsdatascience.com/dbscan-a-density-based-unsupervised-algorithm-for-fraud-detection-887c0f1016e9

19 Sheridan, K., Puranik, T. G., Mangortey, E., Pinon-Fischer, O. J., Kirby, M., Mavris, D. N. (2020). An application of dbscan clustering for flight anomaly detection during the approach phase. *AIAA*: Proceedings of Scitech 2020 Forum, (p. 1851). https://doi.org/10.2514/6.2020-1851

20 Saeedi Emadi, H., & Mazinani, S. M. (2017). A Novel Anomaly Detection Algorithm Using DBSCAN and SVM in Wireless Sensor Networks. *Wireless Personal Communications*, *98*(2), 2025–2035. https://doi.org/10.1007/s11277-017-4961-1

21 Gu, Y., McCallum, A., Towsley, D. (2005). Detecting Anomalies in Network Traffic Using Maximum Entropy Estimation. *SIGCOMM*: Proceedings of the 5th ACM conference on Internet Measurement (p. 32–32).

22 Radivilova, T., Kirichenko, L., Alghawli, A. S. (2019). Entropy Analysis Method for Attacks Detection. *PIC S&T*: Proceedings of 2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology, (p. 443-446). https://doi.org/10.1109/PICST47496.2019.9061451

23 Callegari, C., Giordano, S., Pagano, M. (2017). Entropy-based network anomaly Detection. *ICNC: Proceedings of 2017 International Conference on Computing* (p. 334-340), Networking and Communications. https://doi.org/10.1109/ICCNC.2017.7876150.

24 Shukla, A. S., & Maurya, R. (2018). Entropy-Based Anomaly Detection in a Network. *Wireless Personal Communications*, *99*(4), 1487–1501. https://doi.org/10.1007/s11277-018-5288-2

25 *UGR'16 Dataset*. NESG - Home. https://nesg.ugr.es/nesg-ugr16/

26 Kalita, J. K., Bhuyan, M. H., & Bhattacharyya, D. K. (2017). *Network Traffic Anomaly Detection and Prevention: Concepts, Techniques, and Tools*. Springer.

27 Saad, A., Sisworahardjo, N. (2017). Data analytics-based anomaly detection in smart distribution network. *ICHVEPS*: *Proceedings of the 2017 International Conference on High Voltage Engineering and Power Systems*, IEEE.

28 Fernandes, G., Rodrigues, J. J. P. C., Carvalho, L. F., Al-Muhtadi, J. F., & Proença, M. L. (2018). A comprehensive survey on network anomaly detection. *Telecommunication Systems*, *70*(3), 447–489. https://doi.org/10.1007/s11235-018-0475-8