УДК 004.414.3:004.056.53
DOI https://doi.org/10.32782/2663-5941/2024.3.1/21

Мойсеєнко О.В.
Івано-Франківський національний технічний університет нафти і газу

ПРОГНОЗУВАННЯ КІБЕРАТАК НА ОСНОВІ МОНІТОРИНГУ ІНТЕНСИВНОСТІ ТРАФІКУ В КОМП'ЮТЕРНИХ МЕРЕЖАХ

Численні дослідження мережевих аномалій або інтрузивних мережевих подій не дозволяють розв'язати деякі задачі безпеки мереж. Дослідницькі проблеми все ще залишаються невирішеними через мінливість шаблонів мережевого трафіку та сценаріїв вторгнень. У роботі здійснено аналіз даних мережевого трафіку для виявлення підозрілих мережевих дій (тобто вторгнень) за допомогою методів прогнозування часових рядів. У цьому дослідженні представлено підхід до прогнозування частот мережевих подій (типових і несанкціонованих) шляхом створення моделей прогнозу та оцінки ризиків атак. Вейвлет функції використані для визначення погодинних змін подій у мережі та визначення частоти подій у мережі, різка зміна якої може бути маркером несанкціонованих дій в мережі. Для прогнозування майбутніх подій мережевого трафіку використовується метод багатовимірних часових рядів, векторна автоматична регресія з екзогенними змінними (VARX). Ризики атак для мережевих подій оцінюються за допомогою адаптивного порогового методу та оцінюються шляхом виконання класифікації за допомогою двох методів машинного навчання. Проведено порівняльну оцінку між різними часовими шкалами (одна секунда, п'ять секунд і п'ятнадцять секунд) і трьома вейвлетами у визначенні ризиків атаки. Моделі з VARX демонструють можливість аналізу багатовимірних даних часових рядів мережевого трафіку для прогнозування майбутніх мережевих подій і оцінки ризиків їх атак. Доведено ефективність запропонованої нами моделі прогнозування для прогнозування частоти мережевих подій на одну годину вперед. Під час оцінки частоти нормальних подій в мережі не спостерігалося суттєвої різниці в продуктивності з точки зору точності прогнозування. Подібні результати були отримані при оцінці передбачуваних частот атак з трьома різними часовими масштабами.

Ключові слова: *пастка для хакерів, багатовимірний аналіз часових рядів, оцінка ризику атаки, Honeypot, вейвлет-перетворення, вторгнення в мережу.*

Постановка проблеми. DDoS-епідемія залишається серйозною проблемою для компаній в усьому світі. Дослідники журналу Cybersecurity Ventures прогнозують щорічну глобальну шкоду від кіберзлочинності 10.5 трильйонів доларів до 2025 року. Хакери атакують як великі проекти, так і маловідомі сайти. Причини різні: від випадковості, коли зловмисники тренують свої сили перед масштабним нападом, до цілеспрямованих зловмисних дій з метою вивести з ладу конкретний веб-ресурс.

Моніторинг і прогнозування мережевих подій є обов'язковими для встановлення вторгнень в мережу, а також щоб зрозуміти майбутні тенденції атак. Аналіз мережевого трафіку становить особливу проблему через постійну зміну природи мережевої діяльності в часі. Аналіз часових рядів може відігравати важливу роль у визначенні важливих атрибутів атак під час перевірки мережевої діяльності. Однак він може бути непридатним для безпосереднього аналізу даних мережевого трафіку, враховуючи, що події мережевого трафіку зазвичай відбуваються як серія послідовних спостережень у межах однієї часової позначки. При цьому моніторинг раптових змін у мережі з плином часу може слугувати ключовою характеристикою для виявлення подій атаки.

У сучасних мережевих середовищах за лічені секунди генерується величезна кількість мережевих подій. Отже, задачі аналізу мережевих подій вимагають значних зусиль, особливо при роботі з великою кількістю мережевих змінних.

Аналіз останніх досліджень і публікацій. Наразі запропоновано кілька підходів для моніторингу та виявлення підозрілих або втручальних дій у мережі. Однак точно визначити загрози все ще складно через різноманітні шаблони атак, які постійно змінюються. Останнім часом технологія обману (пастки) привертає все більше уваги завдяки можливості збору та аналізу різноманітних даних для розуміння моделей поведінки зловмисників. Система Honeypot є широко використовуваною оманливою технікою [2–5] у безпеці мережі. Будучи обчислювальним середовищем-приманкою, воно спонукає зловмисників продемонструвати їхні моделі атак, методи та використовувані інструменти [6].

Існує багато досліджень, які використовували технологію обману для моніторингу подій у мережі та прогнозування можливих майбутніх атак. У [7] запропонували процес аналізу даних стохастичної кібератаки, отриманих у honeypot, із шкалою часу 1 хвилина. Також продемонстровано можливість застосування статистичних властивостей, довгострокової залежності (LRD), для виявлення кібератак із хорошою продуктивністю. У [8] представлено проактивну систему безпеки для прогнозування атак розподіленої відмови в обслуговуванні (DDoS). У [9] запропонували техніку прогнозування коливань (тобто збільшення/зменшення) подій атак за допомогою байєсівського висновку. У [10] описаний гібридний підхід прогнозування вторгнень із аналізом часових рядів, аналізом даних та імовірнісним моделюванням. Використане експоненціальне згладжування, кластеризацію та ланцюг Маркова для одночасної оцінки результатів і сповіщення, якщо більше ніж два методи визначають мережеві події як атаки.

Попри численні дослідження для виявлення мережевих аномалій або інтрузивних мережевих подій деякі дослідницькі проблеми все ще залишаються невирішеними через постійну зміну шаблонів мережевого трафіку та сценаріїв вторгнень, зокрема щодо виявлення мережевих аномалій при збереженні високої продуктивності та вдосконаленні можливостей виявлення атак шляхом розуміння моделей атак з часом. Зловмисники часто надсилають велику кількість мережевих підключень або кілька послідовних мережевих підключень, щоб вторгнутися в обчислювальні мережі. Таким чином, аналіз кожного мережевого підключення (тобто мережевої події) окремо може бути некорисним для виявлення втручання атакувальними діями. Для усунення цього обмеження надзвичайно важливим є розуміння варіацій шаблонів атак у часовій області. Однак аналіз часових рядів мережевих атак і моделей поведінки зловмисників широко не вивчався.

Попередні дослідження показали важливість прогнозування можливих майбутніх атак шляхом включення різних методів аналізу даних часових рядів, дослідження мають обмеження, які пов'язані з непроведенням аналізу повного спектру великомасштабних даних мережевого трафіку та різних часових масштабів.

Інші обмеження:

1) проведено недостатня кількість досліджень для прогнозування мережевих атак у мережі honeypot;

2) попередні дослідження використовували однофакторні змінні для прогнозування атак без урахування частоти як звичайних подій, так і подій атаки з часом;

3) не було проведено дослідження для оцінки ризиків атак.

Щоб усунути ці обмеження, наш підхід зосереджений на оцінці ризиків атак за допомогою багатовимірних часових рядів шляхом аналізу даних за тривалий період часу.

Постановка завдання. Метою статті є дослідження застосування методів аналізу часових рядів для прогнозування мережевих подій, побудови моделей атак на основі досліджень та аналізу мережевого трафіку; а також визначення оцінки ризиків атак на основі частот нормальних та несанкціонованих мережевих подій у різних масштабах часу.

Виклад основного матеріалу. У мережевих середовищах набір упорядкованих за часом мережевих подій записується в певний час t_i. Ці дані утворюють часовий ряд – послідовність даних, що складається з послідовних вимірювань, зроблених протягом регулярного інтервалу часу. Аналіз часових рядів здійснює прогнозування рядів даних, які зазвичай не є детермінованими, тобто містять випадковий компонент. Тому аналіз часових рядів мережевих подій можна використовувати для характеристики поведінки атак. Авторегресія (AR), ковзне середнє (MA) та авторегресійне ковзне середнє (ARMA) – це методи, які зазвичай використовуються для аналізу різних типів даних часових рядів. У контексті виявлення вторгнень в [11] запровадили статистичну модель часових рядів, інтегровану в систему кіберзахисту. Вони використовували узагальнене авторегресійне ковзне середнє (GARMA), щоб передбачити частоту атак на години наперед.

Важливим є прогнозування можливих майбутніх атак, оскільки це може дати системним адміністраторам достатньо часу для підготовки до можливих атак шляхом зміни конфігурацій захисту. У деяких працях [12] здійснено аналіз часових рядів даних щодо мережевих атак, зібраних у приманках. Використано метод Бокса-Дженкінса для розробки прогнозних моделей для прогнозування атак. Встановлено, що модель AR(1) і початкове завантаження на основі моделі AR(1) може бути використано для прогнозування атак. Виявлено, що передбачувані атаки збігаються з реальними атаками з ймовірністю понад 95 %.

Формально часовий ряд може бути представлений як $S = \{s_1, s_2, \ldots, s_n\}$, де S – часовий ряд, s_i – зареєстроване значення змінної S у момент часу t_i, а n – кількість спостережень.

Часові ряди мають деякі характеристики, які ускладнюють аналіз даних: великий об'єм, висока розмірність, ієрархія та багатовимірність.

Кількість спостережень у часовому ряді трафіку часто може бути надзвичайно високою, іноді коливатись від порядку сотень чи тисяч до порядку мільйонів чи мільярдів. Великий об'єм даних створює проблему для аналізу даних і алгоритмів інтелектуального аналізу, оскільки більші бази даних потребують більше часу для аналізу даних і методів інтелектуального аналізу для доступу до даних і виконання обчислень.

Висока розмірність є ще однією характеристикою даних часових рядів спостереження трафіку. Під час пошуку подібностей та аномалій під час аналізу даних часових рядів виникає проблема «прокляття розмірності», коли часовий ряд відображається на k-вимірному просторі, де k – кількість часових точок.

Іншою особливістю даних часових рядів є їх ієрархічний характер. Часовий ряд можна аналізувати за його основною часовою ієрархією.

Останньою характеристикою даних часових рядів є багатовимірний характер деяких даних. Аналіз даних часових рядів часто вивчає одну змінну, але іноді має справу з даними часових рядів, що складаються з кількох пов'язаних змінних.

Запропонований нами підхід до аналізу даних мережевого трафіку для виявлення атак складається з трьох основних частин.

1. Дані мережевого трафіку (D) перетворюються на часовий ряд, масштабуються. Потім застосовуються виділення ознак часового ряду і вибір апроксимаційних функцій.

2. З обраними функціями генеруються моделі прогнозування мережевих подій.

3. За допомогою моделей оцінюються ризики атак.

Етап перший: перетворення даних у масштабований часовий ряд, обчислення характеристик ряду та виділення апроксимаційнийх функцій.

Так як багато подій в мережі відбуваються одночасно і проявляються як серія подій, то дані мережевого трафіку повинні бути перетворено на цільову шкалу часу з узгодженими часовими інтервалами. Існують різні підходи для цього перетворення, такі як агрегація або віднімання під час перетворення даних мережевого трафіку у формат часових рядів.

Ми пропонуємо здійснювати перетворювання у часовий ряд шляхом обчислення різниці між сумою та медіаною вхідних даних (тобто оригінальних даних) у кожній змінній у момент часу t.

Тобто, з цільовим масштабом часу (t_s), усі відліки

$$\Omega^t = \{\omega^t_1, \omega^t_2, \ldots, \omega^t_n\},$$

де $\{\omega^t_i | \ \omega^t_i \subset D, t < D_{ts} \le t + t_s\}$) у діапазоні часу ($t \sim t + t_s$) вибираються з вхідних даних D, де D_{ts} вказує часову позначку події мережі, а $\omega^t \cap \omega^{t+ts} = \varnothing$.

З вибраними екземплярами нові дані часового ряду з цільовою шкалою часу

$$\Gamma = \{u^1, u^2, \ldots, u^m\},$$

де $m = \max(D_{ts})/t_s$, обчислюються шляхом вимірювання різниць ($\varphi(\Omega^t)$).

N^t вказує загальну кількість екземплярів у Ω^t.

$$u^t = \frac{\varphi\left(\copyright^t\right)}{N^t}.$$

Оскільки жодні дослідження не запропонували оптимальний часовий масштаб для аналізу мережевих подій, доцільно провести аналіз із різними часовими масштабами, щоб знайти оптимальний цільовий часовий масштаб для оцінки даних мережевого трафіку.

Після перетворення вхідних даних у масштабований часовий ряд ми будуємо описові функції за допомогою дискретного Вейвлет-перетворення (DWT), яке забезпечує збір інформації про час і частоту дискретного ряду на основі нескінченного набору можливих базових функцій (так званих вейвлетів). Коефіцієнти деталізації DWT відображають високочастотні компоненти. Загалом, коефіцієнти апроксимації використовуються для визначення основних тенденцій, тоді як коефіцієнти деталізації корисні для виявлення відхилень. З цієї причини коефіцієнти деталізації широко використовуються для виявлення будь-яких раптових змін.

Техніка вибору ознак для детектування нормальних подій в мережі та атак.

Наступним кроком необхідно вибрати ознаки часового ряду, які б дозволили б статистично значущо оцінити відмінність частот нормальних мережевих подій та атак.

Припустимо, що задана вихідна послідовність роботи мережі складається з серії спостережень часової послідовності $O = \{(T_i, X_i, Y_i)\}$, $i = 1,2,\ldots,N$ містить події мережевого трафіку (N – загальна кількість подій). T позначає час, X позначає змінні вихідної послідовності мережевого трафіку з номінальними, дійсними числами та двійковими змінними, а $Y \in \{\text{normal, attack}\}$ вказує на мережеві події.

Дані часових рядів створюються шляхом відображення набору рядів даних мережевого трафіку t_s до значень у часі шляхом виділення часових ознак.

По-перше, вихідні дані сегментуються за попередньо визначеною шкалою часу t_s. Новий індекс часу створюється залежно від попередньо визначеної шкали часу t_s через деякий час

$$\nabla t_i = (t_s \cdot c) - \nabla t_{i-1},\ c = 2, \cdots, t_N,\ i = 2, \cdots, t_N,\ \nabla t_i = t_{s,}$$

де $t_N = m_t / t_s$

m_t – це максимальний час.

Таким чином, новий індекс часу генерується як $\nabla t_i = \{t'_1, t'_2, \cdots, t'_N\}$.

У кожному часовому відліку ∇t_i містить серію кортежів $\{(X_i, Y_i)\}$, формуючи матрицю M × J ($X_i \in R^{M \times J}$) та матрицю M × D ($Y_i \in R^{M \times D}$), де M ($M \geq 1$) вказує загальну кількість спостережень за певний час ∇t_i, J – загальна кількість змінних, і D ($D \geq 1$) – розмір залежних змінних. Важливо зазначити, що розмір M може змінюватися, оскільки кількість мережевих подій, що відбуваються в часі, різна.

Для залежних змінних з одним гарячим кодуванням частота мережевих подій понад ∇t_i обчислюється як

$$C(Y^k_i) = \Sigma_1^n I_{Yi}(\delta_i),$$

де δ_i – вказує, чи кожна подія мережі є звичайною чи атакою. Частота кожної однократно закодованої змінної понад ∇t_i також визначається для номінальних змінних. Наприклад, для змінних (номера порту джерела та абонента) кількість використаних номерів портів понад ∇t_i визначаються. Для інших змінних, репрезентативне значення кожного розміру M вектор ∇t_i визначається як часовий ряд з рівними інтервалами.

Після виділення ознак часового ряду наступним кроком є виконання кореляційного аналізу (векторна авторегресія) і рангове перетворення, щоб оцінити та знайти найкращий набір ознак (W_ρ), що описує релевантність між атаками та звичайними подіями. Зокрема, ми використовуємо рангову кореляцію Спірмена, яка не вимагає жодних припущень щодо розподілу даних.

Потім проводимо рангову трансформацію для відбору найкращих можливих ознак для диференціювання мережевих подій (тобто нормальних) відносно атак. Рангова трансформація підвищує чутливість методик при ненормальному розподілі даних.

Отримання математичної моделі для прогнозування.

Модель прогнозування створюється для прогнозування k-наперед прогнозованих значень частот мережевих подій, які відносяться до ендогенних змінних.

Нормальна частота подій (NEF) і частота атак (AEF) вимірюються для відображення загальної кількості звичайних подій і подій атаки відповідно. Однією з головних проблем у використанні багатовимірних часових рядів є визначення методу, який фіксує залежності між кількома змінними.

Для прогнозування майбутніх значень шляхом знаходження причинно-наслідкових зв'язків між кількома змінними з часом використовуються векторна авторегресія (VAR) і векторна авторегресія з екзогенними змінними (VARX).

В якості альтернативного методу для аналізу різних типів наборів даних часових рядів нами запропоновано застосування VARX. Даний статистичний метод показав в дослідженнях високу ефективність при виконанні багатовимірного аналізу часових рядів.

При генерації моделі VARX ($\boldsymbol{f_h^d}$) необхідно виконати перевірку на стаціонарність, щоб переконатися в стаціонарності змінних. Тест на одиничний корінь є стандартним методом перевірки стаціонарності.

Для визначення наявності одиничних коренів ми використовуємо критерій Дікі-Фуллера (ADF). Крім того, даний метод також може перевіряти графік автокореляційної функції (ACF); якщо ACF розпадається дуже повільно, можна робити висновки, що дані не стаціонарні.

Оцінка ризику атак на мережу.

Враховуючи мережеві події в момент часу t_i, $i = 1, 2, \ldots, T$, існує e_i кількість мережевих подій (звичайна/атака) у час t_i, де $e_i \geq 0$.

Наприклад, $e_i = 12$ у момент часу t_i вказує, що в момент часу t_i було помічено 12 мережевих подій. Кожну мережеву подію можна визначити як звичайну або як атаку за допомогою різних методів виявлення вторгнень. Якщо не всі мережеві події класифіковані як звичайні або атакуючі, то визначити вразливість мережевих подій у момент часу t_i непросто.

Однак оцінка ризику атаки на кількох рівнях може стати в нагоді експертам із безпеки під час вивчення можливих загроз і визначення відповідних пріоритетів захисту.

Ми пропонуємо визначати три рівні ризику атаки (низький, середній і високий) із зазначенням різних пріоритетів. Порогові значення, що розділяють рівні, визначаються емпіричною інтегральною функцією розподілу (ECDF). За допомогою ECDF можна отримати розподіл даних, викликаний наявністю різних субпопуляцій у даних.

ECDF використовується для вимірювання кумулятивних розподілів для різниці (x_i) між прогнозованими значеннями норми та атаки, пов'язаних із кількістю сеансів.

Тобто як

$$x_i = \frac{\hat{y}_i^n}{s_i} - \frac{\hat{y}_i^a}{s_i},$$

де $\hat{y}_i^n$ і та $\hat{y}_i^a$ – прогнозована нормальна частота та частота нападів відповідно,

s_i – кількість екземплярів даних у кожному сеансі.

$$\tau_н = F(\theta_0),$$

$$\tau_в = F(\theta_1)$$

де $\tau_н$ та $\tau_в$ – верхня та нижня межі для оцінки рівнів ризику атаки відповідно;

$F(\cdot)$ – інтегральна функція розподілу ECDF.

Для θ_0 і θ_1 ми використали 25-й і 75-й процентилі відповідно.

Перевірка працездатності розробленого підходу.

Для дослідження методу прогнозування нам необхідні дані моніторингу трафіку комп'ютерної мережі. Для цього було використано загальнодоступний набір даних Кіотського протоколу (https://www.takakura.com/Kyoto_data/). Набір даних включає 365 файлів (щоденні дані про мережевий трафік) з січня до грудня. Для перетворення даних мережевого трафіку в цільовий масштаб часу використовується змінна «час початку». Змінна 'label' вказує, чи є сеанс атакою чи ні, і він використовується для вимірювання частоти подій шляхом підрахунку загальної кількості звичайних подій і подій атаки.

Оскільки природа мережевого трафіку за своєю суттю є випадковою, у час t_i може виникати велика кількість подій мережевого трафіку, тоді як у певний час трафіку не виникає взагалі.

Застосовуючи підхід, описаний вище, цей набір даних було перетворено на цільову шкалу часу. Зокрема, перетворення даних було виконано з трьома різними масштабами часу (t_s): одна секунда, п'ять секунд і п'ятнадцять секунд, і ми порівняли їх продуктивність. Даний масштаб було обрано інтуїтивно, тому що дослідження щодо ефективної величини масштабування відсутні.

Дискретне Вейвлет перетворення (DWT) застосовано для вилучення ознак із даних мережевого трафіку з рівнем декомпозиції (j = 3). При застосуванні DWT вибір відповідного сімейства вейвлетів відіграє важливу роль, оскільки сімейства вейвлетів використовують різні базисні функції для створення набору коефіцієнтів.

Оскільки чітко не досліджено, який вейвлет підійде для аналізу даних мережевого трафіку, ми дослідили три різні материнські вейвлети та порівняли їх подібності та відмінності.

Три вейвлети: симлет symN (Добеші лінійно-асиметричний) ("s6"), найкраще локалізований (Best-localized Daubechies) ("l4") і койфлет ("c6").

За допомогою вейвлетів було виділено тридцять одну функцію. Потім було виконано метод Квятковського-Філліпса-Шмідта-Шина (KPSS) для підтвердження припущення про стаціонарність часового ряду. Оскільки деякі особливості вейвлетів не задовольняли стаціонарне припущення, для їх перетворення було застосовано диференціювання першого порядку. Повторно було проведено статистичну перевірку, щоб визначити, чи відповідають перетворені ознаки стаціонарному припущенню. Виявлено, що всі трансформовані ознаки задовольняють припущення про стаціонарність з рівнем значущості 95 %.

Далі було виконано вибір функцій, щоб визначити ключові характеристики, що представляють мережеві події.

Зокрема, наш підхід був застосований для визначення статистично значущих ознак для аналізу нормальних частот і частот атак.

І останнім етапом є отримання моделей VARX з використанням вибраних функцій як екзогенних змінних і частот мережевих подій як ендогенних змінних для вивчення взаємодії між змінними.

Дані моделі були використані для виконання прогнозу частоти мережевих подій на k годин вперед. Погодинні моделі VARX були створені з використанням даних за першу годину для прогнозування погодинних значень NEF (частота нормальних подій) і AEF (частота атак) для наступного 23-годинного періоду кожного дня.

Статистичну різницю між фактичними та прогнозованими значеннями перевіряли шляхом застосування знакового рангу Вілкоксона. Як непараметричний тест, він перевіряє будь-які значні відмінності даних. Р-значення більше 0,05 вказує на відсутність істотної різниці нижче 95 % рівня значущості.

При t_s = 1 с значуща різниця між фактичними і прогнозованими значеннями NEF і AEF відсутня.

При t_s = 5 с лише прогноз з вейвлет "s6" (p-value = 0,1316) для значень NEF і модель з вейвлет "c6" (p-value = 0,9455) для значень AEF не показали істотної різниці.

При t_s = 15 с модель з вейвлетом "c6" (р-значення = 0,7796) для значень NEF і модель з вейвлетом "c6" (p-value = 0,4070) і "s6"

(p-value = 0,3296) для значень AEF не показали істотної різниці між прогнозованими та фактичними значеннями.

Аналізуючи односекундний прогноз значень AEF, ми виявили, що моделі мають подібні тенденції зростання і спадання (рис. 3, 4а). Крім того, ми виявили відмінну різницю між моделями прогнозування NEF і AEF. Кожного разу, коли часову шкалу збільшували, у прогнозах NEF зберігався відповідний характер зміни прогнозованих частот відповідно до фактичних. Однак, закономірності зникли для прогнозів AEF. Це вказує на те, що часовий масштаб сильно впливає на якість прогнозу. Чим менше значення часового масштабу, тим ближчі прогнозовані значення до фактичних. Зі зменшенням значенням часового масштабу зникає вплив використовуваних вейвлетів, використаних для побудови моделі прогнозування.

Отже, мінімальне значення часового масштабу позитивно впливає для прогнозування і вивчення подій атаки, тоді як використання довшої шкали часу (тобто п'ять або п'ятнадцять секунд) корисне для дослідження звичайних подій в мережевому трафіку.

Оскільки було згенеровано декілька прогностичних моделей з використанням різних вейвлетів, ми також провели тест статистичної перевірки Diebold-Mariano Test (DM), щоб порівняти їх прогнозну точність.

Залишки між фактичним і прогнозованим значенням (e_i) від кожної моделі були розраховані з 95 % рівнем довіри на основі залишків співвідношення (тобто відсотків) статистичних відмінностей продуктивності. У часовому масштабі ($t_s = 1$ с) спостерігалося приблизно 80 % і 70 % залишків для NEF і AEF у січні, і суттєвих відмінностей між моделями не виявлено. Подібні результати спостерігалися в різних часових масштабах. Ми виявили, що шкала часу в одну секунду показала кращі результати, ніж шкала часу в п'ять або п'ятнадцять секунд.

Кумулятивну функцію розподілу (CDF) значень абсолютних помилок було згенеровано для оцінки моделей з різними вейвлетами. Суттєвої різниці в продуктивності серед вейвлетів із різними масштабами часу не було виявлено. Але ми виявили кращу ефективність прогнозування з односекундною шкалою часу. Крім того, вейвлет "с6" показав відносно кращу продуктивність у прогнозуванні NEF. Всі три вейвлети показали майже однакову продуктивність у прогнозуванні AEF. Нарешті, ми встановили, що спостерігалося збільшення абсолютних похибок із довшими часовими масштабами.

Продуктивність моделі прогнозування для різних вейвлетів також вимірювалася шляхом обчислення різниці між фактичними та прогнозованими значеннями за допомогою тесту DM.

Хоча істотної різниці між вейвлетами не спостерігалося, результат між с6 і l4 показав дещо вищу статистичну значущість ($p < 0,05$), ніж інші в деякі місяці. Подібні результати були встановлені при проведенні тесту DM з різними часовими масштабами. Загалом, не було значної різниці в аналізі NEF і AEF з різними вейвлетами протягом більшості місяців.

Висновки. Наші дослідження встановили можливість застосування теорії аналізу часових рядів для прогнозування трафіку комп'ютерних мереж. Дане прогнозування може бути використане як інструмент визначення ризику майбутніх несанкціонованих втручань в мережу.

В роботі запропонований новий підхід до прогнозування ризиків кібератак на основі даних моніторингу трафіку мереж. Даний підхід складається з послідовних 3 етапів: перетворення трафіку КМ в багатовимірний часовий ряд шляхом масштабування з певним кроком за часом, застосування вейвлет-перетворення та регресійного аналізу з екзогенними змінними.

Для визначення оптимального кроку масштабування під час аналізу даних мережевого трафіку здійснені дослідження для різних часових шкал. Хоча масштаб часу може бути обрано довільною величино в даному підході, встановлено, що збільшення часу масштабування обернено пропорційне ефективності класифікації оцінки ризику атаки.

Застосування вейвлет-перетворення дозволило отримати вейвлет-функції для визначення погодинних змін подій у мережі та оцінити частоту подій у мережі, різка зміна якої може бути маркером несанкціонованих дій в мережі. Під час генерації погодинних моделей прогнозування були використані різні вейвлет-функції через постійну зміну моделей мережевого трафіку.

Встановлено, що різні типи вейвлетів давали статистично значущі результати залежно від визначеного масштабу часу.

Зі шкалою часу в одну секунду застосування трьох типів вейвлетів давали статистично значущо однакові результати щодо продуктивності і точності прогнозу.

Однак, статистично значуща різниця спостерігалася при прогнозуванні ризиків на наступні місяці при п'яти- та п'ятнадцяти секундній шкалах часу. Це може бути пояснено тим, що

загальний об'єм мережевого трафіку (зокрема, аномальні події в мережі) змінюється щомісяця. Додатковий аналіз помісячного мережевого трафіку показав, що великий об'єм мережевого трафіку часто проявляється як високий сплеск при оцінці ризиків атаки, особливо в часовому масштабі п'ять і п'ятнадцять секунд.

Доведена ефективність запропонованої нами моделі прогнозування для прогнозування частоти мережевих подій на одну годину вперед. Під час оцінки частоти нормальних подій в мережі не спостерігалося суттєвої різниці в продуктивності з точки зору точності прогнозування. Подібні результати були отримані при оцінці передбачуваних частот атак з трьома різними часовими масштабами.

Порівнюючи показники класифікації між прогнозованими та фактичними значеннями частот мережевих подій для визначення ризику атаки, відмічені деякі відмінності в точності.

Встановлено, що тип вейвлетів не дає значущої різниці в продуктивності класифікації ризиків атак. Але, шляхом статистичного аналізу, виявлено, що вейвлети c6 і s6 були визначені як значущі вейвлет-характеристики для аналізу даних мережевого трафіку.

Список літератури:

1. Abdullah A. Intrusion detection forecasting using time series for improving cyber defence / A. Abdullah, T. R. Pillai, L. Z. Cai. *International Journal of Intelligent Systems and Applications in Engineering,* 3 (1), 28–33.
2. Ahmed M. A survey of network anomaly detection techniques / M. Ahmed, A. N. Mahmood, J. Hu. *Journal of Network and Computer Applications,* 60, 19–31.
3. Artail H. A hybrid honeypot framework for improving intrusion detection systems in protecting organizational networks / H. Artail, H. Safa, M. Sraj, I. Kuwatly, Z. Al-Masri. *Computers & security,* 25 (4), 2006, 274–288.
4. Awad M. Support Vector Machines for Classification / M. Awad, R. Khanna. Apress, Berkeley, CA, Ch. 3, 2015, pp. 39–66.
5. Barford P. A signal analysis of network traffic anomalies / P. Barford, J. Kline, D. Plonka, A. Ron. *Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment*. 2002, pp. 71–82.
6. Beliakov G. Aggregation Functions: A Guide for Practitioners, 1st Edition / G. Beliakov, A. Pradera, T. Calvo. Springer Publishing Company, Incorporated, 2008.
7. Bernacki J. Anomaly detection in network traffic using selected methods of time series analysis / J. Bernacki, G. Kołaczek. *Computer Network and Information Security,* 9, 2015, p. 10–18.
8. Besharati E. Lr-hids: logistic regression host-based intrusion detection system for cloud environments / E. Besharati, M. Naderan, E. Namjoo. *Journal of Ambient Intelligence and Humanized Computing*, 10 (9), 2019, 3669–3692.
9. Boto-Giralda D. Wavelet-based denoising for traffic volume time series forecasting with self-organizing neural networks / D. Boto-Giralda, F. J. D´ıaz-Pernas, D. Gonzalez-Ortega. *Computer-Aided Civil and Infrastructure Engineering,* 25 (7), 2010, p. 530–545.
10. Bouzoubaa K. Comparative study of features selection methods: Case of denial of service attacks forecasting / K. Bouzoubaa, Y. Taher, B. Nsiri. *International Conference on Algorithms, Computing and Systems*, 2020. pp. 40–44.
11. Bouzoubaa K. Dos attack forecasting: A comparative study on wrapper feature selection / K. Bouzoubaa, Y. Taher, B. Nsiri. *International Conference on Intelligent Systems and Computer Vision*, 2020. IEEE, pp. 1–7.
12. Box G. E. Time series analysis: forecasting and control / G. E. Box, G. M. Jenkins, G. C. Reinsel, G. M Ljung. John Wiley & Sons, 2015.
13. Cao J. An improved network traffic classification model based on a support vector machine / J. Cao, D. Wang, Z. Qu, H. Sun, B. Li, C.-L. Chen. *Symmetry,* 12 (2), 2020, p. 301.
14. Celenk M. Anomaly prediction in network traffic using adaptive wiener filtering and arma modeling / M. Celenk, T. Conley, J. Graham, J. Willis. *IEEE International Conference on Systems, Man and Cybernetics. IEEE*, pp. 2008, p. 3548–3553.
15. Cortez P. Multi-scale internet traffic forecasting using neural networks and time series methods / P. Cortez, M. Rio, M. Rocha, P. Sousa. *Expert Systems,* 29 (2), 2012, p. 143–155.
16. Curiac D.-I. Malicious node detection in wireless sensor networks using an autoregression technique / D.-I. Curiac, O. Banias, F. Dragan, C. Volosencu, O. Dranga. *International Conference on Networking and Services* (ICNS'07). 2007, IEEE, pp. 83–83.
17. Winter J. C. Comparing the pearson and spearman correlation coefficients across distributions and sample sizes / J. C. Winter, S. D. Gosling, J. Potter. A tutorial using simulations and empirical data. *Psychological methods,* 21 (3), 2016,p. 273.
18. Diebold F. X. Comparing predictive accuracy / F. X. Diebold, R. S. Mariano. J*ournal of Business & Economic Statistics,* 20 (1), 2002, pp. 134–144.

19. Dongxia L. An intrusion detection system based on honeypot technology / L. Dongxia, Z. Yongbo. *2012 international conference on computer science and electronics engineering*. Vol. 1. IEEE, 2012, pp. 451–454.

20. Huang C.-T. Wavelet-based real time detection of network traffic anomalies / C.-T. Huang, S. Thareja, Y.-J. Shin. *Securecomm and Workshops*. IEEE, 2006, pp. 1–7.

21. Huang P. A non-instrusive, wavelet-based approach to detecting network performance problems / P. Huang, A. Feldmann, W. Willinger. *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement,* 2001, pp. 213–227.

22. Iglesias F. Analysis of network traffic features for anomaly detection / F. Iglesias, T. Zseby. *Machine Learning,* 101 (1), Oct 2015, pp. 59–84.

Moyseenko O.V. PREDICTION OF CYBER ATTACKS BASED ON TRAFFIC INTENSITY MONITORING IN COMPUTER NETWORKS

Numerous studies of network anomalies or intrusive network events do not allow solving some network security problems. Research challenges still remain unsolved due to the variability of network traffic patterns and intrusion scenarios. The paper analyzes network traffic data to detect suspicious network activities (ie, intrusions) using time series forecasting methods. This study presents an approach to predicting the frequency of network events (typical and unauthorized) by building prediction models and assessing the risks of attacks. Wavelet functions are used to determine the hourly changes of events in the network and to determine the frequency of events in the network, a sudden change of which can be a marker of unauthorized actions in the network. A multivariate time series method, vector automatic regression with exogenous variables (VARX), is used to predict future network traffic events. Attack risks for network events are assessed using an adaptive threshold method and evaluated by performing classification using two machine learning methods. A comparative assessment was made between different time scales (one second, five seconds and fifteen seconds) and three wavelets in determining attack risks. Models with VARX demonstrate the ability to analyze multidimensional network traffic time series data to predict future network events and assess the risks of their attacks. The effectiveness of our proposed forecasting model for predicting the frequency of network events one hour ahead has been proven. When evaluating the frequency of normal events in the network, no significant difference in performance was observed in terms of prediction accuracy. Similar results were obtained when estimating predicted attack frequencies with three different time scales.

Key words: *Deception, Multivariate time series analysis, Attack risk estimation, Honeypot, Wavelet Transform, Network intrusion.*