Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 4 (26)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

УДК 623.618.51
DOI: https://doi.org/10.30837/ITSSI.2023.26.145

К. Бондаренко

АНАЛІЗ І ВИБІР РЕЛЕВАНТНОЇ МЕТРИКИ ВИЯВЛЕННЯ МЕРЕЖНИХ АНОМАЛІЙ

Метою дослідження є аналіз основних аспектів виявлення аномалій мережі та метрик їх оцінювання, що дає змогу вчасно виявляти кібератаки на мережу та значно підвищити рівень її безпеки. **Об'єктом дослідження** є виявлення мережних аномалій. **Завдання дослідження**: сформулювати принципи, що дозволяють здійснити узагальнення різних методів виявлення аномалій; проаналізувати метрики аномалій, зважаючи на заходи близькості формування оцінки поточного стану безпеки; обґрунтувати вибір релевантної міри близькості виявлення мережних аномалій. Сформульовано принципи, що дають змогу узагальнити різні методи виявлення аномалій. Для класифікації та полегшення виявлення мережних аномалій запропоновано метрики, що ґрунтуються на мірах близькості для типів даних, які характеризують аномалії. Визначено компоненти, що характеризують зазначену проблему, а саме: типи вхідних даних, прийнятність заходів близькості, маркування даних, класифікація методів, що ґрунтуються на використанні розмічених даних, виявлення відповідних особливостей та повідомлення про аномалії. Описано підхід, що дає змогу вчасно сформувати необхідний набір метрик, який забезпечить не тільки формування превентивних заходів протидії, а й дозволить оцінювати поточний стан системи безпеки загалом. Крім цього, забезпечується можливість формування багатоконтурних систем безпеки з огляду на вплив (комплексування) цільових (змішаних) атак на елементи інфраструктури, а також здатність їх синтезу з методами соціальної інженерії. **Висновки**: сформульовано принципи, що дають змогу виконати узагальнення різних методів виявлення аномалій; наведено види, показники та приклади мережних аномалій; запропоновано міри близькості для числових, категоріальних і змішаних типів даних з метою полегшення виявлення мережних аномалій; обґрунтовано вибір міри близькості Махаланобіса як основи метрики аномалій.

Ключові слова: мережна аномалія; система виявлення вторгнення; міра близькості; класифікація атак.

Вступ

Виявлення аномалій є важливою здатністю будь-якої схеми класифікації сигналів. Зважаючи на те, що ми ніколи не зможемо навчити систему машинного навчання на всіх класах об'єктів, з даними яких система може зіткнутися, стає важливим, щоб вона могла розрізняти інформацію про відомі та невідомі об'єкти під час тестування. Виявлення мережних аномалій є надзвичайно складним завданням. Саме з цієї причини існує кілька моделей виявлення аномалій, що добре зарекомендували себе на різних даних. Цілком очевидно, що не існує єдиної найкращої моделі для виявлення мережних аномалій, і успіх залежить не тільки від типу використовуваного методу, а й від статистичних властивостей даних, що обробляються. Попри значний прогрес та великий обсяг роботи все ще існує чимало можливостей для розвитку сучасних технологій виявлення та запобігання мережних атак. Спроба вторгнення чи загроза – це навмисна та несанкційна спроба (I) отримати доступ до інформації, (II) маніпулювати нею або (III) зробити систему ненадійною чи непридатною для використання. Наприклад, (а) атака типу "відмова в обслуговуванні" (DoS) намагається позбавити хост ресурсів, необхідних для правильного функціювання під час оброблення; (б) хробаки та віруси застосовують інші хости з допомогою мережі; (в) зломи дають змогу отримати привілейований доступ до хосту, скориставшись перевагами відомих вразливостей.

Термін "виявлення вторгнень у мережі на основі аномалій" належить до проблеми визначення виняткових закономірностей у мережному трафіку, які не відповідають очікуваній нормальній поведінці. Ці невідповідні закономірності часто називають аномаліями, викидами, винятками, аберраціями, сюрпризами, особливостями чи суперечливими спостереженнями в різних галузях застосування. "Аномалії" та "викиди" – два терміни, що найчастіше використовуються в контексті виявлення вторгнень у мережі на основі аномалій.

Можна стверджувати, що швидке виявлення вторгнень (аномалій, відхилень від нормальної роботи) може забезпечити вчасне формування превентивних заходів та/або необхідний рівень безпеки [1].

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 4 (26)

2. Аналіз літературних джерел і постановка проблеми

Одним із типів аналізу даних, який шукає незвичайні стани в системі, є виявлення аномалій, також відомі терміни "виявлення викидів" або "виявлення подій". Алгоритми виявлення аномалій є контрольними точками вхідного трафіку на різних етапах – від рівня мережі до центру оброблення інформації. В останньому випадку існує висока потреба в надійному виявленні для очищення даних [2] та цілей класифікації [3].

Виявлення аномалій у роботі мережі має широкий спектр застосунків, деякі з яких більш розвинені (наприклад, мережна безпека), а інші мають потенціал для зростання.

Аномалія – це точка даних, яка не пов'язана з прогнозованою поведінкою в системі, що моделюється. Аномалії – це рідкісні події або спостереження, які значно відхиляються від звичайної поведінки чи закономірностей, що спостерігаються в одній точці даних, у певному контексті або інтервалі часу (наприклад, сезон чи квартал) або весь набір даних. Здебільшого аномалії виникають унаслідок зовнішніх факторів, таких як відмова датчика або зовнішній напад, і мета алгоритму виявлення полягає в тому, щоб визначити, де відбулася аномалія, та класифікувати / обчислити причину. За умови бінарної класифікації аномалій вирішальне значення має апроксимаційна модель, яка найкраще відповідає очікуваній поведінці даних. Складність багатьох ситуацій потребує окремої стратегії виявлення кожного застосунку [4, 5].

У роботах [6, 7] подано чотири категорії методів виявлення мережних аномалій. Запропоновано класифікацію методів залежно від підходу до проблеми, способу застосування, типу методу та орієнтації алгоритму.

Описані в дослідженні [8] статистичні методи, зокрема метод мінімального обсягу, намагаються моделювати нормальні дані з використанням математичних моделей і розподілів. Підхід мінімального обсягу спрямований на створення *n*-вимірного симплекса навколо певної хмари даних (наземні справжні дані), де цільова функція полягає в тому, щоб мінімізувати займаний обсяг за умови максимізації точок наземних істинних даних. Аномалія визначається як будь-які дані, що не відповідають симплексу. У праці [9] наведено метод прогнозування, що називається експоненційним згладжуванням. Зазначений метод прогнозує майбутню точку даних, використовуючи попередні точки та параметр згладжування. Аномальні дані, отримані статистичними методами, – це ті, що відхиляються від установленої моделі. Традиційні геометричні та статистичні методи підкріплені значним обсягом досліджень і покладаються на глибоке розуміння процесів, що відбуваються. Автори наголошують на необхідності рішень машинного навчання на основі даних та глибокого навчання, які дають змогу здійснювати більш гнучкі модифікації класичного методу прогнозування.

Для підкатегорії методів, пов'язаних із моделями машинного навчання та глибокого навчання, у роботі [10] стверджується, що характер даних, які надаються, визначає вибір моделі. Наприклад, моделі з довгою короткостроковою пам'яттю (*LSTM*) і трансформатори віддають перевагу послідовному введенню даних, зокрема аудіо, відео та часові ряди [11]. До того ж у статті [12] зазначено, що згорткова нейронна мережа (*CNN*) і автоенкодер (AE) віддають перевагу непослідовним типам даних, зокрема введенню зображень. У студіях [13–15] увагу приділено алгоритмам, що намагаються розрізняти нормальну й аномальну поведінку, установлюючи межу прийняття рішення, наприклад, за допомогою класифікатора машини опорних векторів (*SVM*) [13] або майбутніх значень прогнозування потокових даних [14] з мережами *LSTM* [15]. У роботі [16] показано, що залежності від наявності міток навчання ці підходи бувають контрольованими, напівконтрольованими, самоконтрольованими або повністю не контрольованими.

У публікаціях виокремлено три способи категоризації аномалій: конструктивний, руйнівний та очищення даних. Конструктивні застосунки мають продуктивний характер. У дослідженні [17] виконано порівняння продуктивності багатошарового перцептрона (*MLP*) і *k*-найближчих сусідів (*KNN*) та класифікатори *SVM* для широкого кола застосунків. Інші статті описують використання навчання з підкріпленням для різних застосунків безпілотних літальних апаратів (БПЛА) [18], а в роботі [19] використовується підхід машинного навчання для застосунків "розумного будинку". На відміну від конструктивних застосунків, деструктивні призначені для порушення повсякденної роботи з метою отримання сумнівної фінансової вигоди, наміру завдати шкоди мережі та програмам,

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 4 (26) ISSN 2524-2296 (online)

що проходять крізь мережу Інтернету речей, або порушити критично важливі бізнес-практики. У дослідженні [20] розглядаються різні кібератаки Інтернету речей та останні розробки в сфері безпеки Інтернету речей.

У статті [21] описано один із найбільш відомих типів аномалій, що залежить від обставин, – точковий, контекстуальний і колективний. Прикладом може бути виявлення шахрайства з кредитними картками [22].

Затримка й масштабованість алгоритму виявлення визначають, чи виконуватиметься метод "на льоту" на етапі збору даних або на більш пізньому етапі зберігання. Онлайн-алгоритм може послідовно обробляти інформацію за допомогою однієї точки даних або вікна, не маючи доступу до всіх вхідних даних. Традиційні геометричні та статистичні онлайн-підходи передбачають застосування раніше згаданих методів, основаних на відстані, щільності та відхиленні, а також методи, основані на кутах. У роботі [23] наведено приклад онлайн-методу, в якому використовуються нечіткі *C*-середні, а в статті [24] розглянуто ансамблевий підхід для виявлення аномалій.

Отже, аналіз показав, що попри широту охоплення різноманітних методів, предметних сфер і завдань виявлення мережних аномалій менше уваги приділяється ключовому питанню – аналізу метрик мережних аномалій та обґрунтуванню вибору релевантної метрики в різних ситуаціях.

3. Мета й завдання дослідження

Метою цієї роботи є дослідження основних аспектів виявлення аномалій мережі, що дасть змогу сформувати метрики виявлення та забезпечити оцінювання рівня безпеки.

Для досягнення окресленої мети необхідно вирішити такі завдання:

– сформулювати принципи, що дозволять узагальнити різні методи виявлення аномалій;

– проаналізувати метрики аномалій, зважаючи на заходи близькості формування оцінки поточного стану безпеки;

– обґрунтувати вибір релевантної міри близькості виявлення мережних аномалій.

4. Аналіз та вибір метрик аномалій на основі заходів подібності

4.1. Формування принципів узагальнення методів виявлення аномалій

Виявлення аномалій широко застосовується в таких галузях, як виявлення шахрайства з кредитними картками, встановлення вторгнень із метою кібербезпеки та військове спостереження за діями супротивника. Наприклад, аномальний характер трафіку в комп'ютерній мережі може означати, що зламаний комп'ютер надсилає конфіденційну інформацію неавторизованому хосту.

Вторгнення – це комплекс дій, спрямований на порушення безпеки комп'ютерних і мережних компонентів з погляду конфіденційності, цілісності та доступності. Це може бути зроблено внутрішнім або зовнішнім агентом для отримання несанкційного доступу та контролю за механізмом безпеки. Для захисту інфраструктури мережних систем системи виявлення вторгнень (*IDS*) надають механізми, що добре зарекомендували себе, які збирають і аналізують інформацію з різних сфер усередині хосту або мережі для виявлення можливих порушень безпеки.

Функції виявлення вторгнень передбачають:

1) моніторинг та аналіз дій користувача, системи й мережі;

2) налаштування систем для створення звітів про можливі вразливості;

3) оцінювання цілісності системи та файлів;

4) розпізнавання шаблонів типових атак;

5) аналіз аномальної активності та

6) відстеження порушень політики користувача.

IDS використовує оцінку вразливостей для оцінювання безпеки хосту чи мережі. Виявлення вторгнень ґрунтується на припущенні, що дії вторгнень помітно відрізняються від звичайних дій системи, отже, їх можна виявити.

Прийнято визначати два типи зловмисників: зовнішні та внутрішні. Зовнішні є неавторизованими користувачами машин, що зловмисники атакують, тоді як внутрішні мають дозвіл на доступ до системи, але не мають привілеїв для режиму *root* або суперкористувача. Маскарадний внутрішній зловмисник входить до системи як інший користувач, що має законний доступ до конфіденційної інформації, тоді як таємний внутрішній зловмисник найнебезпечніший, має право відключити контроль аудиту для себе.

У комп'ютерних системах існують різні класи вторгнень чи атак. Подамо їх у табл. 1.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 4 (26)

Таблиця 1. *Класи комп'ютерних атак: характеристики та приклад*

Назва атаки	Характеристики	Приклад
Вірус	– Самовідтворювана програма, що заражає систему без відома користувача. – Збільшує ймовірність зараження мережної файлової системи, якщо до системи звертається інший комп'ютер.	*Trivial. 88.D, Polyboot.B, Tuareg*
Хробак	– Самовідтворювана програма, що поширюється через мережні служби комп'ютерних систем без утручання користувача. – Може завдати серйозної шкоди мережі, споживаючи пропускну спроможність мережі.	*SQL Slammer, Mydoom, Code Red, Nimda*
Троян	– Шкідлива програма, що не здатна копіювати себе, але може спричинити серйозні проблеми безпеки в комп'ютерній системі. – Виглядає як корисна програма, але насправді вона має секретний код, що здатний створити бекдор у системі, дозволяючи їй легко робити що-небудь у системі, і може бути викликаний, коли хакер отримує контроль над системою без дозволу користувача.	*Example-Mail Bomb, phishing attack*
Відмова в обслуговуванні (*DoS*)	– Спроби заблокувати доступ до системних або мережних ресурсів. – Втрата обслуговування – це нездатність конкретної мережі або хост-сервісу, наприклад електронної пошти, для роботи. – Реалізується способом примусового перезавантаження цільового комп'ютера (комп'ютерів), або способом споживання ресурсів. – Передбачувані користувачі не можуть адекватно спілкуватися через недоступність послуги або перешкоди засобів зв'язку.	*Buffer overflow, ping of death (PoD), TCP SYN, smurf, teardrop*
Мережна атака	– Будь-який процес, що використовується для зловмисної спроби поставити під загрозу безпеку мережі, починаючи з рівня каналу передачі інформації та завершуючи прикладним рівнем за допомогою різних засобів, зокрема маніпулювання мережними протоколами. – Незаконне використання облікових записів і привілеїв користувачів, виконання дій з вилучення мережних ресурсів та пропускної спроможності, виконання дій, що перешкоджають законним авторизованим користувачам отримати доступ до мережних служб та ресурсів.	*Packet injection, SYN flood*
Фізична атака	Спроба пошкодження фізичних компонентів мереж або комп'ютерів.	*Cold boot, evil maid*
Парольна атака	Метою є отримання пароля протягом нетривалого часу, що проявляється як серія невдалих спроб входу в систему.	Атака за словником, атака за допомогою *SQL*-ін'єкції
Атака для збору інформації	Збирає інформацію або знаходить відомі вразливості з допомогою сканування або перевірки комп'ютерів чи мереж.	Сканування *SYS*, сканування *FIN*, сканування *XMAS*
Атака для отримання *root*-прав користувача (*U2R*)	– Може використовувати вразливості для отримання привілеїв суперкористувача системи під час запуску в системі як звичайний користувач. – До вразливостей належать перехоплення паролів, атака за словником або соціальна інженерія.	Руткіт, завантажувальний модуль, *Perl*
Віддалена атака на локальний комп'ютер (*R2L*)	– Можливість відправляти пакети у віддалену систему з допомогою мережі, не маючи облікового запису в цій системі, або отримати доступ як користувач або *root* у системі та виконувати шкідливі операції. – Здійснення атаки на загальнодоступні служби (наприклад, *HTTP* і *FTP*) або в процесі підключення захищених сервісів (таких як *POP* та *IMAP*).	*Warezclient, warezmaster, imap, ftp write, multihop, phf, spy*
Зонд	– Сканує мережі для визначення дійсних IP-адрес та збору інформації про хост (наприклад, які послуги вони пропонують, використовувана операційна система). – Надає зловмисникові інформацію зі списком потенційних уразливостей, які згодом можуть бути використані для атаки на обрані системи та служби.	перевірка IP та портів

Виявлення вторгнень, зважаючи на неправомірне використання, зазвичай шукає відомі шаблони вторгнень, але виявлення вторгнень з огляду на аномалії намагається виявити незвичайні шаблони. Методи виявлення вторгнень можна поділити на три типи залежно від механізму виявлення. До них належать: (I) основані на неправильному використанні; (II) основані на аномаліях та (III) гібридні (див. табл. 2). Сучасні дослідники здебільшого зосереджуються на виявленні мережних вторгнень на основі аномалій, оскільки вони можуть встановлювати як відомі, так і невідомі атаки.

Щоб забезпечити відповідне рішення щодо виявлення мережних аномалій, необхідно сформувати концепцію нормальності. Ідея нормальності зазвичай вводиться за допомогою формальної моделі, яка

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 4 (26) ISSN 2524-2296 (online)

виражає відношення між фундаментальними змінними, що беруть участь у динаміці системи. Отже, подія чи об'єкт розпізнається як аномальне, якщо ступінь його відхилення щодо профілю або поведінки системи, заданої моделлю нормальності, досить високий.

Таблиця 2. *Характеристики та типи методів виявлення вторгнень*

Метод	Характеристики
Оснований на неправильному використанні	– Виявлення грунтується на наборі правил або сигнатур відомих атак. – Може виявити всі відомі шаблони атак на основі довідкових даних. – Написання сигнатури, що охоплює всі можливі варіанти відповідної атаки, є складним завданням.
Оснований на аномаліях	– Основне припущення: всі дії щодо вторгнення обов'язково є аномальними. – Такий метод створює нормальний профіль активності та перевіряє, чи стан системи відрізняється від встановленого профілю на статистично значущу величину, щоб повідомити про спроби вторгнення. – Аномальні дії, які не є вторгненням, можуть бути позначені як вторгнення. Це хибні спрацьовування. – Необхідно обирати порогові рівні так, щоб жодна з двох вищезгаданих проблем не була необґрунтовано посилена, а вибір функцій для моніторингу не було оптимізовано. – Обчислювально витратно через накладні витрати та можливе оновлення кількох матриць профілів системи.
Гібридний	– Використовує переваги як методів неправильного використання, так і методів виявлення аномалій. – Спроби виявити як відомі, так і невідомі атаки.

Наприклад, візьмемо систему виявлення аномалій S, яка використовує контрольований підхід. Її можна розглядати як пару $S=(M, D)$, де M – модель нормальної поведінки системи, а D – міра близькості, що дозволяє за даними запису активності визначити ступінь відхилення, який ця діяльність має щодо моделі M. Отже, кожна система має здебільшого два модулі: модуль моделювання та модуль виявлення. Системи навчаються отримати модель нормальності M. Отримана модель згодом використовується модулем виявлення для оцінювання нових подій, об'єктів чи трафіку як аномальних або викидів. Саме вимір відхилення дає змогу класифікувати події чи об'єкти як аномальні або викиди. Зокрема модуль моделювання має бути адаптивним, щоб справлятися з динамічними сценаріями.

Існують дві широкі категорії мережних аномалій: (а) аномалії, пов'язані з функціюванням, і (б) аномалії, пов'язані з безпекою. Аномалії, пов'язані з безпекою, поділяються на три типи: точкові, контекстні та колективні. Ця схема класифікації описана в табл. 3.

Таблиця 3. *Аномалії: види, характеристики та приклади*

Типи	Характеристики	Приклад
Точкова аномалія	Зразок окремих даних, який виявився аномальним щодо інших даних.	Ізольований екземпляр мережного трафіку від звичайних екземплярів у певний час.
Контекстна аномалія	– Зразок даних, визнаний аномальним у певному контексті. – Контекст визначається структурою набору даних. – Для визначення контексту використовуються два набори атрибутів: (а) контекстуальні та (б) поведінкові.	Інтервал часу між транзакціями під час шахрайства з кредитними картками.
Колективна аномалія	– Множина пов'язаних екземплярів даних, які виявилися аномальними щодо всього набору даних. – Сукупність подій є аномалією, але окремі події не є аномаліями, якщо вони відбуваються окремо в послідовності.	Послідовність, подібна до такої: ... *http – web*, переповнення буфера, *http – web, http – web, ftp, httpweb, ssh,* *http – web, ssh*, переповнення буфера...

Розгляд основних аспектів виникнення, прояву мережних аномалій, а також характеристик і наявних методів виявлення вторгнень дає змогу сформулювати принципи узагальнення методів виявлення аномалій (див. табл. 4).

Сформульовані принципи можна розглядати як сукупність критеріїв та обмежень, що дозволяють оцінювати наявні та проєктувати нові методи й технології виявлення мережних аномалій.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 4 (26)

Таблиця 4. *Принципи узагальнення методів виявлення аномалій*

№	Принцип	Опис
1	Принцип надійності та компромісу	Метод виявлення новизни має забезпечувати надійну роботу з тестовою інформацією, що дозволяє максимально вилучити нові зразки та звести до мінімуму вилучення відомих зразків. Цей компроміс має бути обмеженою мірою передбачуваним і під експериментальним контролем.
2	Принцип одноманітного масштабування даних	Щоб полегшити виявлення новизни, має бути можливість, щоб усі тестові дані та дані навчання після нормалізації розміщувалися в одному діапазоні.
3	Принцип мінімізації параметрів	Метод виявлення новизни має бути спрямований на мінімізацію кількості параметрів, що встановлюються користувачем.
4	Принцип узагальнення	Система має бути здатною до узагальнення, не плутаючи узагальнену інформацію з новою.
5	Принцип незалежності	Метод виявлення новизни не має залежати від кількості доступних ознак і класів та має демонструвати прийнятну продуктивність у контексті незбалансованого набору даних, невеликої кількості вибірок і шуму.
6	Принцип адаптивності	Важливо, щоб система, яка розпізнає нові зразки під час тестування, мала змогу використовувати цю інформацію для повторного навчання.
7	Принцип обчислювальної складності	Низка застосунків для виявлення аномалій є в мережі, тому обчислювальна складність механізму виявлення аномалій має бути якнайменшою.

4.2. Аналіз метрик аномалій на основі мір близькості

Проблема виявлення мережних аномалій є завданням класифікації або кластеризації, що визначається такими компонентами [25]:

– типи вхідних даних;

– прийнятність мір близькості;

– маркування даних;

– класифікація методів, основаних на використанні розмічених даних;

– виявлення відповідних особливостей;

– повідомлення про аномалії.

Типи вхідних даних. Ключовим аспектом будь-якого методу виявлення вторгнень мереж на основі аномалій є характер вхідних даних, що використовуються для аналізу. Вхідні дані зазвичай є набором екземплярів даних (так званими об'єктами, записами, точками, векторами, шаблонами, подіями, випадками, вибірками, спостереженнями, об'єктами). Кожен екземпляр даних можна описати за допомогою набору атрибутів двійкового, категоріального або числового типу. Кожен екземпляр даних може складатися лише з одного атрибута (одномірний) або кількох атрибутів (багатомірний). У разі екземплярів багатовимірних даних усі атрибути можуть бути одного типу або бути поєднанням типів даних. Природа атрибутів визначає застосовність методів виявлення аномалій.

Доцільність заходів близькості. Заходи близькості (подібності чи відмінності) необхідні для розв'язання багатьох проблем розпізнавання образів для класифікації та кластеризації. Відстань – це кількісний ступінь того, наскільки далеко один від одного є два об'єкти. Заходи відстані, що задовольняють метричні властивості, називаються просто метрикою, тоді як інші неметричні заходи відстані іноді називають дивергенцією. Вибір міри близькості залежить від типу вимірювання чи подання об'єктів.

Як правило, міри близькості – це функції, що приймають аргументи у вигляді пар об'єктів і повертають числові значення, які стають вищими залежно від того, як об'єкти стають більш схожими. Міра близькості зазвичай визначається так.

Визначення: міра близькості D – це функція, $X \times X \to \mathbb{R}$, що має такі властивості:

– позитивність $\forall x, y \in X, S(x, y) \geq 0$;

– симетричність $\forall x, y \in X, S(x, y) = S(y, x)$;

– максимальність $\forall x, y \in X, S(x, x) \geq S(x, y)$,

де X – простір даних (інша назва «всесвіт»), а x, y – пара k-мірних об'єктів.

Найбільш поширені заходи близькості для числових, категоріальних і змішаних типів даних перелічені в табл. 5. Для числових даних передбачається, що вони подані у вигляді векторів, які містять реальні числа. Значення атрибутів належать до безперервної ділянки. Передбачається, що є два об'єкти: $x = x_1, x_2, x_3, \ldots, x_d$, $y = y_1, y_2, y_3, \ldots, y_d$ і Σ^{-1}, що є коваріацією даних з атрибутами, тобто розмірністю.

Для категоріальних даних обчислення мір подібності чи близькості не є простим через те, що немає чіткого поняття впорядкування категоріальних значень. Найпростіший спосіб знайти подібність

Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 4 (26)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

між двома категоріальними атрибутами – присвоїти подібність 1, якщо значення ідентичні, і 0, якщо значення не ідентичні. У табл. 5 функція $S_k(x_k, y_k)$ є подібністю за атрибутами. Вага атрибута w_k для атрибута k обчислюється, як показано у таблиці, *IOF* – зворотна частота виникнення, а *OF* – частота виникнення.

Таблиця 5. *Міра близькості для даних числового, категоріального та змішаного типу*

Numeric			
Name	**Measure,** $S_i(x_i, y_i)$	**Name**	**Measure,** $S_i(x_i, y_i)$
Euclidean	$\sqrt{\sum_{i=1}^{d}\lvert x_i - y_i\rvert^2}$	Weighted Euclidean	$\sqrt{\sum_{i=1}^{d}\alpha_i\lvert x_i - y_i\rvert^2}$
Squared Euclidean	$\sum_{i=1}^{d}\lvert x_i - y_i\rvert^2$	Squared-chord	$\sum_{i=1}^{d}\left(\sqrt{x_i} - \sqrt{y_i}\right)^2$
Squared X^2	$\sum_{i=1}^{d}\frac{(x_i - y_i)^2}{x_i + y_i}$	City block	$\sum_{i=1}^{d}\lvert x_i - y_i\rvert$
Minkowski	$\sqrt[p]{\sum_{i=1}^{d}\lvert x_i - y_i\rvert^p}$	Chebyshev	$\max_i\lvert x_i - y_i\rvert$
Canberra	$\sum_{i=1}^{d}\frac{\lvert x_i - y_i\rvert}{x_i + y_i}$	Cosine	$\frac{\sum_{i=1}^{d}x_i y_i}{\sqrt{\sum_{i=1}^{d}x_i^2}\sqrt{\sum_{i=1}^{d}y_i^2}}$
Jaccard	$\frac{\sum_{i=1}^{d}x_i y_i}{\sum_{i=1}^{d}x_i^2 + \sum_{i=1}^{d}y_i^2 - \sum_{i=1}^{d}x_i y_i}$	Bhattacharyya	$-\ln\sum_{i=1}^{d}\sqrt{(x_i y_i)}$
Pearson	$\sum_{i=1}^{d}(x_i - y_i)^2$	Divergence	$2\sum_{i=1}^{d}\frac{(x_i - y_i)^2}{(x_i + y_i)^2}$
Mahalanobis	$\sqrt{(x-y)^t\,\Sigma^{-1}(x-y)}$	–	–
Categorical			
$w_k,\ k = 1...d$	**Measure,** $S_k(x_k, y_k)$	$w_k,\ k = 1...d$	**Measure,** $S_k(x_k, y_k)$
$1/2$	$Overlap = \begin{cases} 1 & \text{if } x_k = y_k \\ 0 & \text{otherwise} \end{cases}$	$1/d$	$Eskin = \begin{cases} 1 & \text{if } x_k = y_k \\ \frac{n_k^2}{n_k^2 + 2} & \text{otherwise} \end{cases}$
$1/d$	$IOF = \begin{cases} 1 & \text{if } x_k = y_k \\ \frac{1}{1 + \log f_k(x_k) \times \log f_k(y_k)} & \text{otherwise} \end{cases}$	$1/d$	$OF = \begin{cases} 1 & \text{if } x_k = y_k \\ \frac{1}{1 + \log\frac{N}{f_k(x_k)} \times \log\frac{N}{f_k(y_k)}} & \text{otherwise} \end{cases}$
Mixed			
Name	**Measure**	**Name**	**Measure**
General Similarity Coefficient	$s_{gsc}(x,y) = \frac{\sum_{k=1}^{d}w(x_k,y_k)s(x_k,y_k)}{\sum_{k=1}^{d}w(x_k,y_k)}$ Для *числових* атрибутів, $s(x_k,y_k) = 1 - \frac{\lvert x_k - y_k\rvert}{R_k}$, де R_k – діапазон k *-го* атрибуту; $w(x_k,y_k) = 0$, якщо x чи y мають пропущені значення для k *-го* атрибуту; інакше $w(x_k, y_k) = 1$. Для *категоріальних* атрибутів, $s(x_k, y_k) = 1$, якщо $x_k = y_k$; інакше $s(x_k,y_k) = 0$; $w(x_k,y_k) = 0$, якщо точка даних x чи y має пропущене значення k *-го* атрибуту; інакше $w(x_k,y_k) = 1$.	General Distance Coefficient	$d_{gdc}(x,y) = \left(\frac{\sum_{k=1}^{d}w(s_k,y_k)}{\sum_{k=1}^{d}w(x_k,y_k)}d^2(x_k,y_k)\right)^{1/2}$, де $d^2(x_k,y_k)$ – квадратна відстань для k *-го* атрибуту; $w(x_k,y_k)$ – аналогічно як у *General Similarity Coefficient.* Для *числових* атрибутів, $d(x_k,y_k) = \frac{\lvert x_k - y_k\rvert}{R_k}$, де R_k – діапазон значень k *-го* атрибуту. Для *категоріальних* атрибутів, $d(x_k,y_k) = 0$, якщо $x_k = y_k$; інакше $d(x_k,y_k) = 1$.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 4 (26)

Нарешті, дані змішаного типу містять категоріальні й числові значення. Звичайною практикою кластеризації змішаного набору даних є перетворення категоріальних значень на числові значення з подальшим використанням алгоритму числової кластеризації. Інший підхід полягає в прямому порівнянні категоріальних значень, за умови якого два різні значення дають відстань, що дорівнює 1, а ідентичні значення – відстань, що дорівнює 0. Звичайно, можна використовувати й інші заходи для категоріальних даних. Дві добре відомі міри близькості, загальний коефіцієнт подібності й загальний коефіцієнт відстані даних змішаного типу показані в табл. 5. Такі методи можуть брати до уваги інформацію про подібність, укладену в категоріальних значеннях. Отже, кластеризація не спроможна точно виявити структуру подібності множини даних.

Маркування даних. Позначка, пов'язана з екземпляром даних, вказує, чи є цей екземпляр нормальним або аномальним. Необхідно зазначити, що отримання точних даних як нормального, так і аномального типу здебільшого обходиться непомірно дорого. Маркування часто виконується вручну фахівцями-людьми, і, отже, для отримання маркованого набору навчальних даних потрібні значні зусилля. Крім того, аномальна поведінка часто має динамічний характер, наприклад, можуть виникати нові типи аномалій, для яких немає позначених навчальних даних.

Класифікація методів, основана на використанні розмічених даних. Залежно від ступеня доступності міток методи виявлення аномалій можуть працювати в трьох режимах – контрольованому, напівконтрольованому й неконтрольованому.

У контрольованому режимі передбачається наявність набору навчальних даних, у якому зазначені екземпляри як нормального, так і аномального класу. Типовий підхід у таких випадках – побудова прогнозної моделі для класів нормальних та аномальних. Будь-який невидимий екземпляр даних порівнюється з моделлю, щоб визначити, до якого класу він належить. У разі контрольованого виявлення аномалій виникають дві основні проблеми. По-перше, аномальних випадків у навчальних даних набагато менше, ніж звичайних. По-друге, отримання точних і репрезентативних міток, особливо для класу аномалій, зазвичай є складним завданням. Низка методів вводять штучні аномалії у звичайний набір даних для отримання позначеного набору навчальних даних.

Напівконтрольовані методи припускають, що дані навчання позначають екземпляри тільки для класу. Оскільки їм не потрібні мітки класу аномалії, їх легше використовувати, на відміну від контрольованих методів.

Нарешті, методи без учителя не вимагають навчальних даних, отже, потенційно найбільш застосовувані. Методи цієї категорії неявно припускають, що нормальні ситуації трапляються набагато частіше, ніж аномалії тестових даних. Коли це припущення не правильне, такі методи страждають від високого рівня хибних тривог. Багато напівконтрольованих методів можна адаптувати для роботи в неконтрольованому режимі, використовуючи вибірку немаркованого набору даних як навчальних даних. Така адаптація передбачає, що тестові дані містять дуже мало аномалій і навчена модель стійка до цих небагатьох аномалій.

Ідентифікація релевантної функції. Вибір функцій відіграє важливу роль у виявленні мережних аномалій. Методи вибору функцій використовуються у сфері виявлення вторгнень для унеможливлення неважливих або нерелевантних функцій. Вибір ознак знижує обчислювальну складність, усуває надмірність інформації, підвищує точність алгоритму виявлення, полегшує розуміння даних та покращує узагальнення. Процес вибору функцій передбачає три основні етапи: (а) створення підмножини, (б) оцінювання підмножини та (в) перевірка. Три різні підходи до генерації підмножини: повний, евристичний і випадковий. Функції оцінки поділяються на п'ять окремих категорій: на основі оцінок, на основі ентропії чи взаємної інформації, на основі кореляції, на основі узгодженості та на основі точності виявлення. Моделювання та реалізація в реальному світі – два способи перевірки оціненої підмножини.

Алгоритми вибору ознак поділяються на три типи: методи-оболонки, фільтри та гібридні методи. Тоді як методи-оболонки намагаються оптимізувати деякі зумовлені критерії щодо набору функцій у межах процесу вибору, методи фільтрації покладаються на загальні характеристики навчальних даних для вибору функцій, які не залежать один від одного й сильно залежать від вихідних даних. Гібридний метод вибору ознак намагається використати суттєві особливості методів обгортки та фільтра.

Повідомлення про аномалії. Важливим аспектом будь-якого методу виявлення аномалій є спосіб повідомлення про аномалії. Зазвичай вихідні дані, отримані за допомогою методів виявлення аномалій,

Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 4 (26)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

бувають двох типів: (а) оцінка, яка є значенням, що поєднує (I) відстань або відхилення з посиланням на набір профілів або сигнатур; (II) вплив більшості в його околиці та (III) явне домінування відповідного підпростору; (б) позначка, що є значенням (нормальним або аномальним), присвоєним кожному екземпляру тесту. Зазвичай маркування екземпляра залежить від (I) розміру груп, створених за допомогою неконтрольованого методу; (II) компактності групи (груп); (III) голосування більшості на основі результатів, отриманих за допомогою декількох індексів або (IV) явне домінування підмножини ознак.

Аналіз запропонованої таблиці дає змогу зробити висновок, що серед безлічі поданих заходів близькості суттєво виокремлюється міра Махаланобіса. Її особливість пояснюється тим, що вона єдина, яка бере до уваги значення коефіцієнтів кореляції між різними вимірами, отриманими під час моніторингу активності мережі. Це змушує звернути увагу та більш детально проаналізувати саме цей захід близькості для виявлення мережних аномалій.

4.3. Обґрунтування вибору релевантної міри близькості виявлення мережних аномалій

Вибір міри близькості між спостереженнями дає підстави для обґрунтування метрики аномалій мережі, що виявляються в режимі онлайн. Нехай $A=\{a_1,\ \ldots, a_n\}$ – набір атрибутів мережної активності, що відстежуються, в довільний момент часу. Контрольовані атрибути можуть збиратися як внутрішніми, так і зовнішніми датчиками системи виявлення вторгнень (*IDS*). Дані періодично ростуть кожні t мілісекунд. Вхідний вектор $\vec{i}_t=\{i_{t,1},...,i_{t,n}\}$ надається онлайн, де $i_{t,j}\in\mathbb{R}$ позначає значення атрибута a_j за певний час t. З кожним заданим $\vec{i}_t$ рішення має бути прийняте негайно, незалежно від того, чи є $\vec{i}_t$ аномальним, чи ні.

Також доступні попередні дані H (передбачені нормальним функціюванням мережі). H – це $m\times n$ матриця, де стовпці позначають n атрибутів, що відстежуються, а рядки зберігають значення цих атрибутів за m періодів часу. H може бути записаний на основі повної роботи, яка свідомо є номінальною (наприклад, трафік без відомих збоїв), або він може бути створений на основі останніх вхідних даних m, які були отримані в режимі онлайн, тобто $H=\{\vec{i}_{t-m-1},...,\vec{i}_{t-1}\}$.

Розглянемо для виявлення аномалій дві найбільш популярні міри близькості спостережень за умови виявлення мережних аномалій – евклідову метрику й метрику Махаланобіса. Їх принципова різниця полягає в тому, що евклідова відстань може бути визначена для кожної пари спостережень, і, на відміну від відстані Махаланобіса, жодним чином не зважає на кореляцію, що існує між елементами всього вектора спостережень.

Відстань Махаланобіса – це n-мірний Z-показник. Він обчислює відстань між n-вимірною точкою та групою інших точок в одиницях стандартних відхилень. На відміну від звичайної n-вимірної евклідової відстані, відстань Махаланобіса також зважає на розподіл точок. Отже, якщо група точок є спостереженнями, то відстань Махаланобіса вказує, чи є нова точка викидом порівняно зі спостереженням. Точка зі значеннями, аналогічними точкам спостереження, розташована в багатовимірному просторі, в щільній ділянці й матиме меншу відстань Махаланобіса. Однак якщо викид буде розташований за межами щільної ділянки, тоді для нього буде більша відстань Махаланобіса.

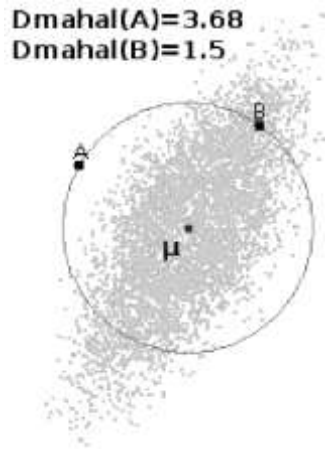

Рис. 1. Евклідова відстань порівняно з відстанню Махаланобіса

Формально відстань Махаланобіса обчислюється так: нехай $\vec{i}_t=\{i_{t,1},i_{t,2},...,i_{t,n}\}$ – вектор поточних вхідних даних n атрибутів, що відстежуються, а H матриця розмірністю $m\times n$ – група номінальних значень цих атрибутів. Визначаємо середнє значення H за значенням вектора $\mu=(\mu_1,\mu_2,...,\mu_n)$, а S – матриця H. Відстань Махаланобіса, D_{mahal}, від $\vec{i}_t$ до H визначається так:

$$D_{mahal}(\vec{i}_t,H)=\sqrt{(\vec{i}_t-\vec{\mu})S^{-1}(\vec{i}_t^{\,T}-\vec{\mu}^{\,T})}.$$

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2023. No. 4 (26)

Приклад зображено на рис. 1, на якому можна побачити, що, хоча A і B мають однакову евклідову відстань від центроїда μ, відстань Махаланобіса для $A(3,68)$ більша, ніж для $B(1,5)$, оскільки екземпляр B імовірніший, ніж екземпляр A щодо інших точок.

Завдяки природі відстані Махаланобіса можемо використовувати його для виявлення аномалій у мережному оточенні. Кожен із n атрибутів домену (наприклад, трафіку) відповідає виміру. Вхідний вектор $\vec{i}_t$ – це n-вимірна точка, яка вимірюється відстанню Махаланобіса щодо H. Відстань Махаланобіса потім використовується для вказівки того, чи кожна нова вхідна точка $\vec{i}_t$ є викидом щодо H.

5. Аналіз досягнутих результатів

Вибір міри близькості Махаланобіса як підстави метрики виявлення мережних аномалій пояснюється тим, що тільки міра близькості Махаланобіса бере до уваги корельованість спостережень і, відповідно до цього зважає на геометрію розкиду спостерігачів нормального режиму роботи, що дає більш обґрунтовані оцінки для віднесення спостереження.

Використовуючи відстань Махаланобіса, можна легко виявити три загальні категорії аномалій.

1) *Точкові аномалії*: неприпустимі екземпляри даних, що відповідають неприпустимим значенням $\vec{i}_t$.

2) *Контекстуальні аномалії*: екземпляри даних, які є аномальними лише щодо певного контексту, але не інакше. У нашому підході контекст забезпечується змінними даними ковзного вікна.

3) *Групові аномалії*, що є пов'язаними екземплярами даних, які допустимі окремо, але аномальні, коли вони зустрічаються разом. Це досягається завдяки багатовимірності точок, що визначаються відстанню Махаланобіса.

Аномалія будь-якого типу може призвести до того, що репрезентативна точка виявиться окремо від номінальних точок у відповідному вимірі, що помістить її за межі щільної ділянки. Це призведе до значної відстані Махаланобіса і, зрештою, спрацьовування тривоги.

Висновок

1. У статті розглянуто основні аспекти виявлення мережних аномалій. Сформульовано принципи, що дають змогу узагальнити різні методи виявлення аномалій. Описано атаки, з якими зазвичай стикаються системи виявлення мережних вторгнень, а також характеристики й типи методів їх виявлення. Мережні аномалії розглянуто як прояви мережних атак, що дозволяє виконати класифікацію аномалій.

2. Подано види, показники та приклади мережних аномалій. Для класифікації та полегшення виявлення мережних аномалій пропонуються міри близькості для числових (що характеризують аномалії), категоріальних і змішаних типів даних.

3. Аргументовано вибір міри близькості Махаланобіса як основи метрики аномалій. Обґрунтовано, що тільки міра близькості Махаланобіса бере до уваги корельованість спостережень і, отже, враховує геометрію розкиду спостережень нормального режиму роботи й відповідно дає більш повні оцінки для визначення спостереження як аномального.

Список літератури

1. Yevseiev S., Zviertseva N., Pribyliev Y., Lezik O., Komisarenko O., Nalyvaiko A., Pogorelov V., Katsalap V., Husarova I. Development of the concept for determining the level of critical business processes security. *Eastern-European Journal of Enterprise Technologies.* 2023. Vol. 1/9. No. 121. P. 21–40. DOI: 10.15587/1729-4061.2023.274301

2. Blazquez-Gartfa, A., Conde A., Mori U., Lozano J. A review on outlier/anomaly detection in time series data, *ACM Comput. Surv.* 2021. Vol. 54. No. 3. DOI: http://dx.doi.org/10.1145/3444690

3. Arif I., Ackovska N. IoT aided smart home architecture for anomaly detection, in: Data Science and Internet of Things: Research and Applications at the Intersection of DS and IoT. *Springer International Publishing, Cham.* 2021. P. 1–19. DOI: http://dx.doi.org/10.1007/978-3-030-67197-6_1

4. Lin X., Yeh E., Lin P. Anomaly detection for IoT systems, in: Encyclopedia of Wireless Networks. *Springer International Publishing, Cham.* 2020. P. 18–20. DOI: http://dx.doi.org/10.1007/978-3-319-78262-1_183

5. Pei J., Zhong K., Jan M., Li J. RETRACTED: Personalized federated learning framework for network traffic anomaly detection. *Computer Networks*. 2022. Vol. 209, P. 1389–1286. DOI: https://doi.org/10.1016/j.comnet.2022.108906

Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 4 (26)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

6. Fahim M., Sillitti A. Anomaly detection, analysis and prediction techniques in IoT environment: A systematic literature review. *IEEE Access 7.* 2019. P. 81664–81681. DOI: http://dx.doi.org/10.1109/ACCESS.2019.2921912

7. Cook A., Misirli G., Fan Z. Anomaly detection for loT time-series data: A survey. *IEEE Internet Things J.* 2020. Vol. 7. No. 7. P. 6481–6494. DOI: http://dx.doi.org/ 10.1109/JIOT.2019.2958185

8. O'Reilly C., Gluhak A., Imran M. Distributed anomaly detection using minimum volume elliptical principal component analysis. *IEEE Trans. Knowl.* 2016. Vol. 28. No. 9. P. 2320–2333. DOI: http://dx.doi.org/10.1109/TKDE.2016.2555804

9. Mahajan S., Chen L., Tsai T. Short-term PM2.5 forecasting using exponential smoothing method: a comparative analysis. *Sensors.* 2018. Vol. 18. No. 10. 3223 p. DOI: http://dx.doi.org/10.3390/s18103223

10. Charles, A. Interpreting deep learning: the machine learning rorschach test? 2018. URL: arXiv:1806.00148

11. Chen Z., Chen D., Zhang X., Yuan Z., Cheng X. Learning graph structures with transformer for multivariate time series anomaly detection in IoT. *IEEE Internet Things J.* 2021. No. 1. P. 1–12. DOI: http://dx.doi.org/10.1109/JIOT.2021.3100509

12. Ukil A., Bandyoapdhyay S., Puri C., Pal A. IoT healthcare analytics: The importance of anomaly detection. *30th International Conference on Advanced Information Networking and Applications. AINA.* 2016. P. 994–997. DOI: http://dx.doi.org/10.1109/AINA.2016.158

13. Yang K., Kpotufe S., Feamster N. An efficient one-class SVM for anomaly detection in the internet of things. 2021. URL: arXiv:2104.11146

14. Dunne M., Gracioli G., Fischmeister S. A comparison of data streaming frameworks for anomaly detection in embedded systems. *Proceedings of the 1st International* Workshop on Security and Privacy for the Internet-of-Things IoTSec. 2018. Orlando. FL. USA, URL: https://uwaterloo.ca/embedded-software-group/publications/comparison-data-streaming-frameworks-anomaly-detection

15. Wu D., Jiang Z., Xie X., Wei X., Yu W., Li R. LSTM learning with Bayesian and Gaussian processing for anomaly detection in industrial IoT. *IEEE Trans. Ind. Inf.* 2020. Vol 16. No 8. P. 5244–5253. DOI: http://dx.doi.org/10.1109/TII.2019.2952917

16. Fahim M., Sillitti A. Anomaly detection, analysis and prediction techniques in IoT environment: A systematic literature review. *IEEE Access.* 2019. Vol. 7. P. 81664–81681. DOI: http://dx.doi.org/10.1109/ACCESS.2019.2921912

17. Galvao Y., Albuquerque V., Fernandes B., Valenka M. Anomaly detection in smart houses: Monitoring elderly daily behavior for fall detecting. *Latin American Conference on Computational Intelligence. la-CCI.* 2017. P. 1–6. DOI: http://dx.doi.org/10.1109/LA- CCI.2017.8285701

18. Lu H., Li Y., Mu S., Wang D., Kim H., Serikawa S. Motor anomaly detection for unmanned aerial vehicles using reinforcement learning. *IEEE Internet Things J.* 2018. Vol. 5. No. 4. P. 2315–2322. DOI: http://dx.doi.org/10.1109/JIOT.2017.2737479

19. Nguyen T., Marchal S., Miettinen M., Fereidooni H., Asokan N., Sadeghi A. Diot: A federated self-learning anomaly detection system for loT. *39th International Conference on Distributed Computing Systems. ICDCS.* 2019. P. 756–767. DOI: http://dx.doi.org/10.1109/ICDCS.2019.00080

20. Alsheikh M., Konieczny L., Prater M., Smith G., Uludag S. State and trends of IoT security: Unequivocal appeal to cybercriminals, onerous to defenders. *IEEE Consum. Electr. Mag.* 2021. Vol. 1. P. 1–17. DOI: http://dx.doi.org/10.1109/MCE.2021.3079635

21. Munir M., Siddiqui S., Dengel A., Ahmed S. DeepAnT: A deep learning approach for unsupervised anomaly detection in time series. *IEEE Access.* 2019. Vol. 7. P. 1991–2005. DOI: http://dx.doi.org/10.1109/ACCESS.2018.2886457

22. Srikanth P. An efficient approach for clustering and classification for fraud detection using bankruptcy data in IoT environment. *Int. J. Inf. Technol.* 2021. P. 1–7. URL: https://www.x-mol.net/paper/article/1442394146737025024

23. Hafeez I., Antikainen M., Ding A., Tarkoma S. IoT-KEEPER: Detecting malicious IoT network activity using online traffic analysis at the edge. *IEEE Trans. Netw. Serv. Manag.* 2020. Vol. 17. No. 1. P. 45–59. DOI: http://dx.doi.org/10.1109/TNSM.2020.2966951

24. Bosman H., Lacca G., Tejada A., Wortche H., Liotta A. Ensembles of incremental learners to detect anomalies in ad hoc sensor networks. *Ad Hoc Netw.* 2015. No. 35. P. 14–36. DOI: http://dx.doi.org/10.1016/j.adhoc.2015.07.013

25. Milov O., Yevseiev S., Opirskyy I., Dunaievska O., Huk O., Pogorelov V., Bondarenko K., Zviertseva N., Yevgen Melenti Y., Tomashevsky B. Development of concepts for the cyber security metrics classification. *Eastern-European Journal of Enterprise Technologies.* 2022. Vol. 4/4. No. 118. P. 6–18, DOI: https://doi.org/10.15587/1729-4061.2022.263416

References

1. Yevseiev, S., Zviertseva, N., Pribyliev, Y., Lezik, O., Komisarenko, O., Nalyvaiko, A., Pogorelov, V., Katsalap, V., Husarova, I. (2023), "Development of the concept for determining the level of critical business processes security", *Eastern-European Journal of Enterprise Technologies*, No. 1/9 (121). P. 21–40. DOI: 10.15587/1729-4061.2023.274301

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2023. No. 4 (26)*

2. Blazquez-Gartfa, A., Conde, A., Mori, U., Lozano, J. (2021), "A review on outlier/anomaly detection in time series data", *ACM Comput*, No. 54 (3). DOI: http://dx.doi.org/10.1145/3444690

3. Arif, I., Ackovska, N. (2021), "IoT aided smart home architecture for anomaly detection, in: Data Science and Internet of Things: Research and Applications at the Intersection of DS and IoT", *Springer International Publishing, Cham*, P. 1–19. DOI: http://dx.doi.org/10.1007/978-3-030-67197-6_1

4. Lin, X., Yeh, E., Lin, P. (2020), "Anomaly detection for IoT systems, in: Encyclopedia of Wireless Networks", *Springer International Publishing, Cham*, P. 18–20. DOI: http://dx.doi.org/10.1007/978-3-319-78262-1_183

5. Pei, J., Zhong, K., Jan, M., Li, J. (2022), "RETRACTED: Personalized federated learning framework for network traffic anomaly detection", *Computer Networks*, Vol. 209, 108906, ISSN 1389–1286. DOI: https://doi.org/10.1016/j.comnet.2022.108906

6. Fahim, M., Sillitti, A. (2019), "Anomaly detection, analysis and prediction techniques in IoT environment: A systematic literature review", *IEEE Access* No.7, P. 81664–81681. DOI: http://dx.doi.org/10.1109/ACCESS.2019.2921912

7. Cook, A., Misirli, G., Fan, Z. (2020), "Anomaly detection for loT time-series data: A survey", *IEEE Internet Things J.*, Vol. 7, No. 7, P. 6481–6494. DOI: http://dx.doi.org/ 10.1109/JIOT.2019.2958185

8. O'Reilly, C., Gluhak, A., A. Imran, M. (2016), "Distributed anomaly detection using minimum volume elliptical principal component analysis", *IEEE Trans. Knowl*, Vol. 28, No. 9, P. 2320–2333. DOI: http://dx.doi.org/10.1109/TKDE.2016.2555804

9. Mahajan, S., Chen, L., Tsai, T. (2018), "Short-term PM2.5 forecasting using exponential smoothing method: a comparative analysis", *Sensors*, Vol. 18, No. 10. 3223 p. DOI: http://dx.doi.org/10.3390/s18103223

10. Charles, A. (2018), "Interpreting deep learning: the machine learning rorschach test?" available at: arXiv:1806.00148

11. Chen, Z., Chen, D., Zhang, X., Yuan, Z., Cheng, X. (2021), "Learning graph structures with transformer for multivariate time series anomaly detection in IoT", *IEEE Internet Things J.* No. 1. P. 1–12. DOI: http://dx.doi.org/10.1109/JIOT.2021.3100509

12. Ukil, A., Bandyoapdhyay, S., Puri, C., Pal, A. (2016), "IoT healthcare analytics: The importance of anomaly detection, in: 2016 IEEE", *30th International Conference on Advanced Information Networking and Applications, AINA*, P. 994–997. DOI: http://dx.doi.org/10.1109/AINA.2016.158

13. Yang, K., Kpotufe, S., Feamster, N. (2021), "An efficient one-class SVM for anomaly detection in the internet of things", available at: arXiv:2104.11146

14. Dunne, M., Gracioli, G., Fischmeister, S. (2018), "A comparison of data streaming frameworks for anomaly detection in embedded systems", Proceedings of the, available at: https://uwaterloo.ca/embedded-software-group/publications/comparison-data-streaming-frameworks-anomaly-detection

15. Wu, D., Jiang, Z., Xie, X., Wei, X., Yu, W., Li, R. (2020), "LSTM learning with Bayesian and Gaussian processing for anomaly detection in industrial IoT", *IEEE Trans*, Vol. 16, No. 8, P. 5244–5253. DOI: http://dx.doi.org/10.1109/TII.2019.2952917

16. Fahim, M., Sillitti, A. (2019), "Anomaly detection, analysis and prediction techniques in IoT environment: A systematic literature review", *IEEE Access* No. 7, P. 81664–81681. DOI: http://dx.doi.org/10.1109/ACCESS.2019.2921912

17. Galvao, Y., Albuquerque, V., Fernandes, B., Valenka, M. (2017), "Anomaly detection in smart houses: Monitoring elderly daily behavior for fall detecting", *IEEE Latin American Conference on Computational Intelligence, la-CCI*, P. 1–6. DOI: http://dx.doi.org/10.1109/LA- CCI.2017.8285701

18. Lu, H., Li, Y., Mu, S., Wang, D., Kim, H., Serikawa, S. (2018), "Motor anomaly detection for unmanned aerial vehicles using reinforcement learning", *IEEE Internet Things J.*, Vol. 5, No. 4, P. 2315–2322. DOI: http://dx.doi.org/10.1109/JIOT.2017.2737479

19. Nguyen, T., Marchal S., Miettinen, M., Fereidooni, H., Asokan, N., Sadeghi, A. "Diot: A federated self-learning anomaly detection system for loT", *39th International Conference on Distributed Computing Systems, ICDCS*, P. 756–767. DOI: http://dx.doi.org/10.1109/ICDCS.2019.00080

20. Alsheikh, M., Konieczny, L., Prater, M., Smith, G., Uludag, S. "State and trends of IoT security: Unequivocal appeal to cybercriminals, onerous to defenders", *IEEE Consum. Electr. Mag.* Vol. 1, P. 1–17. DOI: http://dx.doi.org/10.1109/MCE.2021.3079635

21. Munir, M., Siddiqui, S., Dengel, A., Ahmed, S. (2019), "DeepAnT: A deep learning approach for unsupervised anomaly detection in time series", *IEEE Access*, Vol. 7, P. 1991–2005. DOI: http://dx.doi.org/10.1109/ACCESS.2018.2886457

22. Srikanth, P. (2021), "An efficient approach for clustering and classification for fraud detection using bankruptcy data in IoT environment", *Int. J. Inf. Techno.*, available at: https://www.x-mol.net/paper/article/1442394146737025024

23. Hafeez, I., Antikainen, M., Ding, A., Tarkoma, S. (2020), "IoT-KEEPER: Detecting malicious IoT network activity using online traffic analysis at the edge", *IEEE Trans. Netw. Serv. Manag*, Vol. 17, No. 1, P. 45–59. DOI: http://dx.doi.org/10.1109/TNSM.2020.2966951

Сучасний стан наукових досліджень та технологій в промисловості. 2023. № 4 (26)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

24. Bosman, H., Lacca, G., Tejada, A., Wortche, H., Liotta, A. (2015), "Ensembles of incremental learners to detect anomalies in ad hoc sensor networks", *Ad Hoc Netw.*, No. 35, P. 14–36. DOI: http://dx.doi.org/10.1016/j.adhoc.2015.07.013

25. Milov, O., Yevseiev, S., Opirskyy, I., Dunaievska, O., Huk, O., Pogorelov, V., Bondarenko, K., Zviertseva, N., Melenti, Y., Tomashevsky, B. (2022), "Development of concepts for the cyber security metrics classification", *Eastern-European Journal of Enterprise Technologies.* Vol. 4/4, No. 118, P. 6–18, DOI: https://doi.org/10.15587/1729-4061.2022.263416

Received 07.12.2023

Відомості про авторів / About the Authors

Бондаренко Кирило Олександрович – Національний технічний університет "Харківський політехнічний інститут", аспірант, Харків, Україна; e-mail: bond.kirill.alexandrovich@gmail.com, ORCID ID: https://orcid.org/0000-0002-2168-155X

Bondarenko Kyrylo – National Technical University "Kharkiv polytechnic institute", PhD student, Kharkiv, Ukraine.

ANALYSIS AND SELECTION OF RELEVANT NETWORK ANOMALY DETECTION METRICS

The object of the study is the detection of network anomalies - an important and dynamically developing area of research. The article discusses the main aspects of network anomaly detection. Principles are formulated that allow generalization of various anomaly detection methods. The attacks that network intrusion detection systems typically face are presented, along with the characteristics and types of intrusion detection methods. Network anomalies are considered as manifestations of network attacks, which makes it possible to classify anomalies. The analysis of iterative sources showed that, despite the breadth of coverage of various methods, subject areas and tasks for detecting network anomalies, less attention is paid to the key issue – the analysis of network anomaly metrics and the rationale for choosing the relevant metric in a particular case. The paper presents types, characteristics and examples of network anomalies. To classify and facilitate the detection of network anomalies, metrics are proposed that are based on proximity measures for numeric, categorical, and mixed data types that characterize anomalies. The network anomaly detection problem is presented as a classification or clustering problem. The components that characterize this problem are identified, namely types of input data, acceptability of proximity measures, data labeling, classification of methods based on the use of labeled data, identifying relevant features and reporting anomalies. An approach is described that allows you to timely generate the required set of metrics, which will ensure not only the formation of preventive countermeasures, but also allow you to assess the current state of the security system as a whole. In addition, it provides the possibility of forming multi-circuit security systems, taking into account the influence (integration) of targeted (mixed) attacks on infrastructure elements, as well as the possibility of their synthesis with social engineering methods.

Keywords: network anomaly; intrusion detection system; proximity measure; attack classification.

Бібліографічні описи / Bibliographic descriptions

Бондаренко К. О. Аналіз і вибір релевантної метрики виявлення мережних аномалій. *Сучасний стан наукових досліджень та технологій в промисловості*. 2023. № 4 (26). С. 145–157. DOI: https://doi.org/10.30837/ITSSI.2023.26.145

Bondarenko, K. (2023), "Analysis and selection of relevant network anomaly detection metrics", *Innovative Technologies and Scientific Solutions for Industries*, No. 4 (26), P. 145–157. DOI: https://doi.org/10.30837/ITSSI.2023.26.145