КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 1 (25), 2024
ISSN 2663 - 4023
DOI 10.28925/2663-4023.2024.25.434448
УДК 004.942:004.056

Субач Ігор Юрійович
доктор технічних наук, професор, завідувач Спеціальної кафедри №5
Інститут спеціального зв'язку та захисту інформації
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна
ORCID ID: 0000–0002–9344–713X
igor_subach@ukr.net

Шарадкін Дмитро Михайлович
кандидат технічних наук, доцент, доцент Спеціальної кафедри №5
Інститут спеціального зв'язку та захисту інформації
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна
ORCID ID: 0000–0001–6407–8040
dmsh@ukr.net

Яковів Ігор Богданович
кандидат технічних наук, доцент, доцент спеціальної кафедри №5
Інститут спеціального зв'язку та захисту інформації
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна
ORCID ID: 0000–0001–7432–898X
iyakov52@gmail.net

ВИКОРИСТАННЯ МЕТРИЧНИХ МЕТОДІВ ПОРІВНЯННЯ ГІСТОГРАМ ПРИ ВИЯВЛЕННІ ЗМІН В ШИФРОВАНОМУ МЕРЕЖЕВОМУ ТРАФІКУ

Анотація. З зростанням частки шифрованого трафіку яка передається мережею Internet, і як наслідок — через відсутність доступу до вмісту зашифрованих пакетів, унеможливлюється безпосередня ідентифікація причин, які викликають аномалій в поведінці мережі, кардинально ускладняється задача виявлення загроз кібербезпеці. Доступними для аналізу лишаються лише зовнішні симптоми, які проявляються у вигляді зміни певних, як правило, найпростіших параметрів трафіку — його об'єму, інтенсивності, затримок між пакетами тощо. Таким чином підвищується роль та значення алгоритмів визначення змін у трафіку, які використовуючи сучасні методи, зокрема такі як машинне навчання, здатні ідентифікувати різноманітні, в тому числі невідомі раніше, типи аномалії. Такі алгоритми виконують аналіз доступних для безпосереднього вимірювання параметрів трафіку мережі, представляючи їх розвиток у вигляді часових рядів. На сьогоднішній день одним з найменш вивченим класів таких алгоритмів є алгоритми прямого порівняння гістограм розподілів значень часових рядів на різних проміжках часу, зокрема підклас таких алгоритмів, а саме метричні алгоритми. Ці алгоритми ґрунтуються на припущенні, що різниця між гістограмами значень часового ряду на сусідніх інтервалах спостереження свідчить про зміни в потоках подій, що породжують мережевий трафік. Однак задача виміру відмінності/подібності між гістограмами, які розглядаються як об'єкти в багатовимірному просторі, не має однозначного вирішення. В роботі проведений аналіз існуючих метрик подібності гістограм і описана серія досліджень методом статистичного моделювання, яка дозволила оцінити залежність ефективності алгоритмів від зовнішніх параметрів, а також порівняти алгоритми вказаного класу як між собою, так і з іншими алгоритмами виявлення змін. Це дало можливість проаналізувати якість використання алгоритмів з точки зору практичного застосування. Результати показали, що метричні алгоритми порівняння гістограм можуть демонструвати високу якість роботи і в деяких випадках перевищують результати інших відомих алгоритмів виявлення змін в часових рядах, забезпечуючи зниження кількості хибних спрацьовувань і скорочення затримок між моментом появи зміни в об'єкті спостереження та моментом її фіксації алгоритмом.

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 1 (25), 2024
ISSN 2663 - 4023

Ключові слова: кібербезпека; кіберінцидент; машинне навчання; шифрований мережевий трафік; часові ряди; метрики подібності гістограм; алгоритми виявлення подібності.

ВСТУП

За останнє десятиріччя з'явилися нові чинники, які суттєво впливають на методи виявлення зловмисного втручання в комп'ютерні мережі та системи. Серед чинників, які створюють нові виклики для розробників систем кіберзахисту, як один з найсуттєвіших виділимо значне збільшення об'єму шифрованого трафіку. Хоча перші спроби шифрування інтернет-трафіку з'явилися разом з самою інтернет-мережею як засіб забезпечення конфіденційності користувачів, до 2014 року більша частка трафіку передавалося у нешифрованому вигляді. Це давало можливість використовувати системи DPI (Deep Packet Inspection — глибокий аналіз пакетів) з перевіркою корисного навантаження пакетів на різних рівнях мережевих протоколів. Наразі ситуація стрімко та кардинально змінилася. Якщо в 2016 році за даними Google доля зашифрованого трафіку становила 70% від всього трафіку, в 2022 році цей показник сягнув 95% [1]. Згідно прогнозів, до 2025 року практично весь мережевий трафік буде передаватися в зашифрованому вигляді [2]. З'явившись як засіб кіберзахисту та забезпечення конфіденційності інформації користувача, шифрування трафіку стало використовуватися і зловмисниками. Виявилося, що оскільки зашифрований трафік не може бути проаналізований системами DPI, а суб'єкти загроз продовжують розвивати свої методики (зокрема, приховування атаки у зашифрованому трафіку), поширення шифрування трафіку водночас знову підвищує рівень вразливості систем до кібератак [3].

Постановка проблеми. Описане концептуальне протиріччя ставить перед розробниками систем захисту дуже складне завдання: знайти баланс між наскрізною безпекою, збереженням конфіденційності кінцевих користувачів та необхідністю отриманням з трафіку інформації, необхідної для виявлення загроз. Іншими словами — розробити такі способи виявлення шкідливої поведінки, які не знижують рівень конфіденційності інформації користувачів. Водночас такі системи мають використовувати лише ту інформацію, яку можна отримати з шифрованого трафіку. За вказаних умов найбільш перспективним видається застосування формалізованих методів математичної статистики та машинного навчання (МН) та їх аналіз з точку зору ефективності, точності та розподільчої здатності.

Аналіз останніх досліджень і публікацій. За останні роки проблемам аналізу шифрованого трафіку приділяється все більше уваги дослідників [4] – [6]. Основна ідея більшості робіт полягає в тому, що незважаючи на те, що самі дані зашифровані і не піддаються аналізу за технологією DPI, доступною лишається інформація про заголовки пакетів, IP-адреси та порти джерела та отримувача інформації, статистичні властивості потоків трафіку (такі як інтервали затримки між моментами надходження пакетів, розмір пакетів, кількість пакетів за протоколами) тощо і може слугувати джерелом для подальшого аналізу. Дослідження концентруються на використанні технологій МН для виявлення несправностей, аномалій і загроз у трафіку, спираючись саме на таку інформацію.

Історично першими в якості характеристик, що порівнюються використовувалися точкові оцінки основних статистичних параметрів набору даних, зокрема середнього, дисперсії, медіани, рідше — моментів третього і четвертого порядку, коефіцієнтів автокореляції [7], [8]. Популярність такого підходу підтримується зокрема і тим, що обчислені точкові оцінки легко і зрозуміло для користувача можна відображати на

інформаційних панелях для візуальної перевірки. Спільною особливістю вказаних алгоритмів є те, що вони спершу переходять від поелементного представлення даних у вигляді значень ряду до агрегованої точкової характеристики (оцінки) цих значень. Легко зрозуміти що на цьому кроці втрачається значна частина інформації, яка доступна у вхідному наборі даних. Як і слід очікувати, такі методи не завжди здатні чутливо відображати реальні зміни в розподілі даних. Тому природно виникає ідея використовувати більше інформації, ніж залишається при аналізі на основі точкової оцінки, тобто працювати не з моментами, а безпосередньо з вибірковими розподілами, тобто з менш «урізаним» об'ємом інформації. Один з шляхів реалізації таких ідей полягає в використанні аналізу гістограм розподілів значень часових рядів [9], [10].

Одна група методів порівняння гістограм базується на використання апроксимації гістограм відповідними функціями щільності розподілу. Рішення щодо наявності точки змін приймаються на основі порівняння цих функцій [11]. В якості параметрів які при цьому використовуються, можуть обиратися або параметри функцій розподілу, або параметри ядер (у випадку використання KDE-методів) [12]. Слід, однак, зазначити, що ці методи часто виявляються непридатними для використання в прикладних задачах. Зокрема в задачах аналізу мережевого трафіку, особливо у випадках наявності в ньому складових, що породжуються різноманітними протоколами та застосунками, дані часто виявляються не унімодальними, не регулярними і не відповідними вживаним теоретичним розподілам.

Інша група методів базується на метричному підході до визначення відмінності між гістограмами двох вибірок [13] – [16]. Близькість, подібність гістограм вимірюється за допомогою деякої статистики, що забезпечує кількісний вираз поняття «відстані» (або зворотної до неї «близькості») між гістограмами [17]. Хоча на сьогоднішній день запропоновано багато різноманітних метрик для вимірювання відстані у багатовимірних просторах, порівняльне дослідження їх властивостей при використанні для виміру подібності гістограм значень часових рядів авторам роботи невідомо. Тому навіть завдання порівняння, виявлення та опису сильних і слабких сторін, формування рекомендацій щодо вибору конкретних метрик для часових рядів параметрів об'єктів різної природи і типів зміни їхньої поведінки — є актуальним завданням.

Суттєвою проблемою, яка з огляду на забезпечення конфіденційності інформації значно ускладнює дослідження в галузі виявлення точок змін в мережевому трафіку є відсутність загальновживаних наборів даних, які б дозволяли об'єктивне порівняння результатів різноманітних досліджень. Крім того, створити або зібрати датасет, що здатен охопити всі можливі варіанти змін в трафіку практично неможливо. Нарешті, при розробці систем кіберзахисту, центральною проблемою також є необхідність убезпечення від несанкціонованого користування результатами досліджень зловмисниками. Тому для аналізу алгоритмів шифрованого трафіку видається доцільним їх вивчення на певному наборі змодельованих даних, що імітують найбільш поширені зміни в поведінці об'єктів дослідження. Опис таких наборів можна знайти в роботах [18] – [22]. Цей підхід до аналізу алгоритмів на базових складових, використовується і в представленому дослідженні.

Мета статті. Метою описаних в роботі досліджень є вивчання та порівняння можливостей використання метричних алгоритмів аналізу гістограм, розробка та практична перевірка процедури визначення критичних значень метрик для прийняття рішень щодо наявності/відсутності точки зміни в часових рядах, зокрема таких, які імітують шифрований мережевий трафік.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Виходячи з того, що мережевий трафік є складним динамічним процесом і суперпозицією багатьох потоків з множинними взаємопов'язаними характеристиками що генеруються різними протоколами, кібератака повинна змінити деякі ознаки мережевого трафіку. Припускається, що поява в трафіку складової, яка генерується певним програмним застосунком, мережевим протоколом, або навіть операційною системою, що встановлена на кінцевих пристроях мережі, змінює ознаки мережевого трафіку, що може бути зафіксовано та проаналізовано системою виявлення вторгнень (IDS, Intrusion Detection System). При цьому додаткова інформація про характер таких змін як правило відсутня, що тим більше характерно для випадку новітніх атак, т. з. «атак нульового для» [23]. Єдине, що може зробити дослідник за цих умов — виявити факт та момент зміни ознак трафіку. Іншими словами — задача ставиться не як задача діагностування аномального явища, а як задача виявлення самого факту появи аномалій (змін) в ознаках трафіку. В рамках технології машинного навчання описана прикладна задача зводиться до задачі виявлення аномалії (Anomaly Detection) або виявлення точки зміни (Change Point Detection — CPD) в поведінці об'єкту моніторингу [7], [8], [24], [25].

Для виконання процедури CPD необхідно обрати ознаки, за якими буде проводитися процедура визначення зміни та алгоритм, який буде використовуватися. Різні дослідники виходячи з своїх цілей пропонують різні набори ознак, що характеризують мережевий трафік як об'єкт моніторингу. Серед них — окрім вже зазначених вище первинних ознак — зустрічаються і похідні ознаки, такі як частота появи вхідних та вихідних пакетів, відношення вказаних величин, співвідношення розмірів вхідних та вихідних пакетів, тривалість сесій тощо. Окремо слід відзначити роботу [26] в якій автори наводять близько 250 ознак, які теоретично можна визначити без поглибленого дешифрування мережевих пакетів та використовувати для класифікації характеристик шифрованого потоку. З огляду на різноманітність як типів подій, які призводять до змін в ознаках трафіку, так і мережевих середовищ, протоколів, обладнання тощо, а також цілей, заради яких виконується аналіз, не доводиться сподіватися на існування якогось одного універсального алгоритму, який буде здатен демонструвати прийнятний рівень результатів в усіх випадках. Тому пошук нових алгоритмів та порівняльний аналіз вже існуючих наразі лишається актуальною науковою та прикладною задачею.

Для вивченням властивостей заснованих на метриках алгоритмів аналізу гістограм при вирішенні задач CDP було проведено серію експериментальних досліджень із застосуванням методу статистичного моделювання. Алгоритми CPD гуртуються на концепції ковзаючого вікна, яка полягає в порівнянні наборів значень параметра який обраний для моніторингу на двох інтервалах часу. Інтервал, що починається раніше, разом з набором значень параметру називають «базовим вікном», а інші — «поточним вікном». Однак моменти початку інтервалів мають різнитися не менш ніж на одиницю.

Сам алгоритм CPD є нескінченим циклом моніторингу (тобто після початку роботи закінчитися він може лише примусово). На першому кроці отримують дані з базового вікна та розраховують їх аналітичні характеристики. На другому кроці отримують дані поточного вікна і їх характеристики. На третьому кроці виконується порівняння характеристик базового та поточного вікон. Якщо відмінностей між вказаними характеристиками не виявлено, береться наступне поточне вікно та процедура продовжується з другого кроку. В разі, якщо між характеристиками виявлені значимі зміни (що є наслідком зміни в поведінці об'єкту моніторингу), алгоритм генерує

відповідний сигнал для сповіщення особи, яка відповідальна за прийняття рішення. Після цього поточне вікно приймається в якості нового базового вікна (тобто подальші зміни в об'єкті будуть визначатися відносно цього нового базового вікна) і алгоритм продовжує роботу починаючи з першого кроку.

У формулах нижче позначимо $f_B(i), f_C(i)$ — долю елементів вибірок відповідно з базового (B-base) та поточного (C-currant) вікон спостереження що потрапили до i-ого інтервалу гістограми, K — загальна кількість інтервалів в гістограмі. Множина можливих значень даних, яка розбивається на інтервали, визначається як об'єднання відповідних множин значень базового та поточного вікон.

Метрики для визначення відстані між гістограмами

Серед множини метрик визначення відстані між об'єктами в багатовимірному просторі були обрані шість метрик різних типів. Вибір диктувався з одного боку бажанням охопити широкий спектр зазначених метрик, а з іншого оминути включення до цього переліку метрик, які дуже схожі за визначенням, а отже — будуть завідомо давати сильно корельовані результати.

Евклідова відстань. (EUC). Найбільш традиційна ти широковживана метрика для вимірювання розбіжності між об'єктами в багатовимірному просторі.

$$D_{EUC} = \sqrt{\sum_{i}^{K} \left(f_B(i) - f_C(i) \right)^2}$$

Характерною особливістю цієї метрики для вимірювання відстані між об'єктами є те, що при її розрахунку суттєво більший вклад надають ті виміри (інтервали гістограми), в яких розбіжність між $f_B(i), f_C(i)$ велика, і суттєво менший ті, де ця розбіжність незначна.

Манхетенська (або блочна) відстань. (MNCHT).

$$D_{\text{MNCHT}} = \sum_{i}^{K} \left| f_B(i) - f_C(i) \right|$$

При використанні цієї метрики залежність покоординатних розбіжностей від їх величин нівелюється.

Відстань Бгаттачар'я (BHATT) [27]

$$D_{\text{BHATT}} = -\ln\left(\sum_{i=1}^{k} \sqrt{f_B(i) * f_C(i)} \right)$$

Слід зазначити, що незважаючи на назву «відстань», ця характеристика не є метрикою в строгому математичному сенсі, оскільки вона не забезпечує виконання нерівності трикутника. Позатим, цей показник добре зарекомендував себе в деяких прикладних багатовимірних задачах. Крім того він пов'язаний з цілою групою мір, що засновані на понятті ентропії, тому він також був включений до списку.

Усі зазначені вище метрики розбіжностей між гістограмами наслідуючи традиційний при вивченні об'єктів в метричних просторах підхід ніяк не враховують взаємне положення «координат» цього простору. Однак для гістограм виявляється вкрай важливим врахування взаємно розташування окремих «координат»-інтервалів. Це можна зрозуміти проаналізувавши приклад наведений на рис. 1.

ISSN 2663 - 4023

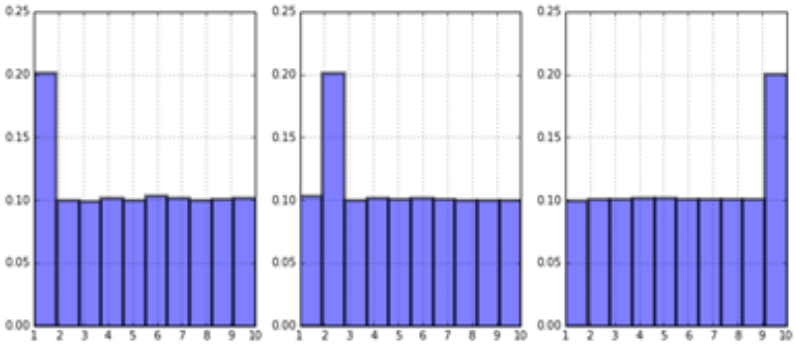

Рис.1. Приклади гістограм. Пояснення в тексті

Легко побачити що всі міри дадуть однакове значення відстані між кожною з пар наведених на малюнку гістограм. Однак враховуючи статистичний характер даних та відмінностей в них, ясно, що відмінність між лівою та центральною гістограмами може бути викликана випадковою флуктуацією даних. В той же час відмінність між лівою та правою — скоріш за все викликана суттєвими відмінностями в даних. Наступна метрика дозволяє врахувати вказані особливості.

Відстань «землечерпалки» (EMD) [28].

$$D_{EMD} = \frac{\sum_{i}^{K}\left|\sum_{j}^{i}\left(f_B(j) - f_C(j)\right)\right|}{K}$$

Інтуїтивно, якщо розглядати кожен розподіл як одиницю деякої маси (наприклад — грунту), метрика є мінімальною «вартістю» перетворення однієї «купи» маси на іншу; тобто є кількістю маси, яку необхідно перемістити, помноженої на середню відстань, на яку її потрібно перемістити. Через цю аналогію метрика відома в інформатиці як відстань «землечерпалки» (Earth Mover's Distance). Легко бачити, що таким чином визначена метрика симетрична, лежить в діапазоні від 0 (повна тотожність двох гістограм) до 1, причому наближення до 1 досягається як при кардинальній розбіжності законів розподілу даних з вікон спостереження, так і при тотожності закону, але наявності зміни в параметру положення даних.

Відстань перекриття гістограм (HI — Histogram Intersection) [13].

$$D_{HI} = 1 - \sum_{i}^{K} \min\left(f_1(i), f_2(i)\right)$$

Ця метрика, що початково була запропонованаа як один з інструментів для аналізу зображень, обчислює долю сукупної вибірки даних, яка тотожна для обох гістограм.

Модіфікована Хі-квадрат відстань (CHIM). Хі-квадрат відстань є загальновживаним і потужним інструментом, що використовується при аналізі гіпотези подібності розподілів даних та базується на використанні однойменного статистичного критерію [17]. Однак базовий критерій Хі-квадрат працює за умови відсутні інтервалів з нульовою кількістю елементів в гістограмі базового вікна.

$$D_{chi2} = \sum_{i}^{K} \frac{\left(f_C(i) - f_B(i)\right)^2}{f_B(i)}$$

При вирішенні практичних завдань, зокрема, коли зміни в моделі приводять до зсуву значень, при порівнянні гістограм виявляється, що інтервалів, які не містять елементів однієї з вибірок, стає все більше. Тому виникає ідея порівнювати не самі

ISSN 2663 - 4023

гістограми, а їх розбіжність з умовною гістограмою, елементи якої визначаються як середнє значення кожного з інтервалів вхідних гістограм.

$$D_{CHIM} = \frac{1}{2}\sum_{i}^{K}\frac{\left(f_B(i) - f_C(i)\right)^2}{\left(f_B(i) + f_C(i)\right)}$$

Вказаний підхід дає можливість використовувати дану метрику навіть в разі, якщо одне або обидва зі значень $f_B(i), f_C(i)$ дорівнюють нулю.

Визначення критичного значення для подібності гістограм

Вимір відстані між двома гістограмами може відповісти на питання «на скільки схожі» гістограми між собою у відповідному метричному просторі. Головною проблемою, що ускладнює використання метричного підходу до аналізу гістограм є відсутність об'єктивного показника, за допомогою якого можна визначити межу (критичне значення), при перетині якої можна вважати дві гістограми настільки різними, що логічним видається відхилення гіпотези про незмінність поведінки об'єкту дослідження. Аналогічною ознакою, на основі якої приймається відповідне рішення в разі застосування точкових статистичних критеріїв є значення статистики p_value [7], [8]. Однією із задач даного дослідження є створення процедури, яка дозволить для кожної з метрик подібності запропонувати таке критичне значення, а також дослідити його залежність від кількості інтервалів, які використовуються при побудові метрики та кількості значень в кожному з вікон спостереження.

Як було вказано, метрики, що описані вище ґрунтуються на аналізі об'єктів в багатовимірному метричному просторі. Такі метрики мають нижню границю своїх значень що дорівнює 0 і відповідає випадку абсолютної тотожності обох об'єктів. Однак в загальному випадку ці метрики не мають верхньої границі. Особливістю застосування метричного підходу до аналізу саме гістограм полягає в тому, що ми маємо додаткове важливе обмеження загального випадку, а саме:

$$\sum_{i}^{K} f_B(i) = \sum_{i}^{K} f_C(i) = 1$$

При цьому для всіх метрик (за виключенням D_{EMD}) виникає натуральне обмеження зверху, що дорівнює 2 і відповідає випадку, коли області значень базового та поточного вікон не перетинаються.

План експериментальних досліджень

Для вирішення задачі визначення критичних значень для метрик, було проведено ряд статистичних експериментів. В якості наборів значень що досліджувалися обрані закони розподілу P, а саме нормальні розподіли з параметрами $N(0,1)$, $N(3,1)$ та $N(0,3)$, безперервні рівномірні розподіли з параметрами $U(0,1)$, $U(1,2)$ та $U(0,2)$, а також логнормальні розподіли з параметрами $LN(0,1)$, $LN(0,1.5)$, $LN(0,2)$. Такий вибір дає змогу визначити, як буде змінюватися критичне значення в залежності як від зміни математичного очікування так і від зміни дисперсії генеральної сукупності.

Також досліджувалася залежність критичного значення від рівня значущості — тобто від прийнятої при дослідженні відповідної прикладної задачі імовірності ухвалити рішення про наявність відмінності в даних базового та поточного вікон спостережень, якщо насправді такої відмінності нема (помилка I роду). Оскільки значення рівня значущості — це позастатистичний параметр вибір якого виконується виключно з точки

зору прийнятності для прикладної задачі, були обрані стандартні значення $\alpha = [01, 0.025, 0.05, 0.075, 0.1]$.

Оскільки відомо, що кількість інтервалів гістограми має обиратися виходячи з загальної кількості елементів у вибірці (тобто для нашої задачі — від довжини вікна спостереження), виконувалися досліди при значеннях цієї величини $N \in [50,100,200,400,800]$. Відповідно, кількість інтервалів гістограм для кожного значення N вибиралися з ряду $n\_bin \in [10, 20, 40, 80, 160]$ за умовою, що максимально значення n_bin має бути $\leq \frac{N}{5}$.

Для кожної з комбінації описаних параметрів (P, N, n_bin) проводилося 10000 експериментів. Результати всіх експериментів використовувалися для визначення такого значення $D_*^{кр}$ для кожної із зазначених метрик, проте лише у α експериментах були отримані значення більші за $D_*^{кр}$. Ці значення можуть трактуватися, як визначені за методом Монте-Крало значення кордону інтервалу прийняття гіпотези про статистичну нерозрізненість розподілів даних базового та поточного вікна спостереження. На рис. 2. отримані результати представлені у вигляді графіку залежності значень $D_*^{кр}$ від N, α, n_bin для різних метрик. Крім того, отримані результати зберігаються у вигляді бази даних для подальшого використання.

Під час проведення експериментів аномалія кожного типу генерувалася 1000 разів, для кожного згенерованого зразка розраховувалися значення для обраних критеріїв CPD. Довжини вікон спостереження обиралися з значень 50, 100 та 200. Для метричних алгоритмів кількість інтервалів гістограм обиралися як 10, 20 та 40. Значення параметру α для всіх експериментів обиралося як 0.05. Використовувався метод ковзаючого вікна спостереження зі зсувом вікна на одну позицію на кожному кроці. Наприклад, при довжині згенерованого зразка часового ряду в 300 значень і довжині вікна спостережень 100, виконувалося максимально 200 кроків порівняння для кожного алгоритму та комбінації параметрів.

Порівняння алгоритмів проводилося за двома метриками, найважливішими у багатьох прикладних задачах, зокрема при онлайн моніторингу мережевого трафіку — кількості випадків виникнення помилок I роду («помилкових тривог»), що допускалося алгоритмом на всій множині наданих зразків, та затримки між моментом настання заміни та моментом її фіксації алгоритмом. Отримані результати представлені в табл. 1. Наведені дані свідчать, що розглянуті метричні алгоритми встановлення відмінностей між гістограмами показують результати співставні за якістю по обраним метрикам з результатами статистичних алгоритмів, а в деяких випадках і перевищують останні. Так, наприклад, при довжині вікна спостереження що дорівнює 50, та зміні моделі типу «sh2» (зсув при логнормальному розподілу значень) практично всі метричні алгоритми показують рівень помилок I роду нижчий за статистичні алгоритми, при цьому час затримки між моментом виникнення зміни та моментом її фіксації практично для всіх алгоритмів метричного аналізу гістограм теж виявився меншим. При збільшенні вікна спостереження до 200 і тому самому типу зміни моделі всі метричні алгоритми за виключенням алгоритму EUC зберігають перевагу над статистичними алгоритмами. Для більшості випадків змін типу «tr» (поява тренду) алгоритми метричного аналізу гістограм хоча і показують кількість помилок I роду більшу за статистичні алгоритмі, але перевершують останні — деякі досить суттєво — за часом затримки у визначенні заміни.

ISSN 2663 - 4023

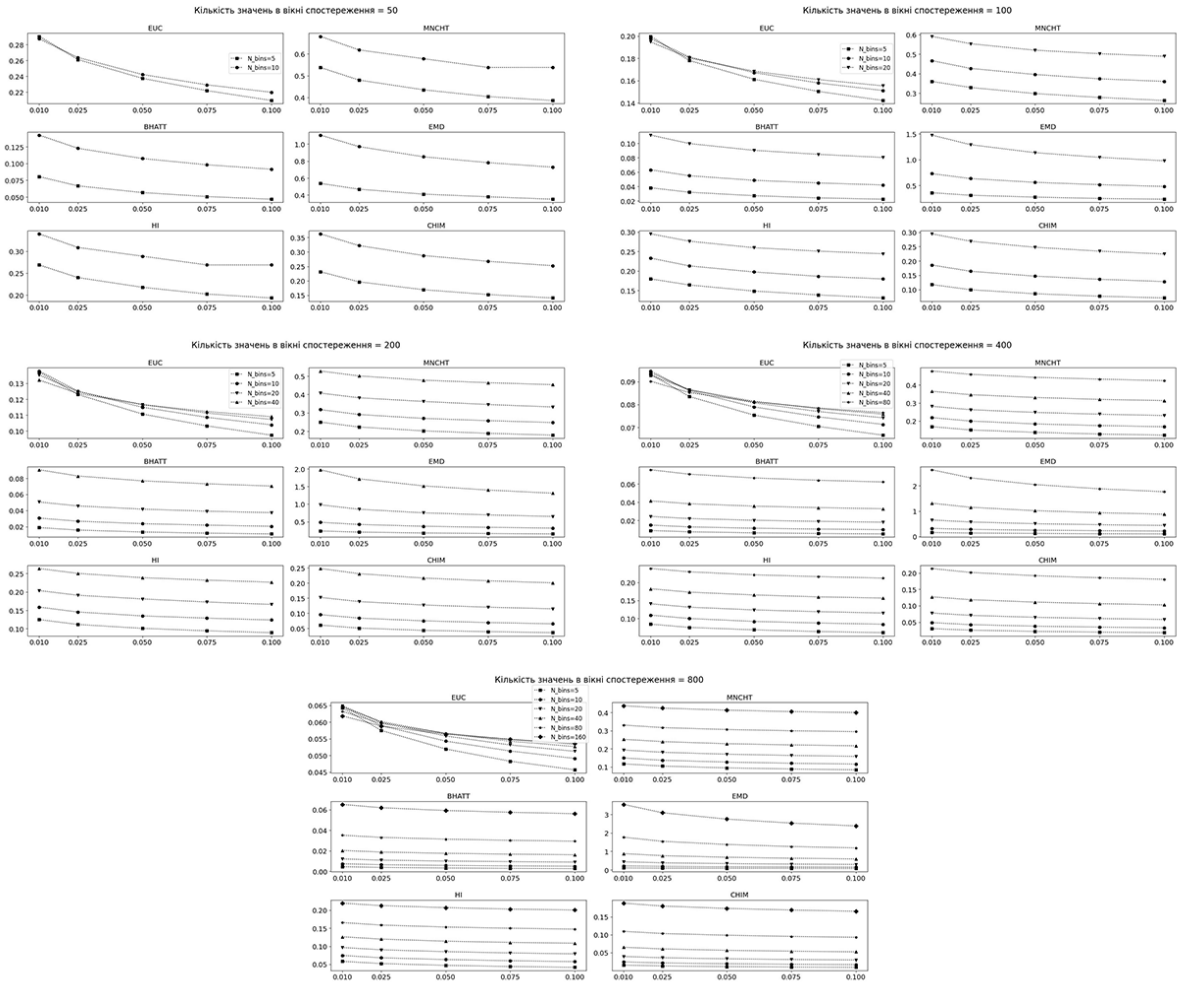

Рис. 2. Залежність критичного значення для метрики від довжини вікна спостереження, рівня значущості та кількості інтервалів гістограм

Реалізація алгоритмів для експериментів виконувалася у вигляді програмних застосунків створених мовою програмування Python.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Задача виявлення точок змін в поведінці об’єктів кібербезпеки шляхом аналізу значень параметрів цих об’єктів представлених у вигляді часових рядів відіграє важливу роль в системах кіберзахисту. Зокрема, з огляду на все більше домінування протоколів шифрування даних, важливість цих алгоритмів зростає і в галузі забезпечення безпеки інформаційно-комунікаційних систем.

Алгоритми виявлення точок змін у моделях часових рядів реалізовані у вигляді застосунків, розроблених мовою програмування Python. Ці алгоритми, як і створена база даних для визначення критичних значень для кожної з метрик, можуть використовуватися при вирішення практичних задач моніторингу, зокрема для задач виявлення змін в параметрах мережевого трафіку.

ISSN 2663 - 4023

Таблиця 1

Результати експерементальних досліджень

Тип тесту	Кількість інтервалів гістограми	Довжина вікна=50		Довжина вікна=100		Довжина вікна=200		Довжина вікна=50		Довжина вікна=100		Довжина вікна=200	
		Доля хибних спрацювань	Затримка розпізнавання	Доля хибних спрацювань	Затримка розпізнавання	Доля хибних спрацювань	Затримка розпізнавання	Доля хибних спрацювань	Затримка розпізнавання	Доля хибних спрацювань	Затримка розпізнавання	Доля хибних спрацювань	Затримка розпізнавання
		Тип зміни моделі — sh1						**Тип зміни моделі — scl1**					
TS		0.080	19.0	0.115	28.3	0.090	40.5	0.085	12.3	0.080	29.1	0.095	66.8
LV		0.045	18.7	0.060	32.7	0.095	60.6	0.045	20.8	0.095	25.2	0.090	35.5
KS		0.065	22.1	0.065	35.1	0.080	48.7	0.060	29.7	0.045	66.1	0.035	104.0
EUC	10	0.165	22.6	0.160	37.1	0.260	43.4	0.120	24.4	0.130	37.6	0.250	48.5
	20			0.130	44.7	0.185	60.4			0.090	42.8	0.180	61.0
	40					0.135	79.7					0.135	77.8
MNCHT	10	0.175	20.5	0.190	31.4	0.285	39.8	0.140	23.5	0.195	33.5	0.285	46.9
	20			0.220	33.2	0.195	52.1			0.190	39.6	0.225	58.2
	40					0.305	46.7					0.355	56.3
BHATT	10	0.085	23.8	0.095	38.0	0.110	53.0	0.085	24.0	0.110	29.2	0.145	38.7
	20			0.190	35.6	0.150	54.1			0.125	34.9	0.145	45.4
	40					0.180	54.6					0.175	57.6
EMD	10	0.085	20.4	0.085	30.6	0.090	43.6	0.060	32.9	0.085	54.3	0.080	81.7
	20			0.090	30.3	0.075	44.0			0.075	55.1	0.075	82.0
	40					0.070	44.5					0.085	83.1
HI	10	0.175	20.5	0.190	31.4	0.255	41.1	0.140	23.5	0.195	33.5	0.285	48.9
	20			0.285	29.7	0.195	52.1			0.245	35.8	0.225	58.2
	40					0.305	46.7					0.355	56.3
CHIM	10	0.120	22.5	0.100	36.6	0.150	48.0	0.100	26.1	0.095	35.7	0.130	46.5
	20			0.165	36.6	0.130	54.6			0.150	40.8	0.150	55.2
	40					0.160	55.0					0.180	68.2
		Тип зміни моделі — sh2						**Тип зміни моделі — tr**					
TS		0.065	25.1	0.055	47.3	0.115	61.3	0.040	32.9	0.085	44.1	0.095	52.7
LV		0.060	25.2	0.055	38.5	0.105	47.9	0.070	23.4	0.090	60.6	0.105	71.9
KS		0.065	28.6	0.055	67.1	0.075	101.3	0.045	34.4	0.040	50.5	0.070	58.0
EUC	10	0.125	18.7	0.190	53.4	0.205	96.4	0.105	24.7	0.260	42.3	0.260	50.2
	20			0.165	42.9	0.185	85.6			0.130	50.1	0.135	60.3
	40					0.175	68.2					0.105	71.9
MNCHT	10	0.050	15.1	0.055	55.6	0.075	122.4	0.150	24.6	0.290	35.9	0.280	45.3
	20			0.030	50.8	0.050	124.6			0.290	39.7	0.215	53.5
	40					0.025	126.3					0.280	48.2
BHATT	10	0.055	25.9	0.075	53.4	0.105	83.8	0.110	29.1	0.165	43.8	0.115	57.3
	20			0.040	58.9	0.105	113.2			0.155	43.3	0.130	56.4
	40					0.030	138.4					0.210	55.7
EMD	10	0.020	32.4	0.010	74.9	0.020	127.5	0.060	34.1	0.070	45.4	0.100	54.3
	20			0.010	70.7	0.015	131.4			0.070	46.8	0.090	55.4
	40					0.015	131.1					0.095	55.4
HI	10	0.050	15.1	0.055	55.6	0.070	124.1	0.150	24.6	0.290	35.9	0.265	46.3
	20			0.040	48.4	0.050	124.6			0.340	35.8	0.215	53.5
	40					0.025	126.3					0.280	48.2
CHIM	10	0.040	21.5	0.050	68.5	0.050	120.5	0.115	28.3	0.175	42.3	0.120	53.6
	20			0.015	58.9	0.040	135.0			0.160	43.1	0.120	57.1
	40					0.015	142.7					0.195	55.8

У результаті проведених досліджень показано, що алгоритми виявлення зміни поведінки об’єкту моніторингу за результатами аналізу його параметрів, які базуються на використанні метричного аналізу гістограм та на використанні методу ковзаючого вікна, дозволяють ефективно розширити набір засобів виявлення аномальних явищ в об’єктах спостереження. Вказані алгоритми у поєднанні із запропонованим в роботі підходом до визначення критичних значень областей прийняття/відхилення гіпотез, при використанні до часових рядів з різноманітними типами змін моделей поведінки, дозволяють отримати результати, які за якістю (кількістю помилок I роду та часу затримки між моментом появи та моментом фіксації зміни) не поступаються, а подекуди — перевищують якісні показники статистичних алгоритмів. Враховуючи обчислювальні витрати, що забезпечують виконання

метричних алгоритмів аналізу гістограм, представляється доцільним сконцентрувати подальші дослідження на вивченні застосування метричних алгоритмів в ансамблях, а також на розширенні і застосуванні інших метрик при аналізі гістограм наборів експериментальних даних.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. *Google Transparency Report*. (n. d.). https://transparencyreport.google.com/https/overview
2. *The role of streaming machine learning in encrypted traffic analysis - Help Net Security*. (2022). https://www.helpnetsecurity.com/2022/05/09/ml-encrypted-traffic-analysis/
3. The Challenges of Inspecting Encrypted Network Traffic. *Fortinet Blog*. (2022). http://www.fortinet.com/blog/industry-trends/keeping-up-with-performance-demands-of-encrypted-web-traffic
4. Alwhbi, I. A., Zou, C. C., & Alharbi, R. N. (2024). Encrypted Network Traffic Analysis and Classification Utilizing Machine Learning. *Sensors, 24(11)*. https://doi.org/10.3390/s24113509
5. Papadogiannaki, E., & Ioannidis, S. (2021). A Survey on Encrypted Network Traffic Analysis Applications, Techniques, and Countermeasures. *ACM Computing Surveys, 54(6),* 1–35. https://doi.org/10.1145/3457904
6. Encrypted Traffic Analysis: Use Cases & Security Challenges. *ENISA Report. European Union Agency for Cybersecurity (ENISA).* (2020). https://www.enisa.europa.eu/publications/encrypted-traffic-analysis
7. Schroth, C., Siebert, J., & Groß, J. (2021). Time Traveling with Data Science: Focusing on Change Point Detection in Time Series Analysis (Part 2). *Analytics, Big Data, Data Science, Fraunhofer IESE-Blog, Künstliche Intelligenz* Published. https://www.iese.fraunhofer.de/blog/change-point-detection
8. Mehrotra, K. G, Mohan, C. K., & Huang, H. M. (2017). Anomaly Detection. Principles and Algorithms. *Springer International Publishing AG* 2017. https://doi.org/10.1007/978-3-319-67526-8
9. Lakhina, A., Crovella, M., & Diot, C. (2005). Mining anomalies using traffic feature distributions. *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications - SIGCOMM '05. Philadelphia, Pennsylvania, USA.* https://doi.org/10.1145/1080091.1080118
10. Chen, L., & Dobra, A. (2013). Histograms as statistical estimators for aggregate queries. *Information Systems, 38(2),* 213–230. https://doi.org/10.1016/j.is.2012.08.003
11. Oliynyk, O., & Taranenko, Y. (2021). Automated system for identification of data distribution laws by analysis of histogram proximity with sample reduction. *Ukrainian metrological journal. NSC "Institute of Metrology", 3,* 31–37. URL: https://doi.org/10.24027/2306-7039.3.2021.241627
12. Rosenberger, J., Müller, K., Selig, A., Bühren, M., & Schramm, D. (2022). Extended kernel density estimation for anomaly detection in streaming data. *Procedia CIRP, 112,* 156–161. https://doi.org/10.1016/j.procir.2022.09.065
13. Cha, S.-H., & Srihari, S. N. (2002). On measuring the distance between histograms. *Pattern Recognition, 35(6),* 1355–1370. https://doi.org/10.1016/s0031-3203(01)00118-2
14. Bityukov, S. I., Krasnikov, N. V., Nikitenko, A. N., Smirnova, V. V. (2013). A method for statistical comparison of histograms. *Discrete and Continuous Models and Applied Computational Science, (2),* 324–330. https://doi.org/10.48550/arXiv.1302.2651
15. Wood, J. C. S. (2018). Non-Parametric Comparison of Single Parameter Histograms. *Current Protocols in Cytometry, 83(1), 2018.* 20p. https://doi.org/10.1002/cpcy.33
16. Lepskiy, A. (2018). On the Preservation of Comparison of Distorted Histograms. *International Journal of Information Technology & Decision Making, 17(01), 2018.* p 339–355. DOI:10.1142/s0219622017400028.
17. Gagunashvili, N. D. Tests for comparing weighted histograms. Review and improvements. *The European Physical Journal Plus, 132(5). 2017.* https://doi.org/10.1140/epjp/i2017-11481-1
18. van den Burg, G. J. J., & Williams, C. K. I. (2022). *An Evaluation of Change Point Detection Algorithms*. https://doi.org/10.48550/arXiv.2003.06222
19. Bharadiy, J. P. (2023). Machine Learning in Cybersecurity: Techniques and Challenges. *European Journal of Technology, 7(2),* 1–14. https://doi.org/10.47672/EJT.1486
20. Sokolov, V. V., Shapoval, O. M., & Sharadkin, D. M. (2020). An ensemble of algorithms for detecting anomalies in time series and its application to real-time monitoring of the state of systems. *Collection of scientific papers of VITI, 3,* 82–93.

21. Ryabtsev, V., Sharadkin, D., & Klyat, Y. (2021). A comparative study of algorithms for detecting change points in regression models of time series. *Information Technology and Security, 9(2),* 137–150. https://doi.org/10.20535/2411-1031.2021.9.2.249887
22. Truong, C., Oudre, L., & Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*. https://doi.org/10.1016/j.sigpro.2019.107299
23. Fesokha, V, Subach, I., Kubrak, V., Mykytiuk, A., & Korotaiev, S. (2020). Zero-Day Polymorphic Cyberattacks Detection Using Fuzzy Infetrence System. *Austrian Journal of Technical and Natural Sciences, 5-6,* 8–14. https://doi.org/10.29013/AJT-20-5.6-8-13
24. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Comput. Surv., 41(3),* 1–58. https://doi.org/10:1145/1541880:1541882
25. Aminikhanghahi, S. (2017). Cook D.J. A Survey of Methods for Time Series Change Point Detection. *Knowledge and information systems, 51(2),* 339–367. https://doi.org/10.1007/s10115-016-0987-z
26. Moore, A. W., Zuev, D., & Crogan, M. L. (2005). Discriminators for use inflow-based classification. *Technical report, RR-05-13, University of Cambridge.*
27. Bi, S., Broggi, M., & Beer, M. (2019). The role of the Bhattacharyya distance in stochastic model updating. *Mechanical Systems and Signal Processing, 117,* 437–452. https://doi.org/10.1016/j.ymssp.2018.08.017
28. Lee, S. M., Xin, J. H., & Westland, S. (2005). Evaluation of image similarity by histogram intersection. *Color Research & Application, 30(4),* 265–274. https://doi.org/10.1002/col.20122

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 1 (25), 2024
ISSN 2663 - 4023

Ihor Subach
Doctor of Technical Science, Professor, Head of the Special Department №5
Institute of Special Communications and Information Security National
Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
ORCID ID: 0000-0002-9344-713X
igor_subach@ukr.net

Dmytro Sharadkin
Candidate of Technical Sciences, Associate Professor,
Associate Professor of the Special Department №5
Institute of Special Communications and Information Security National
Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
ORCID ID: 0000-0001-6407-8040
dmsh@ukr.net

Ihor Yakoviv
Candidate of Technical Sciences, Associate Professor,
Associate Professor of the Special Department №5
Institute of Special Communications and Information Security National
Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
ORCID ID: 0000-0001-7432-898X
iyakov52@gmail.net

APPLICATION OF METRIC METHODS OF HISTOGRAM COMPARISON FOR DETECTING CHANGES IN ENCRYPTED NETWORK TRAFFIC

Abstract. With the increase in the share of encrypted traffic transmitted over the Internet, it has become impossible to directly identify the causes of anomalies in network behavior due to the lack of access to the contents of encrypted packets. This has significantly complicated the task of identifying information security threats. Only external symptoms are available for analysis, which manifest as changes in certain basic traffic parameters, such as volume, intensity, delays between packets, etc. As a result, the role and importance of algorithms for detecting changes in traffic have increased. These algorithms, using modern methods like machine learning, can identify various types of anomalies, including previously unknown ones. They analyze network traffic parameters which are available for direct measurement, presenting their development as time series. One of the least studied classes of such algorithms is the direct comparison of histograms of time series value distributions at different time intervals, particularly a subclass known as metric algorithms. These algorithms are based on the assumption that differences between histograms of time series values at adjacent observation intervals indicate changes in the flow of events that generate network traffic. However, the problem of measuring the difference or similarity between histograms, which are considered as objects in a multidimensional space, does not have a unambiguous solution. The paper analyzes existing histogram similarity metrics and describes a series of studies using statistical modeling. These studies evaluated the dependence of algorithm efficiency on external parameters and compared algorithms within this class to other change detection algorithms. This analysis made it possible to assess the practical application of these algorithms. The results showed that metric algorithms for comparing histograms can demonstrate high performance and, in some cases, outperform other known algorithms for detecting changes in time series. They ensure a reduction in the number of false positives and a decrease in the delay between the moment a change appears in the observed object and the moment it is detected by the algorithm.

Keywords: cybersecurity; cyber incident; machine learning; encrypted network traffic; time series; histogram similarity metrics; similarity detection algorithms.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. *Google Transparency Report.* (n. d.). https://transparencyreport.google.com/https/overview
2. *The role of streaming machine learning in encrypted traffic analysis - Help Net Security.* (2022). https://www.helpnetsecurity.com/2022/05/09/ml-encrypted-traffic-analysis/
3. The Challenges of Inspecting Encrypted Network Traffic. *Fortinet Blog.* (2022). http://www.fortinet.com/blog/industry-trends/keeping-up-with-performance-demands-of-encrypted-web-traffic
4. Alwhbi, I. A., Zou, C. C., & Alharbi, R. N. (2024). Encrypted Network Traffic Analysis and Classification Utilizing Machine Learning. *Sensors, 24(11)*. https://doi.org/10.3390/s24113509
5. Papadogiannaki, E., & Ioannidis, S. (2021). A Survey on Encrypted Network Traffic Analysis Applications, Techniques, and Countermeasures. *ACM Computing Surveys, 54(6),* 1–35. https://doi.org/10.1145/3457904
6. Encrypted Traffic Analysis: Use Cases & Security Challenges. *ENISA Report. European Union Agency for Cybersecurity (ENISA).* (2020). https://www.enisa.europa.eu/publications/encrypted-traffic-analysis
7. Schroth, C., Siebert, J., & Groß, J. (2021). Time Traveling with Data Science: Focusing on Change Point Detection in Time Series Analysis (Part 2). *Analytics, Big Data, Data Science, Fraunhofer IESE-Blog, Künstliche Intelligenz* Published. https://www.iese.fraunhofer.de/blog/change-point-detection
8. Mehrotra, K. G, Mohan, C. K., & Huang, H. M. (2017). Anomaly Detection. Principles and Algorithms. *Springer International Publishing AG* 2017. https://doi.org/10.1007/978-3-319-67526-8
9. Lakhina, A., Crovella, M., & Diot, C. (2005). Mining anomalies using traffic feature distributions. *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications - SIGCOMM '05. Philadelphia, Pennsylvania, USA.* https://doi.org/10.1145/1080091.1080118
10. Chen, L., & Dobra, A. (2013). Histograms as statistical estimators for aggregate queries. *Information Systems, 38(2),* 213–230. https://doi.org/10.1016/j.is.2012.08.003
11. Oliynyk, O., & Taranenko, Y. (2021). Automated system for identification of data distribution laws by analysis of histogram proximity with sample reduction. *Ukrainian metrological journal. NSC "Institute of Metrology", 3,* 31–37. URL: https://doi.org/10.24027/2306-7039.3.2021.241627
12. Rosenberger, J., Müller, K., Selig, A., Bühren, M., & Schramm, D. (2022). Extended kernel density estimation for anomaly detection in streaming data. *Procedia CIRP, 112,* 156–161. https://doi.org/10.1016/j.procir.2022.09.065
13. Cha, S.-H., & Srihari, S. N. (2002). On measuring the distance between histograms. *Pattern Recognition, 35(6),* 1355–1370. https://doi.org/10.1016/s0031-3203(01)00118-2
14. Bityukov, S. I., Krasnikov, N. V., Nikitenko, A. N., Smirnova, V. V. (2013). A method for statistical comparison of histograms. *Discrete and Continuous Models and Applied Computational Science, (2),* 324–330. https://doi.org/10.48550/arXiv.1302.2651
15. Wood, J. C. S. (2018). Non-Parametric Comparison of Single Parameter Histograms. *Current Protocols in Cytometry, 83(1), 2018.* 20p. https://doi.org/10.1002/cpcy.33
16. Lepskiy, A. (2018). On the Preservation of Comparison of Distorted Histograms. *International Journal of Information Technology & Decision Making, 17(01), 2018.* p 339–355. DOI:10.1142/s0219622017400028.
17. Gagunashvili, N. D. Tests for comparing weighted histograms. Review and improvements. *The European Physical Journal Plus, 132(5). 2017.* https://doi.org/10.1140/epjp/i2017-11481-1
18. van den Burg, G. J. J., & Williams, C. K. I. (2022). *An Evaluation of Change Point Detection Algorithms*. https://doi.org/10.48550/arXiv.2003.06222
19. Bharadiy, J. P. (2023). Machine Learning in Cybersecurity: Techniques and Challenges. *European Journal of Technology, 7(2),* 1–14. https://doi.org/10.47672/EJT.1486
20. Sokolov, V. V., Shapoval, O. M., & Sharadkin, D. M. (2020). An ensemble of algorithms for detecting anomalies in time series and its application to real-time monitoring of the state of systems. *Collection of scientific papers of VITI, 3,* 82–93.
21. Ryabtsev, V., Sharadkin, D., & Klyat, Y. (2021). A comparative study of algorithms for detecting change points in regression models of time series. *Information Technology and Security, 9(2),* 137–150. https://doi.org/10.20535/2411-1031.2021.9.2.249887
22. Truong, C., Oudre, L., & Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*. https://doi.org/10.1016/j.sigpro.2019.107299

23. Fesokha, V, Subach, I., Kubrak, V., Mykytiuk, A., & Korotaiev, S. (2020). Zero-Day Polymorphic Cyberattacks Detection Using Fuzzy Infetrence System. *Austrian Journal of Technical and Natural Sciences, 5-6,* 8–14. https://doi.org/10.29013/AJT-20-5.6-8-13
24. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Comput. Surv., 41(3),* 1–58. https://doi.org/10:1145/1541880:1541882
25. Aminikhanghahi, S. (2017). Cook D.J. A Survey of Methods for Time Series Change Point Detection. *Knowledge and information systems, 51(2),* 339–367. https://doi.org/10.1007/s10115-016-0987-z
26. Moore, A. W., Zuev, D., & Crogan, M. L. (2005). Discriminators for use inflow-based classification. *Technical report, RR-05-13, University of Cambridge.*
27. Bi, S., Broggi, M., & Beer, M. (2019). The role of the Bhattacharyya distance in stochastic model updating. *Mechanical Systems and Signal Processing, 117,* 437–452. https://doi.org/10.1016/j.ymssp.2018.08.017
28. Lee, S. M., Xin, J. H., & Westland, S. (2005). Evaluation of image similarity by histogram intersection. *Color Research & Application, 30(4),* 265–274. https://doi.org/10.1002/col.20122