УДК 004.7.056 DOI: 10.31673/2412-4338.2022.025459

Сьомін Д.А.
Державний університет телекомунікацій, Київ

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ КЛАСИФІКАЦІЇ ШИФРОВАНОГО ТРАФІКУ В КОРПОРАТИВНИХ МЕРЕЖАХ ЗА ДОПОМОГОЮ МАШИННОГО НАВЧАННЯ

Анотація. *В даний час зростає інтерес до завдань ефективного управління пакетними мережами, а саме: якість обслуговування, забезпечення інформаційної безпеки, оптимізація використання апаратно-програмних ресурсів мережі. Всі ці завдання значною мірою базуються на аналізі та класифікації мережевого трафіку.*

Класифікація трафіку дозволяє ідентифікувати пакети різних програм і сервісів і забезпечити їх пріоритезацію під час передачі по мережі. Сьогодні завдання є актуальним як з точки зору адміністрування мережі, так і з точки зору забезпечення її безпеки. З огляду на те, що зараз велика кількість додатків мають шифрування, особливий інтерес представляє класифікація трафіку, яка дає можливість опосередковано ідентифікувати аномалії в мережі.

Усе вищесказане підтверджує, що ідентифікація та класифікація трафіку мереж передачі даних є важливою темою дослідження, оскільки визначають основні етапи створення моделі управління трафіком при вирішенні завдань коректного застосування політики безпеки.

У статті розглядається проблема класифікації мережевого трафіку за допомогою методів машинного навчання. Представлені різні постановки задачі, характеристики, використані для її вирішення, описані існуючі підходи та області їх застосування. Аналізуються властивості мережевого трафіку, зумовлені характеристиками середовища передачі, а також використовуваними технологіями, які так чи інакше впливають на процес класифікації.

Ключові слова. *аналіз мережевого трафіку, мережна безпека, класифікація мережевого трафіку, машинне навчання*

Semin D.A.
State University of Telecommunications, Kyiv

INFORMATION TECHNOLOGY FOR CLASSIFICATION OF ENCRYPTED TRAFFIC IN CORPORATE NETWORKS USING MACHINE LEARNING

Abstract*. Currently, there is growing interest in the tasks of effective management of packet networks, namely: quality of service, ensuring information security, optimization of the use of hardware and software resources of the network. All these tasks are largely based on the analysis and classification of network traffic.*

Traffic classification allows you to identify packets of various applications and services and ensure their prioritization during transmission over the network. Today, the task is relevant both from the point of view of network administration and from the point of view of ensuring its security. Given the fact that a large number of applications now have encryption, traffic classification is of particular interest, which makes it possible to indirectly identify anomalies in the network.

All of the above confirms that the identification and classification of traffic of data transmission networks is an important topic of research, as they determine the main steps in creating a traffic management model when solving the problems of correct application of the security policy.

The article considers the problem of network traffic classification using machine learning methods. Various statements of the task are presented, the characteristics used for its solution, existing approaches and areas of their applicability are described. The properties of network traffic are analyzed, due to the characteristics of the transmission environment, as well as the technologies used, which somehow affect the classification process.

Keywords*. Network traffic analysis, network security, network traffic classification, machine learning.*

1. Постановка проблеми.

Кількість шифрованого трафіку в Інтернеті зростає з року в рік, а частка використання HTTPS протоколу наближається до 90% і продовжує зростати. Така популярність шифрування пов'язана з зростанням потреби в забезпеченні інформаційної безпеки, а використання технології VPN також дає користувачам можливість обходити місцеві блокування ресурсів та анонімізувати свій цифровий слід. В результаті шифрування частина даних перестає бути видимою, що ускладнює завдання класифікації трафіку.

Інтенсивне використання інтернет-додатків у різних аспектах життя призвело до збільшення обсягу переданого трафіку та одночасно до збільшення загроз безпеці інформації.

Реакцією на це стало вдосконалення способів захисту даних та користувачів. Найбільш відомим методом захисту є шифрування. Але при використанні шифрування управління трафіком та безпекою мережі стає складніше через неможливість здійснення перевірки вмісту зашифрованих пакетів. Ситуація посилюється ще тим, що шифрування часто використовується в обхід політики безпеки та правил використання мережевих ресурсів. В цьому випадку необхідно дотримуватися балансу між забезпеченням конфіденційності та забезпеченням мережевої безпеки при загрозі з боку зловмисників. Політика безпеки як на рівні мереж інтернет-провайдерів, так і на рівні глобальних мережевих операторів визначає правила поведінки з потенційно небезпечними додатками з точки зору їх впливу на економіку, політику, моральні підвалини нашого суспільства.

Для ефективної реалізації багатьох сценаріїв захисту від загроз інформаційної безпеки у мережах при аналізі мережного трафіку часто потрібна ідентифікація протоколу шифрування та типу програми, до якого цей трафік належить. Така класифікація корисна не тільки для запобігання атак і виявлення аномалій, але й аналізу поведінки користувача, управління трафіком, контролю продуктивності додатків.

2. Аналіз останніх досліджень і публікацій.

Поява нових програм та взаємодії між різними компонентами в Інтернеті різко збільшили складність та різноманітність цієї мережі, що зробило класифікацію трафіку складною проблемою. Нижче наведено деякі з найважливіших проблем класифікації мережного трафіку.

По-перше, збільшені вимоги до конфіденційності та шифрування даних користувача надзвичайно збільшили кількість шифрованого трафіку у сучасному Інтернеті. Процедура шифрування перетворює вихідну інформацію на псевдовипадковий формат з метою ускладнення її розшифровки. Як результат, зашифрована інформація мало містить характерних патернів для ідентифікації мережного трафіку. Тож точна класифікація зашифрованого трафіку стала реальною проблемою в сучасних мережах.

По-друге, багато з запропонованих підходів до класифікації мережного трафіку, такі як перевірка корисного навантаження та методи, засновані на машинному навчанні та статистиці, вимагають, щоб експерти самостійно витягували патерни або ознаки. Цей процес схильний до помилок, вимагає багато часу і витрат.

Нарешті, багато інтернет-провайдерів блокують P2P-додатки для обміну файлами через їхнє високе споживання пропускної спроможності та проблеми з авторським правом. Отже, щоб уникнути цієї проблеми, ці додатки використовують методи вбудовування протоколів і

обфускації для обходу систем управління трафіком. Ідентифікація такого роду додатків є одним із найскладніших завдань у класифікації мережного трафіку.

3. Мета і задачі дослідження.

Класифікація трафіку є важливим завданням у сучасних мережах, у тому числі бездротових. У зв'язку з швидким зростанням вимог до високошвидкісного трафіку для правильного розподілу мережевих ресурсів дуже важливо розпізнавати різні типи додатків, які їх використовують. Точна класифікація трафіку має величезне значення для складних завдань управління мережею, таких як забезпечення належної якості обслуговування (Quality of Service), виявлення аномалій, детектування атак тощо. Класифікація трафіку привернула великий інтерес як у академічних колах, і у промисловості, що з управлінням мережею.

Як приклад важливості класифікації мережного трафіку можна навести асиметричну архітектуру сучасних каналів доступу до мережі, яка була розроблена на основі припущення, що клієнти завантажують більше, ніж вони завантажують. Однак повсюдне поширення додатків симетричного попиту (таких, як point-to-point програми, VoIP (voice over IP) та відеодзвінки) вплинуло на вимоги клієнтів, через що класична асиметрична архітектура стала застарілою. Ключову роль таких ситуаціях має поняття якості досвіду (Quality of Experience). Деякі програми нечутливі до затримки інформації (текстові чати, відвідування сайтів), у той час як для відеозв'язку та потокових програм затримка часто критична. Таким чином, щоб забезпечити задовільну роботу пристрою клієнта, потрібне знання рівня додатків для виділення відповідних ресурсів для кожної програми.

4. Результати дослідження.

Окрім вивчення функціоналу програм для аналізу мережного трафіку, також необхідно враховувати збирані та аналізовані за їх допомогою показники мережевого трафіку, що безпосередньо впливають на якість передачі даних у корпоративній мережі підприємства.

До основних показників мережевого трафіку відносяться:

- тип трафіку;
- кількість підключень та обсяг трафіку.

Тип трафіку – це кількість переданих пакети за одиницю часу для конкретного протоколу. Тип трафіку залежить від протоколу пакета. Кожен пакет, що передається, має в заголовку поле протоколу (TCP/UDP/ICMP тощо). Різні протоколи відповідають за різні завдання, переважно: TCP – за гарантовану доставку пакета, UDP – за негарантовану доставку пакета (потокове відео), ICMP – за передачу службових даних.

Кількість підключень вузла корпоративної мережі – це кількість комп'ютерів, з якими відбувається обмін даними. Цей параметр включає в собі: вихідний трафік, що показує, на яке кількість комп'ютерів іде передача даних; вхідний трафік, що визначає, з якої кількості вузлів йде передача даних. Підозріла велика кількість підключень може означати, що комп'ютер у мережі виконує роль сервера роздачі. Великий обсяг трафіку при одному підключенні можна інтерпретувати як робочий відеодзвінок, або той, що не відноситься до роботи перегляд потокового відео.

Об'єм трафіку – обсяг інформації, переданої та отриманої в корпоративній мережі. Таким чином, метою дослідження є проектування програмного забезпечення для моніторингу характеристик трафіку в корпоративній комп'ютерної мережі.

Для реалізації зазначеної мети необхідно вирішити завдання, пов'язані з аналізом вже існуючих програмних продуктів вивчення мережевого трафіку; з визначенням ключових показників якості мережевого трафіку; з розробкою алгоритму збору характеристик мережі за певні часові відрізки; з організацією процесів зберігання та візуалізації результатів моніторингу мережевого трафіку; з розробкою програмного забезпечення для моніторингу мережного трафіку.

Загалом завдання класифікації мережевого трафіку може бути сформульоване наступним чином: отримання на вхід деяких характеристик мережевого трафіку з видачою на

виході класу, до якого цей вид трафіку відноситься. Вхідні характеристики можуть виступати як дані пакетів, так і різні частотні характеристики, як ідентифікатор конкретної програми, відповідального за генерацію цього трафіку, а також як ідентифікатор виду трафіку.

Це завдання є однією з центральних тем галузі організації мережевої взаємодії. Історично це завдання було найбільш актуальне в галузі керування трафіком для підвищення ефективності використання існуючих каналів зв'язку та якості послуг для кінцевих користувачів. Однак на даний момент актуальність цього завдання значно зросла, у зв'язку з розширенням її сфери застосування, в яку на даний момент входять як системи застосування політик, так і сфера інформаційної безпеки. Практично будь-яка система аналізу трафіку в тому чи іншому вигляді включає себе компонент класифікації.

Виділяють 3 підходи до класифікації мережевого трафіку – на основі аналізу портів, шляхом аналізу корисного навантаження та за допомогою машинного навчання.

Перший зіставляє кожну програму з відповідним номером порту (наприклад, порт 20 для FTP). Актуальність такого підходу помітно знизилася через запровадження динамічного розподілу портів. Другий спосіб слабо підходить для шифрованого трафіку, оскільки виділити корисне навантаження після шифрування практично неможливо. Тому найбільш популярним підходом до класифікації шифрованого трафіку в останні роки стало використання машинного навчання. Його можна розділити на 2 великі групи: класифікація на основі класичних алгоритмів та глибоке навчання (deep learning). Широке застосування вирішення завдання класифікації шифрованого трафіку набули методи глибокого навчання за допомогою нейромереж. З них допомогою досягаються найкращі результати. Проте їх використання має зворотній бік – висока обчислювальна складність. Якщо великі корпорації та компанії можуть собі дозволити використання нейромереж для аналізу трафіку, то для малого і середнього бізнесу найчастіше це стає недозволеною розкішшю. Отже, при виборі методу аналізу трафіку необхідно досягти балансу між якістю класифікації та обчислювальною складністю.

Крім глибокого навчання, завдання класифікації вирішуються за допомогою класичних алгоритмів машинного навчання, таких як логістична регресія, дерева рішень, випадковий ліс, градієнтний бустинг та інші. Моделі на основі цих алгоритмів менш точніші, але й значно менш вимогливі до обчислювальних ресурсів, тому можуть використовуватись користувачами з нижчим порогом входу. Предметом цього дослідження є класифікація шифрованого трафіку за допомогою класичних алгоритмів машинного навчання.

Алгоритми машинного навчання для своєї роботи потребують отримання ознак класифікованих прикладів, на основі яких прийматиметься рішення. Для завдання класифікації мережного трафіку можна виділити два основні способи їх отримання:

- ознаки виділяються на спеціальному етапі підготовки даних на основі деякої зовнішньої інформації (наприклад, знання експерта в галузі); такі ознаки можуть включати як вміст пакетів, так і додаткову інформацію про потік (таку, як загальна кількість пакетів, розмір переданих даних, час отримання пакетів і т.п.);
- ознаки виділяються з даних (зазвичай це вміст пакетів) самою моделлю в процесі так званого глибокого навчання.

Перший підхід вимагає від експериментатора додаткових зусиль, але дозволяє краще контролювати процес і використовувати будь-яку доступну інформацію та її похідні (наприклад, на основі наявних даних статистики). Другий підхід вимагає мінімальних дій при підготовці даних (нормалізувати дані, уніфікувати довжину прикладів, переупорядкувати дані, якщо це необхідно) і може дозволити знаходити неочевидні на перший погляд залежності, але набір ознак, що витягується ним, не завжди є кращим або мінімальним для вирішення поставленого завдання. Крім того, другий підхід зазвичай вимагає навчання більшої кількості даних.

За змістом ознак їх можна умовно розбити на такі класи:

- дані пакета;
- метаінформація про пакет;
- тимчасові характеристики;

- інформація про потік.

Алгоритм виявлення аномалій мережі може бути описаний наступним чином. Даними для аналізу є мережевий трафік, поданий як набір мережевих пакетів, у випадку фрагментованих лише на рівні IP. Зібрані сирі дані надалі слугують джерелом для формування необхідної інформації для подальшого аналізу.

Так, отримані дані можуть бути агреговані за певний часовий інтервал та нормалізовані з метою побудови ознакових атрибутів загального вигляду, які будуть потрібні при побудові поточного профілю активності. Створений набір ознак порівнюється з набором характеристик нормальної діяльності об'єкта (користувача або системи) - шаблоном нормальної поведінки.

Якщо спостерігається суттєва розбіжність порівнюваних параметрів, то фіксується мережева аномалія. В іншому випадку приймається рішення про те, що дані характеристики трафіку відносяться до нормальної поведінки.

Додавання та зміна шаблонів нормальної поведінки може здійснюватися як у ручному режимі, так і автоматично, причому деякі параметри, що задають налаштування поточного нормального профілю мережної активності можуть змінюватися в залежності від часу доби.

При розробці математичних моделей на основі методів машинного навчання обов'язково проводять експериментальні дослідження на модельованих чи зібраних даних. У даній роботі розглядаються питання розробки параметрів моделі, що працює в режимі реального часу, тому до бази даних висуваються такі вимоги:

- база даних має бути досить великою (кілька тисяч потоків і більше для одного класу), щоб дозволити будувати вибірки різної величини;
- у базі даних має міститися деяка кількість різних типів додатків, серед яких рекомендується відбирати як програми, близькі один до одного за форматом пакетів та інтенсивністю їх надходження, наприклад DNS та HTTP;
- потоки, відібрані для класифікації, повинні бути достатньо довгими, у менших масштабах завдання втрачає свій зміст;
- потоки повинні бути або промарковані в самій базі даних, або формат зібраних даних має бути підлягає маркуванню відомими засобами автоматичної класифікації.

Сучасні методи класифікації практично без винятків спираються на механізми навчання машин. При цьому на практиці найчастіше застосовують статистичні методи, коли вирішальне правило будується на основі навчального набору прецедентів, для яких приналежності до класів свідомо відомі. Опис завдання навчання класифікатора за прецедентами:

Нехай X – безліч досліджуваних об'єктів. Будемо називати ознакою (feature) наступне відображення $f: X \to D_f, D_f \subset R$ – множина допустимих значень ознаки f. Насправді зазвичай безліч D_f являє кінцеву безліч (у цьому випадку ознака називається номінальною), або збігається з безліччю дійсних значень (у цьому випадку ознака називається кількісним).

Нехай є набір ознак $f_1, \dots, f_n$. Тоді вектор $\left(f_1(x), \dots, f_n(x)\right)$ називається ознаковим описом об'єкта x $\subseteq$ X.

Сукупність ознакових описів всіх об'єктів підбірки $X^l = \{x_1 \in \mathrm{X}, \dots, x_l \in \mathrm{X}\}$ записану у вигляді матриці, називають матрицею об'єктів ознак:

$$F = \begin{pmatrix} f_1(x_1) & \cdots & f_1(x_1) \\ \vdots & \ddots & \vdots \\ f_n(x_1) & \cdots & f_n(x_1) \end{pmatrix} \qquad (1)$$

Матриця об'єктів-ознак є найпоширенішим способом представлення вихідних даних у прикладних задачах навчання машин.

Далі будемо вважати, що безліч X розбивається на M непересічних класів. Введемо безліч $Y = \{1, \dots, M\}$ – безліч індексів класів. Тоді існує відображення (яке називається цільова функція) $y^*: X \to Y$.

Нехай є підмножина $X^l = \{x_1 \in \mathrm{X}, \dots, x_l \in \mathrm{X}\}$, для якого відомі значення цільової функції $y_i = y^*(x_i)$. Завдання навчання з прецедентів полягає в тому, щоб за вибіркою x^l та

відповідним значенням цільової функції побудувати вирішальну функцію $a: X \to Y$, яка наближала б цільову функцію $y^*(x)$. на всій множині X найкращим чином.

Пари $(x_i y_i)_{i=1}^{l}$ називаються прецедентами. Сукупність всіх прецедентів називається навчальною вибіркою.

Оскільки за змістом вирішальна функція $a(x)$ повинна відповідати на питання «до якого класу належить об'єкт x», то називатимемо її класифікатором.

Як було сказано вище, при постановці завдання навчання з прецедентів необхідно знайти найкращу вирішальну функцію $a(x)$. Найкращою вважається функція, яка доставляє оптимальне значення вибраному функціоналу якості.

Будемо називати функцією втрат (loss function) невід'ємну функцію $L(a, x)$, що характеризує величину помилки вирішальною функцією а на об'єкті x. Якщо $L(a, x) = 0$, то відповідь $a(x)$ називається коректною (достовірною).

Функціонал якості вирішальної функції a на вибірці X^l визначається так

$$Q(a, X^l) = \frac{1}{l} \sum_{i=1}^{l} L(a, x_i) \qquad (2)$$

Функціонал $Q(a, X^l)$ також називають функціоналом середніх втрат або емпіричним ризиком [17], оскільки він обчислюється за емпіричними даними.

Що ж до функції втрат, то завдання класифікації практично найчастіше використовується функція-індикатор помилки:

$$L(a, x) = [a(x) \neq y^*(x)] = \begin{Bmatrix} 1 & a(x) \neq y^*(x) \\ 0 & a(x) = y^*(x) \end{Bmatrix} \qquad (3)$$

5. Висновки і перспективи подальших досліджень.

В даний час зростає інтерес до завдань ефективного управління пакетними мережами, а саме: якості обслуговування, забезпеченню інформаційної безпеки, оптимізації використання програмно-апаратних ресурсів мережі. Всі ці завдання багато в чому опираються на аналіз та класифікацію мережевого трафіку.

Класифікація трафіку дозволяє ідентифікувати пакети різних додатків та служб та забезпечити їх пріоретизацію при передачі по мережі. На сьогоднішній день завдання актуальне як з погляду адміністрування мережі, так і з погляду забезпечення її безпеки.

З огляду на те, що велика кількість додатків зараз має шифрування, особливий інтерес являє класифікація трафіку, яка дозволяє за непрямими ознаками визначити аномалії у роботі мережі.

Все вищесказане підтверджує, що ідентифікація та класифікація трафіку мереж передачі даних є важливою темою дослідження, оскільки визначають собою основні кроки зі створення моделі управління трафіком при вирішенні задач коректного застосування політики безпеки.

Список використаних джерел

1. Manish J., Hassn H.T. A Review of Network Traffic Analysis and Prediction Techniques (2015)

2. Usama M. et al., Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges, vol. 7, pp. 655-659, 2019.

3. Singh K., Guntuku S.C., Thakur A., Hota C. Big data analytics framework for peer-to-peer botnet detection using random forests (2014) pp. 488–497

4. Casas P., D'Alconzo A., Zseby T., Mellia M. Big-DAMA: Big Data Analytics for Network Traffic Monitoring and Analysis (2016) pp. 1–3

6. Chitrakar R., Huang C. Anomaly based Intrusion Detection using Hybrid Learning Approach of combining k-MedoidsClustering and Naive Bayes Classification. 8th International Conference in Wireless Communications, Networking and Mobile Computing (WiCOM). 2012, pp. 1–5.

7. Kesavulu Reddy E. Neural Networks for Intrusion Detection and Its Applications. Proceedings of the World Congress on Engineering. 2013. London, pp. 12–15.

8. Risso F., Baldi M., Morandi O., Baldini A., Monclus P. Lightweight, payload-based traffic classification: An experimental evaluation, in Proc. IEEE ICC, 2008,pp. 869-875.

8. Callado A., Kamienski C., Szabo G., Gero B., Kelner J., Fernandes S., Sadok D. A Survey on Internet Traffic Identification, Communications Surveys & Tutorials, Vol. 11, pp. 37-52.

9. Hong J.W., Park SU., Kang YM. Enterprise Network Traffic Monitoring, Analysis, and Reporting Using Web Technology. Journal of Network and Systems Management 9, 89–111 (2001).

References:

1. Manish J., Hassn H.T. A Review of Network Traffic Analysis and Prediction Techniques (2015)

2. Usama M. et al., Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges, vol. 7, pp. 655-659, 2019.

3. Singh K., Guntuku S.C., Thakur A., Hota C. Big data analytics framework for peer-to-peer botnet detection using random forests (2014) pp. 488–497

4. Casas P., D'Alconzo A., Zseby T., Mellia M. Big-DAMA: Big Data Analytics for Network Traffic Monitoring and Analysis (2016) pp. 1–3

6. Chitrakar R., Huang C. Anomaly based Intrusion Detection using Hybrid Learning Approach of combining k-MedoidsClustering and Naive Bayes Classification. 8th International Conference in Wireless Communications, Networking and Mobile Computing (WiCOM). 2012, pp. 1–5.

7. Kesavulu Reddy E. Neural Networks for Intrusion Detection and Its Applications. Proceedings of the World Congress on Engineering. 2013. London, pp. 12–15.

8. Risso F., Baldi M., Morandi O., Baldini A., Monclus P. Lightweight, payload-based traffic classification: An experimental evaluation, in Proc. IEEE ICC, 2008,pp. 869-875.

8. Callado A., Kamienski C., Szabo G., Gero B., Kelner J., Fernandes S., Sadok D. A Survey on Internet Traffic Identification, Communications Surveys & Tutorials, Vol. 11, pp. 37-52.

9. Hong J.W., Park SU., Kang YM. Enterprise Network Traffic Monitoring, Analysis, and Reporting Using Web Technology. Journal of Network and Systems Management 9, 89–111 (2001).