УДК 004.891

к.т.н., доц., **Джулій В.М.** (ХмНУ)
к.т.н., с.н.с. **Мірошніченко О.В.** (ВІКНУ)
Солодєєва Л.В. (ВІКНУ)

DOI: https://doi.org/10.17721/2519-481X/2022/74-07

МЕТОД КЛАСИФІКАЦІЇ ДОДАТКІВ ТРАФІКА КОМП'ЮТЕРНИХ МЕРЕЖ НА ОСНОВІ МАШИННОГО НАВЧАННЯ В УМОВАХ НЕВИЗНАЧЕНОСТІ

У роботі запропоновано метод класифікації додатків трафіка комп'ютерних мереж на основі машинного навчання в умовах невизначеності. Сучасні методи класифікація додатків трафіка комп'ютерних мереж (таких, як класифікація протоколів транспортного рівня за номерами портів) мають суттєві недоліки, що призводить і є причиною до зростання проведення досліджень в напрямку класифікація додатків трафіка комп'ютерних мереж. Стрімке зростання, за останні роки, типів та кількості мережевих протоколів транспортного рівня підвищують актуальність дослідження в даному напрямку, розробки відповідних алгоритмів та методів класифікації додатків трафіка комп'ютерних мереж, які забезпечують при цьому зниження обчислювальної складності. На сучасному етапі, задача, яка потребує термінового вирішення - класифікації додатків трафіка комп'ютерних мереж з використанням відповідних протоколів та алгоритмів шифрування.

Перспективним напрямком класифікації додатків трафіка комп'ютерних мереж є статистичні методи, які опираються на аналізі та виявленні статистичних характеристик IP-трафіка. Найбільш перспективними є інтелектуальний аналіз потоку даних, а також технології машинного навчання, які на сучасному етапі широко використовуються в суміжних областях науки. Вирішується задача дослідження та навчання по прецедентах - класифікація додатків трафіка комп'ютерних мереж на основі зазделегідь відомої сукупності атрибутів їх ознак, з метою вдосконалення технічної бази комп'ютерних мереж, а також теоретичної бази, при цьому забезпечення високих експлуатаційних та якісних показників мереж, на прикладі використання протоколів транспортного рівня (стека TCP/IP). Результат вирішення поставленої задачі полягає у віднесенні додатка, відповідно до правил навчальної вибірки, до одного з непересічних класів, які заздалегідь визначенні, який містить відповідні, але при цьому вже класифіковані додатки.

Статистичний аналіз та дослідження атрибутів інтернет додатків показав, що найважливіші атрибути, пов'язані зі зміною об'єму інтернет трафіка потоку даних, мають експоненційний вигляд. Для виявлення аномальних змін об'єму інтернет трафіка додатків для розрахунку середніх значень може бути використаний критерій Фішера. Для класифікації інтернет додатків у потоковому режимі даних, при безперервному надходженні потоку даних запропоновано алгоритм виявлення зміщення концепту (дрейфа) трафіка потоку даних. Детектор дрейфа Фішера базується на статистичних характеристиках атрибутів інтернет додатків, аналізуються з використанням ковзаючих вікон, які контролюють зміну трафіка поточних статистичних характеристик атрибутів додатків.

Ключові слова: моделі, класифікація додатків, комп'ютерні мережі, дрейф, трафік, ковзаюче вікно, машинне навчання.

Вступ. Класифікація додатків IP-трафіка комп'ютерних мереж є важливим завданням керуванням потоком даних, покращення експлуатаційних характеристик, техніко-економічних, а також захисту трафіка комп'ютерних мереж. Класифікація IP-трафіка комп'ютерних мереж дозволяє визначити структуру, тип додатка, також джерело програми. Системи класифікації додатків IP-трафіка комп'ютерних мереж використовуються в достатньо широкому спектрі функцій: забезпечення на достатньому рівні якості зв'язку, виконання та забезпечення політик інформаційної безпеки, а також при розробці відповідних алгоритмів, програмних продуктів, що забезпечують достатній рівень стан комп'ютерних мереж, діагностику, контроль, а також надають засоби набору статистичних даних, виявлення проблем комп'ютерних мереж.

Сучасні методи класифікація додатків трафіка комп'ютерних мереж (таких, як класифікація протоколів транспортного рівня за номерами портів) мають суттєві недоліки, що призводить і є причиною до зростання проведення досліджень в напрямку класифікація додатків трафіка комп'ютерних мереж. Стрімке зростання, за останні роки, типів та кількості мережевих протоколів транспортного рівня підвищують актуальність дослідження в даному напрямку, розробки відповідних алгоритмів та методів класифікації додатків трафіка комп'ютерних мереж, які забезпечують при цьому зниження обчислювальної складності. На сучасному етапі, задача, яка потребує термінового вирішення - класифікації додатків трафіка комп'ютерних мереж з використанням відповідних протоколів та алгоритмів шифрування.

Перспективним напрямком класифікації додатків трафіка комп'ютерних мереж є статистичні методи, які опираються на аналізі та виявленні статистичних характеристик IP-трафіка. Найбільш перспективними є інтелектуальний аналіз потоку даних, а також технології машинного навчання, які на сучасному етапі широко використовуються в суміжних областях науки.

Аналіз останніх досліджень та постановка задачі. Методи класифікація об'єктів, широко використовуються переважно в економічних дослідженнях, які можна вбудувати в область комп'ютерних мереж та телекомунікаційних досліджень [1-12]. Однак в даних роботах не отримали на достатньому рівні відображення як теоретичні так і практичні питання класифікації додатків трафіка комп'ютерних мереж, які використовують протоколи транспортного рівня (стек TCP/IP) в умовах невизначеності, при наявності «фонового» трафіка, а також проведення оцінки ефективності алгоритмів, які в основі реалізують методи машинного навчання при наявності режиму потокового надходження даних. Більшість використовуваних алгоритмів машинного навчання з учителем призначені для навчання мультикласових або бінарних класифікаторів. На основі навчального набору даних трафіка, що складається з екземплярів двох класів, біномні (бінарні) класифікатори вибирають між класами об'єктів. Мультикласові (мультиноміальні) класифікатори розподіляють екземпляри на множину класів згідно з тренувальним набором даних, що складається з екземплярів усіх класів. Дані типи класифікаторів засновані на припущеннях: всі класи відомі наперед; для кожного класу існує ефективний і показовий набір потоку даних.

Таким чином, класифікатори з учителем нездатні визначити екземпляри класу, які не представлені у навчальній вибірці простору ознак. Ідентифікація невідомого об’єкта типу трафіка є найважливішою вимогою на сьогоднішньому етапі ідентифікації мережного трафіку, оскільки у зв'язку з розвитком Інтернету з'являються нові інтернет протоколи та додатки, які на даний час невідомі, або представлені не в повній мірі на момент навчання. Також, навіть для існуючих протоколів та мережних додатків дорого і важко отримати повноцінний позначений набір потоку даних, які характеризують відповідні класи. Для того, щоб розробити практичний класифікатор мережевого трафіка з використанням методів машинного навчання з учителем, необхідно бути скурпульозним з визначенням відповідного класу та побудовою даних тренувального набору.

Розглянемо вплив фонового невідомого трафіку на якість класифікації з використанням методів машинного навчання. Будемо розглядати як «корисні» мережні протоколи та додатки DNS, HTTP, BitTorrent, Steam, Skype.

Крім перевірки роботи алгоритму на тестовій вибірці, яка має класовий склад, як і навчальна, оцінка якості класифікації здійснювалася за умов присутності домішок фонового трафіка, в тестовій вибірці присутні екземпляри класів, відсутніх у навчальній вибірці простору ознак. Така ситуація, коли в тестовій вибірці, яка ідентифікується, присутній мережевий фоновий трафік, більш наближена до реальності, в силу множини протоколів, що використовуються в Інтернет мережі. Такий набір даних дозволяє оцінити роботу якості класифікації алгоритму в реальній ситуації.

В умовах відсутності фонового трафіка потоку даних алгоритми C4.5 і Random Forest мають найкращі показники оцінки якості класифікації мережевих додатків. Однак за наявності в наборі даних фонового трафіка оцінка якості класифікації мережевих додатків суттєво

знижується, для алгоритму Random Forest зниження оцінка становить 15%, а для C4.5 оцінка зниження досягає 20%.

Класифікація мережевих додатків в наявності в наборі даних фонового трафіка показала, алгоритми машинного начання з учителем, не здатні визначити фоновий трафік, що призведе до неминучих та критичних помилок класифікації мережевого трафіка.

Розвитком у даному напрямку є застосування інших методів кластеризації (алгоритмів навчання), для визначення та розмежування мережевих невідомих типів трафіка потоку даних, які вже потім класифікуються та аналізуються відповідними системами.

При використанні змішаних даних набору трафіка, що складаються з великої кількості «корисних» екземплярів потоків даних і невеликої кількості фонових екземплярів, в даному випадку можливе застосування технік кластеризації для розподілу трафіка потоків даних в декілька груп відповідно до подібності статистичних показників трафіка.

Технології кластеризації важлива задача класифікації трафіка, на практиці отримання повного «корисного» трафіка набору даних для навчання є трудомістким та складним процесом. Одним з напрямків вирішення задачі – розробка нових шаблонів, які представлятимуть невідомі додатки або зміни в існуючих класах класифікації. Отриманий трафік у формі мережевих інтернет пакетів збирається в мережеві потоки даних, на основі вищезазначених п'яти параметрах для класифікації. Для проведення кластеризації кожен із потоків описується значеннями заздалегідь визначеним набором властивостей, тобто. точкою $x = (x_1, ..., x_d)$ в d – вимірному просторі ознак трафіка, d – кількість ознак в просторі ознак. На даній стадії може проводитись попередня обробка атрибутів, трансформації та їх відбору.

Фонові вектори властивостей можуть бути наперед оброблені алгоритмами кластеризації, які розподіляють трафік на по відстані. Задачею даного етапу створення чистих кластерів.

Основна частина. Для класифікації трафіка потоку даних в режимі on-line кластери трафіка необхідно пов'язати з конкретними класами інтернет додатків, і на їх основі побудувати класифікатори. Простим рішенням задачі є ручна ідентифікація потоків у кожному кластері даних з наступним розподіленням цих кластерів відповідно до потоків даних. Інший підхід полягає в подачі на вхід алгоритму змішаних даних, які складаються з «корисних» екземплярів та невеликої кількості фонових екземплярів потоку. Промарковані потоки даних, які містяться в кластерах, будуть використані для найменування кластерів. Розглянемо отримання максимально чистих кластерів трафіка потоку даних. У методі кластеризації, заснованому на «відстані», кластери представлені центральними точками (центроїдами), а екземпляри, відносяться до найближчої точки відповідно до метрики відстані (Евклідова відстань). Метод K-Means відносить екземпляри трафіка до кластерів з найближчим середнім значенням, а потім перетворює на локальний мінімум суми квадратів відстаней між кожним екземпляром трафіка потоку даних і центром кластера. Метод кластеризації, заснований на ймовірності, екземпляри з певною ймовірністю можуть бути віднесені найбільш можливого кластера. Традиційні методи кластеризації ґрунтуються на шаблонах в Евклідовому просторі ознак інтернет трафіка та припущенні, що всі ознаки, при цьому мають однакову вагу в кластеризації.

Методи неконтрольованого навчання (навчання без вчителя) значно поступаються алгоритмам Random Forest навчання з вчителем. Таким чином метод DBSCAN не придатний в режимі online, може бути критичним для ідентифікації додатків, що генерують мережеві інтернет потоки, у системах забезпечення інформаційної безпеки. У подібних системах мережний трафік потоку даних неоднорідний і може в процесі змінюватися, що призведе за собою постійного перенавчання алгоритму. Якість оцінки роботи алгоритмів DBSCAN і *k*-Means, в значній мірі залежить від мережевого типу трафіка, який класифікується. Для мережевих протоколів DNS і BitTorrent більш придатним - алгоритм DBSCAN, для мережевого трафіка Steam і SSL – *k*-Means, результати для класів додатків Skype та HTTP у розглянутих алгоритмів близькі.

Проведений аналіз характеристик алгоритмів класифікації з використанням фонового трафіка показав зниження якості оцінки класифікації додатків. Застосування різних методів кластеризації також показали низькі оцінки якості кластеризації.

Розглянемо кластеризацію мережного трафіка потоку даних, що базується на методі Random Forest. Даний підхід застосовується в біометричних дослідженнях для пошуку кластерів даних геномної послідовності. Random Forest один із контрольованих алгоритмів навчання, показав найкращі результати у задачах класифікації мережевого трафіка потоку даних. Random Forest забезпечує високу точність, дає незміщену оцінку помилки під час навчання, а також оцінку близькості між парами трафіка вхідних точок потоку даних, що надає можливість використовувати його для кластеризації трафіка вхідного потоку даних. Для побудови «лісу» необхідно визначити два базових параметри: кількість змінних, для розподілу вузлів (m) та кількість дерев (n). Для побудови дерева рішень метод генерує кореневий вузол шляхом випадкового відбору N точок даних трафіка з навчальної вибірки, де N - розмір тренувального набору ознак. Потім ітеративно розділяє вузли на основі m змінних, за критерієм індекса Джіні. Дерева рішень будують настільки великими, наскільки це можливо, без відсікання гілок. На початковому процесу відбору навчання близько третини всіх точок потоку даних залишаються поза набором, Ці дані, в подальшому, можна використовувати для оцінки помилки класифікації. Коли дерева рішень побудовані, дані можуть бути пропущені через отриманий «ліс» і також обчислені міри близькості кожної пари даних точок. Якщо дві точки даних потрапляють в листовий вузол, їхня близькість збільшується на одиницю, близькості нормалізуються шляхом поділу на число дерев. Таким чином створюється симетрична матриця близькості P, кожен елемент матриці має значення в інтервалі [0, 1]. За наявності фонових даних безпосередньо збудувати дерево неможливо. Для виділення показника близькості необхідна штучна класифікації, коли випадковий «ліс» будується шляхом розподілення штучних даних від вхідних. Результуючий ліс і значення близькості точок великою мірою залежить, в даному випадку від складу штучних об'єктів.

Метрики оцінки якості кластеризації простору ознак. Для оцінки ефективності кластеризації простору ознак алгоритмів використовувалися наступні метрики:

1. Однорідність - величина, що приймає максимальне значення - 1, якщо в кластер входять екземпляри одного класу. Близькість результатів кластеризації до визначається шляхом оцінки ентропії класів з урахуванням запропонованих кластерів:

$$H(C \mid K) = -\sum_{k=1}^{|K|}\sum_{c=1}^{|C|}\frac{n_{c,k}}{n}\cdot\log\left(\frac{n_{c,k}}{n}\right) = 0. \quad (1)$$

Отримане значення нормується за допомогою максимальної ентропії, яку може забезпечити, в даному випадку кластеризація - ентропії класу. Таким чином, однорідність кластера визначиться як $h = 1 - \frac{H(C \mid K)}{H(C)}$, де $H(C \mid K)$ – ентропія класу при умові кластера, а $H(C)$ – ентропія класу.

2. Повнота (completeness) – величина, симетрична однорідності кластера $c = 1 - \frac{H(K \mid C)}{H(K)}$,

де $$H(K \mid C) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{n_{c,k}}{n}\cdot\log\left(\frac{n_{c,k}}{n}\right). \quad (2)$$

3. V-міра - визначається обчисленням середнього гармонійного повноти та однорідності: $V = \frac{2\cdot c\cdot h}{c+h}$. Дана метрика не залежить від абсолютних значень міток кластера чи класу: перестановка значень не змінить значення оцінки класифікації.

4. Коефіцієнт силуету обчислюється для кожного екземпляра окремо:

$$s = \frac{b-a}{\max(a,b)},$$

де a - середня відстань від даного екземпляра до інших екземплярів кластера,

b - середня відстань від даного екземпляра всіх екземплярів найближчого кластера.

Коефіцієнт силуету для набору екземплярів визначається як середнє значення коефіцієнта силуету для кожного визначеного зразка. Коефіцієнт силуету приймає значення від -1 до 1. Від'ємне значення - неправильна кластеризаці, екземпляр поміщений не в той кластер. Значення близько нуля кластери перекриваються. Значення коефіцієнта силуету чим ближче до 1, тим щільніше розділені кластери.

5. Незміщений індекс Ранда - обчислює міру подібності між реальними значеннями міток і результатом кластеризації, при цьому розглядаються всі пари екземплярів, підраховуються пари, які призначені при класифікації в одні або різні кластери і класи, розраховується наступним чином:

$$RI = \frac{a+b}{C_2^n},$$

де a – кількість пар екземплярів у вибірці простору ознак, які потрапили в один кластер та клас;

b - кількість пар екземплярів у вибірці, що потрапили в кластер, що не відповідає даному класу, n - кількість елементів у вибірці простору ознак. Індекс Ранда дорівнює 1 при співпаданні результатів кластеризації додатків з істинними значеннями «корисних» об'єктів.

При застосуванні метрики Індекс Ранда навіть випадковий розкид екземплярів за класами (кластерами) матиме додатню оцінку. Для правильної оцінки випадкової кластеризації необхідно нормувати даний індекс:

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]},$$

де $E[RI]$- очікуваний зміщений індекс Ранда. Індекс Ранда може приймати значення від -1 до 1. Випадкове присвоєння міток об'єктам буде мати показник, близький до 0, для будь-яких кількостей класів і кластерів, Від'ємні значення позначають погану кластеризацію, правильна кластеризація має додатні значення Індекса Ранда. При ідеальному співпадані кластерів та класів Індекс Ранда дорівнює 1.

Класифікація додатків комп'ютерних мереж є процесом передбачення невідомого атрибута відповідного класу елемента, використовуючи модель, навченої на тренувальному набору потоку даних. На відміну від традиційних підходів класифікації, потокові методи класифікації не можуть оперувати з об'ємом потоку даних, який поділяється на тестовий та тренувальний набори, таким чином, тестування та побудова моделі необхідно здійснювати на льоту.

Вузьким місцем класифікації мережних потокових даних є необхідність аналізу за один перегляд. Однопрохідний перегляд та аналіз потоку даних не запам'ятовує зміни, що відбулися в моделі з початку обробки потокових даних.

Таким чином, процес класифікації потокових даних може вимагати побудови моделі та її тестування в змінному середовищі трафіка. Процес тестування моделі відбувається у постійній конкуренції з процесом тренування. Обчислювальні методи потокового аналізу даних повинні використовувати статистику та теорію обчислень. Швидкість мережевих потокових даних та великий об'єм, висувають додаткові вимоги до ресурсів у системі кластеризації. На сьогодні розроблено ряд підходів до обробки мережевих потокових даних. Дані методи дозволяють застосовувати алгоритми машинного навчання мережевих потокових

даних. Особливістю мережевого потокового режиму є дрейф (зміщення) концепту, в результаті поточних змін в атрибутах мережевого аналізованого трафіка потоку даних. Зміни виникають при появі зміни інтенсивності атрибутів, нових пристроїв в мережі. Таким чином, зміни відображаються в мережних об'єктах і знижують точність оцінки класифікаторів, побудованих на навчальних об'єктах отриманих раніше.

Зміщення концепту (дрейф) – виникає, коли розподіл вхідних потокових значень x і отриманих результатів y змінюється у часі. У навчанні з учителем дрейф впливає на умовну ймовірність вхідного значення $P(x|y)$, на оцінку ймовірності $P(y|x)$, на результуючий розподіл $P(y)$, на сам розподіл вхідних значення $P(x)$. Зміщення концепту (дрейф) - поняття між моментом часу t_0 і моментом часу t_1 визначається як $\exists X : p_{t_o}(X,y) \neq p_{t_1}(X,y)$, де p_{t_o} спільний розподіл потокових даних у момент t_0 між потоковим набором вхідних змінних X і цільової функцією y. Таким чином можуть змінитися умовні ймовірності класу $p(x|y)$, ймовірності класів $p(y)$, (y); в результаті змінюються ймовірності класів $p(y|x)$, впливаючи на результати кластеризації прогнозування.

Більшість методів зміни концепту використовують, для вирішення задачі, часове вікно, обробляють атрибути в часовому вікні, і «забувають» інформацію про «минуле» даних атрибутів. Методи використовують часові вікна, при цьому припускають, що важлива інформація лише останніх атрибутів. Таким чином, адаптивне навчання проводить оновлення прогнозуючих моделей трафіка у режимі онлайн, щоб адекватно реагувати на дрейф концепту.

Адаптивний алгоритм ADWIN для виявлення змін, використовує часове вікно. Нехай задана послідовність дійсних чисел - $x_1, x_2, x_3, ..., x_t$. На вхід алгоритму ADWIN необхідна вхідна послідовність на інтервалі [0,1], необхідне масштабування вхідних даних. Визначимо μ_t математичне очікування x_t, що підпорядковується D_t. Алгоритм ADWIN не передбачає конкретного розподілу даних і використовує фіксованого розміру W ковзне вікно з новими значеннями x_t. Позначимо $\tilde{\mu}_W$ середнє значення спостережень в W, μ_W - невідоме середнє значення μ_t для $t \in W$. Як тільки дві частини W демонструють середні значення, що відрізняються, алгоритм ADWIN вирішує, що математичні очікування цих частин відрізняються і стара частина вікна відкидається.

Основним обмеженням використовуваних методів виявлення дрейфа, які в основі використовують моніторинг двох розподілів, у порівнянні з детекторами послідовної обробки потоку даних, є вимоги до пам'яті. Основною перевагою методів виявлення дрейфа є точніша локалізація моменту дрейфу (із затримкою не менше W вибірок). Недоліком методу - він не враховує реального розподілу потоку даних, і вимагає, щоб дані знаходилися в інтервалі [0,1].

Алгоритм виявлення дрейфа (зміни концепту) за критерієм Фішера. Нехай заданий спостережуваний потік даних $Y = \{y_0, y_1, y_2, ..., y_{N-1}\}$, де y_t - значення елементів потоку даних (атрибутів і додатків), виміряне в $t \in T = \{0,1,2,...,N-1\}$, N - розмір множини Y. Виявлення зміни концепту (дрейфу) додатків здійснюється з використанням ковзних вікон W_1 і W_2, які контролюють зміну поточних статистичних характеристик додатків і атрибутів, як показано на рис. 1.

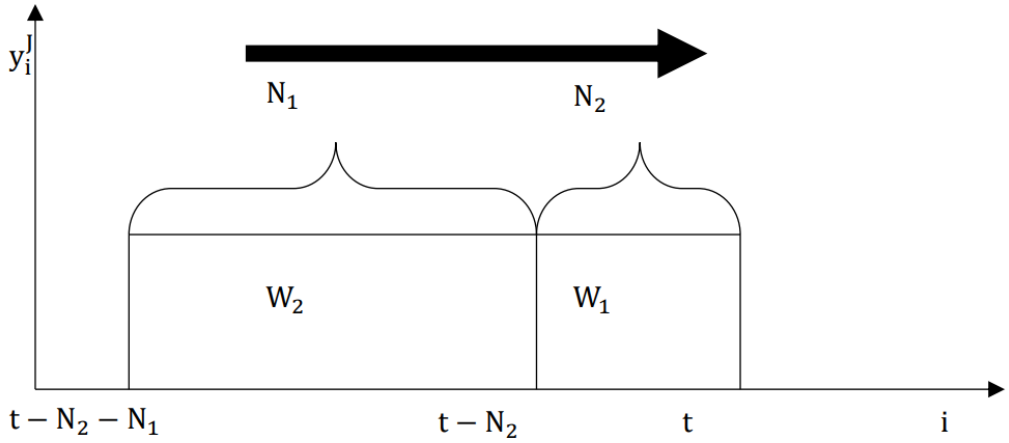

Рисунок 1 – Динаміка аналізу додатків за допомогою вікон W_1 і W_2

Вікно W_1 характеризує статистику атрибутів мережевого додатка в «минулому»:

$$Y_1^J = \left\{ y_{ij}^J ; j = \overline{1,K}; i = \overline{1;N_1} \right\}. \quad (3)$$

Вікно W_2 характеризує статистику атрибутів мережевого додатка на «поточний час»

$$Y_2^J = \left\{ y_{ij}^J ; j = \overline{1,K}; i = \overline{1;N_2} \right\}, \quad (4)$$

де, y_{ij}^J *j*-й атрибут додатка *J*, для *i*-го інтервалу спостереження; *K* - кількість атрибутів додатка *J*; N_1 - об'єм вікна пам'яті; N_2 - об'єм вікна аналізу $0 < N_2 < N_1$..

Використання алгоритму «ковзаючих вікон» дозволяє визначити незначні аномалії в реальному часі для виявлення зміни концепту (дрейфу) *i*-го мережевого додатка пропонується використовувати статистику, на основі статистичних даних приймається рішення про зміни концепту (дрейфу). Пропонується використовувати статистику відповідно до критерію Фішера для середніх значень вікон:

$$R_t^J = \frac{M_{W_2}^J(t)}{M_{W_1}^J(t)} > \lambda, \quad (5)$$

де, $M_{W_1}^J(t) = \frac{1}{N_2 K} \sum_{i=t}^{t-N_2} \sum_{j=1}^{K} y_{ij}^J$, $M_{W_2}^J(t) = \frac{1}{N_1 K} \sum_{i=t-N_2}^{t} \sum_{j=1}^{K} y_{ij}^J$, .

Перевищення порогового рівня вирішальної статистики $R(t) > \lambda$ свідчить про зміну характеристик мережевих додатків, і вказує на необхідність перенавчання поточного класифікатора. Збільшення порогу λ призводить до зменшення помилкових спрацьовувань класифікації трафіка, це можна призвести до пропуску зміни концепту (дрейфу), чи до затримок у виявленні дрейфу.

Для проведення практичної перевірки роботи запропонованого методу виявлення зміни концепту (дрейфу) в якості вхідних даних взято трафік мобільного інтернет додатка «Instagram». Для зміни тренда проведено множення значень отриманих ознак, що характеризують корисне навантаження трафіка на мережному та транспортному рівнях. В результаті запропонованого перетворення отриманий потік, який описується мережевим графіком, наведеним на рис. 2. Кожна точка наведеного графіка описується співвідношенням $D(t) = \sum_{j=1}^{K} y_{tj}$, де y_{tj} - значення *j*-го атрибута аналізованого інтернет додатка на інтервалі

часу - t. В правій частині графіка (рис. 2а) наведено результат множення ознак. Графіки (рис. 2б, рис. 2в) відображають значення, що спостерігаються відповідно у вікнах W_1 і W_2 у процесі проведення аналізу трафіка потоку даних. На рис. 2, через різке наростання тренда після $2{,}5\times10^5$, у вікні W_2 відбувається різке збільшення отриманого середнього значення. Збільшення середнього значення призводить до різкого наростання (стрибка) значення $R(t)$, що наведено на рис. 2г.

Таким чином для визначення різкого збільшення обєму корисного навантаження, може бути використаний запропонований метод.

Для виявлення дрейфу концепції розробленим методом необхідно обчислення для кожного значення t середніх значень $M_{W_1}(t)$ і $M_{W_2}(t)$. При великих розмірах вікон продуктивність методу може різко впасти. Для усунення даного недоліка можна розглядати значення $M_{W_1}(t)$ і $M_{W_2}(t)$ не для всіх значень t, а лише для кратних значень інтервалу S_t. Введення інтервалу впливає тільки на значення $M_{W_1}(t)$, $M_{W_2}(t)$ і $R(t)$, але не на трафік вхідного потоку елементів Y.

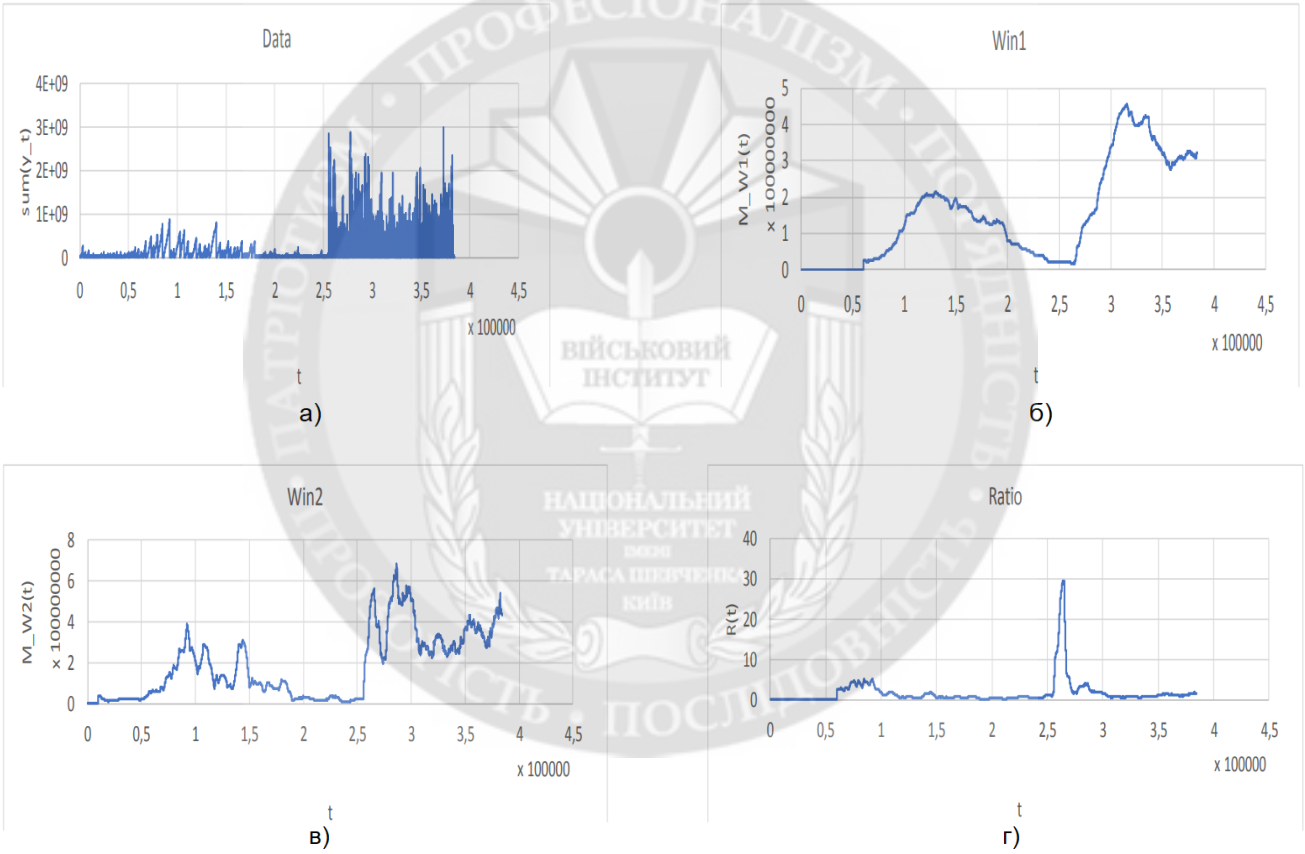

Рисунок 2 - Експериментальні дані
а) додатка «Instagram», б) залежності M_{W_1}, в) залежності, г) Залежності $R(t)$

Висновки. Наявність фонового трафіка значно погіршує оцінку якості і точності класифікації. Для інтернет додатка Skype зниження оцінки точності класифікації для алгоритму Random Forest - 13,2 %, для C4.5 до 35%. Для інтернет додатка BitTorrent зниження оцінки точності класифікації: для Random Forest -5,7%, для C4.5-37%, обумовлено помилковою класифікацією фонових інтернет додатків.

Алгоритми неконтрольованого навчання DBSCAN та *k*-Means значно поступаються алгоритмам які використовують навчання з вчителем (Random Forest). Алгоритм *k*-Means вирішує задачу кластеризацією мережного трафіка, лише за умови, що кількість кластерів

наперед відома, інакше якість класифікації погіршується і для інтернет додатків DNS, HTTP, Skype, Steam досягає 30%. Алгоритм DBSCAN видає значні помилки у змісті та кількості кластерів аналізованих інтернет додатків розкидані по багатьох кластерах.

Алгоритми оцінки якості і точності класифікації C4.5 та Random Forest показують близькі результати. Середня величина оцінки для інтернет додатків становить для алгоритму Random Forest - 0,984, для алгоритму C4.5 - 0,985. За часом тестування та навчання суттєво різняться. Час тестування Random Forest в середньому в 4 рази менше, алгоритму C4.5.

Статистичний аналіз та дослідження атрибутів інтернет додатків показав, що найважливіші атрибути, пов'язані зі зміною об'єму інтернет трафіка потоку даних, мають експоненційний вигляд. Для виявлення аномальних змін об'єму інтернет трафіка додатків для розрахунку середніх значень може бути використаний критерій Фішера.

Для класифікації інтернет додатків у потоковому режимі даних, при безперервному надходженні потоку даних запропоновано алгоритм виявлення зміщення концепту (дрейфа) трафіка потоку даних. Детектор дрейфа Фішера базується на статистичних характеристиках атрибутів інтернет додатків, аналізуються з використанням ковзаючих вікон, які контролюють зміну трафіка поточних статистичних характеристик атрибутів додатків.

ЛІТЕРАТУРА:

1. Ленков, С.В. Модель безпеки поширення забороненої інформації в інформаційно-телекомунікаційних мережах / С.В. Ленков, , В.М. Джулій, В.С. Орленко, О.В. Сєлюков, А.В. Атаманюк // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – К.: ВІКНУ, 2020. – Вип. №68. – С. 53-64.

2. Джулій, В.М. Модель нелегітимного абонента забезпечення безпеки IP-телефонії / О.С. Андрощук, В.М. Джулій, Ю.П. Кльоц, І.В. Муляр // Вимірювальна та обчислювальна техніка в технологічних процесах. – Хмельницький, 2020. – №2. – С. 38–45.

3. Джулій В.М., Кльоц Ю.П., Муляр І.В., Жилевич М.Л., Джулій А.В. Контроль додатків інтернет-трафіка комп'ютерних мереж методами машинного навчання. Вісник Хмельницького національного університету. Технічні науки. 2021. № 5. С. 22-26.

4. Шелухин О.И. Сетевые аномалии. Обнаружение, локализация, прогнозирование/ О.И. Шелухин - М.: Горячая линия -Телеком, 2019. - 448 с.

5. Шелухин О.И. Классификация IP-трафика методами машинного обучения / О.И. Шелухин, С.Д. Ерохин - М.: Горячая линия -Телеком, 2018. - 284 с.

6. Батурин, Ю.М. Компьютерная преступность и компьютерная безопасность / Ю.М. Батурин, А.М. Жодзинский. – М.: Юридическая литература, 2006. – 160 с.

7. Нестеров, С.А. Основы информационной безопасности: учебник / С. А. Нестеров. - СПб. : Лань, 2017. – 423 с.

8. Олифер, В.Г. Безопасность компьютерных сетей / В. Г. Олифер, Н. А. Олифер. - М. : Горячая линия-Телеком, 2017. - 644 с.

9. Бабаш, А.В. Криптографические методы защиты информации: учебник для студетнов вузов / А. В. Бабаш, Є. К. Баранова. - М. : КНОРУС, 2016. - 190 с.

10. Борисов, М.А. Основы для программно-аппаратной защиты информации : учеб. пособие для вузов / М. А. Борисов, И. В. Заводцев, И. В. Чижов. - 4-е изд., переработаное и доп. - М. : ЛЕНАНД, 2016. - 416 с.

11. Васильева, И. И. Криптографические методы защиты информации : практикум и учебник для академ. бакалавриата / И. И. Васильева. - Санкт-Петерб. гос. эконом. университет . - М. : Юрайт, 2017. - 349 с.

12. Нестеров, С.А. Основы информационной безопасности : учебник / С. А. Нестеров. - СПб. : Лань, 2017. – 423 с.

REFERENCES:

1. Lenkov, S.V. Model bezpeky poshyrennia zaboronenoi informatsii v informatsiino-telekomunikatsiinykh merezhakh / S.V. Lenkov, V.M. Dzhulii, V.S. Orlenko, O.V. Sieliukov, A.V. Atamaniuk // Zbirnyk naukovykh prats Viiskovoho instytutu Kyivskoho natsionalnoho universytetu imeni Tarasa Shevchenka. – K.: VIKNU, 2020. – №68. – Pp. 53-64.

2. Dzhulii, V.M. Model nelehitymnoho abonenta zabezpechennia bezpeky IP-telefonii / O.S. Androshchuk, V.M. Dzhulii, Yu.P. Klots, I.V. Muliar // Vymiriuvalna ta obchysliuvalna tekhnika v tekhnolohichnykh protsesakh. – Khmelnytskyi, 2020. – №2. – Pp. 38–45.

3. Dzhulii V.M., Klots Yu.P., Muliar I.V., Zhylevych M.L., Dzhulii A.V. Kontrol dodatkiv internet-trafika kompiuternykh merezh metodamy mashynnoho navchannia. Visnyk Khmelnytskoho natsionalnoho universytetu. Tekhnichni nauky. – Khmelnytskyi, 2021. – №5. – Pp. 22–26.

4. Shelukhyn O.Y. Setevыe anomalyy. Obnaruzhenye, lokalyzatsyia, prohnozyrovanye/ O.Y. Shelukhyn - M.: Horiachaia lynyia -Telekom, 2019. - 448 s.

5. Shelukhyn O.Y. Klassyfykatsyia IP-trafyka metodamy mashynnoho obuchenyia / O.Y. Shelukhyn, S.D. Erokhyn - M.: Horiachaia lynyia -Telekom, 2018. - 284 s.

6. Baturyn, Yu.M. Kompiuternaia prestupnost y kompiuternaia bezopasnost / Yu.M. Baturyn, A.M. Zhodzynskyi. – M.: Yurydycheskaia lyteratura, 2006. – 160 s.

7. Nesterov, S.A. Osnovы ynformatsyonnoi bezopasnosty : uchebnyk / S. A. Nesterov. - SPb. : Lan, 2017. – 423 s.

8. Olyfer, V.H. Bezopasnost kompiuternыkh setei / V. H. Olyfer, N. A. Olyfer. - M. : Horiachaia lynyia-Telekom, 2017. - 644 s.

9. Babash, A.V. and Baranova, Ye. K. (2016), “Kryptohrafycheskye metodы zashchytы ynformatsyy : uchebnyk dlia studetnov vuzov” / M. : KNORUS, 190 p.

10. Borysov, M.A., Zavodtsev, Y.V. and Chyzhov Y.V.(2016), “Osnovы dlia prohrammno-apparatnoi zashchytы ynformatsyy : ucheb. posobye dlia vuzov” / M. : LENAND, 416 p.

11. Vasyleva, Y.Y. (2017),_”Kryptohrafycheskye metodы zashchytы ynformatsyy : praktykum y uchebnyk dlia akadem. Bakalavryata” / M. : Yurait, 349 p.

12. Nesterov, S.A. (2017), “Osnovы ynformatsyonnoi bezopasnosty : uchebnyk” / SPb. : Lan, 423 p.

PhD Dzhuliy V.M., PhD Miroshnichenko O.V., Solodeeva L.V.

METHOD OF CLASSIFICATION OF APPLICATIONS TRAFFIC OF COMPUTER NETWORKS ON THE BASIS OF MACHINE LEARNING UNDER UNCERTAINTY

The paper proposes a method for classifying applications of computer network traffic based on machine learning in conditions of uncertainty. Modern methods of classification of computer network traffic applications (such as the classification of transport layer protocols by port numbers) have significant shortcomings, which leads to and is the reason for the growth of research in the direction of classification of computer network traffic applications. The rapid growth in recent years of the types and number of transport layer network protocols increases the relevance of research in this area, the development of appropriate algorithms and methods for classifying applications of computer network traffic, which reduce computational complexity. At the present stage, the problem that needs to be urgently addressed is the classification of computer network traffic applications using appropriate protocols and encryption algorithms.

A promising area of classification of computer network traffic applications is statistical methods, which are based on the analysis and identification of statistical characteristics of IP traffic. The most promising are the intellectual analysis of data flow, as well as machine learning technologies, which are currently widely used in related fields of science. The problem of research and training according to precedents is solved - classification of computer network traffic applications on the basis of pre-known set of attributes of their features, in order to improve the technical base of computer networks and theoretical base, while ensuring high performance and quality networks. example of using transport layer protocols (TCP / IP stack). The result of solving this problem is to assign the application, in accordance with the rules of the educational sample, to one of the outstanding classes, which are predetermined, which contains the relevant, but already classified applications. Statistical analysis and research of the attributes of Internet applications showed that the most important attributes associated with changes in the volume of Internet traffic flow are exponential. Fisher's criterion can be used to calculate anomalous changes in the amount of Internet traffic of applications to calculate averages.

To classify Internet applications in data streaming mode, an algorithm for detecting the offset of the concept (drift) of data flow traffic is proposed for continuous data flow. Fisher's drift detector is based on the statistical characteristics of the attributes of Internet applications, analyzed using sliding windows that monitor changes in traffic current statistical characteristics of the attributes of applications.

Key words: models, application classification, computer networks, drift, traffic, sliding window, machine learning.