УДК 681.518:004.93.1'

В. В. МОСКАЛЕНКО, А. С. РИЖОВА

Сумський державний університет, Україна

ІНФОРМАЦІЙНО-ЕКСТРЕМАЛЬНИЙ МЕТОД ІДЕНТИФІКАЦІЇ ТИПУ МЕРЕЖЕВОГО ТРАФІКА

Розглядається інформаційний синтез здатного навчатися класифікатора трафіка в рамках інформаційно-екстремальної інтелектуальної технології аналізу даних, що грунтується на максимізації інформаційної спроможності системи в процесі навчання. У статті досліджено алгоритм навчання класифікатора за незбалансованими неоднорідними навчальними вибірками, що має місце в практичних задачах моніторігу мережевого трафіка. Із врахуванням багатовимірності словника ознак і алфавіту класів розпізнавання, які характеризують типи мережевого трафіка, розроблено інформаційно-екстремальний алгоритм навчання класифікатора з ієрархічною структурою вирішальних правил. Запропонований алгоритм реалізовано при розв'язанні задачі ідентифікації зашифрованого мережевого трафіка на прикладі додатків VoIP, SSH та BitTorrent.

Ключові слова: *класифікація мережевого трафіка, потік пакетів даних, машинне навчання, розпізнавання образів, незбалансована навчальна вибірка, оптимізація, інформаційний критерій.*

Вступ

Сучасні телекомунікаційні мережі характеризуються високими і надвисокими показниками пакетообігу, що обумовлює необхідність пріоритезації трафіка відповідно до вимог рівня обслуговування користувачів – Service Level Agreement (SLA) та якості мережевих сервісів – Quality of Service (QoS). Для визначення пріоритету при формуванні смуги пропускання окремого трафіка необхідно мати інструмент точної ідентифікації та класифікації трафіка. При цьому поширення мережевих додатків, які динамічно змінюють порти транспортних протоколів, використовують шифрування та інкапсуляцію трафіка в тунельний протокол, призводить до низької ефективності класифікації трафіка на основі портів чи корисного навантаження (Deep Packet Inspection) [1, 2]. Вирішення цих проблем пов'язується з використанням методів машинного навчання та розпізнавання образів, де, як правило, ознаками розпізнавання виступають статистичні характеристики потоку пакетів (Flow-based Classification) [1-3]. Проте різноманітність та параметрична невизначеність інформаційних процесів, які генерують мережевий трафік, використання методів маскування трафіка призводять до значного перетину класів розпізнавання в просторі ознак, що зменшує достовірність класифікації [3,4]. Особливо критично це проявляється при збільшенні потужності алфавіту класів та незбалансованості навчальних наборів даних [3, 5].

Одним із перспективних шляхів підвищення точності класифікації мережевого трафіка є інформаційний синтез класифікатора в рамках інформаційно-екстремальної інтелектуальної технології (IEI-технологія), яка дозволяє трансформувати апріорно нечітке розбиття простору ознак в чітку еквівалентність класів розпізнавання [6-8]. Крім цього, важливими перевагами IEI-технології, є невисока обчислювальна складність вирішальних правил, що сприяє економії ресурсів мережевого пристрою в умовах високого пакетообігу.

У статті пропонується в рамках IEI-технології класифікатор з ієрархічною структурою вирішальних правил, який навчається за незбалансованими наборами даних, для ідентифікації зашифрованого мережевого трафіка на прикладі додатків VoIP, SSH та BitTorrent.

1. Формалізована постановка задачі

Нехай дано алфавіт $\{ X^{o}_{m} \mid m = \overline{1,M} \}$ класів розпізнавання, які характеризують трафік різних типів мережевого сервісу. Як реалізації кожного класу розпізнавання розглядаються впорядковані вектори статистичних характеристик потоку пакетів. Відома навчальна матриця

$$\{ y^{(j)}_{m,i} \mid m = \overline{1, M}; j = \overline{1, n_m}; \quad i = \overline{1, N} \},$$

де M – потужність алфавіту класів розпізнавання; n_m – кількість реалізацій класу X^{o}_{m}; N – кількість ознак розпізнавання.

Відомі s -ярусна деревоподібна бінарна ієрархічна структура класів розпізнавання і структурований вектор параметрів:

$$g=<\delta_{s_i}, d_s>, s=\overline{1,M\text{-}1}, \quad (1)$$

де $\delta_{i,s}$ – параметр, який визначає півширину симетричного рецептивного поля для i -ї ознаки розпізнавання відносно усередненого значенню i -ї ознаки класу $X_s^o \in \{X_m^o\}$, який відокремлюється у листок дерева на s -му ярусі ;

d_s – кодовий радіус гіперсферичного (вписаного в одиничний гіперпаралелепіпед) контейнера класу X_s^o, який відновлюється в радіальному базисі бінарного простору Хеммінга $\Omega_Б$ з центром, що визначається вершиною одиничного еталонного вектора $x_{s,i}=1$, $i=\overline{1,N}$.

При цьому відомі обмеження на параметри функціонування

$$g: \delta_{s_i} \in [0; 0{,}5\cdot\delta_{H,i}],$$

де $\delta_{H,i}$ – максимальна ширина рецептивного поля для i -ї ознаки розпізнавання; $0<d_s<N$.

Необхідно в процесі навчання класифікатора трафіка визначити оптимальні значення координат вектора параметрів функціонування (1), що забезпечують максимальне значення усередненого за алфавітом класів розпізнавання критерію функціональної ефективності (КФЕ)

$$\overline{E}=\frac{1}{M-1}\sum_{s=1}^{M-1} E_s, \quad (2)$$

де E_s – інформаційний критерій функціональної ефективності навчання класифікатора розпізнавати реалізації класу X_s^o ;

$\{k\}$ – упорядкована множина кроків навчання (відновлення контейнерів класів розпізнавання).

При функціонуванні класифікатора трафіка в режимі розпізнавання необхідно прийняти рішення про належність реалізації мережевого трафіка до одного з класів мережевих сервісів $\left\{X_m^o \mid m=\overline{1,M}\right\}$ і призначити трафіку клас обслуговування, що відповідає вимогам SLA.

2. Алгоритм навчання класифікатора трафіку

Для формування деревоподібної бінарної ієрархічної структури вирішальних правил послідовно на кожному ярусі здійснюється розбиття більших груп класів на дві менші. У найпростішому випадку достатньо здійснювати на кожному ярусі побудову контейнера для одного класу розпізнавання для його відокремлення в листок дерева. При цьому рецептивні поля для ознак розпізнавання визначаються на кожному ярусі окремо [6].

Квазіоптимізація параметра $\delta_s=\delta_{s_i}, i=\overline{1,N}$ рецептивного поля призначена для визначення стартових значень, які відповідають робочій області визначення функції інформаційного КФЕ, і здійснюється за ітераційною процедурою

$$\delta_s^*=\arg\max_{G_\delta}\left\{\frac{1}{M-1}\sum_{s=1}^{M-1}\left[\max_{G_E\cap G_d} E_s\right]\right\}, \quad (3)$$

де G_δ – область допустимих значень параметра рецептивного поля;

G_d – область допустимих значень радіусу гіперсферичного контейнера;

G_E – робоча область визначення функції КФЕ;

E_s – КФЕ навчання класифікатора на s -ярусі ієрархічної структури.

Послідовна оптимізація параметра δ_{s_i} рецептивного поля для i -ї ознаки здійснюється за ітераційною процедурою

$$\delta_{s_i}^*=\arg\left\{\bigotimes_{l=1}^{L}\max_{G_\delta}\left\{\frac{1}{M-1}\sum_{m=1}^{M-1}\left[\max_{G_E\cap G_d} E_s^{(l)}\right]\right\}\right\}, \quad (4)$$

де $E_s^{(l)}$ – КФЕ навчання класифікатора на s -ярусі при l -му прогоні послідовної процедури оптимізації;

$\otimes$ – символ операції повторення;

L – кількість прогонів ітераційної процедури послідовної оптимізації рецептивних полів.

Базовий алгоритм [6,8] є вкладеним в процедури (3) та (4) і здійснює побудову контейнера тільки для базового класу $X_s^o \in \{X_m^o\}$, що відокремлюється у листок деревоподібної структури

$$d_s^*=\arg\max_{\{d_s\}\in G_d} E_s, \quad (5)$$

Вибір базового класу $X_s^o \in \{X_m^o\}$ для s -го ярусу ієрархічної структури базується на ідеї найкращого відокремлення від реалізацій чужих класів і здійснюється за таким алгоритмом (алгоритм LEARNING-1):

Крок 1. Ініціалізація лічильника ярусів дерева рішень: $s := 0$.

Крок 2. Ініціалізація лічильника класів: $k := 1$.

Крок 3. Оптимізація радіусу гіперсферичного контейнера класу X_k^o за ітеративною процедурою (5), прийнявши за сусідній клас X_c^o сукупність найближчих до ядра класу X_k^o реалізацій $\{x_c^{(j)} \mid j = \overline{1, n_k}\} \in \left[\bigcup_{c=1}^{M-h} X_c^o\right] \setminus X_k^o$.

Крок 4. $k := k + 1$.

Крок 5. Порівняння: якщо $k < M - s$, то виконується крок 3, інакше – крок 6.

Крок 6. Прийняти за листок дерева клас, що забезпечує максимальне значення інформаційного КФЕ $X_s^o = \arg\{\max_{\{X_m\}}\{\max_{G_\delta}\{\max_{G_E}\{E_k\}\}\}$ і вилучити його з подальшого розгляду.

Крок 7. $s := s + 1$.

Крок 8. Якщо $s < M - 1$, то перехід на крок 2, інакше – «ЗУПИН».

Як КФЕ навчання класифікатора на s -му ярусі розглянемо модифіковану інформаційну міру Кульбака [7,8], в якій відношення правдоподібності представлено у вигляді відношення повної ймовірності правильного прийняття рішень P_{true} до повної ймовірності помилкового прийняття рішень P_{false}. В цьому випадку для двохальтернативних гіпотез міра Кульбака має вигляд

$$E_s^{(k)} = \left[P_{true,s}^{(k)} - P_{false,s}^{(k)}\right] \log_2 \frac{P_{true,s}^{(k)}}{P_{false,s}^{(k)}} =$$

$$= \left| \begin{array}{l} P_{true,s}^{(k)} = p_1 D_{1,s} + p_2 D_{2,s} \\ P_{false,s}^{(k)} = p_1 \alpha_s + p_2 \beta_s \\ p_1 = \dfrac{n_s}{n_s + n_c}; p_2 = \dfrac{n_c}{n_s + n_c} \\ \alpha_s = 1 - D_{1,s}; D_{2,s} = 1 - \beta_s \end{array} \right| =$$

$$= \frac{\left[n_c - n_s + 2 \cdot (n_s D_{1,s}^{(k)} - n_c \beta_s^{(k)})\right]}{n_s + n_c} *$$

$$* \log_2 \left(\frac{n_c + (n_s D_{1,s}^{(k)} - n_c \beta_s^{(k)})}{n_s - (n_s D_{1,s}^{(k)} - n_c \beta_s^{(k)})} \right), \qquad (6)$$

де $D_{1,s}^{(k)}$ – перша достовірність, обчислена на k -му кроці навчання для s -го ярусу;

$D_{2,s}^{(k)}$ – друга достовірність;

$\alpha_s^{(k)}$ – помилка першого роду;

$\beta_s^{(k)}$ – помилка другого роду;

n_s – кількість реалізацій у навчальній вибірці базового класу X_s^o;

n_c – кількість сусідніх реалізацій, що належать до інших класів s -го ярусу.

Нормовану модифікацію критерію (6) представимо у вигляді

$$\hat{E}_s^{(k)} = \frac{E_s^{(k)}}{E_{max}}, \qquad (7)$$

де E_{max} – значення критерію при $D_{1,h}^{(k)} = 1$ і $\beta_h^{(k)} = 0$.

При цьому робоча (допустима) область визначення функції інформаційного КФЕ обмежена нерівностями $D_1 \geq 0,5$ та $D_2 \geq 0,5$.

Визначення належності тестової реалізації $x^{(j)}$ до контейнера класу X_s^o здійснюється за правилом

$$\text{if } d[x_s \oplus x^{(j)}] \leq d_s \text{ then } x^{(j)} \in X_s^o \text{ else } x^{(j)} \notin X_s^o, \qquad (8)$$

де $d[x_s \oplus x^{(j)}]$ – кодова відстань від вектора $x^{(j)}$ до x_s;

d_s – радіус контейнера класу X_s^o, що відновлюється в бінарному просторі ознак на s -му ярусі.

Як правило дані архівів моніторингу мережевого трафіку мають дуже великий обсяг та характеризуються різноманітністю і незбалансованістю, різним розподілом різних типів трафіку. Навчання класифікатора з використанням повного обсягу навчальних даних займе досить тривалий час, а формування вибірок меншого обсягу не гарантуватиме їх репрезентативності і призведе до втрат інформації. Для підвищення оперативності навчання інформаційно-ектремального класифікатора трафіка запропоновано наступну модифікацію алгоритму навчання (алгоритм LEARNING-2):

Крок 1. Ініціалізація масивів оптимального $\delta_{s_i}^*$ та стартового $\delta_{s_i}^{start}$ параметрів рецептивних полів для ознак розпізнавання: $\delta_{s_i}^* := 0$; $\delta_{s_i}^{start} := 0$ при

$i = \overline{1, N}$.

Крок 2. Поділ великого масиву апріорно-класифікованих векторів-реалізацій на навчальну $Y = \{y^{(j)}_{m,i} \mid m = \overline{1, M}; j = \overline{1, n_{min}}; i = \overline{1, N}\}$, де n_{min} – мінімальний за замовчуванням обсяг вибірки, та тестову $Y_{test} = \{y^{(t)}_i \mid t = \overline{1, T}; \quad i = \overline{1, N}\}$ матриці.

Крок 3. Запуск інформаційно-ектремального навчання за навчальною матрицею Y з квазіоптимізацією параметра рецептивних полів $\delta_s = \delta_{si}$ за процедурою (3) при стартових параметрах $\{\delta^{start}_{s_i}\}$.

Крок 4. $\delta^{start}_{s_i} := \delta^*_s$.

Крок 5. Ініціалізація лічильника векторів-реалізацій тестової матриці Y_{test}: $t := 0$.

Крок 6. Запуск інформаційно-ектремального навчання з послідовною оптимізацією параметра рецептивних полів $\{\delta_{si}\}$ за процедурою (4) при стартових параметрах $\{\delta^{start}_{s_i}\}$.

Крок 7. $\delta^{start}_{s_i} := \delta^*_{s_i}$.

Крок 8. $t := t + 1$.

Крок 9. Якщо $t \le T$, то визначити належність $y^{(t)}$ до одного з класів алфавіту $\{X^o_m\}$ за правилом (7), інакше перехід до кроку 12.

Крок 10. Якщо належність вектора-реалізації $y^{(t)}$ не співпадає з апріорною класифікацією, то додати $y^{(t)}$ до навчальної матриці Y та перейти до кроку 6, інакше до кроку 8.

Крок 11. ЗУПИН.

Таким чином, алгоритм навчання класифікатора трафіку в рамках IEI-технології полягає в ітераційній процедурі наближення глобального максимуму інформаційного КФЕ (2) до його граничного значення шляхом оптимізації параметрів рецептивних полів та геометричних параметрів контейнерів на кожному ярусі ієрархічної структури вирішальних правил. При цьому отримані вирішальні правила за навчальною вибіркою малого обсягу використовуються для сканування тестової вибірки з метою донавчання при невірних класифікаціях.

3. Реалізація системи ідентифікації мережевого трафіку

Розглянемо результати реалізації запропонованого алгоритму навчання інформаційно-екстремального класифікатора для ідентифікації зашифрованого трафіку таких застосувань як SSH (трафік дистанційного керування операційними системами та тунелювання TCP-з'єднань), VoIP (представлений в основному голосовим трафіком Skype) та BitTorrent (зашифрований трафік пірінгової мережі обміну файлами), оскільки вони останнім часом займають значну частину від загального мережевого трафіку і враховуються адміністратором при настройці QoS-механізму.

Навчальні набори даних були сформовані в процесі трасування трафіку утилітою TcpDump [4] з наступним формуванням потоків і обчисленням ознак розпізнавання за допомогою утиліти NetMate [4, 5]. Апріорна класифікація реалізацій навчального трафіку основана на результатах моніторингу сокетів утилітою CurrPorts (для Windows) [5] та Net Activity Viewer (для Linux) [9].

Ознаками розпізнавання є статистичні характеристики двонаправленого потоку пакетів, де як потік розглядається ряд пакетів, що поділяють однаковий кортеж з п'яти елементів: IP-адреса джерела та отримувача, номер портів джерела і отримувача, номер протоколу. При цьому TCP-потоки обмежені тривалістю до 600 с, а UDP-потоки обмежені максимальною тривалістю між прибуттям пакетів, що становить 64 с. Загальна кількість ознак розпізнавання становить $N = 37$, а саме: кількість пакетів та байтів в прямому/зворотному напрямках потоку; відношення кількості пакетів до кількості байтів корисного навантаження в прямому/зворотному напрямках; середнє значення, мінімальне значення, перша та третя квартілі, медіана та дисперсія розміру корисного навантаження (в байтах) для вхідних/вихідних пакетів двонаправленого потоку; відношення кількості пакетів малого розміру (до 50 байтів корисного навантаження) до загальної кількості пакетів в прямому/зворотному та в обох напрямах; відношення кількості пакетів великого розміру (більше 1300 байтів корисного навантаження) до загальної кількості пакетів в прямому/зворотному та в обох напрямах; мінімальне, максимальне та середнє значення тривалості часового інтервалу між прибуттям пакетів в прямому/зворотному напрямах; відношення кількості пакетів без корисного навантаження до загальної кількості пакетів в прямому/зворотному та в обох напрямах; кількість прапорців ACK / PSH в потоці прямого/зворотного напряму.

Збір даних моніторингу мережевого трафіку здійснювався на 10-ти комп'ютерах локальної мережі кафедри комп'ютерних наук Сумського державного університету протягом 5 годин робочого дня. Загальний обсяг накопичених наборів даних становить 5,31 Гбайт (22 271 754 реалізацій). Набір даних незбалансований, проте була здійснена процедура видалення однакових реалізацій, що зменшило обсяг

даних до 4,11 Гбайт (17 238 589 реалізацій). Для побудови в процесі виконання алгоритму LEARNING-1 ієрархічної структури вирішальних правил (рис.1) для 4-х класів розпізнавання використано випадкові вибірки з наявного набору даних по 100 векторів-реалізацій для кожного класу.

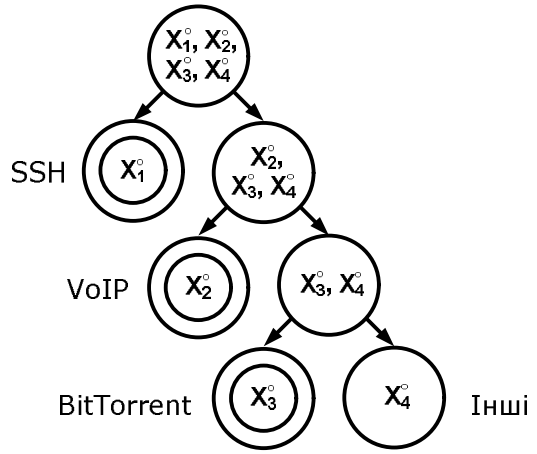


Рис. 1. Ієрархічна структура вирішальних правил для чотирьох класів розпізнавання

Аналіз рис. 1 показує, що на першому ярусі побудовано контейнер для класу X_1^o (трафік SSH), на другому ярусі побудовано контейнер для класу X_2^o (трафік VoIP), а на третьому ярусі побудовано контейнер для класу X_3^o (трафік BitTorrent). Клас X_4^o є сусіднім до класу X_3^o і представлений реалізаціями трафіка, які характеризують веб-серфінг, незашифрований обмін файлами та деякі фонові процеси оновлення системного та користувацького програмного забезпечення, DNS-запити та інше.

З метою підвищення точності класифікатора реалізується його донавчання за алгоритмом LEARNING-2. Графіки зміни максимумів нормованого КФЕ в процесі послідовної оптимізації параметра рецептивних полів на кожному ярусі ієрархічної структури в режимах навчання (до першого максимуму $E_s^o = 1,0$) за алгоритмом LEARNING-1 та донавчання (після першого максимуму $E_s^o = 1,0$) за алгоритмом LEARNING-2 показано на рис. 2. При цьому кожен крок оптимізації збільшує лічильник кроків k і відповідає одній зміні параметра рецептивного поля для будь-якої ознаки розпізнавання.

Аналіз рис. 2. показує, що в процесі навчання на кожному ярусі деревоподібної структури було отримано безпомилкове за навчальною матрицею вирішальне правило. Однак в процесі донавчання за алгоритмом LEARNING-2 на першому ярусі було виявлено дві помилкові класифікації тестових реалізацій, на другому ярусі − три помилкові класифікації, а на третьому ярусі − 15 помилкових класифікацій. Після додавання до навчальної матриці помилково класифікованих реалізацій вдалося побудувати безпомилковий за навчальною та тестовою матрицями класифікатор.

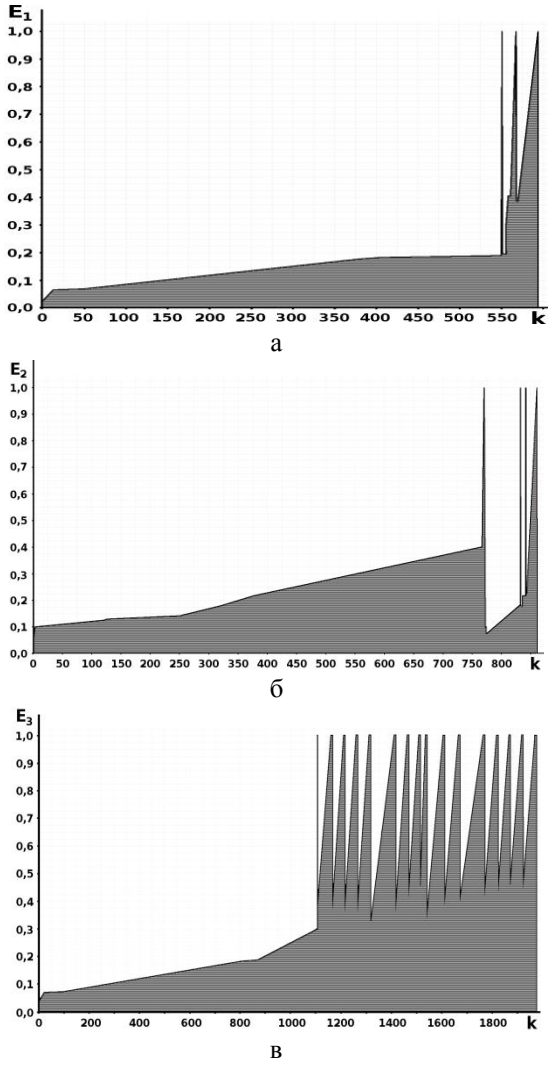
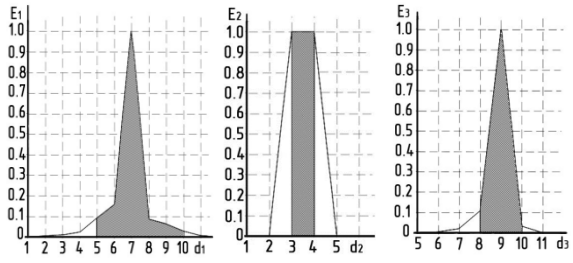


Рис. 2. Графіки зміни максимумів КФЕ в процесі послідовної оптимізації параметра рецептивних полів в режимах навчання та донавчання: а − перший ярус; б − другий; в − третій

На рис. 3 показано графіки залежності радіусу гіперсферичного контейнера для кожного класу, що розпізнається на s-му ярусі при оптимальних рецептивних полях для ознак розпізнавання.

Аналіз рис. 3 показує, що оптимальні значення радіусів контейнерів для класів розпізнавання дорівнюють: для класу X_1^o $-d_1^* = 7$ (тут і далі в кодових одиницях), для класу X_2^o $-d_2^* = 3$, для класу X_3^o $-d_3^* = 9$.



а б в

Рис. 3. Графіки залежності нормованого КФЕ при оптимальних рецептивних полях для ознак розпізнавання від радіусу контейнера класу:

$$а - X_1^o ; б - X_2^o ; б - X_3^o .$$

Таким чином, запропонований алгоритм інформаційно-ектремального навчання дозволяє отримати безпомилкові вирішальні правила за незбалансованими навчальними наборами даних для розпізнавання зашифрованного трафіку додатків SSH, VoIP та BitTorrent.

Висновки

В рамках інформаційно-ектремальної інтелектуальної технології синтезовано класифікатор мережевих трафіків з підвищеною достовірністю та оперативністю. Оперативність навчання класифікатора підвищена за рахунок використання отриманих вирішальних правил в результаті навчання за вибірками малого розміру для сканування апріорно-класифікованої тестової матриці великого розміру для виявлення неправильно класифікованих реалізацій трафіку з метою донавчання. При цьому побудова безпомилкових за навчальною матрицею вирішальних правил у рамках геометричного підходу забезпечує прийняття рішень в робочому режимі в реальному темпі часу.

За результатами фізичного моделювання за даними моніторингу трафіків SSH, VoIP та BitTorrent було доведено високу ефективність розроблених алгоритмів і отримано безпомилкові за навчальними матрицями вирішальні правила.

Література

1. *Zander, S. Automated Traffic Classification and Application Identification using Machine Learning [Text] / S. Zander, T. Nguyen, G. Armitage // Proceedings of the IEEE Conference on Local Computer Networks 30th Anniversary. − Sydney : IEEE Press, 2005. − P. 250-257.*

2. Pedro, M. Santiago del Rio. Internet Traffic Classification for High-Performance and Off-The-Shelf Systems [Text] / Pedro Maria Santiago del Rio // Ph.D. Thesis. − Madrid, Spain : Technical University of Madrid, 2013. − 217 p.

3. Raman, Singh. Issue related to sampling techniques for network traffic dataset [Text] / Raman Sigh, Harish Kumar, R. K. Singla //International Journal of Mobile Network Communications & Telematics. − Sydney, Australla : WSP, 2013. − Vol. 3., No. 4. − P. 75-85.

4. Iacovazzi, A. Network Communication Privacy: Traffic Masking against Traffic Analysis [Text] / Alfonso Iacovazzi // Ph.D. Thesis. − Rome, Italy : Sapienza University of Rome, 2013. − 119 p.

5. Volunteer-Based System for classification of traffic in computer networks [Text] / T. Bujlow, K. Balachandran, M. T. Riaz, J. M. Pedersen // In Proceedings of 19th Telecommunications Forum TELFOR 2011'. − Sydney : IEEE Press, 2011. − P. 210-213.

6. Москаленко, В. В. Ієрархічний інформаційно-екстремальний класифікатор [Текст] / В. В. Москаленко, С. А. С. М. Джулгам // Радіоелектронні і комп'ютерні системи. – 2012. − № 3 (55). – С. 86-93.

7. Moskalenko, V. V. Information-Extreme Algorithm for Optimizing Parameters of Hyperellipsoidal Containers of Recognition Classes [Text] / A. S. Dovbysh, N. N. Budnyk, V. V. Moskalenko // Journal of automation and information sciences. − New York : Begell House Inc., 2012. – Vol. 44, Iss. 10. − P. 35-44.

8. Довбиш, А. С. Основи проектування інтелектуальних систем [Текст] / А. С. Довбиш. – Суми : СумДУ. – 2009. – 171 с.

9. Markowsky, G. Who's Knocking at Your Cybercastle's Gate? [Text] / G. Markowsky, L. Markowsky // In Proceedings of International Conference on Security and Management. – Las Vegas Nevada, USA : DCSREA Press, 2012. – P. 206-212.

Поступила в редакцию 4.11.2014, рассмотрена на редколлегии 18.12.2014

Рецензент: д-р техн. наук, проф., зав. каф. авіаційних приладів та вимірювань М. Д. Кошовий, Національний аерокосмічний університет ім. М. Є. Жуковського «ХАІ», Харків, Україна.

ИНФОРМАЦИОННО-ЭКСТРЕМАЛЬНЫЙ МЕТОД ИДЕНТИФИКАЦИИ ТИПА СЕТЕВОГО ТРАФИКА

В. В. Москаленко, А. С. Рыжова

Рассматривается информационный синтез обучающегося классификатора трафика в рамках информационно-экстремальной интеллектуальной технологии анализа данных, основанной на максимизации информационной способности системы в процессе обучения. В статье исследован алгоритм обучения классификатора по несбалансированным неоднородным обучающим выборкам, что имеет место в практических задачах мониторинга сетевого трафика. С учетом многомерности словаря признаков и алфавита классов распознавания, которые характеризуют типы сетевого трафика, разработан информационно-экстремальный алгоритм обучения классификатора с иерархической структурой решающих правил. Предложенный алгоритм реализован при решении задачи идентификации зашифрованного сетевого трафика на примере приложений VoIP, SSH и BitTorrent.

Ключевые слова: классификация сетевого трафика, поток пакетов данных, машинное обучение, распознавание образов, несбалансированная обучающая выборка, оптимизация, информационный критерий.

INFORMATION-EXTREME METHOD OF IDENTIFICATION OF NETWORK TRAFFIC TYPE

V. V. Moskalenko, A. S. Rizhova

In this paper informational synthesis of machine learning classifier of network traffic within the information-extreme intellectual technology of data analysis which based on maximization of informational capabilities during machine learning is considered. This article explores the classifier learning algorithm for unbalanced mixed training set, which occurs in practical problems of network traffic monitoring. Taking into account multidimensionality of feature set and set of classes that characterize types of network traffic, information-extreme classifier learning algorithm with a hierarchical structure of decision rules is developed. The proposed algorithm is implemented for solving the problem of identification of encrypted network traffic on an example application VoIP, SSH and BitTorrent.

Key words: traffic classification, packet flow, machine learning, pattern recognition, unbalanced training set, optimization, information criterion.

Москаленко В'ячеслав Васильович – канд. техн. наук, асистент каф. комп'ютерних наук, Сумський державний університет, Суми, Україна, e-mail: ai.sys.dev@gmail.com.

Рижова Алена Сергіївна – аспірант каф. комп'ютерних наук, Сумський державний університет, Суми, Україна.