ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

УДК 004.85:519.172:519.816 **doi: 10.32620/reks.2020.2.08**

**В. В. МОСКАЛЕНКО, М. О. ЗАРЕЦЬКИЙ, А. С. МОСКАЛЕНКО, А. М. КУДРЯВЦЕВ, В. А. СЕМАШКО**

***Сумський державний університет, Україна***

# БАГАТОШАРОВА МОДЕЛЬ ТА МЕТОД НАВЧАННЯ ДЛЯ ДЕТЕКТУВАННЯ ШКІДЛИВОГО ТРАФІКУ НА ОСНОВІ АНСАМБЛЮ ДЕРЕВ РІШЕНЬ

*Запропоновано модель і метод навчання багатошарового екстрактора ознак та вирішувальних правил для детектора шкідливого трафіку. Модель екстрактора ознак основана на згортковій розріджено кодуючій мережі, розріджений кодер якої апроксимується моделлю регресійного випадкового лісу згідно принципів дистиляції знань. При цьому розроблено алгоритм зростаючого розріджено кодуючого нейронного газу для навчання екстрактора ознак без вчителя з автоматичним визначення необхідної кількості ознак на кожному шарі. На етапі навчання екстрактора ознак для реалізації розрідженого кодування використано жадібний L0-регуляризований метод ортогонального узгодженого переслідування (Orthogonal Matching Pursuit), а під час дистиляції знань додатково використовувався L1-регуляризований метод найменших кутів (Least angle regression algorithm). Завдяки ефекту редукції причини отримані ознаки є некорельованими, а сформований ознаковий опис є стійким до завад та змагальних (Adveserial) атак. Запропонований екстрактор ознак навчається без вчителя для розділення пояснюючих факторів і дозволяє з максимальною ефективністю використати нерозмічені навчальні дані, обсяг яких, як правило, досить великий. Як модель вирішувальних правил запропоновано використовувати двійковий кодер спостережень на основі ансамблю дерев рішень та інформаційно-екстремальні роздільні замкнені гіперповерхні (контейнери) класів, що відновлюються в радіальному базисі двійкового простору Хеммінга. Додавання кодуючих дерев відбувається за принципом бустінгу, а радіус контейнерів класів оптимізується шляхом прямого перебору. Інформаційно-екстремальний класифікатор характеризується низькою обчислювальною складністю та високою узагальнюючою здатністю для малих наборів розмічених навчальних даних. Результати верифікації навченої моделі на відкритих тестових наборах даних CTU підтверджують придатність запропонованих алгоритмів для практичного застосування, оскільки точність розпізнавання шкідливого трафіку становить 96,1 %.*

***Ключові слова:*** *система детектування загроз; згорткова розріджено кодуюча модель; зростаючий нейронний газ; ансамбль дерев рішень; регресійний випадковий ліс; інформаційний критерій; дистиляція знань; інформаційно-екстремальна машинне навчання.*

## Вступ та пов'язані роботи

Існуючі системи детектування шкідливого трафіку все ще не забезпечують високої достовірності рішень, оскільки кількість та різноманітність нових джерел шкідливого трафіку постійно зростають, а кількість актуальних розмічених даних є досить обмеженою [1, 2]. Таким чином, вручну спроектовані ознаки для опису спостережень є недостатньо інформативними, а побудовані на їх основі вирішальні правила є недостатньо ефективними для виявлення шкідливого трафіку [2, 3]. Найбільш перспективним підходом до синтезу екстрактора ознак вважається використання ідей та методів машинного навчання для ієрархічного ознакового подання спостережень за нерозміченими даними [4, 5].

Згорткові багатошарові нейронні мережі дозволяють формувати інформативний ієрархічний ознаковий опис вхідних спостережень [6]. Крім того, вони вже показали високу ефективність у вирішенні завдань машинного зору та аналізу часових рядів [6, 7]. При цьому навчання з учителем вимагає великої кількості розмічених даних, розмітка яких може бути досить затратною або недоступною в бажаний період часу. Навчання згорткових мереж без учителя спрямоване на ефективне використання нерозмічених зразків, яких зазвичай доступно досить багато. В цьому випадку навчання виконується на основі автоенкодера або обмеженої машини Больцмана, що потребує великої кількості даних та тривалого часу навчання для отримання прийнятного результату [8]. У роботі [9] пропонується використовувати альтернативний підхід, оснований на алгоритмі кластерного аналізу k-середніх, з метою прискорення навчання словника ознак. Однак алгоритм k-середніх характеризується повільною збіжністю та субоптимальністю результатів через жорстку конкурентну схему навчання та чутливість до початкової ініціалізації кластерів.

© В. В. Москаленко, М. О. Зарецький, А. С. Москаленко, А. М. Кудрявцев, В. А. Семашко, 2020

У роботі [10] запропоновано поєднання принципів нейронного газу та розрідженого кодування для навчання екстрактора ознак на нерозмічених даних. Даний підхід характеризується м'якою конкурентною схемою навчання, що забезпечує збіжність до розподілу ознак на навчальній вибірці, що наближений до оптимального. При цьому впровадження методів розрідженого кодування може підвищити завадостійкість та узагальнюючу здатність ознакового подання вхідних даних. Крім того, добре відомий факт, що розріджене подання вхідних даних є потужним інструментом у боротьбі зі змагальними (adversarial) атаками і для отримання некорельованих ознак за рахунок ефекту редукції причини (explaining-away effect). Однак оптимальна потужність словника ознак заздалегідь невідома, і тому обирається розробником, що призводить до збільшення часу оптимізації.

Необхідну потужність словника ознак на кожному шарі ієрархічного подання важко передбачити заздалегідь, тому перспективним підходом навчання екстрактора ознак є використання принципів зростаючого нейронного газу, що автоматично визначає необхідну кількість нейронів (ознак) [11]. Наявність механізму додавання нових нейронів, а також видалення зайвих робить алгоритм більш гнучким порівняно з класичним нейронним газом, але він також має серйозні недоліки. Невеликі значення періоду між ітераціями додавання нових нейронів призводять до нестабільності процесу навчання та викривлення сформованих структур, оскільки спостерігається надмірно часте додавання нових нейронів. Завелике значення періоду забезпечує очікуваний ефект, але в той же час призводить до значного уповільнення алгоритму. Проте у роботах [11, 12] було показано, що досягти стабільності навчання можна за рахунок встановлення «радіусу досяжності» нейронів, що передбачає заміну параметра на поріг максимального віддалення нейрону від кожної з віднесених до нього точок навчальної множини. Однак, досі не були переглянуті механізми оновлення нейронів та оцінки відстані між вузлами мережі та вхідними зразками з метою адаптації процесу навчання до процедури розрідженого кодування спостережень.

Основним недоліком розрідженого кодування в репрезентативному навчанні є використання ітеративної процедури під час екзамену, що уповільнює процес розпізнавання. Один із популярних способів прискорити моделювання - це використовувати принципи дистиляції знань, де надлишкова модель, яка виступає вчителем, може бути замінена легкою моделлю, що виступає студентом [13]. Ансамбль дерев рішень - це гнучка та обчислювально ефективна модель, яка потенційно може бути використана як модель студента для спрощення розрідженого кодера [14]. Однак таких досліджень не проводилося, і ефективність такого підходу невідома, що підкреслює актуальність цього питання.

Крім того, вирішувальні правила є важливими компонентами в системах детектування загроз. Як правило, їх будують у вигляді навченого класифікатора. У той же час ефективність навчання класифікатора часто розглядається як міра ефективності екстрактора ознак [5]. Найпопулярнішим алгоритмом класифікаційного аналізу є метод опорних векторів, де навчання вирішувальних правил відбувається в рамках геометричного підходу шляхом побудови лінійної роздільної гіперповерхні в просторі вторинних ознак [15]. Однак цей алгоритм вимагає багато коригувань гіперпараметрів і його продуктивність залежить від складності функції ядра трансформації простору ознак. У праці [16] було запропоновано будувати вирішувальні правила шляхом адаптивного бінарного кодування вхідних ознак та оптимізації в інформаційному розумінні роздільної гіперповерхні, що відновлюється у радіальному базисі бінарного простору Хеммінга. Такий класифікатор характеризується високою оперативністю, оскільки використовує операції з низькою обчислювальною складністю, такі як порівняння та логічне «виключаюче АБО» XOR.

Проблема оптимального вибору розміру моделі та її параметрів в умовах обмежених обчислювальних ресурсів та малого обсягу розмічених даних остаточно не вирішена. *Метою статті* є розробка нової багатошарової моделі детектора шкідливого трафіку та спосіб її навчання для отримання оптимальних результатів в інформаційному та обчислювальному сенсах.

## Постановка задачі

Задано набори даних CTU-Mixed та CTU-13 із реального мережевого середовища які отримані дослідниками CTU з 2011 по 2015 роки, сформовані у вигляді pcap-файлів [4, 5]. Перший набір даних CTU-Mixed може бути використаний для підготовки екстрактора ознак. Другий набір даних CTU-13 містить розмічені потоки, і його можна використовувати для навчання вирішувальних правил для детектування шкідливого мережевого трафіку.

Необхідно побудувати інформативний екстрактор ознак та надійні вирішувальні правила, використовуючи розмічені та не розмічені набори даних шляхом оптимізації параметрів моделі. У процесі навчання необхідно інформаційний критерій ефективності детектора шкідливого мережевого трафіку.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

$$\bar{E}^{*} = \frac{1}{M}\sum_{m=1}^{M}\max_{\{k\}} E_m^{(k)},$$

де $E_m^{(k)}$ інформаційний критерій ефективності розпізнавання класу $X_m^o$ на $k$ -му кроці навчання;

$\{k\}$ – впорядкована множина навчальних кроків.

Під час роботи, детектора шкідливого трафіку, в режимі екзамену важливо забезпечити обчислювальну ефективність для високо інтенсивного трафіку.

## Модель та метод навчання детектора шкідливого трафіку

Для навчання системи детектування шкідливого трафіку використовуються два набори даних, що зібрані дослідниками CTU з реального мережевого середовища в період 2011 - 2015 роки у вигляді pcap-файлів [4, 5]. Перший набір даних, CTU-Mixed, було запропоновано використовувати для навчання згорткового екстрактора ознак та побудови його апроксимації на основі моделей випадкового лісу. Другий набір даних, CTU-13, пропонується використати для навчання вирішувальних правил детектора шкідливого трафіку. Внутрішні характеристики одиниці трафіку (потоковий пакет або сесія) найкраще відображаються в початковій частині її байтів, яка містить інформацію про з'єднання та деяку інформацію про контент. Процес перетворення pcap-файлу в набір навчальних даних включає три основні етапи: поділ трафіку на дискретні одиниці згідно з одним із рівнів гранулярності, очищення трафіку шляхом видалення порожніх і дублюючих зразків, формування навчальних зображень. Під час поділу трафіку на дискретні одиниці, можна розглянути такі рівні гранулярності : TCP-з'єднання, потоки, сесія, сервіс чи хост. У цій роботі пропонується розділити вхідний трафік на потоки, де ряд пакетів має однаковий кортеж з п'яти елементів: IP-адреса джерела та отримувача, порти джерела та отримувача, номер протоколу. При цьому довжина потоку обмежена 784 байтами, тому довші потоки обрізаються, а коротші – доповнюються нульовими байтами. В результаті у нас є зображення розміром 28x28 пікселів, яке буде надходити на вхід екстрактора ознак. Яскравість кожного пікселя нормалізується до діапазону [0, 1].

Як основа для побудови архітектури екстрактора ознак використовується згорткова мережа, відома як LeNet-5 [5]. Модифікація даної мережі полягає у використанні нефіксованої кількості згорткових фільтрів, яка визначається під час пошарового навчання. Активація пікселів кожного каналу карти ознак пропонується обчислювати на основі жадібного L-0 алгоритму ортогонального узгодженого переслідування (Orthogonal Matching Pursuit) або L1-регуляризованим методом найменших кутів (Least angle regression algorithm) з функцією активації ReLU [17]. Для прискорення моделі в режимі екзамену можна замінити обчислювально інтенсивний пошук розріджених коефіцієнтів на неітераційний апроксимуючий кодер (рис. 1). Відповідно до принципу дистиляції знань навчальний набір для апроксимуючого кодера формується на основі вхідних сигналів розріджено-кодуючого шару та псевдоміток з його виходу. У цьому випадку псевдо-мітки отримують за алгоритмами ортогонального узгодженого переслідування або методом найменших кутів.

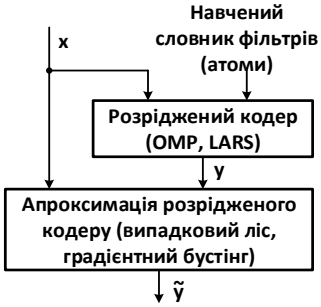



Рис. 1. Діаграма дистиляції знань для кожного шару екстрактора ознак

Пропонується реалізувати розріджене кодування з алгоритмами OMP та LARS, де критерій зупинки базується на досягненні 30% ненульових записів у розрідженому коді. Шар локальної нормалізації контрасту, розміщений після шару підвибірки, перед наступним шаром, посилює інформативні ознаки та послаблює решту пікселів карти ознак.

Набір даних для навчання шару екстрактора ознак формується шляхом декомпозиції зображень або карт ознак на патчі. Ці патчі приводяться до одновимірних векторів, які подають на вхід алгоритму зростаючого розріджено кодуючого нейронного газу, основні етапи якого наведені нижче [18]:

1. Ініціалізація лічильника навчальних векторів $t := 0$.

2. Задаються два початкові вузли (нейрони) $w_a$ та $w_b$, які обраються випадковим шляхом з навчального набору. Вузли $w_a$ та $w_b$ з'єднуються ребром, вік якого дорівнює нулю. Ці вузли вважаються нефіксованими.

3. З набору даних обирається наступний вектор $x$, який приводиться до одиничної довжини (L2-нормалізація).

4. Приведення кожного базисного вектору $w_k, k = \overline{1,M}$ до одиничної довжини (L2-нормалізація).

5. Обчислення міри схожості вхідного вектора $x$ на вузли мережі $w_{s_k} \in W$ для їх сортування

$$-(w_{s_0}^T x)^2 \le ... \le -(w_{s_k}^T x)^2 \le ... \le -(w_{s_{M-1}}^T x)^2 .$$

6. Обирається найближчий вузол $w_{s_0}$ та другий за близькістю вузол $w_{s_1}$.

7. Збільшується на одиницю вік усіх ребер, інцидентних до $w_{s_0}$.

8. Якщо $w_{s_0}$ фіксований, то здійснюється перехід до кроку 9, інакше до кроку 10.

9. Якщо $(w_{s_0}^T x)^2 \ge v$, то перехід до кроку 12. В іншому випадку додається новий нефіксований вузол $w_r$ в точку, що співпадає з вхідним вектором $w_r = x$, також додається ребро, що з'єднує $w_r$ та $w_{s_0}$, потім перехід до кроку 13.

10. Вузол $w_{s_0}$ і його топологічні сусіди (вузли, що з'єднані з ним одним ребром) зміщуються в напрямку до вхідного вектору $x$ згідно правила Ойа [5] за наступними формулами :

$$\Delta w_{s_0} = \varepsilon_b \eta_t y_0 (x - y_0 w_{s_0}),\ y_0 := w_{s_0}^T x,$$

$$\Delta w_{s_n} = \varepsilon_n \eta_t y_n (x - y_n w_{s_n}),\ y_n := w_{s_n}^T x,$$

$$0 < \varepsilon_b << 1,\ 0 < \varepsilon_n << \varepsilon_b,$$

$$\eta_t := \eta_0 (\eta_{final} / \eta_0)^{t/t_{max}} .$$

де $\Delta w_{s_0}$, $\Delta w_{s_n}$ – вектори корекції вагових коефіцієнтів вузла переможця та його топологічних сусідів відповідно;

$\varepsilon_b$, $\varepsilon_n$ – константи сили оновлення вагових коефіцієнтів вузла-переможця та його топологічних сусідів відповідно;

$\eta_0, \eta_t, \eta_{final}$ – початкове, поточне і кінцеве значення коефіцієнту швидкості навчання відповідно.

11. Якщо $(w_{s_0}^T x)^2 \ge v$, то помічаємо нейрон $w_{s_0}$ як фіксований.

12. Якщо $w_{s_0}$ та $w_{s_1}$ з'єднані ребром, то його вік обнуляється, інакше між $w_{s_0}$ та $w_{s_1}$ формується нове ребро з нульовим віком.

13. Всі ребра в графі з віком більше ніж $a_{max}$ видаляються. У випадку коли деякі вузли не мають інцидентних ребер (стають ізольованими), вони також видаляються.

14. Якщо $t < t_{max}$, то переходимо до кроку 15, інакше – збільшуємо лічильник кроків $t := t+1$ і переходимо до кроку 3.

15. Якщо всі нейрони фіксовані, то виконання алгоритму зупиняється, інакше перехід до кроку 3 і починається нова епоха навчання (повторення навчальної множини).

Тонке налаштування екстрактора ознак можна виконати за допомогою алгоритму зворотного розповсюдження з тимчасовим або постійним нейронним класифікатором на виході моделі [17]. Проте в умовах нестаціонарності заздалегідь не може бути відома інформативність ознак, тому тонке налаштування екстрактора не передбачене в нашому алгоритмі. Метою екстрактора ознак є розділення пояснюючих факторів.

Інформаційно-екстремальний класифікатор потребує двійкового представлення вхідного сигналу для побудови вирішальних правил. Ансамбль дерев рішень є обчислювально ефективним методом індукції інформативних бінарних ознак для вхідних спостережень (рис. 2). Для цього вузли дерев рішень нумеруються. Номери ненульових бітів отриманого двійкового коду відповідають номерам вузлів, через які пролягає шлях прийняття рішень.

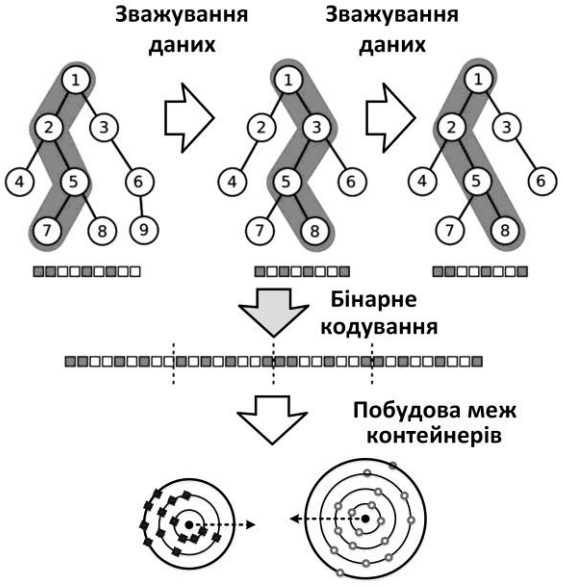



Рис. 2. Архітектура класифікатора

Інформаційно-екстремальний класифікатор в режимі екзамену приймає рішення про належність вхідних даних $x$ з відповідним бінарним поданням $b$ до одного з класів алфавіту $\{X_z^o \mid z = \overline{1,Z}\}$ за максимумом функції належності $\mu_z(b)$ відповідно до формули $\arg\max_z\{\mu_z(b)\}$. При цьому функція належності $\mu_z(b)$ до класу $X_z^o$, оптимальний контейнер якого має опорний вектор $b_z^*$ з розмірністю $N_2$ та

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

радіус $d_z^*$, обчислюється за формулою

$$\mu_z(b) = \exp\left(-\sum_{i=1}^{N_2} b_i \oplus b_{z,i}^* / d_z^*\right). \quad (1)$$

Нехай $D = \{x_j, y_j \mid j = \overline{1,n}\}$ є навчальним набором, де $n$ розмір набору даних та $y_j$ є міткою $j$-го зразка даних, що належить до одного з класів алфавіту $\{X_z^o \mid z = \overline{1,Z}\}$. У цьому випадку класифікатор, що оцінює належність $j$-тої зразка $x_j$ з $N_1$ ознаками до одного $Z$ класів, використовує кодування ознак на основі ансамблю дерев рішень та вирішувальні правила, побудовані у радіально базисному двійковому просторі Хеммінга. Навчання інформаційно-екстремального класифікатора виконується відповідно до наступних кроків:

1. Ініціалізація вагових коефіцієнтів зразків $w_j = 1/n$.

2. Виконувати наступну множину операцій для ітерації $k = 1,...,K$.

3. Формування підвибірки $D_k$ з вибірки $D$ відповідно до розподілу $P(X = x_j) = w_j$.

4. Навчання дерева рішень $T_k$ на підвибірці $D_k$.

5. Бінарне кодування зразка даних $x_j$ з $D$, використовуючи конкатенацію результатів з $T_1,...,T_k$ дерев. Результатом цього кроку є двійкова матриця

$$\{b_{z,s,i} \mid i = \overline{1,N_2}; s = \overline{1,n_z}; z = \overline{1,Z}\},$$

де $N_2$ – кількість індукованих бінарних ознак, а $n_z$ – кількість зразків класу $X_z^o$. Отже, виконується умова рівності $n = \sum_z n_z$.

6. Побудова інформаційно-екстремальних вирішувальних правил в радіальному базисі двійкового простору Хеммінга та обчислення оптимального значення інформаційного критерію:

$$E_z^* = \max_{\{d\}} E_z(d), \quad (2)$$

де $\{d\} = \{0,1,...,\left(\sum_i b_{z,i} \oplus b_{c,i} - 1\right)\}$ – набір концентричних радіусів з центром $b_z$ (опорний вектор) розподілу даних у класі $X_z^o$, який обчислюється за допомогою правила

$$b_{z,i} = \begin{cases} 1, & \text{if } \frac{1}{n_z}\sum_{s=1}^{n_z} b_{z,s,i} > \frac{1}{Z}\sum_{c=1}^{Z}\frac{1}{n_c}\sum_{s=1}^{n_c} b_{z,s,i}; \\ 0, & \text{інакше.} \end{cases} \quad (3)$$

де $E_z$ – критерій ефективності навчання для вирішувального правила класу $X_z^o$, що обчислюється за формулою нормованої модифікації критерія С. Кульбака [7]:

$$E_z = \frac{1-(\alpha_z+\beta_z)}{\log_2(2+\varsigma)-\log_2\varsigma}\cdot\log_2\left[\frac{2-(\alpha_z+\beta_z)+\varsigma}{(\alpha_z+\beta_z)+\varsigma}\right], \quad (4)$$

де $\alpha_z$, $\beta_z$ – частота помилок першого та другого роду щодо належності вхідних векторів до класу $X_z^o$;

$\varsigma$ – будь-яке невелике невід'ємне число, введене для уникнення невизначеності при діленні на нуль.

7. Перевірка отриманих інформаційно-екстремальних правил на наборі даних $D$ та обчислення помилки для кожного зразка з $D$. У режимі екзамену рішення про належність вектору $b$ до одного з класів алфавіту $\{X_z^o \mid z = \overline{1,Z}\}$ приймається відповідно до максимального значення функції належності (1). У цьому випадку функція належності $\mu_z(b)$ для двійкового подання $b$ вхідного зразка $x$ до класу $X_z^o$, оптимальний контейнер якого має опорний вектор $b_z^*$ (3) та радіус $d_z^*$ (2).

8. Оновлення вагових коефіцієнтів $\{w_j\}$ здійснюється пропорційно до помилок класифікації зразків даних $x_j$:

$$w_j = 1 - \mu_{m'}(x_j),\ m' = y_j;$$

$$w_j = \frac{w_j}{\sum_j w_j}.$$

9. Якщо $|E_k^* - E_{k-1}^*| < \varepsilon$ і $k < K/2$, то завершуємо цикл, де $\varepsilon = 0,001$.

Таким чином, отримана модель складається з декількох шарів ансамблів дерев з оптимальними в інформаційному сенсі вирішувальними правилами на виході.

## Результати фізичного моделювання

Навчальний набір даних, що сформований за допомогою CTU-Mixed для навчання екстрактора ознак, містить 10 000 зразків. Для навчання інформаційно-екстремального класифікатора сформовано 1000 зразків на клас, для навчальних та тестових даних. Для алгоритму зростаючого розріджено кодуючого нейронного газу були обрані наступні параметри $\varepsilon_b = 0,5$, $\varepsilon_b = 0,05$, $a_{max} = 100$, $\eta_0 = 1$ та

$\eta_{final} = 0,01$. Параметр порогу фіксації нейрона $\nu$ та параметр максимальної кількості дерев класифікатора K коригуються шляхом прямого перебору значень з кроком. У табл. 1 показано залежність кількості нейронів у першому $M_1$ та другому $M_2$ шарах екстрактора ознак, усереднений за алфавітом класів критерій ефективності навчання $\bar{E}$, та точність на валідаційній вибірці від параметра $\nu$. Для ансамблю дерев рішень максимальна глибина встановлюється рівна 5, а максимальна кількість ознак рівна $\sqrt{N_1}$.

Таблиця 1

Залежність характеристик моделі від параметра алгоритму розріджено кодуючого нейронного газу $\nu$

| $\nu$ | $M_1$ | $M_2$ | $\bar{E}$ | Валідаційна точність |
|---|---|---|---|---|
| 0,10 | 15 | 11 | 0,106 | 74 |
| 0,15 | 17 | 13 | 0,138 | 77 |
| 0,20 | 23 | 13 | 0,138 | 77 |
| 0,25 | 25 | 13 | 0,138 | 77 |
| 0,30 | 27 | 15 | 0,149 | 78 |
| 0,35 | 27 | 15 | 0,220 | 83 |
| 0,40 | 33 | 17 | 0,255 | 85 |
| 0,45 | 34 | 22 | 0,255 | 85 |
| 0,50 | 40 | 25 | 0,366 | 90 |
| 0,55 | 49 | 31 | 0,459 | 93,0 |
| 0,60 | 66 | 43 | 0,466 | 93,2 |
| 0,65 | 70 | 45 | 0,501 | 94,1 |
| 0,70 | 99 | 45 | 0,550 | 95,2 |
| 0,75 | 145 | 57 | 0,554 | 95,3 |
| 0,80 | 161 | 120 | 0.591 | 96,1 |
| 0,85 | 220 | 147 | 0,603 | 95,4 |
| 0,90 | 322 | 238 | 0,611 | 95,0 |

Аналіз табл. 1 показує, що підвищення значення порогу $\nu$ призводить до збільшення кількості нейронів у процесі навчання екстрактора ознак без вчителя. У той же час підвищення порогу від 0,8 до 0,9 практично не впливає на точність прийняття рішення. Це означає, що значення $\nu^* = 0,8$ є оптимальним і дозволяє сформувати більш компактне подання ознакового опису (стиснення), тим часом $\nu = 0,9$ дозволяє сформувати розріджене подання на основі надповного базису.

Дистиляція знань реалізується за допомогою регресійної моделі випадкового лісу як моделі студента, де кількість дерев рішень обмежена кількістю рівній 150. Отримана модель має рівнозначну точність. У цьому випадку час в режимі екзамену скорочується в 65 разів.

На рис. 3 показано графік зміни максимумів інформаційного критерію (4), усередненого за алфавітом класів, залежно від кількості дерев рішень в інформаційно-екстремальному класифікаторі з $\nu^* = 0,8$. У цьому випадку максимальна кількість дерев обмежена значенням $K = 100$.

Аналіз рис. 3 показує, що оптимальне значення гіперпараметра $K^*$ дорівнює 30. Подальше збільшення параметра K не приводить до підвищення точності вирішувальних правил. При оптимальних параметрах екстрактора та класифікатора точність виявлення шкідливого трафіку становить 96,1%. Це свідчить про інформативність сформованого ознакового опису спостережень. На рис. 4 показано залежність інформаційного критерію (4) від кодового радіуса контейнера з класів.

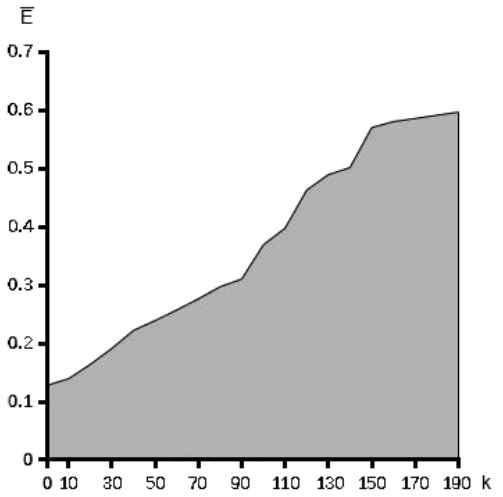
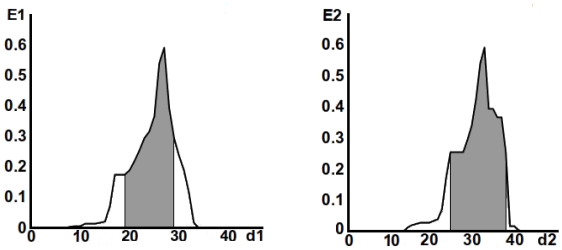



Рис. 3. Графік зміни усередненого інформаційного критерію (4) залежно від кількості дерев рішень в інформаційно-екстремальному класифікаторі




а б

Рис. 4. Графіки залежності інформаційного критерію (4) від радіусів контейнерів класів : а – клас нормального трафіку; б – клас шкідливого трафіку

Аналіз рис. 4 показує, що максимальні значення інформаційного критерію навчання для першого та другого класів дорівнюють $E_1^* = 0,590$ та $E_2^* = 0,597$ відповідно, а оптимальні значення радіу-

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

сів контейнерів класів розпізнавання рівні $d_1^* = 26$, $d_2^* = 32$ (в кодових одиницях) відповідно. У цьому випадку міжцентрова кодова відстань Хеммінга дорівнює 65, що вказує на компактність розподілів векторів ознак та чіткість розбиття у бінарному просторі Хеммінга.

Таким чином, запропонований алгоритм навчання дозволяє автоматично визначати оптимальну кількість нейронів на кожному шарі. При цьому, апроксимація розрідженого кодера неітеративною моделлю регресії випадкового лісу дозволила значно прискорити режим екзамену. Результати моделювання на даних із наборів CTU-Mixed та CTU-13 показують, що побудовані вирішувальні правила є кращими за отримані в [4] та [5] і є прийнятні для практичного використання.

## Висновки

1. Наукова новизна отриманих результатів:

– вперше запропоновано алгоритм зростаючого розріджено кодуючого нейронного газу, який дозволяє без вчителя навчити екстрактор ознак з оптимальною кількістю вузлів для кожного шару моделі;

– вперше було запропоновано застосовувати принцип дистиляції знань для зменшення обчислювальних витрат в алгоритмах з розрідженим кодуванням шляхом застосування апроксимації моделлю випадкового лісу, яка в режимі екзамену є неітеративною та обчислювально ефективною;

– вперше запропоновано інформаційно-екстремальний алгоритм навчання з учителем для побудови вирішувальних правил детектора шкідливого мережевого трафіку.

2. Практична цінність отриманих результатів для систем детектування шкідливого трафіку полягає в розробці нового методу навчання, який ефективно використовує як розмічені, так і нерозмічені навчальні дані. Результати моделювання за допомогою наборів даних CTU-Mixed та CTU-13 підтверджують ефективність отриманих вирішувальних правил щодо детектування шкідливого трафіку на тестових зразках. У цьому випадку точність вирішувальних правил детектора шкідливого трафіку становить 96,1 %.

*Робота виконана на базі лабораторії інтелектуальних систем кафедри комп'ютерних наук Сумського державного університету при фінансовій підтримці МОН України в рамках держбюджетної науково-дослідної роботи* ***ДР № 0117U003934.***

## Література

*1. Skrzewski, M. Flow Based Algorithm for Malware Traffic Detection [Text] / M. Skrzewski // Proceedings of the 18th Conference Computer Networks (Communications in Computer and Information Science). – Ustroń, Poland, 14–18 June, 2011. – Springer, 2011. – Vol. 160. – P. 271-280.*
*DOI: 10.1007/978-3-642-21771-5_29.*

*2. Malware traffic detection using tamper resistant features [Text] / Z. Berkay Celik, R. J. Walls, P. McDaniel, A. Swami // Proceedings of the IEEE MILCOM 2015 – 2015 IEEE Military Communications Conference.*
*– Tampa, FL, 26–28 October 2015. – IEEE, 2015. – P. 330-335. DOI: 10.1109/MILCOM.2015.7357464.*

*3. Ferreira, D. C. A meta-Analysis approach for feature selection in network traffic research [Text] / D. C. Ferreira, F. I. Vázquez, G. Vormayr // Proceedings of the Reproducibility Workshop. – ACM, 2017. – P. 17-20.*

*4. Autoencoder-based feature learning for cyber security applications [Text] / M. Yousefi-Azar, V. Varadharajan, L. Hamey, U. Tupakula // Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN). – Anchorage, Alaska, USA, 14–19 May 2017.*
*– P. 3854-3861. DOI: 10.1109/IJCNN.2017.7966342*

*5. Malware traffic classification using convolutional neural network for representation learning [Text] /*
*W. Wang, M. Zhu, X. Zeng, X. Ye, Y. Sheng // Proceedings of the 31st International Conference on Information Networking (ICOIN 2017). – Da Nang, Vietnam, 5–8 August, 2017. – P. 712-717.*
*DOI: 10.1109/ICOIN.2017.7899588*

*6. Going deeper with convolutions [Text] / C. Szegedy, W. Liu, Y. Jia, P. Sermanet et al. // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – Boston, MA, 7–12 June, 2015. – P. 1-9.*
*DOI: 10.1109/ CVPR.2015.7298594.*

*7. Convolutional neural networks for time series classification [Text] / B. Zhao, H. Lu, S. Chen, J. Liu, D. Wu // Journal of Systems Engineering and Electronics. – 2017. – Vol. 28, N. 1. – P. 162–169.*
*DOI: 10.21629/JSEE.2017.01.18.*

*8. Feng, Q. Compressed auto-encoder building block for deep learning network [Text] / Q. Feng, C. L. P. Chen, L. Chen // Proceedings of the 3rd International Conference on Informative and Cybernetics for Computational Social Systems (ICCSS). – Jinzhou, 26–29 Aug, 2016. – P. 131-136.*
*DOI: 10.1109/ICCSS.2016.7586437*

*9. Weed identification based on K-means feature learning combined with convolutional neural network [Text] / J. Tang, D. Wang, Z. Zhang, L. He, X. Jing, Y. Xu // Computers and Electronics in Agriculture. – 2017. – Vol. 135. – P. 63-70.*
*DOI: 10.1016/j.compag.2017.01.001.*

*10. Labusch, K. Sparse coding neural gas: learning of overcomplete data representations [Text] / K. Labusch, E. Barth, T. Martinetz // Neurocomputing. – 2009. – Vol. 72, Iss. 7-9. – P. 1547-1555. DOI: 10.1016/j.neucom.2008.11.027.*

*11. Mrazova, I. Image Classification with Growing Neural Networks [Text] / I. Mrazova, M. Kukacka // International Journal of Computer Theory and Engineering. – 2013 – Vol. 5, N. 3. – P. 422-427. DOI: 10.7763/IJCTE.2013.V5.722.*

*12. Palomo, E. J. The Growing Hierarchical Neural Gas Self-Organizing Neural Network [Text] / E. J. Palomo, E. López-Rubio // IEEE Transactions on Neural Networks and Learning System. – 2017. – Vol. 28, N. 9. – P. 2000-2009. DOI: 10.1109/ TNNLS.2016.2570124.*

*13. Layer-Level Knowledge Distillation for Deep Neural Network Learning [Text] / H.-T. Li, S.-C. Lin, C.-Y. Chen, C.-K. Chiang // Applied Sciences. – 2019. – Vol. 9. – P. 1966. DOI: 10.3390/app9101966.*

*14. Zhou, Y. Approximation Trees: Statistical Stability in Model Distillation [Text] / Y. Zhou, Z. Zhou, G. Hooker // ArXiv. – 2018. – Vol. abs/1808.07573.*

*15. Deep learning of support vector machines with class probability output networks [Text] / S. Kim, Z. Yu, R. Man Kil, M. Lee // Neural Networks. – 2015. – Vol. 64. – P. 19–28. DOI: 10.1016/ j.neunet.2014.09.007.*

*16. The model and training algorithm of compact drone autonomous visual navigation system [Text] / V. Moskalenko, A. Moskalenko, A. Korobov, V. Semashko // Data. – 2019. – Vol. 4, Iss. 1. – P. 1-14. DOI: 10.3390/data4010004.*

*17. Gwon, Y. Deep Sparse-coded Network (DSN) / Y. Gwon, M. Cha, H. T. Kung // International Conference on Pattern Recognition (ICPR). – 2016. – P. 2610–2615. DOI: 10.1109/ ICPR.2016.7900029.*

*18. Москаленко, В. В. Модель і алгоритм навчання детектора шкідливого трафіку на основі модифікації зростаючого нейронного газу [Текст] / В. В. Москаленко, А. С. Москаленко, М. О. Зарецький // Радіоелектронні і комп'ютерні системи. – 2018. – № 3(87). – С. 11-19. DOI: 10.32620/reks.2018.3.02.*

## References

1. Skrzewski, M. Flow Based Algorithm for Malware Traffic Detection. *Proc. of the 18th Conference Computer Networks (Communications in Computer and Information Science)*, Ustroń, Poland, 2011, vol. 160, pp. 271-280. DOI: 10.1007/978-3-642-21771-5_29.

2. Berkay Celik, Z., Walls, R., McDaniel, P. and Swami, A. Malware traffic detection using tamper resistant features. *MILCOM 2015 – 2015 IEEE Military Communications Conference*, 2015, pp. 330–335. DOI: 10.1109/MILCOM.2015.7357464.

3. Ferreira, D. C., Vázquez F. I., Vormayr, G. A meta-Analysis approach for feature selection in network traffic research. *Proceedings of the Reproducibility Workshop*. ACM, 2017, pp. 17-20

4. Yousefi-Azar, M., Varadharajan, V., Hamey, L., Tupakula, U. Autoencoder-based feature learning for cyber security applications. *Proc. of the 2017 International Joint Conference on Neural Networks (IJCNN)*. Anchorage, Alaska, USA, 2017, pp. 3854-3861. DOI: 10.1109/IJCNN.2017.7966342.

5. Wang, W. Zhu, M., Zeng, X., Ye, X., Sheng, Y. Malware traffic classification using convolutional neural network for representation learning. *Proc. of the 31st International Conference on Information Networking (ICOIN 2017)*. Da Nang, Vietnam, 2017, pp. 712-717. DOI: 10.1109/ICOIN.2017.7899588.

6. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. Going deeper with convolutions. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015, pp. 1-9. DOI: 10.1109/CVPR.2015.7298594.

7. Zhao, B., Lu, H., Chen, S., Liu, J., Wu,D. Convolutional neural networks for time series classification. *Journal of Systems Engineering and Electronics*, 2017, vol. 28, no. 1, pp. 162-169. DOI: 10.21629/JSEE.2017.01.18.

8. Feng, Q. Chen, C. L. P., Chen, L. Compressed auto-encoder building block for deep learning network. *Proc. of the 3rd International Conference on Informative and Cybernetics for Computational Social Systems (ICCSS)*, Jinzhou, 2016, pp. 131-136. DOI: 10.1109/ICCSS.2016.7586437.

9. Tang, J., Wang, D., Zhang, Z., He, L., Jing, X., Xu, Y. Weed identification based on K-means feature learning combined with convolutional neural network. *Computers and Electronics in Agriculture*, 2017, vol. 135, pp. 63-70. DOI: 10.1016/j.compag.2017.01.001.

10. Labusch, K., Barth, E., Martinetz, T. Sparse coding neural gas: learning of overcomplete data representations. *Neurocomputing*, 2009, vol. 72, iss. 7-9, pp. 1547-1555. DOI: 10.1016/j.neucom.2008.11.027.

11. Mrazova, I., Kukacka, M. Image Classification with Growing Neural Networks. *International Journal of Computer Theory and Engineering*, 2013, vol. 5, no. 3, pp. 422-427. DOI: 10.7763/IJCTE.2013.V5.722.

12. Palomo, E. J., López-Rubio, E. The Growing Hierarchical Neural Gas Self-Organizing Neural Network. *IEEE Transactions on Neural Networks and Learning System*, 2017, vol. 28, no. 9, pp. 2000-2009. DOI: 10.1109/TNNLS.2016.2570124.

13. Li, H.-T., Lin, S.-C., Chen, C.-Y., Chiang, C.-K. Layer-Level Knowledge Distillation for Deep Neural Network Learning. *Applied Sciences*, 2019, vol. 9, pp. 1966. DOI: 10.3390/app9101966.

14. Zhou, Y., Zhou, Z., Hooker, G. Approximation Trees: Statistical Stability in Model Distillation. *ArXiv*, 2018, vol. abs/1808.07573.

15. Kim, S., Yu, Z., Man Kil, R., Lee, M. Deep learning of support vector machines with class probability output networks. *Neural Networks*, 2015, vol. 64, pp. 19-28. DOI: 10.1016/j.neunet.2014.09.007.

16. Moskalenko, V., Moskalenko, A., Korobov, A., Semashko, V. The model and training algorithm

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

of compact drone autonomous visual navigation system. *Data*, vol. 4, iss. 1, 2019, pp. 1-14. DOI: 10.3390/data4010004.

17. Gwon, Y., Cha, H., Kung, H.T. Deep Sparse-coded Network (DSN). *International Conference on Pattern Recognition (ICPR)*, 2016, pp. 2610–2615. DOI: 10.1109/ ICPR.2016.7900029.

18. Moskalenko, V. V., Moskalenko, A. S., Zarecz`ky`j, M. O. Model' i alhorytm navchannya detektora shkidlyvoho trafiku na osnovi modyfikatsiyi zrostayuchoho neyronnoho hazu [Model and training algorithm of malware traffic detector based on modification of growing neural gas]. *Radioelektronni i komp'uterni sistemi – Radioelectronic and computer systems*, 2018, no. 3(87), pp. 11-19, 2018. DOI: 10.32620/reks.2018.3.02.

*Надійшла до редколегії 05.03.2020 розглянута на редколегії 15.04.2020*

## МНОГОСЛОЙНАЯ МОДЕЛЬ И МЕТОД ОБУЧЕНИЯ ДЛЯ ДЕТЕКТИРОВАНИЯ ВРЕДОНОСНОГО ТРАФИКА НА ОСНОВЕ АНСАМБЛЯ ДЕРЕВЬЕВ РЕШЕНИЙ

***В. В. Москаленко, Н. А. Зарецкий, А. С. Москаленко, А. М. Кудрявцев, В. А. Семашко***

Предложена модель и метод обучения многослойного экстрактора признаков и решающих правил для детектора вредоносного трафика. Модель экстрактора признаков основана на сверточной разреженно кодирующей сети, разреженный кодер которой аппроксимируется моделью регрессионного случайного леса согласно принципам дистилляции знаний. При этом разработан алгоритм растущего разреженно кодирующего нейронного газа для обучения экстрактора признаков без учителя с автоматическим определение необходимой количества признаков на каждом слое. На этапе обучения экстрактора признаков для реализации разреженного кодирования использовано жадный L0-регуляризированный метод ортогонального согласованного преследования (Orthogonal Matching Pursuit), а при дистилляции знаний дополнительно использовался L1-регуляризированный метод наименьших углов (Least angle regression algorithm). Благодаря эффекту редукции причины полученные признаки являются некоррелированными, а сформированное признаковое описание устойчиво к помехам и состязательным (Adversarial) атакам. Предложенный экстрактор признаков учится без учителя для разделения объясняющих факторов и позволяет с максимальной эффективностью использовать неразмеченные обучающие данные, объем которых, как правило, достаточно большой. Как модель решающих правил предложено использовать двоичный кодер наблюдений на основе ансамбля деревьев решений и информационно-экстремальные разделяющие замкнутые гиперповерхности (контейнеры) классов, которые восстанавливаются в радиальном базисе двоичного пространства Хемминга. Добавление кодирующих деревьев происходит по принципу бустинга, а радиус контейнеров классов оптимизируется путем прямого перебора. Информационно-экстремальный классификатор характеризуется низкой вычислительной сложностью и высокой обобщающей способностью для малых наборов размеченных обучающих данных. Результаты верификации обученной модели на открытых тестовых наборах данных CTU подтверждают пригодность предложенных алгоритмов для практического применения, поскольку точность распознавания вредоносного трафика составляет 96,1 %.

**Ключевые слова:** система обнаружения угроз; свёрточная разреженно кодирующая модель; растущий нейронный газ; ансамбль деревьев решений; регрессионный случайный лес; информационный критерий; дистилляция знаний; информационно-экстремальное машинное обучение.

## MULTI-LAYER MODEL AND TRAINING METHOD FOR MALWARE TRAFFIC DEETECTION BASED ON DECISION TREE ENSEMBLE

***V. V. Moskalenko, M. O. Zaretskyi, A.S. Moskalenko, A. M. Kudryavtsev, V. A. Semashko***

The model and training method of multilayer feature extractor and decision rules for a malware traffic detector is proposed. The feature extractor model is based on a convolutional sparse coding network whose sparse encoder is approximated by a regression random forest model according to the principles of knowledge distillation. In this case, an algorithm of growing sparse coding neural gas has been developed for unsupervised training the features extractor with automatic determination of the required number of features on each layer. As for feature extractor, at the training phase to implement of sparse coding the greedy L1-regularized method of Orthogonal Matching Pursuit was used, and at the knowledge distillation phase, the L1-regularized method at the least angles (Least regression algorithm) was additionally used. Due to the explaining-away effect, the extracted features are uncorrelated and robust to noise and adversarial attacks. The proposed feature extractor is unsupervised trained to separate the explanatory fac-

tors and allows to use the unlabeled training data, which are usually quite large, with the maximum efficiency. As a model of the decision rules proposed to use the binary encoder of input observations based on an ensemble of decision trees and information-extreme closed hyper-surfaces (containers) for class separation, that are recovery in radial-basis of Hemming' binary space. The addition of coding trees is based on the boosting principle, and the radius of class containers is optimized by direct search. The information-extreme classifier is characterized by low computational complexity and high generalization capacity for small sets of labeled training data. The verification results of the trained model on open CTU test data sets confirm the suitability of the proposed algorithms for practical application since the accuracy of malware traffic detection is 96.1 %.

**Keywords:** intrusion detection system; convolutional sparse coding model; growing sparse coding neural gas; decision tree ensemble; regression random forest; information criterion; knowledge distillation; information-extreme machine learning.

**Москаленко В'ячеслав Васильович** – канд. техн. наук, доцент каф. комп'ютерних наук, Сумський державний університет, Україна.

**Зарецький Микола Олександрович** – асп. каф. комп'ютерних наук, Сумський державний університет, Україна.

**Москаленко Альона Сергіївна** – канд. техн. наук, старш. викл. каф. комп'ютерних наук Сумського державного університету, Україна.

**Кудрявцев Антон Михайлович** – студ. каф. комп'ютерних наук Сумського державного університету, Україна.

**Семашко Віктор Анатолійович** – асп. каф. комп'ютерних наук, Сумський державний університет, Україна.

**Viacheslav Moskalenko** – PhD, associate professor of Computer Sciences Department of Sumy State University, Sumy, Ukraine,
e-mail: v.moskalenko@cs.sumdu.edu.ua, ORCID Author ID: 0000-0001-6275-9803. Scopus Author ID: 57189099775.

**Nikolay Zaretskyi** – PhD student of Computer Sciences Department of Sumy State University, Sumy, Ukraine,
e-mail: n.zaretskij@gmail.com, ORCID Author ID: 0000-0001-9117-5604, Scopus Author ID: 57213687285.

**Alona Moskalenko** – PhD, senior lecturer of Computer Sciences Department of Sumy State University, Sumy, Ukraine,
e-mail: a.moskalenko@cs.sumdu.edu.ua, alenarizhova@gmail.com, ORCID Author ID: 0000-0003-3443-3990, Scopus Author ID: 57148522500.

**Anton Kudryavtsev** – student of Computer Sciences Department of Sumy State University, Sumy, Ukraine,
e-mail: kam123ua@gmail.com, ORCID Author ID: 0000-0003-0967-0185.

**Viktor Semashko** – PhD student of Computer Sciences Department of Sumy State University, Sumy, Ukraine,
e-mail: viktor.s.5994@gmail.com, ORCID Author ID: 0000-0002-9765-876X, Scopus Author ID: 57210589608.