ISSN 2226-3780

УДК 004.056.53
DOI: 10.15587/2312-8372.2015.37423

Бурячок В. Л.,
Грищук Р. В.,
Мамарєв В. М.

МЕТОД ПОБУДОВИ КЛАСИФІКАТОРА КІБЕРАТАК НА ДЕРЖАВНІ ІНФОРМАЦІЙНІ РЕСУРСИ

Запропоновано двоетапну схему класифікації стану інформаційно-телекомунікаційної системи, що ґрунтується на бінарному групуванні шаблонів її поведінки. Розроблено метод побудови класифікатора кібератак на базі дерев рішень та оптимізованих потоках вхідних даних. Застосування методу дозволяє в рази скоротити час побудови та функціонування класифікаційної моделі, залишаючи точності класифікації шаблонів поведінки системи у заданих діапазонах.

Ключові слова: *інформаційно-телекомунікаційна система, кібератака, класифікатор, класифікація, дерева рішень, оптимізація.*

1. Вступ

На даний час ефективна робота державних установ та організацій у значній мірі залежить від надійності функціонування інформаційно-телекомунікаційних систем (ІТС) та захищеності їх інформаційних ресурсів. Не зважаючи на впровадження різноманітних рішень, спрямованих на підвищення рівня захищеності державних інформаційних ресурсів в ІТС, динаміка кіберінцидентів пов'язаних з ними, залишається достатньо високою [1].

Однією з проблем, яка стримує впровадження ефективних систем захисту інформації в ІТС, наприклад таких, як системи виявлення атак (СВА), є проблема забезпечення ними достовірної та оперативної класифікації патернів подій в системі. Зважаючи на це, підвищення ефективності виявлення кібератак на державні інформаційні ресурси, залишається актуальним завданням.

2. Аналіз літературних даних та постановка проблеми

Аналіз останніх досліджень і публікацій [2–5] показав, що дослідження пов'язані з підвищенням ефективності функціонування СВА, як правило, проводяться за двома основними напрямами. Перший напрям полягає у розробці нових методів класифікації кібератак на базі парадигми штучного інтелекту [2, 3]. Другий напрям пов'язаний з удосконаленням відомих алгоритмів класифікації [4, 5]. У контексті підвищення ефективності функціонування СВА, не зважаючи на переваги та недоліки кожного з напрямів, вони обидва залишаються актуальними, а тому й надалі інтенсивно розвиваються.

Альтернативою вищезазначеним підходам є подальший розвиток класифікаторів кібератак, в основу яких покладено дерева прийняття рішень. Останні, за умови правильної їх побудови, дають можливість отримати достатньо достовірні результати класифікації та мають відносно низьку обчислювальну складність. Тому удосконалення відомих методів побудови класифікаторів кібератак на державні інформаційні ресурси на основі дерев прийняття рішень, є пріоритетним напрямом наукових досліджень.

3. Об'єкт, ціль та задачі дослідження

Об'єктом дослідження є класифікатор кібератак. При проведені дослідження ставилося за мету удосконалення методів побудови класифікатора кібератак, в основу якого покладено дерева рішень, у напрямі підвищення швидкодії СВА типової конфігурації без втрати точності класифікації.

Для досягнення поставленої мети вирішувалися наступні частинні задачі:

— аналіз архітектури СВА, сучасних підходів та технічних рішень, спрямованих на підвищення ефективності їх функціонування;

— розробка структурної схеми класифікатора, що дозволяє реалізувати двоетапну бінарну схему класифікації стану ІТС;

— розробка методу побудови класифікатора на базі дерев рішень та оптимізованих потоків вхідних даних;

— верифікація розробленого методу.

4. Матеріали та методи дослідження

З метою забезпечення уніфікованості результатів дослідження з іншими авторами, роботи яких присвячені питанням оптимізації розмірності потоків вхідних даних класифікаторів, в якості навчальних і тестових даних обрані загальнодоступні та широко відомі бази шаблонів поведінки ІТС KDD99 та NSL-KDD.

Обчислення проведенні із застосуванням ПЕОМ з наступними параметрами: операційна система Windows 7 Ultimate SP1 (x64); процесор Pentium Dual-Core T4300@2.10Ghz 2.10Ghz; ОЗП 2,00 Gb.

5. Результати досліджень класифікації кібератак з ціллю подальшої ефективної протидії ним

Відомо [6], що СВА типової конфігурації — це спеціалізований програмно-апаратний комплекс, який призначений для виявлення та класифікації кібератак з ціллю подальшої ефективної протидії ним. Так, згідно з [6], типова архітектура СВА включає п'ять основних компонент:

модуль управління компонентами і налаштувань, модулі-сенсори, модуль виявлення/розпізнавання атак, модуль реагування на виявлені атаки та базу шаблонів поведінки системи (рис. 1).

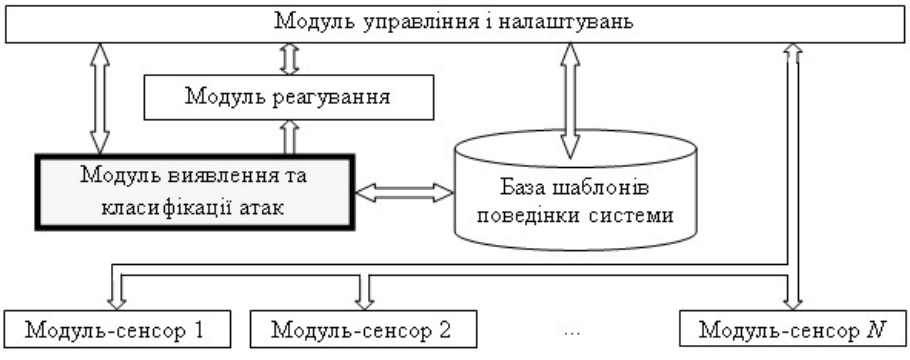

Рис. 1. Типова архітектура системи виявлення атак

Модуль виявлення атак (рис. 1) є одним з основних елементів СВА, від якого залежить ефективність функціонування всієї системи. Технологія виявлення атак СВА у даному модулі передбачає виконання процедур виявлення зловживань відповідною системою — *misuse detection systems* та виявлення аномалій, що покладаються на систему *anomaly detection systems*.

Відомі технічні рішення [7–10], спрямовані на підвищення ефективності функціонування модуля виявлення атак, зводяться до синтезу класифікаторів шляхом групування алгоритмів класифікації. Так, у [8, 9] задачу виявлення та класифікації вирішено за допомогою класифікатора на базі модулярних нейронних мереж. У [10] задача класифікації станів розв'язана за шляхом агрегації ансамблю дерев прийняття рішень. Але, якщо врахувати високу обчислювальну складність, яка виникає при агрегації зазначених алгоритмів, то можна зробити висновок — швидкодія побудованих на їх базі СВА суттєво знижується.

Для забезпечення заданих показників швидкодії СВА при використанні технології дерев прийняття рішень у відповідному модулі, в роботі запропоновано застосовувати оптимізовані потоки вхідних даних [11]. Тоді, з урахуванням [11], структурна схема класифікатора матиме вигляд (рис. 2).

Основною особливістю запропонованої схеми класифікації (рис. 2), на відміну від відомих, є застосування бінарного типу класифікації та двоетапної схеми класифікації потоку вхідних даних (ПВхД). Так, згідно зі схемою, на першому етапі здійснюється бінарна класифікація патернів подій — шаблонів нормальної поведінки (ШНП) та шаблонів атак (ША). На другому етапі, згідно із запропонованою схемою, реалізується процедура бінарної класифікації атак першої та другої груп. В основу принципу групування типів атак («дуетів») покладено припущення про їх максимальну віддаленість у просторі параметрів.

Розглянемо призначення основних елементів класифікатора (рис 2).

Модулі оцінювання інформативності параметрів шаблонів і групування типів атак застосовуються на етапі навчання класифікатора. На етапі виявлення атак модулі реалізують процедури оцінювання інформативності параметрів ШНП та ША, на етапі класифікації використовуються для оцінювання інформативності параметрів ША та їх групування за типами. З цією метою як оцінку ваги для кожного з параметрів шаблонів запропоновано обрати нормований приріст інформації *GainRatio*.

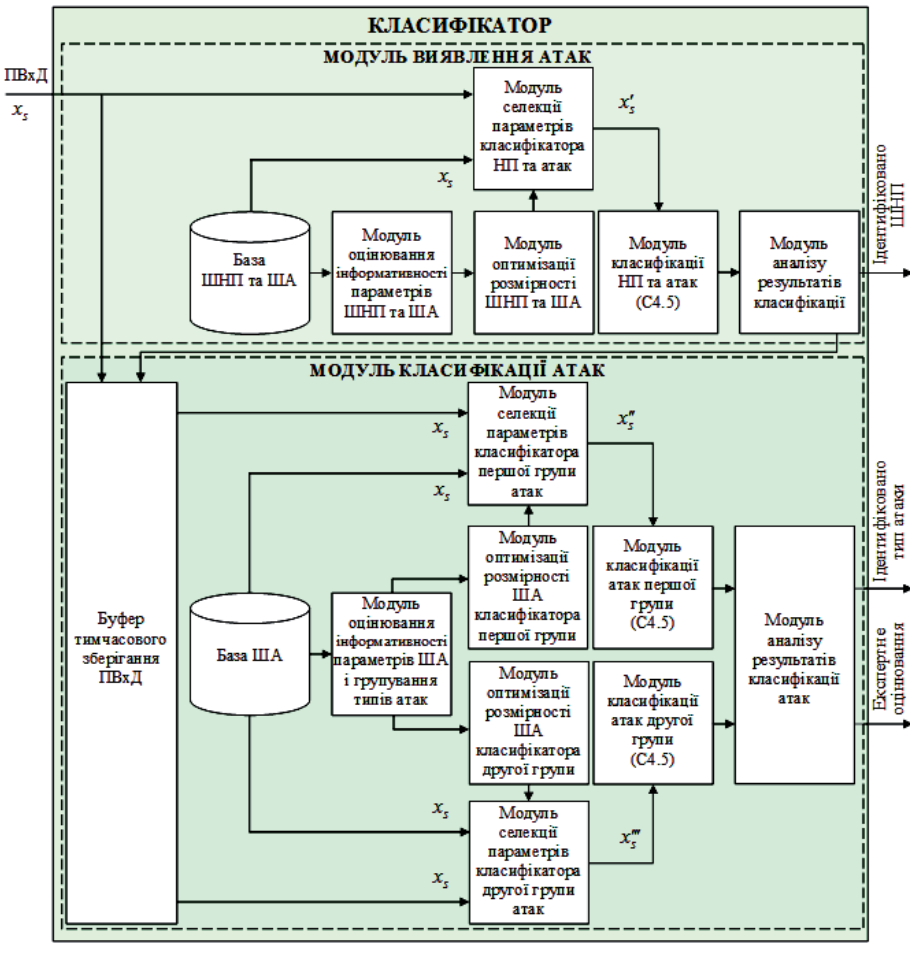

Рис. 2. Принцип побудови класифікатора на основі дерев прийняття рішень та оптимізованих потоків вхідних даних: структурна схема

Дана процедура реалізується згідно з методом, розробленим у [11]. Так зважений приріст інформації *GainRatio* за кожним з параметрів шаблонів поведінки системи визначається згідно з виразом:

$$GainRatio(X,Y)=\frac{H(X)-H(X|Y)}{H(Y)}, \quad (1)$$

де $H(X)$ — ентропія стану ІТС; $H(X|Y)$ — взаємна ентропія, що описує ентропію стану ІТС за умови набуття параметром Y конкретних значень; $H(Y)$ — ентропія Y параметру.

Також даний модуль забезпечує групування ША в «дуети» за принципом забезпечення їх максимальної віддаленості у просторі параметрів. Групування виконується на основі розрахованих сумарних значень приросту інформації шести «дуетів» шаблонів атак. Зі сформованої множини «дуетів» обираються 2 з максимальним значенням $\sum GainRatio$.

Модулі оптимізації розмірності шаблонів на етапі навчання за результатами ранжування параметрів для пари ШНП-ША та двох «дуетів» ША, розраховують оптимізовані розмірності останніх згідно з методом [12].

Модулі селекції параметрів (рис. 2) виконують функцію відбору визначених *модулем оптимізації розмірності шаблонів* оптимальних кількостей параметрів шаблонів під час навчання класифікатора та скорочення розмірності ПВхД під час функціонування класифікатора.

Модулі класифікації здійснюють віднесення патернів ПВхД до попередньо визначеної групи станів. Алгоритм С4.5, що покладено в основу даного модуля, здійснює рекурсивне розбиття ПВхД на асоційовані з класами кібератак підмножини. Процедура розбиття виконується на основі вирішальних правил. Вирішальні правила в свою чергу утворюють ієрархічну деревоподібну структуру (рис. 3), що реалізує процедури класифікації кібератак.

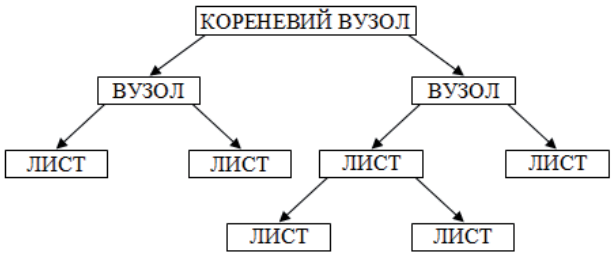

Рис. 3. Структурна схема дерева рішень

Перевагами застосування алгоритму С4.5 на основі оптимізованих ПВхД є:

— висока точність моделей класифікації кібератак, яка є співвимірною з генетичними алгоритмами та нейронними мережами;
— можливість створення класифікаційних моделей слабко формалізованих завдань;
— відсутність вимог до вибору ПВхД на етапі побудови моделі;
— здатність до створення класифікаційних моделей з використанням числових і категоріальних типів даних;
— здатність трансформації отриманих дерев у правила і навпаки;
— здатність до класифікації кібератак у разі пропусків або неповноти потоків вхідних даних;
— непараметричність побудованих моделей;
— лінійна обчислювальна складність;
— стійкість при обробці зашумлених ПВхД;
— відносно низькі, порівняно з нейронними мережами і імунними системами, витрати часу на побудову класифікаційної моделі.

При цьому основним недоліком застосування алгоритму С4.5 на основі оптимізованих ПВхД, є його здатність до перенавчання. Перенавчання — це процес розбиття класифікаційного дерева доти, доки не буде одержано абсолютно чисті листки або доки вони не міститимуть тільки одне спостереження. Якісно ефект перенавчання подано на рис. 4.

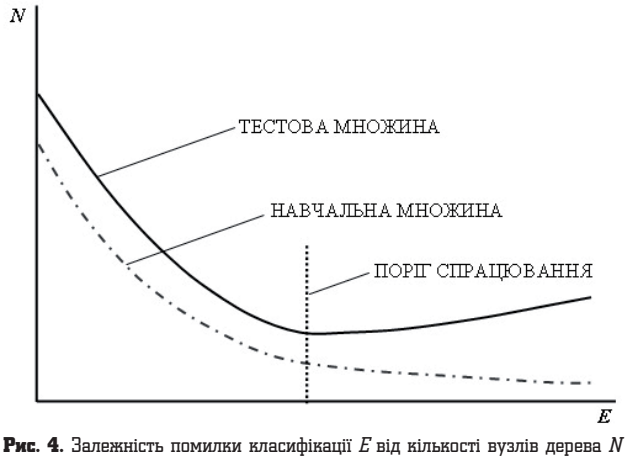

Рис. 4. Залежність помилки класифікації E від кількості вузлів дерева N

Аналіз рис. 4 показує, що збільшення кількості вузлів веде до постійного зменшення помилки класифікації. Помилка навченої «з учителем» класифікаційної моделі на тестовій множині знижується тільки до певного значення порогу. Коли дерево стає занадто складним, модель втрачає стійкість і кількість помилок на тестовій множині починає зростати. Тому для забезпечення заданих показників ефективності класифікаційних моделей побудованих на базі алгоритму С4.5 в модулі застосовуються методи ранньої зупинки та відсікання гілок. Перший здійснює примусову зупинку алгоритму за допомогою умови, при виконанні якої зростання дерева автоматично завершується. Другий виконує побудову повного дерева, а потім здійснює його спрощення шляхом відсікання тих гілок, які є малоінформативними.

Модулі аналізу результатів класифікації атак (рис. 2) виконують інтерпретацію результатів класифікації та формують інформаційні повідомлення у вигляді зрозумілому для користувача ІТС.

На рис. 5 на основі методології ARIS [13] побудовано EPC-діаграму процесу навчання класифікатора на базі дерев рішень та оптимізованих потоках даних. Стартовою подією для початку процесу навчання класифікатора є отримані з навчальних баз шаблони поведінки ІТС. Після цього в модулях виявлення та класифікації атак паралельно виконуються процеси оцінювання інформативності параметрів та групування ША у «дуети», що характеризуються максимальною віддаленістю

у просторі параметрів. Після чого виконуються розрахунки оптимальних розмірностей відповідних шаблонів поведінки та скорочення кількості їх параметрів до заданої. Оптимізовані таким чином патерни забезпечують побудову класифікаційних моделей модулів виявлення та класифікації атак. У разі забезпечення вимог щодо точності класифікації навчання класифікатора вважається закінченим. З метою перевірки адаптивності побудованого класифікатора виконується його перевірка на тестових базах шаблонів поведінки ІТС.

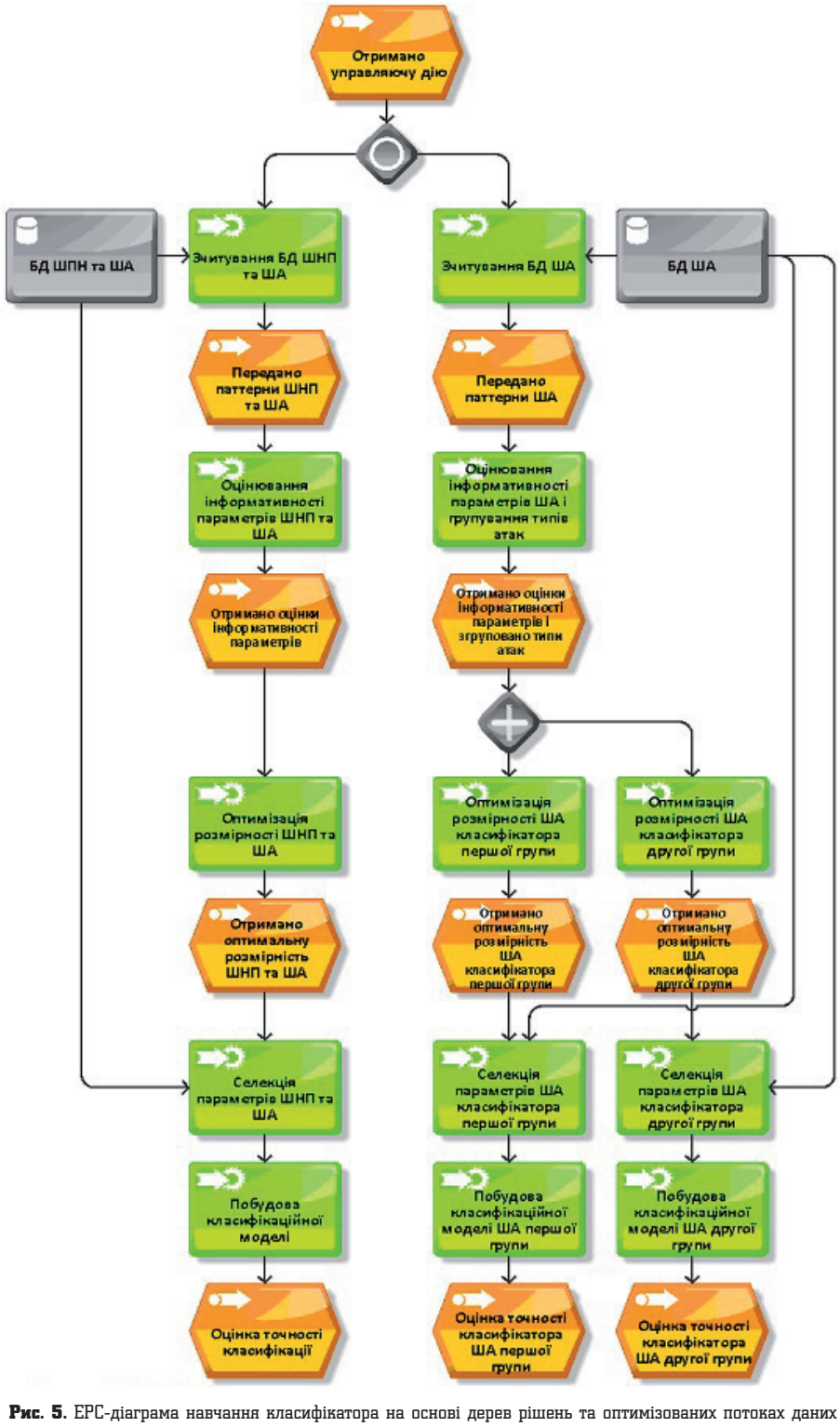

Рис. 5. EPC-діаграма навчання класифікатора на основі дерев рішень та оптимізованих потоках даних

BPMN-діаграму класифікатора на базі дерев рішень та оптимізованих ПВхД, подано на рис. 6. Згідно неї стартовою подією на етапі роботи класифікатора є отримання ПВхД з сенсорів СВА. Одночасно в модулі виявлення атак виконується оптимізація розмірності ПВхД, а у модулі класифікації атак запис потоку у буфер тимчасового зберігання Після цього в модулі виявлення атак виконується процес виявлення відхилень від ШНП. У разі відсутності зазначених відхилень генерується повідомлення про ідентифікацію нормального шаблону. У випадку виявлення аномальної активності в системі (відхилення від ШНП), керування переходить до модуля класифікації атак, при цьому виконується зчитування даних з буферу тимчасового зберігання ПВхД. На наступному етапі, з метою класифікації типу атаки, одночасно виконується оптимізація зчитаного ПВхД для подальшої бінарної класифікації для кожного з «дуетів» атак. Після чого отримані дані подаються для аналізу результатів класифікації. У разі успішної класифікації генерується повідомлення про виявлення певного типу атаки, у іншому випадку — повідомлення про необхідність експертного оцінювання патерну. Отже, спираючись на одержані результати метод побудови класифікатора кібератак на державні інформаційні ресурси на основі дерев прийняття рішень та оптимізованих ПВхД складається з кроків і полягає у такому.

На першому кроці методу виконується попередня обробка баз шаблонів поведінки системи — ШНА та ША атак для модуля виявлення та ША для модуля класифікації атак. Для зазначених баз шаблонів виконується оцінювання інформативності їх параметрів за критерієм *GainRatio* , а для бази ША формування «дуетів» типів атак, що забезпечують виконання вимоги максимальної віддаленості у просторі параметрів.

Другий крок полягає у оптимізації розмірності шаблонів поведінки ІТС. При цьому на основі результатів першого кроку методу розраховуються оптимальні розмірності шаблонів трьох груп — ШНП та ША для модуля виявлення та ША першої та другої груп для модуля класифікації атак. Виконується безпосереднє скорочення розмірностей патернів зазначених баз шаблонів.

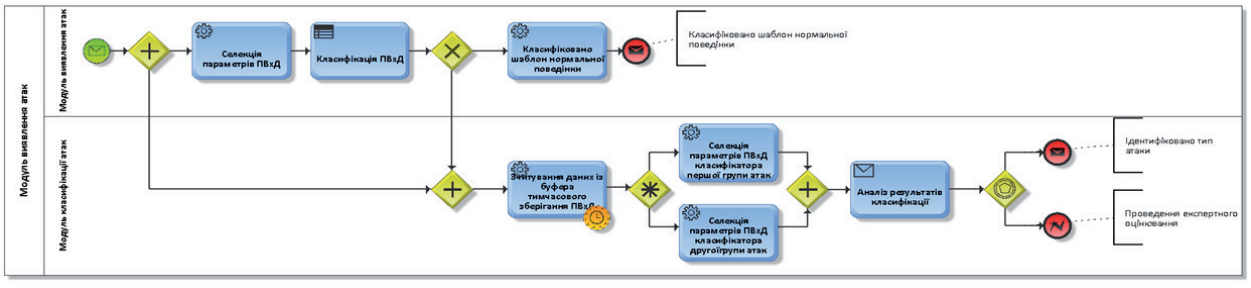

Рис. 6. BPMN-класифікатора на базі дерев рішень та оптимізованих потоків вхідних даних

Третій крок має на меті подову класифікаційних моделей на базі дерев рішень з використанням алгоритму C4.5. Навчання виконується з використання функції перехресної перевірки на шаблонах поведінки оптимізованої розмірності.

На четвертому кроці виконується оцінка точності класифікації.

П'ятий, заключний, крок методу має на меті оцінку адаптивності побудованого класифікатора. Виконання зазначеної операції ґрунтується на перевірці побудованого класифікатора на тестових наборах даних. У разі забезпечення висунутих розробником вимог побудова класифікатора вважається завершеною.

6. Обговорення результатів дослідження функціонування класифікатора

Результати оцінювання ефективності функціонування класифікатора згідно розробленого методу подано у табл. 1.

Результати оцінювання ефективності функціонування модуля виявлення для тестових даних бази NSL KDD (Test, Test^{-21}) наведено в табл. 2. Вибір тестових множин ПВхД бази NSL KDD обумовлений її доступністю, що дає можливість стверджувати про уніфікованість результатів дослідження з іншими авторами, роботи яких присвячені питанням оптимізації розмірності потоків вхідних даних класифікаторів. Застосування тестового набору даних бази NSL KDD для класифікації ШНП та ША обумовлено тим, що зазначена база містить тільки мітки класів нормальної поведінки та атак, без визначення кількості атак кожного з 4 типів.

Таблиця 1

Результати оцінювання ефективності

Шаблон поведінки	Ефективність функціонування					
	до застосування методу			після застосування методу		
	Час побудови моделі, с	Кількість параметрів потоку	Точність класифікації, %	Час побудови моделі, с	Кількість параметрів потоку	Точність класифікації, %
Normal_ Anomaly	75,71	41	99,9018	6,71	11	99,899
DoS_U2R	4,97	41	99,9927	1,09	15	99,9927
Probe_R2L	0,33	41	99,7443	0,06	21	99,7443

Таблиця 2

Результати оцінювання ефективності

Шаблон поведінки NSL KDD	Ефективність функціонування					
	до застосування методу			після застосування методу		
	Час побудови моделі, с	Кількість параметрів потоку	Точність класифікації, %	Час побудови моделі, с	Кількість параметрів потоку	Точність класифікації, %
Test	0,35	41	78,0873	0,18	11	79,52
Test^{-21}	0,19	41	58,3291	0,09	11	61,038

Таким чином, одержані результати оцінювання ефективності функціонування виявлювача/класифікатора побудованого згідно із розробленим методом доводять його працездатність та ефективність. Так, для методу перехресної перевірки час побудови класифікаційної моделі для «дуету» ШНП та ША скоротився у 11,28 разів при зниженні точності класифікації на 0,0028 %, для «дуетів» атак DoS та U2R, Probe та R2L точність класифікації не погіршилась, а час побудови класифікаційної моделі зменшився у 4,5 та 5,5 разів відповідно. При застосуванні тестових наборів даних бази NSL KDD час розпізнавання побудованою класифікаційною моделлю тестового набору NSL KDD Test скоротився у 1,94 рази, а точність розпізнавання ПВхД зросла на 1,43 %. Для тестового набору NSL KDD Test^{-21} час розпізнавання ПВхД скоротився у 2,11 рази, а точність розпізнавання зросла на 2,7 %. Отже, одержані згідно методу результати відповідають поставленому у [14] завданню.

7. Висновки

В результаті проведених досліджень:

1. Встановлено, що підвищення ефективності функціонування СВА досягається за рахунок інтелектуалізації процесу побудови класифікаторів кібератак та шляхом удосконалення відомих алгоритмів класифікації.

2. Показано, що перспективним напрямком наукових досліджень, який забезпечує ефективне виявлення кібеатак у масштабі часу наближеному до реального і позбавлений недоліків відомих підходів є метод класифікації на основі дерев рішень.

3. Набув подальшого розвитку принцип побудови класифікаторів СВА, у напрямі реалізації двоетапної бінарної схеми класифікації станів системи.

4. Набув подальшого розвитку метод побудови класифікаторів кібератак на державні інформаційні ресурси, в основу якого покладено дерева рішень. Метод відрізняється від відомих оптимізацією ПВхД на вході класифікаційної моделі, що забезпечує підвищення оперативності виявлення кібератак за заданих показників точності класифікації.

5. Отримані результати верифікації розробленого методу показали, що його застосування, у порівнянні з класичними, забезпечує зменшення часу побудови класифікаційних моделей для «дуетів» ШНП-ША у 11,28 разів, DoS-U2R та Probe-R2L у 4,5 та 5,5 рази відповідно.

Література

1. Бурячок, В. Політика інформаційної безпеки [Текст]: підручник / В. Л. Бурячок, Р. В. Грищук, В. О. Хорошко; під заг. ред. проф. В. О. Хорошка. — К.: ПВП «Задруга», 2014. — 222 с.
2. Bankovic, Z. A Genetic Algorithm-based Solution for Intrusion Detection [Text] / Z. Bankovic, J. Moya, Á. Araujo, S. Bojanic, O. Nieto-Taladriz // Journal of Information Assurance and Security. — 2009. — V. 4. — P. 192–199.
3. Mukkamala, S. Intrusion Detection Using Neural Networks and Support Vector Machines [Text] / S. Mukkamala, G. Janoski, A. Sung // Proceedings of IEEE Intrnational Joint Conference on Neural Networks. — 2002. — P. 1702–1721. doi:10.1109/ijcnn.2002.1007774
4. Farid, D. M. Anomaly Network Intrusion Detection Based on Improved Self Adaptive Bayesian Algorithm [Text] / D. M. Farid, M. Z. Rahman // Journal of Computers. — 2010. — Vol. 5, № 1. — P. 23–31.doi:10.4304/jcp.5.1.23-31
5. Wee, Y. Y. Causal Discovery and Reasoning for Intrusion Detection using Bayesian Network [Text] / Y. Y. Wee, W. P. Cheah, S. C. Tan, K. Wee // International Journal of Machine Learning and Computing. — 2011. — Vol. 1, № 2. — P. 185–192. doi:10.7763/ijmlc.2011.v1.27
6. Chou, T. Cyber Security Threats Detection Using Ensemble Architecture [Text] / T. Chou // International Journal of Security and Its Applications. — 2011. — Vol. 5, № 2. — P. 17–32. doi:10.14257/ijsia
7. Лукацкий, А. Обнаружение атак [Текст] / А. Лукацкий. — СПб.: БХВ-Петербург, 2001. — 624 с.
8. Комар, М. Метод построения совокупного классификатора трафика информационно-телекоммуникационных сетей для иерархической классификации компьютерных атак [Текст] / М. Комар // Системи обробки інформації. — 2012. — Том 1, Вип. 3(101). — С. 134–138.
9. Panda, M. Ensemble of classifiers for detecting network intrusion [Text] / M. Panda, M. R. Patra // International Conference on Advances in Computing, Communication and Control archive. — 2009. — P. 510–515. doi:10.1145/1523103.1523204
10. Ходашинский, И. Выявление вредоносного сетевого трафика на основе ансамблей деревьев решений [Текст] / И. А. Ходашинский, В. А. Дель, А. Е. Анфилофьев // Доклады ТУСУРа. — 2014. — № 2(32). — С. 202–206.
11. Грищук, Р. Метод оцінювання інформативності параметрів потоку вхідних даних для мережевих систем виявлення атак [Текст] / Р. Грищук, В. Мамарєв // Системи обробки інформації. — 2012. — Том 1, № 4(102). — С. 103–107.
12. Грищук, Р. Метод оптимізації розмірності потоку вхідних даних для систем захисту інформації [Текст] / Р. Грищук, В. Мамарєв // Інформаційна безпека. — 2012. — № 2(8). — С. 27–34.
13. Пількевич, І. Основи побудови автоматизованих систем управління [Текст]: навч. посібник / І. А. Пількевич, К. В. Молодецька, І. І. Сугоняк, Н. М. Лобанчикова. — Житомир: Вид-во ЖДУ ім. І. Франка, 2014. — 226 с.
14. Грищук, Р. Постановка задачі розробки методики скорочення розмірності потоку вхідних даних для мережних систем виявлення атак [Текст] / Р. Грищук, В. Мамарєв // Інформаційна безпека. — 2011. — № 1(5). — С. 74–78.

МЕТОД ПОСТРОЕНИЯ КЛАССИФИКАТОРА КИБЕРАТАК НА ГОСУДАРСТВЕННЫЕ ИНФОРМАЦИОННЫЕ РЕСУРСЫ

Предложено двухэтапную схему классификации состояния информационно-телекоммуникационной системы, основанную на бинарном группировании шаблонов её поведения. Разработан метод построения классификатора кибератак на базе деревьев решений и оптимизированных потоках исходных данных. Применение метода позволяет в разы сократить время построения и функционирования классификационной модели, оставляя точности классификации шаблонов поведения системы в заданных диапазонах.

Ключевые слова: информационно-телекоммуникационная система, кибератака, классификатор, классификация, деревья решений, оптимизация.

Бурячок Володимир Леонідович, *доктор технічних наук, старший науковий співробітник, завідувач кафедри інформаційної та кібернетичної безпеки, Державний університет телекомунікацій, Київ, Україна.*

Грищук Руслан Валентинович, *доктор технічних наук, старший науковий співробітник, провідний науковий співробітник науково-дослідної лабораторії, Науковий центр, Житомирський військовий інститут ім. С. П. Корольова Державного університету телекомунікацій, Україна.*

Мамарєв Віктор Миколайович, *аспірант, кафедра інформаційної та кібернетичної безпеки, Державний університет телекомунікацій, Київ, Україна.*

Бурячок Владимир Леонидович, *доктор технических наук, старший научный сотрудник, заведующий кафедрой информационной и кибернетической безопасности, Государственный университет телекоммуникаций, Киев, Украина.*

Грищук Руслан Валентинович, *доктор технических наук, старший научный сотрудник, ведущий научный сотрудник научно-исследовательской лаборатории, Научный центр, Житомирский военный институт им. С. П. Королева Государственного университета телекоммуникаций, Украина.*

Мамарев Виктор Николаевич, *аспирант, кафедра информационной и кибернетической безопасности, Государственный университет телекоммуникаций, Киев, Украина.*

Burachok Volodymyr, *State University of Telecommunications, Kyiv, Ukraine.*

Hryshchuk Ruslan, *Zhytomyr Military Institute of the State University of Telecommunications, Ukraine.*

Mamarev Viktor, *State University of Telecommunications, Kyiv, Ukraine*