УДК 004.056
DOI: 10.31673/2409-7292.2020.025258

Киричок Р.В., Шуклін Г.В., Бржевська З.М.

МЕТОД КОНТРОЛЮ ПОСЛІДОВНОСТІ РЕАЛІЗАЦІЇ АТАКУЮЧИХ ДІЙ ПІД ЧАС АКТИВНОГО АНАЛІЗУ ЗАХИЩЕНОСТІ КОРПОРАТИВНИХ МЕРЕЖ

У статті запропоновано підхід щодо підвищення ефективності валідації вразливостей під час автоматичного активного аналізу захищеності корпоративних мереж на основі контролю послідовності реалізації атакуючих дій (експлойтів) згідно стратегії вибору дій softmax з використанням ймовірнісного розподілу Гіббса. При цьому, на основі практичного аналізу процесу валідації вразливостей, було введено коефіцієнт хибних рішень щодо реалізації експлойта, який дозволяє динамічно змінювати ключовий параметр з розподілу Гіббса – температуру, що в свою чергу призводить до врівноваження ймовірності вибору наступної атакуючої дії, яка може бути значно ефективнішою при реалізації валідації виявлених вразливостей конкретної цільової системи.

Ключові слова: активний аналіз захищеності, корпоративна мережа, цільова система, навчання з підкріпленням, стратегія вибору дій, валідація вразливостей, експлойт.

Вступ

В сучасному інформатизованому світі, зі зростанням кількості вразливостей комп'ютерних систем та мереж [1], спрощенням проведення кібератак з використанням готових експлойтів даних вразливостей та їх доступності, виникає цілком обґрунтована необхідність у використанні превентивних механізмів захисту інформації. Основна ціль даних механізмів, серед яких найпоширенішим є активний аналіз захищеності, полягає в попередженні самої можливості реалізації зловмисником кібератаки на інформаційну інфраструктуру, що досягається шляхом своєчасного (мається на увазі, раніше зловмисників) виявлення, підтвердження та закриття слабких місць і вразливостей як в самих інформаційно-комунікаційних системах, так і в системах їх захисту. При цьому, активний аналіз захищеності, базуючись на різних методах та методиках проведення тестування на проникнення, дозволяє виконувати контрольовані атаки на інформаційні системи та мережі визначаючи всі ймовірні атакуючі дії зловмисників та встановлювати фактичний стан захищеності. Однак, разом з тим, виявлення та підтвердження можливості реалізації вразливостей, що є основними процедурами активного аналізу захищеності, вимагають значних часових затрат та високої кваліфікації спеціалістів по кібербезпеці, оскільки перевірка виявлених вразливостей здійснюється здебільшого вручну.

Таким чином, питання підвищення ефективності активного аналізу захищеності, особливо великих мереж, таких як корпоративні, де кількість вразливостей злічується тисячами, є досить актуальним.

Аналіз публікацій

На сьогодні, одним із підходів щодо вирішення вищезазначеної проблеми є застосування штучного інтелекту для автоматизації процесу активного аналізу захищеності, зокрема процедури валідації виявлених вразливостей. Під валідацією вразливостей розуміється процедура підтвердження можливості реалізації виявлених вразливостей цільових інформаційних систем шляхом їх експлуатації за допомогою спеціалізованих програмних засобів, за часту, з використанням вже готових експлойтів даних вразливостей (шкідливих скриптів, виконуваних модулів та ін.).

Поточні методи автоматичного активного аналізу захищеності використовують різну математичну базу, однак більшість ґрунтується частково спостережуваних марківських процесах прийняття рішень та звичайних марківских процесах прийняття рішень [5-7].

Використання частково спостережуваних марківських процесів прийняття рішень (POMDP) при моделюванні кібератак на інформаційні системи вводить в симуляцію так звані неповні знання зловмисників. Це дозволило імітувати усунення припущення того, що відомо структуру мережі та конфігурації кожного окремого хоста, замість чого здійснюється моделювання спостережень за конфігураціями по мірі розвитку атаки. Однак, разом з тим,

аналіз останніх публікацій [2-4] показав, що даний підхід добре спрацьовує на практиці при проведенні активного аналізу захищеності в невеликих мережах, оскільки обчислювальні властивості POMDP не дозволяють здійснювати значного масштабування зі зростанням розміру простору станів.

Рішенням даної проблеми стало моделювання активного аналізу захищеності з використанням звичайних марківських процесів прийняття рішень (MDP). В такому підході ігнорується невизначеність щодо стану конфігурації цільових систем, а замість цього вноситься невизначеність в ймовірність успішної реалізації кожної можливої кібератаки. Це зробило даний підхід більш реалістичним в обчислювальному плані, ніж при використанні POMDP, а також позбавило необхідності повного знання конфігурації мережі та її хостів. Натомість, даний тип моделей потребує попередні знання щодо можливості успішної реалізації кожної потенційної атакуючої дії, а цільові системи розглядаються як ідентичні одна одній замість того, щоб використовувати зібрану інформацію щодо заданого хоста задля формування більш спеціалізованих кібератак.

Ще одним, більш загальним, підходом до вирішення питання автоматизації активного аналізу захищеності інформаційних систем та мереж з використанням MDP, коли модель переходу невідома, є навчання з підкріпленням (RL) [8]. RL вимагає представлення простору станів, набору дій, які можуть бути виконані агентом RL та функції винагороди, яка визначає ціль агента (тобто, чого намагається він досягти). Даний підхід передбачає взаємодію з середовищем, на основі чого агент RL вивчає політику дій, які слід вжити в будь-якому заданому стані задля оптимізації продуктивності та підвищення ефективності активного аналізу захищеності корпоративних мереж. Однак, через складну та мінливу природу середовища – в даному випадку, цільова корпоративна мережа, з постійним оновленням програмного забезпечення та появою нових вразливостей, а також беручи до уваги доступність експлойтів вразливостей, підтримка актуальної моделі результатів виконання будь-яких дій є досить важкою задачею.

Постановка задачі

Метою даного дослідження є створення ефективного методу контролю послідовності реалізації атакуючих дій під час проведення активного аналізу захищеності корпоративних мереж на підставі введення коефіцієнту хибних рішень щодо реалізації експлойта, який частково визначає якість процесу валідації виявлених вразливостей. При цьому, основними задачами є аналіз найрозповсюдженіших стратегій дослідження середовища та удосконалення механізму вибору наступних дій агентом RL під час валідації виявлених вразливостей всіх хостів цільової корпоративної мережі.

Основна частина

1. Постановка задачі навчання з підкріпленням

Загалом, під задачею RL розуміється задача досягнення певної мети шляхом навчання через взаємодію з цільовим об’єктом. Виходячи з цього, одиниця яка приймає рішення та навчається називається агентом, а об’єкт з яким агент взаємодіє – середовищем. При цьому, відбувається постійна взаємодія між агентом та середовищем, агент обирає та виконує дію, а середовище відповідає на неї породжуючи підкріплення, спеціальне числове значення, а також надає новий стан агенту. Ціль агента полягає в максимізації числового підкріплення за час своєї роботи. Загальну схему процесу навчання з підкріпленням наведено на рис. 1.

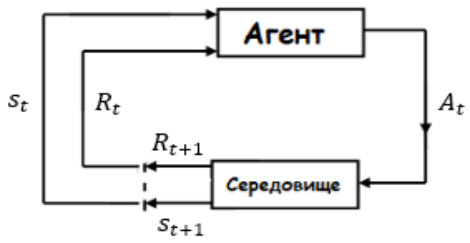

Рис. 1. Узагальнена схема процесу навчання з підкріпленням [8]

Формально, взаємодія між агентом та середовищем відбувається на кожному кроці послідовності дискретних моментів часу $t = 0,1,2...T$, на кожному з яких агент отримує певний стан середовища $S_t \in S$, де S – скінченна множина всіх можливих станів. Після чого, на основі отриманого стану, агент обирає дію $A_t \in A(S_t)$, де $A(S_t)$ – скінченна множина дій, доступних агенту в стані S_t. В результаті виконаної дії, на наступному кроці, за допомогою оціночного зворотного зв'язку, агент отримує числове підкріплення $R_{t+1} \in R \subset \square$, а також новий стан S_{t+1}. Слід відзначити, що числове підкріплення може бути як позитивним, $R_t > 0$ (винагорода), так і негативним, $R_t < 0$ (покарання), на основі чого в агента є можливість формування певного уявлення щодо оптимальності зробленого ним вибору. Окрім цього, виконання певної дії на даному кроці може безпосередньо впливати на значення одержуваного підкріплення не лише для поточного рішення, але і для всіх інших значень підкріплення. Саме тому, агент на кожному кроці зіставляє ймовірності вибору всіх доступних дій конкретному стану. Це зіставлення ще називається політикою агента π_t, де $\pi(s,a)$ описує ймовірність того, що буде обрано дію $A_t = a$ в стані $S_t = s$.

З вищезазначеного випливає, що саме політика визначає поведінку агента в даний момент часу та відповідає за вибір оптимальної дії у відповідності до поточного стану, а ціллю агента в такому разі є підбір політики таким чином, щоб максимізувати сумарну очікувану винагороду $R \to \max$, яку отримує агент за час його роботи приймаючи оптимальні дії. Так, на *t*-му кроці сумарна величина підкріплення визначається за наступним виразом:

$$R_t = \sum_{k=0}^{T} r_{t+1+k} \tag{1}$$

де T відображає останній часовий крок, що символізує завершення епізоду при досягненні певного термінального (завершального) стану.

У випадку коли взаємодію агента з середовищем не можливо явно розбити на епізоди, задачу називають тривалою (не епізодичною), а використання виразу (1) є неможливим, оскільки відсутній термінальний стан. За таких умов, очікувана винагорода, яку агент намагається максимізувати, може досягати безкінечності і щоб цього уникнути вводиться фактор дисконтування (знецінення) винагороди:

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+1+k} \tag{2}$$

де γ – коефіцієнт знецінення, $0 \le \gamma \le 1$, який задає цінність винагороди в майбутньому, тобто чим менший коефіцієнт, тим менше агент замислюється над вигодою від майбутніх дій [66].

2. Функція цінності дії

Оскільки заздалегідь точних значень цінності дій невідомо, інакше вирішення задачі навчання з підкріпленням було б тривіальним та зводилося до обрання дії з найбільшою цінністю, переважна кількість алгоритмів RL включають поняття функції цінності дій. Дана функція являється функцією від пари стану і дії, яка визначає, наскільки для агента цінно застосувати дану дію в даному стані. Саме поняття того, наскільки цінною є якась дія виражається в понятті майбутньої очікуваної винагороди, тобто нагороди на яку агент може розраховувати отримати в майбутньому.

Таким чином, цінність дії a в стані s при дотриманні політики π, що позначається $Q_\pi(s,a)$, визначається як очікуване значення винагороди починаючи з стану s, виконуючи дію a та дотримуючись надалі політики π:

$$Q_\pi(s,a) = E_\pi\left[\sum_{k=0}^{\infty}\gamma^k r_{t+1+k} \,|S_t = s, A_t = a\right] \tag{3}$$

де Q_π – функція цінності дії для політики π.

Оптимальною політикою вважається така політика π для якої очікувана винагорода більша ніж для політики π' для всіх станів, та позначається π_*. Звідси, оптимальна функція цінності дії:

$$Q_*(s,a) = \max_\pi Q_\pi(s,a), \forall s \in S, a \in A(s) \tag{4}$$

Основною перевагою RL є можливість навчання агента не маючи жодних вхідних даних, попередніх знань або моделі щодо ймовірностей результатів дій для будь-якого стану, більш того, часто передбачається, що він навіть не має жодного початкового уявлення про властивості середовища з яким взаємодіє. Натомість, агент має можливість самостійно приймати рішення щодо вибору дії, шляхом спроб та помилок, поступово покращуючи свої знання про середовище з яким взаємодіє. Виходячи з цього, одним із ключових компонентів будь-якого алгоритму RL можна вважати стратегію вибору дій.

3. Стратегії вибору дій

На сьогодні, існує цілий ряд різноманітних доступних стратегій вибору дій, однак найбільш часто використовуваними залишаються дві стратегії, – це ε-жадібна стратегія та стратегія Softmax [9]. Обидві стратегії спрямовані на досягнення балансу між розвідкою та експлуатацією, тому їх ще називають стратегіями дослідження середовища.

a. ε-жадібна стратегія

Згідно з даною стратегією, дія, з певною ймовірністю ε, незалежно від очікуваної цінності, обирається рівномірно випадково. В решті випадків, дія обирається з максимальною оцінкою значення її цінності, що описується наступним чином:

$$\pi_\varepsilon(s,a) = \begin{cases} random\ a, & з\ ймовірністю\ p = \varepsilon \\ \arg\max_a Q(s,a), & з\ ймовірністю\ p = 1-\varepsilon \end{cases} \tag{5}$$

При цьому коефіцієнт ε є регульованим параметром, $0 < \varepsilon < 1$, який, здебільшого, на початку навчального процесу задається великим, заохочуючи агента до дослідження середовища. Після чого поступово значення коефіцієнта, по мірі навчання, зменшується до маленького, близько $0{,}1$, тим самим зменшуючи долю випадкової поведінки агента. Основною перевагою даного методу є те, що при достатньо великій кількості часових кроків, кожна дія буде випробувана, тобто досліджена, тим самим гарантуючи оптимальність вибору наступних дій.

b. Стратегія Softmax

Іншим підходом до вибору дій є стратегія Softmax (метод зваженої дії), яка вирівнює ймовірності дій, в залежності від їхньої ефективності, тобто ймовірність вибору ε-жадібної дії залишається максимальною, однак, всі інші дії проходять ранжування у відповідності до їх передбачуваної цінності за рахунок вагового коефіцієнта.

Найчастіше даний метод здійснює оцінку ймовірність вибору наступної дії a на t-му часовому кроці, використовуючи розподіл Больцмана або, як в даному випадку, розподіл Гіббса:

$$\Pr\{A_t = a\} = \pi_t(s,a) = \frac{e^{\frac{Q_t(s,a)}{\tau}}}{\sum_{b=1}^{n} e^{\frac{Q_t(s,b)}{\tau}}} \quad (6)$$

де: $Q_t(s,b)$ – очікуване значення цінності за вибір іншої дії *b* на *t*-й грі;

τ – температура, додатній параметр, який дозволяє налаштовувати поведінку алгоритму. Так, при високих значеннях температури ймовірності вибору дій врівноважується, в той час як при низьких температурах виникає більш значне розходження в ймовірностях вибору дій, що мають різне очікуване значення винагороди.

Тобто, при $\tau \to 0$, алгоритм стає аналогічним жадібному вибору дії (вибору дії з максимальною оцінкою значення цінності), а при $\tau \to \infty$, алгоритм обирає випадкові дії. Також, слід відзначити, що експонента використовується для того, щоб уникнути нульової ваги дії, навіть втому випадку, коли очікуване значення винагороди за вибір даної дії поки ще нульове.

Результати дослідження

З представленого вище опису стратегій вибору дій слідує, що ε-жадібна стратегія хоча і є досить ефективною в плані балансу між дослідженням та експлуатацією, однак все ж таки має очевидний недолік, який полягає в рівноймовірному виборі серед всіх дій під час дослідження. Мається на увазі, що з однаковою ймовірністю може бути обрана дія як з найнижчою, так і з найвищою цінністю. Це є особливо чуттєвим в задачах, де найгірші дії мають надзвичайно низьку винагороду, зокрема при проведенні валідації вразливостей такі дії можуть призводити до критичної помилки в функціонуванні цільової системи та навіть повного виведення її з ладу.

Саме тому, більш перспективним, задля інтелектуалізації процесу валідації виявлених вразливостей при проведенні автоматичного аналізу захищеності корпоративних мереж є використання методу зваженої дії на основі імовірнісного розподілу Гіббса, що описується формулою (6).

В контексті поставленої задачі інтелектуальної валідації вразливостей, параметр температури визначає наскільки висока ймовірність вибору наступної атакуючої дії з максимальним значенням очікуваного підкріплення. Виходячи з цього, а також беручи до уваги ряд проведених спостережень функціонування засобів експлуатації виявлених вразливостей, пропонується динамічно змінювати, під час безпосередньої валідації вразливостей, параметр τ в залежності від коефіцієнту хибних рішень щодо реалізації експлойта, при цьому початкове значення температури визначено експериментально та встановлено за замовчуванням, а саме:

$$\begin{cases} \tau(E_{fd}) = 5, & при\ E_{fd} \le 1 \\ \tau(E_{fd}) = 5 \cdot k \cdot (E_{fd} - 1) + 5, & при\ E_{fd} > 1 \end{cases} \quad (7)$$

де: $E_{fd} = \frac{N_{fed}}{N_{ted}} \cdot 100$ – коефіцієнт хибних рішень щодо реалізації експлойта, заданий в діапазоні $[0,100]$;

N_{fed} – кількість хибних рішень щодо реалізації експлойта;

N_{ted} – загальна кількість рішень щодо реалізації експлойта;

τ – параметр температури з розподілу Гіббса;

k – коефіцієнт росту, пропонується встановити на значенні $0,5$, за замовчуванням.

Таким чином, чим вище значення температури, тим більша ймовірність вибору альтернативних атакуючих дій, що робить можливим дослідження середовища на наявність інших, значно ефективніших атакуючих дій при реалізації валідації вразливостей конкретної цільової системи. Однак, у випадку, коли якість валідації поліпшується, то інтуїтивно зрозуміло, що використовувані при цьому експлойти повинні обиратися як можна частіше в якості першочергових атакуючих дій, що реалізується за допомогою зменшення параметру τ .

Як приклад, було здійснено імітацію застосування даного підходу щодо контролю послідовності реалізації атакуючих дій, з 5-и доступних, в ході валідації виявлених вразливостей. При цьому, для кожного з експлойтів було встановлено наступні початкові значення цінності дій:

$$Q_1(a) = \{a_1 : 3.5,\ a_2 : 1.75,\ a_3 : 1.4,\ a_4 : 2.33,\ a_5 : 7.0\}$$

Після чого, застосувавши імовірнісний розподіл Гіббса (6) та динамічну зміну значення температури згідно (7), були отримані відповідні ймовірності вибору першочергової реалізації даних експлойтів вразливостей, результати представлено на рис. 2.

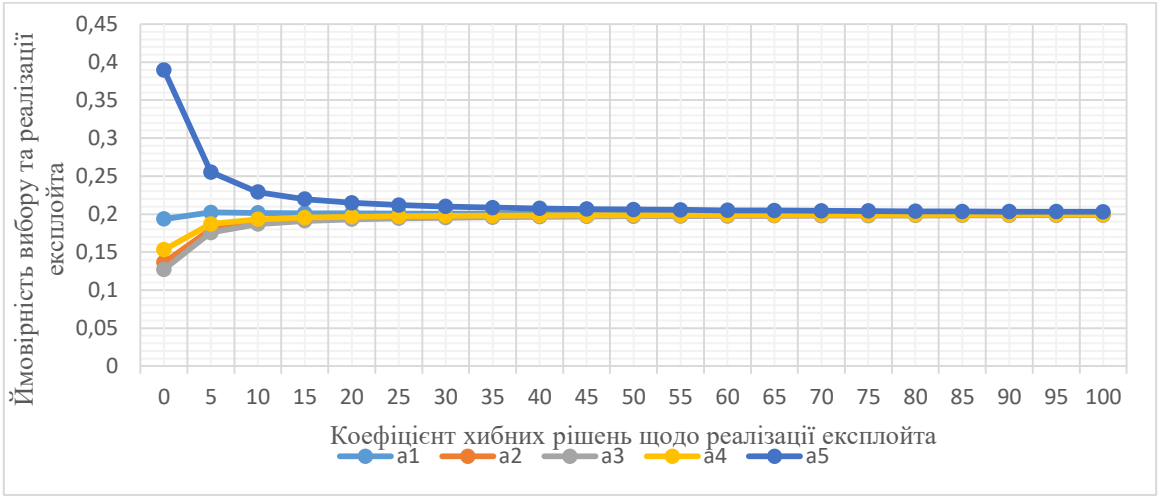

Рис. 2. Розподіл ймовірностей вибору атакуючих дій 1-5 в залежності від E_{fd}

З рисунку 2 видно, що вже при значенні 35 коефіцієнта хибних рішень щодо реалізації експлойтів, ймовірність вибору та реалізації інших атакуючих дій практично врівноважується. При цьому, регулювання початкового значення параметру температури дозволяє змінювати поріг врівноваження ймовірностей.

Висновки

Таким чином, в роботі було розроблено метод контролю послідовності реалізації атакуючих дій під час активного аналізу захищеності, що ґрунтується на стратегії вибору дій softmax з використанням ймовірнісного розподілу Гіббса. При цьому, було введено коефіцієнт хибних рішень щодо реалізації експлойта, що дозволяє динамічно змінювати ключовий параметр даного розподілу – температуру, що в свою чергу допомагає підтримувати баланс між дослідженням та експлуатацією, в залежності від поточних умов, таких як: спрацювання системи захисту на цільовій системі, втрата з нею зв'язку та інше.

Перелік посилань

1. CVSS Severity Distribution Over Time [Електронний ресурс] // National Vulnerability Database – Режим доступу до ресурсу: https://nvd.nist.gov/vuln-metrics/visualizations/cvss-severity-distribution-over-time (03.08.20).

2. Sarraute C. Penetration testing == POMDP solving? / C.Sarraute, O.Buffet, J.Hoffmann. // arXiv. – 2013. - arXiv:1306.4714.

3. Sarraute C. POMDPs make better hackers: Accounting for uncertainty in penetration testing. / C.Sarraute, O.Buffet, J.Hoffmann // In Proceedings of the 26th AAAI Conference on Artificial Intelligence «AAAI'12». Toronto, ON, Canada, July 2012. AAAI Press. - pp. 1816-1824.

4. Shmaryahu D. Partially observable contingent planning for penetration testing / D.Shmaryahu, G.Shani, J.Hoffmann // 2017 1st Int Workshop on Artificial Intelligence in Security. – 2017. – pp. 33-40.

5. Stefinko Ya. Theory of modern penetration testing expert system. / Ya.Ya.Stefinko, A.Z.Piskozub // Information Processing Systems, -2017. - Vol. 2(148), - pp. 129-133.

6. Durkota K. Computing optimal policies for attack graphs with action failures and costs. / K.Durkota, V.Lisy. // In 7th European Starting AI Researchers` Symposium «STAIRS'14». January 2014.

7. Zhou T. NIG-AP: a new method for automated penetration testing. / T.Zhou, Y.Zang, J.Zhu, et al. // Frontiers Inf Technol Electronic Eng 20, - 2019. – pp. 1277–1288.

8. Sutton R.S. Reinforcement Learning: An Introduction second edition. / R.S. Sutton, A.G. Barto // The MIT Press, Cambridge, MA, 2018. - 445 P.

9. McFarlane R. A survey of exploration strategies in reinforcement learning. [Електронний ресурс] / R. McFarlane // McGill University – Режим доступу до ресурсу: http://www.cs.mcgill.ca/~cs526/roger.pdf (03.08.20).

Надійшла: 12.04.2020 Рецензент: д.т.н., професор Вишнівський В.В.