УДК 004.056
DOI: 10.31673/2409-7292.2020.034953

Коростель В.С., Кожухівський А.Д., Луценко І.М., Кітура О.В., Маслова Ю.Ю.

ПРОГНОЗУВАННЯ ЧАСУ ЗДІЙСНЕННЯ КІБЕРАТАКИ НА ОСНОВІ РЕЗУЛЬТАТІВ АНАЛІЗУ НЕСТАЦІОНАРНИХ ПРОЦЕСІВ

Можливість прогнозувати кібератаки до того, як вони відбудуться, безсумнівно, змінить перебіг кібервійн та кіберзлочинності. Проблема прогнозування кібератак полягає у отриманні відповідних та надійних сигналів для протидії кіберзловмисникам. Стаття ґрунтується на методах машинного навчання та розробляє інтегровану систему для трансформації великих обсягів загальнодоступних даних для передбачення кіберінцидентів

Ключові слова: Кібербезпека, прогнозування, нетрадиційні сигнали.

Вступ

За останні роки масштаби та різноманітність кібератак суттєво змінилися, ставши засобом для отримання грошової вигоди, інтелектуальних крадіжок та зміни політичної ситуації в усьому світі. Останні звіти показують, що кількість кібератак продовжує збільшуватись у всьому світі [1], а втрати для суспільства через ці атаки суттєво зростають [2]. Прогнозування кібератак до того, як вони відбудуться, може мати велике значення, але є складним через обмежені можливості методів та недостатність інформації, яку можна знайти у соціальних мережах, новинах та інших публічних форумах. Ця стаття вирішує цю проблему шляхом розробки інтегрованої системи, яка розглядає як звичайні статистичні вибірки, так і неповні та незбалансовані дані.

Ключовою проблемою прогнозування кібератак за неповними даними є те, що не всі дані можуть вказувати на атаку і це додає потенційних помилок у прогнозі. Для роботи з даними, які мають пропуски, необхідний інтелектуальний метод їх введення. Крім того, ці дані можуть мати різну завчасність, тобто час, що пройшов між спостережуваними публічними даними та кіберінцидентом. Спосіб фіксувати різноманітні періоди завчасності є нетривіальним і вимагає розробки нових методів. Успішні кіберінциденти, як очікується, будуть рідкісними подіями для достатньо захищеної організації, що призводить до незбалансованості даних. Незбалансовані дані можуть призвести до упереджених або неточних моделей, коли прогнозований вплив нестандартних даних не фіксується. У цій статті розроблено інтегрований підхід, який охоплює нові та існуючі рішення для ряду дослідницьких завдань, відомих у сфері машинного навчання.

Аналіз робіт попередників

Оскільки ризики кібератаки продовжують зростати, необхідні дослідження та розробки для прогнозування атак замість пасивного виявлення вторгнень. В останні роки дослідники почали використовувати прогностичну аналітику, яка допомагає прогнозувати майбутні кіберінциденти перш ніж вони трапляться. У [3] досліджуються звіти, зібрані від антивірусних агентів McAfee, розміщених на більш ніж 85 000 хостів у багатонаціональному підприємстві. Використовуючи логістичну регресію для прогнозування ризику хостів зіткнутися зі шкідливим програмним забезпеченням, вони виявляють, що хости високого рівня стикалися зі шкідливим програмним забезпеченням у 3 рази більше порівняно з дрібними хостами. У [4] аналізуються 258 зовнішніх вимірюваних параметрів мережі організації, які базуються на неправильно налаштованому DNS (або BGP) у мережі та часових рядах шкідливої діяльності для спаму, фішингу та сканування. Використовуючи класифікатор випадкових дерев на основі повідомлень про кіберінциденти у середовищі VERIS, Hackmageddon та Інтернет-базі даних про випадки хакерської атаки, вони досягають 90% точності прогнозування порушень щодо окремої організації. У [5] використовують журнали зовнішнього вигляду файлів та марковані дані з антивірусів та засобів захисту від вторгнень антивірусної компанії, щоб передбачити, які машини мають високий ризик зараження. За допомогою класифікатора на основі випадкових дерев та підходу щодо

навчання під наглядом вони досягають високої точності (істинний та хибнопозитивний показник 96% та 5% відповідно) у прогнозуванні ризиків зараження для власників.

У той же час, використання неповних чи незбалансованих даних призводить до суттєвих помилок при прогнозуванні. Існують різні підходи до боротьби з проблемою відсутніх (неповних) даних. Одне з них - ігнорувати відсутні значення вибірки; однак тоді кількість навчальних випадків може зменшитися, що призведе до поганої роботи методів. Альтернативним простим підходом є заповнення відсутніх значень середнім значенням існуючих невідсутніх значень. У той же час, при великій кількості осереднених даних якість прогнозування суттєво знижується.

Сучасні методи боротьби з незбалансованими даними включають методи вирівнювання даних, які модифікують розподіл даних або шляхом видалення більшості екземплярів, або через додавання екземплярів меншості, після чого налаштовують існуючі алгоритми навчання на зменшення упередженості до більшості випадків.

Таким чином, для вирішення проблеми прогнозування атак на підставі часових вибірок з неповними чи незбалансованими даними необхідно розробити інший метод, позбавлений зазначених вище недоліків.

Викладення основного матеріалу.

Для одержання універсального методу вирішення задачі доцільним вбачається підхід, заснований на поданні досліджуваного часового ряду виразом [6, 7]

$$X(t) = m(t) + \sum_{\nu} V_\nu \varphi_\nu(t), \tag{1}$$

де $m(t)$ – математичне очікування процесу;

$\varphi_\nu(t)$ – невипадкові (координатні) функції часу;

V_ν – випадкові некорельовані між собою коефіцієнти ($M[V_\nu] = 0, M[V_\nu, V_\mu] = 0, \nu \neq \mu$).

Таке подання дозволяє застосовувати його для будь-якого набору даних у тому числі даних з пропусками та незбалансованих за часом. При цьому розглядається деякий скалярний випадковий процес $X(t)$ заданий випадковою послідовністю $X(t_i) = X(i), i = \overline{1, I}$ на дискретному ряді спостережень t_i.

З самого початку такий процес може бути представлено у вигляді

$$X(i) = m(i) + \sum_{\nu=1}^{i} V_\nu \varphi_\nu(i), i = \overline{1, I}, \tag{2}$$

де V_ν – випадковий коефіцієнт з характеристиками $M[V_\nu] = 0, M[V_\nu, V_\mu] = 0, \nu \neq \mu$; $M[V_\nu^2] = D_\nu$;

$\varphi_\nu(i)$ – невипадкова координатна функція, $\varphi_\nu(\nu) = 1$, $\varphi_\nu(i) = 0$ при $\nu > i$.

Визначення оптимальних значень прогнозу на наступних кроках здійснюється на основі рекурентних співвідношень:

$$V_1 = \overset{\circ}{X}(1),\ V_i = \overset{\circ}{X}(i) - \sum_{\nu=1}^{i-1} V_\nu \varphi_\nu(i),\ i = \overline{2, I}; \tag{3}$$

$$D_1 = D(1),\ D_i = D(i) - \sum_{\nu=1}^{i-1} D_\nu \varphi_\nu^2(i),\ i = \overline{2, I}; \tag{4}$$

$$\varphi_\nu(i) = \frac{1}{D_\nu} M\left[V_\nu \overset{\circ}{X}(i) \right], \nu = \overline{1, I}, i = \overline{\nu, I}. \tag{5}$$

Формула (5) показує, що основним обмеженням, яке накладається на досліджуваний випадковий процес, є скінченність його дисперсії. При дослідженні кібератак це, як правило, виконується, що і забезпечує універсальність методу. Отже, подання (2) здатне забезпечити вирішення задачі прогнозу кібератаки.

Алгоритм роботи методу

Для одержання аналітичного апостеріорного випадкового процесу на базі вибірки випадкових даних необхідно визначити процес $X(t)$ у вигляді (2) на дискретному ряді точок $t_i, i=\overline{1,I}$. Передбачається, що у деякі моменти $t_\mu, \mu=\overline{1,k}, k<I$, які співпадають t_i для $i\le k$, стали відомі значення $x(\mu), \mu=\overline{1,k}$ реалізації процесу $X(t)$. Необхідно одержати аналітичний опис прогнозу $X^{ps}(t)$, який виникає з апріорного $X(t)$ з урахуванням даних спостережень. Визначення прогнозованих значень здійснюється за наступним алгоритмом:

1. Визначення значення $x(1)$ реалізації процесу, одержаної у результаті спостережень. Для цього значення є справедливим (2), яке при $\mu=1$ приводиться до

$$x(1)=m(1)+v_1. \tag{6}$$

Формула (6) конкретизує значення v_1 випадкового коефіцієнта V_1, яке відповідає результату першого спостереження.

2. Визначення коефіцієнтів $V_i, i=\overline{1,I}$ подання (2), які дають змогу конкретизувати значення V_1, що приводить до зміни щільності розподілу решти коефіцієнтів $V_i, i=\overline{2,I}$. Для одержання прогнозу, припустимо, що коефіцієнти вихідного подання (2) попарно незалежні:

$$f_2(v_i,v_j)=f_1(v_i)f_1(v_j), i=\overline{1,I-1}, j=\overline{i+1,I}. \tag{7}$$

3. Підставляючи одержане з (6) значення V_1 до формули (2), одержуємо вираз для прогнозованого процесу, який у момент $i=1$ проходить через точку $x(1)$:

$$X^{(1)}(i)=m(i)+(x(1)-m(1))\varphi_1(i)+\sum_{v=2}^{i}V_v\varphi_v(i), i=\overline{1,I}. \tag{8}$$

з математичним очікуванням

$$m^{(1)}(i)=m(i)+(x(1)-m(1))\varphi_1(i), i=\overline{1,I}, \tag{9}$$

4. Обчислення

$$X^{(1)}(i)=m^{(1)}(i)+\sum_{v=2}^{i}V_v\varphi_v(i), i=\overline{1,I}. \tag{10}$$

Якщо на наступному етапі спостережень одержано наступне значення $x(2)$ тієї ж реалізації процесу, то для цього значення є справедливим (10), де $x(2)=m^{(1)}(2)+v_2$.

5. Повторити операції для випадку $\mu=1$, та одержати

$$m^{(2)}(i)=m^{(1)}(i)+(x(2)-m^{(1)}(2))\varphi_2(i), i=\overline{1,I}; \tag{11}$$

$$X^{(2)}(i)=m^{(2)}(i)+\sum_{v=3}^{i}V_v\varphi_v(i), i=\overline{1,I}\,. \quad (12)$$

6. Повторити ітерації для довільного числа $k<I$ моментів контролю за наступною формулою:

$$m^{(0)}(i)=m(i), i=\overline{1,I}\,,$$
$$m^{(k)}(i)=m^{(k-1)}(i)+\left(x(k)-m^{(k-1)}(i)\right)\varphi_k(i), i=\overline{1,I}\,; \quad (13)$$

$$X^{(k)}(i)=m^{(k)}(i)+\sum_{v=k+1}^{i}V_v\varphi_v(i), i=\overline{1,I}\,. \quad (14)$$

Вирази (13) – (14) повністю описують лінію прогнозу, у якому (14) – математичне очікування у точках t_i.

Дослідження ефективності прогнозування часу здійснення кібератаки на основі результатів аналізу нестаціонарних процесів проведемо на прикладі прогнозу атаки за результатами спостереження активності хакерів перед здійсненням певної атаки. Так, до процесу, реалізації якого наведено на рис. 1, застосовано методику прогнозування (6) – (14), взявши у якості початкових значень спостережень окремі точки часового ряду, які відповідають частковій траєкторії спостереження № 1 (на рис. 1 крива блакитного кольору). Взявши зазначену криву у якості контрольної, у якості початкових даних спостережень було взято перші значення часового ряду, які відповідають першим t = 12 діб спостережень.

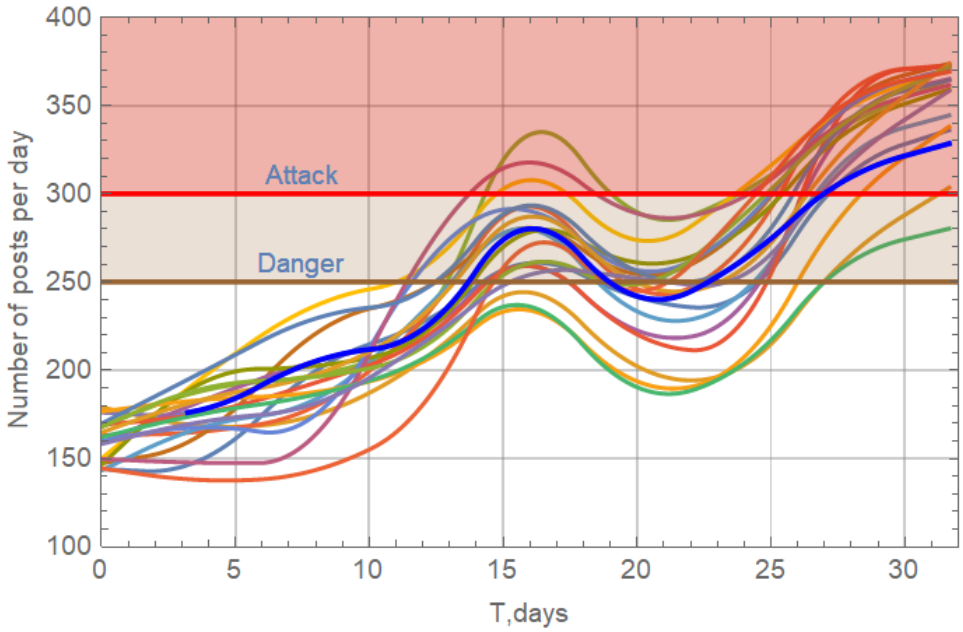

Рис. 1. Результати спостережень за активністю хакерів (кількість постів за добу)

На рис. 2 зображено роботу алгоритму прогнозування за відсутності статистичної інформації. Фізичний зміст зазначеного полягає в тому, що, знаючи середньостатистичні параметри активності хакерів перед атакою та точку входу до прогнозу можна з достатньою точністю передбачити майбутню поведінку системи. Тобто, запропонований математичний апарат сам "підбирає" необхідну траєкторію поведінки системи у залежності від точки входу та середньостатистичної траєкторії. Точність результату, як було зазначено, лежить у межах дисперсії випадкового процесу.

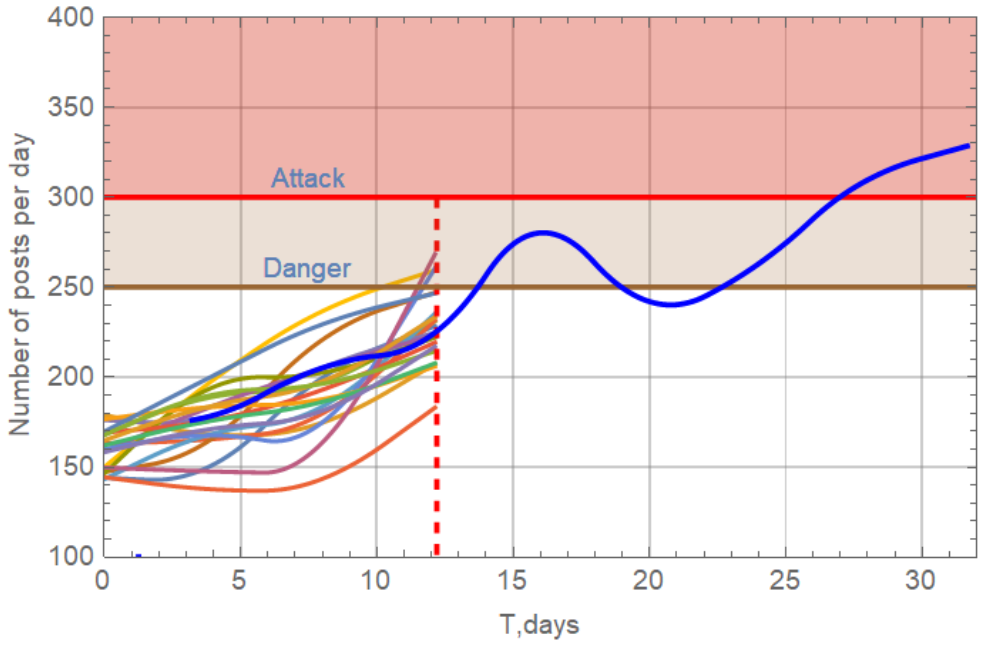

Рис. 2. Результати прогнозування окремої траєкторії

Висновки

1. Прогнозування часу здійснення кібератаки на основі аналізу нестаціонарного процесу активності хакерів під час конференцій є одним з необхідних елементів аналізу та оптимізації систем захисту організацій та установ.

2. Прогнозування атак на основі канонічного розкладання процесу забезпечує вирішення задачі пошуку невідомих майбутніх значень як для скалярних так і для векторних процесів. Таке подання дозволяє застосовувати його для будь-якого набору даних у тому числі даних з пропусками та незбалансованих за часом. Єдиним обмеженням, яке накладається на досліджуваний випадковий процес є скінченність його дисперсії. Запропонований підхід точно визначає випадковий процес у точках контролю та забезпечує мінімум середнього квадрата похибки наближення у проміжках між цими точками.

Напрямком подальших досліджень у цій сфері може бути широке коло питань удосконалення методу для забезпечення можливості прогнозування на інтервалі поза межами доступної статистики, в умовах сильної зашумленості вихідних даних або їх часткової відсутності.

Перелік посилань

1. D. Dagon, C. Zou, and W. Lee. Modeling botnet propagation using time zones. In Proceedings of the 13th Network and Distributed System Security Symposium (NDSS'06), 2006.

2. Neil Daswani, Michael Stoppelman, the Google Click Quality, and Inc Security Teams, Google. The anatomy of Clickbot.A. In USENIX First Workshop on Hot Topics in Understanding Botnets (HotBots), 2007

3. HoneynetProject, "About the honeynet project," 2008. http://www.honeynet.org/about

4. Project Honeypot [Електронний ресурс] – Режим доступу: https://www.projectHoneypot.org/

5. Amit D. Lakhani. Deception Techniques Using Honeypots. Dr. Kenneth G. Paterson Information Security Group Royal Holloway, University of London UK – 75 с.

6. HONEYPOT SECURITY: The Government of the Hong Kong Special Administrative Region. 2008 – 13 с

7. Caleb Townsend. What is a Honeypot? https://www.uscybersecurity.net/Honeypot

8. Securitycode. Что такое Honeypot и от чего защищать виртуальные ИС? 2011 https://habr.com/ru/company/securitycode/blog/119821

Надійшла: 28.07.2020 Рецензент: д.т.н., професор Гайдур Г.І.