УДК 004.414.3:004.056.53
DOI: 10.31673/2409-7292.2023.030808

Кузьменко А. О., Веселков Н. Л.

МЕТОДИКА АНАЛІЗУ ТА ПРОГНОЗУВАННЯ КІБЕРІНЦИДЕНТІВ НА ОСНОВІ МЕТОДУ ГОЛОВНИХ КОМПОНЕНТ

В статті розглянуто теоретичні аспекти методу головних компонент, його застосування для аналізу даних про кіберінциденти та побудови прогностичних моделей. Окрема увага приділена експериментальним результатам, їх аналізу та обговоренню з метою підвищення ефективності методики прогнозування кіберінцидентів. Ця стаття спрямована на вдосконалення інструментів та підвищення рівня кібербезпеки шляхом застосування новітніх методів аналізу даних та прогнозування подій у кіберпросторі. Корисність методу головних компонент при аналізі даних кіберінцидентів ґрунтується на можливості зменшення обсягів аналізу інформації та визначення найбільш суттєвих факторів кіберінцидентів. Перевагою описаного методу аналізу статистики кіберінцидентів є те, що він може застосовуватись незважаючи на характер розподілу випадкових величин – показників інцидентів. Завдяки основним властивостям методу головних компонент він достатньо успішно може бути використаний для прогнозування статистики кіберінцидентів, забезпечуючи при цьому найменшу похибку прогнозу. Загальна модель ризику кіберінцидентів комплексно враховує вплив на кібербезпеку усього спектру технічних, організаційних та людських чинників та будується на основі схеми виникнення кіберінциденту, у якій кожен інцидент пов’язується з передумовою його виникнення. Зазначений підхід дозволяє здійснювати аналіз безпосередніх причинно-наслідкових зв’язків, що мають місце в процесі інциденту та виявляти як основні, так і приховані причини та види подій, що призводять до кіберінцидентів на підставі статистичних даних. Наведений у статті приклад демонструє прикладну спрямованість компонентного аналізу, зокрема для задач прогнозу числа вихідних показників кіберінцидентів за, порівняно малим числом допоміжних (латентних) змінних, що виражають причини цього явища, візуалізації багатовимірних даних та виділення типоутворюючих ознак кіберінцидентів.

Ключові слова: кіберінцидент, кібератака, метод головних компонент, аналіз кіберінцидентів, прогноз кіберінцидентів.

Вступ

У зв’язку з постійним розвитком технологій та поширенням цифрових систем, кібербезпека стала однією з найбільш актуальних проблем сучасного світу. Зростання кількості та складності кіберінцидентів наголошує на необхідності вдосконалення методів їх аналізу та прогнозування для запобігання можливим загрозам. Ця стаття присвячена розгляду методики аналізу та прогнозування кіберінцидентів з використанням методу головних компонент (PCA). Метод головних компонент є потужним інструментом аналізу даних, який дозволяє зменшити розмірність даних, виявити складні зв'язки та визначити ключові фактори, що впливають на події в кіберпросторі.

Постановка проблеми

Традиційно, виявлення кібератак покладається на реактивні методи, коли алгоритми зіставлення шаблонів допомагають експертам сканувати системні журнали та мережевий трафік на наявність відомих вірусів або сигнатур зловмисного програмного забезпечення. Останні дослідження представили ефективні моделі машинного навчання для виявлення кібератак, обіцяючи автоматизувати завдання виявлення, відстеження та блокування зловмисного програмного забезпечення. Набагато менше зусиль було спрямовано на прогнозування кібератак, особливо за межами короткострокового масштабу годин і днів. Для практики потрібні методи, які можуть прогнозувати напади в довгостроковій перспективі, оскільки це дає захисникам більше часу для розробки та обміну захисними діями та інструментами. Сьогодні довгострокові прогнози хвиль атак здебільшого ґрунтуються на суб’єктивному сприйнятті досвідчених експертів-людей, яке може бути порушено через дефіцит досвіду з кібербезпеки. Ця стаття представляє новий підхід, який використовує неструктуровані великі дані та журнали для прогнозування тенденції кібератак у великому масштабі на роки вперед. Маючи можливість такого прогнозування аналітики служб безпеки зможуть організовувати захист більш ефективно, раціонально розподіляючи ресурси та можливості організації.

Аналіз публікацій

Проблема визначення причин кіберінцидентів є ключовою для забезпечення ефективної профілактики цього явища на всіх рівнях управління захистом інформації.

У статті [1] пропонується структура, яка використовує щомісячний набір даних про великі кіберінциденти в 36 країнах за останні 11 років, отриманими з трьох основних категорій джерел великих даних, – наукової літератури, новин, блогів, і твітів. Методика автоматизовано визначає тенденції майбутніх атак і генерує цикл загроз, який розбиває на п’ять ключових фаз, які складають життєвий цикл усіх відомих кіберзагроз. Стаття [2] базується на багатогранних рішеннях машинного навчання та розробляє інтегровану систему для перетворення великих обсягів загальнодоступних даних у сукупні сигнали з імпутацією, які є актуальними та передбачають кіберінциденти. Комплексний аналіз окремих частин і інтегрованого цілого демонструє ефективність і компроміси запропонованого підходу.

Автори [3] використали інформацію з різних веб-форумів, використовуючи структуру відповідної мережі взаємодії користувачів з метою прогнозування корпоративних кібератак. Також автори використовують набір функцій соціальних мереж на додаток до моделей навчання під наглядом і перевіряють їх за допомогою бінарної класифікації, яка намагається передбачити, чи буде атака на організацію в певний день. У статті [4] наведено метод прогнозування атак, які базуються на тактиці, техніці та загальних знаннях з використанням різних аналітичних процедур, як-от кластеризація, аналіз часових рядів і генетичні алгоритми. Метод визначає тенденції у використанні методів атак і створює прогнози щодо майбутнього зловмисного програмного забезпечення та використовуваних методів атак.

У публікації [5] наведено методику ефективного прогнозування кібератак, яка полягає у тому, щоб дізнатися модель поведінки зловмисника, передбачити майбутні атаки та вибрати відповідні заходи протидії. Автори вирішують проблему складності обчислень, розробляючи моделі противника та відповідні методи зменшення складності. Стаття [6] спрямована на використання часових кореляцій між кількістю атак на день, щоб передбачити майбутню інтенсивність кіберінцидентів. Завдяки аналізу даних про атаки, зібраних із Hackmageddon, було виявлено кореляцію між повідомленим обсягом атак протягом послідовних днів. У цьому документі представлено систему прогнозування, яка має на меті передбачити кількість кібератак за певний день лише на основі набору історичних даних про кількість атак. Наша система проводить прогнозування часових рядів ARIMA для всіх раніше зібраних інцидентів, щоб передбачити очікувану кількість атак у майбутньому.

Разом з тим, не зважаючи на значне число публікацій, присвячених прогнозуванню кібератак і та кіберінцидентів, залишається не вирішеним питання, яким чином впливають на кіберінциденти загальні характеристики інформаційних систем та ресурсне забезпечення потреб захисту інформації, що не дозволяє виконувати комплексне оцінювання впливу на захищеність компанії усього спектру інформаційних, управлінських та технічних чинників, а це суттєво збіднює результати аналізу і не дозволяє ураховувати тенденції змін зовнішніх факторів для коригування профілактики кібератак.

Метою статті є розроблення методики аналізу та прогнозування кіберінцидентів на основі методу головних компонент, яка б дозволяла виконувати комплексне оцінювання впливу на захищеність компанії спектру інформаційних, управлінських та технічних чинників.

Визначення ризику кіберінциденту. Під *ризиком кіберінциденту* у статті будемо розуміти кількісну міру прояву небезпеки кіберінциденту в компанії на основі аналізу інцидентів за усією сукупністю їх ознак:

$$R = \sum_{i=1}^{n} S_i P_i , \qquad (1)$$

де S_i – наслідки інциденту; P_i – ймовірність (частота) інцидентів; n – кількість інцидентів.

Для визначення наслідків кіберінциденту S_i можна використовувати економічні показники, що дозволяє оцінювати ризик R у грошових одиницях. При використанні єдиної розмірності при обчисленні наслідків ризик можна подати у вигляді суми складових

$$R = R_o + R_ц + R_к, \tag{2}$$

де R_o, $R_ц$, $R_к$ – ризики порушення доступності, цілісності та конфіденційності інформації відповідно.

Разом з тим, за умови близьких значень наслідків кіберінцидентів або у разі неможливості їх оцінювання, обчислення ризиків можна здійснювати лише за ймовірностями появи кіберінцидентів. Більш актуальним у питанні визначення ризику є визначення причин інцидентів. Для підвищення інформативності у статті досліджуються бінарні поєднання (групи) "причина інциденту – вид інциденту", що багатократно збільшує кількість можливих варіантів причин інцидентів, прихованих у статистичних даних.

Дослідження бінарних груп базується на причинно-наслідковому ланцюжку (рис. 1).

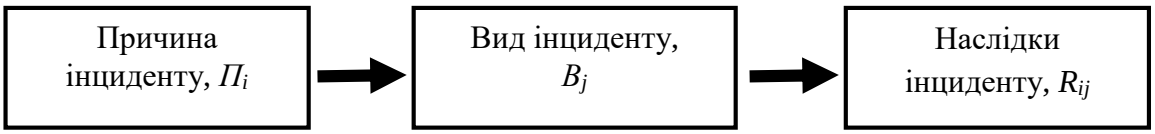

Рис. 1. Схема виникнення кіберінциденту

До причин інцидентів $П_i$, відносяться: 1) слабкі або вкрадені облікові дані, або паролі; 2) бекдори, уразливості програм; 3) шкідливе програмне забезпечення; 4) соціальна інженерія; 5) недосконалі політики безпеки; 6) інсайдерські загрози; 7) фізичні атаки; 8) неправильна конфігурація програм; 9) помилки користувачів та ін.

Інформація про види інцидентів B_j може включати наступні події: 1) атака зловмисним програмним забезпеченням; 2) фішингова атака; 3) атака «людина посередині»; 4) атака SQL Injection; 5) криптоджекінг; 6) експлойт нульового дня; 7) спуфінг; 8) атаки впровадження коду; 9) тунелювання DNS; 10) підробка DNS; 11) програми-вимагачі; 12) розподілені атаки на відмову в обслуговуванні (DDos); 13) спам; 14) захоплення корпоративного облікового запису (CATO); 15) інтерпретація URL; 16) перехоплення сесії; 17) атака грубою силою; 18) Веб-атаки; 19) Троянські коні; 20) міжсайтові сценарії (XSS) атаки; 21) атаки на прикладному рівні; 22) словникові атаки; 23) віруси; 24) черв'яки; 25) руткіти; 26) кейлоггери; 27) розширена постійна загроза (APT) та ін.

Приймемо, що для оцінювання складових схеми (рис. 1) застосовуються кількісні характеристики у вигляді показників ризику. Тобто причини інциденту оцінюються за показниками ризику за кожною з причин $П_i$ (i – індекс причини, $i = 1, 2, ..., n$), а види інцидентів – за показниками ризику, що відповідають кожному інциденту B_j (j – індекс виду інциденту, $j = 1, 2, ..., m$). Показники ризику інциденту загалом R та за окремими причинами чи видами інцидентів визначаються за частотою інцидентів: $R^t = N^t / N_c$, де N^t – кількість інцидентів за окремими причинами $П_i$ чи видами подій B_j, N_c – загальна кількість інцидентів.

Специфіка статистичної інформації про причини та види кіберінцидентів полягає в тому, що виконується умова

$$R^t = \sum_{i=1}^{n} P(П_i^t) = \sum_{j=1}^{m} P(B_j^t), \tag{3}$$

тобто загальний ризик кібератаки R^t дорівнює сумі ризиків за причинами або сумі ризиків за видами інцидентів. Для спрощення будемо розглядати лише ситуацію, коли кожному інциденту відповідає лише одна причина. Тобто ризик інциденту залежить лише від однієї з причин, що наводяться у статистичних звітах:

$$P(B_j^t) = f\left[P(П_i^t)\right]. \tag{4}$$

Оскільки ризик інциденту залежить від ймовірності його настання, з показниками ризику можна виконувати дії, передбачені теорією імовірності. Зокрема, можна визначити поняття умовної ймовірності. Відомо, що умовною імовірністю $P_A(B)$ називають імовірність події B, обчисленої за умови того, що подія A вже настала [7]. Тобто ураховуючи схему причинно-наслідкових зв'язків (рис. 1), приймається, що для розрахунку ймовірності (ризику) інциденту при прояві певної причини інциденту може застосовуватися умовна ймовірність. Для розрахунку умовної ймовірності використовується формула Байєса

$$P_П(B_j) = \frac{P(B_j)P_{Bj}(П_i)}{P(П_i)}. \tag{5}$$

З урахуванням того, що статистична база побудована так, що мають виконуватися умови (3) і (4) одночасно, формула (5) набуває вигляду

$$P_{Пi}(B_j) = \frac{P(B_j)P(П_i)}{\sum_{i=1}^{n} P(П_i)}. \tag{6}$$

За формулою (6) виконуються розрахунки матриці ризиків кіберінцидентів. Така матриця має вигляд:

$$R_{ij}^t = \begin{vmatrix} R_{П1B1} & \cdots & R_{ПnB1} \\ \cdots & \cdots & \cdots \\ R_{П1Bm} & \cdots & R_{ПnBm} \end{vmatrix}, \tag{7}$$

де $R_{П1B1}, ..., R_{ПnBm}$ – значення ризиків інцидентів для бінарних комплексів "причина інциденту – вид інциденту"; $i = 1, 2, ..., n$ – кількість причин кіберінцидентів $П_i$; $j = 1, 2, ..., m$ – кількість видів кіберінцидентів B_j.

Для перевірки результатів, отриманих з використанням формули (7), в статті використано два методи. Перший метод – порівняння розрахованих за формулою (7) матриць ризику з отриманими шляхом безпосереднього заповнення матриць за результатами аналізу актів розслідування кіберінцидентів. Другий метод – це метод аналітичного розв'язку системи

лінійних рівнянь, отриманих з використанням методу головних компонент та регресійного аналізу. Суть другого методу полягає в тому, що виконується *компонентний аналіз* масиву статистичної інформації про причини інцидентів $П_i$.

Використовується така особливість головних компонент, що вони статистично не зв'язані між собою, тобто є за визначенням ортогональними. Така особливість дозволяє отримати регресійні залежності між ризиками інцидентів (залежні змінні) та значеннями головних компонент, отриманих у результаті аналізу причин ризику інциденту (незалежні змінні)

$$B_j = f(\Gamma_{Пр}). \tag{8}$$

Обґрунтування застосовності методу головних компонент для аналізу статистики кіберінцидентів. Метод головних компонент [8] базується на задачі найкращої апроксимації скінченої множини точок прямими та площинами. Дано скінчену множину векторів $x_1, x_2, ..., x_m \in R^n$. Для кожного $k = 0,1,...,n-1$ серед усіх k – вимірних лінійних підпросторів у R^n необхідно відшукати таке $L_k \subset R^n$, що сума квадратів відхилень x_i від L_k буде мінімальною $\sum_{i=1}^{m} dist^2(x_i, L_k) \to min$, де $dist(x_i, L_k)$ – евклідова відстань від точки до лінійного підпростору.

Всякий k – вимірний лінійний підпростір в R^n може бути задано як множину лінійних комбінацій $L_k = \{a_0 + \beta_1 a_1 + ... + \beta_k a_k | \beta_i \in R\}$, де параметри β_i пробігають дійсну пряму R, $a_0 \in R^n$ а $\{a_1, ..., a_k\} \subset R^n$ – ортонормований набір векторів $dist^2(x_i, L_k) = \left\| x_i - a_0 - \sum_{j=1}^{k} a_j (a_j, x_i - a_0) \right\|^2$, де $\|\bullet\|$ – евклідова норма; (a_j, x_i) – евклідовий скалярний добуток.

Або, у координатній формі: $dist^2(x_i, L_k) = \sum_{l=1}^{n} \left(x_{il} - a_{0l} - \sum_{j=1}^{k} a_{jl} \sum_{q=1}^{n} a_{jq} (x_{iq} - a_{0q}) \right)^2$.

Вирішення задачі апроксимації для $k = 0,1,...,n-1$ дається набором вкладених лінійних підпросторів $L_0 \subset L_1 \subset L_2 \subset ... \subset L_{n-1}$, $L_k = \{a_0 + \beta_1 a_1 + ... + \beta_k a_k | \beta_i \in R\}$. Ці лінійні підпростори визначаються ортонормованим набором векторів $\{a_1, ..., a_{n-1}\}$ (векторами головних компонент) і вектором a_0, який відшукується шляхом вирішення задачі мінімізації для L_0: $a_0 = \arg\min_{a_0 \in R^n} \sum_{i=1}^{m} dist^2(x_i, L_0)$.

Корисність методу головних компонент при аналізі даних кіберінцидентів ґрунтується на можливості зменшення обсягів аналізу інформації та визначення найбільш суттєвих факторів кіберінцидентів. При цьому вектори головних компонент можуть бути знайдені як рішення однотипних задач оптимізації за наступним алгоритмом:

1. Центрування даних (шляхом віднімання середніх значень): $x_i := x_i - \overline{X}$. Тепер $\sum_{i=1}^{m} x_i = 0$;

2. Відшукання першої головної компоненти, як вирішення задачі: $a_1 = \underset{\|a_1\|=1}{arg\,min}\left(\sum_{i=1}^{m}\|x_i - a_1(a_1, x_i)\|^2\right)$. Якщо рішення не єдине, то обираємо одне з них.

3. Обчислюємо з даних проекцію на першу головну компоненту: $x_i := x_i - a_1(a_1, x_i)$.

4. Знаходимо другу головну компоненту як вирішення задачі $a_2 = \underset{\|a_2\|=1}{arg\,min}\left(\sum_{i=1}^{m}\|x_i - a_2(a_2, x_i)\|^2\right)$. Якщо рішення не єдине, то обираємо одне з них.

2*k*-1. Знаходимо проекцію на $(k-1)$-у головну компоненту: $x_i := x_i - a_{k-1}(a_{k-1}, x_i)$;

2*k*. Знаходимо *k* – ту головну компоненту як вирішення задачі: $a_k = \underset{\|a_k\|=1}{arg\,min}\left(\sum_{i=1}^{m}\|x_i - a_k(a_k, x_i)\|^2\right)$. Якщо рішення не єдине, то обираємо одне з них.

З урахуванням можливостей сучасних засобів моделювання зазначений алгоритм для статистичного ряду даних $X = \begin{bmatrix} x_{11} & \dots & x_{1m} \\ \cdots & \ddots & \cdots \\ x_{n1} & \dots & x_{nm} \end{bmatrix}$, де присутні *m* ознак і *n* спостережень можна записати наступним чином:

1. Нормуємо складові векторів (рядків) матриці *X* шляхом виконання операції $z_i = \frac{x_{ji} - \overline{x}_i}{\sigma_{x_i}}, j = 1,...,n, i = 1,...,m$, де σ_{x_i} – середньоквадратичне відхилення випадкової величини *X* від середнього значення за стовпчиком матриці *X*. Одержуємо матрицю *Z* розміром $n \times m$.

2. З матриці *Z* відшукуємо кореляційну (коваріаційну) матрицю $R = \lfloor r_{ij} \rfloor_{m \times m}$.

3. Відшукуємо множину власних значень матриці *R* та упорядковуємо її за зменшенням складових $\lambda_i, i = 1,...,m$.

4. Формуємо діагональну матрицю з власних значень матриці *R* $\Lambda = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \cdots & \ddots & \cdots \\ 0 & \cdots & \lambda_m \end{bmatrix}$.

5. З матриці *R* формуємо матрицю власних векторів матрицю $U = \begin{bmatrix} (u_{11},...,u_{1m}) \\ ... \\ (u_{n1},...,u_{nm}) \end{bmatrix}$.

6. Відшукуємо вирішення задачі у вигляді матриці $A = U\sqrt{\Lambda}$, де $\sqrt{\Lambda}$ матриця коренів з кожного елемента матриці Λ.

Знайдені вектори $\{a_1,...,a_{n-1}\}$ ортонормовані просто у результаті вирішення описаної задачі оптимізації, однак щоб не дати похибкам обчислення порушити взаємну

ортогональність векторів головних компонент, можна включити $a_k \perp \{a_1,...,a_{k-1}\}$ до умови задачі оптимізації.

Перевагою описаного методу аналізу статистики кіберінцидентів є те, що він може застосовуватись незважаючи на характер розподілу випадкових величин – показників інцидентів. Однак, цей метод не завжди ефективно знижує розмірність при заданих обмеженнях на точність. Прямі та площини не завжди забезпечують добру апроксимацію. Наприклад, дані можуть з достатньою точністю описуватись якою-небудь кривою, а сама крива може бути складно розташована у просторі даних. Також у випадку ізотропного розподілу розподіл даних еліпсоїд розсіювання являтиме собою гіперкулю і тому зменшити розсіювання методами апроксимації буде неможливо.

Обґрунтування застосовності методу головних компонент для прогнозування кіберінцидентів. Завдяки основним властивостям методу головних компонент він достатньо успішно може бути використаний для прогнозування статистики кіберінцидентів, забезпечуючи при цьому найменшу похибку прогнозу. Покажемо, що з допомогою перших p' головних компонент $z^{(1)}, z^{(2)},..., z^{(p')}$ при $p' < p$, вихідних ознак $x^{(1)}, x^{(2)},..., x^{(p)}$ досягається найкращий прогноз цих ознак серед усіх прогнозів, які можна побудувати за допомогою p' лінійних комбінацій набору з p довільних ознак.

Розглянемо більш детально. Нехай необхідно замінити вихідний досліджуваний p-вимірний вектор спостережень X на вектор $Z = \left(z^{(1)}, z^{(2)},..., z^{(p')}\right)^T$ меншої розмірності p', у якому кожна з компонент була б лінійною комбінацією p вихідних (або допоміжних) ознак, втративши при цьому не надто багато інформації. Інформативність нового вектора Z залежить від того, у якій мірі p' введених допоміжних змінних дають можливість "відновити" p вихідних ознак за допомогою відповідних лінійних комбінацій $z^{(1)}, z^{(2)},..., z^{(p')}$. Можна уявити, що похибка σ прогнозу X по Z буде визначатися залишковою дисперсійною матрицею вектора X при відніманні з нього найкращого прогнозу по Z, тобто матрицею $\Delta = \left[\Delta_{ij}\right]$, де $\Delta_{ij} = E\left\{\left(x^{(i)} - \sum_{l=1}^{p'} b_{il} z^{(l)}\right)\left(x^{(j)} - \sum_{l=1}^{p'} b_{jl} z^{(l)}\right)\right\}$. Тут $\sum_{l=1}^{p'} b_{il} z^{(l)}$ – найкращий у сенсі найменших квадратів прогноз $x^{(i)}$ по компонентам $z^{(1)}, z^{(2)},..., z^{(p')}$. Похибка прогнозу X по Z задається як деяка визначена функція від елементів матриці $\Delta = \left[\Delta_{ij}\right]$, тобто $\sigma = f(\Delta)$, де $f(\Delta)$ визначає деякий критерій якості прогнозування.

Розглянемо наступні міри похибки прогнозу:

1. $f(\Delta) = Tr(\Delta) = \Delta_{11} + \Delta_{22} + ... + \Delta_{pp}$ – на основі сліду матриці $\Delta = \left[\Delta_{ij}\right]$;
2. $f(\Delta) = \|\Delta\| = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p} \Delta_{ij}^2}$ – на основі евклідової норми матриці $\Delta = \left[\Delta_{ij}\right]$.

Як відомо, обидві міри одночасно досягають мінімуму тоді і тільки тоді, коли у якості $z^{(1)}, z^{(2)},..., z^{(p')}$ обрано перші p' головних компонент вектора X, причому величина похибки прогнозу $\sigma = f(\Delta)$ явним чином виражається через останні $p - p'$ власних чисел вихідної коваріаційної матриці C або наближено – через останні $p - p'$ власних чисел $\lambda_{p'+1},..., \lambda_p$ вибіркової коваріаційної матриці $\hat{C}$, побудованої за спостереженнями $X_1, X_2,..., X_n$. Зокрема, при $f(\Delta) = Tr(\Delta): \sigma \approx \lambda_{p'+1} + \lambda_{p'+2} + ... + \lambda_p$; при $f(\Delta) = \|\Delta\|: \sigma \approx \sqrt{\lambda_{p'+1}^2 + \lambda_{p'+2}^2 + ... + \lambda_p^2}$.

Пояснимо ідею прогнозування вихідних ознак $x^{(1)}, x^{(2)}, ..., x^{(p)}$ за допомогою меншого, ніж *p*, числа їх лінійних комбінацій на прикладі.

Приклад. При формуванні типоутворюючих ознак причин кіберінцидентів було досліджено статистичні дані за 5 років ($n = 5$) за трьома основними групами передумов: технічний фактор $x^{(1)}$, організаційний фактор $x^{(2)}$ та людський фактор $x^{(3)}$. За спостереженими даними $\left(x_i^{(1)}, x_i^{(2)}, x_i^{(3)}\right), i = 1,...,5$ було визначено вибіркову коваріаційну матрицю

$$\hat{C} = \begin{bmatrix} 265.524 & 108.468 & 56.233 \\ 135.585 & 68.692 & 34.430 \\ 80.333 & 41.316 & 22.217 \end{bmatrix}.$$

Власні корені такої матриці $\hat{C}$ будуть: $\lambda_1 = 341.798$, $\lambda_2 = 13.440$, $\lambda_3 = 1.194$.

Матриця власних векторів $U = \begin{bmatrix} 0.848 & 0.455 & 0.272 \\ 0.432 & -0.726 & -0.536 \\ -0.018 & -0.426 & 0.905 \end{bmatrix}$.

У результаті у якості головних компонент одержимо: $z^{(1)} = 0.848x^{(1)} + 0.455x^{(2)} + 0.272x^{(3)}$, $z^{(2)} = 0.432x^{(1)} - 0.726x^{(2)} - 0.536x^{(3)}$, $z^{(3)} = -0.018x^{(1)} - 0.426x^{(2)} + 0.905x^{(3)}$.

Тут під $x^{(1)}$, $x^{(2)}$, $x^{(3)}$ вбачаються відхилення кількості інцидентів через технічний фактор $x^{(1)}$, організаційний фактор $x^{(2)}$ та людський фактор $x^{(3)}$ від їх середніх значень. У цьому прикладі $p = 3$. Визначимо за мету знизити розмірність вихідного факторного простору до одиниці $p' = 1$, тобто описати усі три групи ознак за допомогою лінійних комбінацій тільки від однієї допоміжної змінної. У відповідності з розглянутою вище властивістю "автопрогнозу" головних компонент візьмемо у якості цієї єдиної допоміжної змінної першу головну компоненту, тобто змінну $z^{(1)} = 0.848x^{(1)} + 0.455x^{(2)} + 0.272x^{(3)}$.

За методом найменших квадратів невідомі коефіцієнти b_{i1} обчислюються за виразом

$$b_{i1} = \frac{\operatorname{cov}\left(x^{(i)}, z^{(1)}\right)}{Dz^{(1)}} = \frac{0.848\operatorname{cov}\left(x^{(i)}, x^{(1)}\right) + 0.455\operatorname{cov}\left(x^{(i)}, x^{(2)}\right) + 0.272\operatorname{cov}\left(x^{(i)}, x^{(3)}\right)}{Dz^{(1)}}.$$

Підставивши до цієї формули значення $\operatorname{cov}\left(x^{(i)}, x^{(j)}\right)$, взяті з коваріаційної матриці *C* для нашого прикладу одержимо $x^{(1)} = b_{11}z^{(1)} + \varepsilon^{(1)} = 0.848z^{(1)} + \varepsilon^{(1)}$, $x^{(2)} = b_{21}z^{(1)} + \varepsilon^{(2)} = 0.455z^{(1)} + \varepsilon^{(2)}$, $x^{(3)} = b_{31}z^{(1)} + \varepsilon^{(3)} = 0.272z^{(1)} + \varepsilon^{(3)}$, де $\varepsilon^{(i)}$ – випадкові (залишкові) похибки прогнозу вихідних центрованих компонент за першим головним компонентом $z^{(1)}$.

Сумарна відносна похибка прогнозу ознак $x^{(1)}$, $x^{(2)}$, $x^{(3)}$ за $z^{(1)}$ може бути обчислена за виразом $\delta_{сум.} = 100\left(\frac{Tr(\Delta)}{D\left(x^{(1)} + x^{(2)} + x^{(3)}\right)}\right) = 100\frac{\lambda_2 + \lambda_3}{\lambda_1 + \lambda_2 + \lambda_3} = 4.1\%$, що і підтверджує достатню ефективність використання методу головних компонент для прогнозування статистичних характеристик, у т.ч. і для прогнозу ризиків кіберінцидентів.

Наведений приклад демонструє прикладну спрямованість компонентного аналізу, зокрема для задач прогнозу числа вихідних показників кіберінцидентів за, порівняно малим

числом допоміжних (латентних) змінних, що виражають причини цього явища, візуалізації багатовимірних даних та виділення типоутворюючих ознак кіберінцидентів.

Висновки

Загальна модель ризику кіберінцидентів комплексно враховує вплив на кібербезпеку усього спектру технічних, організаційних та людських чинників та будується на основі схеми виникнення кіберінциденту, у якій кожен інцидент пов'язується з передумовою його виникнення. Зазначений підхід дозволяє здійснювати аналіз безпосередніх причинно-наслідкових зв'язків, що мають місце в процесі інциденту та виявляти як основні, так і приховані причини та види подій, що призводять до кіберінцидентів на підставі статистичних даних.

Для забезпечення фільтрації статистичних даних та візуалізації результатів для обробки наявної статистики кіберінцидентів найбільш доцільним є метод головних компонент. Корисність цього методу при аналізі даних кіберінцидентів грунтується на можливості зменшення обсягів аналізу інформації та визначення найбільш суттєвих причин кіберінцидентів. Завдяки основним властивостям методу головних компонент він достатньо успішно може бути використаний для прогнозування значного числа вихідних показників кіберінцидентів за, порівняно малим числом допоміжних змінних, що виражають причини цього явища, забезпечуючи при цьому найменшу похибку прогнозу.

Перелік посилань

1. Almahmoud, Z., Yoo, P.D., Alhussein, O. *et al.* A holistic and proactive approach to forecasting cyber threats. *Sci Rep* 13, 8049 (2023). https://doi.org/10.1038/s41598-023-35198-1
2. Okutan, Ahmet & Werner, Gordon & Yang, Shanchieh & McConky, Katie. (2018). Forecasting cyberattacks with incomplete, imbalanced, and insignificant data. Cybersecurity. 1. 10.1186/s42400-018-0016-5.
3. Sarkar, S., Almukaynizi, M., Shakarian, J., & Shakarian, P. (2019). Predicting enterprise cyber incidents using social network analysis on dark web hacker forums. The Cyber Defense Review, 87–102. https://www.jstor.org/stable/26846122
4. Florian Klaus Kaisera, Tobias Budiga, Elisabeth Goebela, Tessa Fischera, Jurek Muffa, Marcus Wiensa and Frank Schultmanna. Attack Forecast and Prediction. C&ESAR'21: Computer Electronics Security Application Rendezvous, November 16-17, 2021, Rennes, France
5. Jones, M., Kotsalis, G., Shamma, J.S. (2013). Cyber-Attack Forecast Modeling and Complexity Reduction Using a Game-Theoretic Framework. In: Tarraf, D. (eds) Control of Cyber-Physical Systems. Lecture Notes in Control and Information Sciences, vol 449. Springer, Heidelberg. https://doi.org/10.1007/978-3-319-01159-2_4
6. Gordon Werner and Shanchieh Jay Yang and Katie McConky. Time series forecasting of cyber attack intensity. Proceedings of the 12th Annual Conference on Cyber and Information Security Research. 2017. https://api.semanticscholar.org/CorpusID:3590139
7. Карташов, М. В. Імовірність, процеси, статистика. – Київ: ВПЦ Київський університет, 2007. – 504 с.
8. Єременко, В. С., Осінцева М. Б. Застосування методу головних компонент в задачі аналізу спектрів вільних коливань. Вісник Вінницького політехнічного інституту. 2022. № 4. – С. 6–12.

Надійшла: 02.11.2023 Рецензент: д.т.н., професор Кожухівський А.Д.