КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 3 (19), 2023
ISSN 2663 - 4023
DOI 10.28925/2663-4023.2023.19.2033
УДК 681.3.01(075)

Субач Ігор Юрійович
доктор технічних наук, доцент, завідувач кафедри
Інститут спеціального зв'язку та захисту інформації Національного технічного університету України Київський політехнічний інститут імені Ігоря Сікорського, Київ, Україна
ORCID ID: 0000-0002-9344-713X
igor_subach@ukr.net

Микитюк Артем В'ячеславович
аспірант
Інститут спеціального зв'язку та захисту інформації Національного технічного університету України Київський політехнічний інститут імені Ігоря Сікорського, Київ, Україна
ORCID ID: 0000-0002-8307-9978
mukuta8888@gmail.com

МЕТОД ФОРМУВАННЯ АСОЦІАТИВНИХ ПРАВИЛ З БАЗИ ДАНИХ SIEM – СИСТЕМИ НА ОСНОВІ ТЕОРІЇ НЕЧІТКИХ МНОЖИН ТА ЛІНГВІСТИЧНИХ ТЕРМІВ

Анотація. У статті представлено метод формування нечітких асоціативних правил із зваженими атрибутами з бази даних (БД) SIEM – системи для поповнення її бази знань (БЗ) з метою більш ефективного виявлення нею кіберінцидентів, які виникають в ході функціонування спеціальних інформаційно – комунікаційних систем (СІКС). Розглянуто проблеми, які знижують ефективність застосування існуючих методів для вирішення задачі формування асоціативних правил на основі аналізу інформації, що знаходиться у БД систем кіберзахисту. Проведено аналіз публікацій, присвячених методам, в яких здійснено спроби усунення наведених проблем. Сформульовано основну ідею усунення недоліків, що є властивими відомим методам, яка полягає в знаходженні компромісу між зменшенням часу роботи обчислювального алгоритму, який реалізує на практиці метод та зменшенням інформаційних втрат в результаті його роботи. Запропоновано удосконалений метод пошуку асоціативних правил з БД SIEM – систем, в основу якого покладено теорію нечітких множин та лінгвістичних термів. Сформульовано задачу пошуку нечітких асоціативних правил із зваженими атрибутами. Наведено математичний апарат, який покладено в основу реалізації метода. Запропоновано алгоритм пошуку частих наборів елементів, що включають значення ознак кіберінцидентів та класів, до яких вони відносяться та який реалізує перший етап запропонованого методу. Проаналізовано особливості структури тестових наборів даних, які використовуються для навчання та тестування систем кіберзахисту та на основі його результатів зроблено висновок про можливість удосконалення розглянутого алгоритму. Наведено графічну ілюстрацію ідеї удосконалення алгоритму пошуку частих наборів елементів та описано суть його удосконалення. Запропоновано удосконалений алгоритм пошуку частих наборів елементів розглянутого методу та наведено його основні переваги.

Ключові слова: кіберзахист; кіберінцидент; SIEM – система; теорія нечітких множин; інтелектуальний аналіз даних; нечіткі асоціативні правила.

ВСТУП

Постановка проблеми. Значне зростання кількості кібератак, їх різновидів та підвищення складності потребує нових зусиль щодо створення ефективних систем кіберзахисту спеціальних інформаційно – комунікаційних систем.

Основною сучасної системи кіберзахисту СІКС є SIEM – система [1], причому найбільш розповсюдженими методами, що застосовуються в SIEM – системах для виявлення (ідентифікації) кіберінцидентів, які відбуваються під час функціонування СІКС є правило – орієнтовані методи, в основу яких покладено механізм логічного виводу на основі правил – продукцій, що знаходяться у БЗ інтелектуальної SIEM – системи.

Ключовим етапом методики формування нечітких асоціативних правил із зваженими атрибутами з БД SIEM – системи є знаходження, так званих, частих наборів елементів даних, які мають значення нечіткої підтримки не менше, ніж задане аналітиком безпеки порогове значення. У теперішній час, для рішення даної задачі запропоновано застосування декількох методів, наведених у [2] – [5].

Однак існують декілька проблем, які можуть зробити застосування наведених методів не завжди ефективним та які, потребують вирішення.

Проблема “мінімальної підтримки”, яка пов’язана з тим, що під час розбиття діапазонів значень чисельних атрибутів на велику кількість малих інтервалів, існує загроза зменшення підтримки для деяких окремих інтервалів. Виходячи з цього, розбиття, ймовірно, має відбуватися з використанням великих інтервалів.

Проблема “мінімальної вірогідності”, яка безпосередньо пов’язана з попередньою та яка виникає внаслідок розбиття значень чисельних атрибутів на великі інтервали, що приводить до того, що вірогідність деяких правил, може зменшитися.

Проблема “часу виконання”, яка виникає внаслідок того, що при великому числі значень чисельного атрибуту, у середньому, виникає $O(n^2)$ варіантів розбиття діапазону його значень, а це потребує значного часу для роботи обчислювального алгоритму.

Проблема “великої кількості правил”, яка пов’язана з тим, що якщо значення деякого атрибуту має мінімальну підтримку на деякому інтервалі значень, то буде знайдено велику кількість непридатних для використання на практиці правил.

Таким чином, рішення задачі пошуку асоціативних правил з чисельними та категорійними атрибутами, які містять ознаки кіберінцидентів та інформація про які накопичується в БД SIEM – системи, полягає у знаходженні компромісу між зменшенням часу роботи обчислювального алгоритму (завдяки розбиттю значень чисельних атрибутів на більш великі інтервали), з одного боку, та зменшенням інформаційних утрат (що виникають при розбитті на більшу кількість менших інтервалів), з отриманням великої кількості непридатних для використання на практиці правил, з іншого боку.

Аналіз останніх досліджень і публікацій. Вирішенню даних проблем присвячено певну кількість публікацій.

Так у [6] запропоновано метод регулювання розмірів інтервалів розбиття, шляхом комбінування суміжних інтервалів та значень. Проте, дане розбиття приводить до виникнення інших проблем, які пов’язані з необхідністю емпіричного завдання аналітиком безпеки граничних значень для підтримки інтервалів чисельних атрибутів з метою проведення процесу їх квантування.

Рішенню однієї з таких проблем – “проблеми граничних значень”, яка пов’язана з тим, що деякі алгоритми або ігнорують, або занадто виділяють значення навколо границь інтервалів у процесі їх визначення, присвячено публікацію [7]. Крім того, застосування чітких границь інтервалів не завжди є інтуїтивним для розуміння аналітиком безпеки.

На рис. 1. наведено приклад виникнення “проблеми границь” під час розбиття значень чисельного атрибуту “Кількість звернень до Інтернет – ресурсів для оновлення”,

ISSN 2663 - 4023

який використовується для виявлення кіберінциденту, пов'язаного з кібератакою JS (HTML)/ScrInject [8].

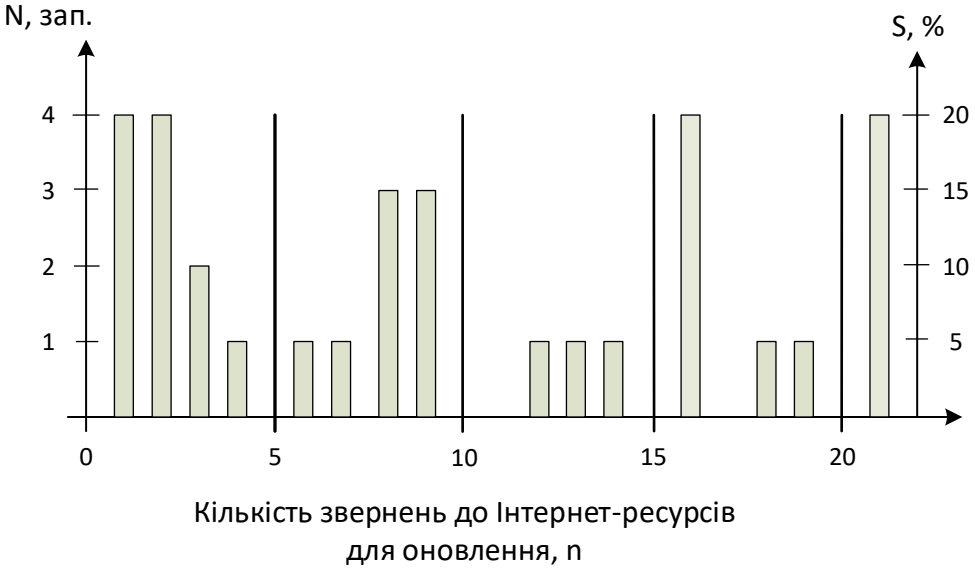

Рис. 1. Приклад виникнення "проблеми границь" під час розбиття значень чисельного атрибуту "Кількість звернень до Інтернет – ресурсів для оновлення"

З рис. 2. видно, що рівномірне розбиття значень даного атрибуту, наприклад, на інтервали: [0,4], [5,9], [10,14], [15,19] і так далі, може привести до того, що у випадку, коли порогове значення підтримки більше за 10%, набір елементів, який відповідає числу звернень до Інтернет – ресурсів в інтервалі від 10 до 14, взагалі не буде присутнім у множині частих наборів елементів. Проте не важко помітити, що значення атрибуту яке відповідає числу звернень – 16 та яке знаходиться на межі даного інтервалу, має високе значення підтримки. Доцільно було б розширити розглянутий інтервал, шляхом включення до нього даного значення.

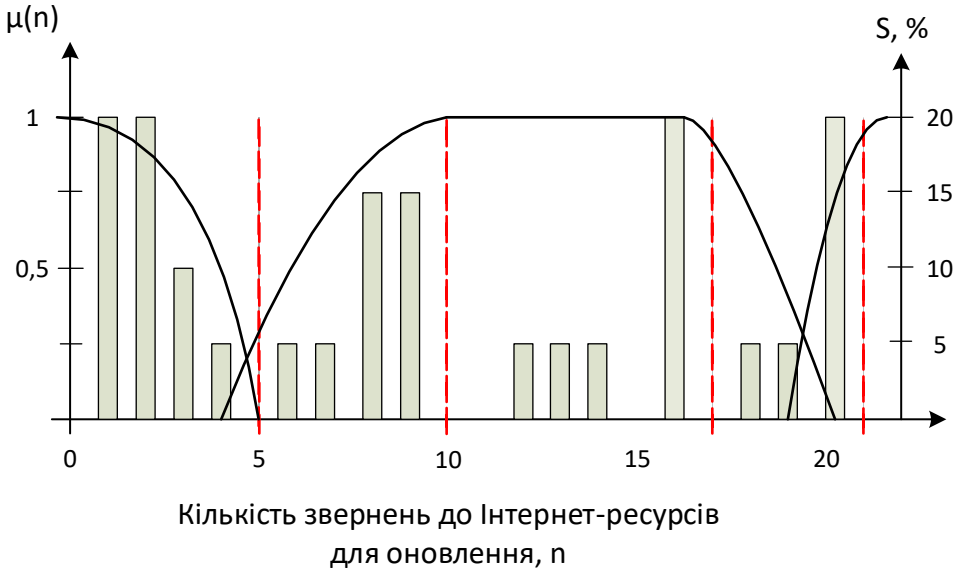

Рис. 2. Приклад рішення "проблеми границь" на основі теорії нечітких множин

Слід також зауважити, що представлення значень чисельних та категорійних атрибутів за допомогою лінгвістичних термів, дозволяє формувати нечіткі асоціативні правила, які є більш придатними для аналізу та зрозумілими (легко інтерпретуємими) для експертів та аналітиків безпеки з метою формування своєчасних та обґрунтованих рішень при вирішенні задач виявлення кіберінцидентів.

Не зважаючи на це, основним недоліком методів, що запропоновано у розглянутих публікаціях, залишається багаторазове сканування БД для підрахунку нечіткої підтримки наборів елементів – кандидатів до нечітких асоціативних правил, яке вимагає значних витрат часових та обчислювальних ресурсів.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Виклад основного матеріалу дослідження. Для усунення наведених недоліків у даній роботі пропонується удосконалений метод пошуку нечітких асоціативних правил на основі аналізу даних з БД SIEM – системи.

Дано: *DBI* – набір записів БД про кіберінциденти, які накопичуються SIEM – системою, що складаються з атрибутів $O = \{o_1, o_2, ..., o_n\}$, де o_j, $j = \overline{1,n}$ – чисельний або категорійний атрибут, який містить ознаку кіберінциденту, що відбуваються в SIEM – системі під час функціонування СІКС. Для будь – якого запису $t \in DBI$, $t[o_j]$ – має деяке значення *v* (значення ознаки кіберінциденту), котре приймає атрибут $o_j \in O$ у запису *t*.

Нехай будь – який чисельний атрибут $o_j \in O$ є заданим на домені $dom(o_j) = [\underline{v}, \overline{v}] \subseteq R$. Тоді $\left\{ l_{o_j k_{o_j}} \right\}, k_{o_j} = \overline{1, K_{o_j}};\ j = \overline{1,n}$ – множина лінгвістичних термів, які відносяться до чисельного атрибута $o_j \in O$. Кожний лінгвістичний терм $l_{o_j k_{o_j}}$ представляється нечіткою множиною $L_{o_j k_{o_j}}$, що є заданою на домені $dom(o_j)$ з функцією належності $\mu_{L_{o_j k_{o_j}}}(v) : dom(o_j) \to [0,\ 1]$.

Відповідно, степінь належності будь – якого значення *v* до деякого лінгвістичного терму $l_{o_j k_{o_j}}$ характеризується функцією належності $\mu_{L_{o_j k_{o_j}}}(v)$.

У випадку будь – якого категорійного атрибута $o_j \in O;\ j = \overline{1,n}$, що є визначеним на домені $dom(o_j) = \{v_1, v_2, ..., v_m\}$, задається множина лінгвістичних термів $l_{o_j k_{o_j}}, k_{o_j} = \overline{1, K_{o_j}};\ j = \overline{1,n}$, де кожний лінгвістичний терм $l_{o_j k_{o_j}}$ представляється за допомогою нечіткої множини $L_{o_j k_{o_j}}$.

Множина атрибутів *O* представляється за допомогою множини лінгвістичних термів: $l = \left\{ l_{o_j k_{o_j}} \mid j = \overline{1,n};\ k_{o_j} = \overline{1, K_{o_j}} \right\}$, де K_{o_j} – кількість лінгвістичних термів атрибута $o_j \in O;\ j = \overline{1,n}$.

ISSN 2663 - 4023

У свою чергу, множина лінгвістичних термів представляється за допомогою нечітких множин: $L = \left\{ L_{o_j k_{o_j}} \mid j = \overline{1,n};\ k_{o_j} = \overline{1,K_{o_j}} \right\}$.

Степінь приналежності значень атрибута $o_j \in O,\ j = \overline{1,n}$ у запису $t \in DBI$ до лінгвістичного терму $l_{o_j k_{o_j}} \in l,\ k_{o_j} = \overline{1,K_{o_j}}$, який представляється нечіткою множиною $L_{o_j k_{o_j}} \in L$, задається за допомогою функції належності $\mu_{L_{o_j k_{o_j}}}(t[o_j])$.

Необхідно: знайти усі нечіткі асоціативні правила із зваженими атрибутами виду:

$$\left\langle X = \left\{ o_1, o_2, ..., o_p \right\}, A = \left\{ l_{o_1 k_{i_1}}, l_{o_2 k_{i_2}}, ..., l_{o_p k_{i_p}} \right\} \right\rangle \Rightarrow \left\langle Y = \left\{ o_{p+1}, o_{p+2}, ..., o_n \right\}, B = \left\{ l_{o_{p+1} q_{o_{p+1}}}, l_{o_{P+2} q_{o_{p+2}}}, ..., l_{o_n q_{o_n}} \right\} \right\rangle, \quad (1)$$

де набори елементів $X \subset O, Y \subset O, X \bigcap Y = \varnothing$;

$l_{o_j k_{o_j}} \in l, l_{o_p q_{o_p}} \in l,\ j \neq p$ – лінгвістичні терми, які є заданими на нечітких множинах $L_{o_j k_{o_j}} \in L$ та $L_{o_p q_{o_p}} \in L$ відповідно, для яких зважена нечітка підтримка (2) та зважена нечітка вірогідність (3) повинні бути не меншими, ніж наперед задані аналітиком безпеки граничні значення, що називаються мінімальною підтримкою $FS_{\min}$ та мінімальною вірогідністю $FC_{\min}$:

$$M = \left\{ \langle X, A \rangle \Rightarrow \langle Y, B \rangle \mid WFC_{(\langle X,A \rangle, \langle Y,B \rangle)} \geq FC_{\min}, WFS_{\langle X,A \rangle} \geq FS_{\min} \right\}.$$

Зважена нечітка підтримка набору елементів <*X*,*A*>, позначається, як $WFS_{<X,A>}$ та розраховується наступним чином [9]:

$$WFS_{<X,A>} = \frac{\sum\limits_{t_h \in D^F} \prod\limits_{o_j \in X} \mu_{L_{o_j} \in A}((t_h[o_j]) \cdot \omega(o_j.l_{o_j k}))}{|DBI^F|}, \quad (2)$$

де $\omega(o_j.l_{o_j k})$ – ваговий коефіцієнт, який характеризує відносну важливість пари $< o_j.l_{o_j k} >$, $o_j \in O$, $j = \overline{1,n}$, $l_{o_j k} \in l$, $k = \overline{1,K_{o_j}}$ та є степеню проявлення інтегральної властивості “важливість атрибута”;

Зважена нечітка вірогідність правила (1) позначається, як $WFC_{<<X,A>,<Y,B>>}$ та розраховується наступним чином [10] – [14]:

$$WFC_{<<X,A>,<Y,B>} = \frac{\sum\limits_{t_h \in D^F} \prod\limits_{o_j \in Z} \mu_{L_{o_j} \in C}((t_h[o_j]) \cdot \omega(o_j.l_{o_j k}))}{\sum\limits_{t_h \in D^F} \prod\limits_{o_j \in X} \mu_{L_{o_j} \in A}((t_h[o_j]) \cdot \omega(o_j.l_{o_j k}))}, \quad (3)$$

де $Z = X \bigcup Y,\ C = A \bigcup B,\ X \bigcap Y = \varnothing,\ A \bigcap B = \varnothing$.

Рішення сформульованої задачі здійснюється у два етапи. На першому знаходяться всі, так звані, часті набори елементів, зважена нечітка підтримка яких є не меншою, ніж наперед задане аналітиком безпеки порогове значення $FS_{\min}$. Відповідно

на другому етапі, із знайдених частих наборів елементів формуються нечіткі асоціативні правила, зважена нечітка вірогідність яких є не меншою за наперед задане порогове значення аналітиком безпеки $FC_{\min}$.

Уведемо допоміжну множину $\overline{C_s}$ – s – елементних наборів, кожний член якої має вигляд:

$$\overline{C_s} = \left\{ \forall \bar{c} \in \overline{C_s} \mid \bar{c} = \left\langle t.TID, \left\{ o_1.l_{o_1 k_{o_1}}, o_2.l_{o_2 k_{o_2}}, \ldots, o_s.l_{o_s k_{o_s}} \mid \mu^{s}{}_{o_1, o_2, \ldots, o_s} \right\} \right\rangle \wedge \right.$$
$$\left. \wedge \left(\forall \left\langle o_1 l_{k_{o_1}}, o_2 l_{k_{o_2}}, \ldots, i_{s-1} l_{o_{i_{s-1}}} \right\rangle \in F_{s-1} \right) \right\}, \tag{4}$$

де $j = \overline{1,n}$; $k_{o_j} = \overline{1, K_{o_j}}$; $s = \overline{1,S}$; $o_1 \neq o_2 \neq \ldots \neq o_s$;

$t.TID$ – номер ідентифікатора транзакції t БД SIEM – системи DBI^F, у якій міститься упорядкований s – елементний набір $\left\langle o_1 l_{k_{o_1}}, o_2 l_{k_{o_2}}, \ldots, i_o l_{k_{o_s}} \right\rangle$ із значенням функції належності (5):

$$\mu^{s}{}_{o_1, o_{o2}, \ldots, o_s} = \prod_{j=1}^{s} \left(\mu_{L_{o_j k_{o_j}}} t.TID[o_j] \cdot \omega(o_j.l_{o_j k}) \right), \tag{5}$$

де $\omega(o_j.l_{o_j k})$ – ваговий коефіцієнт, який характеризує відносну важливість пари $< o_j.l_{o_j k} >$, $o_j \in O$, $j = \overline{1,n}$, $l_{o_j k} \in l$, $k = \overline{1, K_{o_j}}$.

Тоді БД DBI^F у відповідності до виразів (4)–(5) може бути представленою за допомогою множини записів, які складаються з унікальних ідентифікаторів транзакцій $t.TID$ та 1 – елементних наборів виду $\left\langle o_j l_{k_{o_j}} \mid \mu_{o_j k_{o_j}} \cdot \omega(o_j.l_{o_j k}) \right\rangle$, які в них містяться:

$$DBI^* = \overline{C_1} = \left\{ \forall \bar{c} \in \overline{C_1} \mid \bar{c} = \left\langle t.TID, \left\{ \left\langle o_1.l_{o_1 1_{o_1}} \mid \mu_{L_{o_1 1_{o_1}}} t.TID[o_1] \cdot \omega(o_1.l_{o_1 1_{o_1}}) \right\rangle, \right. \right.$$
$$\left\langle o_1.l_{o_1 2_{o_1}} \mid \mu_{L_{o_1 2_{o_1}}} t.TID[o_1] \cdot \omega(o_1.l_{o_1 2_{o_1}}) \right\rangle, \cdots, \left\langle o_1.l_{o_1 K_{o_1}} \mid \mu_{L_{o_1 K_{o_1}}} t.TID[o_1] \cdot \omega(o_1.l_{o_1 k_{o_1}}) \right\rangle,$$
$$\left\langle o_2.l_{o_2 1_{o_2}} \mid \mu_{L_{o_2 1_{o_2}}} t.TID[o_2] \cdot \omega(o_2.l_{o_2 1_{o_2}}) \right\rangle, \left\langle o_2.l_{o_2 2_{o_2}} \mid \mu_{L_{o_2 2_{o_2}}} t.TID[o_2] \cdot \omega(o_2.l_{o_2 2_{o_2}}) \right\rangle, \cdots, \tag{6}$$
$$\left\langle o_2.l_{o_2 K_{o_2}} \mid \mu_{L_{o_2 K_{o_2}}} t.TID[o_2] \cdot \omega(o_2.l_{o_2 k_{o_2}}) \right\rangle, \cdots, \left\langle o_n.l_{o_n 1_{o_n}} \mid \mu_{L_{o_n 1_{o_n}}} t.TID[o_n] \cdot \omega(o_n.l_{o_n 1_{o_n}}) \right\rangle,$$
$$\left. \left. \left\langle o_n.l_{o_n 2_{o_n}} \mid \mu_{L_{o_n 2_{o_n}}} t.TID[o_n] \cdot \omega(o_n.l_{o_n 2_{o_n}}) \right\rangle, \cdots, \left\langle o_n.l_{o_n K_{o_n}} \mid \mu_{L_{o_n K_{o_n}}} t.TID[o_n] \cdot \omega(o_n.l_{o_n k_{o_n}}) \right\rangle \right\} \right\} \right\}.$$

або з використанням операцій $\bigcup$ та $\bigcap$ наступним чином:

$$DBI^* = \overline{C_1} =$$
$$\left\{ \forall \bar{c} \in \overline{C_1} \mid \bar{c} = \left\langle t.TID, \left\{ \bigcup_{j=1}^{n} \left\{ \bigcup_{k_{oj}=1}^{K_{o_j}} < o_j.l_{o_j k_{o_j}} \mid \mu_{L_{o_j k_{o_j}}} t.TID[o_j] \cdot \omega(o_j.l_{o_j k_{o_j}}) > \right\} \right\} \right\rangle \right\}. \tag{7}$$

ISSN 2663 - 4023

Таким чином, значення зваженої нечіткої підтримки $WFS_{<o_j.l_{o_jk_{o_j}}>}$ може бути отриманим, шляхом обробки елементів множини $\overline{C_1}$.

Для $s>1$ допоміжна множина $\overline{C_S}$ має вигляд (7), де кожний член $\overline{C_s}$ складається з унікального ідентифікатора транзакції *t.TID* та упорядкованих s – елементних наборів $\left\langle o_1 l_{k_{o_1}}, o_2 l_{k_{o_2}}, ..., o_s l_{k_{o_s}} \right\rangle$, які є потенційно частими, з відповідними значеннями функцій приналежності $\mu^s{}_{o_1,o_2,...,o_s}$, що розраховуються за виразом (5). Дана підмножина наборів у $\overline{C_s}$ з однаковими *TID,* тобто які містяться в одній транзакції, називається записом. Якщо транзакція не містить жодного s – елементного набору, то $\overline{C_s}$ не буде мати запису для цієї транзакції. Таким чином кількість записів у $\overline{C_s}$ може бути меншою, ніж у БД. Більш того, кожний запис може бути меншим, ніж транзакція, яка йому відповідає, у наслідок того, що у транзакції буде міститися невелике число наборів.

Це надає можливість розрахувати значення зваженої нечіткої підтримки $WFS\left\langle o_1 l_{k_{o_1}}, o_2 l_{k_{o_2}}, ..., o_s l_{k_{o_s}} \right\rangle$ шляхом обробки множини $\overline{C_S}$, для кожного s – елементного набору $\left\langle o_1 l_{k_{o_1}}, o_2 l_{k_{o_2}}, ..., o_s l_{k_{o_s}} \right\rangle$ та сформувати множину s – елементних наборів – кандидатів $C_s = \left\{ \forall c \in C_s \mid c = \left\langle o_1 l_{k_{o_1}}, o_2 l_{k_{o_2}}, ..., o_s l_{k_{o_s}}, FS_c \right\rangle \right\}$.

Причому, до множини частих s – елементних наборів F_s, включаються тільки ті набори з C_s, які задовольняють наступній вимогі:

$$F_s = \left\{ \forall f \in F_s \mid f = \left\langle o_1 l_{o_1 k_{o_1}}, o_2 l_{o_2 k_{o_2}}, ..., o_s l_{o_s k_{o_s}} \in C_s, FS_{f_s} \right\rangle \wedge WFS_{f_s} \geq FS_{min} \right\}. \quad (8)$$

На другому етапі, нечіткі асоціативні правила із зваженими атрибутами виду (1) формуються з частих наборів елементів f множини F_s, на основі розрахунку їхньої зваженої нечіткої вірогідності WFC, при умові:

$$RULE = \left\{ \forall r \in RULE \mid r = \left\langle \langle X, A \rangle, \langle Y, B \rangle \right\rangle \wedge \left(Z = X \cup Y,\ C = A \cup B,\ X \cap Y = \varnothing,\ A \cap B = \varnothing \right) \wedge WFC_{\langle \langle X,A \rangle, \langle Y,B \rangle \rangle} \geq FC_{min} \right\} \quad (9)$$

Алгоритм пошуку частих наборів елементів. Покрокова реалізація першого етапу методу може бути представлена наступним алгоритмом.

Дано: F_s – множина частих s – елементних наборів;

C_s – множина s – елементних наборів – кандидатів;

$\overline{C_s}$ – допоміжна множина s – елементних наборів;

FS_{min} – поріг мінімальної підтримки;

FC_{min} – поріг мінімальної вірогідності.

Крок 1. Присвоїти $s=1$ та сформувати допоміжну множину $\overline{C_1}$ – 1 – елементних наборів.

ISSN 2663 - 4023

Крок 2. Сформувати множину 1 – елементних наборів – кандидатів C_1 та для кожного набору, шляхом обробки множини $\overline{C_1}$, обчислити значення зваженої нечіткої підтримки WFS.

Крок 3. Сформувати множину F_1 – частих наборів елементів, шляхом включення до неї тільки тих наборів з C_1, які мають значення зваженої нечіткої підтримки WFS не меншим, ніж наперед задане аналітиком безпеки значення мінімальної підтримки $FS_{\min}$.

Крок 4. $s = s + 1$.

Крок 5. Якщо не вдається створити s – елементні набори, то завершити алгоритм, інакше виконати наступний крок.

Крок 6. Створити множину s – елементних наборів – кандидатів C_s із частих наборів елементів F_{s-1}, які було знайдено на $(s-1)$ – й ітерації. Кожний кандидат $c \in C_s$ буде формуватися шляхом додавання до частого $(k-1)-$елементного набору останнього, більшого за порядком, елемента з іншого частого $(k-1)-$елементного набору, причому всі $(k-2)$ елементи обох наборів мають бути однаковими.

Крок 7. З побудованої множини C_s вилучити набори, якщо хоча б одна з їх $(k-1)-$підмножин не є частою, тобто є відсутньою у множині F_{s-1}.

Крок 8. Шляхом обробки множин $\overline{C_{s-1}}$ і C_s сформувати допоміжну множину $\overline{C_s}$: $\overline{C_s} = \left\{ c \in C_s \mid (c - c[s]) \in \overline{C_{s-1}} \wedge (c - c[s-1]) \in \overline{C_{s-1}} \right\}$ та для кожного з її наборів обчислити значення функції належності $\mu^s{}_{o_1,o_2,\ldots,o_s}$ за виразом (5).

Крок 9. Для кожного набору множини C_s, шляхом обробки множини $\overline{C_s}$, обчислити значення зваженої нечіткої підтримки WFS.

Крок 10. Сформувати множину F_s – частих наборів, шляхом включення до неї тільки тих наборів з C_s, які мають значення зваженої нечіткої підтримки WFS не меншим, за наперед задане аналітиком безпеки значення мінімальної підтримки $FS_{\min}$.

Крок 11. Перейти до кроку 4.

Результатом роботи алгоритму є об'єднання усіх множин F_s для усіх значень s: $F = \bigcup_s F_s$.

Слід зауважити, що наведений алгоритм є універсальним та може бути застосованим для рішення прикладних задач у багатьох сферах. Проте, аналіз структури наборів даних [15] для навчання та тестування систем виявлення кібератак та кіберінцидентів, дозволяє зробити висновок про можливість удосконалення наведеного алгоритму (рис. 3.).

Фактично, ліва частина нечіткого асоціативного правила виду (1), включає ознаки кіберінциденту, а права частина правила – його клас.

Цим можна скористатися та розбити початкову допоміжну множину 1 – елементних наборів – кандидатів $\overline{C_1}$ на дві підмножини: підмножину $\overline{CO_1} \subset \overline{C_1}$, що включає ознаки кіберінцидентів та підмножину $\overline{CC_1} \subset \overline{C_1}$, яка включає класи кіберінцидентів.

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 3 (19), 2023
ISSN 2663 - 4023

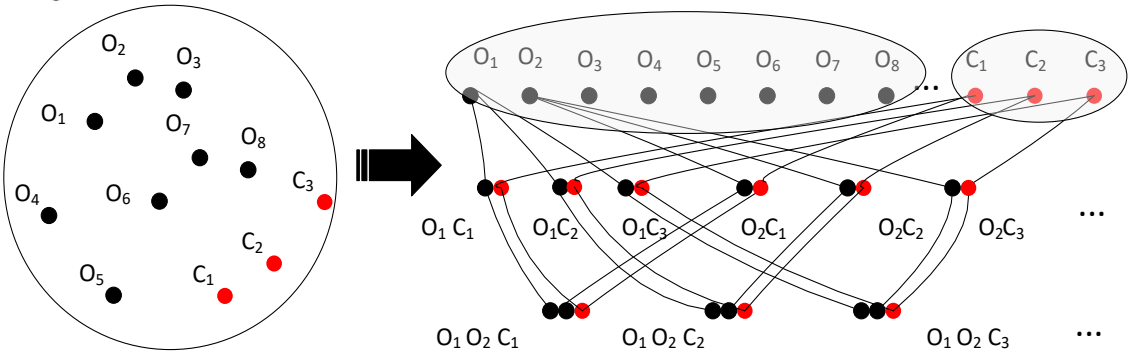

Рис. 3. Графічна ілюстрація удосконаленого методу

Після цього сформувати множину 1 – елементних наборів – кандидатів C_1 та для кожного набору, шляхом обробки множини $\overline{C_1}$, яка включає підмножини $\overline{CO_1} \subset \overline{C_1}$ та $\overline{CC_1} \subset \overline{C_1}$, обчислити значення зваженої нечіткої підтримки WFS для кожного з її елементів.

Тоді, формування множини F_1 – частих наборів елементів, здійснити шляхом включення до неї тільки тих наборів з C_1, які мають значення зваженої нечіткої підтримки WFS не меншим, ніж наперед задане аналітиком безпеки значення мінімальної підтримки $FS_{\min}$.

Відповідно, формування допоміжної множини 2 – елементних наборів – кандидатів $\overline{C_2}$, здійснити наступним чином: першим елементом набору є ознака кіберінциденту з множини C_1 при умові, що $c[1] \in CO_1$, а другим, останнім його елементом – клас кіберінциденту: $c[2] \in CC_1$ (10):

$$\overline{C_2} = \left\{ c \in C_2 \mid (c[1]) \in \overline{CO_1} \wedge (c[2]) \in \overline{CC_1} \right\} \quad (10)$$

Далі, кожний кандидат $c \in C_s$ буде формуватися шляхом додавання до частого $(k-1)$ – елементного набору останнього, більшого за порядком елемента з іншого частого $(k-1)$ – елементного набору, причому всі $(k-2)$ елементи обох наборів мають бути однаковими, а останнім за порядком елементом отриманого набору має бути клас кіберінциденту, який є однаковим в кожному з $(k-1)$ – елементних наборів, що приймають участь у формуванні набору – кандидату $c \in C_s$ (рис. 4.).

$$\overline{C_s} = \left\{ c \in C_s \mid (c - c[s-1]) \in \overline{C_{s-1}} \wedge (c - c[s-2]) \in \overline{C_{s-1}} \wedge \{c[1] \ldots c[s-1]\} \in \overline{CO_1} \wedge c[s] \in \overline{CC_1} \right\} \quad (11)$$

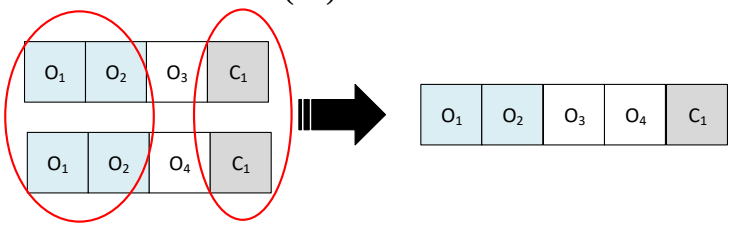

Рис. 4. Приклад формування 4 – елементного набору елементів на основі двох 3 – елементних

ISSN 2663 - 4023

Удосконалений алгоритм пошуку частих наборів елементів.

Дано: F_s – множина частих s – елементних наборів;

C_s – множина s – елементних наборів – кандидатів;

$\overline{C_s}$ – допоміжна множина s – елементних наборів;

$\overline{CO_s} \subset \overline{C_s}$ – підмножина s – елементних наборів, що включає ознаки кіберінцидентів;

$\overline{CC_s} \subset \overline{C_s}$ – підмножина s – елементних наборів, що включає класи кіберінцидентів;

FS_{min} – поріг мінімальної підтримки;

FC_{min} – поріг мінімальної вірогідності.

Крок 1. Присвоїти $s = 1$ та сформувати допоміжну множину $\overline{C_1}$ – 1 – елементних наборів та розбити її на дві підмножини: підмножину $\overline{CO_1} \subset \overline{C_1}$ та $\overline{CC_1} \subset \overline{C_1}$.

Крок 2. Сформувати множину 1 – елементних наборів – кандидатів C_1 та для кожного набору, шляхом обробки множини $\overline{C_1}$, яка включає підмножини $\overline{CO_1} \subset \overline{C_1}$ та $\overline{CC_1} \subset \overline{C_1}$, обчислити значення зваженої нечіткої підтримки WFS для кожного з її елементів.

Крок 3. Сформувати множину F_1 – частих наборів елементів, шляхом включення до неї тільки тих наборів з C_1, які мають значення зваженої нечіткої підтримки WFS не меншим, ніж наперед задане аналітиком безпеки значення мінімальної підтримки $FS_{\min}$.

Крок 4. Сформувати допоміжну множину 2 – елементних наборів – кандидатів $\overline{C_2}$, причому першим елементом набору є ознака кіберінциденту з множини C_1 при умові, що $c[1] \in \overline{CO_1}$, а другим, останнім його елементом – клас кіберінциденту: $c[2] \in \overline{CC_1}$ (10).

Крок 5. $s = s + 1$.

Крок 6. Якщо не вдається створити s – елементні набори, то завершити алгоритм, інакше виконати наступний крок.

Крок 7. Створити множину s – елементних наборів – кандидатів C_s із частих наборів елементів F_{s-1}. Кожний кандидат $c \in C_s$ формується шляхом додавання до частого $(k-1)$ – елементного набору останнього, більшого за порядком елемента з іншого частого $(k-1)$ – елементного набору, причому всі $(k-2)$ елементи обох наборів мають бути однаковими, а останнім за порядком елементом отриманого набору має бути клас кіберінциденту, який є однаковим в кожному з $(k-1)$ – елементних наборів, що приймають участь у формуванні набору – кандидату $c \in C_s$.

Крок 8. З побудованої множини C_s вилучити набори, якщо хоча б одна з їх $(k-1)$ – підмножин не є частою, тобто є відсутньою у множині F_{s-1}.

Крок 8. Шляхом обробки множин $\overline{C_{s-1}}$ і C_s сформувати допоміжну множину $\overline{C_s}$ (11) та для кожного з її наборів обчислити значення функції належності $\mu^s{}_{o_1,o_2,\ldots,o_s}$ за виразом (5).

Крок 9. Для кожного набору множини C_s, шляхом обробки множини $\overline{C_s}$, обчислити значення зваженої нечіткої підтримки WFS.

Крок 10. Сформувати множину F_s – частих наборів, шляхом включення до неї тільки тих наборів з C_s, які мають значення зваженої нечіткої підтримки WFS не меншим, за наперед задане аналітиком безпеки значення мінімальної підтримки $FS_{\min}$.

Крок 11. Перейти до кроку 5.

Результатом роботи алгоритму є об'єднання усіх множин F_s для усіх значень s: $F = \bigcup_s F_s$.

Не важко помітити, що удосконалений метод дозволяє скоротити простір пошуку під час формування множини частих наборів – кандидатів і тим самим підвищити швидкодію алгоритму, який реалізує метод та знизити обчислювальні ресурси для його застосування.

Крім того, завдяки тому, що останнім елементом сформованих частих наборів елементів є клас кіберінциденту, то значно спрощується другий етап методу, який полягає у формуванні нечітких асоціативних правил із зваженими атрибутами виду (1) з частих наборів елементів F множини F_s, на основі розрахунку їхньої зваженої нечіткої вірогідності WFC, при умові виконання (9).

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Застосування запропонованої в роботі удосконаленого методу формування нечітких асоціативних правил із зваженими атрибутами з БД SIEM – системи для виявлення кіберінцидентів, що відбуваються в ході функціонування СІКС, дозволяє на практиці підвищити ефективність застосування SIEM – систем за рахунок покращення таких їхніх характеристик, як оперативність формування асоціативних правил для їхнього завантаження в БЗ системи з метою її адаптування до нових типів кіберінцидентів з високою точністю їх виявлення.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1 Subach, I., Mykytiuk, A., Kubrak, V. (2019). Architecture and functional model of a perspective proactive intellectual siem for cyber protection of objects of critical infrastructure. *Collection "Information technology and security"*, *7*(2), 208–215. https://doi.org/10.20535/2411 – 1031.2019.7.2.190570

2 Horng, S. – J., Su, M. – Y., Chen, Y. – H., Kao, T. – W., Chen, R. – J., Lai, J. – L., Perkasa, C. D. (2011). A novel intrusion detection system based on hierarchical clustering and support vector machines. *Expert Systems with Applications*, *38*(1), 306–313. https://doi.org/10.1016/j.eswa.2010.06.066

3 Mbikayi, H. K. (2012). An Evolution Strategy Approach toward RuleSet Generation for Network Intrusion Detection Systems (IDS). *International Journal of Soft Computing and Engineering*, *2*(5), 1–5.

4 Subach, I., Fesokha, V., Fesokha, N. (2017b). Analysis of existing solutions for preventing invasion in information and telecommunication networks. *Collection "Information technology and security"*, *5*(1), 29–41. https://doi.org/10.20535/2411 – 1031.2017.5.1.120554

5 Lappas, T., Pelechrinis, K. (2007). Data mining techniques for (network) intrusion detection systems. *Department of Computer Science and Engineering UC Riverside, Riverside CA*, (92521). https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=720ec75b12f2e08c5297251e29401c337c251621

6 Srikant, R., Agrawal, R. (1996). Mining quantitative association rules in large relational tables. *ACM SIGMOD Record*, *25*(2), 1–12. https://doi.org/10.1145/235968.233311

7 Gyenesei, A. (2001). A fuzzy approach for mining quantitative association rules. *Acta Cybernetica*, *15*(2), 305–320.

8 Субач, І. Ю., Здоренко, Ю. М., Фесьоха, В. В. (2018). Методика виявлення кібератак типу JS (HTML)/Scrinject на основі застосування математичного апарату теорії нечітких множин. *Збірник наукових праць [Військового інституту телекомунікацій та інформатизації]*, (4), 125 – 131.

9 Gyenesei, A. (2000). *Fuzzy partitioning of quantitative attribute domains by a cluster goodness index*. Turku Centre for Computer Science.

10 Gyenesei, A. (2000). Mining weighted association rules for fuzzy quantitative items. In *Principles of Data Mining and Knowledge Discovery: 4th European Conference, PKDD 2000 Lyon, France, September 13–16, 2000 Proceedings 4* (pp. 416 – 423). Springer Berlin Heidelberg.

11 Субач, І. Ю. (2005). Знаходження нечітких асоціативних правил у реляційних базах даних телекомунікаційного підприємства. *Зв'язок*, (3), 54–57.

12 Герасимов, Б., Субач, І. (2005). Здобуття нечітких асоціативних правил на основі аналізу інформації у базах даних інформаційно – аналітичних систем. *Вісник національного технічного університету "Поділля"*, (4), 266–270.

13 Au, W. H., Chan, K. C. (1998, May). An effective algorithm for discovering fuzzy rules in relational databases. In *1998 IEEE International Conference on Fuzzy Systems Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36228)* (Vol. 2, pp. 1314 – 1319). IEEE.

14 Chan, K. C., Au, W. H. (2001). Mining fuzzy association rules in a database containing relational and transactional data. *Data mining and computational intelligence*, 95 – 114.

15 Newman, D. (1999). KDD Cup'99 Data Sets. *Retrieved February*, *7*, 2010.

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 3 (19), 2023
ISSN 2663 - 4023

Ihor Subach
doctor of technical science, associate professor, head of department
Institute of special communications and information security National technical university of Ukraine
Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, Ukraine
ORCID ID: 0000 – 0002 – 9344 – 713X
igor_subach@ukr.net

Artem Mykytiuk
postgraduate student
Institute of special communications and information security National technical university of Ukraine
Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, Ukraine
ORCID ID: 0000 – 0002 – 8307 – 9978
mukuta8888@gmail.com

METHOD OF FORMING ASSOCIATIVE RULES FROM THE SIEM DATABASE BASED ON FUZZY SET THEORY AND LINGUISTIC TERMS

Abstract. The article presents a method of forming fuzzy associative rules with weighted attributes from the database (DB) of the SIEM to supplement its knowledge base (KB) in order to more effectively detect cyber incidents that occur during the operation of special information and communication systems (SICS). The problems that reduce the effectiveness of the application of existing methods for solving the problem of forming associative rules based on the analysis of information located in the database of cyber protection systems are considered. An analysis of publications devoted to methods in which attempts were made to eliminate these problems was made. The basic idea of eliminating the shortcomings inherent in known methods is formulated, which consists in finding a compromise between reducing the time of the computing algorithm that implements the method in practice and reducing information losses as a result of its operation. An improved method of finding associative rules from SIEM databases is proposed, which is based on the theory of fuzzy sets and linguistic terms. The problem of finding fuzzy associative rules with weighted attributes is formulated. The mathematical apparatus that forms the basis of the implementation of the method is given. An algorithm for finding frequent sets of elements, including the values of the signs of cyber incidents and the classes to which they belong, is proposed, which implements the first stage of the proposed method. The peculiarities of the structure of the test data sets used for training and testing of cyber protection systems were analyzed, and based on its results, a conclusion was drawn about the possibility of improving the considered algorithm. A graphic illustration of the idea of improving the algorithm for finding frequent sets of elements is given and the essence of its improvement is described. An improved algorithm for finding frequent sets of elements of the considered method is proposed and its main advantages are given.

Keywords: cyber protection; cyber incident; SIEM; theory of fuzzy sets; data mining; associative rules.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1 Subach, I., Mykytiuk, A., Kubrak, V. (2019). Architecture and functional model of a perspective proactive intellectual siem for cyber protection of objects of critical infrastructure. *Collection "Information technology and security"*, *7*(2), 208–215. https://doi.org/10.20535/2411 – 1031.2019.7.2.190570

2 Horng, S. – J., Su, M. – Y., Chen, Y. – H., Kao, T. – W., Chen, R. – J., Lai, J. – L., Perkasa, C. D. (2011). A novel intrusion detection system based on hierarchical clustering and support vector machines. *Expert Systems with Applications*, *38*(1), 306–313. https://doi.org/10.1016/j.eswa.2010.06.066

3 Mbikayi, H. K. (2012). An Evolution Strategy Approach toward RuleSet Generation for Network Intrusion Detection Systems (IDS). *International Journal of Soft Computing and Engineering*, *2*(5), 1–5.

ISSN 2663 - 4023

4 Subach, I., Fesokha, V., Fesokha, N. (2017b). Analysis of existing solutions for preventing invasion in information and telecommunication networks. *Collection "Information technology and security"*, *5*(1), 29–41. https://doi.org/10.20535/2411 – 1031.2017.5.1.120554

5 Lappas, T., Pelechrinis, K. (2007). Data mining techniques for (network) intrusion detection systems. *Department of Computer Science and Engineering UC Riverside, Riverside CA*, (92521). https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=720ec75b12f2e08c5297251e29401c337c251621

6 Srikant, R., Agrawal, R. (1996). Mining quantitative association rules in large relational tables. *ACM SIGMOD Record*, *25*(2), 1–12. https://doi.org/10.1145/235968.233311

7 Gyenesei, A. (2001). A fuzzy approach for mining quantitative association rules. *Acta Cybernetica*, *15*(2), 305–320.

8 Subach, I. Y., Zdorenko, Y. M., Fesiokha, V. V. (2018). The method of detecting the JS (HTML)/Scrinject type on the basis of the stop of mathematical aparat theory of non-thematic multiplications. Zbirnik naukovikh prats [Viiskogo institutu telecommunications and informatizatsii], (4), 125 – 131. [9] Gyenesei, A. (2000). *Fuzzy partitioning of quantitative attribute domains by a cluster goodness index*. Turku Centre for Computer Science.

9 Gyenesei, A. (2000). Mining weighted association rules for fuzzy quantitative items. In *Principles of Data Mining and Knowledge Discovery: 4th European Conference, PKDD 2000 Lyon, France, September 13–16, 2000 Proceedings 4* (pp. 416 – 423). Springer Berlin Heidelberg.

10 Subach, I. Y. (2005). Finding fuzzy associative rules in the relational databases of a telecommunications company. *Communications*, (3), 54–57.

11 Gerasimov, B., Subach, I. (2005). In addition to the best social rules on the basis of information in the databases of data information systems. *National Technical University "Podillya"*, (4), 266–270.

12 Au, W. H., Chan, K. C. (1998, May). An effective algorithm for discovering fuzzy rules in relational databases. In *1998 IEEE International Conference on Fuzzy Systems Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36228)* (Vol. 2, pp. 1314 – 1319). IEEE.

13 Chan, K. C., Au, W. H. (2001). Mining fuzzy association rules in a database containing relational and transactional data. *Data mining and computational intelligence*, 95 – 114.

14 Newman, D. (1999). KDD Cup'99 Data Sets. *Retrieved February*, *7*, 2010.