«Системні технології» 2 (139) 2022 «System technologies»
DOI 10.34185/1562-9945-2-139-2022-06
УДК 004.056: 004.94

О.В. Герасіна, В.І. Корнієнко, О.Ю. Гусєв, К.В. Соснін, С.М. Мацюк

ВИЯВЛЕННЯ ФІШИНГОВИХ URL-АДРЕС ЗА ДОПОМОГОЮ АЛГОРИТМІВ НЕЧІТКОЇ КЛАСТЕРИЗАЦІЇ ІЗ ГЛОБАЛЬНОЮ ОПТИМІЗАЦІЄЮ

Анотація. Запропоновано алгоритм виявлення фішингових URL-адрес (класифікатор) із використанням нечіткої кластеризації, який включає вибір типу інтелектуального класифікатора та обґрунтування його параметрів за допомогою методів глобальної оптимізації. Шляхом моделювання оцінено ефективність використання запропонованого алгоритму на прикладі експериментальних даних – набору фішингових і безпечних URL-адрес.

Ключові слова: конфіденційні дані, класифікатор, нечітка кластеризація С-середніх, фішингові атаки, глобальна оптимізація, субтрактивна кластеризація.

Постановка проблеми. Фішингові URL – це шкідливі веб-сайти, які маскують себе під безпечні для отримання конфіденційних даних, таких як номери кредитних карт, інформація щодо імені користувача та паролю для входу тощо. Фішинг використовує соціальну інженерію та технічний обман, щоб отримати приватну інформацію від веб-користувача. Наразі фішинг є однією з найбільших загроз в Інтернеті. Взагалі каналами фішингових атак є електронна пошта, текстові повідомлення та посилання в соціальних мережах. Нерідко цільова сторінка фішингового веб-сайту також робить спробу проникнення в комп'ютер відвідувача та встановлення шкідливого програмного забезпечення [1].

Таким чином, фішингові атаки є значним ризиком як для фізичних осіб, так і для організацій, оскільки представляють собою загрозу розголошення або отримання конфіденційної особистої та корпоративної інформації. За даними звіту компанії APWG (Anti-Phishing Work Group) у першому кварталі 2020 р. загалом зареєстровано 165 772 окремих фішингових операції, зокрема, з використанням електронної пошти [2].

Отже, автоматичне виявлення та класифікація фішингових сайтів є актуальною задачею.

Аналіз останніх досліджень і публікацій. Для вирішення задачі виявлення та класифікація фішингових веб-сайтів перспективним є використання

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

систем штучного інтелекту (нейронних мереж (НМ) та систем з нечіткою логікою) [3-7]. У роботах [3-4] запропоновано використання нечітких методів для ідентифікації фішингових веб-сайтів. При цьому, у роботі [3] зазначається, що основною проблемою використання алгоритмів інтелектуального аналізу даних є вибір набору функцій, які будуть використовуватись. У роботі [5] для вирішення цього завдання запропоновано використовувати адаптивну нейронечітку систему Anfis, у роботах [6-7] – НМ для виявлення фішингових повідомлень електронної пошти. Встановлено [6], що НМ з відповідною кількістю прихованих шарів можуть досягти задовільної точності, навіть якщо прикладів навчання недостатньо. Для підвищення ефективності роботи НМ запропоновано навчати їх алгоритмом рою часток [7].

Таким чином, невирішеним є завдання вибору типу інтелектуального класифікатору та обґрунтування його параметрів для виявлення фішингових URL.

Мета роботи – дослідження алгоритмів нечіткої кластеризації для виявлення фішингових URL-адрес із налаштуванням параметрів нечітких класифікаторів за допомогою методів глобальної оптимізації.

Викладення основного матеріалу дослідження. Класифікація на основі алгоритмів нечіткої кластеризації. Алгоритми нечіткої логіки ґрунтуються на твердженні, що функція належності елемента множині може приймати значення в інтервалі [0, 1]. Переваги таких алгоритмів – прозорість процесу отримання висновків на основі словесного опису експертних знань про процес, а також стійкість до шумів. Недоліки – відсутність автоматичного набуття знань, а також обмежена кількість вхідних змінних [8, 9].

Задача кластеризації полягає в об'єднанні множини об'єктів у групи (кластери) на основі подібності ознак для об'єктів однієї групи і відмінностей між ними. Більшість алгоритмів кластеризації можуть використовуватись в умовах майже повної відсутності інформації про закони розподілу даних. Вихідною інформацією для кластеризації є матриця спостережень [9]: $X = \{x_{i,j}\}$, де x_{ij} – значення ознаки *i* об'єкту кластеризації *j*. $i = \overline{1, p}$; $i = \overline{1, \Xi}$.

Найбільш відомими методами нечіткої кластеризації є: субтрактивна кластеризація (Subtractive Clustering, Genfis 2) та нечітка кластеризація С-середніх (Fuzzy C-means, Genfis 3) [8, 10, 11].

В основі методу субтрактивної кластеризації лежить припущення, що кожна експериментальна точка може бути центром кластеру.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

При субтрактивній кластеризації генерується система нечіткого логічного висновку типу Сугено. Екстракція правил з даних відбувається в два етапи. Спочатку визначається кількість правил і потужностей терм-множин вихідних змінних. Далі за допомогою методу найменших квадратів визначається права частина кожного правила. Результатом є система нечіткого логічного висновку з базою правил, що охоплюють всю предметну область.

Алгоритм субтрактивної кластеризації може бути представлений наступним чином:

1. Розрахувати потенціалу кожної точки x_k (як міри просторової близькості між нею та іншими):

$$\mathrm{E}(x_k) = \frac{1}{K}\sum_{i=1}^{K} e^{-\frac{\|x_k - x_i\|}{(R_c/2)^2}}, \quad (1)$$

де R_c – позитивне число, яке представляє собою радіус центру кластера, K – число точок даних в навчальній послідовності.

2. Встановити кількість кластерів $k_c = 0$.

3. Виявити точку даних з найвищим потенціалом $E(x_p)$, x_p:

$$p = \arg\max_{i=1}^{K} \mathrm{E}(x_i). \quad (2)$$

4. Встановити j-й центр кластера:

$$k_{c_j} = x_p, \quad (3)$$

при цьому $E(j_1)$ – його потенціал, j=j+1 – приріст.

5. Знизити потенціал всіх точок:

$$\mathrm{E}(x_i) = \mathrm{E}(x_i) - \mathrm{E}(k_{cj}) e^{-\frac{\|k_{cj} - x_i\|}{(r/2)^2}}, \quad (4)$$

де r=[1,1.5] R_C – позитивна постійна, що визначає діапазон впливу одного кластера; i=1,...,K.

6. Перевірити значення потенціалу точок відносно встановленого порогу thr:

$$\max_{i=1}^{K} \mathrm{E}(x_i) < thr. \quad (5)$$

Якщо умова (5) виконується, то кінець алгоритму, інакше – перехід до пункту 3.

В алгоритмі субтрактивної кластеризації радіуси кластерів визначають наскільки далеко від центру кластера можуть бути його елементи. Слід зазначити, що вибір радіусу може сильно вплинути на результат. Якщо задати неве-

лике значення радіусу, то база буде повнішою, але чутливою до викидів і неточностей вимірів. Якщо задати радіус занадто великим, то можна втратити деякі правила при синтезі класифікатора.

В основі алгоритму нечіткої кластеризації С-середніх лежить метод невизначених множників Лагранжа, який дозволяє задачі знаходження умовного екстремуму цільової функції на множині допустимих значень перетворитись на задачу безумовної оптимізації функції.

Алгоритм нечіткої кластеризації С-середніх – це ітеративна процедура, в якій виконуються наступні кроки:

1. Завдання нечітких кластерів матрицею розбиття:

$$M_D = [\mu_{\theta i}],\ \mu_{\theta i} \in [0,1],\ \theta = \overline{1,\Theta},\ i = \overline{1,k_c}\ ; \tag{6}$$

при цьому

$$\sum_{i=1}^{k_c} \mu_{\theta i} = 1,\ \ 0 < \sum_{\theta=1}^{\Theta} \mu_{\theta i} < \Theta\ ; \tag{7}$$

де $\mu_{\theta i}$ – ступінь належності об'єкта θ до кластеру i, k_c – кількість кластерів, Θ – кількість елементів.

2. Установка параметрів алгоритму: k_c – кількість кластерів, ϖ – експоненційна вага, яка визначає нечіткість, розмазаність кластерів ($\varpi \in [1,\infty]$), ε – параметр зупинки алгоритму.

3. Генерація випадковим чином матриці нечіткого розбиття з урахуванням умов (7).

4. Розрахунок центрів кластерів Ω_i:

$$\Omega_i = \frac{\sum_{\theta=1}^{\Theta} \mu_{\theta i}^{\varpi} * |X_\theta|}{\sum_{\theta=1}^{\Theta} \mu_{\theta i}^{\varpi}},\ i = \overline{1,k_c}\ . \tag{8}$$

5. Розрахунок відстані між об'єктами з матриці спостережень і центрами кластерів:

$$D_{\theta i} = \sqrt{\|X_\theta - \Omega_i\|^2}\ . \tag{9}$$

6. Перерахунок елементів матриці розбиття.

- якщо $D_{\theta i} > 0$, то $\mu_{\theta i} = 1 \Big/ \left(D_{j\theta}^2 * \sum_{j=1}^{k_c} \frac{1}{D_{j\theta}^2} \right)^{1/(\varpi-1)}$;

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

- якщо $D_{\theta i} = 0$, то $\mu_{\theta i} = \begin{cases} 1, j = i \\ 0, j \neq i \end{cases}$, $j = \overline{1, k_c}$.

6. Перевірка умови (якщо вона виконується, то кінець алгоритму, інакше – перехід до пункту 4):

$$\left\| M_D - M_D^* \right\| < \varepsilon, \tag{10}$$

де M_D^* – матриця нечіткого розбиття на попередній ітерації алгоритму.

У наведеному алгоритмі параметром, який може суттєво вплинути на результат, є число кластерів k_c. Обрати кількість кластерів для реальних завдань без достатньої апріорної інформації про структури даних досить складно, й наразі існують два підходи до цього. Перший підхід заснований на критерії компактності і роздільності отриманих кластерів (наприклад, індекс Хей-Бені). Другий підхід пропонує починати кластеризацію при досить великому числі кластерів, а потім послідовно об'єднувати схожі суміжні кластери. При цьому використовуються різні формальні критерії схожості кластерів.

Також на результат алгоритму нечіткої кластеризації С-середніх може вплинути такий параметр, як експоненційна вага (ϖ), яка задає рівень нечіткості отриманих кластерів. Наразі не існує обґрунтованого правила вибору значення експоненціального ваги, і зазвичай її встановлюють рівною 2.

Оскільки задача вибору нечіткого класифікатора з налаштуванням його параметрів є полімодальною, то це вимагає використання методів глобальної оптимізації, серед яких найбільш ефективними є пошукові методи. У них алгоритм пошуку оптимального рішення пов'язує наступні один за одним рішення $\Psi_s(j+1) = F[\Psi_s(j)]$, де F – алгоритм пошуку, який показує які операції слід зробити на кроці j при рішенні $\Psi_s(j)$, щоб отримати нове рішення $\Psi_s(j+1) \succ \Psi_s(j)$. Тут знак переваги $\succ$ при мінімізації функціоналу має сенс:

$$C[\Psi_s(j+1)] < C[\Psi_s(j)]. \tag{11}$$

В алгоритмах прямого випадкового пошуку (ПВП) задаються напрямки пошуку і визначаються значення функціоналу C в точках $\Psi_s(j) \pm \gamma\varsigma$. Рішення полягає у виборі кроку в напрямку зменшення цього функціоналу:

$$\Psi_s(j+1) = \Psi_s(j) - \omega\varsigma\{C[\Psi_s(j) + \gamma\varsigma] - C[\Psi_s(j) - \gamma\varsigma]\}, \tag{12}$$

де $\omega, \varsigma, \gamma$ – параметри, що визначають сфери прийняття рішення (ω), збору інформації (γ) та одиничний випадковий напрям (ς). У загальному випадку параметри в (12) можуть змінюватися (адаптуватися) до процедури пошуку і виду гіперповерхні прийнятого функціоналу.

Розвитком методу ПВП є метод імітації відпалу (МІВ), який відображає поведінку розплавленого матеріалу при затвердінні із застосуванням процедури керованого охолодження (відпалу). У процесі відпалу кристалізація розплаву супроводжується глобальним зменшенням його енергії, однак допускається її зростання на деякий час. Завдяки цьому можливий вихід з пасток локальних мінімумів енергії, що виникають при реалізації процесу.

В алгоритмах МІВ задаються напрямки пошуку і визначаються значення функціоналу C в точках $\Psi_s(j) \pm \upsilon\tau$. Рішення полягає у виборі кроку в напрямку зменшення цього функціоналу:

$$\Psi_s(j+1) = \Psi_s(j) - \omega\upsilon\{C[\Psi_s(j) + \upsilon\tau] - C[\Psi_s(j) - \upsilon\tau]\}, \quad (13)$$

де ω, υ, τ – параметри, що визначають сфери прийняття рішення (ω), зміну поточного рішення (υ) і зменшення температури (τ).

Метод порогового прийняття (МПП) використовує підхід, схожий з МІВ, але замість того, щоб вибрати нові точки, які оптимізують критерій з певною ймовірністю, цей метод обирає все нові точки нижче встановленого порогу. Таким чином, поріг і температура систематично знижені, що дозволяє уникнути ймовірнісних обчислень, і може визначити оптимум швидше, ніж в МІВ.

Розвитком пошукових методів є еволюційні алгоритми, серед яких найбільш поширені генетичні алгоритми (ГА), які моделюють розвиток біологічної популяції на рівні геномів: мутації структури і параметрів, їх схрещування (розмноження) [8, 9]:

$$\Psi_s(j+1) = \Psi_s(j) + \delta\Psi_s(j), \quad (14)$$

і правило відбору, що дозволяє виявляти їх сприятливі варіації, за допомогою яких будується послідовність поліпшених рішень.

Більшість задач, що вирішуються за допомогою ГА, мають один критерій оптимізації. Багатокритеріальна оптимізація (БО) заснована на знаходженні рішення, одночасно оптимізуючого більш ніж одну функцію. У цьому випадку шукається певний компроміс, в ролі якого виступає рішення, оптимальне в сенсі Парето. При БО, що використовує ГА вибирається не одна хромосома, що представляє собою оптимальне рішення в звичайному сенсі, а безліч хромосом, оптимальних в сенсі Парето. Користувач має можливість вибрати оптимальне рішення з цієї безлічі [8]:

$$k \cdot \Psi_s(j+1) = k \cdot (\Psi_s(j) + \delta\Psi_s(j)), \quad (15)$$

де $k \geq 2$ – число розглянутих критеріїв.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Як критерій глобальної оптимізації використовують комбінований критерій:

$$C_{комб} = 0{,}2 \cdot C_{рег} + 0{,}8 \cdot C_{зм}, \quad (16)$$

де $C_{рег}$ – критерій регулярності, що обчислюється на перевірочній вибірці [12]:

$$C_{рег} = \frac{\left\| Y_B^*[m+n] - \hat{Y}_B[m+n] \right\|}{\left\| Y_B^*[m+n] \right\|}, \quad (17)$$

$C_{зм}$ – критерій незміщенності (мінімуму зсуву), заснований на аналізі рішень [12]:

$$C_{зм} = \frac{\left\| \hat{Y}_A[m+n] - \hat{Y}_B[m+n] \right\|}{\left\| Y^*[m+n] \right\|}. \quad (18)$$

Моделювання процесу виявлення фішингових URL-адрес було проведено в середовищі Matlab за допомогою стандартних та розроблених програм. Узагальнена структура алгоритму виявлення фішингових URL-адрес зображена на рис. 1.

Для моделювання використовувався набір даних із 150 фішингових і 150 безпечних URL-адрес, який включав в себе інформацію про регістратора доменного імені, час життя домену, геолокацію хостинг-сервера, наявність захищеного з'єднання з дійсним сертифікатом. Для збору фішингових URL-адрес використовувалась база валідованих фішингових посилань на Phishtank.org, куди завантажують нові зразки посилань як антивірусні компанії, так і індивідуальні дослідники. Для збору безпечних URL-адрес використовувався сервіс Alexa.com, де є адреси популярних веб-сайтів. При цьому, весь набір даних був розділений на навчальну і тестову вибірки.

Використовувались класифікатори на основі алгоритмів Genfis 2 та Genfis 3.

Як глобальні методи оптимізації використовувались ГА, БО, ПВП, МІВ, МПП. Як критерій глобальної оптимізації використовувався критерій (16).

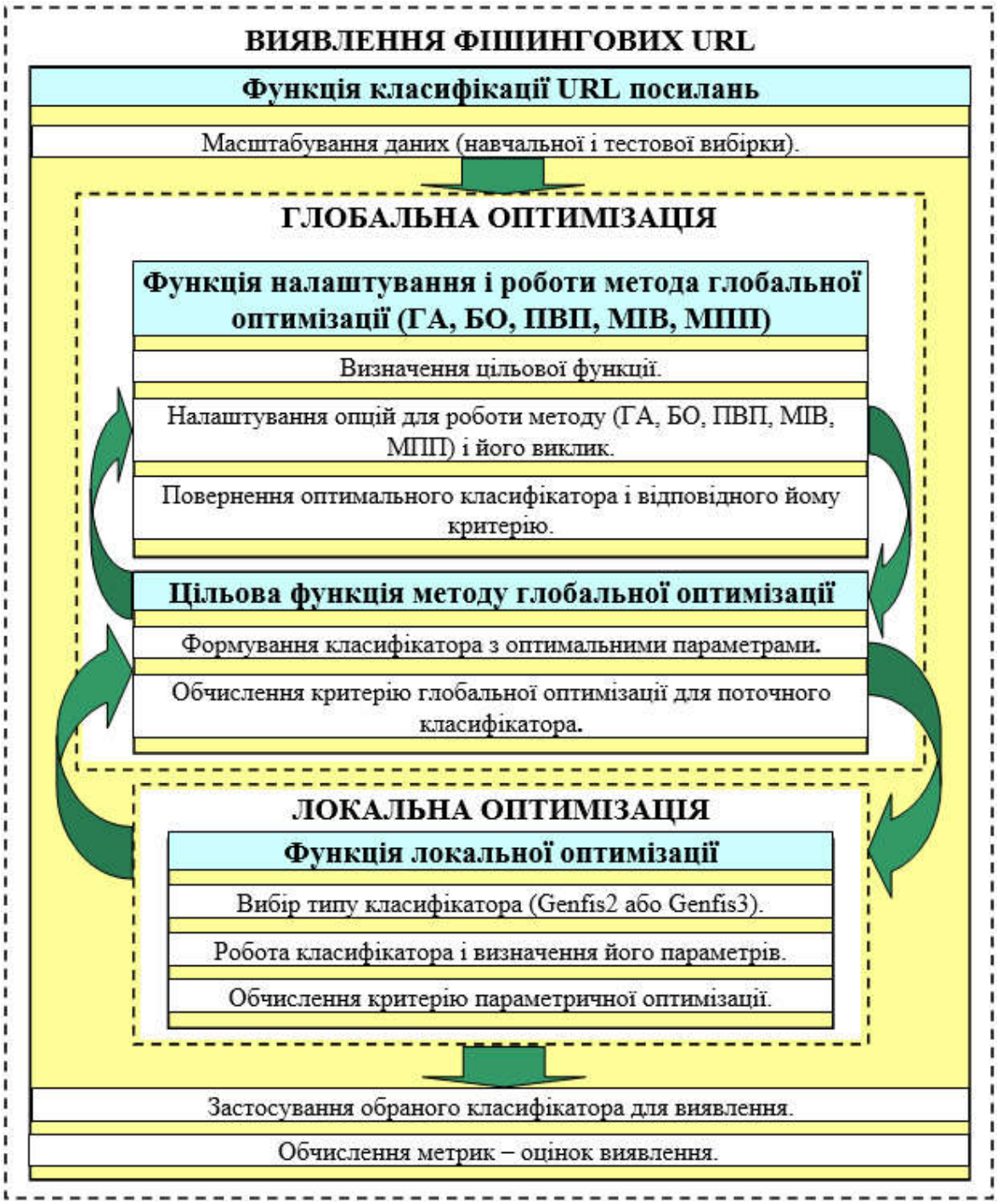


Рисунок 1 - Структура алгоритму виявлення фішингових URL-адрес

Для оцінки класифікації URL-адрес використовувались наступні метрики:

- TP (True Positive) – кількість безпечних посилань які було класифіковано як безпечні;

- TN (True Negative) – кількість фішингових посилань які було класифіковано як фішингові;

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

- FP (False Positive) – кількість безпечних посилань які було класифіковано як фішингові;

- FN (False Negative) – кількість фішингових посилань які було класифіковано як безпечні;

І похідні від них:

- True positive rate:

$$TPR = TP/(TP + FN); \quad (19)$$

- True negative rate:

$$TNR = TN/(TN + FP); \quad (20)$$

- False positive rate:

$$FPR = FP/(FP + TN); \quad (21)$$

- False negative rate:

$$FNR = FN/(FN + TP); \quad (22)$$

- Positive predictive value:

$$PPV = (TP + TN)/(TP + TN + FP + FN); \quad (23)$$

- Negative predictive value:

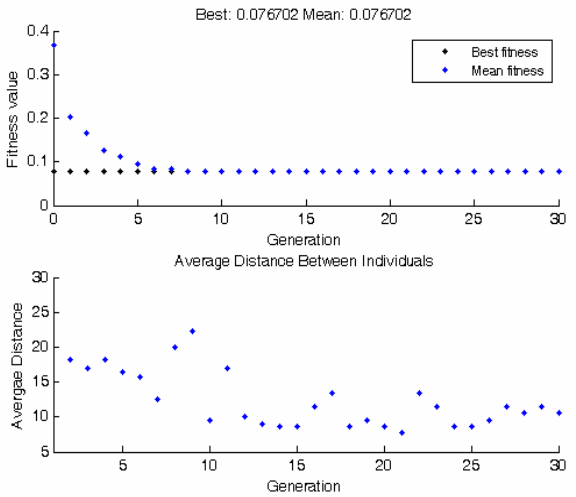
$$NPV = TN/(TN + FN); \quad (24)$$

- F-measure – гармонійне середнє між TPR і PPV:

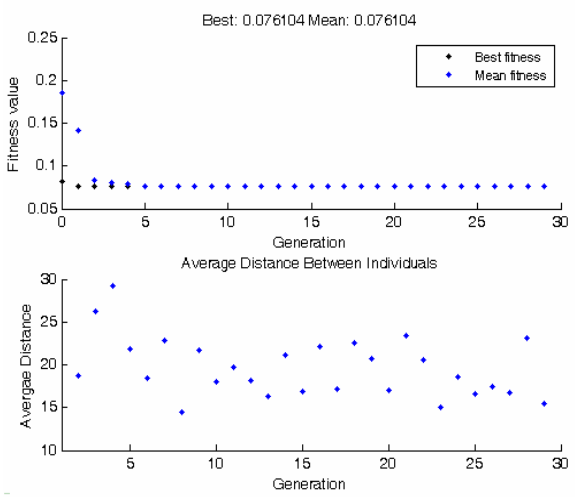
$$F - measure = 2 * PPV * TPR/(PPV + TPR). \quad (25)$$

У контексті завдання виявлення кібератак найбільш важливою є метрика FN / FNR, яка дає кількісну оцінку невиявленим атакам.

Результати глобальної оптимізації для знаходження типу і параметрів нечіткого класифікатора наведено на рис. 2.



а



б

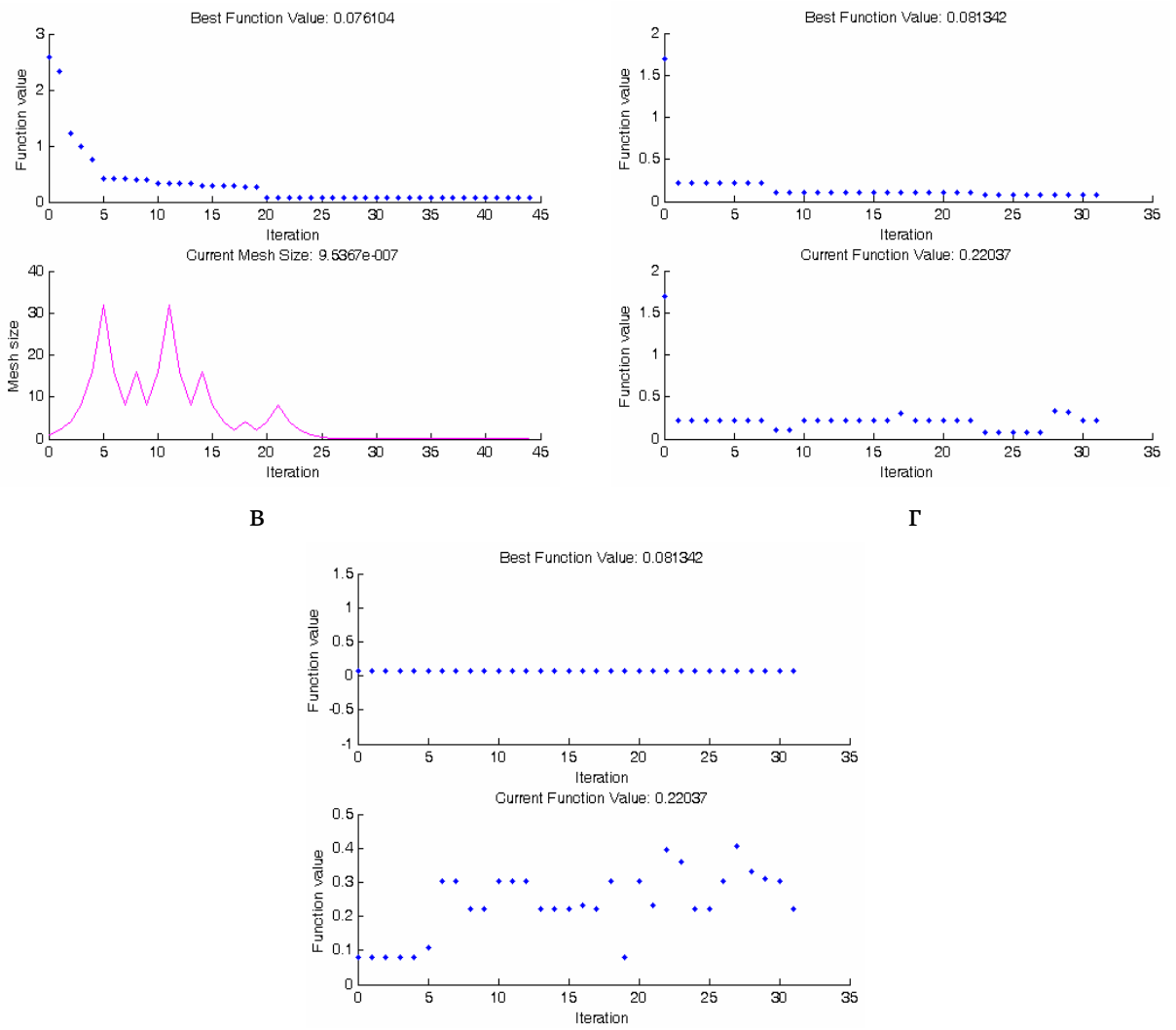


Рисунок 2 - Результати глобальної оптимізації параметрів нечіткого класифікатора для виявлення фішингових URL-адрес за допомогою: а – ГА, б – БО, в – ПВП, г – МІВ, д – МПП

Метод БО використовував ГА для знаходження безлічі оптимальних за Парето рішень. ПВП мав адаптивний крок пошуку і повний пошук навколо поточної ітерації, МІВ і МПП – обмежену область перевідпалу, ГА – одноточкове схрещування, селективний вибір батьків, формування нової популяції із витісненням. Кількість ітерацій для ПВП, МПП і МІВ (для ГА і БО поколінь) обмежувалось на рівні 100, а розмір простору пошуку для ПВП (для ГА і БО розмір популяції, для МПП і МІВ розмір області перевідпалу) – 30.

При глобальній оптимізації варіювались наступні параметри класифікатору:

- тип алгоритму – Genfis2 або Genfis3;

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

- для Genfis2 – діапазон впливу кластерного центру R_c ;

- для Genfis3 – алгоритм нечіткої логіки (Мамдані або Сугено [8]) і число кластерів k_c.

У результаті моделювання (див. рис. 2 (а-д)) встановлено, що БО і МПП мають найвищу швидкість збіжності (БО виходить в область оптимальних рішень на перших поколіннях, МПП – перших ітераціях, ГА – у середньому після 7 поколінь, МІВ – після 7 ітерацій, ПВП –після 20 ітерацій).

Алгоритм ПВП виявив найвищу швидкодію (1 с на ітерацію, при 2 с на ітерацію в МІВ, 3 с на ітерацію в МПП, 9 с на покоління в ГА і 10 с на покоління в БО). При цьому алгоритми ПВП і БО виявили найкращу збіжність (значення критерію (16) при їх використанні склали 0,076, при 0,081 для МІВ і МПП та 0,077 для ГА).

Встановлено, що мінімуму критерію (16) відповідають: тип алгоритму Genfis3 з використанням структури алгоритму Сугено і 6 кластерами.

Значення метрик оцінки виявлення фішингових URL-адрес для цього класифікатора наведено в табл.1. Тут встановлено, що усі фішингові URL-адреси, які були класифіковані як безпечні (FN=30) мали захищене з'єднання з дійсним сертифікатом (https).

Таблиця 1

Значення метрик оцінки виявлення URL-адрес

TP	TN	FP	FN	TPR	TNR	FPR	FNR	PPV	NPV	F-measure
120	150	0	30	0,8	1	0	0,2	0,9	0,833	0,847

Висновки

Розроблено алгоритм виявлення фішингових URL-адрес (класифікатор) із використанням нечіткої кластеризації та методів глобальної оптимізації.

Шляхом моделювання встановлено, що мінімуму комбінованого критерію відповідає нечіткий класифікатор із алгоритмом субтрактивної кластеризації та використанням структури Сугено і 6 кластерами.

Встановлено, що усі фішингові URL-адреси, які помилково були класифіковані як безпечні, мали наявність захищеного з'єднання з дійсним сертифікатом. Таким чином, подальші дослідження мають бути спрямовані на дослідження додаткових інформативних атрибутів (ознак), які могли б дозволити більш якісно розділяти фішингові і безпечні URL-адреси.

ЛІТЕРАТУРА

1. Звіт компанії NSS Labs про тестування захисту від фішингу за другий квартал 2020 р. [Електронний ресурс] – Режим доступу: https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RWAIQX.

2. Звіт компанії APWG про тенденції фішингової діяльності за перший квартал 2020 року. [Електронний ресурс] – Режим доступу: https://docs.apwg.org/reports/apwg_trends_report_q1_2020.pdf.

3. Intelligent phishing detection system for e-banking using fuzzy data mining / M. Aburrous, M.A. Hossain, K. Dahal, F. Thabtah // Expert Systems with Applications. – 2010. – Vol.37, issue 12. – P. 7913–7921.

4. Fuzzy Modelling using Firefly Algorithm for Phishing Detection / N.S. Nordin et al. // Advances in Science, Technology and Engineering Systems Journal. – 2019 – Vol. 4, No. 6. – P. 291-296.

5. Barraclough P.A. Parameter optimization for intelligent phishing detection using Adaptive Neuro-Fuzzy / P.A. Barraclough, M.A. Hossain, G. Sexton, N. Aslam // International Journal of Advanced Research in Artificial Intelligence. – 2014. – 3(10). – P. 16-25.

6. Zhang N. Phishing Detection Using Neural Network / N. Zhang, Y. Yuan // Stanford University – 2012. [Електронний ресурс] – Режим доступу: http://cs229.stanford.edu/proj2012/ZhangYuan-PhishingDetectionUsingNeuralNetwork.pdf.

7. Gupta S. Phishing URL detection by using artificial neural network with PSO / S. Gupta, A. Singhal // 2nd International Conference on Telecommunication and Networks (TEL-NET). – 2017. – P. 1-6.

8. Корнієнко В.І. Інтелектуальне моделювання нелінійних динамічних процесів в системах керування, кібербезпеки, телекомунікацій: підручник / В.І. Корнієнко, О.Ю. Гусєв, О.В. Герасіна; за заг. ред. В.І. Корнієнка ; М-во освіти і науки України, Нац. техн. ун-т «Дніпровська політехніка». – Дніпро : НТУ «ДП», 2020. – 536 с.

9. Рутковская Д. Нейронные сети, генетические алгоритмы и нечеткие системы / Д. Рутковская, М. Пилиньский, Л. Рутковский // пер. с польск. И.Д. Рудинского. – М.: Горячая линия-Телеком, 2006. – 452 с.

10. Штовба С.Д. Проектирование нечетких систем средствами Matlab / С.Д. Штовба. – М.: Горячая линия – Телеком, 2007. – 288 с.

11. Yager R. Essentials of Fuzzy Modelling and Control / R. Yager, D. Filev. – USA: John Wiley & Sons. – 1984. – 387 p.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

12. Bezdek J.C. Pattern Recognition with Fuzzy Objective Function Algorithms / J.C. Bezdek. – New York : Plenum Press. – 1981. – 272 p.
13. Ivakhnenko A.G. Inductive learning algorithms for complex systems modeling / A.G. Ivakhnenko, H.R. Madala – London, Tokyo: CRC Press, 1994. – 384 p.

REFERENCES

1. Zvit kompanii NSS Labs pro testuvannia zakhystu vid fishynhu za druhyi kvartal 2020 r. [Electronic resource] – Access mode: https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RWAIQX.
2. Zvit kompanii APWG pro tendentsii fishynhovoi diialnosti za pershyi kvartal 2020 roku. [Electronic resource] – Access mode: https://docs.apwg.org/reports/apwg_trends_report_q1_2020.pdf.
3. Intelligent phishing detection system for e-banking using fuzzy data mining / M. Aburrous, M.A. Hossain, K. Dahal, F. Thabtah // Expert Systems with Applications. – 2010. – Vol.37, issue 12. – P. 7913–7921.
4. Fuzzy Modelling using Firefly Algorithm for Phishing Detection / N.S. Nordin et al. // Advances in Science, Technology and Engineering Systems Journal. – 2019 – Vol. 4, No. 6. – P. 291-296.
5. Barraclough P.A. Parameter optimization for intelligent phishing detection using Adaptive Neuro-Fuzzy / P.A. Barraclough, M.A. Hossain, G. Sexton, N. Aslam // International Journal of Advanced Research in Artificial Intelligence. – 2014. – 3(10). – P. 16-25.
6. Zhang N. Phishing Detection Using Neural Network / N. Zhang, Y. Yuan // Stanford University – 2012. [Electronic resource] – Access mode: http://cs229.stanford.edu/proj2012/ZhangYuan-PhishingDetectionUsingNeuralNetwork.pdf.
7. Gupta S. Phishing URL detection by using artificial neural network with PSO / S. Gupta, A. Singhal // 2nd International Conference on Telecommunication and Networks (TEL-NET). – 2017. – P. 1-6.
8. Korniienko V.I. Intelektualne modeliuvannia neliniinykh dynamichnykh protsesiv v systemakh keruvannia, kiberbezpeky, telekomunikatsii: pidruchnyk / V.I. Korniienko, O.Yu. Husiev, O.V. Herasina; za zah. red. V.I. Korniienka; M-vo osvity i nauky Ukrainy, Nats. tekhn. un-t «Dniprovska politekhnika». – Dnipro : NTU «DP», 2020. – 536 p.
9. Rutkovskaya D. Neyronnyye seti, geneticheskiye algoritmy i nechetkiye sistemy / D. Rutkovskaya. M. Pilinskiy. L. Rutkovskiy // per. s polsk. I.D. Rudinskogo. – M.: Horiachaia lynyia-Telekom, 2006. – 452 p.

10. Shtovba S.D. Proektyrovanye nechetkykh system sredstvamy Matlab / S.D. Shtovba. – M.: Horiachaia lynyia – Telekom, 2007. – 288 p.

11. Yager R. Essentials of Fuzzy Modelling and Control / R. Yager, D. Filev. – USA: John Wiley & Sons. – 1984. – 387 p.

12. Bezdek J.C. Pattern Recognition with Fuzzy Objective Function Algorithms / J.C. Bezdek. – New York : Plenum Press. – 1981. – 272 p.

13. Ivakhnenko A.G. Inductive learning algorithms for complex systems modeling / A.G. Ivakhnenko, H.R. Madala – London, Tokyo: CRC Press, 1994. – 384 p.

Received 10.02.2022.
Accepted 14.02.2022.

Detecting fishing URLs using fuzzy clustering algorithms with global optimization

An algorithm for detecting phishing URLs (classifier) using fuzzy clustering is proposed, which includes choosing the type of intelligent classifier and justifying its parameters using global optimization methods. The following were studied as intellectual classifiers: subtractive clustering and fuzzy clustering of C-means. To find (adjust) the optimal (for a specific task) parameters of intelligent classifiers, the use of global optimization methods is justified, including genetic algorithm, direct random search, annealing simulation method, multicriteria optimization and threshold acceptance method. As a criterion of global optimization, a combined criterion was used, which includes the definition of the regularity criterion calculated on the test sample and the bias (minimum shift) criterion based on the analysis of solutions. By modeling in the Matlab environment with the help of standard and developed programs, the evaluated efficiency of using the proposed algorithm is evaluated on the example of experimental data – a set of 150 phishing and 150 secure URLs. The set of experimental data included information about the domain name registrar, the lifetime of the domain, the geolocation of the hosting server, the presence of a secure connection with a valid certificate. By simulation it is established that the fuzzy classifier with the subtractive clustering algorithm and using the Sugeno structure and 6 clusters meets the minimum of the combined criterion. All phishing URLs that were mistakenly classified as secure were found to have a secure connection with a valid certificate. Thus, further research should be aimed at exploring additional informative attributes (features) that could allow better separation of phishing and secure URLs.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Герасіна Олександра Володимирівна – доцент кафедри безпеки інформації та телекомунікацій, Національний технічний університет «Дніпровська політехніка», к.т.н., доцент;

Корнієнко Валерій Іванович – завідувач кафедри безпеки інформації та телекомунікацій, Національний технічний університет «Дніпровська політехніка», д.т.н., професор;

Гусєв Олександр Юрійович – професор кафедри безпеки інформації та телекомунікацій, Національний технічний університет «Дніпровська політехніка», к.ф.-м.н., доцент;

Соснін Костянтин Володимирович – доцент кафедри кіберфізичних та інформаційно-вимірювальних систем, Національний технічний університет «Дніпровська політехніка», к.т.н.;

Мацюк Сергій Михайлович – асистент кафедри безпеки інформації та телекомунікацій, Національний технічний університет «Дніпровська політехніка», к.т.н.

Gerasina Oleksandra – Associate Professor of Department of Information Security and Telecommunications, Dnipro University of Technology, Candidate of Technical Sciences, Associate Professor;

Korniienko Valerii – Head of Department of Information Security and Telecommunications, Dnipro University of Technology, Doctor of Technical Sciences, Professor;

Gusev Oleksandr – Professor of Department of Information Security and Telecommunications, Dnipro University of Technology, Candidate of Physics and Mathematics, Associate Professor;

Sosnin Kostiantyn– Associate Professor of Department of Cyberphysical and Information-Measuring Systems, Dnipro University of Technology, Candidate of Technical Sciences;

Matsiuk Serhii – Assistant Lecturer of Department of Information Security and Telecommunications, Dnipro University of Technology, Candidate of Technical Sciences.