КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 3 (23), 2024
ISSN 2663 - 4023
DOI 10.28925/2663-4023.2024.23.258273
УДК 004.94:519.21

Шевченко Світлана Миколаївна
кандидат педагогічних наук, доцент,
доцент кафедри інформаційної та кібернетичної безпеки імені професора Володимира Бурячка
Київський столичний університет імені Бориса Грінченка, Київ, Україна
ORCID 0000-0002-9736-8623
s.shevchenko@kubg.edu.ua

Жданова Юлія Дмитрівна
кандидат фізико-математичних наук, доцент,
доцент кафедри інформаційної та кібернетичної безпеки імені професора Володимира Бурячка
Київський столичний університет імені Бориса Грінченка, Київ, Україна
ORCID 0000-0002-9277-4972
y.zhdanova@kubg.edu.ua

Спасітєлєва Світлана Олексіївна
кандидат фізико-математичних наук, доцент,
доцент кафедри інформаційної та кібернетичної безпеки імені професора Володимира Бурячка
Київський столичний університет імені Бориса Грінченка, Київ, Україна
ORCID 0000-0003-4993-6355
s.spasitielieva@kubg.edu.ua

Мазур Наталія Петрівна
кандидат педагогічних наук, доцент,
доцент кафедри інформаційної та кібернетичної безпеки імені професора Володимира Бурячка
Київський столичний університет імені Бориса Грінченка, Київ, Україна
ORCID 0000-0001-7671-8287
n.mazur@kubg.edu.ua

Складанний Павло Миколайович
кандидат технічних наук, доцент,
завідувач кафедри інформаційної та кібернетичної безпеки імені професора Володимира Бурячка
Київський столичний університет імені Бориса Грінченка, Київ, Україна
ORCID 0000-0002-7775-6039
p.skladannyi@kubg.edu.ua

Негоденко Віталій Петрович
аспірант
кафедри інформаційної та кібернетичної безпеки імені професора Володимира Бурячка
Київський столичний університет імені Бориса Грінченка, Київ, Україна
ORCID 0000-0002-7678-9138
v.nehodenko.asp@kubg.edu.ua

МАТЕМАТИЧНІ МЕТОДИ В КІБЕРБЕЗПЕЦІ: КЛАСТЕРНИЙ АНАЛІЗ ТА ЙОГО ЗАСТОСУВАННЯ В ІНФОРМАЦІЙНІЙ ТА КІБЕРНЕТИЧНІЙ БЕЗПЕЦІ

Анотація. Величезна кількість інформаційних загроз та їх складність спонукає до досліджень та моделювання нових методологій та систем захисту інформації. Розробка та удосконалення систем інформаційної та кібернетичної безпеки включає в себе створення та обробку математичних моделей з використанням інформаційних технологій. Ця стаття є наступним дослідженням щодо застосування математичних методів та технологій в кібербезпеці, а саме: методів кластерного аналізу. Сучасний розвиток комп'ютерної техніки, зростання їх потужності сприяли широкому впровадженню алгоритмів Data Mining для обробки великих обсягів інформації у різних галузях суспільства та науки, зокрема і в сфері кібербезпеки. Кластерний аналіз дозволяє множину розбити на підмножини, щоб елементи кожної підмножини були схожі між собою, а елементи різних підмножин були найбільш відмінними. Це надає можливість

усунити недоліки якісного підходу в оцінці інформаційних ризиків. У роботі здійснено огляд наукових джерел щодо прикладного аспекту застосування методів кластеризації в системах безпеки, адже своєчасне прогнозування можливих інцидентів дозволяє керувати інформаційними ризиками та приймати ефективні рішення в забезпеченні конфіденційності, доступності та цілісності інформації. Охарактеризовані етапи процедури кластеризації, висвітлені питання вибору міри відстані та міри подібності для об'єктів, які вивчаються. Представлена порівняльна характеристика найбільш популярних методів кластерного аналізу: алгоритм «найближчого сусіда», «k-means», «fuzzy c-means», «cosine similarity», визначені їх переваги та недоліки. Це дослідження може бути корисним та використаним у навчальному процесі студентів спеціальності 125 «Кібербезпека та захист інформації».

Ключові слова: математичні методи; кластерний аналіз; інформаційна безпека; кібербезпека; алгоритм «найближчого сусіда», алгоритм «k-means», алгоритм «fuzzy c-means», алгоритм «cosine similarity».

ВСТУП

Постановка проблеми. Обсяги інформаційних загроз постійно зростають, а самі вони стають більш складнішими та більш витонченими. Для протистояння у цьому процесі науковці досліджують та моделюють різні методології та системи захисту інформації. Розробка та удосконалення систем інформаційної та кібернетичної безпеки включає в себе створення та обробку математичних моделей з використанням інформаційних технологій. Математичні методи та технології є підґрунтям для створення та покращення показників у сфері захисту інформації [1] – [8].

Сучасний стан потребує такі методи, які б дозволили прораховувати та прогнозувати можливі ризики використання загрозами уразливостей інформаційних активів з метою забезпечення конфіденційності, цілісності та доступності інформації. Обробка ризиків буде результативною, якщо ефективним буде процес аналізу, ідентифікації та оцінки ризиків. Для інформаційних та кібернетичних систем властива якісна оцінка ризиків, яка здійснюється на основі експертних оцінок фахівців. І хоча ця процедура містить математичну обробку цих оцінок, обчислюється узгодженість експертів, проте існує висока ймовірність отримати даний результат суб'єктивним. Вважаємо, що усунити дані похибки дозволить процес кластеризації загроз та уразливостей, ймовірностей настання ризиків інформаційної безпеки та ідентифікації методів захисту активів. Кластерний аналіз дозволяє множину розбити на підмножини, щоб елементи кожної підмножини були схожі між собою, а елементи різних підмножин були найбільш відмінними, в результаті чого маємо можливість працювати з більш зв'язними, вузькими і конкретними даними. Таким чином, виділення окремих кластерів спрямує класифікувати об'єкти безпеки, а надалі створити бібліотеку, на основі якої навчити штучну мережу.

Аналіз останніх досліджень і публікацій. Термін «кластерний аналіз» вперше запропоновано у 1939 році Тріоном Р., проте в той час не мав зацікавленості у науковців у зв'язку з громіздкими обчисленнями. З розвитком інформаційних технологій та потужності комп'ютерної техніки інтерес до даної технології стрімко зростає. Цей метод можна назвати «інструментом» для створення тієї чи іншої класифікації, що дозволяє упорядкувати дані в більш однорідні групи, класи, кластери. При цьому неабияку увагу приділено структурі та природі досліджуваних даних. На перших етапах свого існування кластерний аналіз займався питаннями, що стосуються біології, медицини, археології, психології, педагогіки, але з розвитком суспільства ця технологія має широкий спектр застосувань у пошукових, експертних та рекомендаційних системах, соціальних мережах та інших.

Алгоритмами кластерного аналізу почали вміло користуватися у сфері кібербезпеки [9] – [22]. Точкові наукові розробки цієї проблеми визначили актуальність даної роботи щодо узагальнення популярних методів кластеризації у системах інформаційної та кібербезпеки та окреслили мету цього дослідження.

Мета статті. Метою статті є порівняльний аналіз найбільш поширених методів кластерного аналізу та висвітлення питань їх застосування в системах безпеки через призму огляду літературних джерел.

ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ

Кластерний аналіз — математична процедура багатовимірного аналізу, яка дає можливість на основі багатьох показників, що характеризують ряд об'єктів, групувати їх у класи (кластери) таким чином, щоби об'єкти, які входять до одного класу, були більш однорідними, схожими в порівнянні з об'єктами, що входять до інших класів [23]. Під кластером розуміють групу однакових або подібних елементів, зібраних разом або близько один до одного.

Процедура кластеризація є неконтрольованим навчанням, навчанням без вчителя, має справу з розділом структури даних у невідомій області та слугує підґрунтям для подальшого навчання. Критерій якості кластеризації тією чи іншою мірою відбиває наступні неформальні вимоги [24]:

- елементи в одному кластері мають бути максимально схожими;
- елементи в різних кластерах повинні якомога більше відрізнятися;
- вимірювання подібності та несхожості має бути чітким і практичним значення.

Нехай множина $X = \{x_1, x_2, \dots x_n\}$ — це множина об'єктів, Y — множина номерів кластерів. Для кожних двох об'єктів множини X задана функція відстані $d_{ij} = d(x_i, x_j)$.

Необхідно розбити множину X на підмножини (кластери), які не перетинаються, і кожен кластер містив об'єкти близькі або подібні між собою, причому можливо не за одним показником, а за декількома водночас. Об'єкти різних класів мають істотно відрізнятися. Кожному об'єкту x_i приписується номер кластера Y_i.

Множина X може складатися із об'єктів, які мають різні одиниці вимірювання або різний діапазон представлених значень, тому потрібно здійснити нормування вхідних даних. При кластерному аналізі є два основні способи нормалізації даних: MinMax-нормалізація та Z-нормалізація.

MinMax-нормалізація здійснюється наступним чином:

$$x' = \frac{x - \min[X]}{\max[X] - \min[\mathrm{X}]},$$

у разі всі значення будуть у діапазоні від 0 до 1; дискретні бінарні значення визначаються як 0 та 1.

Z-нормалізація:

$$x' = \frac{x - \mathrm{M}[\mathrm{X}]}{\sigma[X]},$$

де M[X] — математичне сподівання, $\sigma[X]$ — середньоквадратичне відхилення.

Ключовим питання у кластерному аналізі є проблема визначення схожості об'єктів в одному кластері, яка називається метрикою або мірою подібності між об'єктами. В якості цієї величини можливо використовувати коефіцієнти кореляції, міри відстані

(найчастіше), коефіцієнти асоціативності, ймовірнісні коефіцієнти подібності. У табл. 1 представлені найбільш популярні відстані кластерного аналізу.

Таблиця 1

Відстані кластерного аналізу

Назва	Формула
Евклідова відстань	$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$
Зважена евклідова відстань	$d_{ij}^{*} = \sqrt{\sum_{k=1}^{p} \omega_k (x_{ik} - x_{jk})^2}$
Відстань Мінковського	$d_{ij} = \sqrt[r]{\sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert^r}$
Хеммінгова відстань	$d_{ij} = \sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert$
Косинусна відстань	$d_{ij} = 1 - sim_1(u, v)$, де $sim_1(u, v) = cos(u, v) = \frac{\langle u, v \rangle}{\lVert u \rVert \cdot \lVert v \rVert} = \frac{\sum_{k=1}^{t} u_k v_k}{\sqrt{\sum_{k=1}^{t} u_k^2 \sum_{k=1}^{t} v_k^2}}$ – косинусна міра подібності, де $u = (u_1, u_2, \dots u_t)$, $v = (v_1, v_2, \dots v_t)$ — вектори ознак об'єктів, u_k, v_k — їх компоненти (координати), $\sum_{k=1}^{t} u_k v_k$ — скалярний добуток між цими векторами.

Вибір метрики (міри подібності) залежить від науковця та змістових даних, які потрібно дослідити, оскільки від даної величини залежить об'єктивність результату.

Стандартний процес кластеризації можна розділити на наступні кілька етапів [25]: виділення та вибір найбільш репрезентативних характеристик з вихідного набору даних; вибір алгоритму кластеризації та метрик відповідно до суті проблеми; оцінка результату кластеризації з метою визначення валідності алгоритму; пояснення практичного результату кластеризації. Алгоритм кластеризації можна представити у вигляді наступної схеми, представленої на рис. 1.

ISSN 2663 - 4023

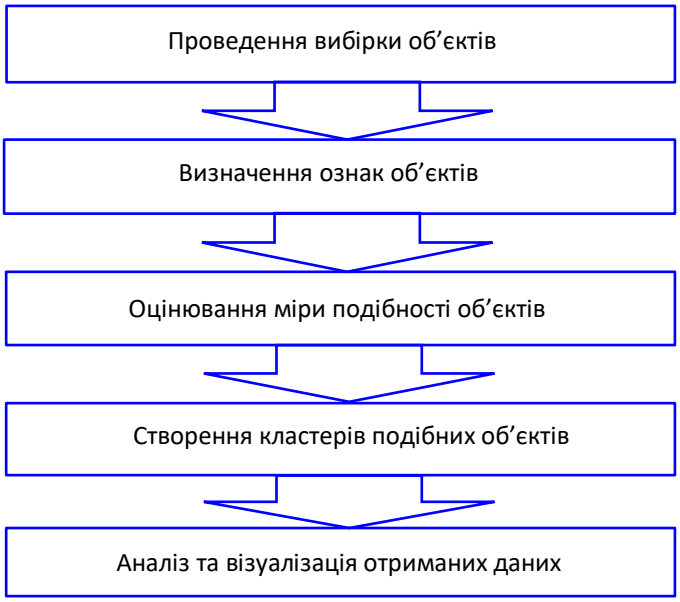

Рис. 1. Стандартний процес кластеризації

Розроблена велика кількість методів та алгоритмів кластерного аналізу [24] – [32]. Серед типів виділяють ієрархічні та неієрархічні, чіткі та нечіткі, алгоритми теорії графів та алгоритми, пов’язані з векторними моделями (рис. 2).

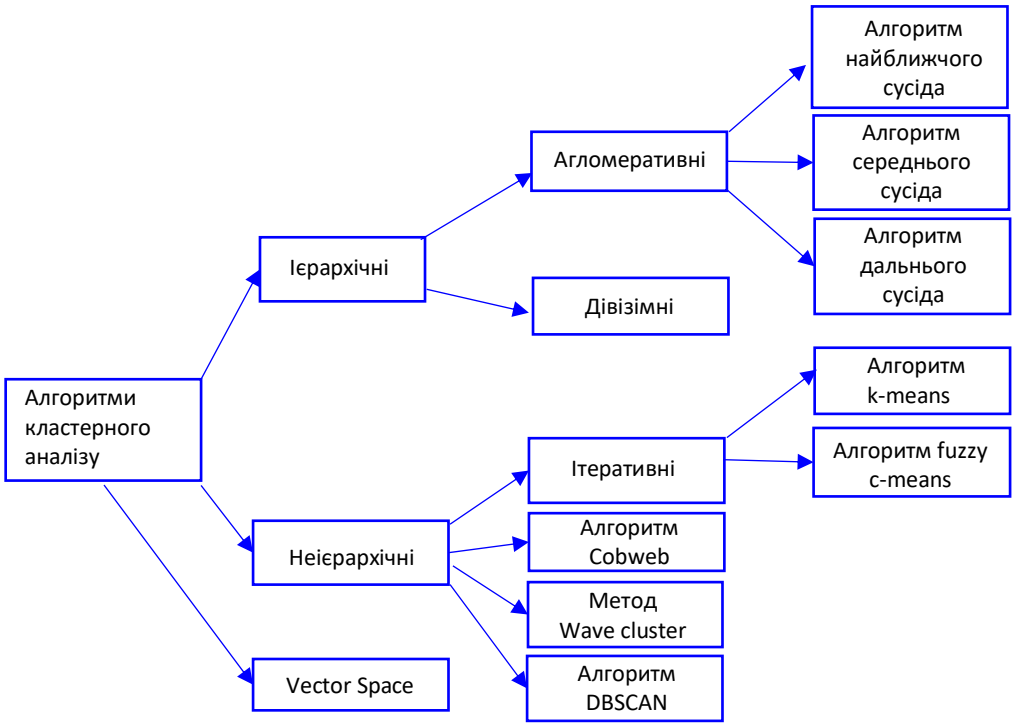

Рис. 2. Найбільш популярні алгоритми кластерного аналізу

У цьому дослідженні не ставилася задача здійснити аналіз усіх типів кластеризації. Надалі розглянемо найбільш поширені методи кластерного аналізу як алгоритм «найближчого сусіда», «k-means», «fuzzy c-means», «cosine similarity» та їх застосування в системах інформаційної та кібербезпеки.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Метод найближчого сусіда

Серед усіх методів кластерного аналізу найбільш поширеними є ієрархічні агломеративні методи. Їх суть полягає у послідовному об'єднанні двох найбільш подібних кластерів в один до тих пір, поки не буде утворено один кластер, що містить в собі всі об'єкти [24] – [27]. Дослідник розділяє цю сукупність на необхідну кількість кластерів на основі значення відстані між кластерами. Для графічної візуалізації об'єднання кластерів будують дендрограму. Серед агломеративних методів виділяють метод найближчого сусіда, метод середнього сусіда, метод дальнього сусіда.

Зупинимося на алгоритмі найближчого сусіда, процес якого містить наступні кроки:

- Крок 0. Нормування даних.
- Крок 1. Побудова матриці відстаней.
- Крок 2. Вибір початкової пари, найближчої одна до одної, їх об'єднання в один кластер і побудова нової матриці відстаней.
- Крок 3. Повторюємо дану операцію до тих пір, поки не будуть задіяні всі елементи. При цьому кожний вибір залишених елементів має здійснюватися за принципом найменшої відстані.
- Крок 4. Побудова дендрограми.

Блок-схема даного методу представлена на рис. 3.

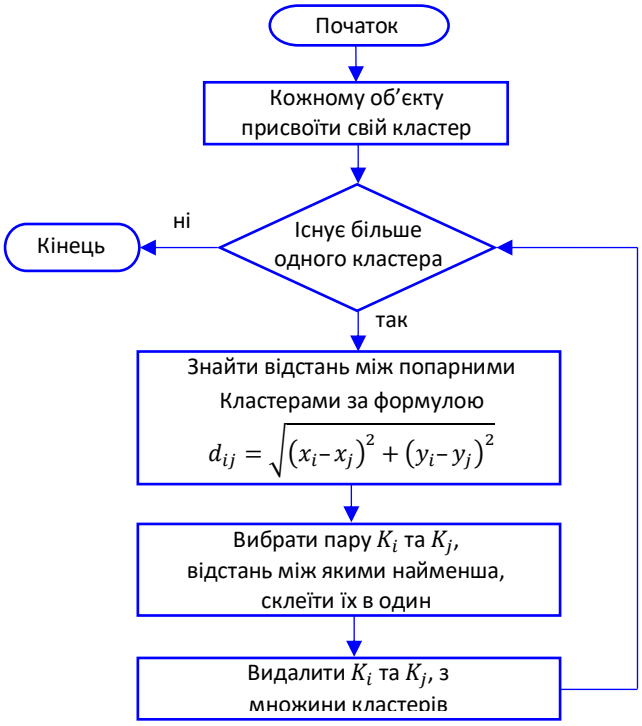

Рис. 3. Блок-схема алгоритму кластеризації методом найближчого сусіда

Як метрику для схожості (відстані) можна застосувати будь-яку, що представлені в табл. 1.

Застосування даного алгоритму в системах безпеки представлено у дослідженні [19] у процесі побудови автоматизованої системи керування кіберзахистом.

Автори у дослідженні [15] представили метод виявлення шкідливого програмного забезпечення на основі алгоритму k-найближчих сусідів, який здійснює класифікацію програмного забезпечення на шкідливе та аномальне. Цей же алгоритм застосовується для виявлення фішингу у роботі [16].

Метод k-means

Поряд з ієрархічними методами існує група ітеративних методів кластеризації. Суть їх полягає в тому, що процес починається з задання деяких початкових умов (кількості утворених кластерів, поріг завершення процесу кластеризації і тому подібне) [24] – [28]. До цих методів відноситься «k-means» та «fuzzy c-means». Алгоритм k-means передбачає таке розбиття множини на підмножини, при якому мінімізуються відстані між елементами однієї підмножини та максимізуються відстані між елементами різних підмножин. Процес виконання даного алгоритму наступний [26].

Нехай є m об'єктів, кожен з яких характеризується n ознаками $x_1, x_2, \dots, x_n$. Об'єкти необхідно розбити на k класів. Спочатку з m об'єктів випадково або виходячи а деяких міркувань вибирають k об'єктів. Ці об'єкти приймаються за еталони (тобто центри класів) $C_1^0, C_2^0, \dots, C_k^0$. Кожному еталону привласнюється порядковий номер, який одночасно є і номером класу. Вага кожного класу (кількість об'єктів, що входили до класу) спочатку дорівнює одиниці: $w_1^0 = 1, w_2^0 = 1, \dots, w_k^0 = 1$.

На першому кроці першої ітерації з $m-k$ об'єктів, що залишилися, вибираємо об'єкт n_{k+1} з координатами $(x_{k+1,1}, x_{k+1,2}, \dots, x_{k+1,n})$ і перевіряємо, до якого з еталонів (центрів) він знаходиться найближче. Для цього використовується одна з відомих метрик, наприклад, евклідова. Після приєднання об'єкту до класу еталон класу і його вага перераховуються таким чином:

$$C_j^1 = \frac{w_j^0 C_j^0 + n_{k+1}}{w_j^0 + 1}, w_j^1 = w_j^0 + 1 .$$

Якщо зустрічаються дві або більше мінімальних відстаней, об'єкт n_{k+1} приєднується до центру є найменшим порядковим номером. Далі вибираємо об'єкт n_{k+2} і для нього повторюємо всі вищезгадані процедури. Через $m-k$ кроків всі об'єкти сукупності будуть віднесені до одного з класів. Отримані еталони та відповідні ваги класів будуть початковими для наступної ітерації. Зазначимо, що перша ітерація містить в собі $m-k$ кроків.

Ітерації, починаючи з другої, відрізняються лише тим, що повторно розглядається вся сукупність об'єктів. Тобто за тим самим правилом всі об'єкти $n_1, n_2, \dots, n_m$ знову приєднуються до одержаних класів. При цьому ваги класів продовжують накопичуватися. Розбиття, отримане на $(i+1)$-й ітерації, порівнюється з розбиттям i-ї ітерації. Якщо вони співпадають, то класифікація об'єктів завершена. Інакше цикл перерозподілу об'єктів між класами повторюється.

Остаточне розбиття мас центри ваги, які можуть не співпадати з еталонами. Нехай центри ваги класів $c_1, c_2, \dots, c_k$. Тоді кожен об'єкт n_i, $(i = 1, 2, \dots, m)$, буде відноситись до того класу l, для якого $d(n_i, \mathrm{c}_l) = \min\limits_{1 \le j \le k} d(n_i, \mathrm{c}_j)$.

Блок-схема цього методу [28] представлена на рис. 4.

ISSN 2663 - 4023

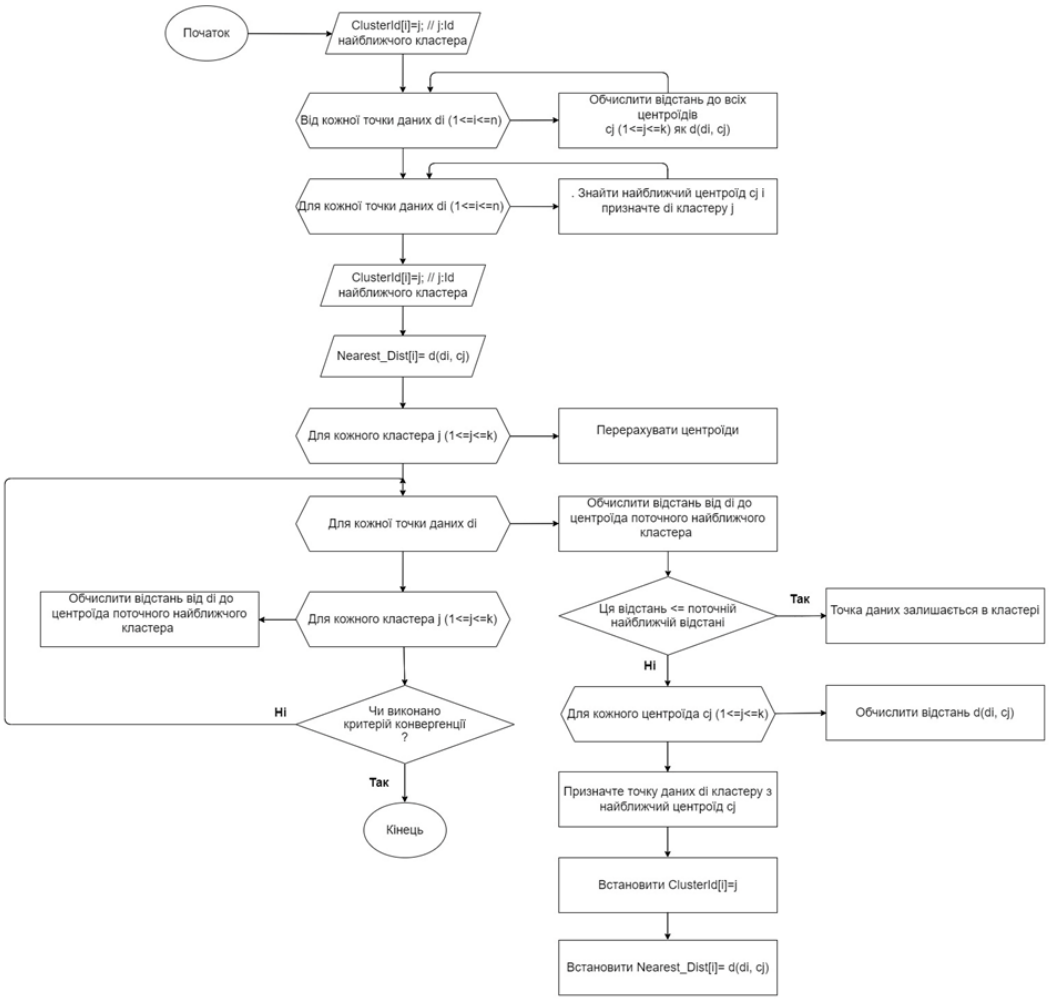

Рис. 4. Блок-схема алгоритму кластеризації методом k-means

Застосування даного методу у системах безпеки представлено у статті [9]. Досліджуючи мережеву безпеку, налаштовано чотири загальні моделі виявлення вторгнень, перша з яких використовує систему збору, гарантує записи з'єднань у процесі використання та збирає дані. На думку авторів кластеризація за допомогою методу k-means дозволить ці дані розрізняти на звичайні та аномальні записи підключення.

У дослідженні [14] науковці запропонували модель для виявлення поведінки зловмисного програмного забезпечення реєстру даних на основі особливостей зловмисного програмного забезпечення за допомогою підходу кластеризації *k*-means. Результати експерименту показують, що запропонована модель здатна кластеризувати звичайні та підозрілі дані у дві окремі групи з високим рівнем виявлення, який становить понад 90 % точності.

Кластеризація інформації про загрози методом *k*-means для зменшення інформаційного перевантаження для команд реагування на надзвичайні ситуації на комп'ютері висвітлено у роботі [17].

Метод fuzzy *c*-means

Одним із недоліків методу k-means є порушення умови зв'язності елементів одного кластера, тому розвиваються різні модифікації цього методу і нечіткі аналоги, до яких і відноситься метод fuzzy *c*-means. Він характеризується тим, що на першій стадії

алгоритму допускається приналежність одного елемента множини до декількох кластерів (із різним ступенем належності).

Даний метод був розроблений Дж. Данном у 1973 р. і вдосконалений Дж. Бездеком у 1981 р. Етапи кластеризації аналогічні попередньому методу з урахуванням того, що для кожного об'єкту випадковим чином визначається ймовірність належності заданим кластерам [29], [30].

Вхідні дані: кількість кластерів, коефіцієнт невизначеності m, коефіцієнт $\partial > 0$, який визначає точність алгоритму.

Крок 0. Обчислення коефіцієнта належності об'єкта k_i до кластеру c_j.

$$u_{ij} = \frac{1}{\sum_{i=1}^{n} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{1}{m-1}}}$$

Крок 1. Обчислення центрів кластерів

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}^m x_{ij}}{\sum_{i=1}^{n} u_{ij}^m}$$

де u_{ij} — коефіцієнт належності об'єкта x_i до кластера c_j.

Крок 2. Обчислення відстані від кожного x_i до центра кожного кластера c_j.

Крок 3. Обчислення та нормалізація коефіцієнтів належності x_i кластерам.

Крок 4. Обчислення значення матриці нечіткого розбиття та порівняння з таким значенням на попередній ітерації

$$\sum_{i=1}^{|x|} \sum_{j=1}^{c} u_{ij}^m \|x_i - c_j\|^2 .$$

Крок 5. Перевірка умови зупинки

$$max\left(\left| u_{ij}^{(k)} - u_{ij}^{(k-1)} \right| \right) < \partial,$$

де m — номер ітерації.

У дослідженні [20] автором представлена самоадаптивна система для забезпечення резильєнтності корпоративних мереж за наявності кібератак-мереж. Резильєнтність забезпечується адаптивним переконфігуруванням мережі, а переконфігурування мережі здійснюється із залученням сценаріїв безпеки, обраних на основі кластерного аналізу зібраних ознак Інтернет-трафіку, притаманних кібератакам. Для вибору необхідних сценаріїв безпеки запропонований метод використовує нечітку кластеризацію fuzzy c-means з частковим навчанням. У результаті експериментів було доведено, що дана модель демонструє здатність забезпечити стійке функціонування мережі в ситуації наявності кібератак бот-мереж на рівні 70 %.

Науковцями наробки [21] запропоновано алгоритм виявлення фішингових URL-адрес (класифікатор) із використанням нечіткої кластеризації, зокрема fuzzy c-means, який включає вибір типу інтелектуального класифікатора та обґрунтування його параметрів за допомогою методів глобальної оптимізації. У результаті моделювання було визначено, що усі фішингові URL-адреси, які помилково були класифіковані як безпечні, мали наявність захищеного з'єднання з дійсним сертифікатом.

Метод cosine similarity

У процесі розв'язання задач інформаційного пошуку та класифікації документів застосовують модель векторного простору (Vector Space Model) [31] – [34]. Модель векторного простору представляє документи як вектори в n-вимірному просторі, тобто

кожному документу ставиться у відповідність числовий вектор ознак $w(d) = \big(x(d,t_1), x(d,t_2), \ldots, x(d,t_n)\big)$. Розмірність кожного вектора визначається кількістю слів (можливо, слово — кількість літер). Найпростішим способом кодування документа є використання бінарних векторів: якщо слово присутнє в документі — 1, якщо ні — 0. Це кодування призведе до простого логічного порівняння або пошуку. Для підвищення точності використовують схеми зважування термінів: чим важливіше термін, тим він важче. Для обчислення ваги пропонується формула:

$$w(d,t) = \frac{tf(d,t)log(N/n_t)}{\sqrt{\sum_{i=1}^{n} tf(d,t_i)^2 log\big(N/n_{t_i}\big)^2}},$$

де N — розмір колекції документів D, n_t — кількість документів у D, які містять t. На основі схеми зважування документа d визначається вектором ваг $w(d) = \big(w(d,t_1), w(d,t_2), \ldots, w(d,t_n)\big)$. Подібність двох документів d_1 і d_2 можна обчислити, використовуючи метрику cosine similarity

$$s(d_1,d_2) = \sum_{k=1}^{n} w(d_1,t_k)\, w(d_2,t_k)\,.$$

Слід зазначити, якщо вектори будуть нормалізовані, то також можна застосувати евклідову метрику. Для текстової кластеризації можливо використати попередньо розглянуті алгоритми. Для формалізації текстових повідомлень, які представлені природньою мовою, існують спеціально розроблені програми WSord2Vec, CBOW та інші.

Метрику cosine similarity широко використовують у різних алгоритмах кластерного аналізу при обробці тексту у системах безпеки. Цій проблемі присвячено дослідження [18].

Переваги та недоліки охарактеризованих методів кластерного аналізу

Аналіз наукових джерел дозволив виділити переваги та недоліки висвітлених методів (табл. 2).

Таблиця 2

Переваги та недоліки найбільш популярних методів кластерного аналізу

Метод	Переваги	Недоліки
Метод найближчого сусіда	1) простота реалізації; 2) можливість наочного представлення результату у вигляді дендрограми.	1) зберігає всю вибірку об'єктів, що провокує витрати пам'яті; 2) працює значно повільніше зі збільшенням числа об'єктів; 3) не створює жодних моделей, які узагальнюють попередній досвід.
Метод k-means	1) зрозумілість та швидкість виконання; 2) можливість перевірки статистичної значимості відмінностей між виділеними кластерами.	1) задання кількості кластерів для розбиття перед кластеризацією; 2) залежність результату від визначення початкових центрів кластерів; 3) якщо серед об'єктів існує викид, то всі об'єкти будуть класифікуватися неправильно.
Метод fuzzy c-means	1) зрозумілість та швидкість виконання при дуже великих наборах числових даних; 2) менш чутливий до викидів; 3) відносно висока точність класифікації.	1) чутливий до початкових значень параметрів і кількості кластерів для розбиття; 2) вимагає великого обсягу обчислювальних ресурсів.

Vector Space Model з метрикою cosine similarity	1) не вимагає додаткової нормалізації даних перед розрахунками, що надає швидкість виконання; 2) зрозуміла інтерпретація.	1) не надає величину відмінностей між кластерами; 2) не враховує розмір вектора, а лише напрям.

У дослідженні [22] здійснено опитування щодо застосування методів кластеризації у системах безпеки. Результати дозволили зробити висновки, що підходи зазвичай переслідують одну або більше з чотирьох основних цілей: огляд і фільтрування, розбір та вилучення сигнатур, статичне виявлення викидів, а також послідовності та динамічне виявлення аномалій. Також викладено концепцію та інструмент, які підтримують вибір відповідних підходів на основі вимог, визначених користувачем.

Розвитку теорії сучасного кластерного аналізу сприяють методи штучного інтелекту. Їх впровадження дозволяє створювати аналітичну базу даних, на основі якої прогнозувати ризики інформаційної безпеки.

Так, в статті [10] автори пропонують практичне рішення для кластерного аналізу із збереженням конфіденційності на основі Homomorphic encryption (гомоморфного шифрування).

У дослідженні [12] науковці представили розроблену практичну систему HinCTI — систему моделювання та ідентифікації кіберзагроз на основі гетерогенної інформаційної мережі. Розроблена подібність інфраструктури загроз (MIIS) на основі металяхів і екземплярів мета-графів між вузлами інфраструктури загроз і здійснено підхід до згорткової мережі гетерогенних графів (GCN) на основі вимірювань MIIS для визначення типів загроз залучених вузлів інфраструктури в CTI.

Вчені у статті [13] описують користувачів на основі їх поведінки фізичних рухів і профілю роботи для ідентифікації користувачів з аномальною поведінкою фізичного доступу за допомогою алгоритму неконтрольованого машинного навчання, відомого як метод двоетапної кластеризації.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У роботі здійснено огляд методів кластеризації та внаслідок аналізу наукової літератури представлено шляхи застосування цих методів у системах безпеки. Дослідникам слід пам’ятати, що:

- процедура кластеризації завжди можлива, але змістова задача може бути не розв’язана;
- результат кластеризації істотно залежить від метрики і визначається експертом у даній галузі;
- не існує однозначно найкращого критерію якості кластеризації.

Автори не ставили за мету подати готові методи для розв’язання проблем в захисті інформації, головне — спонукати спеціалістів інформаційної безпеки до інтерпретації та досліджень в застосуванні даних алгоритмів. Адже дані методи тісно пов’язані з інформаційними технологіями, що сприяє впровадженню даних алгоритмів на основі штучного інтелекту. Розглянуті підходи можуть бути використані у процесі науково-дослідної роботи фахівців спеціальності 125 Кібербезпека та захист інформації. Перспективи подальших досліджень спрямовані на вивчення і впровадження інших методів кластерного аналізу в системи інформаційної та кібернетичної безпеки з метою ефективного ризик-орієнтованого управління.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Шевченко, С., Жданова, Ю., Спасітєлєва, С., Негоденко, О., Мазур, Н., & Кравчук, К. (2019). Математичні методи в кібербезпеці: фрактали та їх застосування в інформаційній та кібернетичній безпеці. *Кібербезпека: освіта, наука, техніка, 1(5)*, 31–39.
2. Шевченко, С., Жданова, Ю., Складанний, П., & Спасітєлєва, С. (2021). Математичні методи в кібербезпеці: графи та їх застосування в інформаційній та кібернетичній безпеці. *Кібербезпека: освіта, наука, техніка*, *1(13)*, 133–144.
3. Шевченко, С., Складанний, П., Негоденко, О., & Негоденко, В. (2022). Дослідження прикладних аспектів теорії конфліктів у системах безпеки. *Кібербезпека: освіта, наука, техніка*, *2(18)*, 150–162.
4. Shevchenko, S., et al. (2023). Conflict Analysis in the Information Security System: Subject – Subject. *CEUR Workshop Proceedings*, *3421.* 56–66.
5. Шевченко, С., Жданова, Ю., & Спасітєлєва, С. (2023). Математичні методи в кібербезпеці: теорія катастроф. *Кібербезпека: освіта, наука, техніка*, *3(19)*, 165–175.
6. Шевченко, С., Жданова, Ю., Складанний, П., & Бойко, С. (2023). Теоретико-ігровий підхід до моделювання конфліктів у системах інформаційної безпеки. *Кібербезпека: освіта, наука, техніка*, *2(22)*, 168–178.
7. Levkin, D., Zhernovnykova, O., & Kotko, Y. (2023). Modern mathematical methods in the cyber security system. Mechanisms for ensuring sustainable development of the economy: problems, prospects, international experience. *Materials of the IV international scientific and practical Internet conference*.
8. Лисенко, Н., Мазуренко, В., Федорович, А., Астахов, Д., & Стаценко В. (2021). Огляд математичних методів у системах виявлення та попередження кіберзагроз. *Актуальні проблеми автоматизації та інформаційних технологій*, *(25)*, 91–102. http://dx.doi.org/10.15421/432110
9. Bu, C. (2018). Network Security Based on K-Means Clustering Algorithm in Data Mining Research. *Advances in Computer Science Research*, *83*, 642–645. https://doi.org/10.2991/snce-18.2018.130
10. Cheon, J., Kim, D., & Park, J. (2019). Towards a Practical Cluster Analysis over Encrypted Data. *Conference: Selected Areas in Cryptography (SAC)*, 1–24.
11. Raptis, G., Katsini, C., & Alexakos, C. (2021). Towards Automated Matching of Cyber Threat Intelligence Reports based on Cluster Analysis in an Internet-of-Vehicles Environment, *2021 IEEE International Conference on Cyber Security and Resilience (CSR)*, 366–371, https://doi.org/10.1109/CSR51186.2021.9527983
12. Gao, Y., et al. (2022). HinCTI: A Cyber Threat Intelligence Modeling and Identification System Based on Heterogeneous Information Network. *IEEE Transactions on Knowledge and Data Engineering*, *34(2)*, 708–722. https://doi.org/10.1109/TKDE.2020.2987019
13. Poh, J., et al. (2020). Physical Access Log Analysis: An Unsupervised Clustering Approach for Anomaly Detection. *DSIT 2020: Proceedings of the 3rd International Conference on Data Science and Information Technology*, 12–18. https://doi.org/10.1145/3414274.3414285
14. Rosli, N., et al. (2019). Clustering Analysis for Malware Behavior Detection using Registry Data. *International Journal of Advanced Computer Science and Applications (IJACSA), 10(12).* http://dx.doi.org/10.14569/IJACSA.2019.0101213
15. Лисенко, С., & Гуменюк, В. (2017). Метод виявлення шкідливих програмних засобів на основі алгоритму найближчих сусідів. *Вісник Хмельницького національного університету, 6, 2017(255)*, 96–101.
16. REDDY K.T. (2023). *Unveiling the Power of k-Nearest Neighbors in Phishing Detection, Insights2Techinfo*. https://insights2techinfo.com/unveiling-the-power-of-k-nearest-neighbors-in-phishing-detection/
17. Kuehn, P., et al. (2022). *Clustering of Threat Information to Mitigate Information Overload for Computer Emergency Response Teams*. https://arxiv.org/abs/2210.14067
18. Patton, R., et al. (2011). Hierarchical clustering and visualization of aggregate cyber data. *2011 7th International Wireless Communications and Mobile Computing Conference*, 1287–1291. https://doi.org/10.1109/IWCMC.2011.5982725
19. Довбиш, А., Ободяк, В., Шелехов, І., & Великодний, Д. (2021). Основи інформаційно-екстремального синтезу автоматизованої системи керування кіберзахистом. *Сучасні інформаційні технології в кібербезпеці*, 7–75.
20. Лисенко, С. (2019). Метод забезпечення резильєнтності комп'ютерних систем в умовах кіберзагроз на основі самоадаптивності. *Радіоелектронні і комп'ютерні системи*, *4(92)*, 4–16.
21. Герасіна, О., Корнієнко, В., Гусєв, О., Соснін, К., & Мацюк, С. (2022). Виявлення фішингових URL-адрес за допомогою алгоритмів нечіткої кластеризації із глобальною оптимізацією. *Системні технології. Регіональний міжвузівський збірник наукових праць*, *2(139)*, 53–67.

22. Landauer, M., et al. (2020). System log clustering approaches for cyber security applications: A survey. *Computers & Security*, *92*, 1–18. https://doi.org/10.1016/j.cose.2020.101739
23. Гончаренко, С. (1997). *Український педагогічний словник*. Либідь.
24. Jain, A., & Dubes, R. (1988). *Algorithms for clustering dat*a. Prentice-Hall, Inc, Upper Saddle River.
25. Xu, R., & Wunsch, D. (2005) Survey of clustering algorithms. *IEEE Transactions on Neural Networks, 16(3),* 645–678. https://doi.org/10.1109/TNN.2005.845141
26. Яровий, А., & Страхов, Є. (2015). *Багатовимірний статистичний аналіз: начально-методичний посібник для студентів математичних та економічних фахів*. Астропринт.
27. Xu, D., & Tian, Y. (2015). Comprehensive Survey of Clustering Algorithms. *Ann. Data. Sci. 2*, 165–193. https://doi.org/10.1007/s40745-015-0040-1
28. Abdul Nazeer, K., & Sebastian, M. (2009). Improving the Accuracy and Efficiency of the *k*-means Clustering Algorithm. *Proceedings of the World Congress on Engineering, I.*
29. Dunn, J. (1973) A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics, 3,* 32–57. http://dx.doi.org/10.1080/01969727308546046
30. Bezdek, J. (1981). *Pattern recognition with fuzzy objective function algorithms*. Plenum Press.
31. Chen, Z. (2022) Research and Application of Clustering Algorithm for Text Big Data. *Comput Intell Neurosci*. https://doi.org/10.1155/2022/7042778
32. Salton, G. (1988). *Automatic text processing*. Addison-Wesley Longman Publishing.
33. Sidorov, G., et al. (2014). Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model. *Computación y Sistemas*, *18(3),* 491–504. https://doi.org/10.13053/CyS-18-3-2043
34. Vijaymeena, M., & Kavitha, K. (2016). A Survey on Similarity Measures in Text Mining. *Machine Learning and Applications: An International Journal*, *3,* 19–28. https://doi.org/10.5121/mlaij.2016.3103

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 3 (23), 2024
ISSN 2663 - 4023

Svitlana Shevchenko
PhD, Associate Professor,
Associate Professor of the Department of Information and Cyber Security
named after Professor Volodymyr Buriachok
Borys Grinchenko Kyiv Metropolitan University, Kyiv, Ukraine
ORCID 0000-0002-9736-8623
s.shevchenko@kubg.edu.ua

Yuliia Zhdanova
PhD, Associate Professor,
Associate Professor of the Department of Information and Cyber Security
named after Professor Volodymyr Buriachok
Borys Grinchenko Kyiv Metropolitan University, Kyiv, Ukraine
ORCID 0000-0002-9277-4972
y.zhdanova@kubg.edu.ua

Svitlana Spasiteleva
PhD, Associate Professor,
Associate Professor of the Department of Information and Cyber Security
named after Professor Volodymyr Buriachok
Borys Grinchenko Kyiv Metropolitan University, Kyiv, Ukraine
ORCID 0000-0003-4993-6355
s.spasitielieva@kubg.edu.ua

Nataliia Mazur
PhD, Associate Professor
Associate Professor of the Department of Information and Cyber Security
named after Professor Volodymyr Buriachok
Borys Grinchenko Kyiv Metropolitan University, Kyiv, Ukraine
ORCID 0000-0001-7671-8287
n.mazur@kubg.edu.ua

Pavlo Skladannyi
PhD, Associate Professor,
Head of the Department of Information and Cybersecurity
named after Professor Volodymyr Buriachok
Borys Grinchenko Kyiv Metropolitan University, Kyiv, Ukraine
ORCID 0000-0002-7775-6039
p.skladannyi@kubg.edu.ua

Vitalii Nehodenko
Graduate Student of the Department of Information and Cybersecurity
named after Professor Volodymyr Buriachok
Borys Grinchenko Kyiv Metropolitan University, Kyiv, Ukraine
ORCID 0000-0002-7678-9138
v.nehodenko.asp@kubg.edu.ua

MATHEMATICAL METHODS IN CYBER SECURITY: CLUSTER ANALYSIS AND ITS APPLICATION IN INFORMATION AND CYBERNETIC SECURITY

Abstract. The huge number of information threats and their complexity prompts research and modeling of new methodologies and information protection systems. The development and improvement of information and cyber security systems includes the creation and processing of mathematical models using information technologies. This article is a follow-up study on the application of mathematical methods and technologies in cyber security, namely: methods of cluster analysis. The modern development of computer technology and the growth of their power have contributed to the wide implementation of Data Mining algorithms for processing large volumes of information in various fields of society and science, in particular in the field of cyber security.

Cluster analysis allows the set to be divided into subsets, so that the elements of each subset are similar to each other, and the elements of different subsets are the most different. This provides an opportunity to eliminate the shortcomings of the qualitative approach in assessing information risks. The paper reviews scientific sources regarding the applied aspect of the application of clustering methods in security systems, because timely forecasting of possible incidents allows you to manage information risks and make effective decisions to ensure confidentiality, availability and integrity of information. The stages of the clustering procedure are characterized, the issues of choosing the distance measure and the similarity measure for the objects under study are highlighted. The comparative characteristics of the most popular methods of cluster analysis are presented: the "nearest neighbor" algorithm, "k-means", "fuzzy c-means", "cosine similarity", their advantages and disadvantages are defined. This study can be useful and used in the educational process of students of the specialty 125 "Cyber security and information protection".

Keywords: mathematical methods; cluster analysis; informational security; cyber security; "nearest neighbor" algorithm, "k-means" algorithm, "fuzzy c-means" algorithm, "cosine similarity" algorithm.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Shevchenko, S., et al. (2019) Mathematical Methods in Cybersecurity: Fractals and their Applications in Information And Cyber Security. *Cybersecurity: education, science, technique, 1(5),* 31–39.
2. Shevchenko, S., et al. (2021). Mathematical Methods in Cibersecurity: Graphs and their Application in Information and Cybernetic Security. *Cybersecurity: education, science, technique, 1(13),* 133–144.
3. Shevchenko, S., et al. (2022). Study of applied aspects of conflict theory in security systems. *Cybersecurity: education, science, technique, 2(18),* 150–162.
4. Shevchenko, S., et al. (2023). Conflict Analysis in the Information Security System: Subject – Subject. *CEUR Workshop Proceedings, 3421.* 56–66.
5. Shevchenko, S., Zhdanova, Yu., & Spasiteleva, S. (2023) Mathematical Methods in Cybersecurity: Catastrophe Theory. *Cybersecurity: education, science, technique, 3(19),* 165–175.
6. Shevchenko, S., et al. (2023) Game Theoretical Approach to the Modeling Of Conflicts in Information Security Systems. *Cybersecurity: education, science, technique, 2(22),* 168–178.
7. Levkin, D., Zhernovnykova, O., & Kotko, Y. (2023). Modern mathematical methods in the cyber security system. Mechanisms for ensuring sustainable development of the economy: problems, prospects, international experience. *Materials of the IV international scientific and practical Internet conference.*
8. Lysenko, N., et al. (2021) Review of Mathematical Methods in Cyber Threat Detection and Prevention Systems. *Actual problems of automation and information technology*, 25, 91–102. http://dx.doi.org/10.15421/432110
9. Bu, C. (2018). Network Security Based on K-Means Clustering Algorithm in Data Mining Research. *Advances in Computer Science Research, 83,* 642–645. https://doi.org/10.2991/snce-18.2018.130
10. Cheon, J., Kim, D., & Park, J. (2019). Towards a Practical Cluster Analysis over Encrypted Data. *Conference: Selected Areas in Cryptography (SAC),* 1–24.
11. Raptis, G., Katsini, C., & Alexakos, C. (2021). Towards Automated Matching of Cyber Threat Intelligence Reports based on Cluster Analysis in an Internet-of-Vehicles Environment, *2021 IEEE International Conference on Cyber Security and Resilience (CSR),* 366–371, https://doi.org/10.1109/CSR51186.2021.9527983
12. Gao, Y., et al. (2022). HinCTI: A Cyber Threat Intelligence Modeling and Identification System Based on Heterogeneous Information Network. *IEEE Transactions on Knowledge and Data Engineering, 34(2),* 708–722. https://doi.org/10.1109/TKDE.2020.2987019
13. Poh, J., et al. (2020). Physical Access Log Analysis: An Unsupervised Clustering Approach for Anomaly Detection. *DSIT 2020: Proceedings of the 3rd International Conference on Data Science and Information Technology*, 12–18. https://doi.org/10.1145/3414274.3414285
14. Rosli, N., et al. (2019). Clustering Analysis for Malware Behavior Detection using Registry Data. *International Journal of Advanced Computer Science and Applications (IJACSA), 10(12).* http://dx.doi.org/10.14569/IJACSA.2019.0101213
15. Lysenko, S., & Humenyuk, V. (2017). Malware detection method based on the nearest neighbor algorithm. *Bulletin of the Khmelnytskyi National University, 6, 2017 (255),* 96–101.
16. REDDY K.T. (2023). *Unveiling the Power of k-Nearest Neighbors in Phishing Detection, Insights2Techinfo*. https://insights2techinfo.com/unveiling-the-power-of-k-nearest-neighbors-in-phishing-detection/

17. Kuehn, P., et al. (2022). *Clustering of Threat Information to Mitigate Information Overload for Computer Emergency Response Teams*. https://arxiv.org/abs/2210.14067
18. Patton, R., et al. (2011). Hierarchical clustering and visualization of aggregate cyber data. *2011 7th International Wireless Communications and Mobile Computing Conference*, 1287–1291. https://doi.org/10.1109/IWCMC.2011.5982725
19. Dovbysh, A., et al. (2021). Fundamentals of information-extreme synthesis of an automated cyber defense control system. *Modern information technologies in cyber security*, 7–75.
20. Lysenko, S. (2019). A method of ensuring the resilience of computer systems in the face of cyber threats based on self-adaptability. *Radioelectronic and computer systems, 4(92),* 4–16.
21. Gerasina, O., et al. (2022).Detecting fishing URLs using fuzzy clustering algorithms with global optimization. *System technologies, 2(139),* 53–67.
22. Landauer, M., et al. (2020). System log clustering approaches for cyber security applications: A survey. *Computers & Security, 92,* 1–18. https://doi.org/10.1016/j.cose.2020.101739
23. Goncharenko, S. (1997). *Ukrainian Pedagogical Dictionary*. Lybid.
24. Jain, A., & Dubes, R. (1988). *Algorithms for clustering dat*a. Prentice-Hall, Inc, Upper Saddle River.
25. Xu, R., & Wunsch, D. (2005) Survey of clustering algorithms. *IEEE Transactions on Neural Networks, 16(3),* 645–678. https://doi.org/10.1109/TNN.2005.845141
26. Yarovy, A., & Strakhov, E. (2015). *Multivariate statistical analysis: an introductory methodological guide for students of mathematics and economics.* Astroprint.
27. Xu, D., & Tian, Y. (2015). Comprehensive Survey of Clustering Algorithms. *Ann. Data. Sci. 2,* 165–193. https://doi.org/10.1007/s40745-015-0040-1
28. Abdul Nazeer, K., & Sebastian, M. (2009). Improving the Accuracy and Efficiency of the *k*-means Clustering Algorithm. *Proceedings of the World Congress on Engineering, I.*
29. Dunn, J. (1973) A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics, 3,* 32–57. http://dx.doi.org/10.1080/01969727308546046
30. Bezdek, J. (1981). *Pattern recognition with fuzzy objective function algorithms*. Plenum Press.
31. Chen, Z. (2022) Research and Application of Clustering Algorithm for Text Big Data. *Comput Intell Neurosci*. https://doi.org/10.1155/2022/7042778
32. Salton, G. (1988). *Automatic text processing.* Addison-Wesley Longman Publishing.
33. Sidorov, G., et al. (2014). Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model. *Computación y Sistemas, 18(3),* 491–504. https://doi.org/10.13053/CyS-18-3-2043
34. Vijaymeena, M., & Kavitha, K. (2016). A Survey on Similarity Measures in Text Mining. *Machine Learning and Applications: An International Journal, 3,* 19–28. https://doi.org/10.5121/mlaij.2016.3103