КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (26), 2024
ISSN 2663 - 4023
DOI 10.28925/2663-4023.2024.26.790
УДК 004.9:336.77:005.74

Редько Денис Вадимович
аспірант кафедри інженерії програмного забезпечення та кібербезпеки
Державний торговельно-економічний університет, Київ, Україна
ORCID ID: 0009-0003-5827-264X
d.redko@knute.edu.ua

Десятко Альона Миколаївна
доктор філософії «Комп'ютерні науки», доцент
завідувач кафедри інженерії програмного забезпечення та кібербезпеки
Державний торговельно-економічний університет, Київ, Україна
ORCID ID: 0000-0002-2284-3418
desyatko@gmail.com

АНСАМБЛЬ АЛГОРИТМІВ КЛАСТЕРИЗАЦІЇ З РІЗНИМИ МЕТРИКАМИ ДЛЯ АНАЛІЗУ МЕРЕЖЕВОГО ТРАФІКУ БАНКУ

Анотація. Актуальність дослідження зумовлена необхідністю підвищення точності аналізу мережевого трафіку банку, представленого гетерогенними даними, що включають логи серверів, мережеві з'єднання, поведінкові дані користувачів і телеметрію трафіку. В умовах зростання обсягів даних і ускладнення структури корпоративних мереж українських банків традиційні методи аналізу втрачають ефективність, роблячи релевантним застосування методів машинного навчання, включаючи ансамблеві алгоритми кластеризації. Сучасні дослідження в цій області зосереджуються на розробці адаптивних методів сегментації, що використовують ймовірнісні моделі та динамічне зважування алгоритмів для підвищення точності розбиття даних. У роботі запропоновано ансамблевий метод КА, що використовує варіації алгоритму K-means з різними метриками відстані. Метод заснований на ймовірнісній моделі, яка враховує латентні класи та динамічно налаштовувані ваги алгоритмів, а для узгодженості розбиття задіяна коасоціативна матриця, що відображає частоту потрапляння пар об'єктів в один кластер. Новизна дослідження полягає в розробці механізму адаптивного зважування алгоритмів ансамблю на основі метрик якості (ARI, NMI), що дозволяє підвищити точність кластеризації мережевого трафіку банку. Запропонований метод компенсує недоліки окремих алгоритмів і враховує особливості різних типів даних (пакети, з'єднання, користувачі, пристрої). Практична значущість роботи підтверджується можливістю застосування методу для аналізу трафіку банківських мереж, виявлення аномалій, оптимізації маршрутизації та підвищення кібербезпеки банківської мережі, а розроблений підхід дозволить враховувати складну структуру даних і адаптуватися до динамічних змін мережевого оточення банку.

Ключові слова: кібербезпека; кластерний аналіз; ансамблевий алгоритм; метрики відстані; коасоціативна матриця; банківські мережі; аналіз мережевого трафіку; ймовірнісні моделі.

ВСТУП

Ансамбль алгоритмів кластеризації з різними метриками — ефективний підхід до аналізу мережевого трафіку банківських систем. Причому в сучасних IT-інфраструктурах банків, у тому числі працюючих на ринку України, великі дані (BD) використовуються для оптимізації та масштабування корпоративних мереж, виявлення аномалій і забезпечення кібербезпеки [1] – [3]. Для точного аналізу трафіку його

необхідно структурувати і класифікувати. У даній роботі розглядається ансамблевий метод кластеризації [4], [5], заснований на варіаціях алгоритму K-means з різними метриками відстані. Об'єктами аналізу можуть бути пакети даних, мережеві з'єднання, пристрої, користувачі і сегменти трафіку. Вибір метрики дозволяє враховувати специфічні характеристики кожного об'єкта, такі як IP-адреси, порти, протоколи, обсяг і частота переданих даних. Використання ансамблю алгоритмів забезпечує більш точне виявлення закономірностей та аномалій у мережевому трафіку банку, що критично для його оптимізації, масштабування ресурсів та підвищення рівня безпеки. Такий підхід компенсує недоліки окремих методів і підвищує надійність кластеризації.

Постановка проблеми. Сучасні банківські мережі генерують великі обсяги гетерогенних даних, що включають журнали подій, дані маршрутизації, телеметрію мережевого трафіку, логи серверів, інформацію від мережевих датчиків (IDS/IPS) та поведінкові дані користувачів. Аналіз цих даних необхідний для виявлення аномалій, запобігання кібератакам, оптимізації ресурсів та підвищення ефективності роботи мережі. Класичні алгоритми кластеризації, такі як K-means, DBSCAN та агломеративна кластеризація, мають обмеження, пов'язані з вибором метрики відстані та чутливістю до параметрів. У результаті можливі помилки групування об'єктів, що знижує точність аналізу. Тому для підвищення надійності сегментації мережевого трафіку потрібен ансамблевий підхід, що дозволяє комбінувати результати декількох методів КА, мінімізуючи помилки розбиття. Проблема полягає в розробці та обґрунтуванні методу ансамблевої кластеризації мережевого трафіку банку, що забезпечує високу точність виявлення закономірностей у гетерогенних даних.

Аналіз останніх досліджень і публікацій. У науковій літературі [6] – [12] описані концептуальні підходи до ансамблевого кластерного аналізу (КА). Перший базується на досягненні консенсусу між результатами окремих алгоритмів, мінімізуючи їх розбіжності. Другий використовує коасоціативні матриці, що відображають частоту потрапляння пар об'єктів в один кластер при різних розбиттях. При аналізі мережевого трафіку банку перший підхід передбачає застосування декількох алгоритмів КА (наприклад, k-means, агломеративної кластеризації та DBSCAN) з подальшою оцінкою метрик схожості та зваженим об'єднанням результатів. У другому підході обчислюються коасоціативні матриці для IP-адрес або інших мережевих об'єктів, після чого підсумкове розбиття формується методами спектральної кластеризації.

Альтернативно, можна використовувати ймовірнісні моделі, такі як суміші розподілів, з EM-алгоритмом для оцінки параметрів. В обох випадках принципову роль відіграють ваги w, що визначають внесок кожного алгоритму в підсумкове розбиття. У запропонованому методі ансамблевого КА враховуються латентні класи та динамічно налаштовувані ваги, що залежать від метрик якості (точності, повноти, специфічності). Такий підхід підвищує адаптивність аналізу мережевого трафіку банку, компенсуючи слабкі сторони окремих алгоритмів і підвищуючи надійність кластеризації великих даних.

Мета статті. Удосконалити ансамблевий метод кластерного аналізу мережевого трафіку банку, заснований на використанні різних метрик відстані та механізму динамічного зважування алгоритмів, для підвищення точності сегментації та виявлення аномалій у гетерогенних даних.

МЕТОДИ ТА МОДЕЛІ

Методи дослідження

Дослідження базується на застосуванні ансамблевого методу КА мережевого трафіку банку, що об'єднує результати декількох алгоритмів. Використовуються модифікації алгоритму K-means з різними метриками відстані (евклідова, косинусна, Мінковського), а також агломеративні алгоритми та DBSCAN, що дозволяє враховувати гетерогенність даних. Для підвищення точності введена ймовірнісна модель з динамічними вагами алгоритмів, визначеними на основі метрик якості (ARI, NMI). Коасоціативна матриця використовується для узгодження розбиття та підвищення стійкості методу. Аналізовані дані включають журнали подій, телеметрію, маршрутизаційні дані та поведінкові характеристики користувачів. Застосовано методи нормалізації ознак, кодування категоріальних даних і виявлення викидів. Оцінка ефективності проведена на реальних даних банківської мережі, що дозволило перевірити точність, надійність і обчислювальну складність алгоритму, а також його застосовність для аналізу великомасштабного трафіку.

Модель

При аналізі мережевого трафіку великої корпоративної мережі, наприклад банківської, розглядаються гетерогенні дані, що характеризуються різноманітністю типів, форматів і джерел. Під гетерогенними даними розуміється сукупність різнорідних структур, включаючи журнали подій, телеметрію мережевого трафіку, відомості від мережевих датчиків (наприклад, IDS/IPS-систем), інформацію про користувачів, маршрутизаційні дані та інші види даних. Гетерогенність мережевого трафіку обумовлена кількома факторами. По-перше, відмінності в рівнях мережевої моделі, на яких збираються дані (від фізичного до прикладного), призводять до варіативності їх структури та змісту. По-друге, різнорідність мережевих протоколів, що включає як класичні (HTTP, SMTP), так і сучасні сервіси (VoIP, відеоконференції), а також внутрішні корпоративні застосунки, призводить до ускладнення аналізу. Ця гетерогенність проявляється в різних аспектах:

- Тип даних, наприклад, числові (обсяг трафіку, час відгуку), категоріальні (протокол, тип пристрою), текстові (вміст пакетів), часові (мітки часу).
- Масштабність від окремих пакетів до агрегованих даних за тривалі часові періоди.
- Джерела, тобто різне розташування датчиків у мережі.
- Якість даних, тобто можливі шуми, пропуски та помилки вимірювань.

Складність аналізу таких даних обумовлена необхідністю обробки багатьох факторів, що робить кластерний аналіз (КА) ефективним інструментом для виявлення прихованих закономірностей. Однак застосування єдиного алгоритму КА може бути недостатньо інформативним, оскільки різні методи мають різну чутливість до особливостей даних. У зв'язку з цим доцільно використовувати ансамблевий підхід, що передбачає комбінацію алгоритмів з різними метриками відстані, що дозволяє враховувати специфіку кожного типу даних, забезпечуючи більш точну сегментацію мережевого трафіку та його оптимізацію.

Одним із можливих варіантів реалізації ансамблевого методу є модифікація базового алгоритму K-means з використанням різних метрик відстані, адаптованих під особливості аналізованих даних.

Введемо такі позначення для формалізації запропонованого методу побудови колективного (ансамблевого) рішення, що враховує ваги різних алгоритмів кластеризації даних про мережевий трафік банку.

Нехай існує множина об'єктів $S = \{o^{(1)},...,o^{(N)}\}$, наприклад, таких як IP-адреси, порти, протокол, розмір пакета, серверні логи, дані про трафік, дані від мережевих датчиків (наприклад, IDS/IPS системи, DLP системи тощо), дані про користувачів, дані маршрутизації, і т.д., випадковим і незалежним чином вибраних з генеральної сукупності. Потрібно розбити ці об'єкти на задану кількість C кластерів. Для розбиття використовуємо критерій якості групування. Кожен об'єкт описується набором дійсних змінних $X_1,...,X_n$.

Пояснимо заданий набір на кількох прикладах. Скажімо, хоча IP-адреси зазвичай і представлені в текстовому форматі (наприклад, 192.168.1.1), їх можна перетворити в числові значення для аналізу. Наприклад, IPv4-адрес може бути представлений як 32-бітне ціле число, що дозволяє використовувати його в розрахунках. Номери портів є цілими числами в діапазоні від 0 до 65535, що робить їх природними дійсними змінними для аналізу. Протоколи мережевого рівня (наприклад, TCP, UDP) можна кодувати числовими значеннями. Наприклад, TCP може бути представлений як 1, а UDP як 2, що дозволяє використовувати ці дані в числових алгоритмах. Розмір пакета — це дійсна змінна, яка вказує кількість байт у пакеті, ця інформація часто використовується для аналізу продуктивності мережі та виявлення аномалій.

Ці дані можуть включати різні метрики, такі як обсяг переданих даних, кількість пакетів і час передачі, які можуть бути представлені у вигляді дійсних змінних. Інформація від мережевих датчиків (наприклад, IDS/IPS, DLP систем) генерується у вигляді даних про події безпеки, які можуть включати часові мітки, типи атак, рівень серйозності та інші параметри, які можуть бути числовими. Дані про користувачів можуть включати ідентифікатори користувачів, кількість сеансів, тривалість сеансів та інші метрики, які можна перетворити в числові значення. Метрики маршрутизації, такі як метрика вартості маршруту, час у дорозі, і т.д. можуть бути представлені числовими значеннями. Ці та інші об'єкти перетворюються в набір дійсних змінних для того, щоб застосовувати алгоритми кластеризації та інші методи аналізу даних. Дійсні змінні дозволять легко використовувати математичні операції та алгоритми, що зробить можливим виявлення прихованих закономірностей, а також сприятиме розробці більш ефективних методів аналізу та обробки BD на основі кластерних колективних рішень.

Позначимо через $x = x(o) = (x_1,...,x_n)$ вектор змінних для об'єкта (o). Тут $x_j = X_j(o),\ j = 1,...n.$ Тобто вектор змінних для кожного об'єкта (o) буде представляти собою числовий опис (o), де кожна змінна відповідає певній характеристиці або ознаці об'єкта. Наприклад, IP-адрес 192.168.1.1 може бути представлений як чотири числових значення. Відповідно, 192, 168, 1, 1. Номер порту, наприклад, 443 для HTTPS представляється одним числовим значенням. Розмір пакета, наприклад, 1500 байт представлений одним числовим значенням. І так далі. Тоді, x_N — це таблиця даних $\left(x\left(o^{(1)},...,o^{(N)}\right)\right)^T$. Ця таблиця, див., наприклад, табл. 1, містить інформацію про всі розглянуті об'єкти (o). Наприклад, IP-адреси, порти, протоколи, розміри пакетів і так далі, де кожен рядок представляє собою вектор змінних для одного (o).

Таблиця 1

Приклад таблиці для кластерного аналізу мережевого трафіку

IP-адреса джерела	IP-адреса призначення	Порт джерела	Порти призначення	Протокол	Розмір пакета	Часовий штамп	Інші параметри
192.168.1.100	8.8.8.8	54321	53	UDP	100	2024-11-22 10:15:01	…
10.0.0.1	192.168.1.1	80	80	TCP	1500	2025-01-22 08:10:02	…
192.168.1.1	172.16.0.5	55432	80	TCP	500	2025-02-22 9:06:03	…
…	…	…	…	…	…	…	…

Припускаємо, що для розгалуженої топології мережі банку може існувати прихована змінна Y . Коли ми говоримо про приховану (безпосередньо ненаблюдаєму) змінну $Y \in (1,...,CL)$, яка визначає належність кожного об'єкта до певного класу (CL), то маємо на увазі, що кожен об'єкт у нашому наборі даних належить до одного з кількох можливих класів. Ці класи не видно безпосередньо в даних і повинні бути визначені на основі спостережуваних характеристик об'єктів.

Як було показано в [6], [12], клас (або кластер) характеризується умовним розподілом. Це справедливо, оскільки дозволяє описувати, як розподілені значення ознак об'єктів, що належать до цього класу. Тобто кожен клас або кластер має унікальні характеристики, наприклад, у задачі аналізу мережевого трафіку нормальний трафік і зловмисний трафік матимуть різні розподіли за такими ознаками, як розмір пакетів, використовувані порти, IP-адреси тощо.

Такі відмінності описуються умовними розподілами [10] – [12]. Тоді, справедливо $p(x|Y = cl) = p_{cl}(x), cl = 1,...,CL.$ Умовний розподіл $p(x|Y = 0)$ може показувати, що порти 80 і 443 частіше використовуються для HTTP/HTTPS, розміри пакетів слідують нормальному розподілу з певними параметрами. А умовний розподіл $p(x|Y = 1)$ може характеризуватися частим використанням незвичайних портів, більшою дисперсією в розмірах пакетів і специфічними IP-адресами, які часто беруть участь в атаках. Розуміння умовних розподілів допоможе релевантно визначати належність нових об'єктів до певних класів. Крім того, для наведених прикладів це дозволить виявляти відхилення від нормальної поведінки, що є значущим, наприклад, для задач безпеки мережі, а також оптимізувати мережу, покращивши процеси маршрутизації та використання ресурсів. Припустимо, що кожному об'єкту присвоюється клас на основі апріорних ймовірностей $P_{cl} = P(Y = cl), cl = 1,...,CL.$ Тут $\sum_{cl=1}^{CL} P_{cl} = 1.$ Це означає, що перед аналізом даних у нас вже є припущення про ймовірності належності об'єктів до різних класів. Ці апріорні ймовірності засновані на попередніх знаннях або даних про те, як розподілені об'єкти між класами. Отже, для нашого завдання це може означати, що вже є початкові припущення щодо того, до яких типів мережевих активностей (або класів) може належати кожен об'єкт. (o) (наприклад, IP-адрес, розмір пакета, дані від мережевих датчиків). Ці ймовірності можуть бути засновані на ретроспективних (історичних) даних про банківський мережевий трафік, статистиці виникнення певних подій або іншій релевантній інформації. Наприклад, ми можемо припускати, що 70% трафіку пов'язано з нормальною діяльністю співробітників, 20% — з автоматичними системами та серверами, і 10% — з підозрілими або аномальними активностями.

Ці припущення допомагають при аналізі та кластеризації даних, оскільки вони задають початкові ймовірності належності об'єктів до різних класів, що, у свою чергу, дозволяє використовувати ймовірнісні методи для більш точного аналізу та обробки даних. Таким чином, апріорні ймовірності служать вихідною точкою для визначення, до якого класу може належати кожен об'єкт, і використовуються в ймовірнісній моделі генерації даних для подальшого аналізу. Таким чином, відповідно до $p_{cl}(x)$ будемо визначати значення (x) незалежно для кожного (o). Далі для пари об'єктів, які вибрані довільно, наприклад, об'єкти $a,b \in s$, визначимо їх відповідність індикаторній функції $I(\cdot)$ [11], [12]. Тобто величина $H = I(Y(a) \neq Y(b))$. Тут $I(true)=1,\ I(false)=0.$

Введемо позначення $P_H = P\left[H=1 \middle| x(a), x(b)\right]$, яке описує ймовірність події a,b. Причому ці події належать до різних класів, за відомих $x(a)$ і $x(b)$. Тоді на підставі викладеного можна записати таке вираз для обчислення P_H:

$$P_H = 1 - \sum_{cl=1}^{CL} W, \tag{1}$$

де $W = \dfrac{p_{cl}\left(x(a)\right) p_{cl}\left(x(b)\right) P_{cl}^2}{p\left(x(a)\right) p\left(x(b)\right)};\ \ p(x(o)) = \sum_{cl=1}^{CL} p_{cl}(x(o)) P_{cl},\quad o = a,b.$

Таким чином, індикаторна функція $I(\cdot)$ (індикатор) — це функція, яка використовується для визначення відповідності певній умові [12]. Наприклад, якщо два пакети даних мають одну і ту ж IP-адресу джерела, індикаторна функція може прийняти значення 1. Якщо два мережевих з'єднання використовують один і той же порт, індикаторна функція може прийняти значення 1. Якщо два пакети даних використовують один і той же мережевий протокол (наприклад, HTTP або FTP), індикаторна функція може прийняти значення 1, див. табл. 2. Таким чином, індикаторна функція $I(\cdot)$ допомагає класифікувати об'єкти та аналізувати їхню належність до різних класів (CL). У випадку кластеризації це дозволить врахувати, наскільки об'єкти схожі один з одним за певними ознаками, що відіграє суттєву роль в оцінці якості кластеризації та побудові кластерних моделей для гетерогенних даних.

Таблиця 2

Приклад демонстрації, як можна використовувати індикаторні функції для виділення загальних характеристик пакетів даних

ID пакета	IP-адреса джерела	Порт джерела	Протокол	Розмір пакета	Індикатор IP = 192.168.1.100	Індикатор порт = 80	Індикатор протокол = TCP
1	192.168.1.100	80	TCP	1024	1	1	1
2	192.168.1.101	443	TCP	512	0	0	1
3	192.168.1.102	80	TCP	2048	1	1	1
4	192.168.1.150	22	SSH	50	0	0	0

Примітки. Індикатор IP = приймає значення 1, якщо IP-адрес джерела дорівнює, і 0 в протилежному випадку.

Індикатор порт = 80 приймає значення 1, якщо порт джерела дорівнює 80 (HTTP), і 0 в протилежному випадку.

Індикатор протокол = TCP приймає значення 1, якщо протокол є TCP, і 0 в протилежному випадку.

Коли ми використовуємо набір алгоритмів КА — $\mu_1, \mu_2, ..., \mu_M$, ми можемо отримати різні варіанти розбиття множини об'єктів S на кластери. Кількість кластерів для кожного варіанту може відрізнятися, оскільки різні алгоритми можуть по-різному групувати дані.

Наприклад, у нас є набір даних про банківський мережевий трафік, що включає IP-адреси, порти, протоколи та розмір пакетів. Для аналізу та оптимізації трафіку банку можемо використовувати різні алгоритми з набору для відповідних підзадач. Наприклад, K-means для групування трафіку за схожими параметрами, такими як порти та IP-адреси. DBSCAN для виявлення щільних регіонів трафіку та виявлення аномалій. Hierarchical Clustering для створення ієрархічної структури кластерів, що може допомогти у виявленні підкластерів. Використовуючи декілька алгоритмів і порівнюючи їх результати, можна отримати більш повне уявлення про структуру мережевого трафіку та виявити приховані закономірності, що в кінцевому підсумку допоможе оптимізувати мережу та підвищити її безпеку.

Оскільки нумерація класів не грає ролі, буде зручніше розглядати відношення еквівалентності. Відношення еквівалентності дозволить нам визначити, чи належать дві довільно вибрані пари об'єктів до одного і того ж класу або до різних класів. Визначаємо згідно з [6] для пари a,b величину $r_m = I[\mu_m(a) \neq \mu_m(b)]$. Як приклад, припустимо, що нам необхідно розглянути два об'єкти a і b, які представляють собою мережеві пакети з певними характеристиками, наприклад: об'єкт a: IP-адрес 192.168.1.1; порт 80; протокол HTTP; розмір пакета 1500 байт; об'єкт b: IP-адрес 192.168.1.2; порт 443; протокол: HTTPS; розмір пакета 1500 байт.

Припустимо, у нас є два алгоритми кластерного аналізу μ_1 і μ_2: алгоритм μ_1 розбиває дані за IP-адресою та портом; алгоритм μ_2 розбиває дані на кластери за протоколом та розміром пакета.

Для кожного алгоритму визначимо, чи потрапляють об'єкти a і b в один і той самий кластер. Для алгоритму μ_1 об'єкти a і b будуть у різних кластерах, оскільки їх IP-адреси та порти відрізняються. Таким чином, $r_1(a,b)=1$. Для алгоритму μ_2 об'єкти будуть у різних кластерах, так як їх протоколи відрізняються. Таким чином, $r_2(a,b)=1$. Таким чином, використовуючи множину алгоритмів КА та визначаючи індикаторні функції $q_m(a,b)=1$ для кожної пари об'єктів, ми можемо створити множину розбиттів об'єктів на кластери, що дозволить більш гнучко та точно аналізувати дані, враховуючи різні критерії та метрики.

Оскільки задача пошуку оптимального розбиття мережевого трафіку за заданим критерієм має експоненційну трудомісткість, на практиці також застосовуються наближені ітеративні алгоритми. Ці алгоритми на кожному кроці модифікують поточне розбиття, досягаючи локального покращення якості. Робота алгоритмів регулюється за допомогою параметрів, заданих користувачем.

Як було показано вище, алгоритми КА не є універсальними, оскільки кожен алгоритм має свою специфічну область застосування. Наприклад, одні алгоритми краще справляються із завданнями, де об'єкти кожного кластера описані «сферичними» областями багатовимірного простору, а інші — для пошуку «стрічкових» кластерів. Коли дані мають різнорідну природу, доцільно застосовувати не один алгоритм, а набір різних

алгоритмів. Колективний (ансамблевий) підхід, на нашу думку, дозволить підвищити стійкість результатів кластеризації, особливо коли невідомо, які параметри алгоритму є найкращими. У цьому випадку будується кілька варіантів результатів (при різних значеннях параметрів), а потім на їх основі формується остаточне рішення.

У роботах [5], [6], [8] розглядається поняття «постійної умовної ймовірності правильного рішення» — q_m. У нашому випадку це означає, що для кожного алгоритму μ_m, використовуваного в ансамблі (або колективному рішенні), ймовірність правильного об'єднання або розділення пари об'єктів (наприклад, IP-адрес або логів) завжди однакова і не залежить від конкретної пари об'єктів. Наприклад, якщо алгоритм використовує евклідову метрику для вимірювання відстані між об'єктами, його точність буде однаковою для всіх пар об'єктів. Іншими словами, параметр q_m — це ймовірність того, що алгоритм правильно об'єднує два об'єкти в один кластер, якщо вони дійсно належать до одного кластера, або правильно розділяє їх на різні кластери, якщо вони належать до різних кластерів. Значення q_m допомагає оцінити, наскільки добре алгоритм справляється із завданням кластеризації, тобто, чим вище значення q_m, тим більш надійним буде алгоритм. Відповідно, умова $q_m > 0{,}5$ називається умовою «слабкої навченості», що означає, що алгоритм приймає рішення краще, ніж випадковий вибір, що значимо для створення ансамблю (колективу) алгоритмів, оскільки гарантує, що кожен алгоритм в ансамблі робить позитивний внесок у загальне рішення. Проілюструємо це прикладом. Нехай є серверні логи і дані від мережевих датчиків. Використовуємо алгоритм — DBSCAN. DBSCAN вирішує, що певний лог і дані від датчика належать до різних кластерів. Якщо $q_m = 0{,}7$, то це означає, що ймовірність того, що DBSCAN правильно визначає, що ці дані належать до різних кластерів, становить 70%.

Таким чином, параметр q_m є ключовим для розуміння та оцінки ефективності кожного алгоритму в ансамблі, показуючи, наскільки кожен алгоритм точний у своїх рішеннях і допомагає визначити, як кожен з них буде впливати на загальний результат кластеризації. Підсумкове рішення по кластеризації отримується на основі зваженого голосування всіх алгоритмів, що дозволяє враховувати внесок кожного з них і покращувати точність та надійність результату.

Одна з основних труднощів КА великих даних для мережевого трафіку великих українських банків — неоднозначна інтерпретація результатів. Алгоритми кластеризації, засновані на різних підходах, можуть давати різні результати, що ускладнює прийняття рішень, наприклад, при оптимізації мережі. Для покращення методів, запропонованих у роботах [6], [7], [9], які враховують поведінку кожного алгоритму в різних умовах, можна використовувати ймовірнісну модель ансамблевої попарної класифікації з латентними класами, що дозволить врахувати вагу кожного алгоритму на основі його продуктивності в різних умовах. Під латентними класами розуміємо приховані або неявні категорії, не спостережувані безпосередньо, але такі, що впливають на поведінку та характеристики даних. Тобто латентні класи [11], [12] — це гіпотетичні категорії, що

допомагають пояснити структуру даних. Наприклад, при аналізі мережевого трафіку це можуть бути групи користувачів з певними шаблонами поведінки, типи пристроїв, типи мережевих атак та інші приховані фактори, що впливають на мережеву активність.

Скажімо, якщо ми аналізуємо дані про мережевий трафік, включаючи IP-адреси, порти, протоколи, розміри пакетів, серверні логи і дані від мережевих датчиків, то латентні класи можуть представляти собою різні типи мережевих пристроїв (наприклад, сервери, робочі станції, мобільні пристрої) або типи мережевих атак (наприклад, DDoS-атаки, фішинг, проникнення).

У ймовірнісній моделі ансамблевої попарної класифікації з латентними класами кожен об'єкт (наприклад, запис мережевого трафіку) передбачається, що належить до одного з прихованих класів. Ця приналежність впливає на його ймовірнісний розподіл. Наприклад, запис мережевого трафіку, що належить до класу «мобільні пристрої», може мати характерні особливості, які відрізняються від записів, що належать до класу «сервери».

Зазначимо, що визначення ваг алгоритмів кластеризації в ансамблі — важливий крок для покращення точності та надійності КА. Для цього можна використовувати, як було показано в [6], [7], [13] – [15], ймовірнісну модель ансамблевої попарної класифікації з латентними класами. Розглянемо математичний опис цієї моделі та визначення ваг.

Вага алгоритму (w) розрахована на основі метрик якості, а ваги нормалізовані, щоб їх сума дорівнювала 1. Ваги показують надійність алгоритму для конкретного набору даних. У результаті, вказується, чи вважає алгоритм пару об'єктів такими, що належать до одного кластера (1) або різних (0). Для визначення внеску в коасоціативну матрицю розраховується як добуток ваги алгоритму на результат кластеризації. Відповідно, підсумкова коасоціативна матриця формується шляхом усереднення внесків усіх алгоритмів ансамблю.

Таким чином, резюмуючи вищесказане, можна стверджувати, що запропонований підхід дає наочне уявлення про те, як різні алгоритми кластеризації, їх метрики та відповідні ваги впливають на процес формування коасоціативної матриці, відображаючи взаємозв'язок між надійністю алгоритмів, вираженою через їх ваги (w), і якістю колективного рішення, що дозволить спеціалісту, наприклад, аналітику банку, глибше зрозуміти їх внесок у загальний результат у процесі вирішення завдання, пов'язаного з аналізом та оптимізацією корпоративного трафіку.

Зауважимо, що відповідна варіація алгоритму K-means може використовувати різні характеристики пакетів, такі як адреса джерела, адреса призначення, тип протоколу, розмір пакета і т.д. У такому випадку, статистична залежність означатиме, що якщо два пакети мають схожі характеристики (наприклад, обидва є HTTP-пакетами з однієї IP-адреси на іншу), то алгоритм з великою ймовірністю віднесе їх до одного кластеру, навіть якщо у них різні випадкові вектори. Ω. І, навпаки, якщо два пакети мають різні характеристики (наприклад, один HTTP-пакет з одного IP-адреса, а інший SSH-пакет з іншого IP-адреса), то алгоритм з великою ймовірністю віднесе їх до різних кластерів, навіть якщо у них однакові випадкові вектори. Ω. Як було показано в роботах [1], [2], [6], [7], [11], [16], [17], досягнення статистичної залежності — важливе завдання при розробці ансамблю алгоритмів кластеризації, оскільки дозволяє алгоритму більш точно відображати реальні класові структури в даних.

Однак, статистична залежність не гарантує ідеальної кластеризації. Адже навіть у випадку, якщо алгоритм μ_m завжди правильно класифікує пари об'єктів з «однаковими» характеристиками, він може помилитися, коли характеристики об'єктів будуть «перекриватися» між класами.

Вважаємо, що можна говорити про те, що для кожного алгоритму μ_m, що входить до ансамблю (колективне рішення), ми можемо використовувати таку формулу для оновлення апостеріорної ймовірності $P(A_i|D)$:

$$P(A_i|D) = \frac{P(D|A_i)P(A_i)}{\sum_{j=1}^{n} P(D|A_j)P(A_j)} \quad (2)$$

де $P(A_i)$ — апріорна ймовірність алгоритму A_i; $P(D|A_i)$ — ймовірність спостереження даних D при дотриманні умови, що використовується алгоритм A_i. $P(D|A_i)$ розраховується на основі метрик якості [6], [7], [11], [12]; $\sum_{j=1}^{n} P(D|A_j)P(A_j)$ — нормалізуючий фактор, що гарантує рівність апостеріорних ймовірностей одиниці 1.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Проілюструємо це прикладом. Припустимо, у нас є три варіанти варіацій алгоритму K-means:

$\mu_1 = A_1$ — варіація K-means з евклідовою відстанню. Підійде, якщо в рамках загального завдання досліджень необхідно виявити аномальний трафік. Зазначимо, що аномальний трафік у ряді випадків свідчить про можливі атаки на мережу або неполадки.

$\mu_2 = A_2$ — варіація K-means з косинусною відстанню. Наприклад, цей варіант підійде для кластеризації співробітників на основі типів їхньої мережевої активності. У подібній підзадачі мета — згрупувати співробітників, чиї профілі використання мережі (наприклад, відвідувані веб-сайти, використовувані додатки) мають схожі патерни. Отже, косинусна відстань буде корисною, оскільки вона дозволить врахувати кутову схожість між векторами активності, незалежно від їх абсолютної величини, що принципово, оскільки співробітники можуть мати різний обсяг мережевої активності, але схожі типи використання.

$\mu_3 = A_3$ — варіація K-means з відстанню Мінковського. Наприклад, K-means з відстанню Мінковського підійде для кластеризації сесій мережевого трафіку для того, щоб виявити аномальні патерни поведінки. Мета — групувати мережеві сесії за характеристиками, такими як тривалість сесії, обсяг переданих даних, кількість звернень до різних вузлів. Відстань Мінковського дозволить нам гнучко налаштувати ступінь впливу різних вимірів, що буде корисно для того, щоб врахувати різні типи аномалій. Скажімо, сюди відносяться короткі, але інтенсивні сплески трафіку або тривалі, але малопотужні з'єднання. Для ілюстрації для даної варіації K-means на рис. 1 показано результат кластеризації.

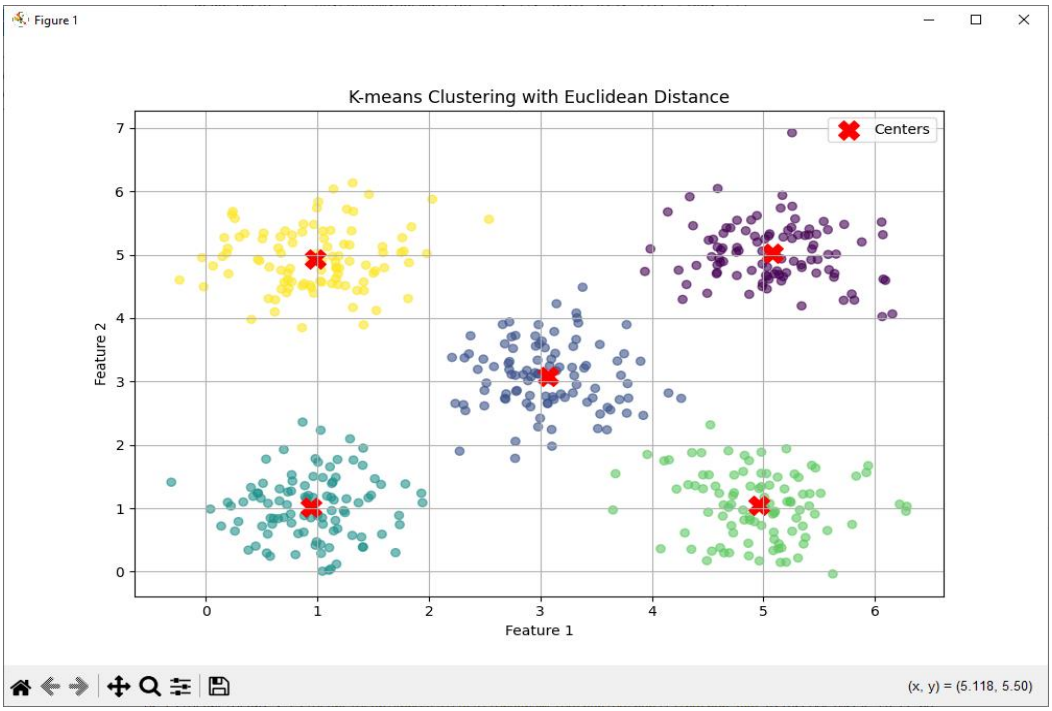

Рис. 1. Приклад візуалізації результатів кластеризації для варіації K-means з відстанню Мінковського для трафіку компанії (банку)

Припускаємо, що спочатку апріорні ймовірності для всіх алгоритмів рівні, тобто можемо записати: $P(A_1, A_2, A_3) = \frac{1}{3}$. Відповідно для більшої кількості варіаційних алгоритмів пропорція буде іншою.

Використовуючи положення робіт [11], [15], [16], [17] і виконавши кластеризацію, можемо далі розрахувати метрики ARI і NMI для кожного варіанта. Припустимо, що отримані такі значення:

$ARI_{A_1} = 0,8, \quad NMI_{A_1} = 0,75, \quad ARI_{A_2} = 0,6, \quad NMI_{A_2} = 0,65, \quad ARI_{A_3} = 0,7, \quad NMI_{A_3} = 0,7.$

Для оцінки $P(D|A_i)$ будемо використовувати нормалізовану суму метрик

$$P(D|A_i) = \frac{ARI_{A_i} + NMI_{A_i}}{\sum_{k=1}^{4}\left(ARI_{A_i} + NMI_{A_i}\right)}. \quad (3)$$

У цьому прикладі була реалізована кластеризація трафіку з використанням алгоритму K-means, що дозволило виявити структурні особливості та закономірності в даних, що представляють різні відділи. Для аналізу були використані дані компанії ITBIZ, для п’яти відділів, кожен з яких представлений нормально розподіленими точками в двовимірному просторі. При застосуванні K-means до даних була встановлена мета розділити їх на п’ять кластерів, що відповідає кількості відділів.

Концептуальна блок-схема реалізації такого алгоритму показана на рис. 2.

ISSN 2663 - 4023

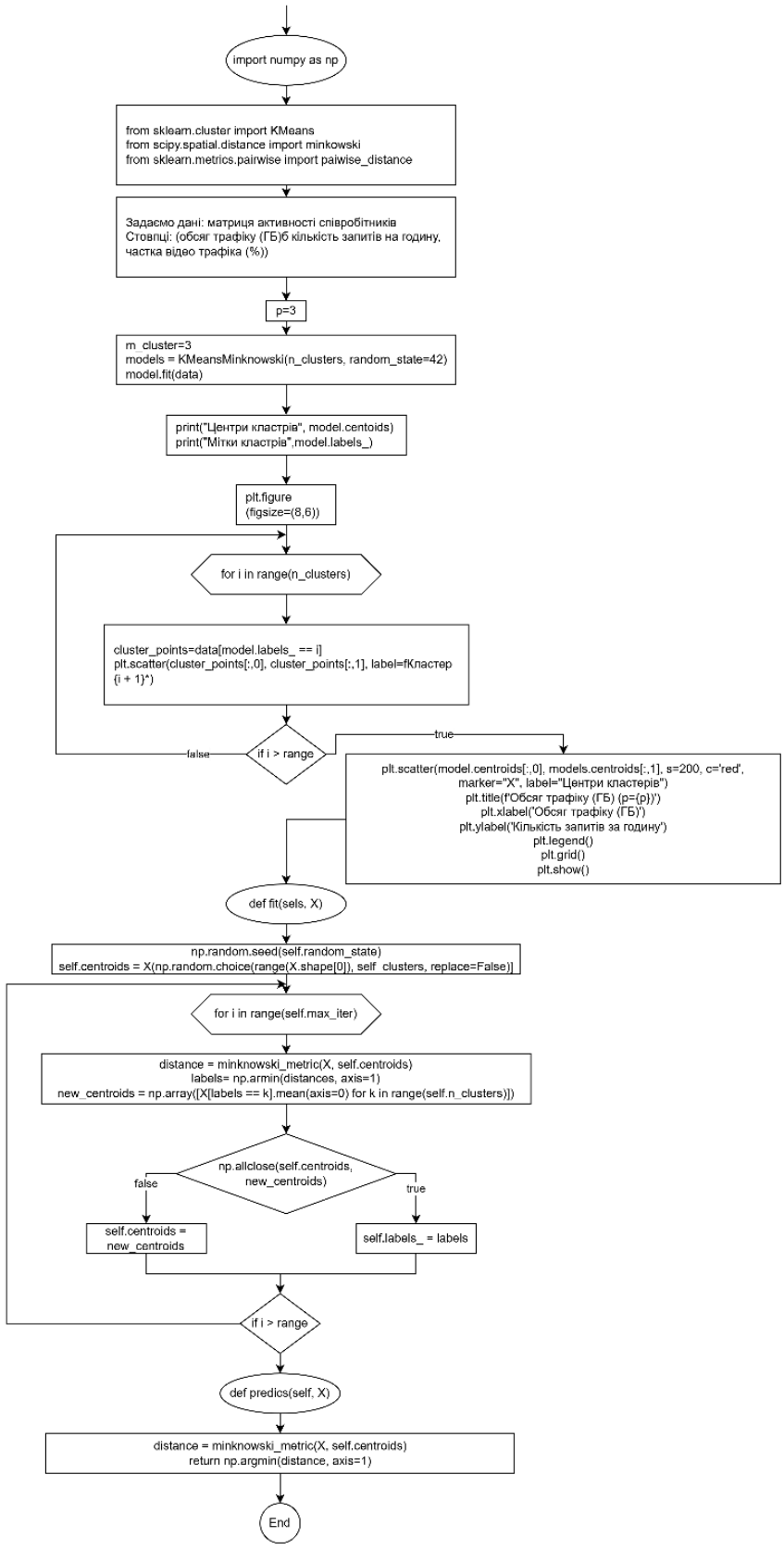

Рис. 2. Концептуальна блок-схема реалізації K-means з відстанню Мінковського для задачі кластеризації співробітників на основі типів їхньої мережевої активності

Алгоритм прагне мінімізувати внутрішньокласову дисперсію, автоматично визначаючи центри кластерів. Результати кластеризації візуалізовані на двовимірному графіку, де кожна точка позначає окремий елемент трафіку і забарвлена залежно від належності до певного кластера, тоді як центри кластерів виділені червоними хрестиками.

Обговорення отриманого результату

На графіку, див. рис. 1, спостерігається чітке розділення даних на п’ять кластерів, що вказує на наявність власних характеристик трафіку кожного відділу, наприклад, банку. Ця різниця в розташуванні та щільності кластерів свідчить про специфічні патерни, властиві різним підрозділам банку. Центри кластерів, позначені червоними хрестиками, і служать для швидкої ідентифікації основних характеристик кожного відділу. Результати цього аналізу демонструють, що метод K-means ефективно виділяє структури в даних про трафік, надаючи цінну інформацію для подальшого аналізу та прийняття управлінських рішень. Виявлені патерни можуть бути використані для оптимізації процесів і більш ефективного розподілу ресурсів у банку, що в кінцевому підсумку сприяє підвищенню операційної ефективності.

Вважаємо, що такий підхід у рамках розвитку колективних кластерних рішень для аналізу BD, що стосуються проблематики мережевого трафіку великих банків, може мати певні переваги. Це пов’язано з тим, що баєсівський підхід дозволить нам, наприклад, використовуючи інтелектуальну інформаційну систему, адаптивно оновлювати ваги алгоритмів на основі нових даних, що релевантно для задач, які вирішуються в умовах динамічно змінюваного трафіку банків. Також використання метрик якості для оновлення ймовірностей покращить точність кластеризації, хоча цей тезис і потребує експериментальної перевірки.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Розглянуто ансамблевий підхід до кластерного аналізу (КА) мережевого трафіку банку, заснований на застосуванні різних метрик відстані. Розроблений метод дозволяє враховувати гетерогенність даних, представлену різними типами мережевої інформації (пакети, з’єднання, пристрої, користувачі), і підвищує точність аналізу за рахунок інтеграції результатів декількох алгоритмів. Запропоновано концептуально ймовірнісну модель ансамблевого КА, що враховує латентні класи та динамічно налаштовувані ваги алгоритмів. Продемонстровано перевагу використання коасоціативних матриць, що дозволяють підвищити узгодженість підсумкового розбиття об’єктів. Виявлено залежність якості кластеризації від вибору метрик відстані, що підтверджує доцільність використання ансамблевого методу. Запропоновано адаптивний механізм зважування алгоритмів на основі метрик якості (ARI, NMI), що підвищує надійність сегментації мережевого трафіку. Показана практична застосовність методу для аналізу корпоративних банківських мереж, включаючи виявлення аномалій та оптимізацію маршрутизації трафіку. Таким чином, ансамблевий метод КА з різними метриками відстані дозволяє компенсувати недоліки окремих алгоритмів і забезпечує більш точну та надійну обробку гетерогенних даних мережевого трафіку банку.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Кучук, Г. А. (2005). Метод дослідження фрактального мережевого трафіку. *Системи обробки інформації,(5)*, 74–84.
2. Бабенко, Т. В. (2013). Дослідження ентропії мережевого трафіку як індикатора DDOS-атак. *Науковий вісник Національного гірничого університету, (2)*, 86–89.
3. Кучук, Г. А., Можаєв, О. О., & Воробйов, О. В. (2006). Метод прогнозування фрактального трафіку. *Радіоелектронні і комп'ютерні системи, (6),* 181–188.
4. Ren, Y., Domeniconi, C., Zhang, G., & Yu, G. (2017). Weighted-object ensemble clustering: methods and analysis. *Knowledge and Information Systems, 51(2),* 661–689. https://doi.org/10.1007/s10115-016-0988-y
5. Zheng, L., Li, T., & Ding, C. (2010). Hierarchical ensemble clustering. In*: 2010 IEEE International Conference on Data Mining*, 1199–1204.
6. Катрич, Д. С. (2021). Задача виявлення аномальних потоків мережевого трафіку на основі кластеризації мережевих з'єднань. *Всеукраїнська науково–практична конференція студентів, аспірантів та молодих вчених. Математичні методи комп'ютерного моделювання та кібернетичної безпеки*, 222–224.
7. Ghosh, J., & Acharya, A. (2011). Cluster ensembles. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 1(4), 305–315.
8. Berikov, V. (2011). A latent variable pairwise classification model of a clustering ensemble. *In Multiple Classifier Systems: 10th International Workshop, MCS 2011, Naples, Italy*, 279–288.
9. Figuera, P., Cuzzocrea, A., García Bringas, P. (2023). Probability Density Function for Clustering Validation. *In International Conference on Hybrid Artificial Intelligence Systems*, 133–144.
10. Arvapally, R. S., Liu, X. (2012). Analyzing credibility of arguments in a web–based intelligent argumentation system for collective decision support based on K–means clustering algorithm. *Knowledge Management Research & Practice*, *10(4)*, 326–341.
11. Tshimanga, R. M., Bola, G. B., Kabuya, P. M., Nkaba, L., Neal, J., Hawker, L., & Wagener, T. (2022). Towards a framework of catchment classification for hydrologic predictions and water resources management in the ungauged basin of the Congo River. *Congo Basin hydrology, climate, and biogeochemistry: a foundation for the future*, 469–498. https://doi.org/10.1002/9781119657002.ch24
12. Shuch, H. P., & Bowyer, S. (2011). SERENDIP: The Berkeley SETI Program. Searching for Extraterrestrial Intelligence. *SETI Past, Present, and Future*, 99–105.
13. Topchy, A. P., Law, M. H., Jain, A. K., & Fred, A. L. (2004). Analysis of consensus partition in cluster ensemble. *In Fourth IEEE International Conference on Data Mining (ICDM'04)*, 225–232.
14. Tsai, C. F. (2014). Combining cluster analysis with classifier ensembles to predict financial distress. *Information Fusion*, *16*, 46–58. https://doi.org/10.1016/j.inffus.2011.12.001
15. Vega–Pons, S., Ruiz–Shulcloper, J., (2009). Clustering Ensemble Method for Heterogeneous Partitions. *CIARP'09*, 481–488.
16. Bissarinov Baituma, & Begaidarova Aida. (2023). Optimizing Network Architecture for Big Data Transmission. *Research Reviews*, *(4).* https://ojs.scipub.de/index.php/RR/article/view/2508
17. Bisarinov, B. J. (2020). Special features of the research of the big data technology. *Scientific Journal-Bulletin of KazATC*, *114(3),* 253–258.

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 2 (26), 2024
ISSN 2663 - 4023

Denys Redko
Postgraduate Student of the Department of Software Engineering and Cybersecurity
State University of Trade and Economics, Kyiv, Ukraine
ORCID ID: 0009-0003-5827-264X
d.redko@knute.edu.ua

Alona Desiatko
Doctor of Philosophy in Computer Science, Associate Professor
Head of the Department of Software Engineering and Cybersecurity
State University of Trade and Economics, Kyiv, Ukraine
ORCID ID: 0000-0002-2284-3418
desyatko@gmail.com

AN ENSEMBLE OF CLUSTERING ALGORITHMS WITH DIFFERENT METRICS FOR ANALYZING THE BANK'S NETWORK TRAFFIC

Abstract. The relevance of the study is due to the need to increase the accuracy of the analysis of bank network traffic, represented by heterogeneous data, including server logs, network connections, user behavioral data and traffic telemetry. In the context of increasing data volumes and the complexity of the structure of corporate networks of Ukrainian banks, traditional analysis methods lose their effectiveness, making the application of machine learning methods, including ensemble clustering algorithms, relevant. Modern research in this area focuses on the development of adaptive segmentation methods that use probabilistic models and dynamic weighting of algorithms to increase the accuracy of data partitioning. The paper proposes an ensemble KA method that uses variations of the K-means algorithm with different distance metrics. The method is based on a probabilistic model that takes into account latent classes and dynamically adjustable algorithm weights, and for the consistency of the partitioning, a coassociative matrix is used that reflects the frequency of pairs of objects falling into one cluster. The novelty of the research lies in the development of a mechanism for adaptive weighting of ensemble algorithms based on quality metrics (ARI, NMI), which allows to increase the accuracy of clustering of bank network traffic. The proposed method compensates for the shortcomings of individual algorithms and takes into account the features of different types of data (packets, connections, users, devices). The practical significance of the work is confirmed by the possibility of applying the method to analyze banking network traffic, detect anomalies, optimize routing and increase cybersecurity of the banking network, and the developed approach will allow to take into account the complex data structure and adapt to dynamic changes in the bank's network environment.

Keywords: cybersecurity; cluster analysis; ensemble algorithm; distance metrics; coassociative matrix; banking networks; network traffic analysis; probabilistic models.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Kuchuk, G. A. (2005). Method for studying fractal network traffic. *Information Processing Systems, (5),* 74–84.
2. Babenko, T. V. (2013). Study of network traffic entropy as an indicator of DDOS attacks. *Scientific Bulletin of the National Mining University, (2),* 86–89.
3. Kuchuk, G. A., Mozhayev, O. O., & Vorobyev, O. V. (2006). Method for predicting fractal traffic. *Radioelectronic and computer systems, (6),* 181–188.
4. Ren, Y., Domeniconi, C., Zhang, G., & Yu, G. (2017). Weighted-object ensemble clustering: methods and analysis. *Knowledge and Information Systems, 51(2),* 661–689. https://doi.org/10.1007/s10115-016-0988-y
5. Zheng, L., Li, T., & Ding, C. (2010). Hierarchical ensemble clustering. *In: 2010 IEEE International Conference on Data Mining,* 1199–1204.
6. Katrych, D. S. (2021). The problem of detecting anomalous network traffic flows based on clustering of network connections. *All-Ukrainian scientific and practical conference of students, postgraduates and young scientists. Mathematical methods of computer modeling and cyber security*, 222–224.

7. Ghosh, J., & Acharya, A. (2011). Cluster ensembles. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, *1(4)*, 305–315.
8. Berikov, V. (2011). A latent variable pairwise classification model of a clustering ensemble. *In Multiple Classifier Systems: 10th International Workshop, MCS 2011, Naples, Italy*, 279–288.
9. Figuera, P., Cuzzocrea, A., & García Bringas, P. (2023). Probability Density Function for Clustering Validation. *In International Conference on Hybrid Artificial Intelligence Systems*, 133–144.
10. Arvapally, R.S., & Liu, X. (2012). Analyzing credibility of arguments in a web-based intelligent argumentation system for collective decision support based on K-means clustering algorithm. *Knowledge Management Research & Practice, 10(4)*, 326–341.
11. Tshimanga, R.M., Bola, G.B., Kabuya, P.M., Nkaba, L., Neal, J., Hawker, L., & Wagener, T. (2022). Towards a framework of catchment classification for hydrologic predictions and water resources management in the ungauged basin of the Congo River. *Congo Basin hydrology, climate, and biogeochemistry: a foundation for the future*, 469–498. https://doi.org/10.1002/9781119657002.ch24
12. Shuch, H. P., & Bowyer, S. (2011). SERENDIP: The Berkeley SETI Program. Searching for Extraterrestrial Intelligence. *SETI Past, Present, and Future*, 99–105.
13. Topchy, A.P., Law, M.H., Jain, A.K., & Fred, A.L. (2004). Analysis of consensus partition in cluster ensemble. *In Fourth IEEE International Conference on Data Mining (ICDM'04)*, 225–232.
14. Tsai, C.F. (2014). Combining cluster analysis with classifier ensembles to predict financial distress. *Information Fusion*, *16*, 46–58. https://doi.org/10.1016/j.inffus.2011.12.001
15. Vega–Pons, S., & Ruiz–Shulcloper, J., (2009). Clustering Ensemble Method for Heterogeneous Partitions. *CIARP'09*, 481–488.
16. Bissarinov Baituma, & Begaidarova Aida. (2023). Optimizing Network Architecture for Big Data Transmission. *Research Reviews, (4).* https://ojs.scipub.de/index.php/RR/article/view/2508.
17. Bisarinov, B. J. (2020). Special features of the research of the big data technology. *Scientific Journal-Bulletin of KazATC*, *114(3)*, 253–258.