УДК 004.056.5

ЗАДАЧА ВИЯВЛЕННЯ АНОМАЛЬНИХ ПОТОКІВ МЕРЕЖЕВОГО ТРАФІКУ НА ОСНОВІ КЛАСТЕРИЗАЦІЇ МЕРЕЖЕВИХ З'ЄДНАНЬ

Д. С. Катрич[1, a]

[1] *Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Фізико-технічний інститут*

Анотація

Протягом останнього десятиліття до систем виявлення вторгнень (IDS) застосовувались різні технології машинного навчання та інтелектуального аналізу даних, які відігравали важливу роль у захисті критичних комп'ютерних систем та мереж від кібератак. Виявленню аномалій на нерозмічених даних приділяють особливу увагу, оскільки вони дозволяють будувати моделі виявлення вторгнень без використання маркованих навчальних даних. У даній статті розглядається нерозмічений набір даних для виявлення аномалій з використанням кластеризації методом k-means та визначення оптимальної кількості кластерів методом «ліктів».

Ключові слова: кластерний аналіз, виявлення аномалій, мережевий трафік, K-середніх, метод «ліктів»

Вступ

Однією з головних проблем у кібербезпеці є забезпечення автоматизованої та ефективної техніки виявлення кіберзагроз. Актуальність задач пошуку та впровадження нових методів забезпечення захисту інформації обумовлюють динамічний розвиток інформаційних технологій в сукупності із збільшенням кількості кібератак. Сьогоденні кіберзагрози спричинені спробами несанкціонованого доступу (НСД), махінаціями, модифікаціями над інформацією та іншими діями які порушують критерії безпеки. Найважливішим кроком для забезпечення захисту від кіберзагроз є ефективність їх виявлення. Саме тому більшість засобів захисту мають підсистеми виявлення задля забезпечення знаходження та ідентифікації кіберзагроз. Модель виявлення аномалій ідентифікує атаку шляхом пошуку поведінки, яка виходить за рамки норми. Основна перевага виявлення аномалій полягає в відсутності попереднього дослідження про вторгнення.

Кластерний аналіз використовується для виявлення і класифікації кібератак. Його суть полягає у визначенні таких характеристик з мережевого трафіку, що дозволяють розбити об'єкти (пакети, з'єднання), які класифікуються на групи, відповідні нормальному функціонуванню мережевої взаємодії. Всі інші екземпляри, які не потрапляють в побудовані області, класифікуються як аномальні.

Мета роботи

Використання кластерного аналізу методом k-середніх (k-means clustering) та перевірка кількості кластерів методом «ліктів"(Elbow method) для подальшого виявлення аномальних потоків мережевого трафіку.

1. Кластерний аналіз

Кластеризація — метод пошуку закономірностей в нерозмічених даних. Кластерний аналіз розподіляє набір точок даних в набір кластерів або груп. Ці точки даних якомога більше схожі в межах однієї групи та віддалені наскільки це можливо від інших груп. Кластерний аналіз відноситься до навчання без вчителя (unsupervised learning) з огляду на те, що на початку немає визначених класів. Це суттєво відрізняє його від класифікації, де потребується навчання з учителем (supervised learning) або завдання міток класу для побудови моделі класифікації [2].

1.1. Кластеризація методом к-середніх (k-means clustering)

Алгоритм k-means — метод кластерного аналізу, метою якого є розподіл m спостережень на k кластерів, при цьому кожне спостереження відноситься до того кластеру, до центру (центроїду) якого воно найближче. Цей алгоритм потребує заздалегідь визначеної кількості кластерів. Етапи процесу кластеризації методом k-means складаються з таких кроків:

1) Визначення кількості кластерів.
2) Ініціалізація центроїду. В даній статті пропонується використати формулу середнього як опис середнього значення кластерів, які показані в (1):

$$\mu_i = \frac{x_1 + x_2 + x_3 + \ldots + x_j}{j} \qquad (1)$$

де x_j — значення об'єкта в кластері та j — кількість об'єктів в кластері, а формула для визна-

[a] katrych.daria@gmail.com

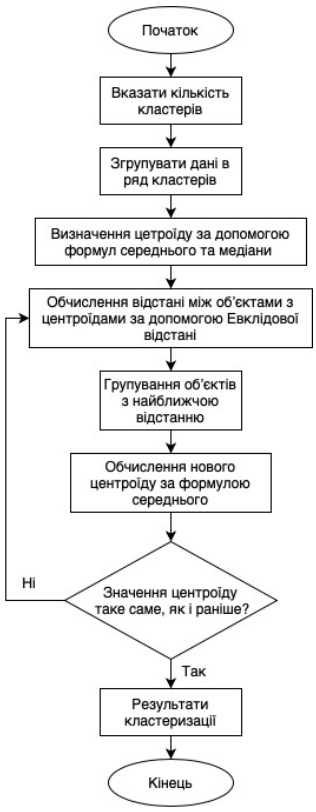

Рис. 1. Етапи алгоритму кластеризації методом k-means [1]

чення медіани є наступною:

$$ME = x_{\frac{n+1}{2}} \tag{2}$$

3) Розподіл даних. Всі дані або об'єкти будуть розташовані у найближчому кластері. Відстань між двума об'єктами визначає близкість об'єкта. Формула Евклідової відстані:

$$d = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2 + \ldots + (x_n - c_n)^2} \tag{3}$$

де d — відстань точок x і c, x_n — дані критерію, c_n — центроїд n-го кластера.

4) Обчислення нового центроїду. Для кожного кластера, після завершення першого розподілу даних ітерацій, розраховується новий центроїд для наступного розподілу даних ітерацій. Математично цей крок можна виразити як мінімізацію суми квадратних помилок від усіх точок даних кластера до цетроїдного кластера. Загальною метою цього кроку є мінімізація суми квадратних помилок кожної групи. Нова формула визначення цетроїду:

$$\mu_i = \frac{1}{ji} \sum_{x \in C_i} X \tag{4}$$

де μ_i — центроїд, X — об'єкт у кластері, ji — кількість об'єктів у клатері.

Етап обчислення нового центроїда та крок присвоєння точок даних новому центроїду повторюється до тих пір, поки не перестануть відбуватися зміни в призначенні точок даних. Останній центроїд використовується як прототип кластера і використовується для опису всієї моделі групування. Етапи алгоритму k-means проілюстровані на Рис. 1.

1.2. Перевірка кількості кластерів «методом ліктя» (Elbow method)

Кластеризація методом k-means з використанням «методу ліктя» застосовується для визначення оптимального числа кластерів у наборі даних.

Цей метод є візуальним методом для перевірки кількості кластерів. Ідея полягає в тому, щоб визначити початкову кількість кластерів, збільшувати її і розрахувати сумарну квадратичну помилку (SSE) на кластер до максимальної кількості кластерів, і потім, порівнявши різницю сумарної квадратичної помилки кожного кластера, побачити найбільш екстремальну різницю формування кута нахилу, яка показує найкраще число кластерів. Формула сумарної квадратичної помилки:

$$SSE = \sum_{i=1}^{k} \sum_{x_j \in C_i} \| x_j - \mu_i \|^2 \tag{5}$$

де x_j — об'єкт в кластері c_i і центроїд кластера.

Алгоритм «методу ліктів» для визначення значення кластерів (k):

1) Ініціалізація k = 1.
2) Запуск.
3) Збільшення значення k.
4) Обчислення сумарної квадратичної помилки (SSE).
5) Значення k при SSE різкого зменшення є оптимальним значенням k.
6) Кінець.

Результати оцінки визначення найкращого числа кластерів представлені на Рис. 2. Оптимальна кількість кластерів на графіку — 2.

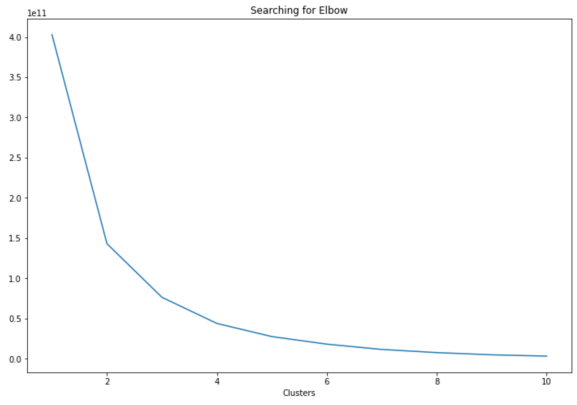

Рис. 2. Elbow method для визначення кількості кластерів.

2. Кластерний аналіз методом к-середніх (k-means) для виявлення аномалій

Кластеризація для виявлення кібератак — метод пошуку закономірностей в нерозмічених даних.

Табл. 1. Кластеризація методом k-means з двома кластерами.

cluster	1	2
length	1184.6	8759.6
protocol_DNS	0.0	0.0
protocol_ESP	0.0	0.0
protocol_HTTP	0.0	0.0
protocol_ISAKAMP	0.0	0.0
protocol_MySQL	0.0	0.0
protocol_NBSS	0.4	0.8
protocol_SMB	0.0	0.0
protocol_SSH	0.0	0.0
protocol_SSHv2	0.0	0.0
protocol_TCP	0.6	0.2
protocol_TLSv1.2	0.0	0.0
protocol_UDP	0.0	0.0
dscp_2	0.0	0.0
dscp_6	0.0	0.0
dscp_Default	1.0	1.0
PSH_Notset	0.9	0.9
PSH_Set	0.1	0.1
hdlen	34.4	32.0

Табл. 2. Розподіл пакетів за протоколами в кластері №1.

protocol_DNS	16
protocol_ESP	196
protocol_HTTP	24
protocol_ISAKAMP	13
protocol_MySQL	48
protocol_NBSS	13063
protocol_SMB	1461
protocol_SSH	435
protocol_SSHv2	272
protocol_TCP	19991
protocol_TLSv1.2	17
protocol_UDP	8

Основною перевагою алгоритму кластеризації є здатність виявляти вторгнення без явних описів вторгнень, які зазвичай надаються експертами по кібербезпеці. Метод кластерного аналізу заключається в визначенні характеризуючого вектора кожного кластера для виявлення відхилення від середньостатистичного значення і на підставі відстані середньостатистичної характеристики певний пакет відноситься до того чи іншого кластеру.

Кластерний аналіз проводиться на основі спостережень мережевого трафіку, зафіксованого в ході змагань між фахівцями з кіберзахисту комп'ютерних систем, що були опубліковані організаторами для подальшого дослідження [3].

На Табл. 1 зображені результати кластеризації мережевого трафіку з двома кластерами. В рядку «cluster» пронумеровано кластери відповідно. В рядку «length» показано ядро кластера, тобто його характеристика — середнє значення для всіх пакетів першого і другого кластера. На Табл. 2 та Табл. 3 зображені розподіли пакетів за протоколами у кластерах №1 та №2. На основі даних отриманих після кластеризації, можна побачити за рахунок чого були відібрані пакети. Перший кластер складає 35544 пакети, другий кластер — 5180 пакетів. Середньостатистичний пакет в першому кластері має довжину 1184.6 байт, у другому кластері довжина 8759.6 байт. Також є значна різниця в розподілі пакетів за протоколами NBSS та TCP в кластері №1 та кластері №2.

Табл. 3. Розподіл пакетів за протоколами в кластері №2.

protocol_DNS	0
protocol_ESP	0
protocol_HTTP	2
protocol_ISAKAMP	0
protocol_MySQL	0
protocol_NBSS	4268
protocol_SMB	65
protocol_SSH	0
protocol_SSHv2	0
protocol_TCP	855
protocol_TLSv1.2	0
protocol_UDP	0

Висновки

Стрімкий розвиток комп'ютерних мереж та інформаційних технологій викликає ряд проблем, що пов'язані з безпекою мережевих ресурсів, які потребують нових підходів. Сучасні методи розпізнання сигнатур, асоціативних правил не дозволяють розпізнавати нові загрози даних які не містяться в базі даних загроз. Задачею виявлення аномалій на нерозмічених даних є спроможність виявляти загрози і атаки як відхилення від нормального використання.

У даній статті розглянуто нерозмічений набір даних і в ньому були виявлені аномалії з використанням кластеризації методом k-means та визначено оптимальну кількість кластерів методом «ліктів».

Перелік використаних джерел

1. Umargono E., Suseno J.E. and Gunawan S. K. K-Means Clustering Optimization using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula. — 01, 2020. — Access mode: `https://www.researchgate.net/publication/346349075_K-Means_Clustering_Optimization_Using_the_Elbow_Method_and_Early_Centroid_Determination_Based_on_Mean_and_Median_Formula`.
2. Chunhui Yuan and Haitao Yang. Research on K-Value Selection Method of K-Means Clustering Algorithm. — 2019. — Access mode: `https://www.mdpi.com/2571-8800/2/2/16/htm`.
3. WRCCDC Archive, Directory: /pcaps/2019/ — Access mode: `https://archive.wrccdc.org/pcaps/2019/regionals/`.