

ПРОБЛЕМА ВИБОРУ СИСТЕМИ ПОКАЗНИКІВ ДЛЯ АНАЛІЗУ РСАР-ФАЙЛІВ В УМОВАХ ПЕРЕХРЕСНОЇ ПЕРЕВІРКИ

Б. С. Білоконь^{1, а}, М. С. Стремецька¹

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

Анотація

В умовах популяризації методів розробки моделей на основі вибірових статистик виникає гостра необхідність реалізації системного підходу до перевірки якості моделювання на незалежних наборах даних. Побудовано модель даних мережевого трафіку на основі МГК та реалізовано алгоритм k-блочної перехресної перевірки ($k=10$) для оцінки якості моделювання розширеного набору лабораторних зразків пакетів мережевого трафіку, захоплених в ході атаки класу DDoS SYN Flood. Моделі з найкращими оцінками якості в середньому були виконані на навчальних наборах A_2, A_4, A_8 , проте суттєвих погіршень якості декомпозиції для решти моделей виявлено не було. Результати оцінювання якості моделювання підтверджують, що отримані впродовж перевірки оцінки дисперсії залишків в навчанні та в перевірці є передбачувано прийнятними і дозволяють якісним чином моделювати поведінку мережевого трафіку за допомогою МГК.

Ключові слова: метод головних компонент (МГК), DDoS SYN Flood, пояснена дисперсія залишків, повна дисперсія залишків, k-блочна перехресна перевірка.

Вступ

Важливим етапом на шляху впровадження простих та ефективних механізмів виявлення підозрілої активності є реалізація системного підходу до перевірки якості моделювання поведінки мережевого трафіку на незалежних наборах даних. Важливо, щоб запропонований підхід дозволяв отримувати оцінки якості моделювання в умовах обмеженого обсягу спостережень з урахуванням характерних особливостей поведінки відповідної інформаційної системи. Тому, спираючись на дослідження вибору системи показників для аналізу захоплених пакетів мережевого трафіку за допомогою МГК [1], розглянемо системні підходи до організації перевірки якості декомпозиції методом тест-валідації на основі алгоритму k-блочної перехресної перевірки.

Мета роботи

Сформувані систему показників для аналізу захоплених пакетів мережевого трафіку. Реалізувати системний підхід до перевірки якості декомпозиції методом тест-валідації шляхом порівняння оцінок якості моделювання в умовах k-блочної перехресної перевірки.

1. Підготовка набору вхідних даних

Дослідження проводиться на основі набору лабораторних зразків пакетів мережевого трафіку, захоплених в ході атаки класу DDoS SYN Flood, що був

опублікований Канадським інститутом кібербезпеки [2, 3] для подальшого вивчення. Спираючись на матеріали публікації Костянтиноса Демерціса, Нікоса Цирітоса та інших [4], було обрано набір базових параметрів, що дозволяють описати поведінку потоків мережевого трафіку між парами $\langle \text{IP-адреса джерела, номер порту джерела} \rangle$, $\langle \text{IP-адреса призначення, номер порту призначення} \rangle$ (Таблиця 1). Впродовж підготовки набору вхідних даних на основі лабораторних зразків захоплених пакетів мережевого трафіку було обчислено значення 21-го параметра для кожної пари мережевих з'єднань.

2. Основні етапи формування моделі мережевого трафіку на основі МГК

Концепція застосування МГК (англ. Principal Component Analysis, PCA) полягає в тому, щоб виконати перехід до нового ортогонального базису (головних компонент), осі якого орієнтовані за напрямками максимальної дисперсії набору вхідних даних. Це перетворення дозволяє скоротити розмірність початкового простору ознак за рахунок нехтування напрямками, що мають мінімальну дисперсію. Таким чином, задача зводиться до пошуку таких лінійних комбінацій початкових показників, що пояснюють основну частку варіативності вхідних даних. Детально ознайомитись з алгоритмом МГК можна у монографії П. Брюса та Е. Брюса [5, с. 264].

В результаті підготовки набору вхідних даних, що включав 360 000 пакетів мережевого трафіку, було виділено 123 020 мережевих з'єднань, кожне з яких характеризується значеннями 21 параметра. Спира-

^аrobin14good@gmail.com

Табл. 1. Вхідні параметри

ID	Ім'я	Інтерпретація
1	total fpackets	Загальна кількість пакетів, що рухаються в прямому напрямку
2	total bpackets	Загальна кількість пакетів, що рухаються в зворотньому напрямку
3	min fpkttl	Мінімальна довжина пакета (у байтах) в прямого напрямку
4	max fpkttl	Максимальна довжина пакета (у байтах) в прямого напрямку
5	min bpkttl	Мінімальна довжина пакета (у байтах) в зворотньому напрямку
6	max bpkttl	Максимальна довжина пакета (у байтах) в зворотньому напрямку
7	min fiat	Мінімальний інтервал (у мікросекундах) між двома пакетами у прямому напрямку.
8	max fiat	Максимальний інтервал (у мікросекундах) між двома пакетами у прямому напрямку
9	min biat	Мінімальний інтервал (у мікросекундах) між двома пакетами у зворотньому напрямку
10	max biat	Максимальний інтервал (у мікросекундах) між двома пакетами у зворотньому напрямку
11	duration	Час, що минув (у мікросекундах) від першого пакета до останнього
12	min active	Мінімальний час, що минув (у мікросекундах) в підпотоці (якщо затримка між пакетами 0.5 секунди, то далі вже йде інший підпоток)
13	max active	Максимальний час, що минув (у мікросекундах) в підпотоці
14	min idle	Мінімальний час (у мікросекундах) затримки потоку
15	max idle	Максимальний час (у мікросекундах) затримки потоку
16	sflow fpackets	Середня кількість пакетів в прямому напрямку в підпотоках
17	sflow fbytes	Середня кількість байтів в прямому напрямку в підпотоках
18	sflow bpackets	Середня кількість пакетів в зворотньому напрямку в підпотоках
19	sflow bbytes	Середня кількість байтів в зворотньому напрямку в підпотоках
20	total fhlen	Загальна довжина заголовка (мережевий і транспортний рівень) пакетів, що рухаються в прямому напрямку
21	total bhlen	Загальна довжина заголовка (мережа та транспортний рівень) пакетів, що рухаються у зворотному напрямку

ючись на ці дані, сформовано матрицю спостережень

$$X = ||x_{ij}||, i = \overline{1, n}, j = \overline{1, m},$$

де $n = 123020$ — кількість мережевих з'єднань <IP-адреса джерела, номер порту джерела>, <IP-адреса призначення, номер порту призначення>, а $m = 21$ — кількість показників спостереження.

На основі матриці X визначено стандартизовану матрицю спостережень Z , значення z_{ij} якої визначаються як відхилення відповідних значень x_{ij} від середнього $\bar{x}_j$ в одиницях стандартного відхилення σ_j :

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}; \sigma_j = \sqrt{\frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{n - 1}}$$

Зі стандартизованої матриці Z розраховано кореляційну матрицю R за формулою $R = \frac{1}{n} Z'Z$, для якої визначено вектор власних чисел $[\lambda_1, \dots, \lambda_m]$ як розв'язок характеристичного рівняння $|R - \lambda E| = 0$ та матрицю відповідних власних векторів U як розв'язок матричного рівняння $(R - \lambda E)U = 0$.

Відомо, що вектор власних чисел вказує на значимість кожної з нових головних компонент. Відберемо мінімальну кількість головних компонент, сумарний відсоток дисперсії яких складатиме не менше 75%. Значення власних чисел, яким відповідають перші 4 головні компоненти, що задовольняють дану умову, наступні: $\lambda_1 \approx 6.7305$; $\lambda_2 \approx 4.0053$; $\lambda_3 \approx 3.2671$; $\lambda_4 \approx 1.9994$.

Таким чином, сумарний відсоток дисперсії, що по-

яснюють перші чотири головні компоненти, складає:

$$\frac{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4}{21} * 100\% \approx 76.2\%.$$

Вважатимемо надалі такий набір головних компонент достатнім для аналізу поведінки мережевого трафіку і перейдемо до встановлення взаємозв'язку відповідних наборів вхідних показників з головними компонентами.

Для цього розрахуємо матрицю факторних навантажень A за формулою:

$$A = V \times \Lambda^{\frac{1}{2}},$$

де V — матриця нормованих власних векторів $V_j = \frac{U_j}{|U_j|}$, $|U_j|$ — евклідова норма вектора U_j , $j = \overline{1, m}$; $\Lambda^{\frac{1}{2}} = \text{diag}\{\sqrt{\lambda_1}, \dots, \sqrt{\lambda_m}\}$, $m = 21$.

Матриця факторних навантажень дозволяє сформувати групи вхідних показників, що найбільшою мірою впливають на головні компоненти PC1, ..., PC4. Визначимо для кожної головної компоненти коефіцієнт інформативності K_U як відношення суми квадратів значимих коефіцієнтів кореляції до суми квадратів всіх коефіцієнтів кореляції. Набір обраних нами значимих параметрів вважається достатнім, якщо коефіцієнт інформативності K_U для кожної з головних компонент складатиме не менше 75% (Таблиця 2).

Спираючись на отримані результати, виконано обчислення матриці значень головних компонент за формулою $F = A^{-1}Z'$.

Табл. 2. Інтерпретація наборів показників для головних компонент

Коефіцієнт інформативності PC1: 0.77185		Коефіцієнт інформативності PC2: 0.75921		Коефіцієнт інформативності PC3: 0.79683		Коефіцієнт інформативності PC4: 0.99619	
Показники	Ваги	Показники	Ваги	Показники	Ваги	Показники	Ваги
total bpackets	0.9531	total fhlen	-0.6498	min active	0.6984	max fpctl	0.9979
total bhlen	0.9531	total fpackets	-0.6495	sflow fpackets	0.6905	min fpctl	0.9979
min bpctl	0.8744	sflow fbytes	-0.6251	sflow fbytes	0.6901		
max bpctl	0.8744	sflow fpackets	-0.6247	max active	0.6517		
sflow bbytes	0.8623	max active	-0.6087	duration	-0.6075		
sflow bpackets	0.8623	min active	-0.5653	max fiat	-0.6069		
total fpackets	0.6008	duration	-0.5116				
		max bpctl	-0.5075				
		max fiat	0.4541				

Табл. 3. Повна та пояснена дисперсії залишків для навчальних та тестових наборів

	(A_0, B_0)	(A_1, B_1)	(A_2, B_2)	(A_3, B_3)	(A_4, B_4)	(A_5, B_5)	(A_6, B_6)	(A_7, B_7)	(A_8, B_8)	(A_9, B_9)
TRVC	0.23937	0.23840	0.23583	0.23724	0.23799	0.23664	0.23756	0.23779	0.23825	0.2386
ERVC	0.76062	0.76159	0.76416	0.76275	0.76201	0.76335	0.76243	0.7622	0.76174	0.76139
TRVP	0.22555	0.25269	0.23518	0.23574	0.23680	0.22986	0.23893	0.23061	0.23594	0.23211
ERVP	0.75069	0.74728	0.76480	0.73941	0.76318	0.74593	0.7359	0.74509	0.76403	0.74343

3. Перевірка якості декомпозиції

Модель мережевого трафіку, побудована на основі методу головних компонент, спирається на низку вибірових статистик. Справедливо, що такий метод формування моделі потребує додаткового підтвердження (валідації) на різних наборах даних. Розглянемо підхід до перевірки якості декомпозиції методом тест-валідації.

При перевірці методом тест-валідації початковий масив даних складається з двох незалежно отриманих наборів. Перший набір називають навчальним та використовують для формування моделі. Другий набір називають перевірочним і застосовують лише для перевірки моделі. Модель, сформована на основі навчального набору даних, застосовується до даних з перевірочного набору, після чого отримані результати порівнюють з перевірочними значеннями. Таким чином, приймається рішення щодо якості та точності запропонованої моделі.

Результатом моделювання МГК є величини $\hat{X}_c$ — оцінки, знайдені за моделлю, що побудована на навчальному наборі X_c . Як результат перевірки використовуються величини $\hat{X}_t$ — оцінки перевірочних значень X_t , обчислені за тією ж моделлю, але як нові зразки. Відхилення оцінки від перевірочного значення обчислюють як матрицю залишків:

- в навчанні:

$$E_c = X_c - \hat{X}_c;$$

- в перевірці:

$$E_t = X_t - \hat{X}_t. \quad (1)$$

Наступні величини характеризують якість моделювання в середньому. Повна дисперсія залишків:

- в навчанні

$$TRVC = \frac{1}{I_c J} \sum_{i=1}^{I_c} \sum_{j=1}^J e_{ij}^2;$$

- в перевірці

$$TRVP = \frac{1}{I_t J} \sum_{i=1}^{I_t} \sum_{j=1}^J e_{ij}^2. \quad (2)$$

Пояснена дисперсія залишків:

- в навчанні

$$ERVC = 1 - \frac{\sum_{i=1}^{I_c} \sum_{j=1}^J e_{ij}^2}{\sum_{i=1}^{I_c} \sum_{j=1}^J x_{ij}^2};$$

- в перевірці

$$ERVP = 1 - \frac{\sum_{i=1}^{I_t} \sum_{j=1}^J e_{ij}^2}{\sum_{i=1}^{I_t} \sum_{j=1}^J x_{ij}^2}. \quad (3)$$

4. Результати перевірки якості декомпозиції

Для перевірки якості декомпозиції методом тест-валідації з початкового масиву даних заздалегідь було виключено тестовий набір, що не брав участі у формуванні моделі. Частка пакетів мережевого трафіку для тестового набору щодо навчального складалась у відношенні 1:9.

В результаті моделювання МГК отримано величини $\hat{X}_c$ — оцінки, знайдені за моделлю, що побудована на навчальному наборі X_c , а також величини $\hat{X}_t$ — оцінки перевірочних значень X_t , обчислені за тією ж моделлю, але як нові зразки. Обчислено матриці залишків в навчанні E_c та в перевірці E_t згідно з формулою (1).

Спираючись на методику оцінки якості моделювання в середньому, повна дисперсія залишків, обчислена за формулами (2), складає в навчанні $TRVC = 0.23799$; в перевірці $TRVP = 0.23680$. Пояснена дисперсія залишків, обчислена за формулами (3), складає в навчанні $ERVC = 0.76201$; в перевірці $ERVP = 0.76318$. Результати перевірки демонструють, що в середньому якість моделювання тестових

вого набору не нижча, ніж начального, і розходиться в межах статистичних похибок.

5. Алгоритм к-блочної перехресної перевірки

Класичні статистичні метричні показники регресії застосовуються до тих же даних, що були використані при розробці моделі, саме тому для незалежної перевірки якості моделювання частину початкового масиву спостережень слід відділити заздалегідь. А також, у випадку коли загальний обсяг даних суттєво обмежений, доцільно скористатись алгоритмом базової к-блочної перехресної перевірки (англ., K-Fold Cross Validation) [5, с. 124]:

- 1) Відкласти $1/k$ даних в якості контрольної вибірки.
- 2) Натренувати модель на даних, що залишились.
- 3) Застосувати модель до контрольної вибірки $1/k$ (оцінити результати) та записати необхідні метричні показники діагностики моделі.
- 4) Відновити перші $1/k$ даних та відкласти наступні $1/k$ (виключаючи будь-які записи, що були вибрані вперше).
- 5) Повторити кроки 2 та 3.
- 6) Повторювати доки кожен запис не буде використаний у відсотковій долі, визначеній для контрольної вибірки.
- 7) Усереднити чи скомбінувати метричні показники діагностики моделі.

6. Валідація моделі в умовах перехресної перевірки

Розглянемо розширений набір лабораторних зразків пакетів мережевого трафіку, захоплених в ході атаки класу DDoS SYN Flood, як універсальну множину B , що складається з 400 000 пакетів. Виконаємо розбиття даного набору на 10 рівних блоків по 40 000 пакетів кожен відповідно до схеми, зображеної в Таблиці 4: $\{B_0, B_1, \dots, B_9\}$. Таким чином, в ході перехресної перевірки елементи B_i , $i = \overline{0,9}$ такого розбиття по черзі виконуватимуть роль тестового набору. Тому визначимо навчальні набори для кожного тестового як абсолютне доповнення до множини B : $A_i = \overline{B_i} = B \setminus B_i$, $i = \overline{0,9}$.

Табл. 4. Розподіл k блоків, $k = 10$

	B_0	B_1	B_2	...	B_9
Ітерація 1	Тест.	Навч.	Навч.	...	Навч.
Ітерація 2	Навч.	Тест.	Навч.	...	Навч.
Ітерація 3	Навч.	Навч.	Тест.	...	Навч.
...	...	...	...	...	...
Ітерація 10	Навч.	Навч.	Навч.	...	Тест.

Варто зазначити, що саме на прикладі пари наборів (A_4, B_4) було продемонстровано підхід до перевірки якості декомпозиції вище. Результати тест-валідації в умовах перехресної перевірки для всіх пар навчальних та тестових наборів (A_i, B_i), $i = \overline{0,9}$ наведено в Таблиці 3. Слід звернути увагу, що моделі з найкращими оцінками якості в середньому були

виконані на навчальних наборах A_2, A_4, A_8 , проте суттєвих погіршень якості декомпозиції в умовах перехресної перевірки для решти моделей виявлено не було.

Висновки

Побудовано модель даних мережевого трафіку на основі МГК та реалізовано системний підхід до перевірки якості моделювання потоків мережевого трафіку, в рамках якого:

- 1) Проведено перевірку якості декомпозиції методом тест-валідації, під час якої виконано оцінку якості моделювання в середньому шляхом порівняння величин повної та поясненої дисперсії залишків для навчального та тестового наборів відповідно;
- 2) Реалізовано алгоритм к-блочної перехресної перевірки для оцінки якості моделювання розширеного набору лабораторних зразків пакетів мережевого трафіку, захоплених в ході атаки класу DDoS SYN Flood.

Результати к-блочної перехресної перевірки демонструють, що відхилення значень повної та поясненої дисперсії залишків для навчальних та тестових наборів є допустимими. Моделі з найкращими оцінками якості в середньому були виконані на навчальних наборах A_2, A_4, A_8 , проте суттєвих погіршень якості декомпозиції для решти моделей виявлено не було.

Таким чином, результати оцінювання якості моделювання підтверджують, що отримані впродовж перевірки оцінки дисперсії залишків в навчанні та в перевірці є передбачувано прийнятними і дозволяють якісним чином моделювати поведінку мережевого трафіку за допомогою МГК.

Перелік використаних джерел

1. Нестеренко Є. С., Стремецька М. С. Проблема вибору системи показників для аналізу PCAP-файлів // Матеріали XVIII Всеукраїнської науково-практичної конференції «Теоретичні і прикладні проблеми фізики, математики та інформатики». — 2020. — С. 97–100.
2. DDoS Evaluation Dataset (CIC-DDoS2019) // UNB. — 2019. — Access mode: <https://www.unb.ca/cic/datasets/ddos-2019.html>.
3. "Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy" / I. Sharafaldin, A. H. Lashkari, S. Hakak, A. A. Ghorbani. — IEEE 53rd International Carnahan Conference on Security Technology. Chennai, India, 2019.
4. The Next Generation Cognitive Security Operations Center: Adaptive Analytic Lambda Architecture for Efficient Defense against Adversarial Attacks / K. Demertzis, N. Tziritas, P. Kikiras et al. // Big Data and Cognitive Computing. — 2019. — Vol. 3, no. 1. — Access mode: <https://www.mdpi.com/2504-2289/3/1/6>.
5. Брюс П., Брюс Э. Практическая статистика для специалистов Data Science. — 2018. — 304 с.