УДК 004.056.5

ФАКТОРНИЙ АНАЛІЗ СИСТЕМИ ПОКАЗНИКІВ ДЛЯ АНАЛІЗУ PCAP-ФАЙЛІВ В УМОВАХ ПЕРЕХРЕСНОЇ ПЕРЕВІРКИ

В. В. Деркач[1, a], М. С. Стремецька[1]

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

Анотація

В умовах різкого зростання попиту на автоматизацію процесів моніторингу мережевого трафіку виникає гостра необхідність пошуку якісних моделей, що дозволяють проводити виявлення аномальної мережевої активності в режимі реального часу. Побудовано факторну модель як систему показників для аналізу перехоплених пакетів мережевому трафіку засобами факторного аналізу. За критерієм Кайзера визначено 5 прихованих факторів, що в сумі пояснюють 85.38% дисперсії даних. Факторні навантаження знайдено як оцінки максимальної правдоподібності на основі SVD підходу. Для закріплення положення матриці навантажень застосовано метод ротації факторів Quartimax. Реалізовано алгоритм k-блочної перехресної перевірки (k=10) для оцінки якості факторних моделей розширеного набору лабораторних зразків пакетів мережевого трафіку, захоплених в ході атаки класу DDoS SYN Flood. За результатами перевірки на основі показників повної та залишкової дисперсії для навчальних та тренувальних наборів виявилось, що факторні моделі з найкращими оцінками якості в середньому були виконані на навчальних наборах C_2, C_4, C_8, проте суттєвих погіршень якості для решти моделей виявлено не було.

Ключові слова: DDoS SYN flood, метод факторного аналізу, метод Кайзера, метод максимальної правдоподібності, Singular Value Decomposition (SVD), ротація факторів, Quartimax, k-блочна перехресна перевірка.

Вступ

Важливим етапом на шляху організації ефективної системи моніторингу подій інформаційної безпеки є розробка якісних факторних моделей мережевого трафіку, що в поєднанні зі статистичними методами дослідження забезпечують виявлення аномальної мережевої активності в режимі реального часу. Застосування факторного аналізу дозволяє зменшити розмірність простору початкових параметрів шляхом їх заміни на меншу кількість нових факторів, обрахованих за допомогою лінійного перетворення. Для утворення факторів використовується спільна дисперсія кожного параметру, що дозволяє детермінувати структуру взаємозв'язків між змінними та виявити наявні приховані зв'язки. Отримані в ході формування моделі значення унікальних дисперсій дозволяють визначити, які з параметрів не корелюють з іншими та потребують окремого дослідження. Зменшення розмірності допомагає пришвидшити роботу алгоритмів аналізу даних, а формування структури взаємозв'язків між змінними — інтерпретувати отримані в ході аналізу результати.

Мета роботи. Сформувати систему показників для аналізу перехоплених пакетів мережевого трафіку засобами дослідницького факторного аналізу. Реалізувати перевірку якості моделювання в умовах k-блочної перехресної перевірки для пошуку факторної моделі з найкращими оцінками якості в середньому.

[a] stopearth72@gmail.com

1. Методика факторного аналізу

Концепція дослідницького факторного аналізу базується на визначенні певної кількості прихованих змінних (факторів), що пояснюють кореляцію між початковими параметрами [1]. Тобто виконується дослідження багатовимірного простору ознак, описаних набором початкових параметрів, в ході якого виконується аналіз дисперсії та кореляцій між початковими параметрами з точки зору меншої кількості прихованих змінних. Дослідницька факторна модель може бути представлена у наступному вигляді:

$$X - \mu = \Lambda F + e$$

де X — початкові параметри (спостережуванні змінні), μ — середні значення спостережень, Λ — матриця факторних навантажень, F — фактори (приховані змінні), e — помилки (залишки).

Матриця навантажень Λ визначається шляхом наближення матриці Σ до матриці коваріацій $Cov(X - \mu)$ з використанням методу оцінки максимальної правдоподібності.

Матриця Σ визначається наступним чином:

$$\Sigma = \Lambda\Lambda^T + \Psi, \text{ де } \Psi = Cov(e). \quad (1)$$

У рівнянні (1) доданок $\Lambda\Lambda^T$ відповідає тій частині дисперсії, що пояснюється лінійною комбінацією факторів, а Ψ — невідома частина загальної дисперсії, яку неможливо пояснити за допомогою лінійної комбінації факторів.

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 fpktl	Мінімальна довжина пакета (у байтах) в прямого напрямку
4	max fpktl	Максимальна довжина пакета (у байтах) в прямого напрямку
5	min bpktl	Мінімальна довжина пакета (у байтах) в зворотньому напрямку
6	max bpktl	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

2. Визначення факторних навантажень як оцінок максимальної правдоподібності на основі SVD підходу

Метод отримання оцінок максимальної правдоподібності спирається на ітеративний алгоритм, в ході якого використовуються наступні припущення:

- Початкова матриця невідомих дисперсій є одиничною: $\hat{\Psi} = E$;
- Бажаний рівень точності алгоритму задається константою δ (приймемо $\delta = 10^{12}$);
- Визначимо, що n_s — кількість спостережень, n_f — кількість параметрів, n_c — кількість факторів, тоді в ході кожної ітерації алгоритму обраховуватимемо наступне:
  1) $\hat{\Psi}_{sqrt} = \sqrt{\hat{\Psi}} + \delta$
  2) $\hat{X} = X/(\hat{\Psi}_{sqrt}\sqrt{n_s})$
  3) Застосуємо SVD розклад до $\hat{X}$: $\hat{X} = USV^T$, де U — матриця лівих сингулярних векторів, S — матриця сингулярних чисел та V^T — транспонована матриця правих сингулярних векторів. Надалі використовуватимемо матрицю сингулярних чисел $\hat{S}$ та транспоновану матриця правих сингулярних векторів $\hat{V}^T$, обмежені обраною кількістю факторів.
  4) Розрахуємо приблизні значення матриці навантажень $\hat{\Lambda}$ та матриці невідомих дисперсій $\hat{\Psi}$:

$$\hat{\Lambda} = \sqrt{\hat{S}^2 - E}\hat{V}^T\hat{\Psi}_{sqrt};$$

$$\hat{\Psi} = max(\textstyle\sum_{j=1}^{n_f} {x_j}^2 - \sum_{j=1}^{n_c} {\lambda_j}^2, \delta).$$

  5) Обчислимо логарифмічну функцію правдоподібності:

$$lnL(x, S, \Psi) = -\frac{n_s}{2}(n_s ln(2\pi) + n_c + \\ + \textstyle\sum_i^{n_c} ln(\hat{s}_i^2) + \sum_i^{n_s}\sum_j^{n_c} x_{ij}^2 - \sum_i^{n_c} \hat{s}_i^2 + \\ + \textstyle\sum_i^{n_f} ln(\hat{\psi}_i)).$$

Після кожної ітерації обчислене значення логарифмічної функції правдоподібності порівнюється зі значенням, отриманим на попередній ітерації. У разі якщо їх різниця менше δ, робота алгоритму завершується, а матрицям Λ та Ψ присвоюються значення, отримані на попередній ітерації алгоритму.

3. Ротація факторів з використання методу Quartimax

Основна проблема при виконанні факторного аналізу полягає в тому, що отримана матриця факторних навантажень Λ може бути не єдиною. В даному випадку для закріплення положення матриці навантажень застосовується метод обертання факторів Quartimax, алгоритм якого є також ітеративним.

Початкова матриця ротації визначається як одинична матриця: $R = E_{n_c}$ На кожній ітерації алгоритму обраховується наступне:

1) Ротація факторів: $\hat{\Lambda} = \Lambda^T(\Lambda R)^3$
2) SVD розклад до $\hat{\Lambda}$: $\hat{\Lambda} = USV^*$
3) Розрахується нова матриця ротації: $R = UV^T$

Після кожної ітерації обчислюється сума сингулярних чисел S, що порівнюється з попереднім значенням (початкове значення приймається за 0). Якщо

Табл. 2. Факторні навантаження

	Фактор 1	Фактор 2	Фактор 3	Фактор 4	Фактор 5
total fpackets	0.011708	-0.207555	-0.12595	0.969954	0.002505
total bpackets	-0.004107	-0.040912	-0.95667	0.253059	0.005152
min fpktl	0.999874	-0.002444	0.003581	0.014989	0.001163
max fpktl	0.999874	-0.002444	0.003581	0.014989	0.001163
min bpktl	-0.001036	0.015442	-0.992048	0.05545	0.000454
max bpktl	-0.001036	0.015442	-0.992048	0.05545	0.000454
min fiat	0.001278	0.016713	0.00365	-0.073928	0.000366
max fiat	-0.005452	0.05224	-0.113543	0.875725	-0.035403
min biat	0.000409	0.014301	-0.268002	-0.02909	-0.005214
max biat	-0.00893	0.020888	-0.159139	0.600242	-0.013718
duration	-0.001909	0.051244	-0.114516	0.886034	-0.033826
min active	-0.002668	-0.829525	-0.030558	0.035044	0.002029
max active	-0.003899	-0.829121	-0.044501	0.118631	0.003991
min idle	0.000858	-0.012827	0.013048	-0.197144	0.971745
max idle	-0.001614	-0.003278	0.029519	-0.151299	-0.97889
sflow fpackets	-0.00399	-0.994909	-0.035706	0.094036	0.003671
sflow fbytes	0.013648	-0.994803	-0.035637	0.094286	0.003691
sflow bpackets	-0.000298	-0.056608	-0.998374	-0.001751	0.005745
sflow bbytes	-0.000298	-0.056608	-0.998374	-0.001751	0.005748
total fhlen	0.035511	-0.20726	-0.125678	0.969458	0.002538
total bhlen	-0.004107	-0.040911	-0.95667	0.253055	0.005157

сума менша за попереднє значення, то алгоритм завершує роботу, а матриці ротації присвоюється значення, отримане на попередній ітерації.

4. Формування моделі мережевого трафіку на основі методу факторного аналізу

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. Обраний набір вхідних даних включав 360 000 пакетів мережевого трафіку, з якого було виділено 123238 мережевих з'єднань виду <IP-адреса джерела, номер порту джерела>, <IP-адреса призначення, номер порту призначення>, описаних значеннями 21 параметра (Табл. 1). На основі наявних даних було сформовано матрицю спостережень $X = ||x_{ij}||$, $i = \overline{1,n}$, $j = \overline{1,m}$, де $n = 123238$ — кількість мережевих з'єднань, $m = 21$ — кількість параметрів.

Над матрицею спостережень X була проведена процедура стандартизації, в ході якої сформовано стандартизовану матрицю спостережень Z:

$$z_{ij} = \frac{x_{ij} - \overline{x_j}}{\sigma_j}, \text{ де}$$

$$\overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \quad \sigma_j = \sqrt{\frac{\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2}{n-1}}$$

Для визначення кількості прихованих факторів було обраховано вектор власних чисел кореляційної матриці $Cor(Z)$. Згідно з методом Кайзера, брались до уваги лише ті фактори, значення власних чисел яких більші за одиницю: $\lambda_1 \approx 6.69$, $\lambda_2 \approx 4.03$, $\lambda_3 \approx 3.28$, $\lambda_4 \approx 2.00$, $\lambda_5 \approx 1.91$

Таким чином, сумарний відсоток дисперсії, що пояснюють відібрані 5 факторів, складає:

$$\frac{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5}{21} * 100\% \approx 85.38\%$$

Отримана дисперсія може здаватися досить значною, однак при проведенні перехресної перевірки було з'ясовано, що при зменшенні кількості факторів до 4, втрати інформативності під час застосування моделі до тестових наборів даних значно зростають. Саме тому, задля збереження рівня інформативності, було прийнято рішення використовувати 5 факторів.

Після вибору кількості шуканих факторів, виконано розрахунок матриці навантажень та одразу застосовано до неї метод ротації Quartimax. В результаті було отримано фінальну матрицю навантажень (Табл. 2), що зокрема описує факторну модель.

Для отриманої моделі було розраховано частки дисперсії, що пояснює кожним з факторів:

- для Фактора 1: 9.529%;
- для Фактора 2: 16.436%;
- для Фактора 3: 28.51%;
- для Фактора 4: 19.059%;
- для Фактора 5: 9.07%.

Таким чином, загальна дисперсія, описана запропонованою моделлю, складає 82.6%. Однак, при аналізі залишкових дисперсії було з'ясовано, що отримані фактори недостатньо інтерпретують два параметри, а саме: min_fiat та min_biat. Частки їх дисперсій, описані знайденими факторами, складають лише 0.576% та 7.291% відповідно.

Тож, якщо параметри min_fiat та min_biat будуть враховуватись на рівні з обрахованими факторами, загальна дисперсія, описана факторною моделлю, складатиме 90.8%. Запропонований підхід доцільно використовувати у випадках, коли критично важливо максимізувати інформативність моделі, проте з метою оптимізації швидкодії алгоритмів у випадку

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

	(C_0, T_0)	(C_1, T_1)	(C_2, T_2)	(C_3, T_3)	(C_4, T_4)	(C_5, T_5)	(C_6, T_6)	(C_7, T_7)	(C_8, T_8)	(C_9, T_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

обробки великих обсягів даних такими факторами допускається знехтувати.

5. Перехресна перевірка

Проведено перевірку якості факторної моделі за методом тест-валідації для розширеного набору лабораторних зразків пакетів мережевого трафіку в умовах алгоритму базової k-блочної перехресної перевірки (англ., K-Fold Cross Validation), що описана в матеріалах [4].

Розширений набір лабораторних зразків пакетів мережевого трафіку, захоплених в ході атаки класу DDoS SYN Flood, розглянемо як універсальну множину T, що складається з 400 000 пакетів. Виконаємо розбиття даного набору на 10 послідовних рівних блоків по 40 000 пакетів відповідно: $\{T_0, T_1, \ldots, T_9\}$. Таким чином, в ході перехресної перевірки елементи T_i, $i = \overline{0,9}$ такого розбиття по черзі виконуватимуть роль тестового набору. Далі визначимо навчальні набори для кожного тестового як абсолютне доповнення до множини T: $C_i = \bar{T}_i = T \backslash T_i$, $i = \overline{0,9}$.

Факторна модель, сформована на основі навчального набору даних, застосовується до даних відповідного перевірочного набору, після чого отримані результати порівнюють з перевірочними значеннями. Таким чином, приймається рішення щодо якості та точності запропонованої моделі.

Наступні величини слугують для оцінки якості моделювання в середньому [5]:

- TRVC — повна дисперсія залишків в навчанні;
- ERVC — пояснена дисперсія залишків в навчанні;
- TRVP — повна дисперсія залишків в тестуванні;
- ERVP — пояснена дисперсія залишків в тестуванні.

Результати тест-валідації в умовах перехресної перевірки для всіх пар навчальних та тестових наборів (C_i, T_i), $i = \overline{0,9}$ наведено в Таблиці 3. Слід зазначити, що найкращі результати якості моделювання були отримані факторними моделями, сформованими на наступних навчальних наборах даних: C_2 з показниками: TRVC = 0.17227, ERVC = 0.82773, TRVP = 0.208, ERVP = 0.79199,
C_4 з показниками: TRVC = 0.17341, ERVC = 0.82659, TRVP = 0.20773, ERVP = 0.79226,
C_8 з показниками: TRVC = 0.17502, ERVC = 0.82498, TRVP = 0.20733, ERVP = 0.79266.

Висновки

Виконано дослідження системи показників для аналізу захоплених пакетів мережевого трафіку за допомогою методу факторного аналізу, в ході якого:

- За критерієм Кайзера визначено 5 прихованих факторів, що в сумі пояснюють 85.38% дисперсії даних;
- За методом максимальної правдоподібності та алгоритмом ротації Quartimax розраховано матрицю навантажень, що описує зв'язок між факторами 1-5 та вхідними параметрами.

Реалізовано системний підхід до перевірки якості моделювання потоків мережевого трафіку, в рамках якого: проведено оцінку якості факторної моделі в середньому шляхом порівняння величин повної та поясненої дисперсії залишків для навчального та тестового наборів відповідно; виконано перевірку якості факторних моделей за методом тест-валідації для розширеного набору лабораторних зразків пакетів мережевого трафіку в умовах алгоритму базової k-блочної перехресної перевірки.

Дослідження показало, що завдяки застосуванню алгоритму ротації Quartimax отримані в ході перехресної перевірки моделі демонструють однаково високі результати як на навчальних, так і на тестових наборах. В свою чергу, моделі з найкращими оцінками якості в середньому були виконані на навчальних наборах C_2, C_4, C_8. Таким чином, факторний аналіз є дієвим методом для скорочення кількості змінних, необхідних для опису поведінки мережевого трафіку, що забезпечує сприятливі умови для формування простих та ефективних механізмів виявлення підозрілої мережевої активності.

Перелік використаних джерел

1. Hartmann K., Krois J., Waske B. Factor analysis. E-Learning Project SOGA: Statistics and Geospatial Data Analysis. Department of Earth Sciences, Freie Universitaet Berlin. — 2018. — Access mode: `https://www.geo.fu-berlin.de/en/v/soga/Geodata-analysis/factor-analysis/index.html`.
2. DDoS Evaluation Dataset (CIC-DDoS2019). — Access mode: `https://www.unb.ca/cic/datasets/ddos-2019.html`.
3. Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy / Sharafaldin I., Lashkari A. H., Hakak S., and Ghorbani A. A. // 2019 International Carnahan Conference on Security Technology (ICCST). — 2019. — P. 1–8.
4. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. — 2001.
5. А. Померанцев. Калибровка. — Режим доступа: `https://rcs.chemometrics.ru/old/Tutorials/calibration.html`.