ВИПАДКОВИЙ ЛІС: ПОШУК АНОМАЛІЙ У КОМП'ЮТЕРНИХ МЕРЕЖАХ

Денис Сташук, Валерій Дубровін, Юлія Тарасова
Національний університет «Запорізька політехніка»
FRESHCODE

RANDOM FOREST: FINDING ANOMALIES IN COMPUTER NETWORKS

Abstract. Random Forest is one of the most popular and commonly used algorithms by Data Scientists. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables, as in the case of regression, and categorical variables, as in the case of classification.

Постановка проблеми. Виявлення аномалій поведінки мережі – це процес моніторингу корпоративних мереж для виявлення ненормальної поведінки. Після виявлення аномалії функція виявлення аномалії поведінки мережі або ініціює автоматичну відповідь, або сповіщає групи безпеки. Проблеми в комп'ютерних мережах виявляються як викликані ними аномалії трафіку[1]. Загалом, аномалія – це те, що суперечить очікуванням. Наприклад, пошкоджений комутатор може створити неочікуваний трафік в іншій частині мережі або нові коди помилок починають з’являтися, коли служба не працює. Усунення несправностей мережі ґрунтується на аномаліях мережі. Перший метод класифікації аномалій заснований на тому, чим вони відрізняються від звичайного спілкування. Аномалії можуть відрізнятися за типом переданих даних (поведінкові), за кількістю переданих даних (за обсягом) або за обома критеріями. Інший спосіб класифікації аномалій - це залежно від їх причини:

1. Нелюдська помилка - збій обладнання або переривання радіозв'язку через погоду.
2. Людська помилка - збій служби мережі, спричинений неправильною конфігурацією або випадковим від’єднанням мережевого кабелю.
3. Зловмисна діяльність людини – внутрішня атака, коли незадоволений співробітник компанії пошкоджує мережевий принтер, або зовнішня атака, коли зловмисник намагається вимкнути мережу та завдати шкоди репутації.

Одним з методів виявлення аномалій у комп’ютерних мережах є випадковий ліс[2]. Для дослідження роботи алгоритму випадкового лісу розглянемо його роботу у гібридній системі виявлення вторгнень на основі комбінації випадкового лісу та автоенкодера.

Модель системи. Оскільки методи мережевих атак ускладнюються, отримати зразки всіх типів атак стає складно. Під час зустрічі з невідомими атаками під час фази виявлення контрольований класифікатор може погано узагальнювати результат, що може призвести до неправильної класифікації для цих зразків і зниження рівня виявлення. Щоб впоратися з цією проблемою, дослідники спробували створити модель системи виявлення вторгнень, яка тренується тільки на звичайних даних, таких як автоенкодер[3]. Він представляє собою спеціальну неконтрольовану нейронну мережу. Оскільки вона може відновити вхідні дані, помилка реконструкції може служити показником аномалії для її виявлення. Автоенкодер зазвичай має структуру симетрії. Його можна розділити на дві частини: кодер і декодер. Така система намагається якомога більше реконструювати оригінальні вхідні дані. І кодер, і декодер складаються з кількох прихованих шарів. Зокрема, кодер генерує приховане представлення для вхідного зразка. Зазвичай латентні представлення мають нижчу розмірність, ніж оригінальні вхідні дані. Декодер намагається відновити оригінальний вхід зі стиснутої форми. Автоенкодер використовує помилку реконструкції як оцінку аномалії, де один зразок із вищим показником може бути виявлений як атака. Однак через відсутність контролю з боку звичайних зразків і зразків атаки він може не отримати такої ж високої продуктивності, як контрольовані алгоритми, які можуть вивчати межі складних рішень між звичайними зразками та зразками атаки.

Розробка надійної моделі системи виявлення вторгнень з вищим рівнем виявлення та нижчим рівнем помилкових позитивних результатів стає критичною. Зважаючи на те, що різні методи можуть вивчати характеристики даних з різних точок зору, краще використовувати гібридну систему[4], яка поєднує випадковий ліс і автоенкодер. Зокрема, гібридна система включає два етапи на етапі виявлення. Першим кроком є застосування випадкового лісу з імовірнісними методами для виявлення атак. Тоді, розглядаючи, як ще більше зменшити частоту помилкових тривог, на другому кроці використовується модуль автоенкодера.

Алгоритм випадкового лісу. Щоб отримати кращу ефективність прогнозування, ансамблеві методи використовують кілька основних класифікаторів для прийняття рішень. Цей підхід загалом демонструє підвищену продуктивність класифікації порівняно з одним базовим класифікатором. Випадковий ліс є потужним класифікатором ансамблю. Він поєднує пакетування та випадковість функцій для навчання кількох дерев рішень. Дерево рішень є часто використовуваним класифікаційним

підходом. Він намагається вивчити набір правил «якщо-тоді» для класифікації даних. Ілюстрація випадкового лісу показана на рис.1.

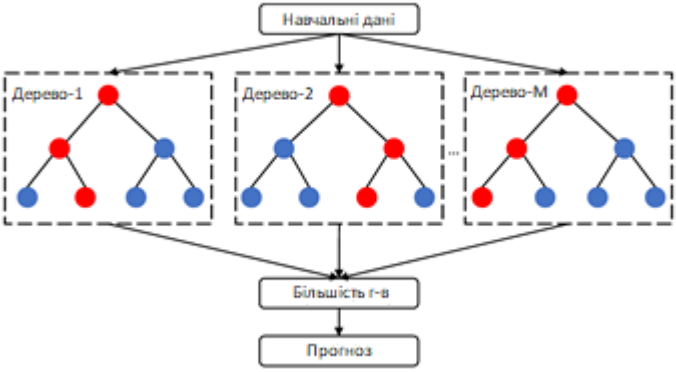


Рис1. - Ілюстрація випадкового лісу

Підхід пакетування є однією з часто використовуваних стратегій ансамблю для побудови кількох класифікаторів. За допомогою початкового завантаження із заміною для вибірки даних велика кількість класифікаторів навчається незалежно. Як показано на рисунку, з урахуванням навчального набору існує *M* різних класифікаторів рішення. Для отримання остаточних результатів передбачення використовується мажоритарне голосування для об'єднання прогнозів від кожного дерева рішень [5].

Щоб побудувати одне дерево рішень, спочатку вибираємо k ознак із j ознак. Значення k встановлюється як $sqrt(j)$. Після цього вибираємо найкращу функцію розділення з вибраних k функцій і ділимо вузол на два дочірні вузли. Домішка Джіні використовується як критерій розщеплення для розділеного вузла. Щоб передбачити вхідну вибірку, випадковий ліс використовує голоси дерев у лісі, зважені за їхніми оцінками ймовірності. Нехай p позначає ймовірність бути передбаченим як атака. Тоді ймовірність бути передбаченим як нормальний дорівнює q і $q = 1- p$. Прогнозовані ймовірності класів вхідної вибірки обчислюються як середні прогнозовані ймовірності класів дерев у лісі. Імовірність класу окремого дерева – це частка зразків одного класу в листовому вузлі [6].

Зазвичай зразки можна віднести до одного класу з найбільшою ймовірністю. Однак у такому підході деякі зразки, що належать до невідомих атак, можуть бути помилково класифіковані як нормальні. У цій частині потрібно визначити поріг T, щоб допомогти процесу

визначення зразків. Якщо ймовірність p, що належить до атаки, перевищує поріг T, вибірки можна класифікувати як атаку. Таким чином, із зразком x_i і відповідною ймовірністю p_i, ми визначаємо вирішувальну функцію $f(\cdot)$. Результати позначені ±1, де +1 - це аномальні зразки:

$$f(x_i) = \begin{cases} -1, & if\ p_i \leq T \\ +1, & if\ p_i > T \end{cases}$$

Переваги алгоритму:

1. Його можна використовувати в задачах класифікації та регресії.

2. Вирішує проблему переобладнання, оскільки вихід базується на більшості голосів або усередненні.

3. Добре працює, навіть якщо дані містять нульові/відсутні значення.

4. Кожне створене дерево рішень є незалежним від іншого; таким чином, він демонструє властивість розпаралелювання.

5. Стабільний, оскільки беруться середні відповіді, надані великою кількістю дерев.

6. Зберігає різноманітність, оскільки не всі атрибути враховуються під час створення кожного дерева рішень, хоча це не так у всіх випадках.

7. Несприйнятливий до розмірності. Оскільки кожне дерево не враховує всі атрибути, простір функцій зменшується.

8. Не потрібно розділяти дані на тренування та тестування, оскільки завжди буде певний відсоток даних, які не видно дереву рішень, створеному за допомогою початкового завантаження.

Недоліки алгоритму:

1. Дуже складний порівняно з деревами рішень, де рішення можна приймати, дотримуючись шляху дерева.

2. Час навчання більше, ніж інші моделі через його складність.

Таким чином, випадковий ліс є прийнятним вибором, якщо хтось хоче створити модель швидко та ефективно, оскільки одна з найкращих переваг випадкового лісу полягає в тому, що він може обробляти відсутні значення. Це один із найкращих методів із високою ефективністю, який широко використовується в різних галузях промисловості завдяки своїй ефективності. Він може обробляти двійкові, безперервні та категоричні дані. Загалом випадковий ліс є швидкою, простою, гнучкою та надійною моделлю з деякими обмеженнями.

Література

1. J. P. Anderson, Computer Security Threat Monitoring and Surveillance, Technical Report, James Anderson Report, Pennsylvania, (1980).

2. L. Breiman, Random Forests, Machine Learning, vol. 45, no. 1, pp. 5–32, (2001).

3. A. Hwoij, M. Al-Kasassbeh, and M. Al-Fayoumi, “Detecting network anomalies using rule-based machine learning within SNMP-MIB dataset,” *Cryptography and Security*, 2020, arXiv:2002.02368.

4. Breiman, L. Random Forests. *Mach. Learn.* **2001**, 45, 5–32.

5. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, 12, 2825–2830.

6. Géron, A. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*; O’Reilly Media, Inc.: Sebastopol, CA, USA, 2022.