Виявлення аномалій методами Machine Learning

УДК 004.056.57 (043.2) Сергій Бондаровець[1], Оксана Коваль[2], Чженбін Ху[3]

Національний авіаційний університет, Центрально-китайський нормальний університет (Китай), [1]bondss29@gmail.com, [2]oksanakoval@mail.ua, [3]hzb@mail.ecnu.edu.cn

Розвиток інформаційних технологій створив усі передумови для справжньої революції цифрових даних. Проте постійне зростання об'ємів таких даних зумовлює також збільшення кількості кіберзагроз, що швидко поширюються і стають серйозною загрозою для організацій. Існуючі системи моніторингу аналізують величезну кількість трафіку, особливо у час популярності Інтернету речей (Internet of Things) і сповіщають про тисячі інцидентів, які можуть становити небезпеку для критичних ресурсів. Але зі зростанням обсягів даних виявити незвичну активність у наявних системах стає все складніше. Для вирішення цих проблем існує група методів, які порівнюють поточну активність з т.з. «базовою» або «нормальною» і шукають відхилення, – методи виявлення аномалій. Враховуючи необхідність у прийнятті рішення в реальному часі, одними із найефективніших методів на сьогодні є Machine Learning. З огляду на це, *метою роботи* є аналіз існуючих методів Machine Learning, які можна використати у процесі виявлення аномалій, а також виділення їх основних переваг та недоліків.

Зазначені методи доцільно поділити на 4 категорії: методи нечіткої логіки, генетичні алгоритми, нейронні мережі та байесові мережі.

Нечітка логіка є похідною від теорії нечітких множин, при якій припущення є наближеними, а не точно виведеними, як у класичній предикативній логіці. Методи нечіткої логіки, таким чином, використовуються для виявлення аномалій, головним чином тому, що ознаки, які слід враховувати, можна розглядати як нечіткі змінні. Втім, хоча методи нечіткої логіки довели свою ефективність, особливо проти таких видів атак, як зондування і сканування портів, їх основним недоліком є використання значної кількості ресурсів.

Генетичні алгоритми є пошуковими евристичними методами, які засновані на біологічних процесах та використовують методи еволюційного алгоритму, такі як: схрещування, наслідування, мутації, вибірка тощо. Таким чином, генетичні алгоритми здатні виводити правила класифікації та обирати оптимальні параметри для процесу виявлення. Застосування генетичного алгоритму до мережевого трафіку здебільшого складається з таких кроків: 1) система виявлення вторгнень збирає дані про трафік, що циркулює у певній мережі; 2) після цього система виявлення вторгнень застосовує генетичні алгоритми, у яких закладені правила класифікації, вивчені із зібраної інформації системою виявлення вторгнень; 3) використовуючи набір правил система виявлення вторгнень класифікує вхідний трафік як аномальний чи нормальний, залежно від шаблону.

Особливістю нейронних мереж є здатність узагальнювати обмежені, «зашумлені» дані, та дані, що не є повними. Ця особливість надає мережам

потенціал для розпізнавання раніше непомічених шаблонів, тобто таких, для яких не вдалося знайти точний збіг серед попередньо визначених структур шаблонів. Нейронні мережі визнаються як найбільш перспективні методи для виявлення аномалій, оскільки система повинна знаходити не тільки атаки, які вже траплялися, але й нові.

Байесові мережі – це модель, яка кодує ймовірнісні відносини між потрібними змінними. Цей метод зазвичай використовується для виявлення вторгнень у поєднанні із статистичними схемами, процедурою, що має декілька переваг, зокрема, можливості кодування взаємозалежностей між змінними та прогнозуванням подій, а також можливість включати в аналіз і враховувати попередньо зібрану інформацію.

Розглянемо тепер переваги та недоліки кожної категорії (табл. 1).

Таблиця 1

Переваги та недоліки методів Machine Learning

Метод	Переваги	Недоліки
Нечітка логіка	• простіші у використанні, ніж більшість альтернатив; • успішні застосування у промислових сферах; • ефективні проти відомих атак.	• споживання великої кількості ресурсів; • недосконалість застосування до великих наборів правил; • багато «ручної роботи».
Генетичні алгоритми	• можливість виводити класифікаційні правила та обирати оптимальні параметри.	• значний час пошуку збіжностей; • складність впровадження.
Нейронні мережі	• легкість планування; • велика кількість академічних досліджень; • успішне використання в індустрії; • велика кількість готових рішень для застосування.	• наявність простіших, легших, швидших альтернативних методів; • важкість тренування багатошарових мереж; • важкість прогнозування роботи алгоритмів.
Байесові мережі	• можливість вирахування чітких ймовірностей для гіпотез; • можливість використання для статистичного навчання; • можливість поєднання актуальної та попередньої інформації для кращої апроксимації результатів.	• необхідність первинного знання багатьох ймовірностей; • споживання великої кількості ресурсів.

Таким чином, у цій роботі проаналізовано методи Machine Learning з позицій виявлення аномалій. Перспективою подальшої роботи є тестування кожної категорії проаналізованих методів, оцінка адекватності їх функціонування та вибір найбільш оптимального методу для практичного застосування у системах виявлення вторгнень, які аналізують велику кількість трафіку у режимі реального часу.

Науковий керівник – к.т.н., доцент, доцент кафедри БІТ, Гнатюк С.О.