Метод фільтрації пакетів в мережах

Вельчик Д.О., Стопчак О.О.
Науковий керівник – к.т.н., доц. Бойчук В.О.
Хмельницький національний університет

З появою мережі Інтернет відбулося безліч змін в житті людей. Інтернет являє собою сховище різноманітної інформації і сервісів. Часто виникає необхідність обмеження доступу до різноманітних ресурсів мережі. Це може бути пов'язано з нераціональним використанням робочого часу і прагненням роботодавця оптимізувати діяльність своїх співробітників, бажанням батьків обмежити дітей від певного контенту і т.д. Також атаки на мережеві ресурси можуть варіюватися від незначних інцидентів, таких як спам до екстремальних атак, які можуть призводить до фізичного ушкодження апаратури.

Одним з основних інструментів, що використовуються для захисту мережі є брандмауер. Мережевий брандмауер являє собою систему, яка забезпечує виконання політик контролю доступу, як правило, у вигляді правил, щоб контролювати трафік, який входить в мережу або комп'ютер [1]. Пакети, які транспортуються, фільтруються на основі певних характеристик, як правило, їх вихідних IP-адрес, портів джерел і призначень, а також використовуваного протоколу. Правила, встановлені в брандмауері налаштовуються вручну адміністратором мережі і встановлюються на основі відомої інформації по вхідному й вихідному трафіку. Пакети, які вступають в мережу від відомих шкідливих джерел заблоковуються і забороняються для проникнення в мережу. З іншого боку, з'єднання, які є обов'язковими для щоденних функцій в організаціях можуть мати правила, наприклад, пов'язані з перевіркою повідомлень електронної пошти,

Фільтрація зазвичай здійснюється на чотирьох рівнях OSI:

1. Канальному (Ethernet).
2. Мережевому (IP).
3. Транспортному (TCP, UDP).
4. Прикладному (FTP, TELNET, HTTP, SMTP і т. Д.)

Штучний інтелект (AI) і машинне навчання (ML) також стали важливим чинником сучасних інформаційних технологій. Їх застосування поширюється на будь-яку професійну сферу, від мистецтва і засобів масової інформації до медицини. Машинне навчання є полем дослідження в області штучного інтелекту, який дозволяє комп'ютерам навчатися з існуючих прикладів без програмування експертами [2].

Міжмережеві екрани стали обов'язковою частиною будь-якої мережі, через їх здатність фільтрувати трафік на основі правил, які встановлюють які пакети повинні бути прийняті або заборонені. Однак правила фільтрації повинні бути налаштовані вручну адміністратором мережі, а також пакети,

які не відповідають ніяким правилам можуть піддаватися неправильним трактуванням брандмауером. Нейронні мережі можуть навчатися правилам фільтрації для того, щоб вирішити долю пакетів які не відповідають будь-яким конкретним правилам. Нейронна мережа навчається з існуючими пакетними даними і дій брандмауера.

У даній роботі розглядається проблема фільтрації IP-трафіку, яка відноситься до таких задач:

- класифікації - поділ об'єктів на задані групи (класи) відповідно до характеристикам об'єктів;

- регресії - пошук функції, що моделює множину досліджуваних об'єктів з найменшою помилкою;

- кластеризації - пошук незалежних груп об'єктів, кластерів і їх характеристик (групи заздалегідь не визначені);

- пошуку асоціативних правил - характерних залежностей між об'єктами або подіями.

Формально методи пошуку закономірностей можна сформулювати наступним чином.

Завдання класифікації і регресії. Є множина досліджуваних об'єктів $X = \{x_1, x_2,\ldots, x_n.\}$. Кожен об'єкт характеризується набором змінних (атрибутів) $X_j = \{ a_1, a_2,\ldots, a_m, y\}$ де a_i - спостережувані змінні, значення яких відомі; y – залежна змінна, значення якої потрібно визначити. При цьому кожна змінна a_i приймає значення з деякого множини $A_i = \{a_{i1}, a_{i2},\ldots\}$. Спостережувані змінні називаються ознаками або атрибутами. Якщо множина $C = \{c_1, c_2,\ldots, c_k\}$ значень змінної у наперед визначена, то завдання називається завданням класифікації. Якщо змінна у приймає значення на множині дійсних чисел R, то завдання називається завданням регресії.

Завдання кластеризації. Є множина досліджуваних об'єктів $X = \{x_1, x_2,\ldots, x_n\}$. Кожен об'єкт характеризується набором змінних $X_j = \{a_1, a_2,\ldots, a_m\}$. Кожна зміна a_i приймає значення з деякої множини Ai ={ai1, ai2,. . . .}. Завдання кластерізації полягає в побудові множини $C = \{c_1, c_2,\ldots, c_k, \}$, де c_i - кластер, який містить подібні об'єкти з множини X, щодо введеної міри близькості $d (x_j,x_p)$, званої відстанню, т. ч. $C_m = \{ x_j,x_p| x_j \in X,x_p\in X \ \&d(x_j,x_p) <\sigma\}$, де σ - величина, що визначає максимальну відстань, на якому можуть перебувати об'єкти одного кластера.

Завдання пошуку асоціативних правил. Є набір вихідних елементів $I = \{ i_1, i_2,\ldots, i_n\}$, а також набір об'єктів $D = \{ d_1, d_2,\ldots, d_m\}$. Кожен об'єкт є підмножиною множини I ($d_i \subseteq I$). Відповідно до термінології, що відноситься до баз даних, d_i є транзакцією, а D - базою даних. Правило - це імплікація виду $X \Rightarrow Y$, де $X,Y \subseteq D$ і $X \cap Y = Ø$. При цьому для виявлення найбільш правдоподібних правил, що відображають залежності між транзакціями в базі даних, що часто зустрічаються, вводяться дві метрики. Підтримка набору X, що позначається як supp (X), - це пропорція набору X щодо всієї множини D.

Підтримка правила supp ($X \Rightarrow Y$) = supp ($X \cup Y$). Довіра до правила визначається за формулою conf ($X \Rightarrow Y$) = supp ($X \cup Y$) / supp (X). Чим більше значення підтримки і довіри, тим більше точно правило відображає залежності.

З алгоритмічної точки зору класифікація або кластеризація - це функція $f: X \to C$, яка кожному об'єкту $x_i \in X$ ставить у відповідність мітку $cj \in C$. У задачі класифікації множина C визначено заздалегідь, в завданню кластеризації заздалегідь не визначено не тільки множина C, але і його розмірність.

Для реалізації даних методів можна використати, як базові статистичні алгоритми, так і генетичні алгоритми, нейронні мережі, та інші алгоритми з області машинного навчання, які здатні навчатися на основі прецедентів. У загальному вигляді постановку задачі можна представити таким чином. Є множина об'єктів (ситуацій) і множина можливих відповідей (відгуків, реакцій). Існує деяка залежність між відповідями і об'єктами, але вона не відома. Відома тільки кінцева сукупність прецедентів - пар "об'єкт - відповідь", звана навчальної вибіркою. На основі цих даних потрібно відновити залежність, тобто побудувати алгоритм, здатний для будь-якого об'єкта видати досить точну відповідь.

За способами навчання алгоритми класифікуються наступним чином: контрольоване (supervised) навчання - навчання на помічених даних, коли для кожного прецеденту задається відображення вхідні дані-бажане рішення і потрібно вивчити функцію відображення, наприклад з метою подальшої диференціації і класифікації будь-яких вхідних даних; неконтрольоване (unsupervised) навчання - навчання, коли помічені дані не надаються і потрібно згрупувати об'єкти в группи (кластери) на підставі "близькості" об'єктів (щодо деякої міри близькості); частково контрольоване (semi-supervised) навчання - навчання, коли вивчення функції відображення здійснюється на комбінації помічених і непомічених даних.

Для визначення ефективності алгоритму фільтрації використаємо наступні метрики: FP (false positive) - частка трафіку, приписаного до класу X, але не належить до X по відношенню до потужності X; FN (false negative) - частка трафіку, що належить до класу X, але не приписаного до класу X; правильність (accuracy) - частка правильно класифікованих одиниць по відношенню до всіх класифікованих одиницям, т. е. (all-FP-FN) / all. Точність (precision) - пропорція правильно класифікованих одиниць (TP) щодо отриманого класу: TP / (TP + FP); повнота (recall) / довіра (trust) - пропорція правильно класифікованих одиниць щодо реального класу: TP / (TP + FN).

Вихідними даними для дослідження є послідовності IP-пакетів, зібраних в точках спостереження. Одиницею розгляду є потік. Послідовність IP-пакетів може бути двобічної або односпрямованої послідовністю пакетів між двома IP-адресами, повної TCP-сесією або односпрямованої

послідовністю IP-пакетів, яка визначається на основі п'яти полів заголовка:

<Src_ip, src_port, dst_ip, dst_port, protocol>,

і правил формування, за якими визначається завершення потоку (зазвичай тайм-аут або прапор "END" в заголовку пакета). Тут src_ip - IP-адреса джерела; src_port - порт джерела; dst_ip - IP-адреса призначення; dst_port - порт призначення, protocol - транспортний протокол. Як правило, в якості протоколів транспортного рівня використовуються TCP і UDP.

При проектуванні нейронної мережі в якості фільтра пакетів треба прийняти до уваги особливості даних для обробки. Фіксується набір змінних (атрибутів), заснованих на статистичних характеристиках, такі як розмір пакетів або інтервали між пакетами, і характеристиках зчитаних з заголовків пакетів, таких як розмір TCP-сегментів або кількість повторних передач. Потоку ставиться у відповідність набір значень атрибутів, згідно з якими проводиться класифікація.

Кожному атрибуту пакета надаємо вагу, яку потім використовуємо в функції гіпотези, як показано нижче :

$$h_\theta = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots \theta_n x_n$$

У цій формулі x значення атрибута n, а θ - вага атрибуту. Гіпотеза буде використовуватися при обчисленні функції вартості, яка є мірою помилки в здатності нейронної мережі, щоб оцінити співвідношення між вхідними значеннями і відомими вихідними значеннями

Для використання великих векторизованних наборів даних в системі фільтрації вартісна функція буде мати вигляд:

$$J(\theta) = (1/m)\,(-y^T \log h - (1 - y) \log (1 - h))$$

y^T в даному випадку є транспонований вектор, що містить вихідні значення навчальних прикладів.

З огляду на характер даних, що подаються для фільтрації в даній роботі, в якості нейронної мережі вибрана нейрона мережа прямого поширення.

Для реалізації НМ була використана Fast Artificial Neural Network (FANN) бібліотека з відкритим вихідним кодом, яка дозволяє користувачам створювати багатошарові НМ. На її основі згідно рекомендаціям була створена нейрона мережа з п'яти входів і одного з одним прихованим шаром з чотирма нейронів зі значенням швидкості навчання рівного 0,5.

Для визначення продуктивності НМ в якості системи фільтрації визначаємо відсоток правильно класифікованих пакетів та кількість правил, які були створені.

Інформаційна мережа, яка використовується для тестування буде складатися з діапазону адрес IP від 1 до 7 і діапазону портів від 2 до 263. Це означає, що в цілому 5,788,104 можливих комбінацій для зразків, використаних в експерименті. Етап тестування буде повторювати тричі зі збільшенням числа генерованих правил: 4, 9, і 15 з варіантами роботи за

замовчуванням: прийняти пакети або відхилити пакети.

Приклади пакетів для перевірки були отримані випадковим чином, використовуючи дані параметрів мережі.

З огляду на результати тестів, проведених з різними параметрами було встановлено, що НМ дійсно може бути корисним в справі поліпшення можливостей фільтрації. Результати показують, що НМ здатна класифікувати мережеві пакети майже так само як брандмауери з великим числом правил. Це може бути використано для розширення можливостей фільтрації брандмауера, особливо за рахунок поліпшення правил по замовчуванню, з неправильною фільтрації пакетів, які не входять в правила брандмауера.

Ці поліпшення не тільки в кінцевому рахунку не тільки розширяють можливості застосування фільтрації пакетів брандмауерів, але можуть також включати в себе механізми безпеки, які запобігають атакам.

Перелік посилань

1. M. Jordan and T. Mitchell, "Machine learning: Trends, perspectives, and prospects", Science, vol. 349, no. 6245, pp. 255-260, 2015.

2. K. Valentin and M. Maly, “Network firewall using artificial neural networks,” Computing and Informatics, vol. 32, pp. 1312–1327, 2013.