Обнаружение аномалий методом моделирования запросов и машинного обучения

Лю Хайлинь

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: hereishailin@outlook.com

Научный руководитель *— Лапонина Ольга Робертовна*

Атаки с использованием SQL-инъекций позволяют злоумышленникам выполнять незаконные SQL-запросы через параметры, передаваемые веб-приложениям, чтобы использовать уязвимости в системе баз данных для подделки или кражи конфиденциальных данных в базе данных.

По сравнению с традиционными методами преимущество использования методов машинного обучения (ML) для предотвращения инъекций заключается в том, что эти алгоритмы могут использовать большой объем данных и меток для автоматического изучения и извлечения характеристик внедрения SQL, а также интеллектуальной классификации SQL-запросов.

В процессе ML необходимы данные. Мы выбрали набор данных SAJID576[1].

Мы сравнили метод предварительной обработки SQLiGoT[2] с традиционным методом векторизации и примененили их для обнаружения аномалий с помощью методов ML. В процессе реализации мы выбрали 6 различных алгоритмов ML.

Была выполнена токенизация и нормализация запросов, любой SQL-запрос, независимо от его длины и сложности, нормализуется в последовательность токенов, сохраняющих его синтаксическую структуру.

Процесс нормализации преобразует запрос в упорядоченную последовательность токенов, $(t_1, t_2, ..., t_N)$. Мы генерировали граф $G = (V, E, w)$ из последовательности, которая отражает его структурные свойства как сеть взаимодействия токенов, что облегчает анализ графа с использованием количественных показателей. Граф имеет n ребер, $n = |V|$ и $n \leqslant N$, соответствующих уникальному токену t_i. Вес ребра отражает степень взаимосвязи между двумя токенами.

Расстояние между двумя токенами в скользящем окне можно определить как промежуток между токенами. Промежуток g между двумя токенами - это количество токенов, присутствующих между

ними в последовательности. Говорят, что два токена встречаются одновременно, если $g \leqslant s-2$, где s — размер скользящего окна.

Граф токенов, состоящий из n узлов$(t_1, t_2, ..., t_N)$, представлен матрицей смежности A размера $n \times n$.

Для взаимодействия между токенами мы рассматриваем примеры неориентированного и ориентированного графов нормальных и аномальных запросов.

Используя как неориентированные, так и ориентированные графы, один и тот же запрос можно обучить по мере центральности, входящей мере центральности и выходящей мере центральности, которые объединяются вместе для создания классификатора тройной модели, как показано на рис. 1.

Во время выполнения выходные данные трех классификаторов будут голосовать за окончательный результат прогнозирования. Если результат внедрения положительный, это означает, что по крайней мере две из трех моделей идентифицируют его как внедренный запрос, поэтому запрос следует отклонить. Окончательный прогноз делается методом большинства голосов. Все системы были протестированы 10 раз с заданными обучающими и тестовыми наборами, чтобы исключить нечастые случаи в процессе обучения. Усилили 4 моделя из 6, и среди них частота ошибок на модель Linear SVC уменьшалась на 18.3%.

Иллюстрации

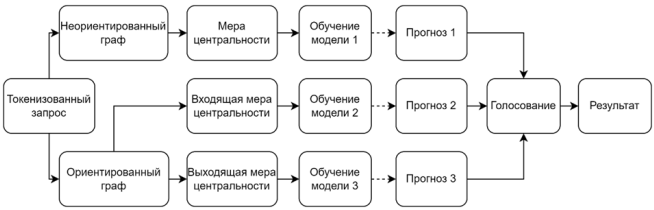

Рис. 1. Структура голосования моделей

Литература

1. SQL Injection Dataset, SAJID576, Version 5.29, URL: www.kaggle.com/datasets/sajid576/sql-injection-dataset/data/
2. Kar, D.; Panigrahi, S.; Sundararajan, S. SQLiGoT: Detecting SQL injection attacks using the graph of tokens and SVM. Comput. Secur. 2016, 60, 206–225.