XÂY DỰNG HỆ THỐNG PHÁT HIỆN XÂM NHẬP DỰA TRÊN HỌC SÂU KHẢ DIỄN GIẢI SỬ DỤNG GIÁ TRỊ SHAPLEY

Huỳnh Thái Thi - 230202032

Tóm tắt

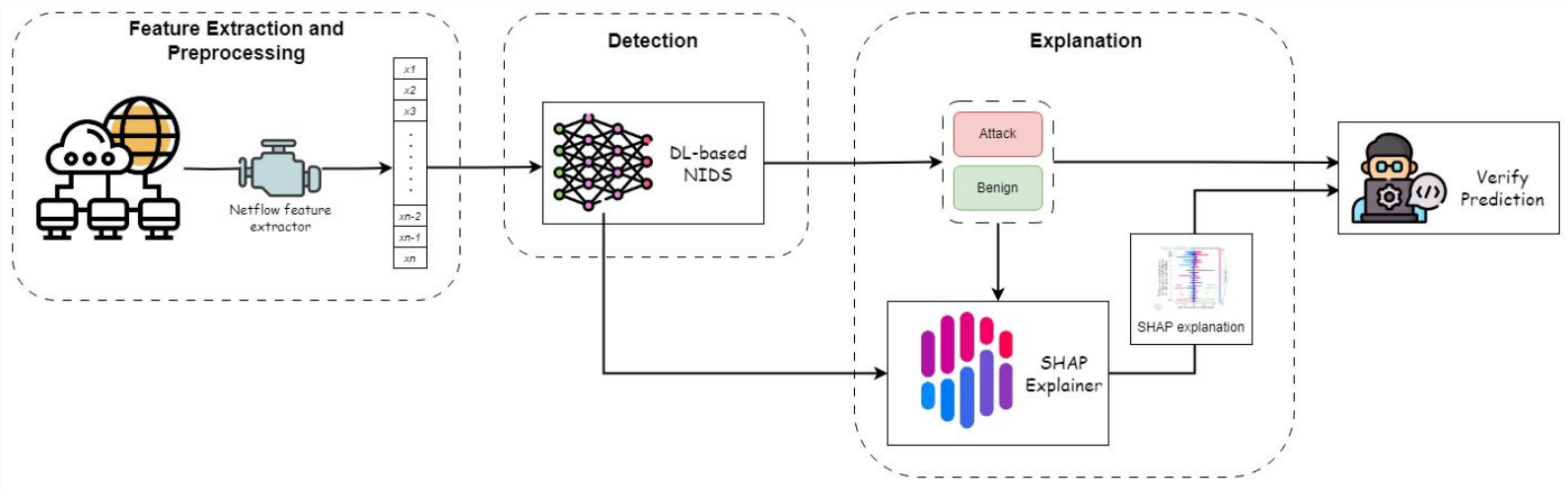
- Lớp: CS2205.MAR2024
- Link Github: https://github.com/ltn0tbug/CS2205.MAR2024/
- Link YouTube video: https://youtu.be/Pw5horT4sdE
- Họ và Tên: Huỳnh Thái Thi

Giới thiệu

- Hệ **thống phát hiện xâm nhập mạng** dựa trên **học sâu** đang là xu hướng phát triển.
- Các **mạng nơ-ron** được xem là “hộp đen” bởi tính **khó diễn giải** của chúng.
- Các **thuật toán học máy khả giải** (**Giá trị Shapley**, LIME, etc.) nhiều tiềm năng nhưng còn hạn chế trong việc ứng dụng.

Giới thiệu

=> Nghiên cứu một **hệ thống phát hiện xâm nhập mạng** dựa trên **học sâu** và được tăng cường khả năng diễn giải dựa trên **giá trị Shapley** thông qua **bộ khung SHAP** [2].



Hình 1: Mô hình đề xuất

Mục tiêu

- Xây dựng các mô hình phát hiện xâm nhập dựa trên các mạng học sâu như CNN, LSTM, GRU, … sử dụng bộ dataset NF-ToN-IoT [4].
- Diễn giải được dự đoán của mô hình phát hiện xâm nhập sau áp dụng bộ khung SHAP theo cách mà các nhà phân tích có thể hiểu được.
- So sánh được ưu và nhược điểm của diễn giải từ bộ khung SHAP với các thuật toán học máy khả giải khác như LIME [1], Anchors [3], ….

Nội dung và Phương pháp

Nội dung 1: Thu thập và tiền xử lý dữ liệu

- Sử dụng bộ dữ liệu NF-ToN-IoTđã được trích xuất thuộc tính dựa trên giao thức NetFlow.
- Tiền xử lý dữ liệu:
  - Chuẩn hóa dữ liệu
  - Xử lý giá trị thiếu
  - Xử lý ngoại lệ: loại bỏ hoặc điều chỉnh các giá trị bất thường.
  - Chia dữ liệu thành tập huấn luyện, tập kiểm tra và tập đánh giá.

Nội dung và Phương pháp

Nội dung 2: Xây dựng mô hình học sâu

- Xây dựng mô hình sử dụng các mạng nơ-ron (CNN, LSTM, GRU, ...)
- Huấn luyện, đánh giá mô hình trên tập dữ liệu đã tiền xử lý.

Nội dung 3: Áp dụng bộ khung SHAP để diễn giải mô hình

- Sử dụng bộ khung SHAP để tính toán giá trị Shapley cho mỗi thuộc tính đầu vào.
- Biểu diễn giá trị Shapley bằng các phương pháp trực quan hóa

Nội dung và Phương pháp

Nội dung 4: So sánh độ chính xác của SHAP với các thuật toán học máy khả giải khác.

- Áp dụng các thuật toán học máy khả giải khác như LIME, Anchors vào mô hình phát hiện xâm nhập.
- So sánh độ chính xác của các thuật toán học máy khả giải trong việc giải thích dự đoán của mô hình.
- Đánh giá ưu và nhược điểm của từng thuật toán để lựa chọn phương pháp phù hợp nhất cho bài toán cụ thể.

Kết quả dự kiến

- Xây dựng được một hệ thống phát hiện xâm nhập mạng dựa trên học sâu và được tăng cường khả năng diễn giải bằng giá trị Shapley.
- Khả năng diễn giải của hệ thống giúp các nhà phân tích bảo mật có thể dễ dàng hiểu được cơ sở cho dự đoán của mô hình.
- So sánh được ưu và nhược điểm của các thuật toán học máy khả giải dựa trên giá trị Shapley so với một số thuật toán khác.

Tài liệu tham khảo

[1]. Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD 2016: 1135-1144

[2]. Scott M. Lundberg, Su-In Lee: A Unified Approach to Interpreting Model Predictions. NIPS 2017: 4765-4774

[3]. Marco Túlio Ribeiro, Sameer Singh, Carlos Guestrin: Anchors: High-Precision Model-Agnostic Explanations. AAAI 2018: 1527-1535

[4]. Mohanad Sarhan, Siamak Layeghy, Nour Moustafa, Marius Portmann: NetFlow Datasets for Machine Learning-based Network Intrusion Detection Systems. CoRR abs/2011.09144 (2020)