HỆ THỐNG PHÁT HIỆN MÃ ĐỘC CÓ KHẢ NĂNG DIỄN GIẢI DỰA TRÊN HỌC TỔNG HỢP

Tô Trọng Nghĩa - 220202019

Tóm tắt

Tô Trọng Nghĩa

220202019

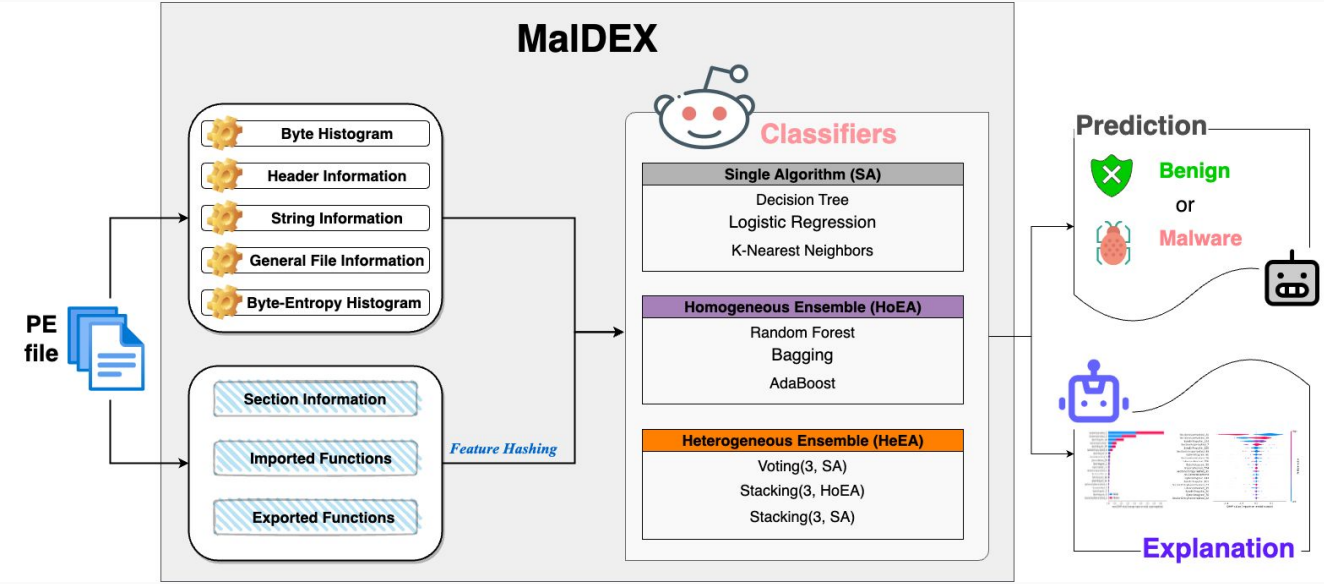
TTNghiaUIT/220202019_CS2205_Final_Project

TTNghiaUIT/220202019_CS2205_video

Giới thiệu

- Đảm bảo an toàn hệ thống luôn là yêu cầu cần thiết và cấp thiết
- Gặp nhiều hạn chế trong việc giải quyết các vấn đề liên quan đến tính tin cậy của các kết quả dự đoán trả về.
- Các mô hình diễn giải thiếu tính thực tế khi sử dụng các bộ dữ liệu đã được xử lý sẵn, ít đặc trưng và chỉ phù hợp với một số mô hình nhất định

Giới thiệu



Hình 1. Tổng quan hệ thống phát hiện mã độc MalDEX

Mục tiêu

- Xây dựng mô-đun để trích xuất các thuộc tính từ mẫu chương trình thực thi định dạng PE trên hệ điều hành Windows.

- Xây dựng các trình phân loại mã độc dựa trên các đặc trưng đã được trích xuất trên (phân loại các mẫu lành tính và độc hại) sử dụng 3 thuật toán chính bao gồm các thuật toán đơn, các thuật toán tổng hợp đồng nhất và các thuật toán tổng hợp không đồng nhất.

- Áp dụng phương thức diễn giải Kernel SHAP cho tất cả các bộ phân loại nhằm mục đích tính toán tầm quan trọng của từng đặc trưng đầu vào (được trích xuất ở trên) đối với kết quả dự đoán (lành tính hay độc hại).

Nội dung và Phương pháp

Nội dung 1:

Xây dựng mô-đun trích xuất các đặc trưng cho hệ thống MalDEX

Phương pháp:

- Tìm hiểu cấu trúc và đặc điểm của chương trình thực thi định dạng PE trên hệ điều hành Windows.
- Tìm hiểu và nghiên cứu các phương pháp trích xuất các đặc trưng của tệp PE để sử dụng cho các trình phân loại.
- Nghiên cứu và thiết kế mô-đun trích xuất các đặc trưng cho hệ thống MalDEX.

Nội dung và Phương pháp

Nội dung 2:

Xây dựng các trình phân loại mã độc cho hệ thống MalDEX

Phương pháp:

- Tìm hiểu và nghiên cứu các thuật toán, phương pháp sử dụng thuật toán học tổng cho các trình phân loại mã độc.
- Nghiên cứu và thiết kế các trình phân loại mã độc dựa trên các đặc trưng đã được trích xuất ở nội dung 1.
- Đánh giá hiệu năng của các trình phân loại trong việc phát hiện các chương trình độc hại. So sánh, đánh giá hiệu năng của các trình phân loại sử dụng thuật toán tổng hợp với các trình phân loại sử dụng thuật toán đơn.

Nội dung và Phương pháp

Nội dung 3:

Áp dụng phương pháp diễn giải cho các trình phân loại

Phương pháp:

- Tìm hiểu, nghiên cứu các phương pháp tăng khả năng giải thích của các mô hình học máy, đặc biệt là Kernel SHAP.
- Nghiên cứu và xây dựng các trình phân loại trong MalDEX sử dụng Kernel SHAP để giải thích các kết quả trả về.
- Đánh giá và nhận xét các kết quả giải thích của các bộ phân loại.

Kết quả dự kiến

- Hệ thống MalDEX có khả năng đưa ra các loại phân loại đúng với các đầu vào tương ứng cùng với hình ảnh biểu thị tầm quan trọng của các đặc trưng cho quyết định đó.

- Bản báo cáo trình bày đánh giá hiệu năng của các thuật toán tổng hợp trong việc phân loại các chương trình thực thi định dạng PE.

- Bài báo khoa học về hệ thống phát hiện mã độc MalDEX có khả năng diễn giải dựa trên học tổng hợp.

Tài liệu tham khảo

- [1]. Gueltoum Bendiab, Stavros Shiaeles, Abdulrahman Alruban, Nicholas Kolokotronis: IoT Malware Network Traffic Classification using Visual Representation and Deep Learning. NetSoft 2020: 444-449
- [2]. Jan Stiborek, Tomás Pevný, Martin Rehák: Multiple instance learning for malware classification. Expert Syst. Appl. 93: 346-357 (2018)
- [3]. Irina Baptista, Stavros Shiaeles, Nicholas Kolokotronis: A Novel Malware Detection System Based on Machine Learning and Binary Visualization. ICC Workshops 2019: 1-6
- [4]. Francesco Paolo Caforio, Giuseppina Andresini, Gennaro Vessio, Annalisa Appice, Donato Malerba: Leveraging Grad-CAM to Improve the Accuracy of Network Intrusion Detection Systems. DS 2021: 385-400
- [5]. Basim Mahbooba, Mohan Timilsina, Radhya Sahal, Martin Serrano: Explainable Artificial Intelligence (XAI) to Enhance Trust Management in Intrusion Detection Systems Using Decision Tree Model. Complex. 2021: 6634811:1-6634811:11 (2021)
- [6]. Rafa Alenezi, Simone A. Ludwig: Explainability of Cybersecurity Threats Data Using SHAP. SSCI 2021: 1-10