**Model Evasion Attack on Intrusion Detection Systems using Adversarial Machine Learning**

Overview

Hai loại IDS là (1) IDS dựa trên chữ ký để phân tích phạm vi mạng dành cho các chữ ký độc hại đã biết và (2) IDS dựa trên bất thường so sánh hệ thống mạng với các mẫu đã biết của người dùng và đưa ra cảnh báo nếu nó đi chệch khỏi mẫu .

Trong nghiên cứu của chúng tôi, chúng tôi tập trung vào IDS dựa trên mạng, còn được gọi là NIDS và Mạng nơ-ron Articial (ANN) là thuật toán học máy của chúng tôi. Trọng tâm của chúng tôi trong nghiên cứu này là chứng minh The model evasion attack for IDS, theo đó kẻ thù có thể né tránh mô hình ML cho IDS dựa trên mạng được xây dựng trong cài đặt hộp trắng bằng cách tạo các mẫu đối thủ.

The model evasion attack có khả năng làm giảm đáng kể độ chính xác của IDS, tức là, phát hiện ra phương thức độc hại là lành tính.

Công việc được trình bày trong bài báo này gồm hai phần: (1) chúng tôi phát triển phương pháp tiếp cận ML để phát hiện xâm nhập bằng cách sử dụng mạng Multilayer Perceptron (MLP) (một cấu trúc liên kết Mạng Nơron phổ biến, để thực hiện phân loại nhị phân đối với dữ liệu lành tính và tấn công trong 1 network-based anomaly IDS) và chứng minh tính hiệu quả của mô hình của chúng tôi với hai bộ dữ liệu IDS dựa trên mạng khác nhau; và (2) chúng tôi thực hiện một cuộc tấn công né tránh mô hình chống lại mạng MLP được xây dựng cho IDS bằng cách sử dụng kỹ thuật học máy đối phương được gọi là phương pháp tấn công bản đồ Saliency dựa trên Jacobian (JSMA - Jacobian-based Saliency Map Attack method).

Target Misclassiﬁcation: Kẻ tấn công làm cho lớp đầu ra phân loại của một mẫu đối thủ cụ thể thành một lớp mục tiêu cụ thể.

* Hộp trắng. Đối thủ có kiến ​​thức hoàn hảo về mô hình phân loại mục tiêu bao gồm loại phân loại được sử dụng và cấu trúc của nó(type of the classiﬁer used and its structure.), tham số cũng như tất cả hoặc một phần của tập dữ liệu huấn luyện và các tính năng của nó.
* Hộp đen. Đối thủ không có kiến ​​thức về mô hình mục tiêu, có thể biết rằng mô hình thực hiện phân loại, nhưng đối thủ không có quyền truy cập vào dữ liệu huấn luyện, cấu trúc hoặc kiểu mô hình hoặc bất kỳ thông số nào của mô hình, có thể bù đắp sự thiếu thông tin này bằng cách truy vấn mô hình về khả năng rò rỉ thông tin.
* Hộp màu xám. Đối thủ có kiến ​​thức chưa đầy đủ về mô hình mục tiêu và biết các tính năng được xem xét bởi mô hình và kiểu của nó. Anh ấy / anh ấy không có bất kỳ phần nào của bộ đào tạo hoặc trọng số trong mô hình.

Cuộc tấn công của chúng tôi phụ thuộc vào kiến ​​thức về các tham số nhất định được sử dụng trong mô hình được đào tạo nhưng không phụ thuộc vào tập dữ liệu đào tạo.

Dataset

* Dataset CICIDS 2017 (<https://www.unb.ca/cic/datasets/ids-2017.html>)

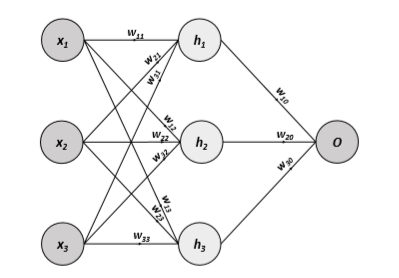
Hành vi trừu tượng của 25 người dùng dựa trên các giao thức HTTP, HTTPS, FTP, SSH và email. Với 12 máy nạn nhân và 2 máy tấn công khác nhau, tập dữ liệu được gắn nhãn này có các tính năng tấn công phổ biến, chẳng hạn như Brute Force dựa trên web, XSS và SQL Injection, DoS, DDoS, In ﬁ ltration, Heart-bleed, Bot và Scan. Chúng tôi đã lấy mẫu tổng cộng 950.000 bản ghi. Mỗi bản ghi bao gồm 80 tính năng liên tục với một lớp có nhãn nhị phân là lành tính hoặc tấn công.

* Dataset TRAbID 2017 (<https://secplab.ppgia.pucpr.br/?q=trabid>)

Bao gồm giao tiếp máy khách-máy chủ. Truy cập hợp pháp được tạo ra bởi máy khách yêu cầu các dịch vụ có sẵn trong máy chủ, chẳng hạn như HTTP, SMTP, SSH, SNMP và DNS, trong khi kẻ tấn công từ máy khách khởi động các cuộc tấn công đến cùng một máy chủ. Loại tấn công chính bao gồm các danh mục DoS khác nhau (ví dụ: SYN ﬂ ood, ICMP ﬂ ood, v.v.) và Scan (ví dụ: quét SYN, quét ACK, v.v.). Đối với thử nghiệm của chúng tôi, chúng tôi đã thu thập tổng cộng 18.000 bản ghi. Không giống như CICIDS 2017, mỗi bản ghi trong trường hợp này bao gồm 43 tính năng liên tục. Tập dữ liệu cũng được gắn nhãn với một lớp nhị phân là lành tính hoặc tấn công.

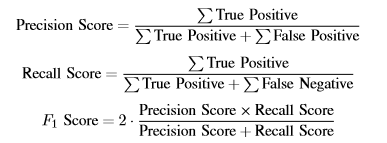
The ML Model for IDS (Multilayer Perceptron (MLP) Network)

Với w là vectơ thực của trọng số, b ∈ R là độ lệch và h là hàm truyền, hàm quyết định của MLP có thể chính thức được định nghĩa là: f (x) = w · h (vi · x + di) + b trong đó, (vi, di) ∈ Rn × R là biểu diễn trọng số của ẩn đơn vị thứ i. Một mạng MLP thường được xây dựng với ba hoặc nhiều lớp, nghĩa là, một lớp đầu vào, một hoặc nhiều lớp ẩn và một lớp đầu ra. Chúng tôi chọn một lớp ẩn và xây dựng một mạng được kết nối đầy đủ (tức là mỗi nút trong lớp đầu vào được kết nối với một trọng số nhất định với mọi nút trong lớp ẩn)



Để kiểm soát khả năng khái quát của một nhận thức, chúng tôi kết hợp sử dụng việc dừng sớm trong quá trình đào tạo. Sau đó, chúng tôi theo dõi việc mất xác thực để kích hoạt hành động này. Chúng tôi cũng chọn 10% hồ sơ đào tạo làm bộ xác nhận để thực hiện nhiệm vụ này. Chúng tôi phân tích dựa trên cả tập dữ liệu CICIDS 2017 và TRAbID 2017 và điều chỉnh mô hình của chúng tôi tương tự như những gì chúng tôi đã thảo luận trước đây. Trước tiên, chúng tôi mở rộng quy mô tập dữ liệu để tất cả các giá trị tính năng nằm trong khoảng từ 0 đến 1, còn được gọi là MinMax Scalar. Sau đó, chúng tôi chia dữ liệu trong các trường hợp đào tạo 80% và thử nghiệm 20%. Do đó, chúng tôi có khoảng 760.000 và 14.400 hồ sơ đào tạo từ tập dữ liệu CICIDS và TRAbID tương ứng. Ngoài ra, chúng tôi thực hiện chia nhỏ các phiên bản huấn luyện và thử nghiệm để duy trì tỷ lệ phần trăm giống nhau cho mỗi lớp mục tiêu như trong bộ hoàn chỉnh được cung cấp trong tập dữ liệu.

Experimental Results



Precision Score: tỉ lệ các trường hợp được gán nhãn Positive là đúng so với tổng các trường hợp được dự đoán nhãn Positive.

Recall: tỉ lệ các trường hợp được gán nhãn Positive là đúng so với tổng các trường hợp thực sự là Positive.

F1 Score: là trung bình hài hòa của độ chính xác và độ thu hồi(Precision Score & Recall).

Mô hình hoạt động với độ chính xác 99,5% đối với tập dữ liệu CICIDS 2017, 99.8% đối với tập dữ liệu TRAbID 2017.

Chúng tôi đào tạo cả hai mô hình cho 100 epochs với batch size là 64, nhận thấy rằng mạng MLP có xu hướng tăng quá mức sau 13 epochs đối với CICIDS 2017 và 16 epochs đối với TRAbID 2017. Tuy nhiên, chúng tôi bảo toàn tính tổng quát của mô hình bằng cách kích hoạt tác vụ dừng sớm dựa trên mất xác thực(validation loss).

Adversarial Machine Learning to Evade ID

Cuộc tấn công dựa trên giả định rằng kiến ​​thức duy nhất mà đối thủ có là về các tham số mà mô hình MLP sử dụng để dự đoán lớp nhị phân.

Adding a perturbation to the legitimate sample X (i.e., X∗ ← X + δX, where δ is denoted as perturbation), such that F(X∗)=Y ∗ <> Y.

Tạo adversarial test samples based on Jacobian-based Saliency Map Attack (JSMA) [26].

