TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

KHOA CÔNG NGHỆ THÔNG TIN

LÊ ĐỨC DŨNG

NGHIÊN CỨU BỘ DỮ LIỆU NSL-KDD VÀ ỨNG DỤNG PHƯƠNG PHÁP HỌC MÁY TRONG PHÁT HIỆN VÀ PHÂN LOẠI MỘT SỐ HÌNH THỨC TẤN CÔNG MẠNG

ĐỒ ÁN TỐT NGHIỆP ĐẠI HỌC

NGÀNH CÔNG NGHỆ THÔNG TIN

Thái Nguyên – 2024

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN

TỐT NGHIỆP ĐẠI HỌC

NGÀNH CÔNG NGHỆ THÔNG TIN

Đề tài:

NGHIÊN CỨU BỘ DỮ LIỆU NSL-KDD VÀ ỨNG DỤNG PHƯƠNG PHÁP HỌC MÁY TRONG PHÁT HIỆN VÀ PHÂN LOẠI MỘT SỐ HÌNH THỨC TẤN CÔNG MẠNG

Sinh viên thực hiện: Lê Đức Dũng

Lớp/Khóa: CNTTK18L

Giáo viên hướng dẫn

(Ký và ghi họ tên)

Thái Nguyên – 2024

LỜI CAM ĐOAN

Em xin cam đoan rằng đồ án tốt nghiệp với đề tài "Nghiên cứu bộ dữ liệu NSL - KDD và ứng dụng phương pháp học máy trong phát hiện và phân loại một số hình thức tấn công mạng" là nghiên cứu của em. Kết quả của đồ án là sự tổng hợp tri thức từ nhiều nguồn tài liệu khác nhau và có trích dẫn đầy đủ, ghi rõ nguồn gốc.

Em xin chịu hoàn toàn trách nhiệm trước nhà trường nếu trường hợp phát hiện ra bất cứ sai phạm hay vấn đề sao chép nào trong đề tài này.

TP.Thái Nguyên , ngày ..tháng ..năm 2024

(SV ký và ghi rõ họ tên)

Lê Đức Dũng

LỜI CẢM ƠN

Đồ án này hoàn thành tại Trường Đại học Công nghệ thông tin và Truyền thông - Đại học Thái Nguyên. Em xin gửi lời cảm ơn đến các thầy cô đã dành tâm huyết của mình truyền đạt vốn kiến thức quý báu cho em trong suốt quá trình học tập. Em xin gửi lời cảm ơn sâu sắc đến TS. Nguyễn Đức Bình đã trực tiếp hướng dẫn em hoàn thành đồ án này với sự nhiệt tình và ân cần chỉ bảo, đồng thời cung cấp cho em những kiến thức chuyên môn để em có thể hoàn thiện đồ án tốt nghiệp này.

Cuối cùng, em xin gửi lời cảm ơn chân thành đến gia đình, bạn bè và người thân, những người đã bên cạnh và động viên em trong suốt quá trình học tập và hoàn thành đồ án. Mặc dù em đã nỗ lực cố gắng nhưng trong quá trình làm đồ án sẽ khó tránh khỏi những thiếu sót. Rất mong nhận được sự góp ý quý báu của quý thầy cô và các bạn sinh viên để đồ án được hoàn thiện hơn.

Thái Nguyên, ngày tháng năm 2024

(Sinh viên thực hiện)

Lê Đức Dũng

MỤC LỤC

DANH MỤC HÌNH ẢNH

MỞ ĐẦU

An ninh mạng ngày càng trở thành mục tiêu đối tượng của các mối đe dọa và tấn công an ninh mạng vì nó được sử dụng thường xuyên hơn trên mọi lĩnh vực của nền kinh tế. Sắp xếp dữ liệu mạng thành các danh mục thường xuyên hoặc nghi ngờ phát hiện tấn công là một bước quan trọng bước vào việc cố gắng ngăn chặn những cuộc tấn công như vậy ta thấy được việc phát hiện sự bất thường đề cập đến nhiệm vụ bảo vệ an toàn, liên quan đến những vấn đề khó xảy ra tấn công trong giao tiếp mạng. Các cuộc tấn công sẽ được phát hiện chính xác là bất thường và được xử lý bằng nghi ngờ nếu có sự sai lệch đáng kể so với tiêu chuẩn. Điểm số sẽ phản ánh thế nào mới sự xuất hiện khác với bình thường.

Trong Machine Learning thường tìm kiếm một giải pháp tổng quát không gian cho mô hình tốt nhất phù hợp với dữ liệu. Mạng nơ-ron (ANN) đề cập đến không gian của tất cả các hàm gần đúng hoặc chính xác mà mạng nơ-ron (ANN) có thể đại diện.

Các mạng hiện đại có nguy cơ gặp phải nhiều mối đe dọa khác nhau do lưu lượng truy cập dựa trên internet tăng trưởng mạnh mẽ. Bằng cách tiêu tốn thời gian và tài nguyên, lưu lượng xâm nhập cản trở hoạt động hiệu quả của cơ sở hạ tầng mạng. Một chiến lược hiệu quả để ngăn chặn, phát hiện và giảm thiểu các sự cố xâm nhập sẽ làm tăng năng suất. Một yếu tố quan trọng của lưu lượng mạng an toàn là Hệ thống phát hiện xâm nhập (IDS). Một hệ thống IDS có thể dựa trên máy chủ hoặc dựa trên mạng để giám sát hoạt động mạng xâm nhập.

Việc phát hiện lưu lượng truy cập internet bất thường đã trở thành một vấn đề nghiêm trọng rủi ro của các bảo mật cho các thiết bị thông minh. Các hệ thống này bị ảnh hưởng tiêu cực bởi một số cuộc tấn công, làm chậm tính toán. Ngoài ra, các vi phạm và bất thường trong giao tiếp nối mạng phải được phát hiện bằng học máy Machine Learning. Do đó mục đích em sử dụng bộ dữ liệu NSL-KDD để đề xuất IDS mới dựa trên Mạng nơ-ron (ANN). Kết quả là mô hình Machine Learning có khả năng khái quát hóa đủ để hoạt động tốt trên dữ liệu chưa được thử nghiệm.

Xử lý trước dữ liệu của bộ dữ liệu được theo sau bởi nhập dữ liệu. Việc thu thập dữ liệu NSL-KDD bao gồm thuộc tính khác. Trong số 41 thuộc tính đó, một thuộc tính không đóng vai trò gì cả, còn thuộc tính đóng vai trò phát hiện các cuộc tấn công. Các hoạt động thu nhỏ và chuẩn hóa tính năng bắt đầu. Kiến trúc mạng nơ-ron

(ANN) được chọn đầu tiên. Sau đó, mạng nơ-ron (ANN) được huấn luyện bằng dữ liệu huấn luyện phù hợp. Sau đó, tập dữ liệu thử nghiệm. Một số các phát hiện sẽ được sử dụng để đánh giá hiệu quả hoạt động để đưa ra các số liệu điển hình bao gồm phân loại, phát hiện tỷ lệ, độ chính xác và tỷ lệ.

Ngoài phần mở đầu và tài liệu kham khảo , nội dung chính của đồ án em chia thành 4 chương . Chương 1 em trình bày về tổng quan lịch sử ra đời của bộ dữ liệu NSL-KDD. Chương 2 em trình bày phân tích về bộ dữ liệu NSL-KDD các thuộc tính của bộ dữ liệu NSL-KDD. Chương 3 em trình bày về học máy và phương pháp học máy bộ dữ liệu NSL-KDD. Chương 4 em phân loại phân tích các hình thức tấn công mạng của bộ dữ liệu NSL-KDD.

Bộ dữ liệu NSL-KDD sẽ được sử dụng cho cả việc huấn luyện và kiểm tra. Trong bài báo báo đồ án này, em trình bày và phân tích các thuộc tính tấn công mạng của các tập dữ liệu trong bộ dữ liệu NSL-KDD bằng cách sử dụng phương pháp mô hình học máy.

CHƯƠNG I. TỔNG QUAN VỀ NSL- KDD

1.1 Tổng quan.

Hệ thống thông tin liên lạc đóng một vai trò tất yếu trong cuộc sống hàng ngày của con người bình thường. Mạng máy tính được sử dụng hiệu quả để xử lý dữ liệu kinh doanh, giáo dục và học tập, cộng tác, thu thập dữ liệu trên diện rộng và giải trí. Các ngăn xếp giao thức mạng máy tính đang được sử dụng ngày nay được phát triển với mục đích tạo ra nó minh bạch và thân thiện với người dùng. Điều này dẫn đến phát triển một chồng giao thức truyền thông mạnh mẽ. Tính linh hoạt của giao thức đã khiến nó dễ bị tấn công các cuộc tấn công được phát động bởi những kẻ xâm nhập. Điều này làm cho yêu cầu mạng máy tính phải hoạt động liên tục được giám sát và bảo vệ. Quá trình giám sát là được tự động hóa bởi hệ thống phát hiện xâm nhập (IDS). Các IDS có thể được tạo thành từ sự kết hợp giữa phần cứng và phần mềm.

Tại bất kỳ thời điểm nào, một máy chủ web có thể được nhiều người truy cập khách hàng và họ tự nhiên tạo ra lưu lượng truy cập lớn. Mỗi kết nối mạng có thể được hiển thị dưới dạng một tập hợp các thuộc tính. Dữ liệu có thể được ghi lại và sử dụng để nghiên cứu và phân loại bình thường và bất thường. Để xử lý cơ sở dữ liệu khổng lồ, học máy kỹ thuật có thể được sử dụng. Khai thác dữ liệu là quá trình trích xuất dữ liệu quan tâm từ các tập dữ liệu khổng lồ bằng cách sử dụng máy học kỹ thuật.

Và thấy được trong thời đại hiện nay, mọi người đều được kết nối qua mạng Internet để chia sẻ thông tin số dữ liệu. Dữ liệu số thông tin được lưu trữ bằng công nghệ đám mây. Tuy nhiên, công nghệ đám mây đang nhanh chóng gia tăng khối lượng thông tin số và sự xâm nhập tấn công mạng. Trong trường hợp này, việc bảo vệ dữ liệu đám mây là điều cần thiết vì một số mục đích. Do đó, các nghiên cứu hiện tại nhấn mạnh việc phát triển hệ thống phát hiện xâm nhập mạng bằng cách sử dụng một bộ dữ liệu NSL-KDD chuẩn. Thuật toán ngẫu nhiên hỗ trợ học tập tổng hợp đã được đề xuất và triển khai để chọn ra các tính năng phù hợp nhất với việc nghiên cứu. Việc phát hiện và phân loại xâm nhập mạng đã được thực hiện bằng ba mô hình học máy: học máy vectơ hỗ trợ (SVM), hồi quy logistic và K-nearest neighbour’s (KNN) với độ chính xác xác thực là 87,58%, 88,86% và 98,24%. Do đó, phương pháp nghiên cứu bộ dữ liệu NSL-KDD có thể được sử dụng để phát hiện và giám sát tấn công mạng theo thời gian thực.

Trong này việc phân tích bộ dữ liệu NSL-KDD là được thực hiện bằng cách sử dụng các thuật toán phân cụm khác nhau có sẵn trong công cụ khai phá dữ liệu WEKA . Bộ dữ liệu NSL-KDD là được phân tích và phân loại thành bốn cụm khác nhau mô tả bốn loại tấn công phổ biến khác nhau. Một nghiên cứu phân tích chuyên sâu được thực hiện trong bài kiểm tra và đào tạo tập dữ liệu. Tốc độ thực hiện của các phân cụm khác nhau thuật toán được phân tích. Đây là tập dữ liệu kiểm tra (kddtrain.arff) và tập dữ liệu huấn luyện 20% (kddtest.arff) Được sử dụng.Sử dụng bộ dữ liệu NSL-KDD để khám phá các giao thức dễ bị tấn công nhất thường được sử dụng bởi những kẻ tấn công xâm nhập để khởi động các cuộc tấn công và xâm nhập dựa trên mạng.

1.2 Lịch sử ra đời NSL – KDD

Phương pháp để các hệ thống hoặc mạng có thể tránh bị malware hoặc lưu lượng mạng xấu từ Internet tấn công là triển khai các hệ thống ở các vị trí nhằm bảo vệ các thông tin quan trọng trong các máy tính hoặc hệ thống mạng. Những hệ thống phát hiện các lưu lượng mạng độc hại như vậy được gọi là các hệ thống phát hiện xâm nhập (IDS) và được huấn luyện với các dữ liệu lưu lượng mạng Internet. Tập dữ liệu phổ biến nhất là NSL-KDD, cũng là bộ dữ liệu chuẩn cho dữ liệu Internet hiện nay.

Bộ dữ liệu NSL-KDD được tạo ra bởi nhóm nghiên cứu của Giáo sư Tavallaee tại Đại học New Brunswick, Canada. Bộ dữ liệu này được tạo ra bằng cách sử dụng tập dữ liệu KDD Cup 1999, nhưng đã được loại bỏ một số bản ghi bị trùng lặp và lỗi.

Tập dữ liệu KDD Cup 1999 được tạo ra bởi Cơ quan An ninh Quốc gia Hoa Kỳ (NSA). Tập dữ liệu này chứa khoảng 5 triệu bản ghi mô tả các phiên giao dịch mạng. Các phiên giao dịch mạng này có thể là một cuộc tấn công hoặc một hoạt động hợp pháp.

Nhóm nghiên cứu của Giáo sư Tavallaee đã thực hiện một số thay đổi đối với tập dữ liệu KDD Cup 1999 để tạo ra bộ dữ liệu NSL-KDD. Những thay đổi này bao gồm:

- Loại bỏ các bản ghi bị trùng lặp.
- Loại bỏ các bản ghi lỗi.
- Sửa đổi các thuộc tính để cải thiện độ chính xác.