

Phát hiện và phân loại Botnet IoT bằng cách sử dụng thuật toán học máy

Phạm Văn Quân¹ , Ngô Văn Úc¹ , Đỗ Phúc Hào^{2,3} , Nguyễn Năng Hùng Văn⁴

¹ Đại học Đồng Á, Đà Nẵng, Việt Nam
² Đại học Viễn thông Bang Bonch-Bruevich Saint-Petersburg, Saint-Petersburg, Liên bang Nga
³ Đại học Kiến trúc Đà Nẵng, Đà Nẵng, Việt Nam
⁴ Đại học Bách khoa Đà Nẵng, Đà Nẵng, Việt Nam

Tóm tắt: Bài báo nghiên cứu học thuật này đề cập đến thách thức quan trọng và phức tạp trong việc phát hiện và phân loại Internet of Things (IoT) botnet thông qua việc sử dụng các thuật toán học máy. Nghiên cứu tập trung vào việc tiến hành phân tích tỉ mỉ và thao tác dữ liệu mạng botnet IoT, với một sự nhấn mạnh cụ thể được đặt vào bộ dữ liệu IoT 23 được thừa nhận rộng rãi. Mục đích chính là tuyển dụng những người được công nhận rộng rãi và các thuật toán học máy được sử dụng rộng rãi, bao gồm Cây quyết định (DT), k-Hàng xóm gần nhất (KNN), Ngẫu nhiên Forests (RF) và eXtreme Gradient Boosting (XGBoost), với mục đích phân loại và phát hiện botnet một cách hiệu quả trong giới hạn của bộ dữ liệu IoT-23. Bằng cách thực hiện các thuật toán này, bài báo tìm cách nâng cao hiểu biết của chúng ta về hiệu suất và hiệu quả của chúng trong miền phát hiện và phân loại mạng botnet IoT. Việc thực hiện một phân tích so sánh, đối chiếu các kết quả thu được từ các thuật toán đa dạng, sẽ cung cấp những hiểu biết vô giá thành những ưu điểm và hạn chế tương ứng của chúng, do đó cho phép các nhà nghiên cứu và các học viên để đưa ra quyết định sáng suốt liên quan đến thuật toán phù hợp nhất để đạt được thành công Phát hiện và phân loại botnet IoT.

Thuật ngữ chỉ mục: Internet vạn vật, Học có giám sát, Hướng dẫn phát hiện sion, IoT Botnet, An ninh mạng.

I. GIỚI THIỆU

Khái niệm về Internet of Things (IoT) ban đầu được đề xuất bởi K.Ashton (1999) như một giải pháp cho sự gia tăng số lượng thiết bị yêu cầu kết nối internet. Các Cơ quan Liên minh Châu Âu về Mạng và Thông tin Bảo mật định nghĩa IoT là một hệ sinh thái vật lý không gian mạng bao gồm các cảm biến và bộ truyền động được kết nối với nhau để tạo điều kiện thuận lợi cho quá trình tạo ra quyết định. Đáng chú ý, IoT là một công nghệ quan trọng trong Công nghiệp 4.0 [1]. IoT công nghiệp (IIoT) đại diện cho một triển khai chuyên biệt IoT kết nối các động cơ và các thành phần công nghiệp để nâng cao năng suất và hiệu quả hoạt động công nghiệp. IIoT đạt được mục tiêu này bằng cách cung cấp giám sát thời gian thực, quản lý hiệu quả, và kiểm soát các quy trình công nghiệp, tài sản và hoạt động

thời gian. Hơn nữa, IIoT nhằm mục đích giảm chi phí hoạt động trong khi nâng cao hiệu quả tổng thể [2].

Tuy nhiên, việc phát triển các hệ thống IoT mang lại ý nghĩa quan trọng những thách thức về an ninh. Các thiết bị IoT thường có bảo mật hạn chế và quản lý tài nguyên. Bên cạnh việc phát triển các hệ thống IoT, cũng cần có các biện pháp bảo mật hiệu quả để bảo vệ chống lại các mối đe dọa mạng và vi phạm dữ liệu [3].

Ngăn chặn các cuộc tấn công và xâm nhập khác nhau và bảo vệ dữ liệu là rất cần thiết. Hiện nay, có rất nhiều xâm nhập hệ thống phát hiện (IDS) đã được phát triển để đáp ứng đến vấn đề này. IDS có nhiệm vụ phát hiện và báo cáo các hệ thống xâm nhập. Ngoài ra, IDS ngăn chặn mã độc tấn công và duy trì hiệu suất trong thời gian tấn công. IDS là một thành phần quan trọng của hệ thống an ninh hiện đại[4].

Mạng botnet IoT là một mạng gồm các thiết bị IoT bị xâm nhập được kiểm soát bởi các tác nhân độc hại cho các hoạt động độc hại. Các botnet này khai thác lỗ hổng trong các thiết bị IoT để có được quyền truy cập trái phép và tạo ra một kết nối lớn mạng của các thiết bị bị xâm nhập. Các tính năng chính của IoT botnet bao gồm đa dạng thiết bị, khai thác lỗ hổng, cơ sở hạ tầng chỉ huy và kiểm soát, tấn công DDoS, khả năng spam và lừa đảo, khai thác tiền điện tử và đánh cắp dữ liệu và vi phạm quyền riêng tư. Để giải quyết các mối đe dọa đặt ra bởi các botnet IoT, điều quan trọng là tăng cường bảo mật thiết bị IoT, triển khai các cơ chế xác thực mạnh, thường xuyên cập nhật phần mềm, phân đoạn mạng và giáo dục người dùng về thực hành bảo mật tốt nhất.

Động lực đằng sau việc tạo ra bài báo học thuật tập trung vào về việc xác định và phân loại Internet of Things (IoT) botnet thông qua các kỹ thuật học máy phát sinh từ sự cần thiết cấp thiết phải giải quyết cảnh quan nguy hiểm ngày càng leo thang do các mạng ác độc này gây ra. Các không thể phủ nhận sự phổ biến của các thiết bị IoT được kết nối với nhau khuếch đại khả năng tấn công botnet, do đó gây nguy hiểm cho sức khỏe của cả người dùng cá nhân và cơ sở hạ tầng thiết yếu. Bằng cách khai thác sức mạnh của máy móc phương pháp học tập, chúng ta có thể nâng cao trình độ của mình trong

nhận biết và phân loại các mạng botnet IoT, từ đó tạo điều kiện thuận lợi cho việc thực hiện các biện pháp chủ động để ngăn chặn và giảm bớt hậu quả có hại của chúng. Việc sử dụng các thuật toán máy học cho phép chúng tôi xem xét kỹ lưỡng khối lượng dữ liệu lưu lượng truy cập mạng đáng kể, tạo điều kiện thuận lợi cho việc xác định các mẫu thông thường và các điểm bất thường, đồng thời phân biệt hành vi thông thường của thiết bị IoT với các hoạt động của mạng botnet. Bằng cách đắm mình trong lĩnh vực tìm hiểu này, chúng ta có thể đóng góp vào việc xây dựng các cơ chế phòng thủ mạnh mẽ và hiệu quả, từ đó củng cố tính bảo mật của hệ sinh thái IoT.

Mục tiêu chính của tài liệu nghiên cứu học thuật này là giải quyết thách thức then chốt trong việc phát hiện và phân loại các botnet IoT thông qua việc sử dụng các thuật toán máy học. Nghiên cứu nhấn mạnh đáng kể vào việc phân tích và thao tác tỉ mỉ dữ liệu mạng botnet IoT, với trọng tâm cụ thể là bộ dữ liệu IoT-23 nổi tiếng.

Mục đích chính là triển khai các thuật toán học máy được sử dụng rộng rãi và được công nhận rộng rãi, cụ thể là Cây quyết định (DT), k-Hàng xóm gần nhất (KNN), Rừng ngẫu nhiên (RF) và Tăng cường độ dốc eXtreme (XGBoost), với mục đích phân loại hiệu quả và phát hiện các botnet trong giới hạn của bộ dữ liệu IoT-23. Việc áp dụng các thuật toán này nhằm mục đích nâng cao hiểu biết của chúng ta về hiệu suất và hiệu quả của chúng khi áp dụng cho nhiệm vụ phát hiện và phân loại mạng botnet IoT. Việc thực hiện phân tích so sánh, kết hợp các kết quả thu được từ các thuật toán đa dạng, sẽ mang lại những hiểu biết vô giá về các điểm mạnh và hạn chế tương ứng của chúng, cho phép các nhà nghiên cứu và các học viên đưa ra quyết định sáng suốt liên quan đến thuật toán phù hợp nhất để phát hiện và phân loại thành công mạng botnet IoT.

II. CÔNG TRÌNH LIÊN QUAN

Trong những năm gần đây, nghiên cứu về phát hiện xâm nhập IoT đã được chú ý nhiều hơn. Nghiên cứu lại về chủ đề này đang trở nên cần thiết hơn. Dựa trên các kết quả tìm kiếm trước đó, có nhiều giải pháp để phát triển hệ thống phát hiện xâm nhập IoT. Việc hiểu rõ những điểm mạnh cũng như những vấn đề của các nghiên cứu trước giúp người nghiên cứu tìm ra hướng nghiên cứu mang lại lợi ích lớn cho việc xây dựng hệ thống IDS. Do đó, các nghiên cứu trước đây là nền tảng trong việc cung cấp kiến thức cho sự phát triển trong tương lai.

Sự phổ biến của các thiết bị IoT đã mở ra một mối đe dọa leo thang do các mạng botnet IoT gây ra, mang đến cho bọn tội phạm mạng những con đường mới để dàn dựng các cuộc tấn công quy mô lớn. Nhiệm vụ cấp thiết là phát hiện và phân loại các mạng botnet này có tầm quan trọng đặc biệt trong việc hạn chế rủi ro đang phát triển. Tận dụng khả năng của các kỹ thuật học máy đã nổi lên như một chiến lược hiệu quả để

tăng cường các biện pháp an ninh mạng bằng cách tạo điều kiện thuận lợi cho việc xác định các mẫu đặc biệt, điểm bất thường và hành vi ác ý đã ăn sâu vào lưu lượng mạng mẽ cung của các botnet IoT. Điều này nhấn mạnh tính tất yếu của việc triển khai các thuật toán máy học để giải quyết những thách thức phức tạp vốn có trong mạng botnet IoT.

J. Hajji et al. [5] đã tiến hành một nghiên cứu để áp dụng các thuật toán học máy không giám sát, bao gồm K-means, PCA và Autoencoder, để phát hiện lưu lượng mạng bất thường. Một. Rahim và cộng sự. [6] đã sử dụng phân tích thống kê để xác định các tính năng quan trọng nhất và sau đó áp dụng một số thuật toán học máy, chẳng hạn như DT, RF và Naive Bayes (NB), để phân loại lưu lượng truy cập thông thường và độc hại.

SMZ Hoi giáo et al. [7] đã nghĩ ra các mô hình tính trung bình và xếp chồng dựa trên các thuật toán Support Vector Machine (SVM), RF và Gradient Boosting. Y. Li và cộng sự. [8] đã phát triển một mô hình đóng gói từ bốn mô hình học máy dựa trên DT, KNN, Hồi quy logistic (LR) và RF để phân loại lưu lượng mạng là bình thường hoặc độc hại.

PH Do [9] đã đề xuất một phương pháp trích xuất đặc trưng bằng cách chia các tập đặc trưng thành các lớp khác nhau, tiếp theo là ứng dụng thuật toán học máy cho các tập đặc trưng ở các lớp chồng nập. Những nghiên cứu này đã sử dụng một loạt các thuật toán, bao gồm một số kết hợp, để đạt được độ chính xác cao trong các nhiệm vụ phân loại của chúng.

Một số nhà nghiên cứu đã khám phá tiềm năng của các mô hình học sâu để phát hiện hành vi bất thường trong lưu lượng mạng. Alotaibi et al. [10] đã phát triển mô hình dựa trên Mạng thần kinh chuyển đổi (CNN) để phân loại lưu lượng mạng là bình thường hoặc độc hại. Lý và cộng sự. [11] đã sử dụng Bộ nhớ dài hạn ngắn hạn (LSTM) và Mạng nơ-ron dày đặc để phân loại lưu lượng mạng là bình thường hoặc độc hại. Tương tự, Abdallah et al. [12] đã sử dụng mô hình học sâu dựa trên CNN và LSTM.

Kiani et al. [13] đã đề xuất Mạng thần kinh Deep Autoencoder để tìm hiểu hành vi bình thường của mạng IoT

lưu lượng truy cập và sử dụng nó để phát hiện hành vi bất thường trong thực tế thời gian. Ngoài ra, Rasool et al. [14] đã thử nghiệm với các mô hình học chuyển đổi như VGG16, ResNet50 và InceptionV3. Những nghiên cứu này cho thấy tiềm năng của các mô hình học sâu trong việc phát hiện sự bất thường trong lưu lượng mạng.

Bộ dữ liệu IoT-23 [15], được phát hành vào năm 2020, đã trở thành tâm điểm cho các nhà nghiên cứu về bảo mật IoT và máy học. Nó đã được sử dụng rộng rãi để phát triển và đánh giá các kỹ thuật phát hiện và giảm thiểu lây nhiễm phần mềm độc hại IoT. Một nghiên cứu đáng chú ý đã kết hợp các thuật toán máy học với các phương pháp học sâu để cải thiện độ chính xác của việc phát hiện phần mềm độc hại bằng bộ dữ liệu. Một dự án khác đã sử dụng bộ dữ liệu để tạo ra một phát hiện xâm nhập phù hợp

hệ thống cho các thiết bị IoT, tập trung vào việc phát hiện các hoạt động bất thường và độc hại theo thời gian thực. Ngoài ra, lại

những người tìm kiếm đã sử dụng bộ dữ liệu IoT-23 để đánh giá nhiều phương pháp trích xuất đặc trưng và thuật toán phân loại cho Phát hiện phần mềm độc hại IoT. Tính khả dụng của bộ dữ liệu này đã đóng góp đáng kể vào những tiến bộ trong bảo mật IoT nghiên cứu và phát triển các thuật toán hiệu quả và các hệ thống để bảo vệ các thiết bị IoT khỏi các hoạt động độc hại. Việc khám phá bộ dữ liệu IoT-23 đang diễn ra tiếp tục tăng cường bảo mật và khả năng phục hồi của IoT đang mở rộng hệ sinh thái.

Mỗi nghiên cứu này trình bày các phương pháp khác nhau để phát hiện và phân loại các cuộc tấn công mạng và tất cả đã cho thấy kết quả tốt. Tuy nhiên, một số nghiên cứu chưa đề cập đến Hiệu suất của mô hình cũng như việc thực hiện thời gian của mô hình, một vài người khác chỉ trình bày phát hiện hoặc phân loại.

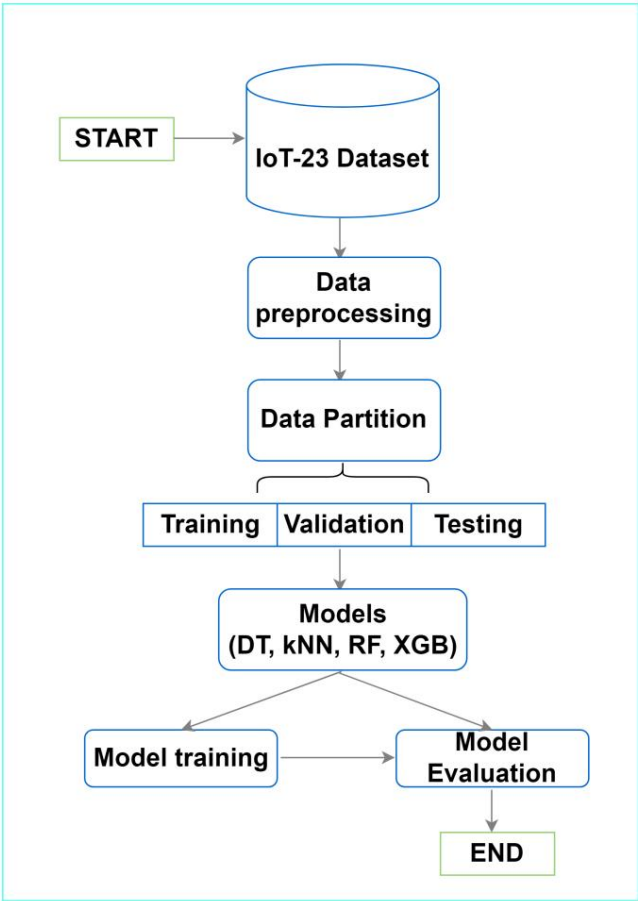
III. PHƯƠNG PHÁP

Phần này sẽ bắt đầu với phần trình bày về bộ dữ liệu sẽ được sử dụng trong phân tích tiếp theo. Các kỹ thuật tiền xử lý sau đó sẽ được áp dụng để đảm bảo độ chính xác và độ tin cậy của dữ liệu. Sau đó, chúng tôi sẽ đi sâu vào quá trình lựa chọn dữ liệu, trực quan hóa, để làm mờ và chia tách. Phân tích của chúng tôi sẽ lên đến đỉnh điểm với một đánh giá hiệu quả của các thuật toán được sử dụng, kèm theo một phân tích so sánh các kết quả. Luồng thực thi chính của nghiên cứu này được minh họa trong Hình 1.

Để đánh giá hiệu quả của các mô hình học máy để phân loại IoT-Botnet, trước tiên chúng tôi thực hiện xử lý trước dữ liệu. Tiền xử lý dữ liệu giúp cải thiện việc phân loại hiệu quả của quy trình. Tiếp theo, chúng tôi chia tập dữ liệu thành tập huấn luyện, kiểm tra và xác nhận tương ứng. Cuối cùng, chúng tôi triển khai các mô hình máy học phổ biến như cây quyết định, k-láng giềng gần nhất (kNN), rừng ngẫu nhiên và XGBoost (XGB) để đánh giá hiệu quả của từng mô hình.

1. Tập dữ liệu

Bộ dữ liệu IoT-23, được phát hành vào tháng 1 năm 2020, là một trọng tâm đáng kể của việc tìm kiếm lại trong các lĩnh vực bảo mật IoT và học máy. Các nhà nghiên cứu đã sử dụng rộng rãi bộ dữ liệu này để phát triển và đánh giá các kỹ thuật khác nhau cho việc phát hiện và giảm thiểu lây nhiễm phần mềm độc hại IoT. Bộ dữ liệu bao gồm các lưu lượng truy cập mạng được ghi lại từ một mảng đa dạng gồm 23 Internet of Things (IoT) riêng biệt thiết bị, thiết bị mở rộng như phích cắm thông minh, máy ảnh, khóa thông minh và bộ điều nhiệt thông minh, trong số những thứ khác. Những cái này lưu lượng truy cập được chụp một cách tỉ mỉ trong một môi trường được kiểm soát, trong đó mỗi thiết bị được kết nối với một mạng Wi-Fi tách biệt và tiếp xúc với một phạm vi



Hình 1: Mô hình đề xuất

kịch bản tấn công, bao gồm các cuộc tấn công vũ phu, Mirai tấn công botnet, tấn công tiêm nhiễm, v.v.

Bảng 1. Nội dung lưu lượng của bộ dữ liệu IoT-23 theo các cuộc tấn công được thực hiện.

Tên tấn công	Chạy
Part-Of-A-Horizontal-PortScan	213, 852, 924
Okiru	47, 381, 241
Okiru-Attack	13, 609, 479
DDoS	19, 538, 713
C&C-Heart Beat	33,673
C&C	21,995
Tấn công	9398
C&C-	888
C&C-Heart Beat Attack	883
C&C-Tải xuống tệp	53
C&C-Torii	30
Tập tin tải về	18
Tải xuống tệp C&C-Heart Beat	11
Tấn công Part-Of-A-Horizontal-PortScan 5	
C&C-Mirai	2

Siêu dữ liệu chi tiết liên quan đến mỗi lần chụp được tích hợp trong tập dữ liệu, bao gồm thông tin như

loại thiết bị, phiên bản phần sụn và loại tấn công. Một mô tả toàn diện về phân phối lưu lượng truy cập được cung cấp trong Bảng 1, trong khi một danh sách phức tạp, trình bày mô tả của tất cả các tính năng, được trình bày trong Bảng 2.

Bảng 2. Các tính năng của bộ dữ liệu IoT-23.

Tính năng	Tên	Mô tả
trường-ts		Dòng thời gian bắt đầu
uid		ID duy nhất
id.orig-h		Nguồn Địa chỉ IP
id.orig-p		Cổng nguồn
id.resp-h		Địa chỉ IP đích
id.resp-p		cổng đích
dịch		giao thức
vụ proto		Loại dịch vụ (http, dns, v.v.)
khoảng thời gian		Tổng thời lượng lưu lượng
orig-bytes		Byte giao dịch nguồn-đích
resp-bytes		Byte giao dịch nguồn đích
Liên kết trạng thái		trạng thái kết nối
local-orig		Nguồn địa chỉ địa phương
local-resp		Địa chỉ đích
resp-pkts		gói đích
orig-ip-bytes		Dòng byte nguồn
lịch sử		Lịch sử của các gói nguồn
byte bị thiếu		Thiếu byte trong quá trình giao dịch
ban-pkts		gói tin nguồn
resp-ip-bytes		Luồng byte đích
nhãn		Tên kiểu tấn công

Để hỗ trợ phát hiện lưu lượng truy cập độc hại, phòng thí nghiệm hình cầu Strato đã phát triển các nhãn cho các loại khác nhau của các luồng mạng, dựa trên phân tích của chúng về phần mềm độc hại chụp. Các nhãn được sử dụng để phát hiện luồng độc hại là Tấn công, C&C, DDoS, FileDownload, HeartBeat, Mirai, Okiru, PartOfAHorizontalPortScan và Torii. Cuộc tấn công" nhãn đã được gán cho các luồng cố gắng khai thác các dịch vụ để bị tổn thương, chẳng hạn như brute-force đăng nhập telnet hoặc đưa một lệnh vào tiêu đề của yêu cầu GET. Các Nhãn "lành tính" chỉ ra rằng không có độc hại hoặc đáng ngờ các hoạt động đã được phát hiện. Nhãn "C&C" chỉ ra rằng thiết bị bị nhiễm đã được kết nối với máy chủ CC, trong khi nhãn "DDoS" chỉ ra rằng thiết bị bị nhiễm tham gia vào một cuộc tấn công từ chối dịch vụ phân tán. Các Nhãn "Tái xuống tệp" đã được gán cho các kết nối liên quan đến việc tải tệp xuống thiết bị bị nhiễm.

Nhãn "Heart-Beat" được gán cho các kết nối đã được sử dụng để theo dõi máy chủ bị nhiễm bởi máy chủ C&C. Các Nhãn "Mirai" được gán cho các kết nối thể hiện đặc điểm của một cuộc tấn công botnet Mirai, trong khi "Okiru" nhãn đã được sử dụng cho các kết nối có mẫu tương tự với mạng botnet Okiru ít phổ biến hơn. Phần " PartOfAHorizontal Nhãn PortScan" đã được gán cho các kết nối được sử dụng cho quét cổng ngang để thu thập thông tin để biết thêm

các cuộc tấn công. Cuối cùng, nhãn "Torii" chỉ ra các kết nối mà thể hiện các đặc điểm của một cuộc tấn công botnet Torii.

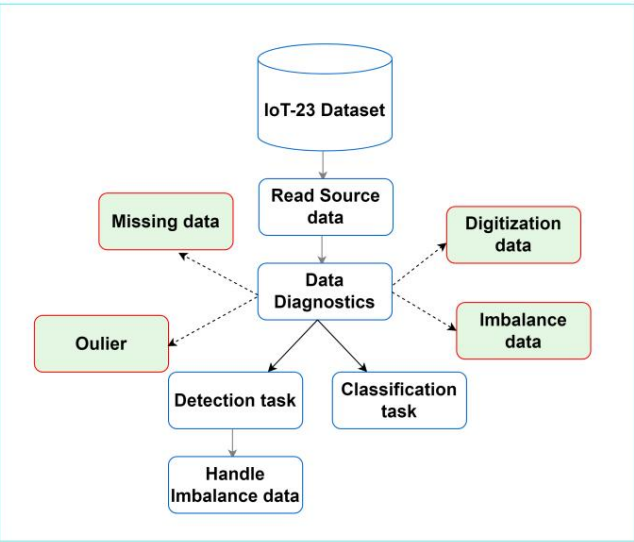
Tính khả dụng của bộ dữ liệu IoT-23, có nhãn chụp phần mềm độc hại và lưu lượng truy cập thiết bị IoT thực, đã đóng một vai trò then chốt trong việc thúc đẩy nghiên cứu về bảo mật IoT. Nó có phục vụ như một nguồn tài nguyên quý giá để phát triển và đánh giá các thuật toán học máy, phát hiện xâm nhập hệ thống và các giải pháp bảo mật khác nhằm mục đích bảo vệ thiết bị IoT khỏi các hoạt động độc hại. Các nhà nghiên cứu tiếp tục để khám phá và mở rộng dựa trên những hiểu biết được cung cấp bởi Bộ dữ liệu IoT-23 để tăng cường hơn nữa tính bảo mật và khả năng phục hồi của hệ sinh thái IoT đang mở rộng nhanh chóng.

2. Tiền xử lý

Giai đoạn tiền xử lý dữ liệu đóng một vai trò quan trọng trong chuẩn bị dữ liệu cho đào tạo mô hình, bao gồm nhiều các bước và cân nhắc. Quá trình này, như được mô tả trong Hình 2, là điều cần thiết để đảm bảo chất lượng dữ liệu và giải quyết những thách thức cụ thể có thể phát sinh trong quá trình phân tích.

Bước đầu tiên trong tiền xử lý dữ liệu liên quan đến việc đọc nguồn dữ liệu, cho phép truy cập vào tập dữ liệu để biết thêm kiểm tra và thao tác. Sau đó, trọng tâm chuyển sang đánh giá chất lượng dữ liệu thông qua toàn diện kiểm tra và chẩn đoán. Bước quan trọng này nhằm xác định và giải quyết các vấn đề tiềm ẩn, chẳng hạn như thiếu dữ liệu, ngoại lệ, hoặc nhu cầu chuyển đổi số.

Để giải quyết những thách thức này, các giải pháp phù hợp được đưa ra cho từng vấn đề cụ thể gặp phải, đồng thời đảm bảo tác động tối thiểu đến tập dữ liệu tổng thể. Cách tiếp cận này cho phép can thiệp có mục tiêu và bảo vệ tính toàn vẹn của dữ liệu.



Hình 2: Các bước tiền xử lý dữ liệu

Tạp chí Công nghệ thông tin & Truyền thông

Hơn nữa, điều quan trọng là phải thừa nhận rằng sự khác biệt nhiệm vụ yêu cầu kỹ thuật xử lý dữ liệu phù hợp. Đặc biệt, mất cân bằng dữ liệu nổi lên như một mối quan tâm nổi bật do sự khác biệt đáng kể về số lượng hồ sơ trên các lớp khác nhau. Để giảm thiểu vấn đề này, các chiến lược khác biệt được thông qua dựa trên nhiệm vụ hiện tại.

Bảng 3. Loại dữ liệu của các tính năng IoT-23		
STT	Tên tính năng	Kiểu dữ liệu
1	trường-ts	Trôi nổi
2	uid	Sự vật
3	id.orig-h	Sự vật
4	id.orig-p	Trôi nổi
5	id.resp-h	Sự vật
6	id.resp-p	Trôi nổi
7	dịch	Sự vật
8	vụ proto	Sự vật
9	khoảng thời gian	Sự vật
10	byte gốc	11 Sự vật
12	byte tương ứng	12 Sự vật
13	trạng thái kết nối	Sự vật
14	local-orig	Sự vật
15	local-orig	Sự vật
16	resp-pkts	16 orig- Trôi nổi
17	ip-byte	17 byte lịch Trôi nổi
18	sử bị mất	19 Sự vật
19	orig-pkts	20 Trôi nổi
20	resp-ip-byte	21 Trôi nổi
21	nhãn	Trôi nổi
22		Sự vật

Sau khi kiểm tra và đánh giá tỉ mỉ bộ dữ liệu IoT 23, người ta xác định rằng bộ dữ liệu này không có của bất kỳ điểm dữ liệu bị thiếu. Tuy nhiên, một sự hiện diện nhỏ của điểm dữ liệu ngoại lệ đã được quan sát. Hơn nữa, việc phân tích mang lại một cái nhìn tổng quan về tập dữ liệu, tiết lộ thông tin như tổng số 3.000.779 bản ghi, bao gồm 20 tính năng riêng biệt. Bảng 3 cung cấp một đại diện toàn diện của các loại dữ liệu liên quan với mỗi tính năng.

Để xử lý các tính năng với các loại dữ liệu không phải là số, một quá trình phân loại đã được tiến hành, kết quả là chia các đặc điểm này thành hai nhóm riêng biệt. Các nhóm đầu tiên bao gồm các loại dữ liệu không phải là số không thể dễ dàng chuyển đổi thành định dạng số. Cái này nhóm bao gồm các tính năng như uid, id.orig h, id.resp h, — nguồn gốc địa phương, và resp địa phương. Ngược lại, nhóm thứ hai bao gồm các loại dữ liệu không phải là số có thể thuận tiện được chuyển thành dạng số. Nhóm này bao gồm các tính năng như proto, dịch vụ, trạng thái kết nối, lịch sử, thời lượng, byte gốc và byte tương ứng. Để tạo thuận lợi cho quá trình chuyển đổi này, mỗi giá trị riêng biệt trong các tính năng nói trên là được gán một giá trị nguyên duy nhất, bắt đầu từ 0. The Labe

lớp lEncoder tử thư viện sklearn đã được sử dụng cho Mục đích này. Sau đó, với việc xử lý kiểu dữ liệu hoàn thành trên tập dữ liệu gốc, thử nghiệm đã được thực hiện trên 15 tính năng được chọn như một phần của phần tiếp theo Phân tích.

Bảng 4. Nhân chi tiết cho nhiệm vụ đa phân loại	
Tên nhãn	Số nhãn
Đang ở	0
C&C-Heart Beat	1
C&C	1
C&C-	1
C&C-Heart Beat Attack	1
C&C-Tải xuống tệp	1
C&C-Torii	1
Tập tin tải về	1
Tải xuống tệp C&C-Heart Beat 1	
C&C-Mirai	1
Tấn công	2
DDoS	3
Okiru	4
Okiru-Attack	4
Part-Of-A-Horizontal-PortScan	5

Để căn chỉnh loại dữ liệu cột nhãn với yêu cầu của nghiên cứu của chúng tôi, nó là bắt buộc để thực hiện một quá trình phân nhánh. Bộ phận này phục vụ một mục đích quan trọng, đặc biệt là trong bối cảnh các nhiệm vụ phát hiện của chúng tôi. Trong khoảng này khuôn khổ, hồ sơ mang nhãn lạnh tính sẽ được quy cho giá trị 0, trong khi tất cả các bản ghi khác sẽ được gán giá trị của 1. Trong lĩnh vực nhiệm vụ phân loại, các lớp thể hiện bản ghi hạn chế và chia sẻ các loại tấn công giống hệt nhau sẽ được hợp nhất thành các thực thể lớp cố kết. sự phức tạp chi tiết của lược đồ ghi nhãn này được phác thảo tỉ mỉ trong Bảng 4, cung cấp một cái nhìn tổng quan toàn diện về nhiệm vụ phân loại và phát hiện.

Trong bối cảnh môi trường mạng, sự xuất hiện các sự kiện diễn ra với tốc độ đáng kinh ngạc, đòi hỏi hoạt động dữ liệu nhanh chóng, liên tục và tự động. Việc thu thập dữ liệu được sắp xếp liền mạch bởi các hệ thống IoT, encom truyền các thông số cần thiết và sau đó được phân phối theo kênh vào các đường ống xử lý tự động và các mô hình được đào tạo để tạo kết quả thời gian thực. Mục tiêu cơ bản của những các hành động năng động xoay quanh việc phát hiện kịp thời và ngăn chặn sự bất thường của mạng, đảm bảo ứng phó kịp thời và can thiệp.

Trong ngữ cảnh của các nhiệm vụ phát hiện, nơi phân biệt giữa các lớp độc hại và lành tính là rất quan trọng, việc lấy mẫu kỹ thuật được sử dụng để giải quyết sự mất cân đối đáng kể. Điều này liên quan đến việc điều chỉnh cẩn thận sự phân phối của nhóm độc hại để đảm bảo đại diện tối ưu.

Đối với các nhiệm vụ phân loại, trong đó nhiều lớp với một số lượng bản ghi hạn chế có liên quan, một ứng dụng hợp nhất

phương pháp được thực hiện để giảm sự mất cân bằng dữ liệu. Sự hợp nhất này tạo điều kiện cho các biểu diễn cân bằng hơn của các lớp khác nhau, cải thiện quá trình phân loại.

3. Phương pháp phân tích

a) Cây quyết định

Trong lĩnh vực phân loại mạng botnet IoT, cây quyết định đại diện cho một kỹ thuật học máy cơ bản và hiệu quả. Những cây này cung cấp một khuôn khổ minh bạch và toàn diện để ra quyết định, tận dụng các thuộc tính riêng biệt được tìm thấy trong dữ liệu lưu lượng truy cập mạng IoT.

Việc sử dụng rộng rãi của họ trong lĩnh vực an ninh mạng thể hiện năng lực của họ trong việc phát hiện và phân loại các botnet IoT.

Về cốt lõi, nguyên tắc làm nền tảng cho các cây quyết định xoay quanh việc phân chia lặp đi lặp lại các tập dữ liệu, được hướng dẫn bởi các tính năng khác nhau, để tạo ra một cấu trúc dạng cây. Mỗi nút bên trong cấu trúc này biểu thị một quyết định dựa trên một tính năng cụ thể, trong khi mỗi nút lá tương ứng với một kết quả phân loại xác định. Xây dựng một cây quyết định đòi hỏi phải lựa chọn tỉ mỉ những thông tin hữu ích nhất tính năng và xác định tiêu chí phân chia tối ưu tại mỗi nút nhằm tối đa hóa sự khác biệt giữa các lớp riêng biệt.

Cây quyết định sở hữu những lợi thế đáng chú ý trong bối cảnh phân loại mạng botnet IoT, xuất phát từ khả năng nắm bắt các mối quan hệ phức tạp và tương tác tính năng. Tính linh hoạt của chúng trong việc xử lý cả các tính năng phân loại và số làm cho chúng phù hợp với nhiều loại dữ liệu lưu lượng truy cập mạng. Hơn nữa, cây quyết định thể hiện sự mạnh mẽ khi đối mặt với sự không hoàn hảo của dữ liệu trong thế giới thực, hỗ trợ các giá trị bị thiếu và ngoại lệ một cách hiệu quả.

Yếu tố khả năng diễn giải đảm nhận một vai trò quan trọng trong lĩnh vực phân loại mạng botnet IoT với cây quyết định. Các các con đường vốn có của cấu trúc cây cung cấp những hiểu biết vô giá về các đặc điểm và hành vi của các mạng botnet IoT. Thông qua việc kiểm tra cẩn thận các quy tắc quyết định, các nhà phân tích bảo mật có thể hiểu sâu hơn về các đặc điểm và kiểu phân biệt liên quan đến các loại hoạt động botnet khác nhau.

Thuật toán cây quyết định là một loại mô hình học có giám sát được sử dụng rộng rãi và phổ biến, có hiệu quả trong việc giải các bài toán phân loại và hồi quy. Cấu trúc của nó bao gồm các nút đại diện cho các biến và các nhánh thể hiện mối quan hệ giữa các biến này. Tại nút gốc, thuật toán xem xét toàn bộ tập dữ liệu và thông qua một loạt các quyết định nhị phân dựa trên các biến đầu vào, nó phân vùng đệ quy dữ liệu thành các tập con nhỏ hơn. Các phân vùng này tạo thành các nút bên trong cây và

các lá tương ứng với giá trị dự đoán của biến mục tiêu cho mỗi tập hợp con.

Để xây dựng cây quyết định, thuật toán tuân theo cách tiếp cận từ trên xuống để chọn thuộc tính tốt nhất để phân chia dữ liệu dựa trên một tiêu chí như mức tăng thông tin, chỉ số Gini hoặc entropy. Sau khi dữ liệu được phân tách, thuật toán sẽ áp dụng đệ quy quy trình tương tự cho từng tập hợp con kết quả cho đến khi đáp ứng tiêu chí dừng, chẳng hạn như độ sâu của cây được xác định trước, số lượng phiên bản tối thiểu trên mỗi lá hoặc không cải thiện thêm về hiệu suất dự đoán.

Để cập nhật và tính toán các giá trị nút trong cây quyết định, hai công thức được sử dụng: thước đo tạp chất và tiêu chí để chọn phân chia tốt nhất. Phép đo tạp chất định lượng mức độ đồng nhất hoặc không đồng nhất của biến mục tiêu trong một nút, trong khi tiêu chí để chọn phân tách tốt nhất xác định thuộc tính nào cung cấp mức tăng thông tin cao nhất hoặc tạp chất thấp nhất sau khi tách nút.

Các entropy của xác suất) phân bố
= (1, 2, 3, . . . , hai lòng = 1 =1 1

() = × nhất ký (1)
=1

Mức tăng thông tin là một công thức kiểm tra nút mang lại lượng thông tin nút đó còn lại sau khi chuyển sang nút con.

Tăng () = () — × () (2)
=1

b) Rừng ngẫu nhiên (RF)

Random Forest là một thuật toán học máy nổi bật đã thể hiện hiệu quả rõ rệt trong lĩnh vực phân loại mạng botnet IoT. sử dụng của nó vì một kỹ thuật học tập đồng bộ cho phép kết hợp nhiều cây quyết định, đạt đến đỉnh cao trong các dự đoán vừa mạnh mẽ vừa chính xác. Trong lĩnh vực an ninh mạng, Random Forest đã thu hút được sự chấp nhận đáng kể nhờ sự tinh thông trong việc xác định và phân loại các mạng botnet IoT.

Trọng tâm của thuật toán Rừng ngẫu nhiên nằm ở việc xây dựng một tập hợp bao gồm các cây quyết định. Mỗi cây quyết định được huấn luyện trên một tập con ngẫu nhiên của dữ liệu đào tạo, kết hợp một tập hợp con ngẫu nhiên các tính năng tại mỗi nút. Bằng cách giới thiệu tính ngẫu nhiên trong cả quy trình lựa chọn dữ liệu và tính năng, Random Forest tăng cường tính đa dạng trong mô hình, hạn chế hiệu quả các rủi ro liên quan đến trang bị thừa. Sự đa dạng hóa này trao quyền cho thuật toán để nắm bắt các khía cạnh khác nhau và các mẫu phức tạp bên trong dữ liệu lưu lượng truy cập mạng IoT, từ đó nâng cao hiệu suất phân loại.

Tạp chí Công nghệ thông tin & Truyền thông

Trong suốt giai đoạn đào tạo, các cây quyết định riêng lẻ trong Rừng ngẫu nhiên được xây dựng một cách tỉ mỉ thông qua phân vùng dữ liệu đệ quy. Việc phân vùng này được thực hiện dựa trên các ngưỡng tính năng riêng biệt, với mục đích chính là giảm thiểu tạp chất hoặc tối đa hóa mức thu được thông tin ở mỗi lần phân tách. Do đó, một sự sắp xếp có thứ bậc của các nút và các lá trở thành hiện thực, phục vụ như một đại diện tập thể của các mẫu thu được và các mối quan hệ qua lại có trong dữ liệu.

Khi đạt đến giai đoạn dự đoán, Random Forest hợp nhất các kết quả đầu ra được tạo bởi từng cây quyết định riêng lẻ thông qua cơ chế bỏ phiếu, được thiết kế riêng cho việc phân loại. Mỗi cây quyết định trình bày phiếu bầu của nó cho lớp được dự đoán và lớp tích lũy được đa số phiếu bầu cuối cùng được coi là dự đoán cuối cùng. Bằng cách tận dụng cách tiếp cận dựa trên tập hợp này, Random Forest đã giảm thiểu thành công các sai lệch hoặc sai sót tiềm ẩn vốn có trong các cây quyết định riêng lẻ, từ đó củng cố độ bền và độ chính xác tổng thể của phân loại.

Random Forest bao gồm một số lợi thế đáng chú ý trong lĩnh vực phân loại mạng botnet IoT. Nó xử lý thành thạo dữ liệu nhiều chiều và phức tạp, khiến nó đặc biệt phù hợp để phân tích các tính năng và mẫu đa diện vốn có trong lưu lượng mạng IoT. Ngoài ra, Random Forest vượt trội trong việc ước tính tầm quan trọng của tính năng, cung cấp thông tin chi tiết vô giá về các yếu tố then chốt góp phần vào các hoạt động của mạng botnet.

c) K-Láng giềng gần nhất (KNN)

Thuật toán k-Láng giềng gần nhất (k-NN) là một kỹ thuật học máy quan trọng và linh hoạt được sử dụng trong phân loại mạng botnet IoT. Được đánh giá cao về tính đơn giản, k-NN trình bày một cách tiếp cận trực quan để phân biệt và phân loại các botnet IoT dựa trên dữ liệu lưu lượng mạng của chúng.

Về cốt lõi, thuật toán k-NN cố gắng gán một lớp cho một điểm dữ liệu mới bằng cách đánh giá các lớp của k hàng xóm gần nhất của nó trong không gian đối tượng. Đánh giá này dựa trên các số liệu khoảng cách như Euclidean hoặc khoảng cách Manhattan, tạo thuận lợi cho việc xác định k hàng xóm gần nhất từ tập dữ liệu huấn luyện. Nhãn lớp của phiên bản mới sau đó sẽ xuất hiện thông qua quy trình bỏ phiếu đa số được tiến hành giữa những người hàng xóm được chọn này.

Trong bối cảnh phân loại mạng botnet IoT, thuật toán k-NN mang lại rất nhiều lợi thế. Trước hết, nó thể hiện khả năng đóng gói các mối quan hệ phức tạp và các mẫu phi tuyến tính phổ biến trong dữ liệu lưu lượng mạng.

Khả năng thích ứng này trao quyền cho k-NN để đáp ứng các biểu hiện đa dạng của các hoạt động mạng botnet IoT và các hoạt động liên quan của chúng. đặc trưng. Hơn nữa, k-NN thể hiện sự thành thạo trong việc xử lý cả các đặc trưng phân loại và số, tạo điều kiện thuận lợi cho

ứng dụng của nó đối với dữ liệu không đồng nhất gặp phải trong môi trường IoT. Cuối cùng, tính đơn giản vốn có của k-NN không cần thiết phải đào tạo rõ ràng, khiến việc triển khai và hiểu nó trở nên đơn giản.

Tuy nhiên, thuật toán k-NN có chứa các phép lặp giới hạn nhất định. Khi kích thước của tập dữ liệu mở rộng, độ phức tạp tính toán của việc xác định các hàng xóm gần nhất tăng lên đáng kể. Do đó, thời gian dự đoán kéo dài, đặc biệt là trong không gian nhiều chiều. Hơn nữa, hiệu suất của k-NN xoay quanh việc lựa chọn đúng tham số k, vì một lựa chọn không phù hợp có thể dẫn đến trang bị thiếu hoặc thừa.

d) Tăng cường độ dốc cực cao (XGBoost)

Extreme Gradient Boosting (XGBoost) là một thuật toán học máy đáng gờm đã nhận được sự hoan nghênh đáng chú ý trong lĩnh vực phân loại mạng botnet IoT

sự. Là một phương pháp học tập đồng bộ, XGBoost kết hợp liền mạch nhiều mô hình dự đoán yếu, điển hình là cây quyết định, để xây dựng một bộ phân loại chính xác và mạnh mẽ.

Hiệu suất và khả năng vượt trội của nó để giải quyết các nhiệm vụ phân loại mạng botnet IoT phức tạp đã giúp XGBoost trở thành một lựa chọn được ưa chuộng trong lĩnh vực an ninh mạng.

Chức năng cơ bản của thuật toán XGBoost nằm trong quy trình đào tạo lặp đi lặp lại của nó, trong đó các mô hình dự đoán yếu được tích hợp tuần tự vào tập hợp trong khi đồng thời giảm thiểu lỗi dự đoán tổng thể. Mỗi mô hình tiếp theo tập trung vào việc nắm bắt các lỗi còn sót lại của các mô hình trước đó, dần dần tinh chỉnh các khả năng dự đoán của XGBoost. Bằng cách kết hợp khéo léo các kỹ thuật tăng cường độ dốc và chính quy hóa, XGBoost cân bằng một cách khéo léo độ phức tạp của mô hình và khả năng tổng thể hóa, đạt đến đỉnh cao về hiệu suất vượt trội.

Trong bối cảnh phân loại mạng botnet IoT, XGBoost mang lại một số lợi thế đáng chú ý. Nó dễ dàng điều hướng các thách thức đặt ra bởi dữ liệu đa chiều và không đồng nhất thường gặp trong môi trường IoT, từ đó tạo điều kiện kết hợp các tính năng và mẫu đa dạng nội tại với các hoạt động của mạng botnet. Ngoài ra, XGBoost vượt trội trong việc nắm bắt các mối quan hệ và tương tác phức tạp giữa các tính năng này, nâng cao khả năng phân loại chính xác các botnet IoT. Ngoài ra, XGBoost Seam tích hợp một cách dễ dàng các kỹ thuật chính quy giúp giảm thiểu hiệu quả nguy cơ trang bị quá mức, tăng cường độ bền và khả năng khái quát hóa của nó.

Tuy nhiên, XGBoost gặp phải những thách thức về khả năng diễn giải. Là một tập hợp các cây quyết định, việc làm sáng tỏ những đóng góp riêng lẻ của từng cây trong mô hình XGBoost có thể rất khó khăn. Tuy nhiên, các kỹ thuật tập trung vào phân tích tầm quan trọng của tính năng có thể được sử dụng để hiểu rõ hơn về tầm quan trọng tương đối của

đặc điểm trong quá trình phân loại.

Extreme Gradient Boosting (XGBoost) nổi lên như một thuật toán mạnh mẽ và đáng khen ngợi trong lĩnh vực phân loại mạng botnet IoT. Thông qua phương pháp học tập đồng bộ, được kết hợp với các kỹ thuật tăng cường độ dốc và điều chỉnh độ dốc, XGBoost nắm bắt một cách thành thạo các mối quan hệ phức tạp và xử lý dữ liệu nhiều chiều một cách khéo léo. Bất chấp những thách thức tiềm ẩn liên quan đến khả năng dự đoán lẫn nhau, hiệu suất kiên quyết và khả năng phân loại chính xác các botnet IoT của XGBoost đã nhấn mạnh vai trò không thể thiếu của XGBoost trong lĩnh vực an ninh mạng.

4. Đánh giá hiệu suất

Khi đánh giá các mô hình thuật toán được mô tả trước đó, các kỹ thuật đa dạng đã được sử dụng để đánh giá độ chính xác của kết quả và rút ra kết quả toàn diện cho từng mô hình. Nghiên cứu này liên quan đến một số khái niệm cơ bản, chẳng hạn như TP, TN, FP và FN. TP biểu thị số lượng dương tính thực sự đã được xác định chính xác, trong khi TN là số lượng âm tính thực sự đã được xác định chính xác. FP biểu thị số trường hợp dương tính thực tế đã bị phân loại nhầm là âm tính, trong khi FN biểu thị số trường hợp âm tính bị phân loại nhầm là dương tính.

a) Độ chính xác (PRE)

Độ chính xác là thước đo hiệu suất được sử dụng để đánh giá hiệu quả của mô hình bằng cách xác định tỷ lệ các trường hợp tích cực được xác định chính xác. Biện pháp này có thể được tính toán bằng toán học thông qua việc sử dụng một công thức, có tính đến số lượng dương tính thật và dương tính giả. Cụ thể, độ chính xác được tính bằng cách chia số lượng dương tính thực cho tổng số dương tính thật và dương tính giả.

= /(+) (3)

b) Độ chính xác (ACC)

Độ chính xác là thước đo hiệu suất được sử dụng để đánh giá hiệu quả của mô hình bằng cách xác định tỷ lệ dự đoán chính xác trong tổng số dự đoán. Biện pháp này có thể được tính toán bằng toán học thông qua việc sử dụng một công thức, xem xét số lượng dương tính thực và âm tính thực. Cụ thể, độ chính xác được tính bằng cách chia tổng số dương thực và âm thực cho tổng số dự đoán.

= (TP + TN)/(TP + TN + FP + FN) (4)

c) Điểm nhớ lại (RE)

Điểm thu hồi là chỉ số hiệu suất được sử dụng để đánh giá hiệu quả của mô hình trong việc xác định chính xác các trường hợp tích cực thực tế. Biện pháp này có thể được tính toán bằng toán học thông qua việc sử dụng một công thức, kết hợp số lượng dương tính thực và âm tính giả. Cụ thể, điểm thu hồi được tính bằng cách chia số lượng dương tính thực cho tổng số dương tính thật và âm tính giả.

= /(+) (5)

d) Điểm F1 cho lớp nhị phân

Điểm F1 là thước đo hiệu suất thu được bằng cách tính trung bình cả điểm chính xác và điểm thu hồi. Biện pháp này được sử dụng rộng rãi để đánh giá hiệu quả tổng thể của một mô hình vì nó cung cấp một cái nhìn toàn diện về các kết quả dương tính giả và âm tính giả. Cụ thể, điểm F1 được tính là giá trị trung bình hài hòa của độ chính xác và khả năng thu hồi. Công thức tính điểm F1 tính đến cả số lượng kết quả dương tính thật, kết quả dương tính giả và âm tính giả, từ đó đưa ra đánh giá cân bằng hơn về hiệu suất của mô hình.

1 điểm = 2 × $\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$ (6)

e) Điểm F1 cho Đa lớp

Điểm F1 [16] là một thước đo được công nhận rộng rãi và thường được sử dụng để đánh giá hiệu suất của các thuật toán phân loại nhiều lớp. Nó phục vụ như một phép đo toàn diện có tính đến cả độ chính xác và khả năng thu hồi, do đó cho phép đánh giá cân bằng về hiệu quả tổng thể của bộ phân loại.

Trong bối cảnh phân loại nhiều lớp trong IoT

vấn đề botnet, điểm số F1 cung cấp thông tin chi tiết có giá trị về khả năng của thuật toán để phân loại chính xác các phiên bản trên các lớp khác nhau. Bằng cách xem xét sự xuất hiện của kết quả dương tính giả (trường hợp được phân loại sai) và kết quả âm tính giả (trường hợp bị bỏ sót) đối với mỗi lớp, điểm F1 cung cấp đánh giá mạnh mẽ về hiệu suất của thuật toán.

Việc tính toán điểm F1 liên quan đến việc xác định giá trị trung bình điều hòa của độ chính xác và khả năng thu hồi cho mỗi lớp. Cách tiếp cận này đánh giá hiệu quả trạng thái cân bằng giữa việc xác định chính xác các trường hợp tích cực (độ chính xác) và nắm bắt tất cả các trường hợp tích cực (thu hồi). Điểm F1 cao hơn biểu thị sự cân bằng vượt trội giữa độ chính xác và khả năng thu hồi, qua đó thể hiện năng lực của trình phân loại trong việc xác định chính xác các trường hợp trên tất cả các lớp.

Tóm lại, điểm F1 đóng vai trò then chốt như một thước đo quan trọng trong việc đánh giá hiệu suất của các thuật toán phân loại nhiều lớp trong vấn đề mạng botnet IoT. Nó tạo điều kiện cho việc đánh giá toàn diện bằng cách xem xét

Tạp chí Công nghệ thông tin & Truyền thông

độ chính xác và thu hồi cho từng lớp, do đó cho phép đánh giá toàn diện về hiệu quả tổng thể của bộ phân loại trong việc phân loại chính xác các trường hợp trên nhiều lớp.

IV. KẾT QUẢ VÀ BÀN LUẬN

1. Phân loại nhị phân

Bảng 5 trình bày một phân tích toàn diện về các kết quả đánh giá thu được từ phân loại nhị phân trong vấn đề mạng botnet IoT. Bốn thuật toán học máy riêng biệt, cụ thể là Cây quyết định (DT), Rừng ngẫu nhiên (RF), k Hàng xóm gần nhất (KNN) và Tăng cường độ dốc cực cao (XGB), đã được sử dụng để phân biệt hiệu suất của chúng bằng cách sử dụng các chỉ số khác nhau.

Bảng 5. Một số thước đo phân loại nhị phân				
đánh giá DT	RF KNN XGB			
ACC	99,9%	89,5%	99,6%	99,8%
TRƯỚC	100,0%	88,0%	100,0%	100%
NÓT RẺ	100,0%	86,0%	99,0%	100%
F1	100,0%	87,0%	100,0%	100%
Thời gian	5,9	40.2	58.2	99,4

Về độ chính xác (ACC), thuật toán Cây quyết định trưng bày độ chính xác vô song, đạt được một đặc biệt tỷ lệ chính xác 99,9%. Theo sát gót, Extreme Gra dient Boosting đã thể hiện độ chính xác ấn tượng của 99,8%. k-Nearest Neighbor thể hiện độ chính xác đáng khen ngợi ở mức 99,6%, trong khi Random Forest đạt được mức đáng nể độ chính xác 89,5%.

Độ chính xác (PRE), đo lường khả năng chính xác phân loại các trường hợp tích cực trong số tất cả các trường hợp được dự đoán tích cực, mang lại kết quả đáng chú ý. cả quyết định Tree và Extreme Gradient Boosting đạt được hiệu quả hoàn hảo điểm chính xác 100,0%. k-Nearest Neighbor cũng thể hiện độ chính xác tuyệt vời, phù hợp với số điểm hoàn hảo của 100,0%. Mặc dù thấp hơn một chút, Random Forest đã thực hiện đáng khen ngợi với số điểm chính xác là 88,0%.

Việc đánh giá Thu hồi (RE), định lượng tỷ lệ phần trăm của các trường hợp tích cực được phân loại chính xác trong số tất cả trường hợp tích cực thực tế, tiết lộ kết quả mẫu mực. Cả hai Cây quyết định và Extreme Gradient Boosting xuất sắc với điểm thu hồi hoàn hảo là 100,0%. k-Hàng xóm gần nhất đã chứng minh tỷ lệ thu hồi đáng khen ngợi là 99,0%, trong khi Random Forest đạt được tỷ lệ thu hồi đáng nể là 86,0%.

Điểm F1, đại diện cho giá trị trung bình hài hòa của độ chính xác và khả năng thu hồi, cung cấp đánh giá tổng thể về hiệu suất của bộ phân loại. Cả cây quyết định và cực đoan Gradient Boosting đạt được điểm số F1 hoàn hảo là 100,0%. k-Nearest Neighbor thể hiện hiệu suất tuyệt vời với điểm F1 là 100,0%, trong khi Random Forest đạt được điểm số điểm đáng nể là 87,0%.

Xét về thời gian tính toán, thuật toán Cây quyết định nổi lên là hiệu quả nhất, chỉ cần một 5,9 giây để xử lý. Rừng ngẫu nhiên theo sau với thời gian xử lý là 40,2 giây, trong khi k-Hàng xóm gần nhất yêu cầu 58,2 giây. Tăng cường độ dốc cực cao thể hiện thời gian xử lý lâu nhất, đạt 99,4 giây.

2. Đa phân loại

Bảng 6 minh họa việc đánh giá toàn diện và kết quả phân tích liên quan đến đa phân loại trong bối cảnh của vấn đề IoT-botnet. Nghiên cứu có sự tham gia của đánh giá bốn thuật toán học máy riêng biệt, cụ thể là Cây quyết định (DT), Rừng ngẫu nhiên (RF), k Hàng xóm gần nhất (KNN) và Tăng cường độ dốc cực cao (XGB), sử dụng một loạt các chỉ số hiệu suất.

Bảng 6. Một số thước đo của đa phân loại				
Đánh giá DT RF KNN XGB				
ACC	99,9%	78,2%	99,8%	99,9%
TRƯỚC	99,9%	59,6%	99,7%	100,0%
NÓT RẺ	99,7%	46,1%	98,7%	99,8%
F1	99,6%	49,8%	99,2%	99,9%
Thời gian	11,5	206.1	85.2	956.2

Số liệu chính về độ chính xác (ACC) cho thấy hiệu suất đáng kể đạt được bởi cả Quyết định Các thuật toán Tree và Extreme Gradient Boosting, đạt được tỷ lệ chính xác cao 99,9%. Theo sát, k-Gần nhất Neighbor thể hiện độ chính xác đáng khen ngợi ở mức 99,8%, trong khi Random Forest thể hiện độ chính xác giảm nhẹ là 78,2%.

Độ chính xác (PRE), đánh giá khả năng chính xác phân loại các trường hợp trong mỗi lớp, nhấn mạnh điểm số chính xác đặc biệt mà Cây quyết định đạt được (99,9%) và thuật toán Extreme Gradient Boosting (100,0%). k-Nearest Neighbor đã thể hiện một kết quả đáng khen ngợi chính xác 99,7%, trong khi Random Forest hiển thị độ chính xác tương đối thấp hơn là 59,6%.

Số liệu thu hồi (RE), phản ánh độ nhạy để phân loại chính xác các trường hợp trong mỗi lớp, được giới thiệu kết quả nổi bật cho Cây quyết định (99,7%) và Thuật toán tăng cường độ dốc cực cao (99,8%). k-gần nhất Hàng xóm thể hiện tỷ lệ nhớ lại là 98,7%, trong khi Ngẫu nhiên Forest đạt được mức thu hồi 46,1%, cho thấy tương đối hiệu suất thấp hơn trong khía cạnh này.

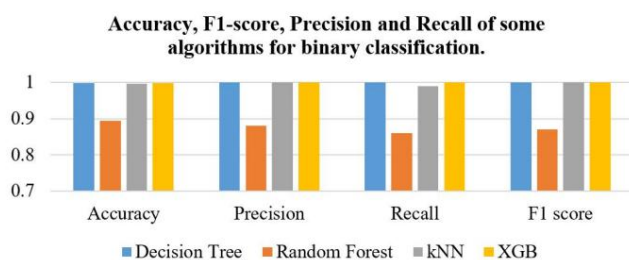
Điểm số F1, phục vụ như một đánh giá toàn diện về cả độ chính xác và khả năng thu hồi, làm nổi bật hơn nữa khả năng đáng chú ý hiệu suất của Cây quyết định (99,6%) và Extreme Thuật toán tăng cường độ dốc (99,9%). k-Hàng xóm gần nhất đạt được số điểm F1 đáng nể là 99,2%, trong khi Ngẫu nhiên Rừng mang lại số điểm tương đối thấp hơn là 49,8%.

Về mặt thời gian tính toán, Cây quyết định algorithm đã thể hiện thời gian xử lý ngắn nhất là 11,5 giây, cho thấy tính hiệu quả của nó. k-Hàng xóm gần nhất yêu cầu 85,2 giây, trong khi Khu rừng ngẫu nhiên tiêu tốn thời gian xử lý lâu hơn là 206,1 giây. Mặt khác, thuật toán Extreme Gradient Boosting thể hiện nhu cầu tính toán cao nhất, với thời gian xử lý là 956,2 giây.

Những phát hiện của đánh giá này làm nổi bật hiệu suất vượt trội của các thuật toán Cây quyết định và Tăng cường độ dốc cực cao trên nhiều chỉ số, bao gồm độ chính xác, độ chính xác, thu hồi và điểm F1. Ngoài ra, thuật toán Cây quyết định nổi bật về hiệu quả tính toán của nó. Tuy nhiên, điều quan trọng là phải xem xét các yêu cầu cụ thể và mức độ ưu tiên của ứng dụng khi chọn thuật toán phù hợp nhất, có tính đến việc xem xét cân bằng độ chính xác, độ chính xác, khả năng thu hồi và hiệu quả tính toán. Kết quả của nghiên cứu này cung cấp những hiểu biết có giá trị về các đặc tính hiệu suất của các thuật toán học máy khác nhau để phân loại đa dạng trong các vấn đề về botnet IoT.

3. Đánh giá mô hình

Để xác định sự phù hợp của một mô hình nhất định đối với một vấn đề cụ thể, các kỹ thuật đánh giá hiệu suất được sử dụng. Một loạt các phương pháp có thể được sử dụng, bao gồm nhưng không giới hạn ở các phép đo độ chính xác, độ chính xác, thu hồi và điểm F1. Những phương pháp này phục vụ như đáng tin cậy các chỉ số về khả năng của một mô hình để giải quyết vấn đề ban đầu một cách hiệu quả.

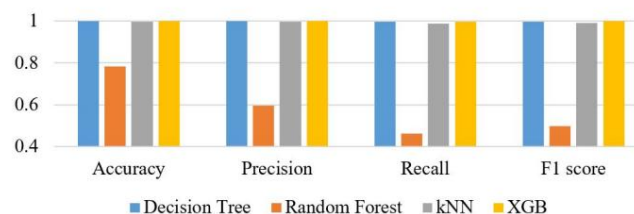


Hình 3: Accuracy, F1-score, Precision và Recall của một số thuật toán phân loại nhị phân

Đánh giá hiệu suất của các thuật toán DT, RF, KNN và XGB để phân loại nhị phân được trình bày trong Hình 3, cung cấp tổng quan toàn diện về độ chính xác, độ chính xác, thu hồi và số liệu điểm F1. Các kết quả thể hiện rõ ràng hiệu suất vượt trội được thể hiện bởi cả hai thuật toán DT và XGB, vượt qua các đối tác tương ứng của chúng. Ngoài ra, về mặt thời gian đào tạo, thuật toán Cây quyết định nổi lên khi càng nhiều

tùy chọn hiệu quả, do đó mang lại hiệu suất tổng thể vượt trội cây chùy.

Accuracy, F1-score, Precision and Recall of some algorithms for multi-classification.

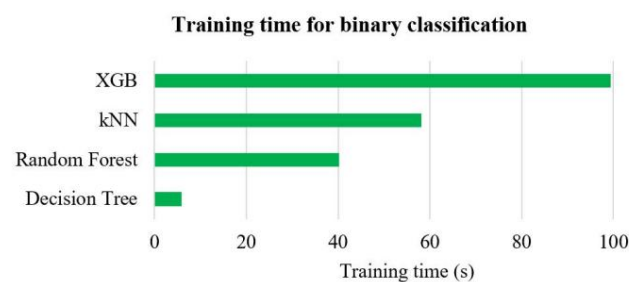


Hình 4: Accuracy, F1-score, Precision và Recall của một số thuật toán cho đa phân loại

Bài toán đa lớp được đánh giá để đánh giá hiệu suất của bốn thuật toán riêng biệt, cụ thể là Cây quyết định, Rừng ngẫu nhiên, KNN và XGB, như được mô tả trong Hình 4. Đánh giá bao gồm phân tích toàn diện về độ chính xác, độ chính xác, khả năng thu hồi và chỉ số điểm F1.

Đáng chú ý, kết quả chứng minh rằng thuật toán Cây quyết định vượt qua các thuật toán khác về các chỉ số hiệu suất này, thể hiện hiệu suất vượt trội.

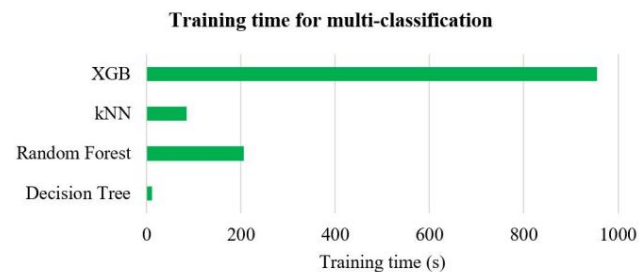
Hình 5 trình bày thời gian đào tạo của các thuật toán cho bài toán phân loại nhị phân. Kết quả cho thấy thuật toán XGB có thời gian đào tạo lâu hơn đáng kể so với các thuật toán khác. Ngược lại, thuật toán cây quyết định nổi lên là hiệu quả nhất về mặt thời gian đào tạo, vượt trội so với tất cả các thuật toán khác về mặt này.



Hình 5: Thời gian đào tạo để phân loại nhị phân

Trong bài toán đa phân loại, thời gian đào tạo của các thuật toán được trình bày trong hình 6. Kết quả cho thấy thuật toán XGB mất nhiều thời gian hơn để đào tạo so với các thuật toán khác, với thời gian đào tạo gần gấp 100 lần so với thuật toán cây quyết định. Mặt khác, thuật toán cây quyết định thể hiện hiệu suất tốt nhất về thời gian đào tạo.

Tạp chí Công nghệ thông tin & Truyền thông



Hình 6: Thời gian đào tạo đa phân loại

V. KẾT LUẬN

Tóm lại, bài nghiên cứu học thuật này đã thực hiện nhiệm vụ quan trọng là phát hiện và phân loại các botnet IoT bằng thuật toán học máy. Nghiên cứu đã nhấn mạnh đáng kể vào việc phân tích và thao tác kỹ lưỡng dữ liệu mạng botnet IoT, với trọng tâm cụ thể là bộ dữ liệu IoT-23 được đánh giá cao. Bằng cách triển khai các thuật toán học máy được công nhận và sử dụng rộng rãi như Cây quyết định (DT), k-Hàng xóm gần nhất (KNN), Rừng ngẫu nhiên (RF) và Tăng cường độ dốc cực đại (XGBoost), mục tiêu là phân loại và phát hiện botnet một cách hiệu quả trong giới hạn của bộ dữ liệu IoT-23.

Việc áp dụng các thuật toán này nhằm mục đích nâng cao hiểu biết của chúng ta về hiệu suất và hiệu quả của chúng trong lĩnh vực phát hiện và phân loại mạng botnet IoT.

Thông qua phân tích so sánh các kết quả thu được từ các thuật toán đa dạng này, người ta đã thu được những hiểu biết có giá trị, làm sáng tỏ những điểm mạnh và hạn chế tương ứng của chúng. Những hiểu biết sâu sắc như vậy trang bị cho các nhà nghiên cứu và người thực hành kiến thức cần thiết để đưa ra quyết định sáng suốt khi chọn thuật toán phù hợp nhất để phát hiện và phân loại mạng botnet IoT thành công. Cái này nghiên cứu đóng góp đáng kể cho lĩnh vực này bằng cách cung cấp đánh giá toàn diện về hiệu suất của các thuật toán học máy này, từ đó tạo điều kiện phát triển các phương pháp phân loại và phát hiện linh hoạt và hiệu quả hơn.

Bằng cách giải quyết thách thức then chốt trong việc phát hiện và phân loại mạng botnet IoT, nghiên cứu này có ý nghĩa mở rộng sang lĩnh vực tăng cường các biện pháp an ninh mạng trong hệ sinh thái IoT. Những phát hiện và thông tin chi tiết thu được từ nghiên cứu này có khả năng cung cấp thông tin cho việc phát triển các biện pháp chủ động nhằm ngăn chặn và giảm thiểu tác động bất lợi của các cuộc tấn công mạng botnet IoT. Người ta dự đoán rằng công việc này sẽ truyền cảm hứng cho những người đam mê nghiên cứu sâu hơn và thúc đẩy sự hợp tác trong lĩnh vực này, cuối cùng dẫn đến sự tiến bộ của các hệ thống IoT an toàn hơn và bảo vệ các cơ sở hạ tầng quan trọng.

NGƯỜI GIỚI THIỆU

- [1] Williams, P., Dutta, IK, Daoud, H, và Bayoumi, M.: Một cuộc khảo sát về bảo mật trong internet vạn vật tập trung vào tác động của các công nghệ mới nổi. Elsevier. (2022).
- [2] Khan, WZ, Rehman, MH, Zangoti, HM, Afzal, MK, Armi, N., và Salah, K.: Internet vạn vật trong công nghiệp: Những tiến bộ gần đây, hỗ trợ công nghệ và thách thức mở. Elsevier. (2020).
- [3] Vitorino, J., Andrade, R., Prac, a, I., Sousa, O., và Maia, E.: Phân tích so sánh các kỹ thuật học máy cho phát hiện xâm nhập IoT. Nền tảng và Thực hành Bảo mật (trang 191- 207). Springer. (2022).
- [4] Haq, NF, Onik, AR, Hridoy, MAK, Rafni, M., Shah, FM và Farid, DM: Ứng dụng các phương pháp học máy trong Hệ thống phát hiện xâm nhập: Khảo sát. Bài báo được đăng trên Tạp chí Quốc tế về Nghiên cứu Tiên tiến về Trí tuệ Nhân tạo (IJARAI), Tập 4 Số 3. (2015).
- [5] Hajji, J., Khalily, M., Moustafa, N., và Nelms, T. IoT-23: Bộ dữ liệu để phân tích lưu lượng mạng IoT. lờ xo. (2019).
- [6] Rahim, A., Razzaque, MA, Hasan, R. và Hossain, MF Bảo mật mạng IoT hiệu quả thông qua lựa chọn tính năng và kỹ thuật máy học. IEEE. (2020).
- [7] Islam, SMZ, Bhuiyan, MZH và Hasan, R. Sự kết hợp của các mô hình học máy để phát hiện xâm nhập trong mạng IoT bằng Bộ dữ liệu IoT-23. IEEE. (2020)
- [8] Li, Y., Qiu, L., Chen, Y. và Chen, Y. Hệ thống phát hiện xâm nhập dựa trên Ensemble cho các mạng IoT sử dụng Bộ dữ liệu IoT-23. IEEE. (2020)
- [9] PH Do, TD Dinh, DT Le, VD Pham, L. Myrova và R. Kirichek, " An Efficient Feature Extraction Method for Attack Classification in IoT Networks," 2021 13th International Congress on Ultra Modern Telecommunications and Control Systems và Cửa hàng làm việc (ICUMT)
- [10] Alotaibi, F., Al-Qaness, MA, Abunadi, A. và Al ghazzawi, MA Một phương pháp tiếp cận ứng dụng học sâu để phát hiện xâm nhập trong mạng IoT bằng cách sử dụng Bộ dữ liệu IoT 23. IEEE. (2020).
- [11] Li, J., Hu, C., Yang, K., Zhang, X. và Lu, J. Một hệ thống phát hiện xâm nhập IoT dựa trên IoT-23 sử dụng Deep Learning. IEEE. (2020).
- [12] Abdallah, A., Khalil, I., Al-Emadi, N., Almohameed, A., và Kim, H. Phát hiện Botnet IoT thời gian thực bằng cách sử dụng Deep Learning trên Bộ dữ liệu IoT-23. IEEE. (2020)
- [13] Kiani, AT, Abbas, RA, Abbasi, AZ và Khan,

Phát hiện sự bất thường dựa trên học sâu MK cho các mạng IoT bằng cách sử dụng Bộ dữ liệu IoT-23. IEEE. (2020)

[14] Rasool, S., Saeed, S., Farooq, F., và Madani, A. Nghiên cứu so sánh các phương pháp học tập chuyển đổi để phát hiện phần mềm độc hại IoT bằng bộ dữ liệu IoT-23. IEEE. (2021).

[15] Sebastian Garcia, Agustin Parmisano, và Maria Jose Erquiaga. (2020). IoT-23: Tập dữ liệu được gắn nhãn có lưu lượng truy cập mạng IoT độc hại và lành tính (Phiên bản 1.0.0) [Tập dữ liệu]. Zenodo. <http://doi.org/10.5281/zenodo.4743746>

[16] Stoian, NA Máy học để phát hiện bất

thường trong mạng IoT : Phân tích phần mềm độc hại trên bộ dữ liệu IoT-23. EEMCS: Kỹ thuật Điện, Toán học và Khoa học Máy tính. (2020)



Kết nối mạng.

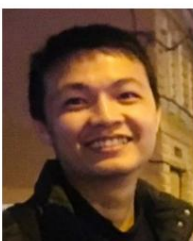
Nguyễn Năng Hùng Vân nhận bằng Tiến sĩ. bằng Khoa học Máy tính của Đại học Đà Nẵng, Việt Nam, năm 2021. Ông hiện là giảng viên tại Đại học Đà Nẵng - Đại học Khoa học và Công nghệ. Lĩnh vực nghiên cứu của anh bao gồm Trí tuệ nhân tạo, Học máy, Đại số hình học, Máy tính



Phạm Văn Quân hiện đang là sinh viên năm cuối chuyên ngành Khoa học dữ liệu và Trí tuệ nhân tạo trường Đại học Đông Á. Lĩnh vực nghiên cứu của ông bao gồm DS, ML, AI và ứng dụng của nó trong các lĩnh vực khác nhau như Tài chính, Mạng và Xử lý ngôn ngữ tự nhiên.



Ngô Văn Úc trở thành sinh viên Đại học Đông Á năm 2020, chuyên ngành Trí tuệ nhân tạo và Khoa học dữ liệu. Mối quan tâm nghiên cứu của anh ấy bao gồm máy học
ing, học sâu, khoa học dữ liệu, trí tuệ nhân tạo, xử lý hình ảnh và các ứng dụng của chúng



Đỗ Phúc Hào nhận bằng Thạc sĩ Khoa học Máy tính của Đại học Đà Nẵng - Đại học Bách khoa năm 2017. Anh hiện là Tiến sĩ Khoa học máy tính. sinh viên khoa Mạng Truyền thông và Truyền số liệu tại

Đại học Viễn thông Bang Bonch-Bruevich Saint-Peterburg, Nga.

Lĩnh vực nghiên cứu của anh bao gồm Trí tuệ nhân tạo, Học máy và ứng dụng của nó trong các lĩnh vực khác nhau như mạng, chuỗi khối.