BỘ GIÁO DỤC VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

TỐNG ANH TUẤN

NGHIÊN CỨU CẢI TIẾN MỘT SỐ MÔ HÌNH HỌC MÁY VÀ HỌC SÂU ÁP DỤNG CHO BÀI TOÁN PHÂN LOẠI DGA BOTNET

Chuyên ngành: Hệ thống thông tin

Mã số: 9 48 01 04

TÓM TẮT LUẬN ÁN TIẾN SĨ MÁY TÍNH

Hà Nội - 2023

Công trình được hoàn thành tại: Học viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam

Người hướng dẫn khoa học 1: PGS. TS. Hoàng Việt Long
Người hướng dẫn khoa học 2: PGS. TS. Nguyễn Việt Anh

Phản biện 1:
Phản biện 2:
Phản biện 3:

Luận án được bảo vệ trước Hội đồng chấm luận án tiến sĩ, họp tại Học viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam vào hồi giờ, ngày tháng năm 2023.

Có thể tìm hiểu luận án tại:
- Thư viện Học viện Khoa học và Công nghệ
- Thư viện Quốc gia Việt Nam

Danh mục các công trình của tác giả

CT1 Nguyễn Văn Căn, Đoàn Ngọc Tú, **Tống Anh Tuấn**, Hoàng Việt Long, Lê Hoàng Sơn, Nguyễn Thị Kim Sơn. (2020). A new method to classify malicious domain name using Neutrosophic sets in DGA Botnet detection. *Journal of Intelligent & Fuzzy Systems*, 38(4), 4223-4236. *(ISI Q2, IF = 1.737)*

CT 2 **Tống Anh Tuấn**, Hoàng Việt Long, Lê Hoàng Sơn, Raghvendra Kumar, Ishaani Priyadarshini, Nguyễn Thị Kim Sơn. (2020). Performance evaluation of Botnet DDoS attack detection using machine learning. *Evolutionary Intelligence*, 13(2), 283-294. *(SCOPUS, ESCI Q2)*

CT 3 **Tống Anh Tuấn**, Nguyễn Việt Anh, Hoàng Việt Long. (2021, December). Assessment of Machine Learning Models in Detecting DGA Botnet in Characteristics by TF-IDF. In *2021 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)* (pp. 1-5). IEEE. *(SCOPUS)*

CT 4 **Tống Anh Tuấn**, Hoàng Việt Long, David Taniar. (2022). On Detecting and Classifying DGA Botnets and their Families. *Computers & Security*, 113, 102549. *(ISI Q1, IF = 5.105)*

CT 5 **Tống Anh Tuấn**, Nguyễn Việt Anh, Trần Thị Lượng, Hoàng Việt Long. (2023). UTL_DGA22-a dataset for DGA botnet detection and classification. *Computer Networks*, 221, 109508. *(ISI Q1, IF = 5.493)*

CT 6 **Tống Anh Tuấn**, Nguyễn Ngọc Cương, Nguyễn Việt Anh, Hoàng Việt Long. (2022). Đề xuất ứng dụng giải pháp phân lớp nhị phân trong bài toán DGA Botnet cho phát hiện địa chỉ IP độc hại. *Hội thảo Quốc gia lần thứ XXV "Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông" (VNICT 2022)*, trang 55-60.

MỞ ĐẦU

1. Tính cấp thiết của luận án

Botnet là một mối đe dọa thường trực trên Internet [1]. Chúng liên tục phát triển, cải tiến mã nguồn, đổi mới phương thức lây nhiễm và khả năng phá hoại ngày càng lớn. Các hệ thông thông tin khi triển khai trên Internet luôn phải đối mặt với nguy cơ bị tấn công bởi Botnet, mang đến những thiệt hại rất lớn về kinh tế, danh tiếng, dịch vụ và thậm chí cả ảnh hưởng chính trị.

Một số nghiên cứu đã làm rõ mối nguy hiểm và đề xuất kỹ thuật phát hiện Botnet như: Ghafir và cộng sự [3], Alieyan và cộng sự [7], Kwon và cộng sự với giải pháp PsyBoG [8], Wang và cộng sự với giải pháp DBod [9], Bisio và cộng sự [10], Trung và cộng sự [11] với PGS-Graph.

Theo hướng tiếp cận dựa trên học máy và học sâu, một số nghiên cứu tiêu biểu gồm: Hiếu và cộng sự [12], Khan và cộng sự [13], Xuân và cộng sự [14], Đức và cộng sự [15] với LSTM.MI, Curtin và cộng sự [16], Vinayakumar và cộng sự [17]. Zago và cộng sự công bố một bộ dữ liệu mới là UMUDGA Dataset [18] chuyên dùng cho đánh giá bài toán DGA Botnet.

Các kết quả nghiên cứu trên cho thấy: Trong các hướng tiếp cận được đề cập, hướng tiếp cận phân tích lưu lượng, sử dụng học máy, học sâu nói chung hay mạng LSTM nói riêng cho kết quả cao từ 96.3% trở lên trong bài toán phát hiện DGA Botnet. Tuy nhiên, các kết quả này hoàn toàn có thể được cải tiến thêm và đánh giá toàn diện hơn trên các bộ dữ liệu mới đầy đủ hơn. Một vấn đề khác đặt ra là các nghiên cứu về phân loại hay nhận diện họ DGA Botnet còn hạn chế, ít được đề cập hoặc độ chính xác đạt được chưa cao (LSTM đạt 53%, LSTM.MI đạt 49%), một số họ DGA Botnet khả năng nhận diện kém. Cuối cùng, việc đánh giá trên các bộ dữ liệu chính thức còn hạn chế.

Từ các vấn đề trên, NCS đặt ra các câu hỏi nghiên cứu cho luận án như sau:

Câu hỏi 1: Đối với bài toán phát hiện DGA Botnet, các hướng tiếp cận mới bao gồm sử dụng thuật toán phân cụm trên tập mờ, sử dụng mô hình học máy kết hợp có hiệu quả hay không?

Câu hỏi 2: Mạng LSTM có thể được cải tiến để nâng cao hiệu quả của việc phát hiện và phân loại DGA Botnet không và giải pháp cụ thể là gì? Trong đó trọng tâm là giải pháp để phân loại DGA Botnet.

Câu hỏi 3: Các bộ dữ liệu về DGA Botnet hiện nay có những đặc điểm gì gây hạn chế cho việc thử nghiệm thuật toán, đối sánh các kết quả nghiên cứu hay tính cập nhật. Có thể xây dựng bộ dữ liệu mới để giải quyết các hạn chế trên hay không?

Kết quả nghiên cứu của luận án có thể được ứng dụng vào các module phòng chống Botnet trên các thiết bị bảo mật truyền thống như Firewall, IDS hay trên các giải pháp tiên tiến như NGFW và UTM.

2. Mục tiêu nghiên cứu

Đề tài đặt ra mục tiêu chung là nghiên cứu, cải tiến các mô hình học máy, học sâu để nâng cao độ chính xác của giải pháp phân loại DGA Botnet, với các mục tiêu cụ thể như sau:

- Nghiên cứu về đặc điểm của DGA Botnet. Trình bày nền tảng lý thuyết, các kỹ thuật, nghiên cứu liên quan, là cơ sở để phát triển các thuật toán phát hiện và phân loại DGA Botnet.

- Nghiên cứu, đánh giá hiệu quả của hai hướng tiếp cận là thuật toán phân cụm trên tập mờ, kỹ thuật học máy kết hợp để giải quyết bài toán phát hiện DGA Botnet.

- Đề xuất mô hình học sâu mới trên nền tảng kế thừa mạng LSTM để phát hiện và phân loại DGA Botnet. Trong đó, trọng tâm chính là bài toán phân loại DGA Botnet với mục tiêu nâng cao đáng kể độ chính xác so với các giải pháp trước đó.

3. Các nội dung nghiên cứu chính của luận án

Để giải quyết các câu hỏi nghiên cứu đặt ra, NCS nghiên cứu tổng quan các kỹ thuật phát hiện DGA Botnet và các nghiên cứu liên quan. Đề xuất giải pháp để nâng cao độ chính xác của thuật toán phát hiện và phân loại DGA Botnet. Bên cạnh các hướng tiếp cận truyền thống, NCS cũng thực hiện các hướng tiếp cận mới như sử dụng thuật toán phân cụm trên tập mờ, sử dụng kỹ thuật học kết hợp. NCS cũng xây dựng một bộ dữ liệu mới về DGA Botnet với những cải tiến, cập nhật mới.

Một số nội dung chi tiết mà NCS sẽ tập trung nghiên cứu như sau:

- Nghiên cứu đặc điểm, các kỹ thuật phát hiện và phân loại DGA Botnet;

- Nghiên cứu, thuật toán phân cụm trên tập Neutrosophic Set, học máy và các mô hình học kết hợp để áp dụng cho phát hiện DGA Botnet.

- Nghiên cứu mạng LSTM và các biến thể, trên cơ sở đó cải tiến, đề xuất giải pháp phát hiện, phân loại DGA Botnet. Trọng tâm là bài toán phân loại DGA Botnet.

- Nghiên cứu các bộ dữ liệu chuyên dùng về DGA Botnet, bao gồm: Botnet DGA Dataset [19], Andrey Abakumov [20], UMUDGA Dataset [21] [18], DGArchive [22], OSINT DGA feed [23], 360NetLab Dataset [24], Johannes Bader [25] và xây dựng bộ dữ liệu mới.

4. Các đóng góp của luận án

Các đóng góp của luận án đạt được qua quá trình nghiên cứu như sau:

- *Đóng góp 1:* Đề xuất ba giải pháp phát hiện và phân loại DGA Botnet, bao gồm NCM, LA_Bin07, LA_Mul07 nhằm nâng cao độ chính xác so với các giải pháp trước đó.

- *Đóng góp 2:* Đề xuất một bộ dữ liệu mới UTL_DGA22 chuyên dụng cho bài toán DGA Botnet phục vụ cho các nghiên cứu cùng hướng trong tương lai.

CHƯƠNG 1. CƠ SỞ LÝ THUYẾT VỀ DGA BOTNET

Chương 1 trình bày cơ sở kiến thức về Botnet nói chung và DGA Botnet nói riêng. NCS cũng trình bày hai bài toán trong DGA Botnet là phân lớp nhị phân và phân lớp đa lớp, tương ứng với phát hiện và phân loại DGA Botnet. Đây cũng là vấn đề được NCS tập trung nghiên cứu, giải quyết và trình bày kết quả trong các chương tiếp theo của luận án này.

1.1. Tổng quan chung về Botnet

1.1.1. Khái niệm Botnet

Theo Provos & Holz, Botnet là một "mạng gồm rất nhiều máy tính bị xâm nhập và có thể bị kẻ tấn công điều khiển từ xa".

1.1.2. Các bước phát triển của công nghệ Botnet

Các con Bot ban đầu được thiết kế như một công cụ hữu ích hoạt động trên giao thức IRC. Sau đó các tính năng mới dần được cập nhật như cho phép điều khiển từ xa, chiếm quyền điều khiển, thiết kế dạng module, khả năng gián điệp, các cơ chế lây nhiễm và ẩn mình mới tinh vi và hiện nay là mở rộng lây nhiễm sang thiết bị IoT.

1.1.3. Một số đặc điểm của Botnet

Botnet có những đặc trưng riêng về vòng đời hoạt động, phương thức lây nhiễm và các hành vi độc hại.

1.1.4. Phân loại Botnet

Botnet có thể được phân loại theo các tiêu chí như: Giao thức, thiết bị lây nhiễm hoặc kiến trúc.

1.2. Kỹ thuật phát hiện Botnet

Có kỹ thuật chính được sử dụng để phát hiện Botnet:

- (1) Các kỹ thuật dựa trên honeynet (mạng bẫy tin tặc).
- (2) Các kỹ thuật dựa trên hệ thống phát hiện xâm nhập:
  + Phát hiện Botnet dựa trên sự bất thường.
  + Phát hiện Botnet dựa trên chữ ký.
  + Phát hiện Botnet dựa trên tên miền.

1.3. Bài toán DGA Botnet

1.3.1. Khái quát về DGA Botnet

DGA Botnet là khái niệm chỉ một dạng Botnet được triển khai theo mô hình Client-Server. Trong đó, các Bot đóng vai trò là Client sẽ liên kết trở lại máy chủ C&C - đóng vai trò là Server - thông qua các tên miền DNS được sinh một cách tự động và được thống nhất trước đó. Các thuật toán sinh tên miền được thiết kế để cố gắng sinh ra các tên miền có thể qua mặt được các hệ thống bảo mật.

1.3.2. Bài toán phân lớp nhị phân trong DGA Botnet

Bài toán phân lớp nhị phân: Là bài toán với mục tiêu phát hiện các tên miền được sinh ra bởi DGA Botnet. Bộ dữ liệu gồm có hai nhãn là 0 và 1.

1.3.3. Bài toán phân lớp đa lớp trong DGA Botnet

Bài toán phân lớp đa lớp: Là bài toán nhằm mục tiêu phát hiện họ/chủng loại của DGA Botnet, với đầu vào là các tên miền đã được gán nhãn là DGA Botnet. Bộ dữ liệu gồm có n nhãn, tương ứng với n họ DGA Botnet được xem xét.

1.3.4. Phân biệt với bài toán phát hiện URL giả mạo

Bài toán phát hiện DGA Botnet có sự khác biệt so với bài toán phát hiện các Uniform Resource Locator - URL giả mạo (Bảng 1.3).

Bảng 1.3. So sánh bài toán phát hiện Website giả mạo và bài toán DGA Botnet

	Đầu vào	**Mục tiêu**	**Nhãn phân lớp nhị phân**	**Nhãn phân lớp đa lớp**	**Thuật toán DGA**
Phát hiện Website giả mạo	URLs / HTML Code	Trang Web giả mạo	0: Website lành tính 1: Website giả mạo	NULL	Không
Phát hiện và phân loại DGA Botnet	Tên miền	Địa chỉ IP của C&C Server	0: Tên miền lành tính 1: Tên miền độc hại	n nhãn tương ứng với n họ DGA Botnet	Có

1.3.5. Bộ dữ liệu đánh giá cho bài toán DGA Botnet

NCS lựa chọn 04 bộ dữ liệu được xem là phù hợp nhất cho các đánh giá thuật toán được trình bày ở các chương sau, bao gồm: Andrey Abakumov's DGA Repository [20], OSINT DGA feed [23], UMUDGA Dataset [18] và 360NetLab Dataset [24] (Bảng 1.4).

Bảng 1.4. Mô tả về 04 bộ dữ liệu được sử dụng trong các đánh giá

Bộ dữ liệu	**Phân lớp nhị phân**	**Phân lớp đa lớp**	**Số mẫu lành tính**	**Số mẫu DGA Botnet**	**Số họ DGA Botnet**
AADR	X	X	1.000.000	801.667	08
OSINT	X		1.000.000	495.186	
UMUDGA	X	X	1.000.000	500.000	50
360NetLab	X		1.000.000	1.513.524	

1.3.6. Thông số đánh giá bài toán

NCS đánh giá qua các tham số gồm Accuracy, Precision, Recall và F_1-score.

1.3.7. Ý nghĩa bài toán DGA Botnet

Botnet ngày càng phát triển, các con Bot được thiết kế ngày càng tinh vi và có khả năng phá hoại lớn hơn.

Vận dụng cơ chế hoạt động của DGA Botnet có thể mang lại một giải pháp hiệu quả và mang lại nhiều ưu điểm như không đòi hỏi quá nhiều năng lực thu thập và xử lý của hệ thống; việc phát hiện hoạt động của DGA Botnet có thể diễn kể ra khi chúng đã lây nhiễm vào thiết bị.

Các giải pháp để giải quyết bài toán phân lớp nhị phân và phân lớp đa lớp trong DGA Botnet có thể được áp dụng vào các giải pháp bảo mật như Firewall, IDS, NGFW hay UTM.

1.4. Một số nghiên cứu giải quyết bài toán DGA Botnet

Một số hướng tiếp cận đã được đề xuất để phát hiện Botnet nói chung và DGA Botnet nói riêng, bao gồm các phương pháp phát hiện dựa trên chữ ký, phát hiện dựa trên sự bất thường, sử dụng các thuật toán học máy và học sâu.

1.4.1. Hướng tiếp cận sử dụng các kỹ thuật phân tích DNS

Trong nghiên cứu [8], tác giả Kwon và cộng sự giới thiệu giải pháp PsyBoG, dùng để phát hiện hành vi độc hại dựa trên phân tích một lượng lớn lưu lượng DNS. Giải pháp PsyBoG có các ưu điểm bao gồm: (1) Phát hiện các mạng Botnet có khả năng ẩn mình tinh vi; (2) Cho phép xử lý các truy vấn DNS trên quy mô lớn và (3) Có khả năng phát hiện các nhóm máy chủ có hành vi độc hại.

Wang và cộng sự phát hiện Botnet dựa trên phân tích tên miền [9]. Họ đề xuất một sơ đồ phát hiện Botnet dựa trên DGA, gọi là DBod. Giải pháp này dựa trên phân tích hành vi truy vấn DNS.

Chowdhury và cộng sự [42] đề xuất một phương pháp phát hiện Botnet mới dựa trên đặc điểm cấu trúc liên kết của các nút trong đồ thị, được đánh giá trên bộ dữ liệu CTU-13 [43]. Phương pháp đề xuất có thể phát hiện các con Bot một cách hiệu quả, dù cho các hành vi của chúng là khác nhau.

Trong nghiên cứu [10], Bisio và cộng sự đã trình bày báo cáo về thuật toán phát hiện DGA Botnet dựa trên một Single Network Monitoring. Bộ dữ liệu thử nghiệm bao gồm 40 họ DGA Botnet khác nhau, với khả năng phát hiện hầu hết các họ DGA Botnet với độ chính xác cao từ 92,67% trở lên, duy nhất một trường hợp đạt 88,85%.

Wang và cộng sự giới thiệu một cách tiếp cận bao gồm: (1) Phát hiện sự hiện diện của Botnet và (2) Xác định các nút bị lây nhiễm [44]. Tuy nhiên, các kết quả đánh giá chưa được thảo luận một cách kỹ lưỡng.

Trung và cộng sự [11] trình bày các nghiên cứu mở rộng của Botnet trên các thiết bị IoT. Nhóm tác giả đề xuất phương pháp phát hiện IoT Botnet dựa trên việc trích xuất các thuộc tính từ đồ thị PSI (PGS-Graph). Giải pháp này có thể khắc phục được vấn đề đa kiến trúc trong thiết bị IoT, đồng thời giảm thiểu sự phức tạp tính toán. Kết quả thử nghiệm cho thấy, giải pháp đề xuất đạt độ chính xác 98,7%.

1.4.2. Hướng tiếp cận dựa trên học máy

Hiếu và cộng sự đã đánh giá độ hiệu quả các các thuật toán học máy có giám sát trong việc phát hiện DGA Botnet [12]. Với dữ liệu đầu vào là các tiên miền, nhóm nghiên cứu xây dựng các mô hình bao gồm: Hidden Markov, C4.5 Decision Tree, Extreme Learing Machine. Đồng thời cũng thí nghiệm trên các mô hình SVM, Recurrent SVM, CNN kết hợp LSTM và Bidirectional LSTM. Các bộ phân loại hoạt động hiệu quả với bài toán phân lớp nhị phân nhưng còn hạn chế với bài toán phân lớp đa lớp.

Khan và cộng sự xem xét tới việc phát hiện các mạng Botnet ngang hàng (Peer-to-Peer) [13]. Họ đề xuất một phương pháp phân loại lưu lượng đa lớp bằng cách áp dụng các mô hình học máy. Độ chính xác trung bình đạt được là 98,7%. Các thí nghiệm được thực hiện trên bộ dữ liệu CTU-13 và ISOT Dataset.

Zago và cộng sự trình bày một nghiên cứu về bộ dữ liệu mới cho phát hiện DGA Botnet. Nhóm nghiên cứu tổng hợp và xây dựng thành bộ dữ liệu mới với tên gọi là UMUDGA Dataset [18][21]. Bộ dữ liệu này bao gồm 1.000.000 tên miền lành tính kết hợp với 50 họ DGA Botnet khác nhau.

Xuân và cộng sự [14] đã đề xuất những cải tiến cho mô hình học máy. Họ đề xuất sử dụng các thuộc tính mới và áp dụng trên thuật toán rừng ngẫu nhiên. Mô hình có thể lệ cảnh báo sai dưới 3,02% và điểm F1-score đạt 97,03% trong các đánh giá.

Alauthman [50] và cộng sự đề xuất một cơ chế giúp giảm thiểu các lưu lượng phức tạp, tích hợp với một kỹ thuật học tăng cường. Kết quả thử nghiệm cho thấy phương pháp mới đạt tỉ lệ phát hiện đúng là 98,3%, tỉ lệ dương tính giả thấp với 0,012%. Nghiên cứu được đánh giá trên sự tổng hợp của ba bộ dữ liệu gồm ISOT Dataset, P2P Botnet và Information Security Centre of Excellence Dataset.

1.4.3. Hướng tiếp cận dựa trên học sâu

Trong hướng tiếp cận dựa trên học sâu, Đức và cộng sự [15] đã sử dụng mạng bộ nhớ ngắn hạn dài LSTM để giải quyết cả hai bài toán DGA Botnet. Nhóm nghiên cứu đề xuất một thuật toán mới với tên gọi là LSTM.MI. Kết quả thử nghiệm cho thấy thuật toán đề xuất giúp nâng cao ít nhất 7% độ chính xác so với mô hình LSTM truyền thống và đạt độ chính xác cao trong bài toán phân lớp nhị phân với F1-score đạt 98,49%, đồng thời có khả năng nhận ra 05 họ Botnet bổ sung.

Curtin và cộng sự đã sử dụng mạng RNN để phát hiện và phân loại DGA Botnet [16]. Nhóm nghiên cứu đã đề xuất một khái niệm mới, gọi là thang điểm Smashword. Thực nghiệm cho thấy mô hình mới có tiềm năng ứng dụng cao cải tiến được độ chính xác hơn so với các mô hình trước đó.

Vinayakumar và nhóm cộng sự nghiên cứu bài toán phát hiện tên miền độc hại được sinh bởi Botnet hay thư điện tử, URL độc hại [17]. Bộ dữ liệu để đánh giá bao gồm các tên miền lành tính và độc hại được thu thập từ OpenDNS, Alexa và OSINT Feeds. Kết quả so sánh thấy mô hình CNN-LSTM là hiệu quả nhất với F_1-score đạt 96,3% cho bài toán phân lớp nhị phân.

1.5. Kết luận Chương 1

Trong Chương 1, NCS trình bày tổng quan chung về Botnet nói chung, bài toán DGA Botnet nói riêng và các nghiên cứu liên quan. Trong đó, xác phạm vi nghiên cứu của luận án là bài toán DGA Botnet, tập trung vào bài toán phân loại . Chương 1 cũng trình bày sự khác nhau giữa bài toán phát hiện DGA Botnet với bài toán phát hiện URL độc hại và ý nghĩa của bài toán này.

Trong các chương tiếp theo, NCS trình bày các kết quả nghiên cứu, đánh giá và đề xuất giải pháp dựa trên học sâu để giải quyết hai bài toán gồm phân lớp nhị phân và phân lớp đa lớp, với ý nghĩa phát hiện và phân loại DGA Botnet.

Một phần kết quả nghiên cứu được trình bày tại Chương 1 được công bố tại [CT2] [CT6] trong danh mục công trình của tác giả.

CHƯƠNG 2. ĐÁNH GIÁ GIẢI PHÁP PHÁT HIỆN DGA BOTNET SỬ DỤNG LÝ THUYẾT TẬP MỜ VÀ HỌC MÁY

Trong chương 2, NCS tiếp cận để giải quyết bài toán phân lớp nhị phân - phát hiện DGA Botnet sử dụng lý thuyết tập mờ và học máy. NCS cũng đề xuất hai mô hình học máy kết hợp là VEA và HEA để nâng cao độ chính xác so với các mô hình đơn. Mục tiêu của chương này là đánh giá hiệu quả của hai hướng tiếp cận gồm dựa trên lý thuyết tập mờ và học máy trong bài toán phân lớp nhị phân.

2.1. Phát hiện DGA Botnet dựa trên lý thuyết tập mờ

2.1.1. Cơ sở thuật toán phân cụm mờ

Thuật toán phân cụm trên tập mờ của Zadeh là Fuzzy C-Means - FCM đưa ra bởi Bezdek và cho đến nay đã được ứng dụng trong nhiều lĩnh vực khác nhau.

Thuật toán phân cụm NCM trên tập Neutrosophic Set được đề xuất bởi Gou và cộng sự [57]. Cho X là một tập khác rỗng, với một phần tử của X ký hiệu là $x \in$ X, tập mờ trung lập A xác định trên không gian X được đặc trưng bởi ba hàm số:

- Hàm T_A(x) đo độ thuộc chỉ ra rằng sự kiện x sẽ xảy ra;
- Hàm I_A(x) đo độ trung lập, tức là không có ý kiến gì về việc sự kiện x có xảy ra hay không;
- Hàm F_A(x) đo độ không thuộc, tin rằng sự kiện x sẽ không xảy ra.
- Hàm mục tiêu:

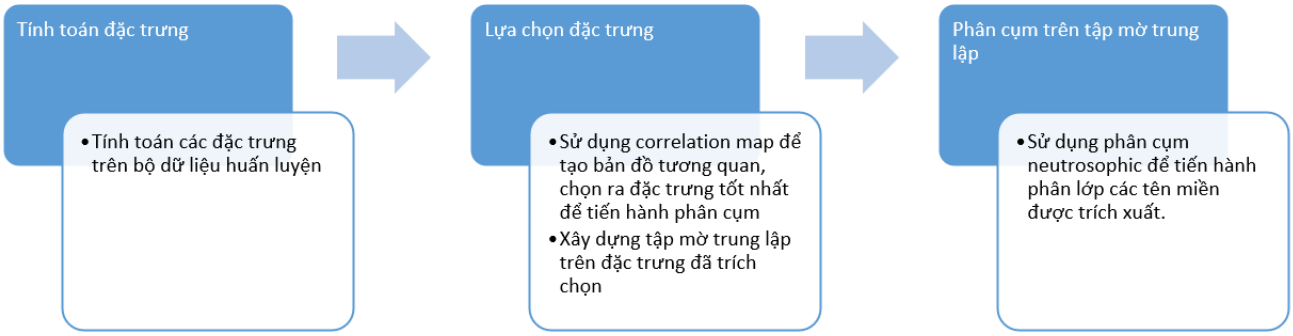
$$J_{NCM}(T,I,F,c) = \sum_{i=1}^{N}\sum_{j=1}^{C}(\omega_1 T_{ij})^m \|x_i - c_j\|^2 + \sum_{i=1}^{N}(\omega_2 I_i)^m \|x_i - \bar{c}_{i_{max}}\|^2 + \delta^2 \sum_{i=1}^{N}(\omega_3 F_i)^m \quad (2.1)$$

Thuật toán Neutrosophic C-Means Clustering được tóm tắt như sau:

Thuật toán: $NCM(X, \varepsilon)$
Input X, ε
Output k
$init(T^{(0)}, I^{(0)}, F^{(0)})$
$init(C, m, \varepsilon, \delta, \omega_1, \omega_2, \omega_3)$
do:
$calculate(c_i^{(k)})$:
$calculate(\bar{c}_{i_{max}})$
$update(T^{(k+1)})$
$update(I^{(k+1)})$
$update(F^{(k+1)})$
while $\left|T_{ii}^{(k+1)} - T_{ii}^{(k)}\right| > \varepsilon$
$TM = [T, I, F]$ $với$ $x(i) \epsilon$ k^{th}
$k = argmax(TM_{ij})$ $với$ $j = 1, 2, \ldots, C+2$
return k

2.1.2. Thuật toán phát hiện DGA Botnet với NCM

Các giai đoạn trong giải pháp phân cụm mờ NCM để phát hiện DGA Botnet được thể hiện tại Hình 2.1.



Hình 2.1. Mô hình áp dụng thuật toán NCM để phát hiện DGA Botnet

NCS đề xuất các đặc trưng dựa trên tên miền và sử dụng hệ số tương quan Pearson để lựa chọn các đặc trưng ảnh hưởng nhất. Kết quả trích chọn đặc trưng được cho tại Bảng 2.4.

Bảng 2.4. Các đặc trưng có ảnh hưởng cao nhất trong các bộ dữ liệu

STT	AADR	360NetLab	OSINT	UMUDGA
1	CIPA	RCC	DNL	RCC
2	HVTLD	VR	NoDistinct	VR
3	VR	Entropy	VR	Entropy
4	RCC	NoDistinct	RCC	NoDistinct
5	NoDistinct	DNL	Entropy	DNL
6	Entropy	NR	NR	NR
7	RCN	CD	CD	CD

2.1.3. Đánh giá và thảo luận

Lần lượt đánh giá trên các bộ dữ liệu AADR, 360NetLab, OSINT và UMUDGA. Kết quả được thể hiện ở Bảng 2.5:

Bảng 2.5. Kết quả phân lớp nhị phân của thuật toán NCM

	Precision	Recall	F_1-Score
AADR	0,87	0,76	0,79
360NetLab	0,87	0,81	0,84
OSINT	0,77	0,61	0,54
UMUDGA	0,87	0,81	0,84

Nhìn chung, thuật toán NCM phân loại tên miền lành tính và độc hại, hoạt động tổng thể tốt nhất trên bộ dữ liệu 360NetLab và UMUDGA với F_1-score là 0,84; cho kết quả thấp nhất trên bộ dữ liệu OSINT khi chỉ đạt F_1-score là 0,54.

Thực hiện so sánh thuật toán NCM đề xuất với các thuật toán khác dựa trên lý thuyết mờ của Sahin, K-means, FCM, SVM, TSVM và FSVM, ta thu được các kết quả được thể hiện tại Bảng 2.6.

Bảng 2.6. So sánh NCM với một số thuật toán tương tự

Phương pháp	Precision	Recall	F_1-score
*NCM**	*0,85*	*0,75*	*0,75*
Sahin	0,48	0,80	0,60
K-means	0,74	0,86	0,79
FCM	0,76	0,83	0,79

SVM	0,82	0,72	0,77
TSVM	0,80	0,81	0,81
FSVM	0,83	0,73	0,78

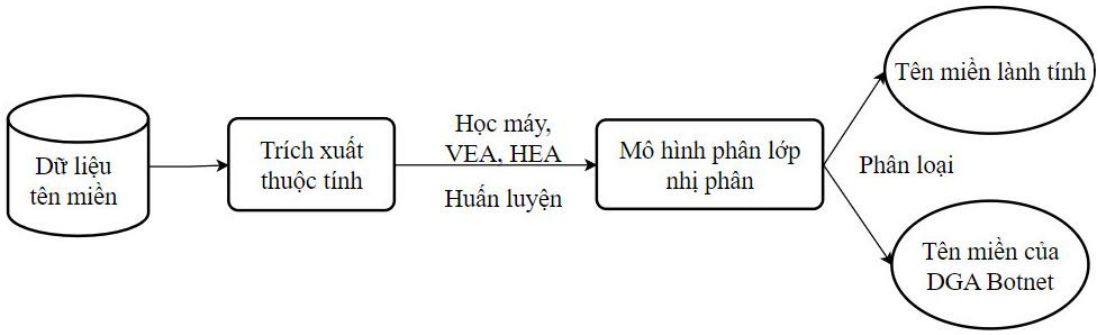
** Giá trị trung bình cộng khi đánh giá trên 04 bộ dữ liệu*

Mô hình sử dụng thuật toán NCM có sự cân bằng tốt giữa thời gian chạy và độ chính xác.

2.2. Phát hiện DGA Botnet dựa trên học máy

2.2.1. Mô hình đánh giá các thuật toán học máy

Các giai đoạn trong quá trình đánh giá thuật toán học máy đối với bài toán phân lớp nhị phân được thể hiện ở Hình 2.8.



Hình 2.8. Sơ đồ mô hình huấn luyện, đánh giá

NCS tiến hành thực nghiệm với các thuật toán bao gồm: Support Vector Machines (SVM); Logistic Regression (LR); Naive Bayes (NB); Neural Networks (NN); Decision Trees (DT); Random Forests (RF); k-Nearest Neighbour (kNN); Adaptive Boosting (AB). Đánh giá được thực hiện trên bộ dữ liệu UMUDGA Dataset.

2.2.2. Kết quả đánh giá và thảo luận

Bảng 2.8 thể hiện kết quả các độ đo Precision, Recall và F_1-score cho bài toán phân lớp nhị phân.

Bảng 2.8. Kết quả phát hiện DGA Botnet sử dụng học máy trên bộ dữ liệu UMUDGA

Method	**Precision**	**Recall**	**F_1-score**
Logistic Regression	0,92	0,97	0,97
Naive Bayes	0,98	0,84	0,90
Decision Tree	0,93	0,95	0,94
Neural Network	0,97	0,97	0,97
Support Vector Machine	0,97	0,97	0,97
Random Forrest	0,74	0,82	0,77
K-Nearest Neighbor	0,97	0,66	0,78
Adaptive Boosting	0,83	0,85	0,84

Hầu hết các thuật toán học máy đạt được độ chính xác cao trong bài toán phân lớp nhị phân, bao gồm LG, NN và SVM. Mô hình NN cho kết quả tổng thể cao nhất với F_1-score đạt 0,97 và mô hình có kết quả thấp nhất là RF với 0,78.

2.2.3. Mô hình học máy kết hợp

NCS đề xuất hai mô hình là VEA và HEA, là các bộ phân loại mạnh hơn, hoạt động dựa trên cơ chế học kết hợp. Kết quả thử nghiệm bài toán phân lớp nhị phân với hai mô hình VEA và HEA được thể hiện tại Bảng 2.9.

Bảng 2.9. Kết quả phát hiện DGA Botnet của mô hình VEA và HEA trên bộ dữ liệu UMUDGA

Thuật toán	Precision	Recall	F_1-score
Trung bình các mô hình đơn	0,92	0,88	0,89
Neural Network (mạnh nhất)	0,97	0,97	0,97
Random Forrest (yếu nhất)	0,74	0,82	0,77
VEA	0,98	0,99	0,98
HEA	0,97	0,97	0,97

Nhận xét, cả hai mô hình VEA và HEA đều cải thiện được độ chính xác so với giá trị trung bình của từng mô hình đơn lẻ.

Hạn chế của giải pháp NCM và học máy:

- Độ chính xác vẫn có thể tiếp tục được nâng cao;
- Đòi hỏi nhiều thời gian chạy huấn luyện bởi chạy trên CPU;
- Chưa phù hợp với bài toán phân lớp đa lớp.

2.3. Kết luận Chương 2

Trong chương 2, NCS trình bày các kết quả nghiên cứu và đánh giá cách tiếp cận sử dụng lý thuyết tập mờ và học máy cho bài toán phát hiện DGA Botnet.

- Với hướng tiếp cận sử dụng lý thuyết tập mờ, NCS đã đề xuất ứng dụng thuật toán phân cụm trên tập Neutrosophic Set - NCM để phát hiện DGA Botnet. NCS đã đưa vào các điều chỉnh từ thuật toán gốc để phù hợp với bài toán DGA Botnet. Đánh giá trên 04 bộ dữ liệu tiêu chuẩn cho thấy, thuật toán NCM có độ chính xác tương tự các giải pháp cùng hướng tiếp cận sử dụng lý thuyết mờ. Đồng thời, có ưu điểm là cải thiện đáng kể về Precision so với các bộ phân loại này. Thời gian tính toán của mô hình cũng có sự cân bằng giữa thời gian tiêu hoa và độ chính xác đạt được. Mô hình NCM cũng phát hiện các phần tử nhiễu hay trung tính để từ đó đưa ra các đề nghị kiểm tra bổ sung. Mặc dù có tốc độ thực thi nhanh, hạn chế của NCM là độ chính xác thấp hơn đáng kể so với các thuật toán học máy.

- Với hướng tiếp cận sử dụng học máy, các đánh giá trên 04 bộ dữ liệu tiêu chuẩn cho thấy, các mô hình học máy mang lại độ chính xác cao hơn rất đáng kể, với cao nhất là thuật toán Neural Network với F1-score đạt 0,97. Độ chính xác này được cải thiện khi sử dụng các mô hình kết hợp dựa trên cơ chế vote là VEA và HEA mà NCS đề xuất.

Cả hai hướng tiếp cận trên đều có những ưu điểm và hạn chế về mặt thời gian hoặc độ chính xác. Đồng thời, việc áp dụng vào bài toán phân lớp đa lớp là còn hạn chế. Cả hai yếu tố này vẫn có thể được giải quyết bằng các mô hình học sâu. Nội dung này được NCS đề xuất và trình bày ở Chương 3 của luận án.

Một phần kết quả nghiên cứu được trình bày tại Chương 2 được công bố tại [CT1] [CT3] trong danh mục công trình của tác giả.

CHƯƠNG 3. GIẢI PHÁP PHÁT HIỆN VÀ PHÂN LOẠI DGA BOTNET SỬ DỤNG KỸ THUẬT HỌC SÂU

Dựa trên các nghiên cứu và đánh giá trước đó, NCS đề xuất một giải pháp dựa trên kỹ thuật học sâu để nâng cao độ chính xác của bài toán phát hiện và phân loại DGA Botnet. NCS đề xuất hai mô hình học sâu mới là LA_Bin07 và LA_Mul07 trên cơ sở cải tiến so với mạng LSTM truyền thống. Các đánh giá cho thấy, mô hình đề xuất có độ chính xác được cải thiện đáng kể so với các nghiên cứu trước đó, đặc biệt là trong bài toán phân loại DGA Botnet.

3.1. Nền tảng kỹ thuật học sâu

3.1.1. Mạng Recurrent Neural Network

Recurrent Neural Network - RNN hay mạng nơ-ron hồi quy, được thiết kế để huấn luyện với các dữ liệu đầu vào có dạng chuỗi.

3.1.2. Mạng Long-Short Term Memory

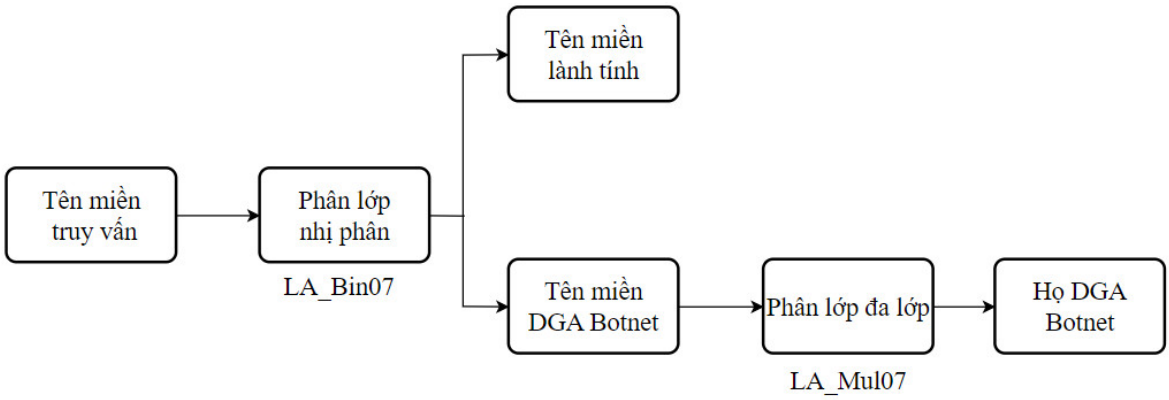
Mạng LSTM, được phát triển từ mạng RNN, với khả năng học được các phụ thuộc xa hơn so với RNN. LSTM lần đầu tiên được đề xuất vào năm 1996 bởi Hochreiter [64] và liên tục được cải tiến.

3.1.3. Cơ chế Attention

Attention là một thành phần gồm 3 nhân tố: Query, Key và Value. Chúng được đề xuất để giúp mô hình huấn luyện tập trung hơn vào một đặc điểm nào đó của dữ liệu.

3.2. Hai mô hình học sâu mới trong phát hiện và phân loại DGA Botnet

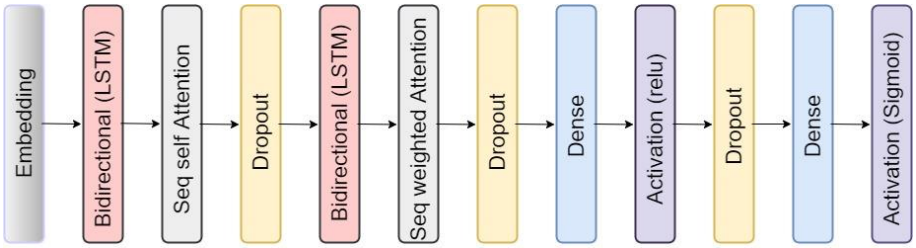
Giải pháp sử dụng hai mô hình LA_Bin07 và LA_Mul07 được thể hiện tại Hình 3.8.



Hình 3.8. Giải pháp phát hiện và phân loại DGA Botnet với hai mô hình học sâu mới LA_Bin07 và LA_Mul07

3.2.1. Mô hình LA_Bin07 cho phát hiện DGA Botnet

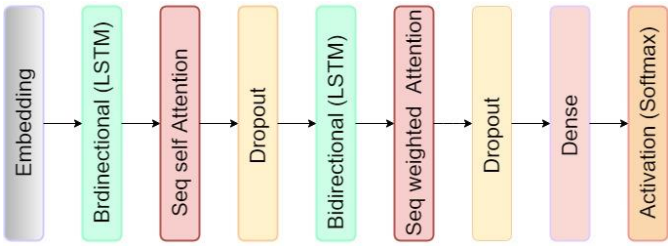
Mô hình LA_Bin07 được thiết kế theo dạng Sequence-to-Sequence, với cấu trúc được cho ở Hình 3.9:



Hình 3.9. Kiến trúc của mô hình LA_Bin07

3.2.2. Kiến trúc mô hình LA_Mul07 cho phân loại DGA Botnet

Cấu trúc các lớp trong mô hình LA_Mul07 được thể hiện ở Hình 3.10:



Hình 3.10. Kiến trúc của mô hình LA_Mul07

3.2.3. Cải tiến so với LSTM truyền thống

Bổ sung Attention giúp mô hình xác định các tham số cần học trọng tâm hơn các tham số khác, giúp nâng cao độ chính xác của mô hình so với LSTM gốc. Ngoài ra, thứ tự, trọng số và kích thước các lớp cũng được tối ưu để mô hình đạt hiệu quả cao nhất.

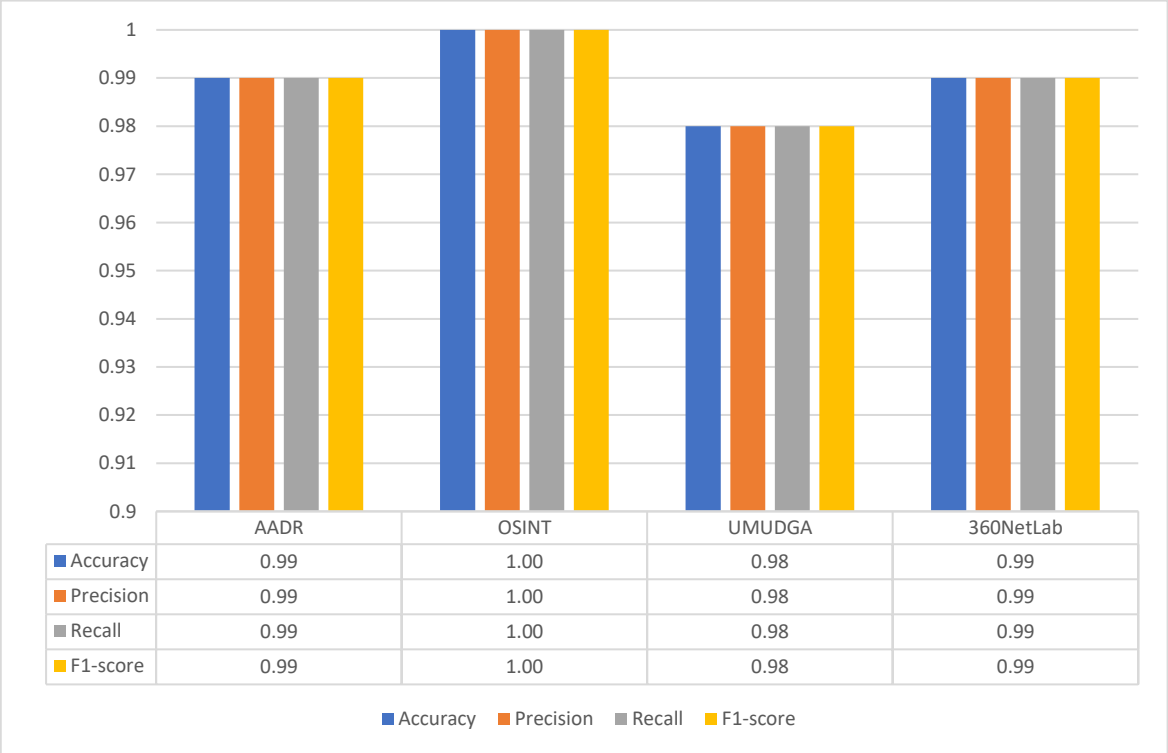
3.3. Đánh giá hai mô hình học sâu đề xuất

3.3.1. Bộ dữ liệu và môi trường đánh giá

NCS sử dụng cả 04 bộ dữ liệu cho bài toán BinaryClassification và sử dụng 02 bộ dữ liệu gồm Andrey Abakymov's DGA Repository và UMUDGA Dataset cho bài toán MultiClassification.

3.3.2. Đánh giá mô hình LA_Bin07 cho bài toán phát hiện DGA Botnet

Kết quả đánh giá mô hình LA_Bin07 qua các tham số Accuracy, Precision, Recall và F_1-Score được cho ở Hình 3.13.



	AADR	OSINT	UMUDGA	360NetLab
Accuracy	0.99	1.00	0.98	0.99
Precision	0.99	1.00	0.98	0.99
Recall	0.99	1.00	0.98	0.99
F1-score	0.99	1.00	0.98	0.99

Hình 3.13. Kết quả đánh giá của mô hình LA_Bin07 cho bài toán phân lớp nhị phân trên 04 bộ dữ liệu tiêu chuẩn

Mô hình LA_Bin07 có độ chính xác rất cao, với Accuracy đều đạt từ 0,98 trở lên trên cả 04 bộ dữ liệu được đánh giá. Đặc biệt, mô hình LA_Bin07 dự đoán chính xác 1,00 trên bộ dữ liệu OSINT.

3.3.3. Đánh giá mô hình LA_Mul07 cho bài toán phân loại DGA Botnet

Mô hình LA_Mul07 được NCS đánh giá lần lượt trên hai bộ dữ liệu là AADR và UMUDGA, bởi chúng được gán nhãn các họ DGA Botnet.

Kết quả đánh giá trên bộ AADR được thể hiện tại Bảng 3.3.

Bảng 3.3. Kết quả đánh giá mô hình LA_Mul07 trên bộ dữ liệu AADR

STT	DGA Botnet	Precision	Recall	F_1-Score
1	cryptolocker	1,00	0,98	0,99
2	zeus	1,00	1,00	1,00
3	pushdo	1,00	1,00	1,00
4	rovnix	1,00	1,00	1,00
5	tinba	0,98	1,00	0,99
6	conficker	1,00	1,00	1,00
7	matsnu	1,00	1,00	1,00
8	ramdo	1,00	1,00	1,00
Avg Accuracy		***1,00***		

Kết quả đánh giá trên bộ UMUDGA được cho tại Bảng 3.4.

Bảng 3.4. Kết quả đánh giá mô hình LA_Mul07 trên bộ dữ liệu UMUDGA

STT	DGA Botnet	Pre	Re	F_1	STT	DGA Botnet	Pre	Re	F_1
1	alureon	0,45	0,92	0,60	26.	pizd	0,97	0,86	0,91
2	banjori	0,99	1,00	1,00	27.	proslikefan	0,82	0,65	0,73
3	bedep	0,96	0,47	0,63	28.	pushdo	0,99	0,99	0,99
4	ccleaner	1,00	1,00	1,00	29.	pykspa	0,39	0,57	0,47
5	china	1,00	0,99	1,00	30,	pykspa_noise	0,35	0,16	0,22
6	corebot	1,00	1,00	1,00	31.	qadars	0,99	0,99	0,99
7	cryptoloker	0,70	0,66	0,68	32.	qakbot	0,84	0,55	0,67
8	dircrypt	0,52	0,42	0,47	33.	ramdo	1,00	1,00	1,00
9	dyre	1,00	1,00	1,00	34.	ramnit	0,44	0,66	0,52
10	fobber_v1	0,88	1,00	0,93	35.	ranbyus_v1	0,76	0,98	0,86
11	fobber_v2	0,48	0,08	0,14	36.	ranbyus_v2	0,76	0,88	0,82
12	gozi_gpl	0,96	0,99	0,98	37	rovnix	0,97	0,94	0,95
13	gozi_luther	0,97	0,95	0,96	38	shiotob	1,00	0,90	0,95
14	gozi_nase	0,89	0,97	0,93	39	simda	1,00	1,00	1,00
15	gozi_rfc4343	0,91	0,98	0,90	40	aaron	1,00	1,00	1,00
16	kraken_v1	0,72	0,96	0,83	41	suppobox_1	0,87	0,97	0,92
17	kraken_v2	0,82	0,41	0,55	42	suppobox_2	0,98	1,00	0,99
18	locky	0,84	0,62	0,71	43	suppobox_3	0,99	1,00	1,00
19	matsnu	0,98	0,94	0,96	44	symmi	1,00	1,00	1,00
20	murofet_v1	0,99	1,00	1,00	45	tempedreve	0,58	0,86	0,69
21	murofet_v2	0,94	0,96	0,95	46	tinba	0,77	0,97	0,86
22	murofet_v3	1,00	1,00	1,00	47	vawtrak_v1	1,00	1,00	1,00

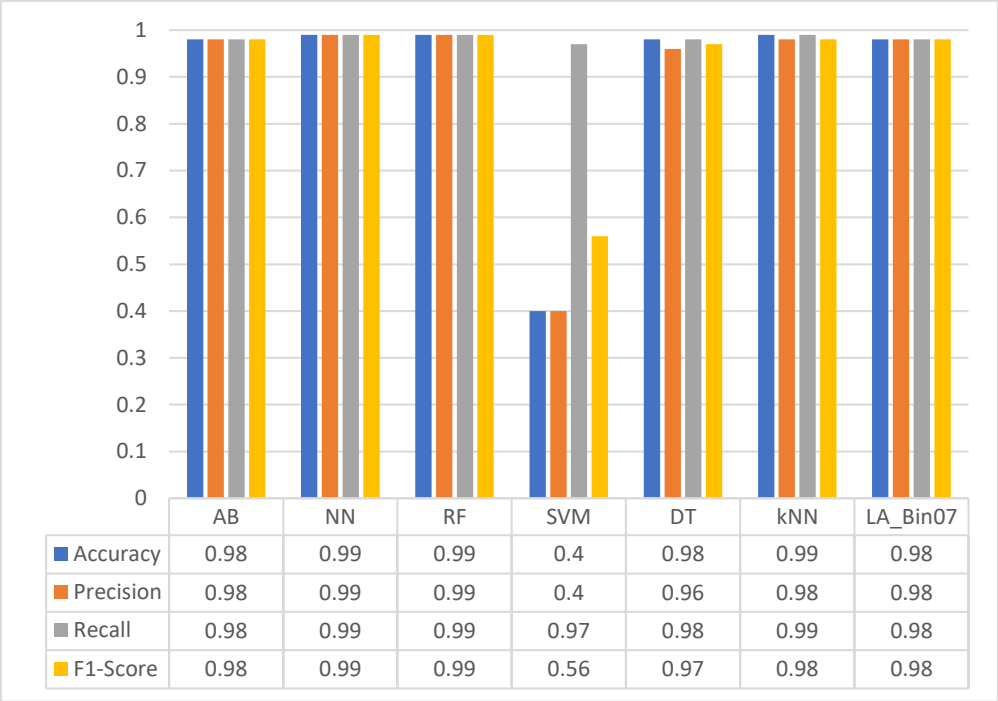
23	necurs	0,99	0,80	0,89	48	vawtrak_v2	0,99	1.00	1.00
24	maim	0,95	0,94	0,95	49	vawtrak_v3	1.00	1.00	1.00
25	padcrypt	1.00	1.00	1.00	50	zeus_newgoz	1.00	1.00	1.00
Avg Accuracy					***0,86***				

Mô hình LA_Mul07 có độ chính xác cao trong phân lớp các họ DGA Botnet, kể cả trong trường hợp số lượng họ DGA Botnet cần phân lớp là nhiều, cụ thể đạt 1,00 trên bộ dữ liệu AADR và 0,86 trên bộ dữ liệu UMUDGA.

3.4. Đánh giá với các nghiên cứu liên quan

3.4.1. Đánh giá chung trên bộ dữ liệu UMUDGA

Đối với nhiệm vụ phân lớp nhị phân, kết quả so sánh được thể hiện tại Hình 3.21.

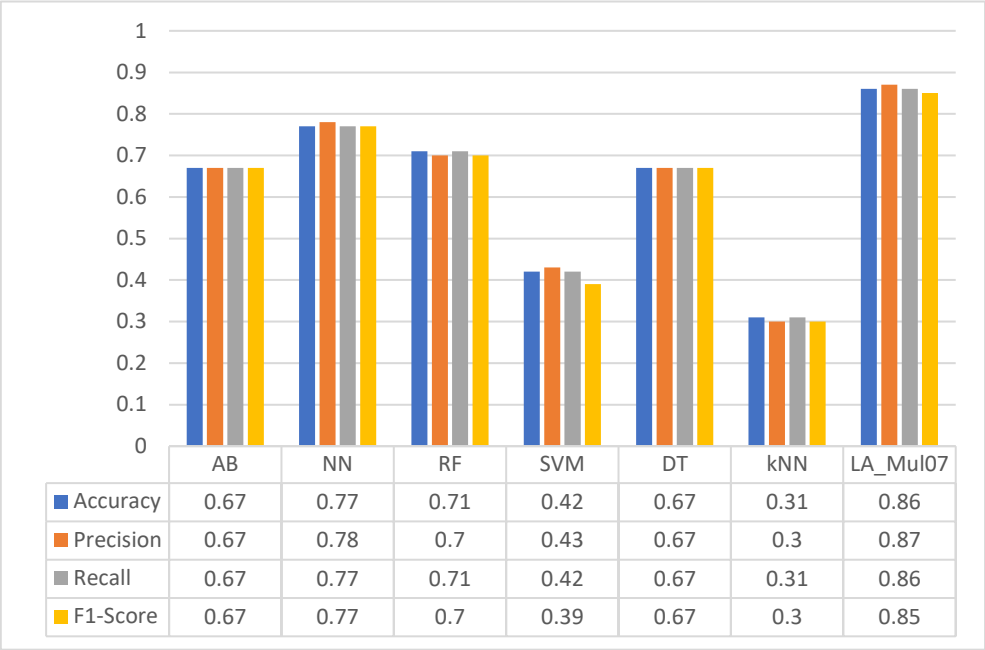


	AB	NN	RF	SVM	DT	kNN	LA_Bin07
Accuracy	0.98	0.99	0.99	0.4	0.98	0.99	0.98
Precision	0.98	0.99	0.99	0.4	0.96	0.98	0.98
Recall	0.98	0.99	0.99	0.97	0.98	0.99	0.98
F1-Score	0.98	0.99	0.99	0.56	0.97	0.98	0.98

Hình 3.21. So sánh bộ phân loại LA_Bin07 với các thuật toán học máy trên bộ dữ liệu UMUDGA

Kết quả cho thấy mô hình LA_Bin07 đề độ chính xác tốt hơn rất nhiều so với mô hình SVM. Đồng thời, cho kết quả gần như tương đương với các mô hình AB, NN, RF, DT hay kNN.

Đối với nhiệm vụ phân lớp đa lớp, kết quả so sánh được tổng hợp và thể hiện tại Hình 3.22.



	AB	NN	RF	SVM	DT	kNN	LA_Mul07
Accuracy	0.67	0.77	0.71	0.42	0.67	0.31	0.86
Precision	0.67	0.78	0.7	0.43	0.67	0.3	0.87
Recall	0.67	0.77	0.71	0.42	0.67	0.31	0.86
F1-Score	0.67	0.77	0.7	0.39	0.67	0.3	0.85

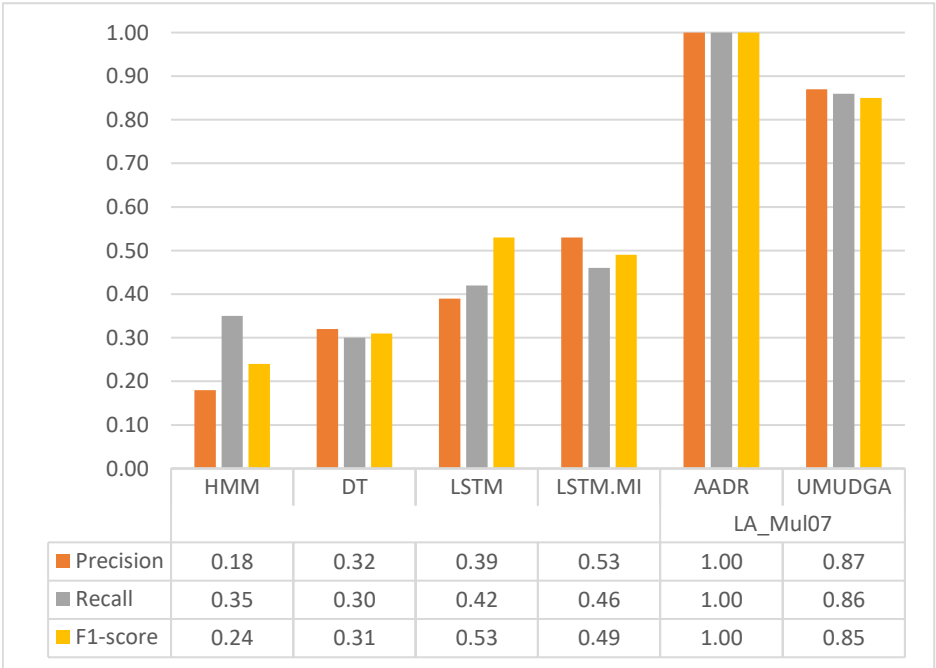
Hình 3.22. So sánh bộ phân loại LA_Mul07 với các thuật toán học máy trên bộ dữ liệu UMUDGA

Đối với bài toán phân lớp đa lớp, mô hình LA_Mul07 cho Accuracy cao hơn nhiều so với các mô hình học máy còn lại.

3.4.2. Đánh giá với một số mô hình học sâu khác

NCS cũng đồng thời đánh giá với một số kiến trúc học sâu khác mà NCS xây dựng trên cơ sở CNN và LSTM bao gồm: Basic CNN, Basic LSTM, Bi-LSTM và CNN-LSTM. Kết quả cho thấy, mô hình LA_Bin07 và LA_Mul07 đạt kết quả tốt nhất trong các mô hình thử nghiệm.

3.4.3. Đánh giá với một số nghiên cứu khác trong bài toán phân lớp đa lớp



	HMM	DT	LSTM	LSTM.MI	AADR	UMUDGA
					LA_Mul07	
Precision	0.18	0.32	0.39	0.53	1.00	0.87
Recall	0.35	0.30	0.42	0.46	1.00	0.86
F1-score	0.24	0.31	0.53	0.49	1.00	0.85

Hình 3.23. Kết quả so sánh mô hình LA_Mul07 với các mô hình khác

Mô hình LA_Mul07 cũng cho kết quả tốt hơn so với các mô hình học sâu được đề xuất trước đó.

3.5. Kết luận Chương 3

Trong Chương 3, NCS đã trình bày các kết quả nghiên cứu về cải tiến mô hình học sâu để giải quyết bài toán DGA Botnet. NCS đề xuất hai mô hình học sâu mới là LA_Bin07 và LA_Mul07 lần lượt giải quyết bài toán phát hiện và phân loại DGA Botnet. Mô hình mới được cải tiến trên cơ sở mạng LSTM truyền thống. Hai mô hình được đánh giá một cách đầy đủ trên 04 bộ dữ liệu bao gồm: Andrey Abakumov's DGA Repository, OSINT DGA feed, UMUDGA Dataset và 360NetLab Dataset.

Các đánh giá cho thấy, mô hình LA_Bin07 có độ chính xác rất cao, đạt từ 0,98 trên bộ UMUDGA và cho đến 1,00 trên bộ dữ liệu OSINT. Mô hình LA_Mul07 cho khả năng phân loại các họ DGA Botnet cao, cải thiện đáng kể so với các mô hình trước đó, với Accuracy lần lượt đạt 1,00 và 0,86 trên hai bộ dữ liệu là AADR và UMUDGA.

Việc giải quyết bài toán DGA Botnet mang lại nhiều ý nghĩa trong vấn đề đảm bảo an ninh mạng, đặc biệt là bài toán phân loại DGA Botnet. Thứ nhất, hướng tiếp cận này có thể nhanh chóng đưa ra các cảnh báo phát hiện và phân loại về DGA Botnet khi được tích hợp trên các thiết bị Firewall/IDS. Thứ hai, giải pháp này đòi hỏi ít tài nguyên tính toán hơn so với các giải pháp phân tích gói tin truyền thống và tận dụng được năng lực tính toán của GPU. Thứ ba, giải pháp có thể mở rộng áp dụng cho mã độc, phần mềm độc hại, phần mềm gián điệp có cơ chế truy vấn tên miền tương tự. Cuối cùng, module phát hiện tên miền độc hại hoàn toàn có thể được tích hợp trên các giải pháp bảo mật tiên tiến, hiện đại như tường lửa thế hệ mới, giải pháp an ninh hợp nhất.

Một phần kết quả nghiên cứu được trình bày tại Chương 3 được công bố tại [CT4] trong danh mục công trình của tác giả.

CHƯƠNG 4. BỘ DỮ LIỆU MỚI UTL_DGA22 CHUYÊN DÙNG CHO BÀI TOÁN DGA BOTNET

Trong chương 4, NCS đề xuất một bộ dữ liệu mới chuyên dùng cho bài toán DGA Botnet là UTL_DGA22. Bộ dữ liệu này kế thừa kết quả của những bộ dữ liệu trước đó, đồng thời có thêm các điểm mới, cải tiến bao gồm: Bổ sung các họ DGA Botnet mới, chuẩn hóa dữ liệu và gán nhãn, đề xuất và trích chọn sẵn các thuộc tính mới, tài liệu mô tả chi tiết. NCS cũng đánh giá các giải pháp đề xuất ở Chương 2 và Chương 3 bao gồm NCM, VEA, HEA, LA_Mul07 và LA_Bin07 trên bộ dữ liệu mới cho kết quả tốt. Bộ dữ liệu UTL_DGA22 được kỳ vọng sẽ là một cơ sở tin cậy, công khai, khách quan, đầy đủ cho các nhà khoa học thử nghiệm, so sánh, đánh giá các giải pháp của họ trong thời gian tới.

4.1. Đặt vấn đề bộ dữ liệu DGA Botnet

4.1.1. Khái quát vấn đề

Giải pháp đề xuất trong các nghiên cứu trước đó thường được đánh giá trên những bộ dữ liệu do nhóm nghiên cứu thu thập vào những thời điểm khác nhau, số lượng mẫu không đồng đều, tính công bố rộng rãi không cao và thường không thuận tiện cho việc đối sánh.

4.1.2. Bộ dữ liệu về Botnet nói chung

Một số bộ dữ liệu về Botnet nói chung như CTU-13 [86], UGR16 [87], DreLAB [88], UNSW-NB15 [89], ISCX-Bot-2014 [40]. Cả 5 bộ dữ liệu ở trên đều không được thiết kế để đánh giá chuyên biệt cho bài toán DGA Botnet bởi chúng thiếu tên miền của các họ DGA Botnet và nhãn tương ứng.

4.1.3. Bộ dữ liệu chuyên dùng về DGA Botnet

Bảng 4.3. tóm tắt các đặc điểm chính của các bộ dữ liệu phổ biến về DGA Botnet đã nêu ở trên.

Bảng 4.3. Đặc điểm chính của các bộ dữ liệu phổ biến hiện nay về DGA Botnet

STT	Bộ dữ liệu	Viết tắt	Số lượng tên miền lành tính	Số lượng tên miền của DGA Botnet	Sọ họ DGA Botnet	Định dạng	Năm công bố
1	Andrey Abakumov's DGA Repository [20]	AADR	1.000.000	801.667	08	txt	2016
2	Johannes Bader's Domain Generation Algorithms Repository [25]	JBR	Null	N/A	48	txt	2018
3	Alexa Top 1 milion domains [47]	AT1D	1.000.000	0	0	csv	2019
4	Botnet DGA Dataset [92]	BDD	1.000.000	1.803.333	10	csv	2020
5	UMUDGA Dataset [18]	UMU	1.000.000	Trên 30.000.000	50	arff, csv, txt	2020
6	DGArchive by Fraunhofer FKIE [22]	DFF	Null	N/A	86	csv	2020
7	OSINT DGA feed [23]	OSINT	1.000.000	495.186	0	txt	2021

8	360NetLab Dataset [24]	360NL	0	Không cố định	Không cố định	txt	2021
9	The Majestic Million [96]	TMM	1.000.000	0	0	csv	2021

4.1.4. Đặt vấn đề nghiên cứu

Có sự khác nhau về cấu trúc và mục đích giữa các bộ dữ liệu cho Botnet nói chung và bộ dữ liệu cho DGA Botnet nói riêng. NCS phân nhóm và thể hiện chi tiết tại Bảng 4.4.

Bảng 4.4. Đánh giá về đặc điểm các nhóm bộ dữ liệu cho Botnet

Bộ dữ liệu	**Nhóm**	**Phát hiện Botnet**	**Phát hiện DGA Botnet**	**Phát hiện tấn công lưu**	**Lưu lượng mạng**	**Định dạng**		
						PCAP	**SCV**	**TXT**
CTU	Botnet	✓	×	×	✓	✓	✓	×
UGR	Botnet/IDS	✓	×	✓	✓	✓	✓	×
DLAB	Botnet	✓	×	×	✓	✓	✓	×
UNSW	Botnet/IDS	✓	×	✓	✓	✓	✓	×
ISCX	Botnet/IDS	✓	×	✓	✓	✓	✓	×
AADR	DGA Botnet	✓	✓	×	×	×	✓	✓
JBR	DGA Botnet	✓	✓	×	×	×	✓	✓
AT1D	DGA Botnet	✓	✓	×	×	×	✓	✓
BDD	DGA Botnet	✓	✓	×	×	×	✓	✓
UMU	DGA Botnet	✓	✓	×	×	×	✓	✓
DFF	DGA Botnet	✓	✓	×	×	×	✓	✓
OSINT	DGA Botnet	✓	✓	×	×	×	✓	✓
360NL	DGA Botnet	✓	✓	×	×	×	✓	✓
TMM	DGA Botnet	✓	✓	×	×	×	✓	✓
UTL	*DGA Botnet*	✓	✓	×	×	×	✓	✓

Các bộ dữ liệu Botnet/IDS có thể dùng cho đánh giá các kỹ thuật tấn công mạng khác nhưng không phù hợp để đánh giá cho các giải pháp DGA Botnet bởi chúng không được đánh nhãn tương ứng và số lượng mẫu cũng rất hạn chế. Trong khi bộ dữ liệu DGA Botnet chỉ dành cho đánh giá các giải pháp phát hiện DGA Botnet và được đánh nhãn tương ứng. Các bộ dữ liệu về Botnet chung bao gồm cả các lưu lượng mạng thô, nên có dung lượng lớn hơn rất nhiều (tính bằng GB) so với các bộ dữ liệu dành cho DGA Botnet đã được trích xuất các thông tin cần thiết (tính bằng MB). Cuối cùng, số lượng mẫu tên miền và số họ DGA Botnet trong các bộ dữ liệu chung về Botnet là rất ít so với các bộ dữ liệu về DGA Botnet chuyên dụng. Điều này tạo nên các hạn chế cho các đánh giá thuật toán. Các vấn đề trên cho thấy sự cần thiết của một số dữ liệu về DGA Botnet chuyên dụng.

Bảng 4.5 khái quát các ưu điểm và hạn chế của những bộ dữ liệu về DGA Botnet chuyên dụng.

Bảng 4.5. Khái quát ưu điểm và hạn chế của các bộ dữ liệu DGA Botnet hiện có và bộ dữ liệu UTL_DGA22 đề xuất

Bộ dữ liệu	**Phân lớp nhị phân**	**Phân lớp đa lớp**	**Tên miền gốc**	**Trích xuất thuộc tính**	**Công khai**	**Tài liệu**
AADR	✓	✓	✓	×	✓	N/A

JBR	✓	✓	✓	×	✓	✓
AT1D	✓	×	✓	×	✓	N/A
BDD	✓	×	×	✓	✓	✓
UMU	✓	✓	✓	✓	✓	✓
DFF	✓	✓	✓	×	✓	✓
OSINT	✓	×	✓	×	✓*	N/A
360NL	✓	✓	✓	×	✓	N/A
TMM	✓	×	✓	×	✓	N/A
UTL	✓	✓	✓	✓	✓	✓

Bộ dữ liệu UTL_DGA22 đề xuất sẽ đáp ứng đầy đủ các yêu cầu trên, và cập nhật thêm các dữ liệu mới, thuộc tính mới đã trích xuất.

4.2. Bộ dữ liệu UTL_DGA22 đề xuất

NCS giới thiệu một bộ dữ liệu mới về DGA Botnet là UTL_DGA22 [38].

4.2.1. Xây dựng bộ dữ liệu

Bộ dữ liệu UTL_DGA22 được xây dựng trải qua các bước như sau:

- Tổng hợp dữ liệu tên miền lành tính và tên miền DGA Botnet từ các nguồn đáng tin cậy.
- Tổng hợp và đánh giá, xác định các họ DGA Botnet độc lập.
- Tiến hành làm giàu cơ sở dữ liệu tên miền;
- Đề xuất 02 nhóm thuộc tính mới, bao gồm nhóm Base-Features và TF-IDF Features và rút trích đặc trưng.

4.2.2. Các thuộc tính đề xuất

NCS đề xuất 02 nhóm thuộc tính gồm nhóm thuộc tính trên tên miền (Base) và nhóm thuộc tính trên họ tên miền (TF-IDF),

4.2.3. Cấu trúc lưu trữ của bộ dữ liệu

Bộ dữ liệu DGA_UTL22 được cấu trúc gồm hai phần tương ứng với hai thư mục, gồm DGA_Botnets_Domains và DGA_Botnets_Features_Extraction.

4.3. Các họ DGA Botnet trong bộ dữ liệu UTL_DGA22

Bộ dữ liệu UTL_DGA22 bao gồm 76 họ DGA Botnet riêng biệt, được liệt kê ở Bảng 4.9.

4.4. Đánh giá bộ thuộc tính đề xuất

NCS sử dụng các thuộc tính đề xuất làm đầu vào của các mô hình học máy cơ bản cho bài toán phân lớp nhị phân và phân lớp đa lớp. Cả hai bộ thuộc tính Base và TF-IDF đều chứng tỏ sự phù hợp khi làm đầu vào cho các thuật toán học máy để giải quyết bài toán phân lớp nhị phân và phân lớp đa lớp.

4.5. Đánh giá các giải pháp đề xuất trên bộ dữ liệu UTL_DGA22

NCS tiến hành đánh giá các giải pháp đề xuất sử dụng bộ dữ liệu mới UTL_DGA22, bao gồm: Thuật toán NCM, học máy, VEA, HEA, LA_Bin07 và LA_Mul07. Kết quả được thể hiện tại Bảng 4.11, 4.13 và 4.14.

Bảng 4.11. Kết quả đánh giá thuật toán NCM trên bộ dữ liệu UTL_DGA22

Nhãn	**Precision**	**Recall**	**F_1-Score**
0	0,66	0,93	0,77

1	0,91	0,48	0,62
Avg*	*0,79*	*0,70*	*0,70*

Bảng 4.13. Kết quả đánh giá các thuật toán học máy đề xuất trên bộ dữ liệu UTL_DGA22

Method	**Acc.**	**Pre.**	**Re.**	**F_1.**	**Thời gian huấn luyện**	**Thời gian đánh giá**
Logistic Regression	0,96	0,97	0,95	0,96	75,59	0,04
Naive Bayes	0,90	0,91	0,86	0,86	1,34	0,14
Decision Tree	0,90	0,90	0,88	0,89	16.004,87	0,71
Neural Network	0,97	0,97	0,96	0,96	4.777,21	0,96
Support Vector Machine	0,96	0,97	0,94	0,96	11,04	0,03
Random Forrest	0,60	0,99	0,07	0,14	62,57	4,06
K-Nearest Neighbor	0,80	0,97	0,56	0,71	0,47	19.462,80
Adaptive Boosting	0,84	0,84	0,79	0,81	2.917,49	12,22
VEA	0,97	0,97	0,96	0,96	16.803,67	18,45
HEA	0,97	0,97	0,95	0,96	7.425,03	10,10

Bảng 4.14. Kết quả đánh giá mô hình LA_Bin07 trên bộ dữ liệu UTL_DGA22

Nhãn	**Precision**	**Recall**	**F_1-Score**
Lành tính	0,98	0,98	0,98
DGA Botnet	0,98	0,97	0,97
Accuracy	*0,98*		

Mô hình LA_Mul07 có độ chính xác 0,86 khi đánh giá trên bộ dữ liệu UTL_DGA22.

Các kết quả trên cho thấy, (1) bộ dữ liệu UTL_DGA22 hoàn toàn phù hợp để đánh giá bài toán DGA Botnet và (2) các thuật toán đề xuất vẫn có độ chính xác cao tương tự khi đánh giá trên bộ dữ liệu mới.

4.6. Kết luận Chương 4

Trong Chương 4, NCS trình bày kết quả nghiên cứu về các bộ dữ liệu cho phân lớp nhị phân và phân lớp đa lớp trong bài toán DGA Botnet. Trong đó, trình bày chi tiết các điểm hạn chế cần cải tiến của các bộ dữ liệu hiện có, sau đó đề xuất một bộ dữ liệu mới UTL_DGA22, được xây dựng dựa trên cơ sở kế thừa các thành quả trước đó, bổ sung những cải tiến, dữ liệu và thuộc tính mới.

Các đóng góp chính bao gồm:

- Đề xuất một bộ dữ liệu DGA_UTL22, bao gồm 76 họ DGA Botnet phổ biến hiện nay, mỗi họ gồm 20.000 mẫu tên miền. Các tên miền thể hiện dưới dạng nguyên bản, lưu trữ dưới dạng tệp tin CSV, ARFF và TXT.

- Đề xuất 02 nhóm các thuộc tính, bao gồm 36 thuộc tính thuộc nhóm Base-Features cho tên miền và các thuộc tính TF-IDF cho họ tên miền DGA Botnet. Các thuộc tính này được trích xuất và lưu trữ dưới dạng *CSV, *ARFF và được chia sẻ đi kèm với bộ dữ liệu. Các đánh giá thử nghiệm sử dụng bộ thuộc tính này với các thuật toán học máy cho kết quả tích cực.

NCS cũng tiến hành các đánh giá trên bộ dữ liệu mới UTL_DGA22 với các thuật toán đề xuất trong luận án, bao gồm: NCM, học máy, VEA, HEA, LA_Bin07 và LA_Mul07. Kết quả cho thấy các thuật toán đề xuất đều phù hợp và có độ chính xác tương đồng như đánh giá trên các bộ dữ liệu trước đó.

NCS kỳ vọng UTL_DGA22 là cơ sở chung đáng tin cậy, được sử dụng rộng rãi cho phép các nhà khoa học đánh giá hiệu năng giữa các giải pháp đề xuất mới một cách thuận lợi, khách quan và dễ dàng đối sánh, tham chiếu.

Một phần kết quả nghiên cứu được trình bày tại Chương 4 được công bố tại [CT5] trong danh mục công trình của tác giả.

KẾT LUẬN

Sau thời gian học tập và nghiên cứu tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, NCS đã hoàn thành luận án "Nghiên cứu cải tiến một số mô hình học máy và học sâu áp dụng cho bài toán phân loại DGA Botnet", đạt được các kết quả chính như sau:

- Trình bày một cơ sở lý thuyết về DGA Botnet; bài toán phát hiện và phân loại DGA Botnet. Đánh giá hiệu quả của hai hướng tiếp cận sử dụng thuật toán phân cụm mờ NCM, học máy với hai mô hình học kết hợp là VEA và HEA trong phát hiện DGA Botnet.

- Đề xuất hai mô hình học sâu mới là LA_Bin07 và LA_Mul07 áp dụng cho phát hiện và phân loại DGA Botnet. Các đánh giá cho thấy, hai mô hình mới đạt được độ chính xác cao trong phân lớp, đặc biệt là trong bài toán phân loại họ DGA Botnet.

- Đề xuất bộ dữ liệu UTL_DGA22 chuyên dùng cho đánh giá bài toán DGA Botnet. Đây là có thể coi là bộ dữ liệu mới, đầy đủ mã nguồn và tài liệu nhất về DGA Botnet tính đến thời điểm hiện tại.

Các kết quả nghiên cứu trên đã trả lời các câu hỏi nghiên cứu đặt ra, mang lại ý nghĩa thiết thực trong bài toán phát hiện và phân loại DGA Botnet.

- *Đối với câu hỏi 1*: Trong bài toán phát hiện DGA Botnet, các đánh giá cho thấy việc áp dụng thuật toán phân cụm mờ NCM để phát hiện DGA Botnet tuy có tốc độ nhanh hơn nhưng độ chính xác thấp hơn các mô hình học máy. Ở chiều ngược lại, các mô hình học máy có độ chính xác cao, với Precision cao nhất đạt 0,97 nhưng hạn chế là tốn kém nhiều thời gian huấn luyện mô hình. Mô hình học sâu LA_Bin07 được đề xuất đã nâng cao độ chính xác phân loại, đạt từ 0,98 đến 1,00 khi đánh giá trên từng bộ dữ liệu khác nhau.

- *Đối với câu hỏi 2*: Mạng LSTM truyền thống có thể được cải tiến để nâng cao hiệu quả phát hiện và phân loại DGA Botnet. NCS đã đề xuất hai mô hình học sâu mới là LA_Bin07 và LA_Mul07 giải quyết được vấn đề trên. Trong đó trọng tâm là mô hình LA_Mul07 đã giải quyết bài toán phân loại DGA Botnet với độ chính xác được cải thiện rất đáng kể so với các giải pháp trước đó. Mô hình đạt độ chính xác thấp nhất là 0,86 và cao nhất là 1,00 khi đánh giá trên 03 bộ dữ liệu chuyên dụng. Giải pháp học sâu cũng có ưu điểm là thời gian huấn luyện nhanh hơn rất nhiều so với học máy, bởi tận dụng được năng lực tính toán của GPU.

- *Đối với câu hỏi 3*: Các bộ dữ liệu về DGA Botnet trước đó có những hạn chế về khả năng áp dụng cho từng bài toán, độ chính xác, số lượng và sự cân bằng mẫu, chưa có các họ DGA Botnet mới, thiếu tài liệu, thuộc tính đề xuất. NCS đã đề xuất bộ dữ liệu mới UTL_DGA22 chuyên dùng cho đánh giá bài toán DGA Botnet, được kế thừa đầy đủ các ưu điểm của những bộ dữ liệu trước đó, đồng thời bổ sung các mẫu dữ liệu, thuộc tính và mô tả đầy đủ. Tính từ thời điểm công bố bài báo vào đầu năm 2023, NCS đã nhận được 06 đề nghị truy cập để phục vụ nghiên cứu từ các nhà khoa học ở ngoài nước. NCS kỳ vọng UTL_DGA22 sẽ trở thành một bộ dữ liệu tiêu chuẩn được sử dụng rộng rãi bởi các nhà khoa học, chuyên gia an ninh mạng trong thời gian tới cho công tác đánh giá của họ.

Các kết quả nghiên cứu của NCS được công bố tại 04 bài báo trên tạp chí khoa học chuyên ngành uy tín thuộc danh mục ISI/Scopus, 02 báo cáo tại hội thảo khoa học chuyên ngành quốc gia, quốc tế uy tín, được thể hiện tại "Danh mục công trình của tác giả" và "Phụ lục" đính kèm.

Bên cạnh những kết quả đạt được, NCS đặt ra một số hướng phát triển trong thời gian tới để tiếp tục cải tiến mô hình. Cụ thể, cải tiến mô hình LA_Mul07 để nâng cao độ chính xác so

với mức 0,86 hiện tại trên các bộ dữ liệu có nhiều hơn 50 nhãn. Xây dựng cơ chế phát hiện riêng cho các họ DGA Botnet có sự tương đồng cao hoặc là các phiên bản kế thừa của nhau. Về hướng ứng dụng, NCS dự kiến kế thừa mô hình LA_Bin07 và LA_Mul07, tích hợp thành một module để phát hiện và phân loại DGA Botnet dựa trên tên miền. Module này có thể được tích hợp vào các giải pháp bảo vệ an ninh mạng tiên tiến, hiện đại như Tường lửa thế hệ mới hoặc Giải pháp an ninh hợp nhất.