Một số bộ dữ liệu kiểm thử phổ biến cho phát hiện xâm nhập mạng và đặc tính phân cụm

Bùi Công Thành[1*], Nguyễn Quang Uy[2] , Hoàng Minh[3]

[1]Binh chủng Thông tin liên lạc
[2]Học viện Kỹ thuật Quân sự
[3]Học viện Khoa học, Công nghệ và Đổi mới sáng tạo

Ngày nhận bài 24/5/2019; ngày chuyển phản biện 28/5/2019; ngày nhận phản biện 25/6/2019; ngày chấp nhận đăng 28/6/2019

Tóm tắt:

Những năm qua, đã có rất nhiều nghiên cứu về học máy (Machine learning), học sâu (Deep learning) cho lĩnh vực phát hiện xâm nhập mạng máy tính (IDS - Intrusion Detection System), sử dụng các bộ dữ liệu để đánh giá, phân tích. Do sự đa dạng, phức tạp của các bộ dữ liệu nên vấn đề phân cụm, chia nhỏ bộ dữ liệu ra thành các tập con nhưng vẫn giữ được đặc trưng của chúng là rất cần thiết. Trong nghiên cứu này, các tác giả tập trung phân tích đặc điểm của các tập dữ liệu kiểm thử phổ biến. Đồng thời, tiến hành thực nghiệm để đánh giá tính phân cụm, xác định số cụm tối ưu mà một bộ dữ liệu nên được chia ra. Thực nghiệm được tiến hành trên 6 tập dữ liệu huấn luyện của NSL-KDD, UNSW-NB15, CTU-13 phiên bản 08, 09, 10 và 13. Kết quả theo phương pháp Elbow, Silhouetee khá đồng nhất và cho thấy một số bộ dữ liệu nên được tách thành 2, 3 cụm, tuy nhiên cũng có những bộ nên để nguyên.

Từ khóa: **bộ dữ liệu, hệ thống phát hiện xâm nhập, K-Means.**

Chỉ số phân loại: **1.2**

Đặt vấn đề

Sự phát triển nhanh chóng của mạng máy tính (sau đây gọi tắt là mạng) và các dịch vụ mạng đang làm cho hoạt động của con người trở nên bị lệ thuộc. Hệ thống IDS là công nghệ an ninh mạng chủ động, cho phép giải quyết được vấn đề tấn công mạng cả từ bên trong, bên ngoài và phát hiện, ngăn chặn các hình thức tấn công mới lạ; các công việc này được thực hiện theo thời gian thực. Theo đánh giá, nghiên cứu về IDS phải luôn được cập nhật, cải tiến [1]. Trong những năm gần đây, nhiều công trình nghiên cứu về học máy (Machine learning), học sâu (Deep learning) cho lĩnh vực IDS đã được thực hiện. Khi đánh giá hiệu quả các công trình, các bộ dữ liệu lưu lượng mạng đã được sử dụng, mỗi bộ dữ liệu chứa nhiều bản ghi với các trường dữ liệu đặc trưng ứng với nhãn được gán. Nhiều bộ dữ liệu kiểm thử đã được các tổ chức, nhà khoa học nghiên cứu xây dựng (sau đây gọi là các bộ dữ liệu IDS dataset).

Thuộc tính của IDS dataset cơ bản được chia làm 2 nhóm: số (numerical) và tập hợp (catagorical). Việc xác định các thuộc tính của lưu lượng mạng có ý nghĩa hết sức quan trọng trong lĩnh vực nghiên cứu về IDS [2, 3], ví dụ như giảm số chiều dữ liệu sẽ tăng hiệu năng thuật toán; tăng chất lượng thuộc tính, từ đó tăng hiệu quả thuật toán; tăng tỷ lệ cảnh báo đúng, giúp cho việc biểu diễn dữ liệu được tường minh hơn. Khi thiết lập các bộ IDS dataset, các thuộc tính lưu lượng mạng được tính toán trên cơ sở giá trị tương ứng trong gói tin, tiêu đề gói tin và phiên kết nối mạng [2]. Ngoài thuộc tính, các tham số đặc trưng khác cho bộ dữ liệu như: kiểu dữ liệu, tính sẵn có; kích thước cho tập huấn luyện, kiểm tra; số mẫu tấn công, loại tấn công mạng; các hạn chế mang tính thời sự cũng cần được quan tâm trước khi lựa chọn để đánh giá các công trình nghiên cứu.

Trong lĩnh vực khám phá dữ liệu, phân cụm là phương thức chia dữ liệu thành các nhóm đối tượng có tính tương đương [4], giúp một số bài toán nâng cao hiệu suất, cân đối tài nguyên phần cứng... Mục tiêu của mô hình phân cụm là gán nhãn cho dữ liệu theo số cụm cho trước hoặc số cụm tối ưu nhất có thể theo từng bài toán. Việc xác định số cụm tối ưu cho một tập dữ liệu cụ thể đã được nhiều nhà nghiên cứu quan tâm, phổ biến như các phương pháp Elbow, Silhouete…

Việc nghiên cứu, tìm hiểu sâu về các bộ IDS dataset đã có nhiều công bố gần đây, tuy vậy mới tập trung phân tích một bộ dữ liệu cụ thể [5-8] mà không đưa ra được bức tranh khái quát về các bộ dữ dữ liệu phổ biến đang được sử dụng cho kiểm thử các thuật toán Machine learning, Deep learning trong lĩnh vực an ninh mạng. Thêm vào đó, với hiệu quả mang lại của tính phân cụm [4, 9], việc đánh giá tính phân cụm cho các bộ dữ liệu phổ biến này cần được quan tâm đúng mức. Từ các vấn đề đã phân tích ở trên, trong phạm vi nghiên cứu này, chúng tôi phân tích tổng quan các bộ IDS dataset phổ biến, tính phù hợp khi sử dụng, đặc biệt

*Tác giả liên hệ: Email: congthanhttmt@gmail.com

Some common datasets of an intrusion detection system and clustering properties

Cong Thanh Bui[1*], Quang Uy Nguyen[2] , Minh Hoang[3]

[1]*Communications Command*
[2]*Institute of Military Technology*
[3]*Institute of Science Technology and Innovation*

Received 24 May 2019; accepted 28 June 2019

Abstract:

In recent years, machine learning and deep learning based methods for intrusion detection systems (IDSs) have received great attention from many researchers. IDS datasets have been used to evaluate and analyse these methods. Because of the popularity and complication, the requirement to deeply explore the optimisation of clustering, which is known as one of the most useful techniques, not only reducing the amount of data but also keeping its characteristics, is necessary for these datasets. In this paper, we focus on analysing the characteristics of IDS common datasets. In addition, we also evaluate the clustering properties and discover the optimal number of clusters which should be divided from a dataset. The experiment has been conducted on six datasets NSL-KDD, UNSW-NB15, and four versions of CTU-13 (08, 09, 10, and 13). Using Elbow and Silhouette methods to determine the optimisation of clustering a dataset has revealed that some datasets should be divided into two or three clusters while some should keep their original forms.

Keywords: **dataset, intrusion detection system, K-Means.**

Classification number: **1.2**

tập trung sử dụng một số phương pháp để đánh giá tính phân cụm và đề xuất số cụm tối ưu cho tập huấn luyện của mỗi bộ dữ liệu này.

Một số bộ dữ liệu phổ biến

Bộ dữ liệu DARPA

Dữ liệu DARPA ra đời năm 1998, được tạo bởi Phòng thí nghiệm Lincoln (Viện Công nghệ Massachusetts) theo dự án tài trợ của Cục Dự án nghiên cứu cao cấp thuộc Bộ Quốc phòng Mỹ (Defence Advanced Research Project Agency). Bộ dataset được tạo bằng cách thu thập lưu lượng mạng (sử dụng tcpdump) của một hệ thống mạng mô phỏng các loại tấn công khác nhau [10]. Dataset DARPA được chia thành bộ dữ liệu huấn luyện và bộ dữ liệu kiểm thử: bộ dữ liệu huấn luyện được thu thập trong 7 tuần vận hành hệ thống, với mỗi tuần dữ liệu được thu thập trong 5 ngày, từ thứ 2 đến thứ 6; bộ dữ liệu kiểm thử được thu thập trong 2 tuần chạy hệ thống thử nghiệm, với mỗi tuần dữ liệu cũng được thu thập trong 5 ngày từ thứ 2 đến thứ 6. Bộ dữ liệu hiện có sẵn tại địa chỉ website chính thức của Phòng thí nghiệm Lincoln. Kích thước dữ liệu khoảng 4 GB với trên 5 triệu bản ghi cho bộ dữ liệu huấn luyện và khoảng 2 triệu bản ghi cho bộ dữ liệu kiểm thử.

Các loại tấn công mạng: dataset DARPA 1998 bao gồm 54 loại xâm nhập được phân làm 4 nhóm: R2L (Remote to Local), U2R (User to Root), DoS (Deniel of Service), Probe [5].

Một số hạn chế của bộ dữ liệu DARPA [5]: tính đúng đắn của dữ liệu thu thập gây nhiều tranh cãi; việc lưu trữ dữ liệu lưu lượng mạng dạng thô nên kích thước lớn và dẫn đến khó khăn cho các thử nghiệm; ngoài ra, vì hiện trạng dịch vụ, tốc độ mạng hiện nay đã khác rất nhiều so với năm 1998 nên không còn nhiều nghiên cứu sử dụng bộ dữ liệu này cho thử nghiệm, đánh giá. Đó là lý do chúng tôi không đặt trọng tâm phân tích cho bộ dữ liệu này.

Bộ dữ liệu KDD Cup 1999

Đây từng là bộ dữ liệu phổ biến cho kiểm thử các công trình nghiên cứu về lĩnh vực IDS trong hai thập kỷ qua. Dataset KDD Cup 1999 là một phiên bản của bộ dữ liệu DARPA 1998 [5], được sử dụng trong cuộc thi "Các công cụ khai phá dữ liệu và nghiên cứu tri thức quốc tế lần thứ 3 (The Third International Knowledge Discovery and Data Mining Tools Competition)". Để tạo ra bộ dữ liệu này, các thuộc tính từ bộ dữ liệu thô của dataset DARPA được trích ra thành các đặc trưng theo các thuật toán riêng biệt, độ lớn và số thuộc tính của bộ dữ liệu cũ vẫn được giữ nguyên [7]. Bộ dữ liệu hiện nay sẵn có tại website chính thức của cuộc thi và trên kho dữ liệu UCU Machina Learning Repository. Bộ dữ liệu có 24 loại tấn công, thêm 14 loại tấn công cho tập dữ liệu kiểm thử.

KDD Cup 1999 gồm hai bộ dữ liệu con: một bộ dữ liệu

đầy đủ và một bộ dữ liệu bằng 10% so với bộ dữ liệu đầy đủ. Với mỗi bộ lại có một bản không có nhãn và một bản có nhãn (label) đi kèm. Các bộ dữ liệu đều được lưu dưới dạng file text (txt). Mỗi bản ghi chứa 41 trường thông tin và một nhãn, nhãn được đánh là bình thường hoặc là một loại tấn công cụ thể. Các thuộc tính được chia làm 3 nhóm: 1) Basic features: bao gồm các thuộc tính có thể thu thập được từ một kết nối TCP/IP, hầu kết các thuộc tính này dẫn đến độ trễ trong phát hiện; 2) Traffic features: là các thuộc tính được tính toán dựa trên giá trị trường window trong gói tin TCP/IP; 3) Content features: với các tấn công R2L, U2R thường thì các kết nối và tần suất các kết nối rất khác với các tấn công dạng DoS hay Probe. Thông tin về các loại tấn công này cơ bản chứa trong phần nội dung (content) của TCP/IP, ví dụ như số lần login lỗi… Một phiên bản mở rộng, gần giống với bộ dữ liệu này có tên là gure KDD Cup [11], được xem là bộ dữ liệu (KDDCup99+payload).

Hạn chế của dataset KDD [5] là: bộ dữ liệu có rất nhiều bản ghi trùng lặp, cụ thể trên bộ dữ liệu huấn luyện và kiểm thử tương ứng có 78% và 75% bản ghi trùng; thêm vào đó, sự không đồng đều trong phân bố giữa tập huấn luyện và tập kiểm thử làm ảnh hưởng đến kết quả đánh giá cho các thuật toán phân lớp. Theo các đánh giá [5], khi sử dụng các bộ phân lớp phổ biến J48, Decision Tree Learning, Naive Bayes, NBTree, Random Forest, Support Vector Machine (SVM)… để huấn luyện và kiểm thử trên bộ dữ liệu KDD cho độ chính xác rất cao, tất cả đều từ 96-98%, do vậy việc sử dụng bộ dữ liệu này cho kiểm thử các thuật toán mới hơn sẽ không còn thực sự phù hợp nữa (bảng 1).

Bảng 1. Phân bố theo loại tấn công của các bộ KDD.

Dataset	Tổng số	DoS	Probe	R2L	U2R	Normal	Số chiều
Tập huấn luyện	1.074.992	247.267	13.860	999	52	812.814	42
Tập kiểm thử	311.029	229.853	4.166	16.189	228	60.593	42

Bộ dữ liệu NSL-KDD

NSL-KDD là bộ dữ liệu được Tavallaee và cộng sự công bố năm 2009 [5], là một phiên bản được định nghĩa lại từ bộ KDD Cup 1999 trên cơ sở loại bỏ một số bản ghi bị thừa, trùng lặp thông tin [6]. Hiện tại, bộ dữ liệu được sử dụng trong rất nhiều công trình nghiên cứu, giúp phát hiện sự bất thường khi kiểm thử, đánh giá. So với bộ dữ liệu gốc, bộ dữ liệu này có các đặc điểm mới như: không bao gồm các bản ghi dư thừa trong tập huấn luyện, do vậy kết quả phân lớp sẽ không theo hướng của các bản ghi xuất hiện nhiều hơn; không còn bản ghi trùng lặp trong bộ dữ liệu kiểm thử; xử lý vấn đề khi vùng kết quả đánh giá hẹp hiệu quả hơn so với bộ dữ liệu KDD; cân đối hợp lý số lượng bản ghi giữa tập huấn luyện và kiểm thử. Bộ dữ liệu hiện sẵn có tại website của nhóm nghiên cứu dưới dạng tệp tin .csv, với tập huấn luyện gồm hơn 125 nghìn bản ghi, tập kiểm thử hơn 22 nghìn bản ghi.

Mỗi bản ghi trong bộ dữ liệu có 42 thuộc tính được liệt kê giống như với bộ dữ liệu KDD Cup 1999, được mô tả ở bảng 2. Bộ dữ liệu này cho hiệu quả khá tốt khi sử dụng để đánh giá các thuật toán học máy. Hạn chế lớn nhất của bộ dữ liệu đó là không thể hiện được vết của các cuộc tấn công ở mức độ thấp, tinh vi [12].

Bảng 2. Phân bố theo loại tấn công của các bộ NSL-KDD.

Dataset	Tổng số	DoS	Probe	U2R	R2L	Normal	Số chiều
Tập huấn luyện	125.972	45.927	11.656	52	995	67.342	42
Tập kiểm thử	22.542	7.457	2421	200	2.754	9.711	42

Bộ dữ liệu UNSW-NB15

Bộ dữ liệu UNSW-NB15 [8] được công bố năm 2015, được tạo thông qua việc thu thập lưu lượng mạng bởi Phòng thí nghiệm Cyber Range của Australian Centre for Cyber Security (ACCS). Hệ thống mạng và giả lập tấn công được đánh giá là sát với thực tế hoạt động của mạng và các mã độc hiện nay thông qua công cụ giả lập tấn công của hãng IXIA. Sau khi sử dụng Tcpdump để thu thập hơn 100 GB lưu lượng thô (dạng tệp .pcap), với 9 mẫu tấn công (Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode và Worms), họ sử dụng công cụ Argus, Bro-IDS với 12 thuật toán khác nhau để tạo ra 49 thuộc tính dữ liệu. Bộ dữ liệu hiện sẵn có trên mạng Internet với số bản ghi của tập huấn luyện và tập kiểm thử tương ứng là trên 175 nghìn và 82 nghìn [8].

Bộ dữ liệu UNSW-NB15 được nhiều công trình nghiên cứu sử dụng để kiểm thử các thuật toán phân lớp trong những năm gần đây [12] nhờ khắc phục được hạn chế thiếu mẫu tấn công mới; lưu lượng mạng thể hiện được dịch vụ mạng đương thời; có sự phân bố đồng đều giữa tập huấn luyện và kiểm thử (được phân bố theo tỷ lệ 40/60 tương ứng giữa tập kiểm thử và tập huấn luyện) [13]. Mỗi bản ghi trong bộ dữ liệu có 49 thuộc tính được mô tả ở bảng 3.

Bảng 3. Phân bố theo loại tấn công của các bộ UNSW-NB15.

Loại tấn công	Tập huấn luyện		Tập kiểm thử	
	Số bản ghi	*Tỷ lệ %*	*Số bản ghi*	*Tỷ lệ %*
Analysis	2.000	1,141	677	0,822
Backdoor	1.746	0,996	583	0,708
DoS	12.264	6,994	4.089	4,966
Exploit	33.393	19,045	11.132	13,521
Generic	40.000	22,813	18.871	22,921
Fuzzers	18.184	10,371	6.092	7,363
Reconnaissance	10.491	5,983	3.496	4,246
Shellcode	1.133	0,646	378	0,439
Worms	130	0,074	44	0,053
Dữ liệu Normal	56.000	31,938	37.000	44,942

Bộ dữ liệu CTU-13

Bộ dữ liệu CTU-13 được nghiên cứu bởi Đại học Kỹ thuật Séc và được công bố năm 2011 [14]. Đây là bộ dữ liệu chứa thông tin bao gồm cả lưu lượng Botnet, dữ liệu bình thường và dữ liệu lưu lượng của hạ tầng dịch vụ mạng. Bộ dữ liệu gồm 13 tập dữ liệu con theo các tình huống hoạt động khác nhau ứng với từng mẫu mã độc. Các gói tin sau khi được thu thập (dạng .pcap) sẽ được xử lý bởi công cụ Argus (Audit Record Generation and Utilization System) để tạo thành các thuộc tính cho bộ dữ liệu huấn luyện và kiểm thử. Các bộ dữ liệu con có số các thuộc tính khác nhau và được đánh tên theo ký hiệu từ CTU-13_01 đến CTU-13_13. Bộ dữ liệu có tại website của đơn vị chủ quản. Trong phạm vi bài viết, chúng tôi tập trung vào các bộ dữ liệu con CTU-13_08, CTU-13_09, CTU-13_10, CTU-13_13. Các bộ dữ liệu này đang được nhiều nghiên cứu đưa ra để đánh giá kết quả trong lĩnh vực Machine learning, Deep learning [15], tuy vậy hạn chế lớn nhất là bộ dữ liệu chỉ chứa các tấn công mạng dạng Botnet.

Khi tách số lượng mẫu theo 2 loại dữ liệu bình thường và bất thường, thu được số lượng bản ghi tương ứng và số chiều của mỗi tập con CTU-13 như trên bảng 4.

Bảng 4. Phân bố theo loại tấn công của các bộ CTU-13 (08, 09, 10, 13).

Bộ dữ liệu	Số mẫu bất thường	Số mẫu bình thường	Tổng số	Số chiều
CTU-13_08	6.127	72.822	78.949	16
CTU-13_09	184.987	29.967	214.954	16
CTU-13_10	106.352	15.847	122.199	16
CTU-13_13	40.003	31.939	71.942	16

Tính phân cụm dữ liệu, số cụm tối ưu

Phân cụm dữ liệu

Phân cụm là chia dữ liệu thành các nhóm đối tượng tương đương [4] để giúp giảm kích thước dữ liệu mà vẫn giữ được đặc trưng của chúng, khi đó dữ liệu được mô tả bằng từng cụm riêng lẻ. Việc phân cụm góp phần quan trọng trong cải thiện, nâng cao hiệu quả giải quyết các vấn đề trong các lĩnh vực toán học, thống kê và phân tích số liệu. Trong lĩnh vực học máy, phân cụm thuộc bài toán học không giám sát, mục tiêu của mô hình phân cụm là gán nhãn cho dữ liệu theo số cụm cho trước hoặc số cụm tối ưu nhất có thể theo từng bài toán.

K-Means là một trong những thuật toán phân cụm phổ biến và được ứng dụng rộng rãi nhất, từ tập dữ liệu đầu vào với N điểm, thuật toán thực hiện trên cơ sở xác định K trung tâm là đại diện cho K cụm dữ liệu được tạo ra, K trung tâm được xác định dựa vào trung bình khoảng cách của các điểm tương ứng thuộc cụm đó đến các trung tâm. Thuật toán có thể mô tả như sau:

Input: *N điểm dữ liệu là* $X=[x_1, x_2, \dots x_N] \in R^{dxN}$*, số cụm mong muốn* $K<N$.

Output: *các* $c_1, c_2, \dots c_K \in R^{dx1}$ *và nhãn của từng điểm dữ liệu* x_i*, với* $i<N$.

Bước 1: chọn ngẫu nhiên K điểm làm các giả trung tâm C_j *ban đầu,* $j<K$.

Bước 2: xác định tương quan dx_iC_j *từ* x_i *đến mỗi* C_j*, với* dx_iC_j *bé nhất, gán điểm* x_i *về thuộc* C_j.

Bước 3: lặp lại bước 2 đến khi không còn x_i *cần cập nhật lại về* C_j.

Bước 4: lấy trung bình cộng của tất cả các khoảng cách dx_iC_j *ứng với cụm đó, cập nhật giá trị này cho các giả trung tâm* C_j.

Bước 5: thực hiện lại từ bước 2. Thuật toán dừng khi các giả trung tâm C_j *không còn thay đổi.*

Phương pháp đánh giá tính phân cụm

Trong bài toán phân cụm, việc dữ liệu có nên phân thành cụm nhỏ hơn hay không và nên chia thành bao nhiêu cụm là một vấn đề rất quan trọng. Việc đánh giá vấn đề này sẽ trả lời được câu hỏi là tập dữ liệu có tính phân cụm không, số cụm tối ưu K nên phân ra là bao nhiêu, đây là cơ sở để hỗ trợ cho các kỹ thuật xử lý dữ liệu tiếp theo. Phân dữ liệu thành các cụm, việc xác định số cụm K tối ưu nhất khi phân cụm đóng vai trò quan trọng đối với việc biểu diễn tốt nhất đặc trưng của toàn bộ dữ liệu và đặc trưng của từng cụm [9]. Có nhiều phương pháp để xác định số cụm tối ưu, trong đó Elbow và Silhouette là các phương pháp xác định số K tối ưu dựa vào trực quan trên biểu đồ.

Theo phương pháp khủy tay (Elbow), một đồ thị 2D sẽ được biểu diễn bởi trục x là số cụm dự kiến sẽ chia (ví dụ từ 1-5), trục y biểu diễn tổng bình phương khoảng cách tất cả các điểm đến trung tâm cụm Cj. Số K tối ưu được xác định theo công thức sau, ứng với điểm tại đó trục x và đồ thị tạo nên khủy tay:

$$Y(K)=\sum_{j=1}^{K}\sum_{i=1}^{len(C_j)} Dx_iC_j$$

Theo phương pháp bóng mờ (Silhouette), cũng là một phương pháp phổ biến cho xác định số K tối ưu thông qua biểu đồ, bóng mờ của mỗi điểm được tính theo công thức:

$$S(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))}$$

Trong đó, a(i) là khoảng cách từ điểm x_{ij} tới tất cả các điểm trong cùng cụm Cj, b(i) là trung bình của khoảng cách của tất cả các điểm trong cụm gần nhất. Phương pháp chứng minh được rằng điểm x_{ij} được phân cụm tốt nếu giá trị S(i)

tiến tới giá trị cao nhất là 1, và ngược lại thì việc phân cụm cho xij là không tốt. Bóng của mỗi trường hợp K khác nhau được biểu diễn trên biểu đồ với x là giá trị bóng S(i) của mỗi điểm, y thể hiện mật độ số điểm tương ứng với giá trị bóng đó. Với mỗi trường hợp số K khác nhau, S là trung bình giá trị bóng của tất cả các điểm, được tính theo công thức:

$$S_K = (1/N) \sum S(i)$$

Theo phương pháp Silhouette, trong trường lý tưởng, biểu đồ sẽ thể hiện bóng của N cụm tương ứng với N hình chữ nhật có chiều dài +1, không có bóng được vẽ trên khoảng (-1, 0). Việc đánh giá tính phân cụm của một tập dữ liệu được chia theo K cụm dựa vào nguyên tắc sau: i) Phương án trong đó có ít giá trị S(i) âm nhất (được vẽ trên phần x<0 trên sơ đồ); ii) Các cụm có xu hướng gần tới +1 nhất; iii) Các cụm có bóng gần giá trị Silhouette trung bình; iv) Trong trường hợp số cụm bằng M, M+1 đều cho kết quả như nhau, thì phương án được chọn thường là M, tuy nhiên chọn M+1 nếu phương án này chia các tập đều nhau hơn và tốt hơn cho các bước xử lý tiếp theo.

Thực nghiệm

Thiết lập thực nghiệm

Chúng tôi tiến hành 2 thực nghiệm để đánh giá tính phân cụm của 6 tập dữ liệu, trên cơ sở đó xác định số cụm tối ưu theo mỗi phương pháp đưa ra, kiểm tra tính đồng nhất về kết quả giữa các phương pháp. Lần thử nghiệm thứ nhất theo phương pháp Elbow, lần thử nghiệm thứ hai theo phương pháp Silhoutte.

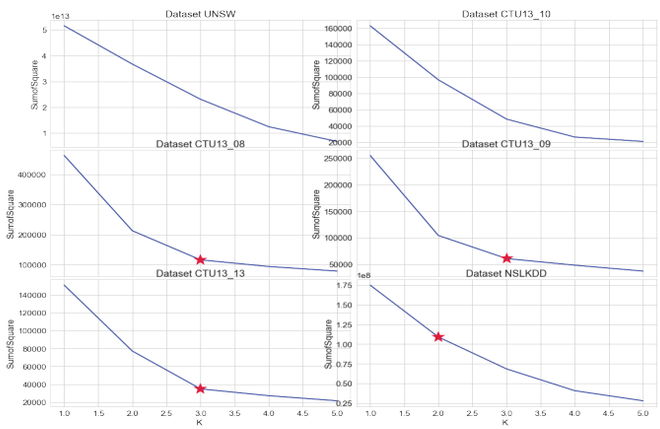
Trừ bộ dữ liệu DARPA và KDD1999, do không được đánh giá cao, các tập dữ liệu còn lại đã nêu ở trên đều được sử dụng cho thực nghiệm. Các bộ dữ liệu sau khi tiền xử lý [15] với kỹ thuật one-hot-encoding, bộ dữ liệu NSL-KDD, UNSW-NB15, CTU13-08, 09, 10, 13 có số chiều tương ứng là 122, 196, 40, 41, 38 và 40. Các bộ dữ liệu sau đó được trích rút phần dữ liệu bình thường của tập huấn luyện và không sử dụng nhãn, đối với các tập CTU-13 vì không có tập huấn luyện và tập dữ liệu kiểm thử riêng được chia theo tỷ lệ 40/60 ứng với tập huấn luyện/kiểm thử của tập dữ liệu tương ứng. Do nền tảng phần cứng không cho phép, các thuật toán như Silhouette có lượng tính toán lớn nên quá trình thực nghiệm sẽ được tiến hành 5 lần với 10 và 20% dữ liệu được lấy ngẫu nhiên từ các bộ dữ liệu kiểm thử tương ứng với nhóm NSL-KDD, UNSW-NB15 và nhóm CTU-13. Việc cài đặt thử nghiệm được tiến hành trên ngôn ngữ Python 3.0, công cụ phát triển Jupyter Notebook, sử dụng thư viện Sklearn, Numpy và Pandas cho việc cài đặt các thuật toán. Kết quả thử nghiệm được thực hiện trên máy tính sử dụng hệ điều hành MAC OS 10.14.3, cấu hình Intel(R) Core (TM) i5, 8 GB DDR3. Thông số tập dữ liệu kiểm thử được trình bày ở bảng 5.

Bảng 5. Thông tin chi tiết dữ liệu thử nghiệm.

Bộ dữ liệu	Số chiều nguyên bản	Số chiều sau tiền xử lý dữ liệu [15]	Số bản ghi kiểm thử	Tỷ lệ lấy mẫu	Tập huấn luyện
UNSW-NB15	49	196	5.600	10%	56.000
CTU-13_08	16	40	5.825	20%	29.128
CTU-13_09	16	41	2.397	20%	11.986
CTU-13_10	16	38	1.267	20%	6.338
CTU-13_13	16	40	2.555	20%	12.775
NLS-KDD	42	122	6.734	10%	67.3400

Kết quả và đánh giá

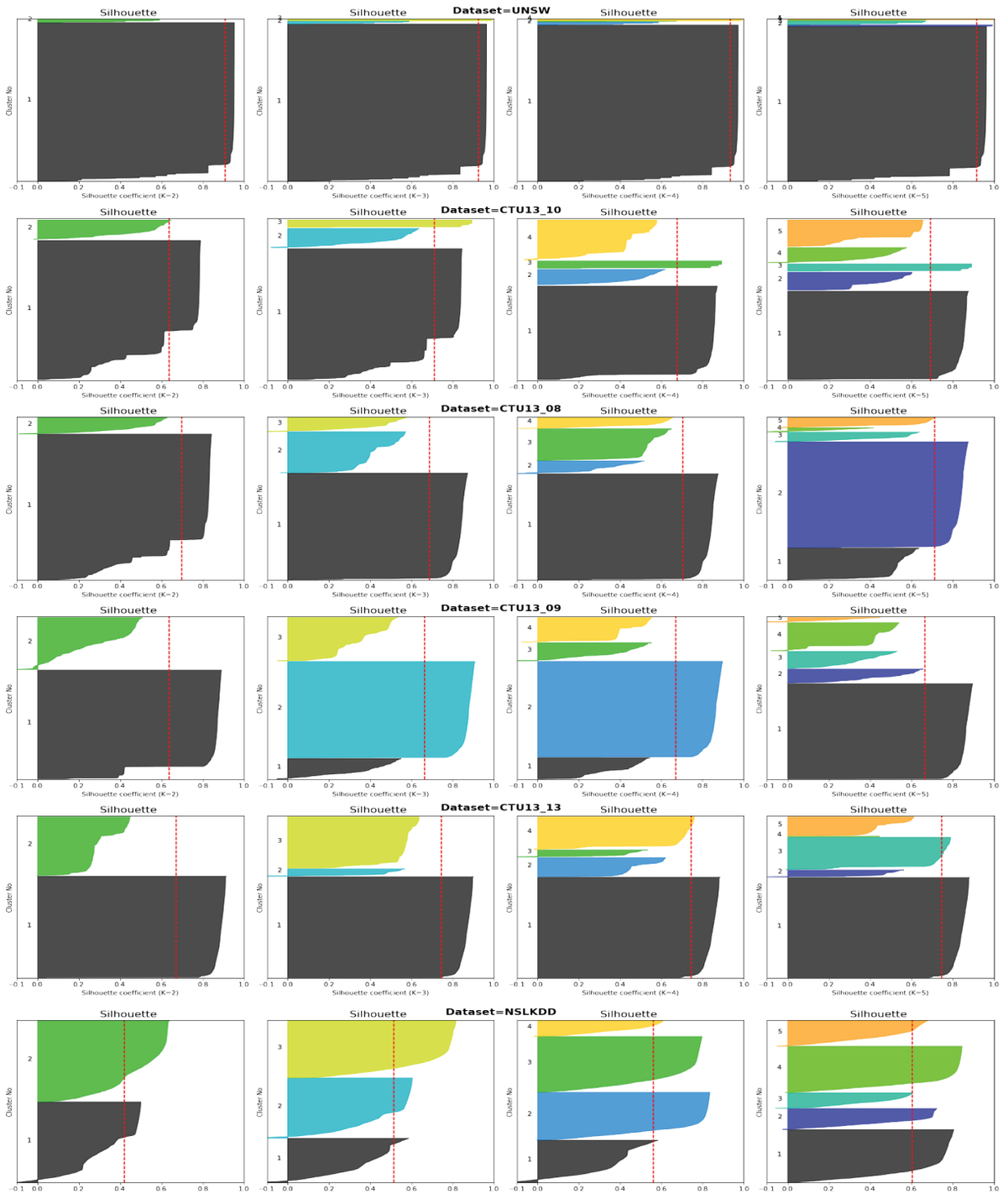
Kết quả thực nghiệm theo phương pháp Elbow: thực nghiệm theo phương pháp Elbow trên 6 bộ dữ liệu, qua 5 lần thử, mỗi lần thử tính K=(1-5) trên 10% dữ liệu của tập huấn luyện NSL-KDD, UNSW-NB15 và 20% trên các bộ dữ liệu thuộc nhóm CTU-13, cho kết quả ổn định ở các lần thử khác nhau; theo đó với bộ dữ liệu UNSW-NB15, việc tách bộ dữ liệu thành K cụm khác nhau (với K từ 1 đến 5) đều thể hiện không rõ bởi phương pháp Elbow. Tương tự với bộ dữ liệu CTU-13_10. Còn với các bộ dữ liệu còn lại, CTU-13_08, 09, 13 thể hiện rất rõ Elbow tại vị trí K=3, còn NLS-KDD thể hiện Elbow ở K=2. Sơ đồ thể hiện vị trí Elbow khi thử nghiệm các bộ dữ liệu trong lần thử thứ nhất thể hiện ở hình 1.



Hình 1. Kết quả thực nghiệm lựa chọn K tối tưu theo phương pháp Elbow (lần 1).

Kết quả thực nghiệm theo phương pháp Silhouette: thực nghiệm phương pháp Silhouette trên 6 bộ dữ liệu, qua 5 lần thử, mỗi lần thử tính K=(2-5) trên 10% dữ liệu của tập huấn luyện NSL-KDD, UNSW-NB15 và 20% trên các bộ dữ liệu thuộc nhóm CTU13, kết quả cho thấy sự ổn định ở các lần thử khác nhau, theo đó với bộ dữ liệu UNSW-NB15 và CTU-13_10, việc tách bộ dữ liệu thành nhiều cụm đều không tốt vì có một cụm chiếm đa số điểm dữ liệu, trong các phương án K khác nhau thì các điểm của cụm này vẫn gần

như không bị tách ra. Với bộ dữ liệu CTU-13_8 cho giá trị Silhouette khá đồng đều tại K=5, tuy nhiên việc phân cụm tạo ra số điểm các cụm vẫn khá cao, theo biểu đồ thì chia số cụm bằng 3 sẽ hợp lý hơn cả. Bộ dữ liệu CTU-13_09, 13 cho giá trị Silhouette với trường hợp K=3, 4, 5 gần như nhau, tuy nhiên giữa việc chia thành 3 cụm hay nhiều cụm hơn không làm cho các cụm có số điểm dữ liệu đều hơn, nên chia thành 3 cụm là phù hợp nhất. Còn bộ dữ liệu NLS-KDD, việc chia thành 2 cụm cho thấy có ít điểm có giá trị Silhouette âm và đa số giá trị hướng đến +1, do vậy chia thành 2 cụm là phù hợp hơn trong số các phương án đưa ra. Hình 2 là biểu đồ thể hiện các bóng Silhouette của các cụm trong mỗi phương án khi thử của lần 1.



Hình 2. Kết quả thử nghiệm theo phương pháp Silhouette (lần thử 1) trên 6 Dataset.

Kết luận

Trong phạm vi bài viết, chúng tôi đã phân tích, đánh giá một số bộ dữ liệu phục vụ cho kiểm thử trong lĩnh vực nghiên cứu IDS, một số bộ dữ liệu chính như NSL-KDD, UNSW-NB15, CTU-13 phiên bản 08, 09, 10 và 13, đây là các bộ dữ liệu thường được sử dụng cho kiểm thử các công trình về ứng dụng học máy cho phát hiện bất thường trên mạng máy tính [12, 15] trong những năm gần đây. Qua phân tích cho thấy, các bộ dữ liệu DARPA, KDD1999 hiện đã không còn phù hợp cho đánh giá các kết quả nghiên cứu trong lĩnh vực học máy, học sâu. Bộ dữ liệu NSL-KDD có những tiến bộ vượt trội so với bộ dữ liệu gốc, tuy vậy vẫn còn thiếu tính thời sự khi không chứa các cuộc tấn công mạng gần đây. Bộ dữ liệu UNSW-NB13 và các phiên bản của CTU-13 được tạo ra gần đây đã cơ bản khắc phục được hạn chế của các bộ dữ liệu có trước. Tuy vậy, các tập CTU-13 thường chỉ được sử dụng cho kiểm thử các tấn công Botnet.

Tính phân cụm của dữ liệu có vai trò quan trọng vì giúp cho dữ liệu có kích thước nhỏ hơn nhưng vẫn cơ bản giữ được các đặc trưng vốn có. Chúng tôi đã tiến hành thực nghiệm đánh giá tính phân cụm của các bộ dữ liệu, nội dung chính của thực nghiệm là tiền xử lý các bộ dữ liệu, cài đặt bài toán phân cụm dữ liệu sử dụng thuật toán K-Means và đánh giá tính phân cụm theo 2 phương pháp thường được sử dụng là Elbow và Silhouette. Qua thử nghiệm trên 6 tập dữ liệu khác nhau (NSL-KDD, UNSW-NB15, CTU-13_08, 09, 10 và 13), với 5 lần thử, dữ liệu được lấy mẫu theo tỷ lệ cố định, ngẫu nhiên từ tập huấn luyện cho thấy kết quả đánh giá tính phân cụm theo 2 phương án là đồng nhất. Qua đó có thể đưa ra đánh giá, với bộ dữ liệu huấn luyện của UNSW-NB15, CTU-13_10 thì không rõ tính phân cụm, các bộ dữ liệu CTU-13_09, 13 thì việc phân thành 3 cụm là phù hợp nhất, còn bộ dữ liệu NSL-KDD thì việc phân thành 2 cụm cho thấy tốt hơn so với phân thành nhiều cụm.

Kết quả nghiên cứu là cơ sở để lựa chọn tập dữ liệu kiểm thử cho các công trình nghiên cứu về học máy, học sâu trong lĩnh vực IDS. Ngoài ra, số cụm tối ưu được đề xuất theo kết quả thực nghiệm có thể là cơ sở để sử dụng cho chia nhỏ dữ liệu thành nhiều cụm, giúp phát triển các thuật toán lai ghép phân cụm với thuật toán đã có.

TÀI LIỆU THAM KHẢO

[1] C. Manasa, M.V. Panduranga Rao, S. Basavaraj Patil (2012), "A Survey on Intrusion Detection System", *International Journal of Computer Application and Managerment Research*, DOI: 10.1109/WSWAN.2015.7210351.

[2] D.K. Bhattacharyya, J.K. Kalita (2013), *Network anomaly detection: A machine learning perspective*, CRC Press.

[3] M.H. Bhuyan, D.K. Bhattacharyya, J.K. Kalita (2014), "Network anomaly detection: methods, Systems and Tools", *IEEE Communications Surveys & Tutorials*, **16(1)**, pp.303-336.

[4] P. Berkhin (2002), "Grouping Multidimensional Data", *Springer*, https://doi.org/10.1007/3-540-28349-8_2.

[5] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, Ali A. Ghorbani (2009), "A Detailed Analysis of the KDD CUP 99 Data Set", *Proceedings of the Second IEEE International Conference*, DOI: 10.1109/CISDA.2009.5356528, pp.53-58.

[6] L. Dhanabal, S.P. Shantharajah (2015), "A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms", *International Journal of Advanced Research in Computer and Communication Engineering*, **4(6)**, pp.446-452.

[7] Atilla Ozgur, Hamit Erdem (2016), "A review of KDD99 dataset usage in intrusion detection and machine learning between 2010 and 2015", *PeerJ Preprints,* DOI:10.7287/peerj.preprints.1954v1.

[8] Nour Moustafa, Jill Slay (2015), *NSW-NB15 A Comprehensive Data set for Network Intrusion Detection Systems*, School of Engineering and Information Technology, University of New South Wales at the Australian Defence Force Academy Canberra, Australia.

[9] Suneel Kumar Kingrani, Mark Levene, Dell Zhang (2018), "Estimating the number of clusters using diversity", *Artificial Intelligence Research*, **7(1)**, DOI: https://doi.org/10.5430/air.v7n1p15.

[10] Richard Lippmann (1999), "Summary and Plans for the 1999 DARPA Evaluation", *MIT Lincoln Laboratory*, DOI: 10.1007/3-540-39945-3_11.

[11] Inigo Perona, Olatz Arbelaitz, Ibai Gurrutxaga, Jose I. Martın, Javier Muguerza, Jesus M. Perez (2017), *Generation of the database gurekddcup*, Department of Education, Universities and Research of the Basque Government.

[12] Abhishek Divekar, Meet Parekh, Vaibhav Savla, Rudra Mishra (2018), "Benchmarking datasets for Anomaly-based Network Intrusion Detection: KDD CUP 99 alternatives", *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS).*

[13] Nour Moustafa, Jill Slay (2016), "The evaluation of Network Anomaly Detection Systems: statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 dataset", *Information Security Journal: A Global Perspective 2016*, **25(1-3)**, pp.18-31.

[14] S. Garcia, M. Grill, H. Stiborek, A. Zunino (2014), "An empirical comparison of botnet detection methods", *Computers and Security Journal*, **45**, pp.100-123.

[15] Van Loi Cao, Miguel Nicolau, James McDermott (2018), "Learning neural rep- resentations for network anomaly detection", *IEEE Transactions on Cybernetics,* **49(8)**, pp.3074-3087.