HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

BÙI CÔNG THÀNH

PHÁT TRIỂN MỘT SỐ MÔ HÌNH PHÁT HIỆN BẤT THƯỜNG MẠNG DỰA TRÊN HỌC SÂU VÀ TỔNG HỢP DỮ LIỆU

Chuyên ngành : Hệ thống thông tin
Mã số : 9.48.01.04

TÓM TẮT LUẬN ÁN TIẾN SĨ KỸ THUẬT

Hà Nội - 2021

Công trình được hoàn thành tại
Học viện Công nghệ Bưu chính Viễn thông

Người hướng dẫn khoa học:

Phản biện 1:

Phản biện 2:

Phản biện 3:

Luận án sẽ được bảo vệ trước Hội đồng chấm luận án cấp Học viện

Họp tại: Học viện Công nghệ Bưu chính Viễn thông

Vào hồi giờ ngày tháng năm 2021

Có thể tìm hiểu luận án tại:

- Thư viện Học viện Bưu chính Viễn thông
- Thư viện Quốc gia Việt Nam

CÁC CÔNG TRÌNH CÓ LIÊN QUAN ĐẾN LUẬN ÁN

- [CT1] **Thanh Cong Bui,** Loi Van Cao, Minh Hoang, and Quang Uy Nguyen. A clustering-based shrink autoencoder for detecting anomalies in intrusion detection systems. In 2019 11th International Conference on Knowledge and Systems Engineering (KSE), pp. 1–5. IEEE, (2019).
- [CT2] **Thanh Cong Bui**, Minh Hoang, Quang Uy Nguyen, and Cao Loi Van. Data fusion-based network anomaly detection towards evidence theory. 2019 6th NAFOSTED International Conference on Information and Computer- Science (NICS'19). pp. 33–38. IEEE (2019) (The Best Paper Award).
- [CT3] **Bùi Công Thành**, Vũ Tuấn Anh, Hoàng Trung Kiên. Ứng dụng lý thuyết Dempster Shafer trong xây dựng mô hình suy luận. Tạp chí Nghiên cứu Khoa học và Công nghệ Quân sự, 50(08) 08.2017, 144–157 (2017).
- [CT4] **Bùi Công Thành**, Nguyễn Quang Uy, Hoàng Minh. Một số bộ dữ liệu kiểm thử phổ biến cho phát hiện xâm nhập mạng và đặc tính phân cụm. Tạp chí Khoa học và Công nghệ Việt Nam, Bộ Khoa học và Công nghệ, 62(1) 1.2020:1–7, (2020), (Series B), ISSN 1859-4794.
- [CT5] **Thanh Cong Bui**, Loi Van Cao, Minh Hoang, and Quang Uy Nguyen. Double-shrink autoencoder for network anomaly detection. Tạp chí Tin học điều khiển, Viện Hàn lâm Khoa học và Công nghệ Việt Nam V.36, N.2 (2020).
- [CT6] **Thanh Cong Bui**, Van Loi Cao, Quang Uy Nguyen, and Minh Hoang. One-class Fusion-based Learning Model for Anomaly Detection. Journal of Computer in Industry: Classification, Machine learning, pp. ...--.... (ISI-SCIE, IF=3.954) (2021) (Under Review).

KẾT LUẬN

Như vậy, luận án đã nghiên cứu và giải quyết các vấn đề theo phát biểu bài toán đặt ra; kết quả nghiên cứu chính đã được công bố trong các công trình khoa học uy tín trong và ngoài nước.

Chương một, luận án trình bày các nội dung cơ sở liên quan đến luận án, tập trung vào làm rõ về khái niệm, mô hình tổng thể phát hiện bất thường mạng, các thành phần chính của mô hình. Chương hai, luận án trình bày kết quả nghiên cứu để giải quyết vấn đề thứ nhất mà luận án đã đặt ra trong phát biểu bài toán, đề xuất giải pháp có tên KSAE và DSAE. Chương thứ ba, luận án trình bày mô hình NAD dựa trên tổng hợp, có tên OFuseAD, sử dụng lý thuyết D-S theo hướng kết hợp nhiều phương pháp đơn OCC, để tạo ra phương pháp tổng thể có khả năng phát hiện bất thường mạnh hơn, mô hình NAD dựa trên tổng hợp dữ liệu đã đề xuất còn có khả năng tự động ước lượng ngưỡng ra quyết định.

1. Một số giới hạn của luận án

Hiện đang giả định các nguồn cung cấp thông tin (các phương pháp đơn) trong OFuseAD đều đang quan sát cùng một đối tượng thông tin gốc như nhau. Trong thực tế vẫn có nhiều bài toán, việc nhiều nguồn thông tin gốc khác nhau nhưng đều tham gia đóng góp cho cùng một giả định của hệ thống.

Thêm vào đó, độ phức tạp tính toán của OFuseAD phụ thuộc lớn vào các phương pháp đơn, trong khi các OCC truyền thống thường cho chi phí tính toán rất lớn.

2. Hướng nghiên cứu trong tương lai

Tiếp tục nghiên cứu cải tiến mô hình DSAE để có thể áp dụng cho các bài toán phát hiện bất thường khác, không chỉ dừng lại ở lĩnh vực an ninh mạng.

Nghiên cứu mở rộng OFuseAD theo hướng sử dụng các nguồn thông tin gốc khác nhau cho các phương pháp đơn.

Phát triển kết quả đề xuất mở rộng lý thuyết D-S, giải thuật DRC_AD, cho các bài toán trong các lĩnh vực khác nhau./.

PHẦN MỞ ĐẦU

1. Tính cấp thiết của luận án

Nghiên cứu phát hiện bất thường mạng về cơ bản là tìm kiếm giải pháp hiệu quả để xác định độ lệch của dữ liệu đầu vào so với các mẫu dữ liệu sử dụng cho biểu diễn hoạt động thông thường của hệ thống đã được thiết lập trước, qua đó đánh dấu các xâm nhập (các bất thường hay tấn công mạng).

Trong xây dựng các phương pháp phát hiện bất thường mạng, nhãn của tấn công được cho là không sẵn có trong quá trình huấn luyện mô hình. Các phương pháp NAD được khuyến nghị chỉ sử dụng dữ liệu bình thường cho xây dựng mô hình. Các kỹ thuật cho xây dựng các bộ phân lớp từ một lớp dữ liệu được gọi là phân đơn lớp (One-class classifications - OCC).

Các phương pháp OCC truyền thống đã chứng minh rất hiệu quả. Gần đây, học sâu đã cho thấy những ưu điểm và phạm vi ứng dụng rộng hơn. Học sâu là thuật ngữ liên quan đến học cách biểu diễn dữ liệu (representation learning) với nhiều tầng, nhiều mức xử lý, là một nhánh của học máy, cho phép tự học đặc tính dữ liệu (feature engineering).

Trong số đó, các phương pháp học sâu dựa trên kiến trúc AutoEncoder (AE) được cho là kỹ thuật tiên tiến (the state-of-the-art) cho phát hiện bất thường mạng. Tiêu biểu như mô hình Shrink AE (SAE), hoạt động dựa trên huấn luyện để chỉ dữ liệu bình thường được co về gốc toạ độ trong không gian vector lớp ẩn. Tuy nhiên cơ chế hoạt động cũng cho thấy SAE vẫn cần được cải tiến, phát triển ở cả ở phần tiền xử lý dữ liệu trước SAE và lõi của SAE: (1) không đạt hiệu quả tốt khi tập dữ liệu cho huấn luyện tồn tại ở dạng nhiều cụm (cluster); (2) SAE gặp khó khăn với một số nhóm tấn công.

Xác định ngưỡng ra quyết định là một bài toán khó khăn với các bộ phân đơn lớp OCC, đây là yêu cầu đối với mô hình NAD khi triển khai trong thực tế.

Các phương pháp phát hiện xâm nhập đơn được cho là thường chỉ hoạt động tốt với một loại tấn công mạng cụ thể.

Vấn đề kết hợp các ưu điểm từ các phương pháp đơn đã được nhiều nghiên cứu thực hiện. Trong dó, Data Fusion (DF), tạm dịch là tổng hợp dữ liệu, trong phạm vi luận án có nghĩa là tổng hợp quyết định từ đa máy phát hiện NAD,

được cho là phù hợp.

Với sự phức tạp của DF, việc xây dựng phương pháp khung cho tổng hợp dữ liệu để giải pháp có tính linh hoạt, khả năng mở rộng và dễ ứng dụng được cho là cần thiết.

Theo đó, phát biểu bài toán luận án gồm ba vấn đề đặt ra: (1) phát triển phương pháp tiêu biểu dựa trên học sâu; (2) nghiên cứu được phương pháp khung NAD dựa trên tổng hợp dữ liệu từ các phương pháp đơn OCC; (3) giải pháp đề xuất cần tiến tới tự động thiết lập ngưỡng ra quyết định.

2. Mục tiêu của luận án

[1] Phát triển phương pháp học sâu NAD.

[2] Phát triển được mô hình khung của NAD dựa trên tổng hợp dữ liệu sử dụng lý thuyết D-S, tổng hợp cả OCC học sâu và truyền thống. Giải pháp có khả năng tự ước lượng ngưỡng quyết định.

3. Đóng góp của luận án

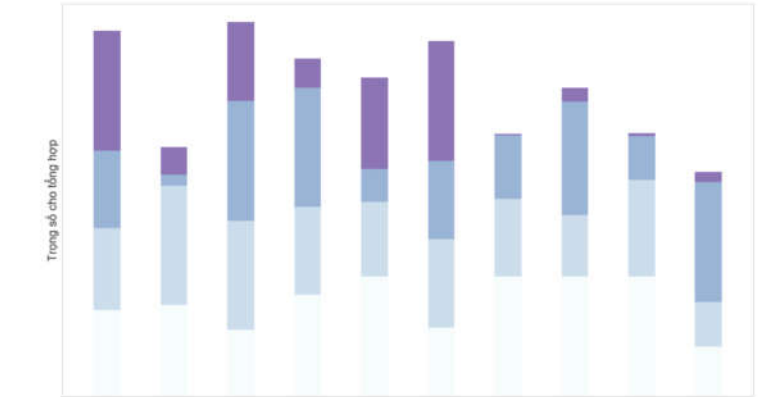
[1] Đề xuất được mô hình NAD dựa trên học sâu có tên Clustering-Shrink AutoEncoder (KSAE) và Double-Shrink AutoEncoder (DSAE).

[2] Luận án đã đề xuất được một mô hình khung OFuseAD, cho phép xây dựng các mô hình NAD dựa trên tổng hợp từ các bộ phân lớp OCC cả truyền thống và học sâu. Giải pháp tự động thiết lập ngưỡng ra quyết định.

[3] Ngoài ra, luận án đã đề xuất mở rộng được hàm DRC của lý thuyết Dempster-Shafer để áp dụng phù hợp hơn cho các bài toán AD.

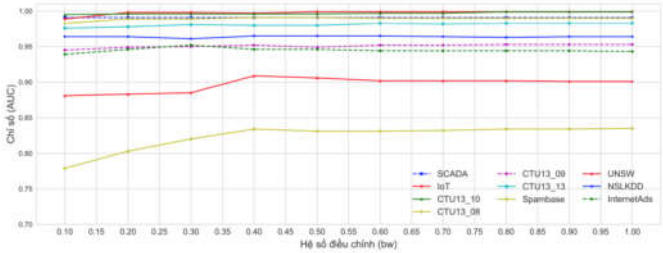
4. Bố cục luận án

Phần mở đầu, giới thiệu tổng quan về luận án. Chương 1, trình bày các kiến thức cơ sở và liên quan đến luận án. Chương 2, trình bày kết quả đề xuất mô hình NAD dựa trên học sâu KSAE và DSAE. Chương 3, trình bày kết quả đề xuất mô hình khung OFuseAD. Phần kết luận, tóm tắt kết quả, một số hạn chế và hướng nghiên cứu.

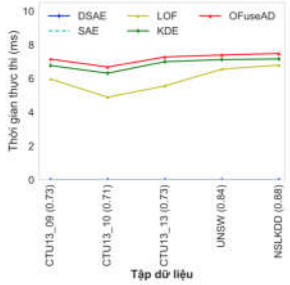


	SCADA	IoT	CTU13_10	CTU13_08	CTU13_09	CTU13_13	Spambase	UNSW	NSLKDD	InternetAds
KDE	1.00000	0.23106	0.66129	0.24444	0.76667	1.00000	0.01072	0.11377	0.02430	0.08265
LOF	0.65079	0.09242	1.00000	1.00000	0.27711	0.65385	0.53333	0.95000	0.36420	1.00000
DSAE_Z1	0.68333	1.00000	0.91111	0.73333	0.62162	0.73913	0.64516	0.51351	0.80822	0.37249
DSAE_Z2	0.71930	0.75934	0.55405	0.84615	1.00000	0.57303	1.00000	1.00000	1.00000	0.41629

Hình 3.8: *Trọng số tham gia tổng hợp của các OCC được tính cho mười tập dữ liệu*

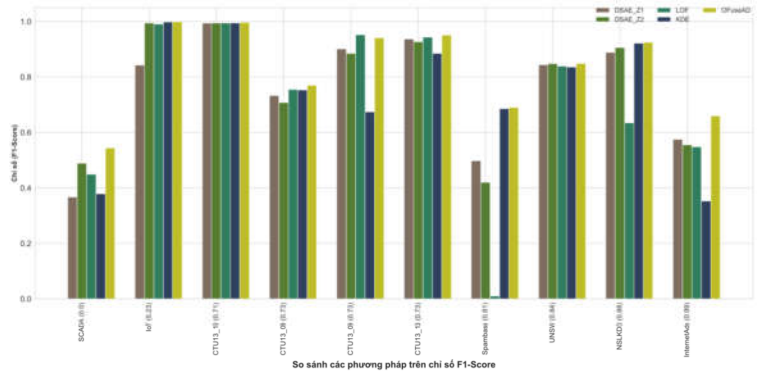


Hình 3.9: *Ảnh hưởng bw đến hiệu quả của OFuseAD.*

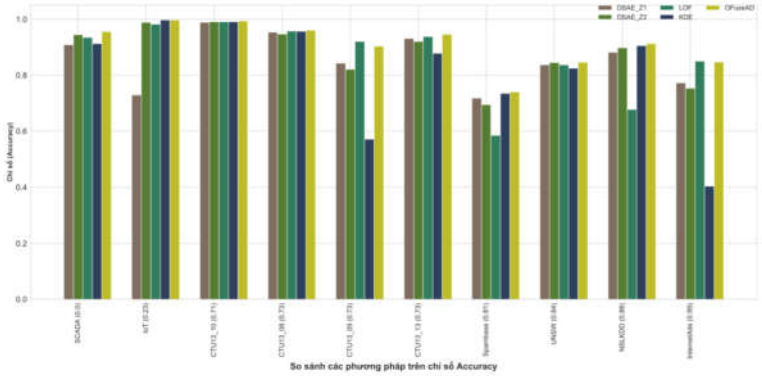


Hình 3.10: *Thời gian truy vấn của các phương pháp khác nhau*

cách và dựa trên mật độ thường cho độ phức tạp lớn, do vậy OFuseAD phải chịu trả giá cho độ phức tạp tính toán.



Hình 3.5: *Biểu đồ so sánh F1-score giữa các phương pháp trên mười tập dữ liệu*



Hình 3.6: *Biểu đồ so sánh ACC giữa các phương pháp trên mười tập dữ liệu*

phân tách giữa bình thường và bất thường khó; OFuseAD có thời gian thực thi phụ thuộc chính vào các phương pháp OCC, gần tương đương với LOF và KDE như trên hình 3.10.

3.5 Kết luận

Mô hình khung OFuseAD hoạt động khả thi, có thể khắc phục được hạn chế của một phương pháp đơn được cho là thường chỉ tốt trên bài toán (tập dữ liệu) cụ thể mà không thực sự hiệu quả trên các bài toán khác. OFuseAD cũng có khả năng tự thiết lập ngưỡng quyết định.

Thực nghiệm trên các bộ dữ liệu phổ biến, OFuseAD cho độ ổn định và khả năng phát hiện bất thường hiệu quả hơn các phương pháp đơn OCC trên hầu hết (9 trên 10) các bộ dữ liệu được thực nghiệm.

Tuy nhiên, việc OFuseAD sử dụng các phương pháp đơn dựa trên khoảng

Chapter 1

TỔNG QUAN VỀ PHÁT HIỆN BẤT THƯỜNG MẠNG

1.1 Hệ thống phát hiện bất thường mạng

1.1.1 Khái niệm

Phát hiện bất thường (Anomaly Detection - AD) là việc tìm ra các mẫu dữ liệu có sự khác biệt so với các mẫu dữ liệu còn lại, các mẫu dữ liệu được phân biệt này thường được gọi là bất thường (anomaly). Trong lĩnh vực an ninh mạng, AD được biết đến với thuật ngữ phát hiện bất thường mạng (Network Anomaly Detection - NAD).

1.1.2 Mô hình phát hiện bất thường mạng

Kiến trúc tổng thể của mô hình phát hiện bất thường mạng có thể được mô tả gồm ba thành phần chính: dữ liệu đầu vào; máy phát hiện bất thường; đầu ra.

Theo đó, lưu lượng mạng sau khi được thu thập, xử lý và trích chọn đặc trưng sẽ được thực hiện tiền xử lý, đưa vào máy phát hiện (bộ phân lớp) bất thường, đây là thành phần chính của hệ thống NAD. Quá trình kiểm thử, kết quả đầu ra của mô hình là độ lệch nhau trên không gian biểu diễn mới giữa mẫu dữ liệu đầu vào và dữ liệu đã được huấn luyện, giá trị này được sử dụng làm cơ sở để phân tách bất thường và bình thường.

1.1.3 Lưu lượng mạng

1.1.4 Đầu ra của mô hình NAD

Cơ bản, có hai dạng đầu ra cho mô hình NAD là: độ đo bất thường (anomaly score - AS); và nhãn nhị phân (binary label - BL). Trong đó, các mô hình phát hiện bất thường hướng đến mục tiêu cho đầu ra là nhãn nhị phân.

Đầu ra AS: mô hình dự đoán sẽ cung cấp một xác suất ứng với mỗi điểm dữ liệu đầu vào, được gọi là độ đo bất thường có giá trị trong khoảng (0,1). Đầu ra BL: mô hình cho dữ liệu đầu ra loại này thường gắn 1 cho trạng thái bất thường và 0 cho trạng thái bình thường của hệ thống mạng đang giám sát; từ độ đo AS, với một ngưỡng quyết định để đạt được đầu ra BL.

1.2 Một số phương pháp đơn cho phát hiện bất thường mạng

Các phương pháp phát hiện bất thường chủ yếu dựa trên thống kê, khai phá dữ liệu và học máy. Việc phân loại các kỹ thuật có nhiều quan điểm khác nhau và các thuật toán cho AD thường có những phần chồng lấn. Các kỹ thuật này thường được phân thành hai nhóm chính là có khả năng tự học (self-learning) hay được lập trình (trang bị kiến thức) rõ từ đầu. Trong số đó, các kỹ thuật phát hiện bất thường dựa trên tự học theo hướng phân đơn lớp OCC được đánh giá là phù hợp và tiềm năng cho lĩnh vực an ninh mạng. Điều này vì các mô hình NAD được cho là phù hợp, có tiềm năng hơn khi chỉ sử dụng mỗi dữ liệu bình thường của hệ thống mạng cho huấn luyện.

Các phương pháp OCC được cho là có thể giải quyết được các vấn đề với không gian thuộc tính dữ liệu quá nhiều chiều (high-dimensional), có thể giúp ước lượng bộ siêu tham số (hyper-parameters) cũng như nâng cao khả năng phân lớp, giúp phát hiện ra các yếu tố mới (chưa từng biết). Các phương pháp OCC có thể được phân thành hai nhóm, phương pháp OCC truyền thống và phương pháp OCC học sâu. Trong các mô hình phát hiện bất thường mạng, các phương pháp OCC thể đóng vai trò như là các phương pháp độc lập, thực thi từ nguyên bản dữ liệu thuộc tính đầu vào hay được đặt phía sau một phương pháp giảm chiều dữ liệu (feature reduction).

1.2.1 Một số phương pháp OCC truyền thống

Các phương pháp OCC truyền thống đã chứng minh rất hiệu quả trong lĩnh vực NAD, trong số đó, một số phương pháp nổi tiếng có thể giải quyết được các vấn đề của dữ liệu mạng như: Local Outlier Factor (LOF) hoạt động hiệu quả trên dữ liệu không gian rất nhiều chiều; Kernel Density Estimation (KDE) có thể tự học mà không cần giả định về phân bố của dữ liệu; One-Class Support Vector Machine (OCSVM) hoạt động phù hợp cho nhiều lĩnh vực ứng dụng khác nhau.

Các phương pháp OCC truyền thống có thể được chia thành các nhóm chính là: phương pháp dựa trên khoảng cách và phương pháp dựa trên mật độ. Ngoài ra, các phương pháp dựa trên vector hỗ trợ có thể được xem là phổ biến và nổi tiếng nhất, phương pháp Centroid (CEN) đơn giản, dễ cài đặt và không cần tham số.

Bảng 3.3: *Kết quả F1-score của các phương pháp trên mười tập dữ liệu*

Phương pháp	Tập dữ liệu									
	SCADA (0.0)	IoT (0.38)	CTU13_10 (0.71)	CTU13_08 (0.73)	CTU13_09 (0.73)	CTU13_13 (0.73)	Spambase (0.81)	UNSW (0.84)	NSLKDD (0.88)	InternetAds (0.99)
DSAE_Z1	0.367	0.843	0.994	0.733	0.901	0.937	0.498	0.844	0.889	0.575
DSAE_Z2	0.489	0.995	0.995	0.708	0.885	0.927	0.420	0.848	0.906	0.555
LOF	0.449	0.991	0.995	0.755	**0.952**	0.944	0.01	0.839	0.635	0.548
KDE	0.379	0.998	0.995	0.753	0.674	0.885	0.686	0.836	0.922	0.353
OFuseAD(ORG) bw=0.5	0.490	.995	0.996	0.758	0.930	0.945	0.687	0.845	0.918	0.508
OFuseAD bw=0.5	**0.544**	**0.999**	**0.997**	**0.770**	0.941	**0.951**	**0.690**	**0.849**	**0.925**	**0.660**

Bảng 3.4: *Kết quả ACC của các phương pháp trên mười tập dữ liệu*

Phương pháp	Tập dữ liệu									
	SCADA (0.0)	IoT (0.38)	CTU13_10 (0.71)	CTU13_08 (0.73)	CTU13_09 (0.73)	CTU13_13 (0.73)	Spambase (0.81)	UNSW (0.84)	NSLKDD (0.88)	InternetAds (0.99)
DSAE_Z1	0.909	0.729	0.989	0.954	0.843	0.931	0.718	0.837	0.882	0.772
DSAE_Z2	0.945	0.989	0.991	0.947	0.821	0.921	0.694	0.845	0.898	0.753
LOF	0.935	0.982	0.991	0.958	**0.921**	0.938	0.585	0.837	0.677	0.850
KDE	0.913	0.997	0.991	0.957	0.571	0.879	0.735	0.825	0.906	0.404
OFuseAD(ORG) bw=0.5	0.942	0.989	0.991	0.957	0.901	0.930	0.725	0.838	0.903	0.824
OFuseAD bw=0.5	**0.956**	**0.997**	**0.994**	**0.961**	0.904	**0.946**	**0.740**	**0.846**	**0.913**	**0.847**

(tập dữ liệu). Mô hình OFuseAD(ORG) thể hiện khá tốt khả năng so với các phương pháp đơn đơn, F1-score này lớn hơn trung bình của tất cả các phương pháp đơn OCC trên cả 10 tập dữ liệu. Thêm vào đó, OFuseAD cho hiệu quả hơn OFuseAD(ORG) trong hầu hết các chỉ số và các tập dữ liệu. Điều này cho thấy hàm DRC_AD đề xuất có lợi thế nhất định với hàm DRC nguyên bản. OFuseAD cho khả năng (theo các chỉ số F1-Score và ACC) tốt hơn trên hầu hết tập dữ liệu (9 trên 10, ngoại trừ tập dữ liệu CTU13_09) khi so sánh với các phương pháp đơn. Kết quả này cũng được thể hiện trên Hình 3.5, 3.6, khối màu vàng chỉ kết quả của mô hình đề xuất, OFuseAD, cho thấy xu hướng cao hơn so với các khối còn lại, và điều này thể hiện ở hầu hết (9/10) tập dữ liệu quan sát. Một phương pháp đơn sẽ cho kết quả phát hiện bất thường cao tại một ngưỡng quyết định của nó t^1; tuy nhiên các SglAD khác sẽ cho kết quả tốt trên ngưỡng khác, là t^2. Giải pháp OFuseAD thực hiện tổng hợp các quyết định từ tất cả các quyết định cục bộ, tại các ngưỡng cục bộ khác nhau. Kết quả thực nghiệm thể hiện, mô hình phát hiện bất thường mạng được xây dựng từ OFuseAD có khả năng tự thiết lập ngưỡng quyết định.

Hình 3.8 thể hiện, cùng một mô hình OCC, nhưng với các bài toán khác nhau thì trọng số tham gia mô hình tổng hợp OFuseAD rất khác nhau.

Khi thực nghiệm với *bw* khác nhau, Hình 3.9 cũng cho thấy rằng, việc thay đổi giá trị của tham số này không ảnh hưởng nhiều đến kết quả tổng thể, khi minh hoạ các đơn vị đo (gồm: AUC, F1-Score, ACC, DR). Khi *bw* ở giá trị tối thiểu 0.5, mô hình hoạt động cho thấy hiệu quả trên các tập dữ liệu đã kiểm thử.

Phương pháp OFuseAD có lợi thế rõ hơn đối với các tập dữ liệu mà việc

Thuật toán 3.3 Giải thuật DRC_AD cho phát hiện bất thường

INPUT: Các danh sách trọng số niềm tin của tập các ADs tham gia l_mA, l_mN, l_mNA, các danh sách trọng số ứng với A, N l_w^A, l_w^N.
OUTPUT: $m(A), m(N), m(NA)$.

Khai báo một danh sách rỗng $l\_mass\_ADs \leftarrow [...]$
Khởi tạo $K \leftarrow$ số phương pháp đơn ADs
Khởi tạo $i \leftarrow 0$
while $i < K$ **do**
  Tính lại trọng số niềm tin ứng với trạng thái:
  Cho trạng thái N $l\_mN[i] \leftarrow l\_mN[i] * l\_w^N[i]$
  Cho trạng thái A $l\_mA[i] \leftarrow l\_mA[i] * l\_w^A[i]$
  Cho trạng thái NA $l\_mNA[i] \leftarrow 1 - (l\_mN[i] + l\_mA[i])$
  $l\_mass\_ADs \leftarrow l\_mN[i], l\_mA[i], l\_mN[i]$
  $i++$
end while
Trọng số niềm tin kết hợp $mass(\Theta) \leftarrow drc\_combine(l\_mass\_ADs)$ 1.15
$m(A), m(N), m(NA) \leftarrow mass(\Theta)(["A"], ["N"], ["NA"])$
Trả về $m(A), m(N), m(NA)$.

Bảng 3.2: *Kết quả AUC của các phương pháp trên mười tập dữ liệu*

Phương pháp	Tập dữ liệu									
	SCADA (0.0)	IoT (0.38)	CTU13_10 (0.71)	CTU13_08 (0.73)	CTU13_09 (0.73)	CTU13_13 (0.73)	Spambase (0.81)	UNSW (0.84)	NSLKDD (0.88)	InternetAds (0.99)
DSAE_Z1	0.991	0.920	0.994	0.985	0.942	0.966	0.840	0.882	0.966	0.958
DSAE_Z2	0.991	0.956	0.992	0.986	0.931	0.972	0.827	0.903	0.963	0.959
LOF	0.993	0.967	0.999	0.982	0.973	0.988	0.743	0.893	0.850	0.831
KDE	0.991	0.999	0.999	0.985	0.808	0.944	0.818	0.888	0.938	0.927
OFuseAD(ORG) bw=0.5	0.993	0.999	0.995	0.985	0.940	0.971	0.828	0.895	0.962	0.944
OFuseAD bw=0.5	0.991	0.999	0.996	0.991	0.949	0.980	0.831	0.906	0.965	0.946

chế trên bởi Thuật toán 3.3.

Thiết lập bw là hệ số điều chỉnh toàn cục, theo điều kiện tại Thuật toán 3.2, bw thuộc [0.5,1].

3.2.2 Cơ chế hoạt động của OFuseAD

Cơ chế hoạt động của OFuseAD có thể được mô tả theo ba công đoạn: (1) Huấn luyện các phương pháp đơn; (2) Huấn luyện phương pháp tổng hợp; (3) Quá trình kiểm tra, như mô tả tại hình 3.1.

3.3 Thực nghiệm

3.3.1 Tập dữ liệu thực nghiệm

3.3.2 Thiết lập tham số thực nghiệm

3.4 Kết quả và đánh giá

Quá trình thử được tiến hành với hai phiên bản: ký hiệu là OFuseAD(ORG), mô hình NAD trong trường hợp này sử dụng hàm kết hợp DRC nguyên bản của lý thuyết D-S; Phiên bản thứ hai sử dụng hàm kết hợp DRC_AD, là mở rộng của DRC theo như luận án đề xuất. Có thể thấy từ Bảng 3.2, 3.3 và 3.4, các phương pháp đơn OCC có hiệu năng khá khác nhau trên cùng một bài toán

Phương pháp OCC dựa trên khoảng cách

Phương pháp OCC dựa trên mật độ

Phương pháp OCC dựa trên vector hỗ trợ

Phương pháp Centroid

1.2.2 Phương pháp OCC học sâu

Học sâu

Học sâu là một nhánh nghiên cứu của học máy, thuật ngữ được nhiều học giả quan tâm trong những năm gần đây, với nhiều định nghĩa khác nhau như tại các nghiên cứu. Các mô hình học sâu có thể được phân làm ba nhóm chính: (1) mô hình sinh (generative model); (2) mô hình phân biệt, (3) mô hình kết hợp. Các phương pháp học sâu cho thấy nhiều lợi thế hơn phương pháp truyền thống trong lĩnh vực phát hiện bất thường. Các mô hình OCC học sâu (Deep - OCC) thuộc nhánh nghiên cứu mô hình generative model, như mạng niềm tin sâu (Deep Belief Network - DBN), mạng nơ-ron hồi quy (Recurrent Neural Network - RNN), và AutoEncoder. Trong số đó, học sâu sử dụng kiến trúc AutoEncoder được nhiều các nghiên cứu gần đây ứng dụng cho lĩnh vực an ninh mạng, được cho là phương pháp tiên tiến về phát hiện bất thường mạng.

AutoEncoder (AE) là một mạng nơ-ron nhân tạo (Artificial Neural Network -ANN), có cấu trúc gồm hai khối: mã hoá (lớp đầu vào) và giải mã (lớp đầu ra). Khối mã hoá ánh xạ dữ liệu đầu vào sang không gian lớp ẩn trung tâm (bottleneck hay vector lớp ẩn). Giả sử f_θ là hàm mã hoá, và $X = \{x_1, x_2, ..., x_n\}$ là tập dữ liệu. Quá trình mã hoá, f_θ sẽ tạo các ánh xạ $x_i \subseteq X$ sang không gian lớp ẩn trung tâm $z_i = f_\theta(x_i)$. Quá trình giải mã, g_θ học để tái tạo dữ liệu đầu ra giống như đầu vào X, $\hat{x}_i = g\theta(z_i)$ từ vector z_i.

Quá trình mã hoá và giải mã thường được trình bày ở dạng hàm số sau: $f\theta(x) = s_f(Wx + b)$ và $g\theta(z) = s_g(W'z + b')$, trong đó $W, W^{'}$ là các ma trận trọng số, b và $b^{'}$ là các ma trận độ lệch, còn s_f và s_g là các hàm kích hoạt tương ứng với quá trình mã hoá và giải mã. Hàm mất mát hay hàm mục tiêu (loss function hay cost function) cho AE.

$$Loss_{AE}(\theta) = \frac{1}{m}\sum_{i=1}^{m}(x_i - \hat{x_i})^2 \tag{1.1}$$

trong đó θ là tập tham số cho AE, m là số mẫu dữ liệu cho huấn luyện.

Một số nghiên cứu liên quan AutoEncoder

Thời gian gần đây, Cao và cộng sự đề xuất một mô hình gọi là (Shrink AutoEncoder - SAE) cho phát hiện bất thường, mô hình cho kết quả tốt trên nhiều tập dữ liệu kiểm thử (datasets), được cho là mô hình tiêu biểu trong lĩnh vực NAD.

Mô hình Shrink AutoEncoder (SAE)

Trong mô hình SAE, một thành phần điều chuẩn (regularizer), điều hướng vector lớp ẩn hội tụ về gốc toạ độ (tâm). Hàm mất mát AE như sau:

$$Loss_{SAE}(\theta) = \frac{1}{m}\left(\sum_{i=1}^{m}(x_i - \hat{x_i})^2 + \alpha\sum_{i=1}^{m}\|z_i\|^2\right) \quad (1.2)$$

trong đó $\hat{x_i}$ và z_i là giá trị tái tạo và vector lớp ẩn ứng với điểm dữ liệu quan sát x_i; m là số mẫu huấn luyện, α là tham số điều chỉnh mức độ cân bằng giữa hai thành phần của hàm mất mát.

1.3 Phát hiện bất thường dựa trên tổng hợp, kết hợp

Việc tổng hợp hay kết hợp các bộ phân lớp đơn để tạo ra bộ phân lớp mới đã được nhiều các nghiên cứu thực hiện, có ba hướng nghiên cứu chính cho việc kết hợp các bộ phân lớp đơn bao gồm: (1) tổng hợp theo lai ghép (hybrid); (2) Tổng hợp theo học cộng đồng (ensemble learning); (3) tổng hợp dữ liệu (data fusion).

1.3.1 Tổng hợp theo lai ghép

Hai hướng đi, lai ghép để giảm chiều dữ liệu; lai ghép để kết hợp một phương pháp signature-based và một phương pháp anomaly-based. Tuy vậy, vấn đề cải tiến khả năng cho phương pháp anomaly-based vẫn là bài toán bỏ ngõ.

1.3.2 Tổng hợp theo học cộng đồng

Có ba chiến lược cho kết hợp: 1) đóng bao (bagging); 2) tăng cường (boosting); 3) xếp chồng (stacking). Mặc dù học cộng đồng đã cho nhiều thành tựu rất lớn, kỹ thuật này được cho là thường phụ thuộc vào nhãn của các phương pháp đơn để kết hợp ra quyết định.

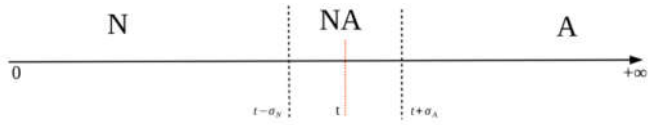
1.3.3 Tổng hợp dữ liệu

Tổng hợp dữ liệu (Data fusion - DF) được định nghĩa như là một công nghệ cho phép kết hợp thông tin từ nhiều nguồn khác nhau để tạo thành một nguồn duy nhất. Trong lĩnh vực phát hiện xâm nhập mạng, DF được định nghĩa là

Thuật toán 3.2 Thiết lập tham số cơ sở OBPA

INPUT: Độ đo AS s_i, ngưỡng t, trọng số cơ sở ($b0$, $p0$, $u0$), độ lệch cực đại (d_N, d_A), hệ số điều chỉnh bw.
OUTPUT: mA, mN, mNA.

Tính giá trị biên phía N $\sigma_N \leftarrow d_N * bw$.
Tính giá trị biên phía A $\sigma_A \leftarrow d_A * bw$.
if $s_i \leqslant t$ **then**
  Tính hệ số ứng với 'Vùng N' $b1 \leftarrow (t - s_i)/\sigma_N$, with $b1 \leq 2$.
  Tính trọng số ứng với trạng thái N, $mN \leftarrow b1 * b0$.
  Trọng số ứng với trạng thái A, $mA \leftarrow u0$.
  Trọng số ứng với trạng thái NA, $mNA \leftarrow (1 - mN - mA)$.
else
  Tính hệ số ứng với 'Vùng A', $b1 \leftarrow s_i/(\sigma_A - t)$, with $b1 \leq 2$.
  Tính trọng số ứng với trạng thái A, $mA \leftarrow bb * b0$.
  Tính trọng số ứng với trạng thái N, $mN \leftarrow u0$.
  Tính trọng số ứng với trạng thái NA, $mNA \leftarrow (1 - mN - mA)$.
end if
Trả về mA, mN, mNA



Hình 3.2: *Ba vùng trên trục độ đo bất thường N, A và NA*

trong đó m là số lượng các OCC tham gia mô hình tổng hợp. Giá trị trọng số này càng lớn, đóng góp của OCC tương ứng vào quá trình tổng hợp càng lớn.

Ứng dụng lý thuyết Dempster-Shafer: Định nghĩa FoD, từ thực tế lĩnh vực phát hiện bất thường mạng, giải pháp phát hiện phải chỉ ra hai trạng thái của hệ thống bình thường hay bất thường, do vậy ký hiệu N để chỉ trạng thái bình thường của hệ thống; và A để chỉ trạng thái bất thường của hệ thống. Hàm FoD được định nghĩa, $\Theta = \{A, N\}$, và tập giả thuyết đầy đủ $P(\Theta) = (A, N, NA, \emptyset)$, trong đó $N \cap A = \emptyset$.

Vấn đề xây dựng hàm BPA theo lý thuyết D-S là tìm giải pháp để $m(A) + m(NA) + m(N) = 1$, luận án đề xuất hàm OBPA như được mô tả tại Thuật toán 3.2.

Trong đó, s_i chỉ giá trị AS cho điểm dữ liệu đầu vào x_i do phương pháp đơn tương ứng tạo ra. Thuật toán sử dụng các khái niệm như, tham số cơ bản của niềm tin (belief) $b0$, độ hợp lý (plausibility) $p0$, và không tin (unbelief) $u0$, các tham số cơ bản này ứng với trạng thái của hệ thống là N, A hay N/A, được mô tả theo các vùng trạng thái như trên Hình 3.2.

Khi áp dụng hàm kết hợp DRC truyền thống của lý thuyết D-S, tất cả phương pháp đơn AD nên có cùng vai trò, nhưng thực tế các phương pháp phát hiện bất thường thường có năng lực khác nhau. Đề xuất khắc phục hạn

Thuật toán 3.1 Thiết lập ngưỡng tự động cho OCC Gen_Threshs

INPUT: Bộ phân lớp OCC f_j, tập kiểm thử va, cận dưới vùng xem xét p_{lower} và cận trên p_{upper}..
OUTPUT: $auto\_thresh$.

Kiểm tra với bộ phân lớp OCC $S^{va} \leftarrow f_i(va)$
Sắp xếp lại $S^{va} \leftarrow sort(S^{va}, ascending\ order)$
Giá trị AS ứng với cận dưới $s_{lower} \leftarrow S^{va}[p_{lower}]$
Giá trị AS ứng với cận trên $s_{upper} \leftarrow S^{va}[p_{upper}]$
Thiết lập mật độ khởi tạo $vol \leftarrow size(S^{va})$
Thiết lập số lần chạy $num\_interval \leftarrow 2$
repeat
  kích thước một vùng $interval \leftarrow (s_{upper} - s_{lower})/num\_interval$
  $k \leftarrow 0$
  for $k \leftarrow 0$ to $num\_interval - 1$ **do**
    $s_1 \leftarrow s_{lower} + interval * (k)$
    $s_2 \leftarrow s_{upper} + interval * (k + 1)$
    $vol \leftarrow size[s_1, s_2]$
    if $vol == 0$ **then**
      hoàn thành việc xác định ngưỡng, break
    end if
  end for
  $num\_interval + +$
until $vol! = 0$
$thresh \leftarrow s_{lower} + k * interval$
Trả về $auto\_thresh$

Chỉ số này tạm gọi là mức độ sinh lỗi (generalization error, ký hiệu gen_error), được định nghĩa như tại Công thức 3.1.

$$gen\_error_j = \frac{1}{k}\sum_{i=1}^{k}\left|TN^i_{train} - TN^i_{va}\right| \tag{3.1}$$

Với k là số ngưỡng được lấy, t_i là ngưỡng thứ i-th, và TN^i_{train}, TN^i_{va} là giá trị TN của bộ phân lớp thứ j trên tập huấn luyện và tập kiểm thử tương ứng. TN^i_{va} được tính theo Công thức 3.2.

$$TN^i_{va} = \frac{\sum(f_j(va) \leqslant t_i)}{|va|} \tag{3.2}$$

Trọng số phương pháp đơn khi tham gia tổng hợp: Các phương pháp đơn khi tham gia tổng hợp có mối liên hệ chặt chẽ với nhau vì chúng cùng được huấn luyện và quan sát cùng một đối tượng dữ liệu.

Trọng số của OCC thứ j có thể được tính toán theo Công thức 3.3, và co dãn về $[0, 1]$.

$$w_j = \frac{min\left[gen\_error_1, gen\_error_2.., gen\_error_m\right])}{gen\_error_j} \tag{3.3}$$

việc xử lý của một nguồn hoặc nhiều nguồn dữ liệu được thu thập từ mạng để cho kết quả đánh giá tốt hơn. Ba mức mô hình tổng hợp: tổng hợp mức dữ liệu (data fusion layer), tổng hợp mức thuộc tính(feature fusion), và tổng hợp mức quyết định (decision fusion layer). Mức quyết định: Hoạt động ở mức này giúp cho hệ thống DF có tính linh động, giảm chi phí tính toán, tận dụng được sức mạnh của các bộ phát hiện đơn, các kỹ thuật đã hiện hữu.

1.3.4 Tổng hợp dữ liệu dựa trên lý thuyết Dempster-Shafer

Lý thuyết Dempster-Shafer

Tập hữu hạn giả thuyết hay tập các nhận định không chắc chắn có thuật ngữ là Frame of Discernment (FoD) và được ký hiệu Θ, đây là tập toàn bộ m trạng thái hay giả thuyết độc lập có thể xảy ra đối với hệ thống đang xem xét.

$$\Theta = \{H_1, H_2, ...H_m\} \tag{1.13}$$

Hàm gán xác suất cơ bản (Basic Probability Assignment - BPA) qua tập Θ là hàm $m : 2^{\Theta} \rightarrow [0, 1]$ với điều kiện sau:

$$\sum\{m(H))|H \subseteq \Theta\} = 1, m(\emptyset)) = 0 \tag{1.14}$$

Khi áp dụng DRC kết hợp hai nguồn E_i và E_j, trọng số niềm tin kết hợp $m(H)$ thu được, $m(H)$= $(E_i \oplus E_j)(H)$, chi tiết như Công thức sau:

$$m(H) = \frac{\sum_{(B\cap C=H;B,C\subseteq\Theta)}\left[m_i\left(B\right)m_j\left(C\right)\right]}{1 - \sum_{(B\cap C=\emptyset;B,C\subseteq\Theta)}\left[m_i\left(B\right)m_j\left(C\right)\right]} \tag{1.15}$$

DRC là công cụ để kết hợp trọng số được gắn từ nhiều nguồn quan sát (hay các dẫn chứng).

Một số nghiên cứu ứng dụng lý thuyết D-S

1.4 Đánh giá giải pháp

Để thực nghiệm đánh giá một giải pháp đề xuất trong lĩnh vực phát hiện bất thường, hai yếu tốt chính cần quan tâm gồm bộ dữ liệu cho kiểm thử và các chỉ số đánh giá được sử dụng.

1.4.1 Bộ dữ liệu cho kiểm thử (datasets)

Các bộ dữ liệu được sử dụng cho thực nghiệm trong luận án đều phổ biến trong lĩnh vực an ninh mạng, gồm: Bộ dữ liệu NSL KDD; UNSW-NB15; các

phiên bản của bộ CTU13; BoT-IoT; Spambase; InternetADs; và WUSTL-IIOT-2018 ICS (SCADA)

Kỹ thuật mã hoá one-hot (One-hot) được sử dụng cho các trường dữ liệu tập hợp (categorical features). Để phù hợp cho kiểm thử các thuật toán OCC, tập huấn luyện như đã trình bày đảm bảo chỉ chứa dữ liệu một lớp (dữ liệu bình thường). Quá trình huấn luyện, toàn bộ nhãn trong các tập huấn luyện được bỏ ra. Trong tập kiểm tra, tất cả các nhóm tấn công của các tập dữ liệu tương ứng đều được gán nhãn là bất thường, giá trị bằng 1, dữ liệu bình thường có nhãn bằng 0. Toàn bộ nhãn trong tập kiểm tra đều được bỏ đi tại thời điểm kiểm tra, nhãn này chỉ sử dụng sau khi hoàn tất quá trình thực nghiệm,

1.4.2 Các chỉ số đánh giá

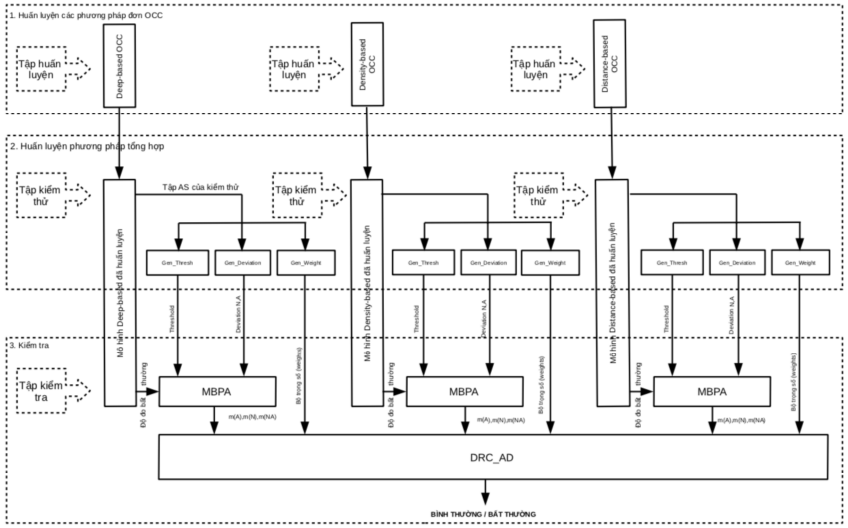
Chỉ số đánh giá với đầu ra là nhãn nhị phân (gồm: Độ chính xác (Accuracy - ACC), Ma trận lỗi (Confusion Matrix), Tỉ lệ phát hiện và tỉ lệ cảnh báo sai, Độ đo F1-Score). Trong đó, F1-score được cho là phù hợp cho đánh giá hiệu quả của mô hình trong điều kiện bài toán. Chỉ số đánh giá với đầu ra là độ đo bất thường, gồm(Đường cong ROC và AUC); về cơ bản AUC thường được sử dụng cho đánh giá các mô hình học máy khi chưa xác định được ngưỡng quyết định. Chỉ số Độ ổn định (nếu chỉ số quan sát được có giá trị ổn định hơn, thì tốt hơn).

1.5 Kết luận

Trong lĩnh vực NAD, phương pháp học sâu dựa trên AutoEncoder được cho là tiên tiến; mô hình SAE tiêu biểu, tuy vậy SAE vẫn có thể gặp phải một số hạn chế cần được cải tiến.

Phương pháp tổng hợp dữ liệu (Data Fusion) là phù hợp cho mục tiêu luận án đề ra vì có thể gom được lợi thế từ các phương pháp đơn khác nhau; lý thuyết Dempster-Shafer (D-S) được đánh giá là phù hợp cho bài toán phát hiện bất thường nhờ sự linh hoạt và không yêu cầu tri thức tiền định khi xây dựng mô hình.

Các mô hình NAD có thể được kiểm thử bởi các bộ dữ liệu phổ biến trong lĩnh vực an ninh mạng; các chỉ số đánh giá được phân nhóm theo dạng đầu ra của mô hình NAD, ngoài ra có chỉ số cho đánh giá sự ổn định của mô hình.



Hình 3.1: *Kiến trúc của giải pháp OFuseAD*

3.2 Giải pháp đề xuất

Mô hình khung NAD dựa trên tổng hợp dữ liệu được đề xuất có tên OFuseAD (One-class Fusion-based Anomaly Detection Framework). Hình 3.1 minh hoạ kiến trúc của OFuseAD.

3.2.1 Các thành phần của phương pháp OFuseAD

Lựa chọn các phương pháp đơn AD: OFuseAD đề xuất chọn các phương pháp đơn từ ba kỹ thuật phát hiện bất thường rất khác nhau: dựa trên học sâu (deep learning-based); dựa trên khoảng cách (distance-based); và dựa trên mật độ (density-based).

Xác định ngưỡng cho phương pháp đơn OCC: Mô hình phát hiện bất thường trên OCC thường cho đầu ra là AS từ các điểm dữ liệu đầu vào, ngưỡng được đưa vào để quyết định mẫu đầu vào là bất thường hay không. Luận án giới thiệu một phương pháp tự động ước lượng ngưỡng cho các phương pháp đơn OCC khi tham gia mô hình tổng hợp dữ liệu, chỉ sử dụng dữ liệu bình thường. Như mô tả tại Thuật toán 3.1.

Độ tin cậy của phương pháp đơn OCC: Mỗi bộ phân lớp đơn trong mô hình tổng hợp được gán trọng số niềm tin, chỉ khả năng của bộ phân lớp về mức độ hiệu qua trong phân tách các lớp dữ liệu. Luận án đề xuất một giải pháp cho ước lượng trọng số niềm tin này từ tập kiểm thử chứa chỉ dữ liệu bình thường.

Chapter 3

PHÁT HIỆN BẤT THƯỜNG DỰA TRÊN TỔNG HỢP DỮ LIỆU

3.1 Giới thiệu

Mặc dù có các phương pháp đơn rất mạnh cho lĩnh vực phát hiện bất thường mạng thì vẫn nổi lên hai hạn chế chung theo sau: (i) các phương pháp đơn nhìn chung thường hoạt động rất hiệu quả trên một số bài toán (tập dữ liệu cụ thể) mà lại không hiệu quả trên các bài toán khác; (ii) thiếu cơ sở để chọn ngưỡng phù hợp cho ra quyết định đối với mô hình NAD phát triển theo hướng OCC.

Vấn đề phương pháp đơn có thể được giải quyết trên cơ sở kết hợp điểm mạnh từ các phương pháp đơn khác nhau. Data Fusion (DF) được cho là phù hợp trong điều kiện bài toán OCC. Các yếu tố cần phải được xem xét khi phát triển một hệ thống DF gồm: (1) xác định mức hoạt động của DF, có ba mức tổng hợp là: mức dữ liệu, mức thuộc tính, và mức quyết định. Trong đó, tổng hợp mức quyết định thường được các nhà nghiên cứu lựa chọn nhờ tính linh hoạt và tính phù hợp cho các bài toán thực tế, có thể tận dụng được lợi thế từ các phương pháp đơn vốn đã rất hiệu quả; (2) cơ sở nào để chọn được các phương pháp đơn; (3) lựa chọn thuật toán cho tổng hợp (fusion algorithm) nào?, đây là nội dung được cho là cơ bản nhất khi xây dựng hệ thống DF.

Lý thuyết Dempster-Shafer (D-S) có nhiều tiềm năng cho phát triển các mô hình tổng hợp dữ liệu trong lĩnh vực an ninh mạng. Khi áp dụng D-S, việc đề xuất hàm BPA thường khó và phức tạp. Thêm vào đó, để xử lý với vấn đề tổng hợp dữ liệu từ các phương pháp đơn có độ tin cậy khác nhau, việc áp dụng D-S đòi hỏi phải cải tiến hàm kết hợp DRC của lý thuyết. Trong điều kiện bài toán OCC, việc ứng dụng DF gặp thách thức lớn như: vấn đề xác định ngưỡng quyết định; làm thế nào để xác định trọng số độ tin cậy của từng phát biểu.

Việc ứng dụng lý thuyết D-S cho cho phép gán trọng số niềm tin cho từng trạng thái, do vậy có thể cung cấp đầu ra là BL thay vì AS như các phương pháp đơn OCC.

Chapter 2

PHÁT HIỆN BẤT THƯỜNG DỰA TRÊN HỌC SÂU AUTOENCODER

2.1 Giới thiệu

Mô hình học sâu dựa trên kiến trúc AutoEncoder (Deep AutoEncoder - DeAE), hình thành từ việc sử dụng AE với nhiều lớp ẩn.

Mô hình Shrink AE (SAE) được cho là mô hình tiêu biểu trong phát hiện bất thường mạng. Phương pháp này hiện vẫn có thể gặp những hạn chế nhất định. (i) SAE có thể đạt hiệu quả không cao với trường hợp đối tượng quan sát có dữ liệu trạng thái bình thường tồn tại ở dạng nhiều cụm. (ii) Mô hình SAE có thể gặp khó khăn với một số loại tấn công, các tấn công này tạo ra các vector lớp ẩn có xu hướng gần gốc toạ độ hơn, có thể do mẫu dữ liệu có nhiều điểm giống với mẫu dữ liệu bình thường.

2.2 Giải pháp đề xuất

Hạn chế thứ nhất nằm ở việc vấn đề xử lý dữ liệu trước khi đẩy vào SAE; hạn chế thứ hai hoàn toàn nằm trong phần lõi SAE, luận án đề xuất hai giải pháp tương ứng có tên KSAE, DSAE.

2.2.1 Giải pháp Clustering-Shrink AutoEncoder

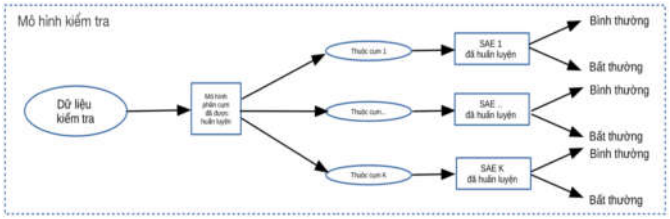
Mô hình KSAE được hình thành từ kết hợp thuật toán phân cụm và SAE. Phân cụm là chia dữ liệu thành các nhóm đối tượng tương đương, luận án chọn K-means đại diện cho bước phân cụm trong mô hình học sâu KSAE.

Quá trình huấn luyện mô hình KSAE gồm hai công đoạn: Thứ nhất, dữ liệu đầu vào được phân cụm sử dụng thuật toán phân cụm (TTPC), thuật toán này được huấn luyện để chia tập dữ liệu theo số cụm K, cho trước. Thứ hai, ứng với số cụm K được chia tách, các mô hình SAE được huấn luyện bởi chỉ dữ liệu ứng với cụm dữ liệu tương ứng thu được từ bước thứ nhất. Thuật toán 2.1 trình bày chi tiết quá trình huấn luyện của KSAE. Quá trình kiểm tra KSAE như mô tả tại Hình 2.3, trong mô hình kiểm tra này, các mẫu dữ liệu đầu vào đầu tiên được kiểm tra để xác định số cụm bởi mô hình phân cụm đã được huấn luyện, kết quả trả về là nhãn $C_j \leq K$, ứng với cụm của dữ liệu đầu vào. Mô hình SAE_j tương ứng sau đó được sử dụng cho kiểm tra để xác định độ đo

Thuật toán 2.1 Huấn luyện mô hình KSAE

INPUT: Tập huấn luyện D_n, số cụm cho trước K.
OUTPUT: $trained^{TTPC}$, K $trained^{SAE}$.

1: $trained^{TTPC} \leftarrow$ huấn luyện thuật toán phân cụm với đầu vào D_n, K.
2: K tập huấn luyện $D^j \leftarrow$ kiểm tra $trained^{TTPC}$ với đầu vào D_n.
3: $j \leftarrow 0$.
4: **while** $j < K$ **do**
5: $\quad trained_j^{SAE} \leftarrow$ huấn luyện SAE với tập dữ liệu D^j.
6: **end while**
7: Trả về $trained^{TTPC}$, K $trained^{SAE}$.



Hình 2.3: *Mô hình kiểm tra theo phương pháp KSAE*

bất thường ứng với điểm dữ liệu đầu vào. Độ đo AS được sử dụng cho đánh giá mô hình theo như cách SAE vẫn thực hiện.

2.2.2 Giải pháp Double-shrink AutoEncoder

Ý tưởng cho giải pháp DSAE (Double-Shrink AutoEncoder), với các bất thường mà SAE gặp khó, có hai trường hợp cho vector tái tạo đầu ra: (1) cho lỗi tái tạo (RE) nhỏ, khi đó mẫu bất thường đầu ra (được tái tạo) sẽ gần giống với mẫu bất thường đầu vào, và là bất thường. So với mẫu bất thường đầu vào, mẫu dữ liệu được tái tạo này có thể khác xa hơn mẫu dữ liệu bình thường; (2) nếu RE lớn thì mẫu bất thường được tái tạo, X-out, có xu thế khác xa hơn so với mẫu bất thường đầu vào, nghĩa là khác xa hơn so với mẫu dữ liệu bình thường. Do vậy, sử dụng thêm dữ liệu của vector tái tạo đầu ra, X-out, có thể giúp cho việc phân tách dữ liệu bình thường và bất thường hiệu quả hơn. Hình 2.4 mô tả quá trình huấn luyện và kiểm tra mô hình DSAE. Mô hình được huấn luyện để đồng thời đạt được các mục tiêu giảm thiểu lỗi tái tạo lần co 1 (RE_1) và lần co 2 (RE_2) đồng thời điều hướng các vector z^1 và z^2 về gốc toạ độ trong không gian lớp ẩn. Quá trình kiểm tra, z^1, z^2 được sử dụng như vector đặc trưng đại diện cho dữ liệu đầu vào gốc. Mô hình DSAE mặc định sử dụng vector z^2 cho biểu diễn dữ liệu đầu vào.

Hàm mất mát DSAE có thể được mô tả

$$L_{RE_1}(\theta, x_i) = \frac{1}{m}\sum_{i=1}^{m}\left(x_i - x_i^{out1}\right)^2 \tag{2.1}$$

Khi minh hoạ trên biểu đồ hai chiều (2-D), các toạ độ x_i và y_i của mỗi vector lớp ẩn z_i. Từ các Hình 2.7, 2.8, 2.9 và 2.10 cho thấy: các mẫu tấn công khó với SAE có vector lớp ẩn xu hướng bị đẩy ra xa gốc toạ độ hơn khi thực thi bởi DSAE; nhóm tấn công R2L cho phân bố vector lớp ấn khác hơn, mật độ dày đặc theo hướng gần gốc toạ độ hơn, điều này cũng phù hợp với nhận định, R2L có dữ liệu giống với lưu lượng mạng bình thường và làm cho R2L thường khó bị phát hiện hơn.

Khi quan sát riêng nhóm tấn công R2L, Bảng 2.5 cho thấy, DSAE thể hiện hiệu quả hơn SAE ở cả hai chỉ số DR và FAR. Kết quả trong thực nghiệm cho thấy DSAE có hiệu quả tương đồng với SAE trên các tập dữ liệu loại tấn công mạng phổ biến. Nhưng DSAE cho thấy khả năng phát hiện bất thường hiệu quả hơn SAE, trên loại tấn công R2L. Tuy vậy, DSAE vẫn có hạn chế, một số mẫu dữ liệu bình thường vốn dĩ đã được lần co thứ nhất thực hiện đúng, nhưng ở lần co sau đã bị SAE phân tách sai. Nhưng về tổng thể, đối với nhóm tấn công R2L kết quả cho thấy số lượng bị phân tách sai của DSAE ít hơn khá nhiều so với số lượng bị phân lớp sai bởi SAE. Theo hoạt động của DSAE, việc kết hợp cả hai đầu ra của là z_1 và z_2 (ký hiệu hai mô hình tương ứng là DSAE_Z1 và DSAE_Z2) vẫn còn cho thấy tiềm năng.

Khi thực thi, độ phức tạp tính toán so sánh thông qua thời gian kiểm tra của SAE và DSAE trên các tập dữ liệu cho kết quả gần tương đồng nhau.

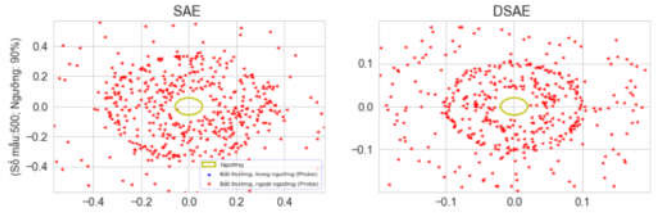
2.5 Kết luận

Kết quả nghiên cứu đã đề xuất được giải pháp khắc phục hai hạn chế mà phương pháp tiêu biểu đang gặp phải, cụ thể: Giải pháp KSAE, để khắc phục hạn chế của SAE được cho là không hiệu quả với các tập dữ liệu hiện hữu ở dạng nhiều cụm; giải pháp DSAE, phát triển mở rộng nội tại nhân SAE để có thể phân tách tốt hơn các bất thường mà SAE gặp khó. Các giải pháp đã được kiểm chứng qua thực nghiệm trên các tập dữ liệu phổ biến trong lĩnh vực an ninh mạng.

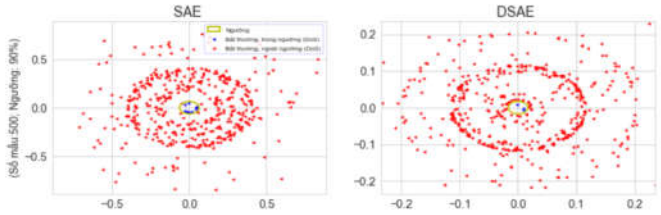
Hai giải pháp KSAE và DSAE độc lập để xử lý hai vấn đề tại hai bước khác nhau của SAE (tiền xử lý trước SAE và nhân SAE), do vậy hai giải pháp có thể triển khai đồng thời để cải tiến hiệu quả của SAE.

Bảng 2.6: *Kết quả DSAE phân tách các nhóm tấn công SAE có thể gặp khó*

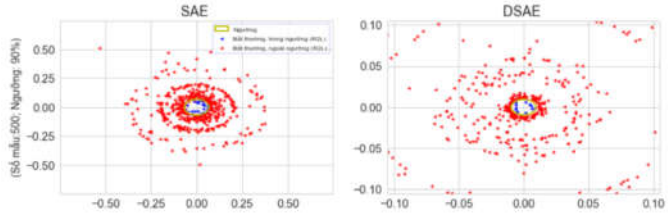
Nhóm tấn công	Tấn công SAE gặp khó đúng			Các tấn công SAE khó, được DSAE phân tách đúng	
	Tổng	RE bé	RE lớn	Re bé	RE lớn
Probe	0	0	0	0	0
DoS	434	434	0	327	0
R2L	146	120	26	95	26
U2R	6	6	0	3	0
Tổng	586	560	26	418	26



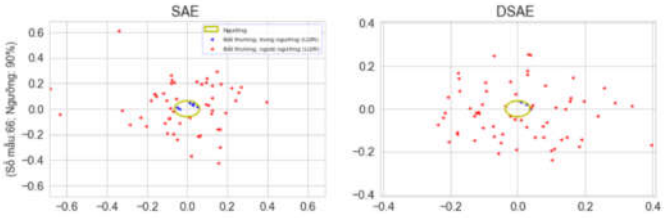
Hình 2.7: *Không gian lớp ẩn nhóm tấn công Probe trên SAE, DSAE*



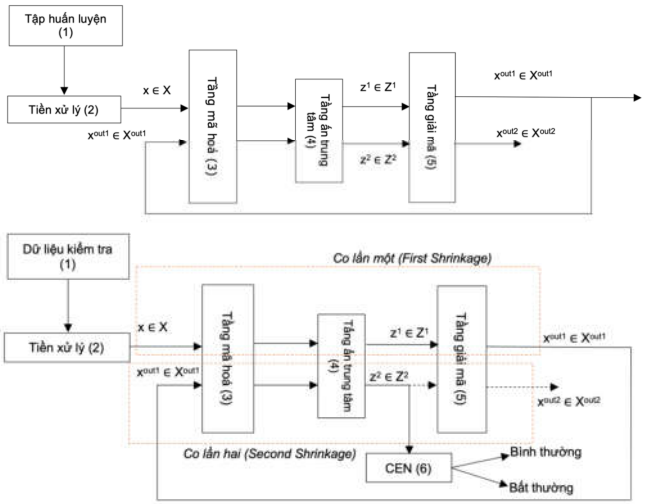
Hình 2.8: *Không gian lớp ẩn nhóm tấn công DoS trên SAE, DSAE*



Hình 2.9: *Không gian lớp ẩn nhóm tấn công R2L trên SAE, DSAE*



Hình 2.10: *Không gian lớp ẩn nhóm tấn công U2R trên SAE, DSAE*



Hình 2.4: *Mô hình Double-shrink AutoEncoder*

$$L_{RE_2}(\theta, x_i) = \frac{1}{m}\sum_{i=1}^{m}\left(x_i^{out1} - x_i^{out2}\right)^2 \tag{2.2}$$

Theo đó, thành phần REs của hàm mất mát theo phương pháp DSAE có thể trình bày theo Công thức 2.3.

$$L_{RE}(\theta, x_i) = L_{RE_1}(\theta, x_i) + L_{RE_2}(\theta, x_i) \tag{2.3}$$

Còn thành phần điều chuẩn co trong hàm mất mát của phương pháp DSAE là tổng của hai lần co và được biểu diễn bởi Công thức 2.4, 2.5 và 2.6:

$$L_{Z_1}(\theta, x_i) = \frac{1}{m}\sum_{i=1}^{m}\left\|z_i^1\right\|^2 \tag{2.4}$$

$$L_{Z_2}(\theta, x_i) = \frac{1}{m}\sum_{i=1}^{m}\left\|z_i^2\right\|^2 \tag{2.5}$$

$$L_Z(\theta, x_i) = L_{Z_1}(\theta, x_i) + L_{Z_2}(\theta, x_i) \tag{2.6}$$

trong đó z_i^1 và $z_i 2$ là các vector ẩn của dữ liệu đầu vào x_i tương ứng tại lần co thứ nhất và thứ hai, m là số điểm dữ liệu đầu vào. Từ trình bày trên, hàm mất mát của DSAE được viết lại như sau:

$$L_{DSAE}(\theta, x_i) = L_{RE}(\theta, x_i) + \alpha * L_Z(\theta, x_i) + \beta * L_W(\theta, x_i) \tag{2.8}$$

Bảng 2.2: *Kết quả AUC của KSAE trên các tập dữ liệu*

K	Tập dữ liệu				
	NSL-KDD	**UNSW**	**CTU13-09**	**CTU13-10**	**CTU13-13**
K=1	0.941	**0.887**	0.923	**0.998**	0.931
K=2	**0.962**	0.885	0.935	0.989	0.933
K=3	0.879	0.858	**0.946**	0.965	**0.962**

trong đó, thành phần cuối cùng để kiểm soát việc suy giảm trọng số (weight decay regularizer) cho bộ trọng số của mạng nơ-ron, W; α và β điều khiển sự cân bằng giữa ba thành phần của hàm mất mát DSAE.

2.3 Thực nghiệm

2.3.1 Dữ liệu thực nghiệm

2.3.2 Phương pháp xác định số cụm tối ưu

Có nhiều phương pháp để xác định số cụm tối ưu, phổ biến trong đó là Elbow, cho phép xác định số K tối ưu dựa vào trực quan trên biểu đồ 2D. Số K tối ưu được xác định ứng với điểm tại đó trục x và đồ thị tạo nên khủy tay.

2.3.3 Thiết lập tham số thực nghiệm

2.4 Kết quả và đánh giá

Đánh giá dựa trên kết quả thực nghiệm KSAE: Với bước thực nghiệm đầu tiên để so sánh KSAE và SAE, vì chưa có cơ sở xác định ngưỡng quyết định do vậy chỉ số AUC (Area Under the Curve) được sử dụng để đánh giá. Các bộ dữ liệu được chia thành số cụm $K = (1, 2, 3)$ khi thực nghiệm; trường hợp K=1, mô hình KSAE hoàn toàn đồng nhất với SAE.

Có thể nhận thấy rằng, với một giá trị K phù hợp, mô hình đề xuất có khả năng phát hiện cải tiến so với SAE. Bảng 2.2 cho thấy, mô hình đề xuất có kết quả tốt hơn trên 3 tập dữ liệu; trên hai bộ dữ liệu còn lại (UNSW-NB15 và CTU13_10), hiệu năng của KSAE không tốt bằng SAE.

Kết quả thực nghiệm theo phương pháp Elbow trên 5 bộ dữ liệu, qua 5 lần thử cho thấy độ ổn định ở các lần thử khác nhau. Phương pháp Elbow cho thấy, kết quả tương đối đồng nhất với kết quả thể hiện tại Bảng 2.2.

Đánh giá dựa trên kết quả thử nghiệm DSAE:

Thực nghiệm thứ nhất để so sánh khả năng phát hiện bất thường của DSAE với các mô hình tiên tiến NAD sử dụng mạng nơ-ron học sâu, bao gồm SAE và Denoising AutoEncoder (DAE). Thực nghiệm thứ hai để đánh giá mức độ hiệu quả của hai mô hình DSAE và SAE trên các nhóm tấn công mạng có thể

Bảng 2.3: *AUC từ các mô hình DAE, SAE, DSAE trên sáu tập dữ liệu*

Phương pháp	Tập dữ liệu					
	NSLKDD	**UNSW**	**CTU13-08**	**CTU13-09**	**CTU13-10**	**CTU13-13**
DAE* + CEN	0.854 ±0.002	0.690 ±0.001	0.938 ±0.015	0.655±0.031	0.951±0.006	0.711±0.002
DAE+RE	0.930±0.090	0.873±0.004	0.960±0.011	0.903±0.002	0.958±0.004	0.952±0.010
SAE**+CEN	0.960 ±0.002	0.896 ±0.006	0.982 ±0.009	0.940 ±0.010	0.997 ±0.001	0.964 ±0.012
SAE+RE	0.920 ±0.000	0.810 ±0.001	0.951 ±0.013	0.703 ±0.020	0.997 ±0.000	0.887 ±0.005
DSAE + CEN	**0.963** ±0.004	0.895 ±0.015	**0.986** ±0.012	0.929 ±0.054	0.992 ±0.008	**0.971** ±0.006

* DAE: Denoising AutoEncoder; ** SAE: Shrink AutoEncoder [20]

Bảng 2.4: *AUC từ SAE, DSAE trên bốn nhóm tấn công của tập dữ liệu NSL-KDD*

Phương pháp	Tập dữu liệu			
	Probe	**DoS**	**R2L**	**U2R**
SAE + CEN *	0.977 ±0.003	0.967 ±0.002	0.924 ±0.010	0.956 ±0.005
DSAE + CEN	0.979 ±0.006	0.966 ±0.007	**0.936** ±0.011	0.960 ±0.010

* SAE + CEN: Shrink AutoEncoder và Centroid [20]

Bảng 2.5: *Kết quả DR, FAR giữa SAE và DSAE trên nhóm tấn công R2L*

Phương pháp	Dữ liệu nhóm tấn công R2L					
	TP	**FP**	**FN**	**TN**	**FAR**	**DETECTION RATE**
SAE + CEN *	1892	1008	995	8702	0.104	0.655
DSAE + CEN	2011	989	876	8721	**0.102**	**0.697**

* SAE + CEN: Shrink AutoEncoder và Centroid [20]

SAE gặp khó.

Nhìn chung, chỉ số AUC (Area Under the ROC Curve) được sử dụng cho đánh giá. Ngoài ra, khi đánh giá mô hình ở các ngưỡng cụ thể, các chỉ số TP, FP, FN, TN và cặp chỉ số DR, FAR được sử dụng để so sánh tính hiệu quả của các mô hình trong khả năng phát hiện đối với nhóm tấn công cụ thể.

Kết quả thực nghiệm thứ nhất, kiểm tra DSAE, SAE và DAE trên 6 bộ dữ liệu được trình bày tại Bảng 2.3. Có thể thấy rằng, AUC của SAE và DSAE tương đương nhau trên hầu hết các tập dữ liệu được kiểm thử, và giá trị này tốt hơn DAE.

Kết quả thực nghiệm thứ hai tại Bảng 2.4 chứng tỏ DSAE cho khả năng phát hiện so sánh được với SAE trên ba nhóm tấn công (Probe, DoS, và U2R). Tuy nhiên, trên nhóm tấn công khó nhất là R2L, DSAE cho kết quả ấn tượng hơn với SAE.

Tính hiệu quả của DSAE so với SAE trên các nhóm tấn công khác nhau có thể được thể hiện thông qua sự chuyển dịch của các vector lớp ẩn

Bảng 2.6 thể hiện kết quả quan sát tấn công khó với SAE, cho thấy: các tấn công cho "RE lớn" mà SAE phân tách lỗi thì đều được DSAE phân tách đúng (26/26 mẫu với R2L); đa số mẫu tấn công khó với SAE mà cho "RE bé" thì được DSAE phân tách đúng.