PHÁT HIỆN LƯU LƯỢNG MẠNG BẤT THƯỜNG TRONG ĐIỀU KIỆN DỮ LIỆU HUẤN LUYỆN CHỨA NGOẠI LAI

Nguyễn Hà Dương*, Hoàng Đăng Hải+
*Khoa Công nghệ thông tin, Trường Đại học Xây Dựng Hà Nội
+Học viện Công nghệ Bưu chính Viễn thông

Tóm tắt: Phát hiện lưu lượng mạng bất thường đối mặt với nhiều khó khăn, thách thức như: xác định mức ngưỡng dùng để so sánh phát hiện bất thường, trích chọn đặc trưng dữ liệu, giảm số dữ liệu cần xử lý, độ chính xác cần thiết... Ngoài ra, ngoại lai có thể gây ra sai lệch đáng kể trong quá trình phát hiện. Bài báo này đề cập các vấn đề phát hiện lưu lượng mạng bất thường trong điều kiện dữ liệu huấn luyện chứa ngoại lai và đề xuất một phương pháp cải tiến dựa trên thuật toán phân tích thành phần chính PCA gọi tên là dPCA. Kết quả thử nghiệm được đánh giá dựa trên tập dữ liệu Kyoto Honeypot.

Từ khóa: Phát hiện lưu lượng mạng bất thường, phát hiện ngoại lai, an ninh mạng.

I. MỞ ĐẦU

Tính mở và sự đa dạng của hạ tầng mạng, dịch vụ và ứng dụng đã tạo ra biến động, thăng giáng đáng kể của lưu lượng mạng. Mặt khác, hoạt động tấn công của tin tặc trên mạng cũng góp phần không nhỏ trong việc tạo ra lưu lượng đột biến so với lưu lượng bình thường trên mạng. Phát hiện lưu lượng mạng bất thường đã là một chủ đề nghiên cứu được quan tâm nhiều trong thời gian qua và đang trở thành một hướng nghiên cứu được đặc biệt quan tâm trong sự phát triển của lĩnh vực an ninh mạng [1]. Lưu lượng mạng bất thường là lưu lượng có sự biến đổi không bình thường, có những thăng giáng đáng kể so với lưu lượng bình thường của mạng. Sự biến đổi bất thường này có thể do nhiều nguyên nhân, ví dụ điển hình là tấn công của tin tặc trên mạng (như DoS, Scan) và lỗi mạng. Ví dụ, tấn công DoS thường tạo ra một lượng lưu lượng đột biến so với lưu lượng bình thường trên mạng.

Phát hiện nhanh và sớm lưu lượng mạng bất thường có thể giúp sớm phát hiện dấu hiệu tấn công mạng. So với các phương pháp truyền thống phát hiện tấn công mạng dựa trên dấu hiệu (signature-based) thường dùng trong các hệ thống phát hiện xâm nhập (Intrusion Detection System - IDS) [2,3,4], các phương pháp dựa trên sự kiện bất thường (anomaly-based detection) có ưu thế vì cho phép phát hiện được những kiểu tấn công mới. Nếu phát hiện chỉ dựa trên các mẫu dấu hiệu tấn công đã biết, hệ thống sẽ không thể phát hiện nếu tin tặc thay đổi một vài chi tiết để biến tấn công trở thành một kiểu mới. Vì vậy, các hệ thống ADS (Anomaly Detection System) đã được phát triển dựa trên phương pháp phát hiện hành vi bất thường (ví dụ [1,2]).

Triển khai các ADS khó khăn hơn nhiều so với các IDS truyền thống. Các IDS thường dựa trên việc so sánh mẫu lưu lượng mạng thu được với các mẫu dấu hiệu biết trước lưu trong cơ sở dữ liệu tập mẫu. Ngược lại, ADS không đòi hỏi mẫu dữ liệu tấn công biết trước. Đối với ADS, cần xác định một tập hợp lưu lượng mạng bình thường. Lưu lượng mạng thu được sẽ được so sánh với tập hợp được coi là bình thường nêu trên. Dữ liệu

Tác giả liên hệ: Nguyễn Hà Dương,
email: nghaduong@gmail.com
Đến tòa soạn: 12/2/2016, chỉnh sửa: 12/4/2016, chấp nhận đăng: 12/5/2016.
Một phần kết quả của bài báo này đã được trình bày tại hội thảo quốc gia ECIT'2015.

không nằm trong tập bình thường sẽ bị coi là bất thường. Các phương pháp phát hiện lưu lượng mạng bất thường cho ADS phải đối mặt với một số vấn đề chủ yếu như sau:

1) Cần xác định tập mẫu dữ liệu không chứa bất thường để từ đó phát hiện ra những sự kiện bất thường trong các tập dữ liệu thu được từ mạng.

2) Để tăng độ chính xác, tập mẫu dữ liệu thường rất lớn với số lượng biến (thuộc tính dữ liệu) lớn dẫn đến tốn tài nguyên hệ thống, thời gian xử lý dài, tốc độ phát hiện chậm. Vấn đề là cần trích chọn đặc trưng dữ liệu sao cho giảm yêu cầu về lượng dữ liệu phải xử lý trong khi vẫn bảo đảm độ chính xác cần thiết, tốc độ xử lý và phát hiện nhanh.

3) Thực tế các tập mẫu dữ liệu bình thường vẫn có thể chứa một phần dữ liệu bất thường (gọi chung là ngoại lai) có thể làm sai lệch quá trình huấn luyện và kết quả phát hiện. Do vậy cần phương pháp loại bỏ ngoại lai khỏi tập dữ liệu huấn luyện.

Đã có nhiều công trình nghiên cứu về phát hiện lưu lượng mạng bất thường đã được đề xuất tới nay, song các phương pháp phát hiện theo mô hình thống kê, khai phá dữ liệu, học máy vẫn được coi là hiệu quả và khả thi hơn (xem [5-12]). Một số nghiên cứu áp dụng thuật toán PCA (Principle Component Analysis) [13-17] đã cho thấy khả năng giảm lượng dữ liệu cần xử lý, độ chính xác tương đối cao, khả năng phát hiện nhanh. Mặc dù vậy, vấn đề phát hiện trong điều kiện dữ liệu huấn luyện có chứa ngoại lai vẫn chưa được quan tâm đúng mức. Ngoại lai là những phần tử bất thường lẫn vào tập dữ liệu dùng để huấn luyện. Những phần tử này gây ra sự sai lệch trong các tham số khi huấn luyện và ảnh hưởng đến hiệu suất hoạt động của hệ thống.

Bài báo này đề xuất một phương pháp phát hiện lưu lượng mạng bất thường trong điều kiện dữ liệu huấn luyện chứa ngoại lai. Phương pháp được xây dựng dựa trên nền tảng thuật toán PCA với một số cải tiến: giảm thiểu thành phần chính thứ yếu để tính đường cơ sở, khử ngoại lai với chế độ không giám sát và phân cụm, phân cấp phát hiện. Bài báo được bố cục thành ba phần như sau. Phần II trình bày một số nghiên cứu liên quan. Phần III trình bày phương pháp đề xuất của bài báo. Phần IV là kết quả thử nghiệm. Phần V là kết luận.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Các công trình nghiên cứu về phát hiện lưu lượng mạng bất thường tới nay chủ yếu dựa trên một nguyên lý căn bản nhất, đó là chỉ ra các đặc tính lưu lượng mạng trong điều kiện hoạt động bình thường theo một cách nào đó và xác định sự khác biệt của lưu lượng mạng trong so sánh với lưu lượng mạng bình thường đã nêu. Ranh giới của sự khác biệt đó là mức ngưỡng (Threshold) thường có biến động theo thời gian. ADS thường được xây dựng theo mô hình thống kê, khai phá dữ liệu, học máy,... [1-12]. Mục tiêu đặt ra là tạo ra một đường cơ sở (Baseline) bao gồm các mức ngưỡng biến động theo thời gian. Tuy nhiên, do số lượng dữ liệu và số thuộc tính dữ liệu lớn nên việc tạo ra đường cơ sở và so sánh với đường cơ sở vẫn là vấn đề phức tạp, phải xử lý nhiều và khó khả thi.

Một số nghiên cứu tập trung vào lựa chọn đặc trưng dữ liệu nhằm giảm yêu cầu về lượng dữ liệu phải xử lý. Thuật toán phân tích thành phần chính (PCA-Principle Component Analysis) [13-16] đã được đề xuất áp dụng nhằm chuyển đổi tập dữ liệu (p chiều) sang một miền dữ liệu mới (m chiều, với $m < p$) nhằm giảm số chiều dữ liệu.

Trong phần này, bài báo tóm tắt một số công trình điển hình nhất [12-16] sử dụng phương pháp PCA do có liên quan đến nội dung bài. PCA là thuật toán thường sử dụng để giảm số chiều dữ liệu nhưng vẫn giữ được phần lớn đặc tính của dữ liệu. Mỗi trị riêng của thành phần chính tương ứng một phần với sự biến thiên của các thuộc tính hay biến trong dữ liệu. Trị riêng càng lớn thì càng chứa nhiều biến thiên và vector riêng tương ứng phản ánh quy luật biến thiên càng lớn nên càng quan trọng. Do vậy, những thành phần chính quan trọng nhất cần được xếp trước các thành phần không quan trọng [1-3][7][13-16].

Trong [5,12,13,14], các tác giả theo dõi sự thay đổi các giá trị thành phần chính và phát hiện sự

thay đổi bất thường trên các thành phần chính nhất định. Các thành phần chính (Principal Component – PC) có thể phân chia thành những thành phần chủ yếu phản ánh quy luật biến thiên của lưu lượng $y_{(m)}$ trong trạng thái bình thường của hệ thống và những thành phần dư thừa phản ánh sự biến thiên không theo quy luật $y_{(p-m)}$. Trong [5,12], độ lớn của phần dư tái tạo tương ứng với $y_{(p-m)}$ được phân tích từ đó phát hiện ra những dấu hiệu bất thường dựa trên mức ngưỡng. Một cách tương tự là tính khoảng cách Euclidean giữa dữ liệu chuẩn hóa z và dữ liệu tái tạo từ những thành phần chính $y_{(m)}$ [13]. Tuy nhiên sự tái tạo lại z từ những thành phần chính $y_{(m)}$ làm tăng mức độ xử lý của hệ thống. Trong [14], khoảng cách Mahalanobis dựa trên thành phần chính chủ yếu và thứ yếu được sử dụng để phân tích dấu hiệu bất thường. Hiệu quả của phương pháp phụ thuộc vào số lượng và tỷ lệ các PC chủ yếu và thứ yếu. Tác giả trong [15] sử dụng phương pháp Histogram. Phương pháp này đơn giản hơn song đòi hỏi lượng dữ liệu phải lớn để đạt được tỷ lệ phát hiện đúng cao. Công trình [16] đề xuất giảm bớt tập thuộc tính dữ liệu nhằm giảm độ phức tạp của thuật toán phát hiện.

Qua nghiên cứu các công trình liên quan, ta rút ra một số nhận xét như sau:

1) Các nghiên cứu áp dụng PCA đều sử dụng cách so sánh biến thiên của lưu lượng với một đường cơ sở, song giảm được dữ liệu cần xử lý qua việc biến đổi sang miền dữ liệu chỉ sử dụng các thành phần chính. Tuy nhiên, sử dụng các thành phần chính nào vẫn là vấn đề chưa được nghiên cứu cụ thể. Các thành phần chính được chia thành các thành phần chính chủ yếu (những thành phần chính đầu tiên, có trị riêng lớn nhất) và thành phần chính thứ yếu (những thành phần chính cuối có trị riêng nhỏ nhất).

2) Các thành phần chính chủ yếu có xu hướng phản ánh sự biến thiên bình thường của lưu lượng. Trong điều kiện dữ liệu huấn luyện sạch, bất thường có xu hướng xuất hiện ở các thành phần chính thứ yếu. Điều này phù hợp với phương pháp phân tích phần dư. Do vậy, lựa chọn các thành phần chính phù hợp có thể mang lại hiệu quả phát hiện.

3) PCA rất nhạy cảm với dữ liệu ngoại lai. Do đó, cần giảm thiểu tác động của ngoại lai, hoặc cần lọc bớt dữ liệu đầu vào ngoại lai cho tập huấn luyện. PCA cũng có thể phát sinh ngoại lai không mong muốn. Vì vậy, cần đánh giá tác động của các thành phần chính đến việc phát sinh ngoại lai, từ đó lựa chọn thành phần chính hoặc các đặc tính lưu lượng mạng cần thiết để giảm tác động của yếu tố này.

III. PHƯƠNG PHÁP DPCA

A. Cơ sở thuật toán PCA

PCA là phương pháp chuyển đổi tập dữ liệu (p chiều) sang một miền dữ liệu mới (m chiều, với $m < p$) nhằm giảm số chiều dữ liệu [13-16]. Thuật toán PCA cơ sở như sau.

Gọi X là một tập dữ liệu gồm n quan sát với p biến $X_1, X_2, \ldots X_p$ được tổ chức thành ma trận nxp (n hàng, p cột). Mỗi biến biểu thị một thuộc tính của dữ liệu ban đầu. Mỗi quan sát $x=(x_1, x_2, \ldots, x_p)^T$ chứa p thuộc tính khác nhau. Gọi R là ma trận tương quan pxp tính được từ X, (λ_k, e_k) là các cặp trị riêng và vector riêng của R được sắp xếp theo thứ tự giảm dần của trị riêng ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$).

Phép biến đổi của thuật toán PCA cơ sở là sự chuyển các điểm dữ liệu ban đầu sang kết quả thành phần chính. Khi đó thành phần chính thứ i của một quan sát x sẽ là

$$y_i = e_i^T z = e_{i1}z_1 + e_{i2}z_2 + \ldots + e_{ip}z_p \qquad (1)$$

trong đó: y_i là thành phần chính thứ i của quan sát x ban đầu, $i=1 \ldots p$, $e_i=(e_{i1}, e_{i2}, \ldots, e_{ip})^T$ là vectơ riêng thứ i, $z = (z_1, z_2, \ldots, z_p)^T$ là vectơ đã chuẩn hóa của x, z_k của biến thứ k được tính theo công thức

$$z_k = \frac{x_k - \overline{x_k}}{\sqrt{s_k}} \qquad (2)$$

với $\overline{x_k}$ là giá trị trung bình, s_k là phương sai của biến thứ k, $k = 1 \ldots p$.

Bài toán phát hiện bất thường với PCA được đưa về bài toán tính toán khoảng cách ***d*** giữa dữ liệu chuẩn hóa *z* và dữ liệu tái tạo từ các thành phần chính y_i của các quan sát. Khoảng cách được so sánh với mức ngưỡng để xác định tập dữ liệu là bình thường hay bất thường. Nhiều phương pháp tính khoảng cách có thể được áp dụng, điển hình như: Euclidean, Manhattan [13], Mahalanobis [14],… Việc xử lý một lượng dữ liệu lớn nhiều biến sẽ làm tăng thời gian xử lý dữ liệu và tốn tài nguyên của hệ thống. Vì vậy, áp dụng thuật toán PCA có thể giảm thiểu số chiều không cần thiết và tăng hiệu quả tận dụng tài nguyên hệ thống.

B. Phương pháp dPCA

Trong phần này, bài báo đề xuất phương pháp dPCA (Distance-based anomaly detection method in PCA subspace) trên nền tảng thuật toán PCA cơ sở với một số cải tiến: giảm thiểu thành phần chính thứ yếu để tính đường cơ sở, khử ngoại lai với chế độ không giám sát và phân cụm, phân cấp phát hiện. Về cơ bản, phương pháp dPCA cũng sử dụng thuật toán PCA cơ sở để giảm số chiều dữ liệu (*p* chiều), song giữ phần lớn đặc tính dữ liệu ban đầu bằng cách giữ lại *m* thành phần chính.

Tương tự [18], ta chia *m* thành phần chính thành *r* thành phần chính chủ yếu và *m=p-q+1* thành phần chính thứ yếu. Từ kết quả nghiên cứu đã nêu ở phần 2, không nhất thiết phải tính khoảng cách cho toàn bộ các thành phần chính. Những dữ liệu bất thường có xu hướng xuất hiện ở những thành phần chính cuối cùng (thành phần chính thứ yếu). Theo cách này, ta chỉ cần quan sát dữ liệu ở các thành phần chính thứ yếu (miền con của PCA), qua đó giảm thiểu được lượng dữ liệu cần xử lý. Các kết quả ở phần thử nghiệm chứng minh phương pháp này vẫn bảo đảm độ chính xác cần thiết trong khi giảm thiểu được độ phức tạp, tăng được tốc độ xử lý.

Công thức tính khoảng cách để phát hiện dấu hiệu bất thường trong miền con PCA trong phương pháp dPCA được đề xuất như sau:

$$d = \sum_{i=r}^{q} w_i |y_i|^c \qquad (3)$$

trong đó: $1 \le r < q \le p$, w_i là trọng số cho thành phần chính y_i, d là độ lệch hình thành từ các thành phần chính y_i và trọng số tương ứng w_i, c là số mũ của y_i. c là hằng số, có thể là số thực hoặc số nguyên. w_i, c được lựa chọn dựa trên thực nghiệm.

Một giá trị ngưỡng d_N được xác định dựa vào hàm phân bố tích lũy thực nghiệm của độ lệch *d* (empirical cumulative distribute function - ecdf) và được tính trên dữ liệu huấn luyện. Khi có một quan sát mới, giá trị *d* sẽ được tính dựa trên những tham số huấn luyện như sau:

Chuẩn hóa dữ liệu dựa trên giá trị trung bình và căn bậc hai của phương sai cho mỗi thuộc tính (biến đầu vào).

- Sử dụng vectơ riêng để chuyển mỗi quan sát mới sang các trục của miền con PCA.
- Tính giá trị *d* dựa trên công thức (3) và so sánh với ngưỡng đã thiết lập d_N khi huấn luyện. Nếu $d > d_N$, quan sát mới được coi là bất thường. Ngược lại quan sát đó được coi là bình thường.

Phương pháp dPCA có thể hoạt động trong hai chế độ: bán giám sát và không giám sát.

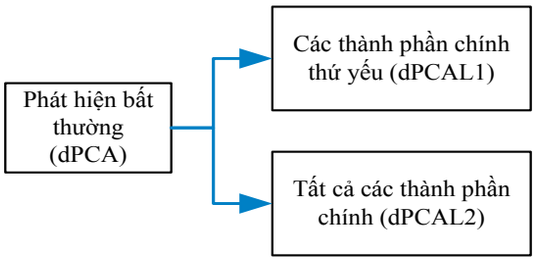
dPCA trong chế độ không giám sát không yêu cầu huấn luyện mà phát hiện trực tiếp với dữ liệu có được. Chế độ này có yêu cầu bổ sung là lượng dữ liệu bình thường phải lớn hơn nhiều so với lượng dữ liệu bất thường. Nếu điều này bị vi phạm sẽ không thể tạo được trạng thái bình thường của mạng để so sánh và phát hiện ra những sự khác biệt của các quan sát có dấu hiệu bất thường.

dPCA chế độ bán giám sát hoạt động theo hai pha:

- Pha huấn luyện (Training phase): Pha này hoạt động ngoại tuyến (offline). Hệ thống được huấn luyện trước với dữ liệu "sạch" (không chứa bất thường hay ngoại lai). Dữ liệu sau khi chuyển đổi PCA tạo thành hồ sơ trạng thái bình thường (normal profile) của hệ thống. Hồ sơ trạng thái chứa các tham số huấn luyện như vectơ riêng, trị riêng và giá trị ngưỡng. Tập hợp các giá trị của *d* được

tính trên tất cả các quan sát với dữ liệu huấn luyện sạch sẽ tạo nên đường cơ sở để phát hiện ngoại lai. Vì vậy có thể coi *d* là độ lệch của mỗi quan sát để xét quan sát đó là bình thường hay bất thường.

- Pha phát hiện (Detection phase): Pha này hoạt động trực tuyến (online). Mỗi quan sát mới là một vectơ chứa các thuộc tính dữ liệu cần chuyển sang miền con của PCA với các vectơ riêng và trị riêng đã có trong pha huấn luyện. Bộ phát hiện áp dụng phương pháp dPCA để tính độ lệch và so sánh với giá trị ngưỡng. Nếu độ lệch vượt quá giá trị ngưỡng, quan sát mới được coi là bất thường và ngược lại là bình thường.



Hình 1. Phương pháp dPCA trong chế độ bán giám sát

Hình 1 mô tả phương pháp dPCA trong chế độ bán giám sát. Khi dPCA chỉ được thực hiện với các thành phần chính thứ yếu (gọi là dPCAL1) được coi là sự kiểm tra nhanh xem có hiện tượng bất thường trên mạng hay không. Tuy nhiên, mặc dù các bất thường có xu hướng xuất hiện tại thành phần chính thứ yếu, chúng vẫn có thể xuất hiện tại các thành phần chính khác. Vì vậy dPCA cũng sẽ kiểm tra với tất cả với các thành phần chính nhưng với tần suất thấp hơn để phát hiện những bất thường tại đây (dPCAL2). Cứ sau *x* lần kiểm tra với dPCAL1 thì lại có một lần kiểm tra với dPCAL2. Số lần *x* tùy thuộc nhu cầu giám sát mạng. Thực nghiệm cho thấy dPCAL2 chỉ cần thực hiện với các PC (chiếm 70%-80% mức biến thiên của tổng các trị riêng) mà vẫn bảo đảm độ chính xác cần thiết. Do vậy có thể bỏ qua một số PC khác trong công thức tính khoảng cách của *d* để tăng tốc độ xử lý và phát hiện với dPCAL2. Nếu thấy số lượng bất thường phát hiện với dPCAL2 tăng đột biến có thể tăng tần suất của dPCAL2. Mỗi khi xuất hiện bất thường ở dPCAL1 có thể kiểm tra lại bằng dPCAL2. Lý do là vì sử dụng tất cả các PC thường có độ ổn định phát hiện cao.

C. Phương pháp dPCA với các thành phần chính thứ yếu trong chế độ bán giám sát (dPCAL1)

Trong nghiên cứu của Lakhina [5] và Wang [13], thực chất phần dư chính là khoảng cách giữa *z* và z_N. z_N được tái tạo từ các thành phần chính chủ yếu. Khi chuyển sang không gian con PCA, công thức này trở thành:

$$d = \left\| z_{(A)} \right\|^2 = (z - z_{(N)})^T (z - z_{(N)}) = \sum_{i=r}^{q} y_i^2 \qquad (4)$$

Trong công thức trên, *d* bằng tổng của các bình phương thành phần chính thứ yếu ($1 < r < q \leq p$). Như vậy bằng cách thiết lập $w_i = 1$ và hằng số $c = 2$ trong công thức (3), phương pháp dPCA sẽ đạt được kết quả tương tự như các nghiên cứu của Lakhina [5] và Wang [13]. Khi tính *d* với các thành phần chính thứ yếu, phương pháp dPCA sẽ tương đương với phương pháp tính phần dư đã nêu trong [5,13] song thay vì phải chuyển đổi dữ liệu trở lại *z* trong không gian ban đầu như trong [13], dPCA cho phép thực hiện trực tiếp với *y* trong miền con PCA, do đó giảm bớt được độ phức tạp.

Nếu thiết lập *d* với trọng số $w_i = 1/\lambda_i$ và hằng số $c = 2$ trong công thức (4), ta được kết quả tương tự với nghiên cứu của Shyu [14]. Thực chất của phương pháp này là chuẩn hóa các bình phương của giá trị y theo trị riêng của mỗi thành phần chính. Nếu sự chênh lệch giá trị của các thành phần chính thứ yếu là đáng kể, chuẩn hóa theo trị riêng sẽ làm giảm sự khác biệt trong công thức tính khoảng cách. Thực chất, phương pháp của Shyu [14] sử dụng song song hai khoảng cách với cả thành phần chính chủ yếu và thứ yếu. Điều này có ưu điểm là làm tăng khả năng phát hiện bất thường nhưng có nhược điểm cơ bản là làm tăng tỷ lệ cảnh báo sai do dữ liệu bình thường bị phát hiện là bất thường. Ngoài ra việc sử dụng hai mức ngưỡng song song cũng làm tăng độ phức tạp của thuật toán. Phương pháp dPCA chỉ cần tính khoảng cách với thành phần chính thứ yếu.

Với trọng số $w_i = 1/\lambda_i$ và hằng số c =2, công thức (3) cho kết quả:

$$d = \sum_{i=r}^{q} \frac{y_i^2}{\lambda_i} \quad (5)$$

Để chuẩn hóa cho các giá trị thành phần chính thứ yếu, có thể thiết lập $w_i = 1/\lambda_i^{1/2}$ và hằng số c =1. Kết quả đạt được sẽ tương tự như (5) song công thức sẽ đơn giản hơn vì không cần tính bình phương của giá trị các thành phần chính, căn bậc hai của trị riêng chỉ phải tính một lần trong pha huấn luyện.

$$d = \sum_{i=r}^{q} \frac{|y_i|}{\sqrt{\lambda_i}} \quad (6)$$

Nếu thiết lập trọng số $w_i = 1/\lambda_i$ và hằng số c =1, công thức tính d sẽ đạt kết quả gần tương đương với (5), (6) song không cần tính căn bậc hai của trị riêng trong pha huấn luyện.

$$d = \sum_{i=r}^{q} \frac{|y_i|}{\lambda_i} \quad (7)$$

Trong thực tế nếu trị riêng của các thành phần chính thứ yếu không có sự khác biệt đáng kể thì có thể thiết lập $w_i = 1$ và hằng số $c = 1$ cho các thành phần chính thứ yếu. Kết quả tính d sẽ tương đương với (4), (5), (6), (7) song công thức sẽ đơn giản hơn nhiều.

$$d = \sum_{i=r}^{q} |y_i| \quad (8)$$

Độ phức tạp của thuật toán tính khoảng cách d với công thức (4) và (5) là $O(n^2)$ tương đương với độ phức tạp trong [5,13,14]. Các công thức (6), (7), (8) có độ phức tạp O(n), giảm được độ phức tạp so với (4), (5). Lưu ý d là độ phức tạp của thuật toán tính khoảng cách trong dPCA chưa tính đến độ phức tạp của chính thuật toán PCA. Trong pha huấn luyện, thuật toán PCA có độ phức tạp $O(np^2)$ khi tính ma trận tương quan và $O(p^3)$ khi tính các cặp trị riêng/vectơ riêng. Độ phức tạp của thuật toán PCA không thay đổi được (trong các công trình nghiên cứu trước cũng phải chấp nhận điều này) nên giảm số chiều dữ liệu p là rất cần thiết. Trong pha phát hiện, dPCAL1 chỉ sử dụng các thành phần chính thứ yếu nên giảm được yêu cầu tính toán.

D. Phương pháp dPCA với tất cả các thành phần chính trong chế độ bán giám sát (dPCAL2)

Khi sử dụng đầy đủ các thành phần chính, chỉ số r của công thức (3) bằng 1. Thường trong trường hợp dữ liệu huấn luyện chứa ngoại lai, dùng tất cả các thành phần chính trong công thức tính khoảng cách sẽ tốt hơn. Nếu tính d với toàn bộ giá trị của p thành phần chính trong công thức (5), kết quả cho lại sẽ tương đương với khoảng cách Mahalanobis hoặc thống kê T^2, χ^2 (Chi-square). Tuy nhiên, vấn đề là độ phức tạp cao hơn do phải tính toán nhiều hơn. Bài báo đề xuất một cách giảm độ phức tạp tính toán là sử dụng công thức (6) hoặc (7). So với (5), công thức (6), (7) vẫn có được hiệu quả tương đương nhưng lại đơn giản hơn. Do không phải tính bình phương cho y_i mỗi khi tính d nên phép tính đơn giản hơn (w_i chỉ phải tính một lần trong pha huấn luyện, trong pha phát hiện w_i là hằng số). Khi thiết lập cặp giá trị này, cần lưu ý là không áp dụng phương pháp thống kê tham số theo phân bố biết trước (T^2, χ^2) để tìm mức ngưỡng bằng cách tra bảng của phân bố tương ứng.

E. Khử ngoại lai trong dữ liệu huấn luyện với dPCA trong chế độ không giám sát

dPCA chế độ bán giam sát đòi hỏi dữ liệu sạch hay nói cách khác là cần tập dữ liệu huấn luyện không chứa ngoại lai do ngoại lai dẫn đến sai lệch kết quả phát hiện. Vì vậy, dPCA chế độ không giám sát mặc dù vẫn có thể sử dụng để phát hiện bất thường nhưng còn có mục đích sử dụng để lọc bỏ bớt ngoại lai trong dữ liệu huấn luyện cho chế độ bán giám sát.

Để loại được ngoại lai, cần thiết lập ngưỡng ở mức thấp hơn so với ngưỡng thường đặt trong chế độ bán giám sát vì nếu đặt mức ngưỡng cao sẽ bỏ qua nhiều ngoại lai. Khi đó, không chỉ ngoại lai mà cả những dữ liệu bình thường vượt quá ngưỡng cũng có thể bị loại bỏ khỏi tập huấn luyện. Điều này nghĩa là tỷ lệ FPR (False Positive Ratio) tức là số dữ liệu bình thường bị phát hiện sai có thể

tăng lên. Do vậy, đặt mức ngưỡng phù hợp là cần thiết. Việc loại bỏ cả những dữ liệu bình thường có khoảng cách lớn hơn những dữ liệu bình thường khác rõ ràng là cũng cần thiết vì chúng ảnh hưởng đến ma trận hiệp phương sai, giá trị trung bình, tập dữ liệu trong miền con PCA, bao gồm cả vectơ riêng, trị riêng và giá trị các thành phần chính. Việc khử ngoại lai trong dữ liệu huấn luyện với dPCA được thực hiện như sau:

- Dữ liệu đầu vào được ánh xạ sang miền con PCA.
- Tính khoảng cách d dựa trên một trong các công thức (5), (6), (7) với tất cả các thành phần chính.
- Xác định mức ngưỡng dựa trên hàm phân bố tích lũy thực nghiệm hoặc phân bố biết trước (phân bố F, χ^2).
- Loại bỏ tất cả những quan sát có khoảng cách lớn hơn mức ngưỡng.
- Những dữ liệu còn lại là tập dữ liệu dùng cho huấn luyện.

F. Khử ngoại lai trong dữ liệu huấn luyện bằng phương pháp K-Means

Một phương pháp khử ngoại lai khác được đề xuất trong bài báo này là sử dụng kỹ thuật phân cụm dựa trên thuật toán K-means. Trước khi thực hiện phân cụm với K-means, dữ liệu đầu vào được chuẩn hóa theo công thức (2). Quá trình phân cụm của thuật toán K-means bao gồm các bước chính sau:

Bước 1: Chọn ngẫu nhiên K tâm (centroid) cho K cụm (cluster). Mỗi cụm được đại diện bằng các tâm của cụm C_1, C_2, ... C_K.

Bước 2: Tính khoảng cách giữa các điểm đến K tâm (thường dùng khoảng cách Euclidean).

Bước 3: Nhóm các đối tượng vào cụm gần nhất.

Bước 4: Xác định lại tâm mới cho các cụm.

Bước 5: Thực hiện lại các bước trên cho đến khi sai số bình phương không thay đổi.

Việc xác định giá trị K ảnh hưởng nhiều đến kết quả phân cụm và phát hiện bất thường. Phát hiện ngoại lai dựa trên phân loại dữ liệu vào các cụm. Phân loại cho biết điểm dữ liệu nào thuộc cụm nào. Để phát hiện ngoại lai bằng phân loại, cần thiết lập cụm dữ liệu bình thường và ngoại lai. Trong trường hợp $K=2$, chỉ có một cụm là bình thường và cụm còn lại là ngoại lai.

Để phát hiện được các điểm ngoại lai, cần thiết lập khoảng cách tối đa (d_{max}). Khi khoảng cách từ mỗi điểm đến tâm cụm bình thường vượt quá d_{max}, điểm đang xét sẽ được coi là ngoại lai. Những điểm này sẽ bị loại bỏ khỏi tập dữ liệu huấn luyện.

IV. THỬ NGHIỆM

Mục tiêu của thử nghiệm là đánh giá khả năng phát hiện của phương pháp đề xuất, khả năng loại bỏ ngoại lai và ảnh hưởng của ngoại lai đến hiệu suất của hệ thống trước và sau khi khử ngoại lai. Quá trình thử nghiệm được thực hiện dựa trên phần mềm *Matlab R2013a.*

A. Dữ liệu dùng cho thử nghiệm

Cách thức chung để thử nghiệm các hệ thống phát hiện lưu lượng mạng bất thường (hay rộng hơn là phát hiện tấn công mạng) là: 1) Thu thập dữ liệu mạng trong điều kiện hoạt động bình thường (dữ liệu sạch, chưa có tấn công hay bất thường); 2) Thiết lập các tập dữ liệu mẫu cho lưu lượng bình thường, tạo đường cơ sở; 3) Thu thập dữ liệu mạng thực tế và so sánh với tập mẫu bình thường (đường cơ sở) để phát hiện.

Các chuẩn thu thập lưu lượng phổ biến là tcpdump, flowdump, netflow, IPFIX. Dữ liệu thu được thường bao gồm những thông tin cơ bản như địa chỉ IP nguồn và đích, cổng nguồn và đích, giao thức… Để thiết lập các tập dữ liệu mẫu, những thuộc tính quan trọng của luồng tin được tách ra và được tổng hợp, chuẩn hóa thành các thuộc tính (attribute) hay đặc trưng (feature) [19-22]. Các thuộc tính thường được thống kê từ các giá trị thu được với các tham số khác nhau như loại giao thức, kết nối, thời gian,…[1,2,5,6,12].

Do việc thu thập, tổng hợp dữ liệu qua các công cụ như tcpdump, flowdump,… và chuyển đổi thành các thuộc tính đòi hỏi nhiều thời gian, công sức nên hầu hết các nghiên cứu tới nay đều sử dụng các tập dữ liệu có sẵn đã thu thập trên mạng thực tế để thử nghiệm. Điển hình là các tập dữ liệu KDD, NSL-KDD, Kyoto Honeypot [19-22]. Đây thực chất là dữ liệu thực thu được từ mạng đang hoạt động. Để giúp các nhà nghiên cứu đánh giá, so sánh các phương pháp đã đề xuất, các tập dữ liệu này thường đã được đánh nhãn để phân biệt là bình thường (phục vụ cho thiết lập tập mẫu bình thường) và bất thường hay có tấn công (phục vụ cho việc kiểm nghiệm). Đây là dữ liệu đo được từ thực tế, nên việc sử dụng các tập dữ liệu để kiểm nghiệm không ảnh hưởng đến chất lượng của phương pháp phát hiện. Tương tự các nghiên cứu [1,4,5-10,12-16], bài báo sử dụng các tập dữ liệu nêu trên đã thu được từ mạng thực tế để kiểm nghiệm.

B. Tập dữ liệu Kyoto Honeypot

Đây là tập dữ liệu thực tế thu được tại hệ thống "bẫy" tổ ong (Honeypot) của đại học Kyoto (Nhật Bản) từ năm 2006 đến năm 2009 [22]. Honeypot được sử dụng với mục đích đánh lừa tin tặc tấn công vào hệ thống này để thu thập dữ liệu cho việc phân tích dấu vết.

Bảng I. Thuộc tính dùng trong thử nghiệm của tập dữ liệu Kyoto Honeypot

No	Thuộc tính	Ý nghĩa
1	duration	Thời gian của kết nối
2	service	Dịch vụ (ví dụ HTTP)
3	src_bytes	Số lượng byte gửi từ nguồn đến đích
4	dst_bytes	Số lượng byte gửi từ đích về nguồn
5	count	Số lượng kết nối đến cùng địa chỉ đích đang xét trong 2s
6	same_srv_rate	Số lượng kết nối trong count có cùng kiểu dịch vụ
7	serror_rate	Số kết nối có lỗi đồng bộ SYN %
8	srv_serror_rate	Số kết nối có lỗi đồng bộ % SYN và cùng kiểu dịch vụ trong thời gian 2 s
9	dst_host_count	Số lượng địa chỉ đích
10	dst_host_srv_ count	Số lượng kết nối đến cùng địa chỉ đích đang xét và cùng dịch vụ đích
11	dst_host_same_ src_port_rate	số kết nối có cùng cổng % nguồn với kết nối đang xét trong trường dst_host_count
12	dst_host_serror_ rate	số kết nối có lỗi SYN trong % dst_host_count
13	dst_host_srv_ serror_rate	số kết nối có lỗi SYN trong % dst_host_srv_count
14	destination Port Number	Số hiệu cổng đích của kết nối

Lưu lượng thu được từ hệ thống Honeypot có điểm đặc biệt là phần lớn các tấn công có nguồn gốc từ Internet. Các thuộc tính của tập dữ liệu này tương tự như của KDDCUP 99 nhưng lược bỏ bớt những thuộc tính được coi là không cần thiết. Số thuộc tính tương đương với KDDCUP 99 là 14. Ngoài ra tập này còn bổ sung thêm 10 thuộc tính khác. Kyoto Honeypot có ưu điểm là phản ánh chính xác hơn quy luật biến thiên của mạng trong điều kiện bình thường cũng như tính khách quan của các sự kiện bất thường trong lưu lượng mạng Internet.

C. Các thông số đánh giá

- True Positive (TP): Sự kiện một mẫu bất thường được phát hiện chính xác.
- False Positive (FP): Sự kiện phát hiện một mẫu là *bất thường* song thực tế là *bình thường*.
- True Negative (TN): Sự kiện một *mẫu bình thường* được phát hiện chính xác.
- False Negative (FN): Sự kiện phát hiện một mẫu là *bình thường* song thực tế *bất thường*.
- Precision (PR): Tỷ lệ số mẫu phát hiện bất thường chính xác và tổng số mẫu phát hiện là bất thường trong tập dữ liệu kiểm tra:

$$\mathrm{PR} = \frac{TP}{TP + FP} \qquad (9)$$

- True Positive Rate (TPR) còn gọi là Recall: Tỷ lệ giữa số *mẫu bất thường* phát hiện chính

xác và số mẫu bất thường thực tế trong tập dữ liệu kiểm tra:

$$TPR = \frac{TP}{TP + FN} \quad (10)$$

- False Positive (FPR): Tỷ lệ giữa số *mẫu bất thường* phát hiện sai và số *mẫu bình thường* trong tập dữ liệu kiểm tra.

$$FPR = \frac{FP}{TN + FP} \quad (11)$$

Total Accuracy (TA): Độ chính xác tổng bằng số mẫu phát hiện chính xác của cả bất thường và bình thường trên số mẫu của tập dữ liệu:

$$TA = \frac{TP + TN}{TP + FP + TN + FN}. \quad (12)$$

Trong các thông số trên, TPR và FPR là hai thông số quan trọng nhất. TA cho biết tỷ lệ phát hiện đúng tổng cộng. PR cũng là thông số hay được sử dụng tuy nhiên giá trị PR trong một số trường hợp không biểu thị hoàn toàn đúng độ chính xác. Ví dụ: nếu số lượng mẫu bình thường gấp 100 lần số lượng mẫu bất thường, chỉ cần tỷ lệ 1% FPR cũng làm cho PR rất thấp.

D. Kết quả thử nghiệm

Bảng II thống kê kết quả phát hiện khi thử nghiệm với dữ liệu huấn luyện sạch Kyoto Honeypot với các trọng số và số thành phần chính khác nhau. Dữ liệu pha huấn luyện sử dụng 5.000 kết nối đánh nhãn bình thường. Dữ liệu pha phát hiện có số lượng trong khoảng 100.000 - 120.000 kết nối. Giá trị k trong các bảng II - VI là số thành phần chính. Các ngày khảo sát được lựa chọn ngẫu nhiên để có kết quả khách quan.

Theo công thức (3), khi $w_i = 1$, c = 2: Các kết quả tương tự nhau với số các thành phần chính k = 2, k = 3, k = 4, k = 5. Kết quả này dựa trên nghiên cứu trong [5, 13] do công thức (4) tương đương công thức (3) khi $w_i = 1$, c = 2.

Nếu sử dụng tất cả các thành phần chính (k = 14), do sự chênh lệch về giá trị giữa các thành phần chính chủ yếu và thứ yếu, giá trị *d* sẽ chứa các giá trị thành phần chính chủ yếu nhiều hơn. Điều này làm mất đi những ngoại lai có xu hướng xuất hiện tại thành phần chính thứ yếu. Vì vậy khi k = 14, có sự suy giảm rõ rệt tỷ lệ TPR trong kết quả phát hiện so với lựa chọn sử dụng các thành phần chính thứ yếu. Do đó cần thiết có sự chuẩn hóa theo mức biến thiên của trị riêng cho mỗi thành phần chính khi cần sử dụng khoảng cách với các thành phần chính chủ yếu như các công thức (5),(6),(7).

Bảng II. Thử nghiệm với dữ liệu huấn luyện sạch

w_i	c	k	PR (%)	TPR (%)	FPR (%)	TA (%)
1	2	2	98.4	91.8	3	93.5
1	2	3	98.4	89.2	2.9	91.7
1	2	14	98.5	57	1.8	70.3
$1/\lambda_i$	2	2	98.7	90.8	2.5	93
$1/\lambda_i$	2	3	98.5	91.1	2.9	93
$1/\lambda_i$	2	4	98.5	91.9	2.8	93.6
$1/\lambda_i$	2	5	98.5	92	2.8	93.7
$1/\lambda_i$	2	14	98.8	87.9	2.1	91.2
$1/\sqrt{\lambda_i}$	1	3	98.5	91.3	2.89	93.2
$1/\sqrt{\lambda_i}$	1	14	98.9	88.6	2.1	91.6
$1/\lambda_i$	1	3	98.6	91.4	2.7	93.3
$1/\lambda_i$	1	14	98.8	91.5	2.3	93.5
1	1	3	98.4	90.1	2.9	92.4
1	1	14	98.5	62.9	1.9	74.4

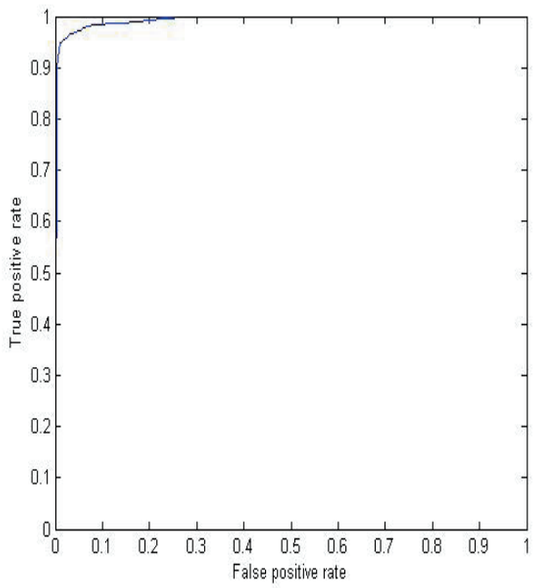
Theo công thức (5): Công thức (5) chuẩn hóa tất cả các thành phần chính với $w_i = 1/\lambda_i$ sẽ làm cho các giá trị phân phối đồng đều hơn trong *d*. Khi k=14, giá trị *d* tương đương với khoảng cách Mahalanobis hoặc thống kê T^2. Có thể thấy khi sử dụng những thành phần chính thứ yếu (k=2,3,4,5), kết quả phát hiện đạt được tương tự khi k=14. Điều này cho thấy bất thường có xu hướng xuất hiện tại những thành phần chính thứ yếu. Do vậy không cần thiết phải sử dụng tất cả các thành phần chính trong công thức tính *d*, do đó giảm được số chiều dữ liệu cần tính toán.

Tỷ lệ của TPR khi chỉ sử dụng các thành phần chính thứ yếu còn cho thấy không cần thiết phải thực hiện tính cả hai khoảng cách và so sánh hai mức ngưỡng song song như trong nghiên cứu của Shyu [14].

Theo công thức (6), khi $w_i = 1/\sqrt{\lambda_i}$, c = 1: Các kết quả phát hiện cũng tương đương như công thức (5) nhưng công thức (6) đơn giản hơn vì không phải tính bình phương cho mỗi thành phần chính trong *d*.

Theo công thức (7), khi $w_i = 1/\lambda_i$, c = 1: Các kết quả phát hiện cũng tương đương như công thức (5), (6) nhưng không phải tính căn bậc hai của trị riêng trong pha huấn luyện.

Theo công thức (8), khi $w_i = 1$, c = 1: Đây là công thức đơn giản nhất nhưng kết quả phát hiện cũng tương đương như các công thức (4), (5), (6), (7) khi sử dụng các thành phần chính thứ yếu. Với k = 14, cũng giống như công thức (4), sự chênh lệch về giá trị của các thành phần chính làm giảm tỷ lệ TPR.



Hình 1. Đồ thị ROC của d biểu diễn quan hệ giữa tỷ lệ FPR và TPR với dữ liệu huấn luyện sạch khi wi =1, c=1, k=3

Hình 1 là đồ thị ROC [23] với khoảng cách d biểu thị mối quan hệ giữa tỷ lệ cảnh báo sai (FPR) trên trục hoành và tỷ lệ cảnh báo đúng (TPR) trên trục tung khi w=1, c=1, k=3. Điểm hoàn hảo là điểm góc trên bên trái với tọa độ (0,1) khi TPR là 100% và FPR là 0%. Trên thực tế không thể đạt được kết quả như vậy. Việc lựa chọn điểm tối ưu rất khó thực hiện được vì điểm này liên tục thay đổi với lưu lượng mạng. Trong các thử nghiệm, bài báo lựa chọn mức ngưỡng cố định theo hàm phân bố tích lũy thực nghiệm với tỷ lệ sai số ước tính (FPR) trong khoảng 2-5%.

Từ kết quả thử nghiệm trên có thể thấy, trong điều kiện dữ liệu huấn luyện sạch, có thể lựa chọn các thành phần chính thứ yếu với k = 2 hoặc k = 3 để giảm số chiều dữ liệu.

Khi dữ liệu huấn luyện của dPCA bán giám sát chứa ngoại lai, kết quả TPR rất thấp (bảng III). Số lượng kết nối trước khi loại bỏ ngoại lai là nhỏ hơn hoặc bằng 10000. Số lượng ngoại lai trong dữ liệu huấn luyện là 10% trên tổng số kết nối bình thường.Lưu ý là TPR cho biết khả năng phát hiện bất thường trên tổng số bất thường được thử nghiệm. Nếu khả năng phát hiện bất thường thấp sẽ làm cho hiệu quả của phương pháp đề xuất suy giảm. Ngoại lai làm ảnh hưởng đến các thông số huấn luyện và làm sai lệch kết quả phát hiện.

Bảng III. Kết quả phát hiện của dPCA trước khi loại bỏ ngoại lai trong dữ liệu huấn luyện

w_i	c	k	PR (%)	TPR (%)	FPR (%)	TA (%)
1	2	3	98.9	4.66	1	35.7
$1/\lambda_i$	2	14	97.7	16.4	0.8	43.4
$1/\sqrt{\lambda_i}$	1	14	98.2	16	0.6	43.2

Bảng IV. Kết quả phát hiện và loại bỏ ngoại lai của dPCA ở chế độ không giám sát

w_i	c	k	PR (%)	TPR (%)	FPR (%)	TA (%)
$1/\lambda_i$	2	14	50	100	11.1	90
$1/\sqrt{\lambda_i}$	1	14	49.7	99.3	11.1	89.9
$1/\lambda_i$	1	14	49.3	98.6	11.2	89.7

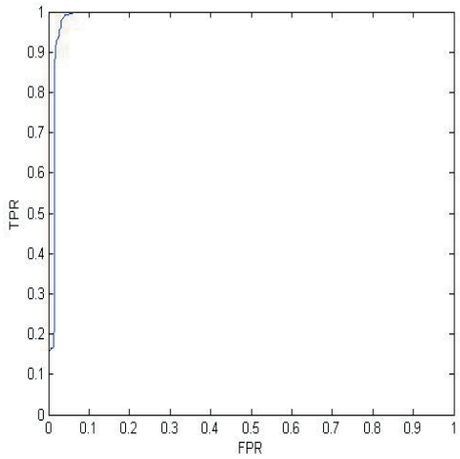
Bảng IV thống kê tỷ lệ phát hiện với phương pháp dPCA chế độ không giám sát (trước khi huấn luyện) với mục đích chính là loại bỏ ngoại

lai cho dữ liệu huấn luyện. Tất cả những ngoại lai phát hiện được sẽ bị loại bỏ tập dữ liệu huấn luyện của chế độ bán giám sát. Ngưỡng của *d* được đặt bằng 75% đến 80% của hàm phân bố tích lũy thực nghiệm. Chế độ không giám sát của dPCA vẫn có thể sử dụng để phát hiện bất thường khi cần thiết.

Bảng V. Kết quả phát hiện của dPCA bán giám sát sau khi loại bỏ ngoại lai trong dữ liệu huấn luyện

w_i	c	k	PR (%)	TPR (%)	FPR (%)	TA (%)
1	2	3	78.9	99.1	5.1	95.6
$1/\lambda_i$	2	3	73.2	97.4	6.9	93.8
$1/\lambda_i$	2	14	78	100	5.4	95.4
$1/\sqrt{\lambda_i}$	1	3	75.2	100	6.4	94.6
$1/\sqrt{\lambda_i}$	1	3	84	98.3	3.6	96.7
$1/\sqrt{\lambda_i}$	1	14	76.6	100	6	95
1	1	3	84.3	93.3	3.4	96.1

Bảng V là kết quả phát hiện của dPCA chế độ bán giám sát sau khi đã loại bỏ ngoại lai. Hình 3 là đồ thị ROC của *d* sau khi khử ngoại lai trong dữ liệu huấn luyện với dPCA chế độ không giám sát khi w_i =1, c=1, k=3.



Hình 2. Đồ thị ROC của d sau khi khử ngoại lai với dPCA chế độ không giám sát trong dữ liệu huấn luyện với wi =1, c=1, k=3

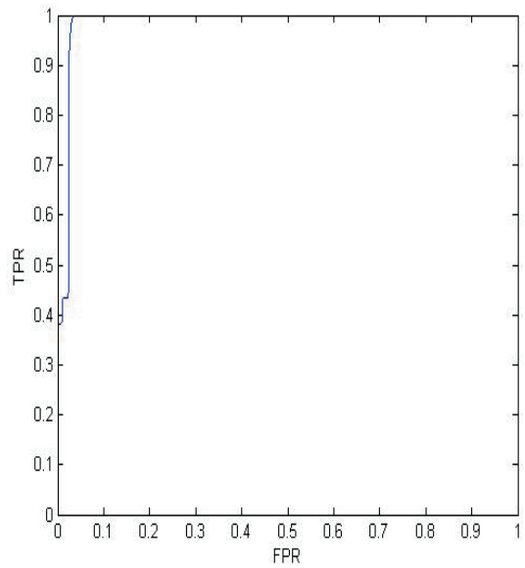
Bảng VI thống kê kết quả phát hiện và loại bỏ ngoại lai bằng K-means. Bảng VII là kết quả phát hiện sau khi loại bỏ ngoại lai bằng K-means trong dữ liệu huấn luyện cho chế độ bán giám sát của dPCA. Hình 4 là đồ thị ROC của d sau khi khử ngoại lai trong dữ liệu huấn luyện với K-means khi $w_i = 1/\lambda_i$, c = 1, k = 3.

Bảng VI. Kết quả phát hiện và loại bỏ ngoại lai bằng K-Means trong dữ liệu huấn luyện

Test	PR (%)	TPR (%)	FPR (%)	TA (%)
1	42.1	100	15.2	86.3
2	39.6	100	16.9	84.8
3	44.2	98.5	14	87.4
4	30.5	100	25.3	77.1
5	34.7	83.5	17.5	82.6

Bảng VII. Kết quả phát hiện của dPCA sau khi loại bỏ ngoại lai bằng K-Means trong dữ liệu huấn luyện

w_i	k	PR (%)	TPR (%)	FPR (%)	TA (%)
2	3	79.7	100	15.7	90
2	3	77.5	100	18.3	88.8
2	14	77.7	100	18.1	88.88
1	3	79.8	100	15.6	90.3
1	14	79.6	84.3	13.4	85.7
1	3	91.8	100	5.6	96.6
1	14	86.4	100	9.9	93.9
1	3	77.8	97.5	17.5	88.3



Hình 4. Đồ thị ROC của d sau khi khử ngoại lai dữ liệu huấn luyện với K-means $w_i = 1/\lambda_i$, c=1, k=3

Sự chính xác của dPCA chế độ bán giám sát phụ

thuộc chất lượng dữ liệu huấn luyện bao gồm số lượng ngoại lai vẫn còn lẫn vào dữ liệu huấn luyện, mức độ biến thiên của lưu lượng mạng trong điều kiện bình thường và sự khác biệt giữa các kết nối bình thường với bất thường được tính trong *d*. Những kết quả thống kê ở trên đạt được trong điều kiện dữ liệu mạng bình thường khác biệt đáng kể với điều kiện bất thường. Tuy nhiên, có những khoảng thời gian sự khác biệt này bị thu hẹp làm cho kết quả phát hiện có sai số lớn. Bảng VIII là kết quả thống kê của dPCA với dữ liệu bình thường có đột biến trong một ngày với tập dữ liệu Kyoto Honeypot.

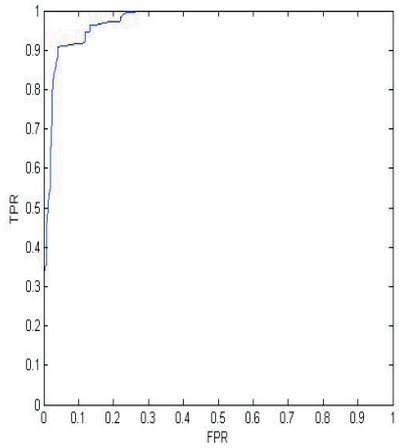
Bảng VIII. Kết quả phát hiện của dPCA khi dữ liệu bình thường có sự thay đổi đột biến

w_i	c	k	PR (%)	TPR (%)	FPR (%)	TA (%)
1	2	3	94.4	54.4	2	81.3
λi/1	2	3	94.5	55.5	2	81.78
λi/1	2	14	95	65.4	2.1	85.4
$1/\sqrt{\lambda_i}$	1	3	94.3	53.9	2	81.2
$1/\sqrt{\lambda_i}$	1	14	95.6	70.3	2	87.4
1	1	3	94.3	54	2	81.2

Kết quả bảng VIII cho thấy, tỷ lệ TPR của k = 14 cao hơn k = 3. Lý do là vì bất thường trong trường hợp này không chỉ xuất hiện ở PC thứ yếu mà cả trong các thành phần chính khác. Do vậy bên cạnh dPCAL1, sử dụng dPCAL2 với nhiều thành phần chính hơn (k = 14) để phát hiện bất thường là cần thiết.

Để có kết quả TPR tốt hơn, ta có thể thay đổi mức ngưỡng với giá trị phù hợp với đồ thị ROC của *d*. Thí dụ với trường hợp $w_i=1/\lambda_i$, c=1, k=14 (đồ thị ROC ở Hình 5), nếu hạ mức ngưỡng xuống 95% hàm ecdf thì kết quả sẽ là PR = 91.8%, TPR = 91%, FPR = 5%, TA = 93.5%. Tương tự với trường hợp $w_i=1/\sqrt{\lambda_i}$, c=1, k=14 nếu hạ mức ngưỡng xuống 92% của hàm ecdf thì kết quả sẽ là PR = 87.5%, TPR = 90.2%, FPR = 8%, TA = 91.3%.

Nhưng như đã trình bày ở phần trên, chọn mức ngưỡng phù hợp với điểm tối ưu giữa TPR và FPR là rất khó vì lưu lượng mạng thực tế làm cho điểm này thay đổi thường xuyên. Một mức ngưỡng cho kết quả tốt với thời điểm này lại không phù hợp với thời điểm khác. Một giải pháp cho vấn đề này là coi những dữ liệu bình thường gây ra đột biến lưu lượng cũng là ngoại lai. Từ đó, thay vì đặt lại mức ngưỡng cho *d*, có thể áp dụng phương pháp khử ngoại lai đã trình bày ở trên để làm sạch dữ liệu huấn luyện.



Hình 5. Đồ thị ROC của d với $w_i=1/\lambda_i$, c=1, k=14

Bảng IX. Kết quả phát hiện tốt hơn (TPR) của dPCA so với Bảng VIII khi khử ngoại lai trong dữ liệu huấn luyện

w_i	c	k	PR (%)	TPR (%)	FPR (%)	TA (%)
1	2	3	81.2	92.6	13.2	89
$1/\sqrt{\lambda_i}$	1	3	81.7	93.5	13	89.5
$1/\sqrt{\lambda_i}$	1	14	79.7	83.9	13.2	85.7
λi/1	1	3	86.6	99.8	9.5	94.1
λi/1	1	14	80	93	14.4	88.5
1	1	3	80.1	91.8	14.1	88.17

Kết quả phát hiện của dPCA sau khi khử những ngoại lai này được thống kê Bảng IX. Như vậy, việc khử ngoại lai có thể áp dụng cho cả trường hợp dữ liệu bình thường nhưng gây ra sự thay đổi đột biến về lưu lượng hoặc không giống với đại đa số dữ liệu bình thường khác.

Các kết quả thử nghiệm cho thấy, việc áp dụng dPCA với dPCAL1 và dPCAL2 chấp nhận được trong thực tế. Đây không phải là những kết quả phát hiện tốt nhất mà phương pháp đề xuất đạt

được nhưng các tác giả đưa vào bài báo để đảm bảo tính khách quan với sự thay đổi của lưu lượng mạng. Những biến động trong trạng thái bình thường của lưu lượng mạng và sự phức tạp của sự kiện bất thường, dPCAL1 có thể không phát hiện được hết những bất thường có thể xảy ra. Vì thế dPCAL2 là giải pháp hỗ trợ cho dPCAL1 khi bất thường xuất hiện tại những thành phần chính khác.

V. KẾT LUẬN

Các phương pháp phát hiện lưu lượng mạng bất thường khó và phức tạp hơn nhiều so với phương pháp phát hiện dựa trên dấu hiệu truyền thống do không biết trước mẫu dấu hiệu. Ngoài ra, có nhiều vấn đề thách thức như cần mô hình hóa trạng thái bình thường, trích chọn đặc trưng dữ liệu sao cho giảm độ phức tạp trong khi vẫn bảo đảm độ chính xác và tốc độ phát hiện, loại bỏ ngoại lai gây sai lệch trong dữ liệu huấn luyện.

Kết quả nghiên cứu cho thấy có thể áp dụng thuật toán PCA cơ sở để chuyển dữ liệu sang miền con PCA nhằm giảm chiều dữ liệu nhằm khắc phục nhược điểm trên. Việc lựa chọn các thành phần chính chủ yếu và thứ yếu giúp giảm độ phức tạp, tăng được độ chính xác khi cần. Bài báo đã đề xuất phương pháp dPCA cải tiến từ PCA cơ sở với phương pháp tính khoảng cách mới là tổng hợp cho các công thức tính trước đây. Khi dữ liệu chuyển sang miền con PCA, dPCA có thể phát hiện một số loại ngoại lai hiệu quả hơn. dPCA dùng phương pháp tính khoảng cách mới để chọn đường cơ sở và phân cụm để khử ngoại lai trong dữ liệu huấn luyện. Kết quả thử nghiệm cho thấy sự thay đổi độ chính xác của phương pháp đề xuất với những tham số khác nhau của thuật toán tính khoảng cách cũng như số lượng thành phần chính tùy theo yêu cầu thực tế. Kết quả phân tích và thử nghiệm cũng cho thấy dPCA phát hiện được lưu lượng mạng bất thường trong điều kiện dữ liệu huấn luyện chứa ngoại lai. Đây cũng là một đóng góp của bài so với các công trình nghiên cứu trước đây.

TÀI LIỆU THAM KHẢO

[1]. M.Bhuyan, D.Bhattacharyya, J.Kalita, "Network anomaly detection: Methods, systems and tools," *IEEE Communications Surveys Tutorials*, vol. 16, no. 1, pp. 303-336, 2014

[2]. S. Myers, J. Musacchio, N. Bao, "Intrusion Detection Systems: A Feature and Capability Analysis," *Tech. Report UCSC-SOE-10-12*, Jack Baskin School of Engineering, 2010.

[3]. K. Wankhade, S. Patka, R. Thool, "An Overview of Intrusion Detection Based on Data Mining Techniques," *Proc. of IEEE CSNT*, 2013.

[4]. C. Kacha, K. A. Shevade, "Comparison of Different Intrusion Detection and Prevention Systems," *Intl. Journal of Emerging Technology and Advanced Engineering*, vol. 2, no.12, pp.243-245, 2012

[5]. A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," *Proc. of ACM SIGCOMM*, pp. 219–230, 2004

[6]. A. Patcha, J.M. Park, "An Overview of Anomaly Detection Techniques: Existing Solutions and Latest Technological Trends" *The International Journal of Computer and Telecom-munications Networking*, vol. 51, no. 12, pp. 3448-3470, Aug. 2007.

[7]. W. Zhang, Q. Yang, Y. Geng, "A Survey of Anomaly Detection Methods in Networks," *Proc. of International Symposium on Computer Network and Multimedia Technology*, Jan. 2009, pp. 1–3.

[8]. M. Thottan, G. Liu, C. Ji, *Anomaly Detection Approaches for Communication Networks: Algorithms for Next Generation Networks*, G. Cormode, Ed. London: Springer, 2010, pp. 239-261.

[9]. V. Jyothsna, V. V. Rama Prasad, K. M. Prasad, "A Review of Anomaly based Intrusion Detection Systems," *International*

Journal of Computer Applications, vol. 28, no. 7, pp. 28-34, 2011.

[10]. A. Jain, B. Verma, J. L. Rana, "Anomaly Intrusion Detection Techniques: A Brief Review," *International Journal of Scientific & Engineering Research*, vol. 5, no. 7, pp. 17-23, 2014

[11]. Y. Bouzida. Efficient intrusion detection using principal component analysis. *Proc. of 7th World Multiconference on Systemics, Cybernetics and Informatics*, 2003.

[12]. A. Lakhina, M. Crovella, and C. Diot, "Mining anomalies using traffic feature distributions," *Proc. of ACM SIGCOMM*, 2005.

[13]. W. Wang and R. Battiti, "Identifying Intrusions in Computer Networks with Principal Component Analysis," *Proc. of IEEE ARES*, 2006.

[14]. M. Shyu, S. Chen, K. Sarinnapakorn, L. Chang. *Principal Componentbased Anomaly Detection Scheme. Foundations and Novel Approaches in Data Mining*, vol. 9, pp. 311-329, 2006.

[15]. D. Brauckhoff, K. Salamatian, M. May, "Applying PCA for Traffic Anomaly Detection: Problems and Solutions," *Proc. of IEEE INFOCOM*, 2009.

[16]. L. Mechtri, F. D. Tolba, N. Ghoualmi, "*Intrusion detection using principal component analysis*," Proc. of IEEE ICESMA, 2010.

[17]. L. Ertoz, E. Eilertson, A. Lazarevic, P. Tan, V. Kumar, and J. Srivastava, "*Data Mining-Next Generation Challenges and Future Directions*," MIT Press, 2004

[18]. Nguyễn Hà Dương, Hoàng Đăng Hải, "Phát hiện lưu lượng mạng bất thường sử dụng phương pháp PCA trong lựa chọn đặc tính dữ liệu," *Chuyên san các công trình nghiên cứu về điện tử, viễn thông và công nghệ thông tin, Tạp chí Khoa học công nghệ*, Tập 53, Số 2C, 2015, tr.52-64.

[19]. M. Tavallaee, E. Bagheri, W. Lu, A. A. Ghorbani. A Detailed Analysis of the KDD CUP 99 Data Set., *Proc. of IEEE CISDA* 2009.

[20]. The KDD99 cup data, http://kdd.ics.uci.edu/databases/kddcup 99/kddcup99.html, 1999.

[21]. The NSL-KDD data, http://nsl.cs.unb.ca/nsl-kdd, 2009.

[22]. J. Song, H. Takakura, Y. Okabe, "Description of Kyoto University Benchmark Data," 2006, pp. 1-3. (http://www.taka kura.com/Kyoto data/BenchmarkData-Description-v5.pdf).

[23]. https://vi.wikipedia.org/wiki/Duong_cong_ROC

NETWORK TRAFFIC ANOMALY DETECTION |WITH OUTLIER IN TRAINING DATA

***Abstract*:** Network traffic anomaly detection has many challenges: adjust threshold, extract data features, reduce data dimension, precision parameters, etc. Besides that, outliers can significantly impact the performance of detection. This paper describes the issues of network traffic anomaly detection with outliers in training data and proposes an enhanced method (called dPCA) based on principal component analysis algorithm. The experiment was evaluated witth Kyoto Honeypot dataset.

Nguyễn Hà Dương, KS (2001), ThS. (2003) tại ĐH Bách Khoa Hà Nội. Giảng viên Khoa CNTT, Trường ĐH Xây dựng Hà Nội. Lĩnh vực nghiên cứu: Mạng và hệ thống thông tin, an ninh mạng, viễn thông.

Hoàng Đăng Hải, PGS.TSKH., TS. (1999), TSKH. (2003) tại Đại học Tổng hợp Kỹ thuật Ilmenau, CHLB Đức. Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: Chất lượng dịch vụ, giao thức truyền thông, hiệu năng mạng, mạng và hệ thống thông tin, an ninh mạng, viễn thông.