

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG



NGUYỄN HOÀNG GIANG

**NGHIÊN CỨU CÁC PHƯƠNG PHÁP HỌC MÁY VÀ ỨNG DỤNG
TRONG PHÁT HIỆN XÂM NHẬP TRÁI PHÉP**

CHUYÊN NGÀNH : HỆ THỐNG THÔNG TIN

MÃ SỐ: 60.48.01.04

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI - 2016

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: PGS. TS. Trần Quang Anh

Phản biện 1: TS. Hoàng Xuân Dậu

Phản biện 2: TS. Nguyễn Vĩnh An

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại
Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: 09 giờ 30 ngày 20 tháng 08 năm 2016

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỞ ĐẦU

Tính cấp thiết của đề tài

Vấn đề đảm bảo an toàn thông tin cho các hệ thống mạng, hệ thống thông tin là một vấn đề không mới nhưng ngày càng trở thành một đòi hỏi có tính cấp bách hiện nay. Cùng với sự phát triển một cách nhanh chóng của hệ thống mạng Internet và các tiện ích của nó, việc bảo đảm an toàn thông tin càng trở nên cấp thiết hơn bao giờ hết. Đặc biệt trong bối cảnh một vài năm trở lại đây, các hệ thống mạng của Việt Nam liên tục trở thành các mục tiêu tấn công phá hoại, xâm nhập trái phép, lấy cắp thông tin của Hacker nước ngoài, trong đó đặc biệt là từ các Hacker Trung Quốc. Đây thực sự là tình trạng báo động bởi rất nhiều các máy bị truy cập trái phép thuộc các mạng của các Bộ, ban, ngành, tổ chức lưu giữ, soạn thảo các thông tin quan trọng quốc gia hoặc các thông tin tài chính, kỹ thuật quan trọng.

Có nhiều phương pháp cả về kỹ thuật lẫn chính sách đã được đề xuất, áp dụng nhằm đảm bảo an toàn thông tin như: triển khai các hệ thống tường lửa (Firewall) nhiều lớp, hệ thống phát hiện xâm nhập trái phép (IDS), các hệ thống xác thực, các hệ thống bảo mật thiết bị đầu cuối (Endpoint). Tuy vậy, giải pháp phát hiện xâm nhập trái phép (IDS) vẫn luôn là một trong những giải pháp quan trọng, được quan tâm triển khai.

Có nhiều cách để phát hiện các xâm nhập trái phép hệ thống mạng. Thông thường được phân làm 2 loại chính: Misuse Detection (phát hiện dựa trên sự lạm dụng) và Anomaly Detection (phát hiện bất thường). Cách thức phát hiện dựa trên sự lạm dụng phân tích các hoạt động của hệ thống, tìm kiếm các sự kiện giống với các mẫu tấn công đã biết trước. Các mẫu tấn công biết trước này gọi là các dấu hiệu tấn công. Do vậy cách thức này còn được gọi là cách thức phát hiện dựa trên dấu hiệu. Kiểu phát hiện tấn công này có ưu điểm là phát hiện các cuộc tấn công nhanh và chính xác, không đưa ra các cảnh báo sai làm giảm khả năng hoạt động của mạng và giúp các người quản trị xác định các lỗ hổng bảo mật trong hệ thống của mình. Tuy nhiên, cách thức này có nhược điểm là không phát hiện được các cuộc tấn công không có trong cơ sở dữ liệu, các kiểu tấn công mới, do vậy hệ thống luôn phải cập nhật các mẫu tấn công mới. Trong khi đó cách thức phát hiện bất thường là kỹ thuật phát hiện thông minh, nhận dạng ra các hành động không bình thường của mạng. Quan niệm của cách thức này về các cuộc tấn công là khác so với các hoạt động thông thường. Ban đầu, chúng lưu trữ các mô tả sơ lược về các hoạt động bình thường của hệ thống. Các cuộc tấn công sẽ có những hành động khác so với bình thường và cách thức phát hiện này có thể nhận dạng. Do đó, cách thức phát hiện xâm nhập trái phép dựa trên bất thường mạng hiện nay đang trở thành hướng nghiên cứu chủ yếu đối với các hệ thống phát hiện xâm nhập trái phép.

Học máy (Machine learning) là kỹ thuật cho phép giải quyết vấn đề hoặc ra quyết định dựa trên dữ liệu và kinh nghiệm. Với học máy, chương trình máy tính sử dụng kinh nghiệm, quan sát, hoặc dữ liệu trong quá khứ để cải thiện công việc của mình trong tương

lai thay vì chỉ thực hiện theo đúng các quy tắc đã được lập trình sẵn. Chính vì thế, việc ứng dụng Học máy trong phát hiện xâm nhập trái phép, đặc biệt đối với phát hiện bất thường, là phù hợp và cần thiết trong bối cảnh hiện nay. Chính vì vậy, học viên chọn đề tài luận văn “Nghiên cứu các phương pháp học máy và ứng dụng trong phát hiện xâm nhập trái phép”, trong đó tập trung nghiên cứu ứng dụng học máy trong phát hiện xâm nhập mạng bất thường.

Mục đích nghiên cứu

- Nghiên cứu các phương pháp học máy;
- Nghiên cứu một số cách thức phát hiện xâm nhập trái phép. Từ đó ứng dụng phương pháp học máy phát hiện bất thường mạng.

Đối tượng nghiên cứu và phạm vi nghiên cứu

- Đối tượng nghiên cứu: Các phương pháp học máy; phương pháp phát hiện xâm nhập trái phép đặc biệt là phát hiện bất thường mạng; các luồng dữ liệu trên mạng.
- Phạm vi nghiên cứu: Phát hiện bất thường trong mạng máy tính.

Cấu trúc luận văn

Nội dung của luận văn được trình bày trong ba phần chính như sau:

1. Phần mở đầu
2. Phần nội dung: bao gồm ba chương

Chương 1: Tổng quan về các phương pháp học máy

Trong chương này, luận văn sẽ trình bày khái niệm về học máy; phân loại các phương pháp học máy chủ yếu là phương pháp học máy có giám sát và phương pháp học máy không có giám sát; tiếp đó sẽ trình bày một số thuật toán học máy tiêu biểu của các phương pháp học máy, trong đó đi sâu vào 3 thuật toán học máy là SVM, K-NN và One-class SVM sẽ được sử dụng trong ứng dụng phát hiện xâm nhập mạng bất thường nêu ở chương 3.

Chương 2: Phát hiện xâm nhập trái phép và cách tiếp cận bằng phương pháp học máy

Trong chương này, luận văn sẽ trình bày khái niệm về xâm nhập trái phép; phân loại phát hiện xâm nhập trái phép dựa trên nguồn dữ liệu (Network-based, Host-based) và dựa theo phương pháp xử lý (Misuse Detection, Anomaly Detection); cách tiếp cận phát hiện xâm nhập trái phép dựa trên bất thường bằng phương pháp học máy; Mô tả bài toán đề xuất trong luận văn.

Chương 3: Ứng dụng phương pháp học máy KNN, SVM và One-class SVM để phát hiện bất thường

Trong chương này, luận văn sẽ trình bày về việc thử nghiệm ứng dụng phương pháp học máy để phát hiện xâm nhập mạng bất thường: giới thiệu về mô hình thử nghiệm, trong đó giới thiệu về cách thức xây dựng các bộ dữ liệu thử nghiệm ở dạng Netflow (từ bộ dữ

liệu thử nghiệm dạng Tcpdump của DARPA, ISCX); cách thức cài đặt thử nghiệm; kết quả thử nghiệm và đánh giá kết quả thử nghiệm.

3. Phần kết luận

CHƯƠNG 1. TỔNG QUAN VỀ CÁC PHƯƠNG PHÁP HỌC MÁY

1. Giới thiệu về học máy

Học máy, còn gọi là Máy học, (tiếng Anh: Machine Learning) là một lĩnh vực của trí tuệ nhân tạo liên quan đến việc nghiên cứu và xây dựng các kỹ thuật cho phép các hệ thống "học" tự động từ dữ liệu để giải quyết những vấn đề cụ thể.

Học máy là khả năng của chương trình máy tính sử dụng kinh nghiệm, quan sát, hoặc dữ liệu trong quá khứ để cải thiện công việc của mình trong tương lai thay vì chỉ thực hiện theo đúng các quy tắc đã được lập trình sẵn.

2. Phân loại các phương pháp học máy

2.1. Phương pháp học máy có giám sát

Học có giám sát (supervised learning): là một kỹ thuật của ngành học máy để xây dựng một hàm (function) từ dữ liệu huấn luyện. Dữ liệu huấn luyện bao gồm các cặp gồm đối tượng đầu vào (thường dạng vec-tơ), và đầu ra mong muốn. Đầu ra của một hàm có thể là một giá trị liên tục (gọi là hồi qui), hay có thể là dự đoán một nhãn phân loại cho một đối tượng đầu vào (gọi là phân loại). Nhiệm vụ của chương trình học có giám sát là dự đoán giá trị của hàm cho một đối tượng bất kỳ là đầu vào hợp lệ, sau khi đã xem xét một số ví dụ huấn luyện (nghĩa là, các cặp đầu vào và đầu ra tương ứng). Để đạt được điều này, chương trình học phải tổng quát hóa từ các dữ liệu sẵn có để dự đoán được những tình huống chưa gặp phải theo một cách "hợp lý".

Học có giám sát có thể tạo ra 2 loại mô hình. Phổ biến nhất, học có giám sát tạo ra một mô hình toàn cục (*global model*) để ánh xạ đối tượng đầu vào đến đầu ra mong muốn. Tuy nhiên, trong một số trường hợp, việc ánh xạ được thực hiện dưới dạng một tập các mô hình cục bộ.

Các phương pháp học máy có giám sát thông dụng:

- K-Nearest Neighbor (K-NN);
- Support Vector Machine (SVM)

2.2. Phương pháp học máy không có giám sát

Học không có giám sát (unsupervised learning) là một phương pháp của ngành học máy nhằm tìm ra một mô hình mà phù hợp với các quan sát. Khác với phương pháp học máy có giám sát, phương pháp học không có giám sát là dạng học máy trong đó các ví dụ được cung cấp nhưng không có giá trị đầu ra hay giá trị đích (hay nói cách khác, dữ liệu huấn luyện không được gán nhãn phân loại). Thay vì xác định giá trị đích, thuật toán học máy dựa trên độ tương tự giữa các ví dụ để xếp chúng thành những nhóm, mỗi nhóm gồm

các ví dụ tương tự nhau. Hình thức học không giám sát như vậy gọi là phân cụm (clustering). Ngoài phân cụm, một dạng học không giám sát phổ biến khác là phát hiện luật kết hợp (association rule). Luật kết hợp có dạng $P(A | B)$, cho thấy xác suất hai tính chất A và B xuất hiện cùng với nhau.

Các phương pháp học máy không có giám sát thông dụng:

- One-class SVM;

3. Một số thuật toán học máy

3.1. Thuật toán học máy có giám sát: *Support Vector Machine (SVM)*

SVM là một công cụ mạnh và phổ biến để phân lớp dữ liệu lớn và nhiều chiều. Máy vector hỗ trợ (SVM) được đề xuất bởi V.Vapnik và các đồng nghiệp vào những năm 1970 ở Nga, và sau đó đã trở nên nổi tiếng trong những năm 90 của thế kỉ trước.

3.1.1. Tổng quan về SVM

Cho một tập dữ liệu huấn luyện biểu diễn trong không gian vector, trong đó mỗi dữ liệu được biểu diễn là một điểm. Đặc điểm cơ bản của thuật toán SVM là tìm ra một siêu phẳng quyết định tốt nhất để chia các điểm trong không gian thành hai lớp riêng biệt được đánh số là +1 và -1. Chất lượng của siêu phẳng được quyết định bởi khoảng cách (được gọi là biên – margin) của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng này. Khoảng cách biên càng lớn thì siêu phẳng quyết định càng tốt và việc phân loại càng chính xác. Mục đích của SVM là tìm được khoảng cách biên lớn nhất và lỗi tách sai là bé nhất.

Như vậy việc tìm ra biên cực đại là vấn đề cốt lõi quyết định đến chất lượng phân loại của phương pháp SVM.

Cho tập mẫu $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ với $x_i \in \mathbb{R}^n$ thuộc vào hai lớp nhãn $y_i \in \{-1, +1\}$ là nhãn lớp tương ứng của các x_i với -1 là nhãn biểu thị lớp âm, +1 là nhãn biểu thị lớp dương.

Ta có phương trình siêu phẳng chứa vector \vec{x}_i trong không gian như sau:

$$\vec{x}_i \cdot \vec{w} + b = 0 \quad (1.1)$$

Trong đó w là vector pháp tuyến n chiều và b là giá trị ngưỡng.

Vector pháp tuyến w xác định chiều của siêu phẳng $h(x)$, còn giá trị ngưỡng b xác định khoảng cách giữa siêu phẳng và gốc.

$$\text{Đặt } h(\vec{x}_i) = \text{sign}(\vec{x}_i \cdot \vec{w} + b) = \begin{cases} +1 & \text{ khi } \vec{x}_i \cdot \vec{w} + b > 0 \\ -1 & \text{ khi } \vec{x}_i \cdot \vec{w} + b < 0 \end{cases} \quad (1.2)$$

Bài toán đặt ra là làm thế nào để tìm ra mặt phẳng phân tách $h(\vec{x})$ tối ưu, tức là tìm w và b sao cho biên của mặt phẳng phân tách là cực đại.

3.1.2. Phân lớp tuyến tính

Hình thức đơn giản của việc phân lớp là phân lớp nhị phân: chỉ gồm hai lớp dương (+1) hoặc âm (-1). Phân lớp nhị phân sử dụng hàm giá trị thực $h: X \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ là hàm tuyến tính, tương ứng đầu ra $y \in \{-1, 1\}$. Có đầu vào $x = (x_1, x_2, \dots, x_n)$ được gán nhãn $y = (-1, +1)$.

$$\text{Với } \begin{cases} X^+ = \{x^i | y^i = 1\} \\ X^- = \{x^i | y^i = -1\} \end{cases} \text{ thì } \begin{cases} h(x) > 0, \forall x \in X^+ \\ h(x) < 0, \forall x \in X^- \end{cases} \quad (1.3)$$

Bởi $h(x)$ là hàm tuyến tính nên

$$h(x) = \langle w, x \rangle + b = \sum_{i=1}^n w_i x_i + b \quad (1.4)$$

Từ đó rút ra siêu phẳng phân chia thỏa mãn công thức sau:

$$y_i(w^i x_i + b) \geq 1 \text{ với } i = 1, 2, \dots, n \quad (1.5)$$

Về mặt hình học, các phần tử của không gian đầu vào X sẽ thuộc một trong hai phần được phân tách bởi siêu phẳng xác định bởi biểu thức $\langle w, x \rangle + b = 0$ với $\langle w, x \rangle$ là tích vô hướng.

Trong không gian hai chiều thì các điểm có phương trình $\langle w, x \rangle = 0$ tương ứng với một đường đi qua gốc tọa độ, còn trong không gian ba chiều thì nó là một mặt phẳng qua gốc tọa độ. Biên b dịch chuyển mặt phẳng đi một khoảng so với mặt phẳng gốc tọa độ.

Như đã đề cập ở trên, vấn đề cốt lõi là làm sao để tìm được siêu phẳng $h(x)$ tối ưu, tức là tìm w và b để khoảng cách biên là cực đại.

Gọi m là độ rộng của biên. Ta phải tìm w và b sao cho m là lớn nhất.

Với mỗi mặt phẳng phân tách thì ta định nghĩa ra được hai đường nằm về hai phía tách các điểm có dấu dương và dấu âm gọi là đường cộng và đường trừ. Đường cộng là đường nằm về phía lớp có dấu dương, đường trừ nằm về phía dấu âm. Các điểm nằm trên hai đường này được gọi là các vector hỗ trợ.

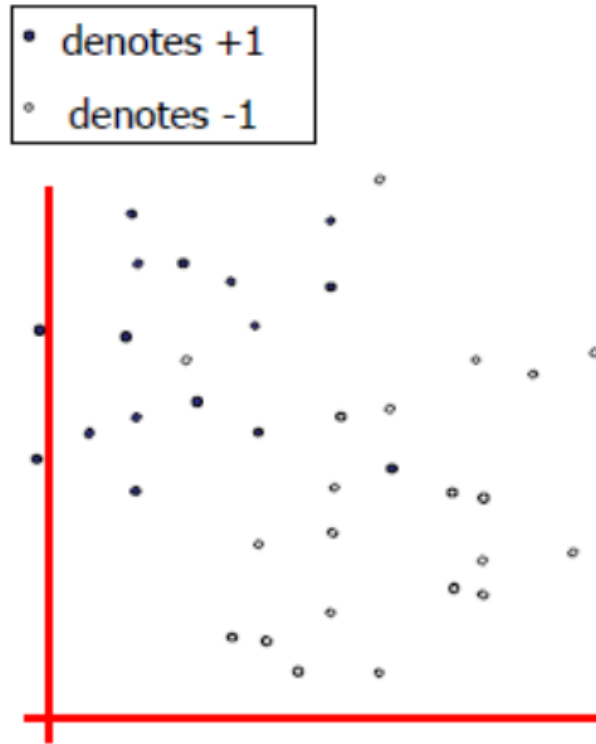
Phương trình của đường cộng là $w^T x + b = 1$

Phương trình của đường trừ là $w^T x + b = -1$

Tính được w và b ta có được phương trình của siêu phẳng.

3.1.3. Phân lớp tuyến tính với trường hợp không phân tách được

Thuật toán phân lớp tuyến tính trong trường hợp không phân tách được còn gọi là thuật toán C-SVM. Giả sử có bài toán như sau:



Hình 1.1: Ví dụ bài toán phân loại tuyến tính không tách biệt

Theo hình trên, tồn tại các điểm có nhãn dương nhưng nằm về phía nhãn âm và ngược lại. Trong thực tế, các dữ liệu thường không phân chia tuyến tính. Trường hợp phân lớp tuyến tính và phân tách được là lý tưởng và ít xảy ra. Nếu tập dữ liệu chứa nhiều thì các điều kiện có thể không được thỏa mãn, do đó có thể không tìm được giá trị tối ưu của w và b . Để chấp nhận một số lỗi, ta thay thế ràng buộc dạng bất đẳng thức ở (1.5) thành:

$$y_i(w \cdot x_i + b) \geq 1 - \zeta_i \text{ với } i = 1, 2, \dots, n \quad (1.6)$$

trong đó ζ_i là các biến phụ không âm.

Đối với một ví dụ bị lỗi thì $\zeta_i > 1$ và $\sum_i \zeta_i$ là giới hạn trên của lỗi của các ví dụ huấn luyện.

Ngoài ra ta phải tích hợp lỗi vào hàm mục tiêu bằng cách gán giá trị chi phí cho các lỗi.

Hàm tối ưu hóa trở thành:

$$\frac{1}{2} \|w\|^2 + C \left(\sum_{i=1}^n \zeta_i \right)^k \quad (1.7)$$

với ràng buộc:

$$\begin{cases} y_i(w \cdot x_i + b) \geq 1 - \zeta_i \\ \zeta_i \geq 0 \end{cases}$$

Trong đó $C > 0$ là tham số xác định mức độ chi phí lỗi (penalty degree).

Giá trị của C càng lớn thì mức độ chi phí càng cao đối với các lỗi. Nó thiết lập mức độ quan trọng của việc cực đại biên và giảm số lượng biến phụ ζ_i . Đây chính là công thức SVM biên mềm (Cortes và Vapnik, 1995). Giá trị k thường được sử dụng là 1 để thu được biểu thức đối ngẫu đơn giản hơn.

Như vậy theo như hàm tối ưu hóa ở trên, chúng ta vẫn sẽ phải tìm cực tiểu của $\|w\|^2$, ngoài ra phải thêm cả khoảng cách của các điểm lỗi đến vị trí đúng của nó

Để giải quyết bài toán cực tiểu hóa này, ta cũng sử dụng biến đổi Lagrange như trên, với biểu thức Lagrange như sau:

$$L = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \zeta_i - \sum_{i=1}^n \lambda_i [y_i(w \cdot x_i + b) - 1 + \zeta_i] - \sum_{i=1}^n \mu_i \zeta_i \quad (1.8)$$

trong đó $\lambda_i \geq 0$ và $\mu_i \geq 0$ là các hệ số nhân Lagrange.

Từ đó ta có tập điều kiện Karush Kuhn Tucker như sau:

$$L'_w = w - \sum_{i=1}^n \lambda_i y_i x_i = 0 \quad (1.9)$$

$$L'_b = - \sum_{i=1}^n \lambda_i y_i = 0 \quad (1.10)$$

$$L'_{\zeta_i} = C - \lambda_i - \mu_i = 0 \quad (1.11)$$

$$y_i(w \cdot x_i + b) - 1 + \zeta_i \geq 0 \quad (1.12)$$

$$\lambda_i (y_i(w \cdot x_i + b) - 1 + \zeta_i) = 0 \quad (1.13)$$

$$\mu_i \zeta_i = 0 \quad (1.14)$$

Chú ý rằng ζ_i và các hệ số nhân Lagrange không xuất hiện trong biểu thức đối ngẫu và hàm mục tiêu giống như trong trường hợp SVM biên cứng ở trên, chỉ khác ở điều kiện ràng buộc. w cũng được tính theo (1.9). Tuy nhiên giá trị của b phụ thuộc vào ζ_i , mà ta lại chưa có giá trị của ζ_i .

Như vậy ta có thể sử dụng một mẫu x_k nào đó thỏa mãn điều kiện trên để tính được giá trị của b .

Từ đó ta có kết luận sau:

- Nếu $\lambda_i = 0$ thì $y_i(w \cdot x_i + b) \geq 1$ và $\zeta_i = 0$
- Nếu $0 < \lambda_i < C$ thì $y_i(w \cdot x_i + b) = 1$ và $\zeta_i = 0$
- Nếu $\lambda_i = C$ thì $y_i(w \cdot x_i + b) < 1$ và $\zeta_i > 0$

Kết luận trên cho thấy: các điểm nằm ngoài viên thì có giá trị $\lambda_i = 0$ và chúng chiếm số lượng lớn trong tập huấn luyện. Các điểm nằm trên viên là các vector hỗ trợ thì có λ_i khác 0. Còn lại các điểm bị lỗi chính là các điểm có $\lambda_i = C$.

3.1.4. Phân lớp phi tuyến tính

Trong nhiều trường hợp thì phân lớp phi tuyến có độ chính xác cao hơn. Tuy nhiên phân lớp tuyến tính thì thuật toán đơn giản hơn. Vì thế người ta nghĩ ra cách để phân lớp tuyến tính có thể áp dụng sang cho phân lớp phi tuyến.

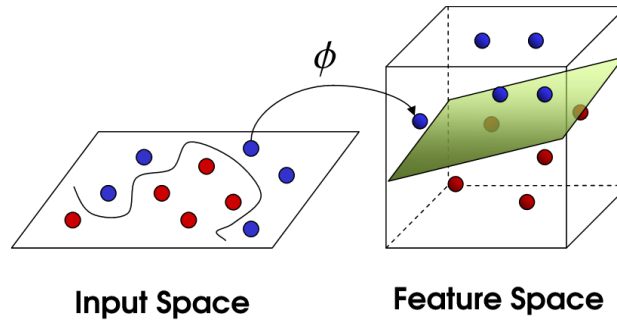
Hướng giải quyết là ánh xạ dữ liệu sang một không gian rộng hơn, để cho chúng trở thành có thể phân loại tuyến tính. Trong không gian này, các điểm dữ liệu trở thành khả tách tuyến tính hoặc có thể được phân tách với ít lỗi hơn so với trường hợp sử dụng không gian ban đầu. Một mặt quyết định tuyến tính trong không gian mới sẽ tương ứng với một mặt quyết định phi tuyến tính trong không gian ban đầu.

$$x \rightarrow \phi(x)$$

Từ đó:

$$f(x) = w \cdot \phi(x) + b$$

Miền ánh xạ được thể hiện như trong hình dưới:



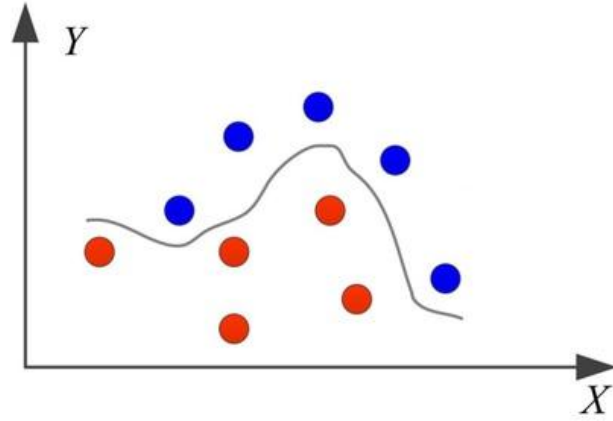
Hình 1.2: Chuyển đổi không gian biểu diễn

Không gian biểu diễn ban đầu được gọi là không gian đầu vào (input space);

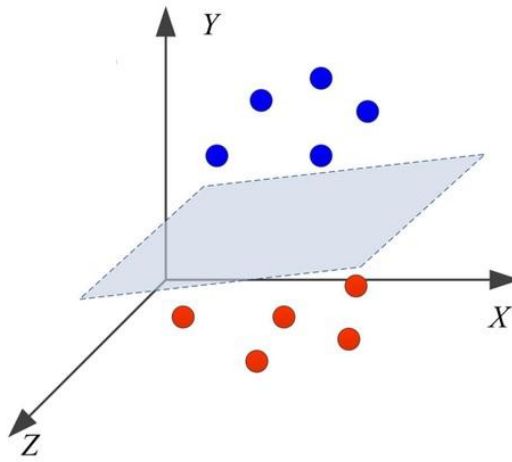
Không gian biểu diễn sau chuyển đổi được gọi là không gian đặc trưng (feature space);

Chiều của không gian đặc trưng liên quan kích thước không gian đầu vào. Thông thường số chiều của không gian đặc trưng lớn hơn nhiều số chiều của không gian ban đầu. Một điểm dữ liệu sẽ được đặc trưng bởi tọa độ (x_i, y_i) . Khi ánh xạ nó vào không gian đặc trưng nhiều chiều hơn, giả sử là không gian 3 chiều, thì phép biến đổi sẽ là:

$$\phi(x_1, x_2) \rightarrow (x_1^2, x_2^2, \sqrt{2}x_1x_2)$$



Hình 1.3: Không gian đầu vào



Hình 1.4: Không gian đặc trưng

Từ đó việc tính toán w và b tương tự như trong trường hợp tuyến tính.

Bài toán cực đại hóa sẽ áp dụng với:

$$L_D(\lambda) = \sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{i,j=1}^n \lambda_i \lambda_j y_i y_j k(x_i, x_j) \quad (1.15)$$

với các ràng buộc:

$$\begin{cases} \sum_{i=1}^n \lambda_i y_i = 0 \\ C \geq \lambda_i \geq 0, \forall i = 1, \dots, n \end{cases}$$

trong đó:

$$k(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \quad (1.16)$$

được gọi là hàm nhân (kernel).

Phương trình của siêu phẳng:

$$f(z) = \sum_{i=1}^n \lambda_i y_i k(x_i, z) + b = 0$$

Việc chuyển đổi không gian trực tiếp có thể gặp vấn đề về số chiều không gian quá lớn. Ngay cả khi không gian ban đầu có số chiều không lớn thì thông qua việc ánh xạ vẫn có thể trả về một không gian mới có số chiều rất lớn. Từ đó chi phí cho việc chuyển đổi cũng là rất lớn. Hàm nhân được sinh ra để giải quyết vấn đề đó, khi chúng ta có thể chuyển đổi không gian không theo một cách trực tiếp, mà theo cách gián tiếp. Theo đó, ta chỉ cần tính được tích vô hướng của hai vector $\Phi(x)$ và $\Phi(z)$ mà không cần biết giá trị của từng vector đó.

Tuy nhiên làm sao để biết một hàm là hàm nhân hay không thì ta phải sử dụng điều kiện Mercer, được định nghĩa như sau:

Tiêu chuẩn đầu tiên để chọn một hàm nhân k là phải tồn tại ϕ để $k(x,y) = \phi(x) \cdot \phi(y)$

Ta có một số hàm nhân cơ bản như sau:

- Hàm nhân đa thức có dạng: $K(x, y) = (x \cdot y + 1)^d$ với d là bậc đa thức
- Hàm bán kính cơ bản Radial Basis Function $K(x, y) = e^{-\|x-y\|^2 / 2\sigma^2}$
- Linear Kernel: $K(x, y) = (x^T y)^d$
- Sigmoid Kernel: $K(x, y) = \tanh(\kappa x^T y + \theta)$ trong đó \tanh là hàm tang hyperbol, dẫn tới mạng nơ ron sigmoid hai lớp (một lớp nơ ron ẩn và một nơ ron đầu ra)

3.2. Thuật toán học máy có giám sát: K-Nearest Neighbor (K-NN)

3.2.1. Tổng quan về K láng giềng gần nhất

Thuật toán K láng giềng gần nhất - K Nearest Neighbors (KNN) được sử dụng rất phổ biến trong lĩnh vực Data Mining. KNN là phương pháp để phân lớp các đối tượng dựa vào khoảng cách gần nhất giữa đối tượng cần phân lớp (Query point) và tất cả các đối tượng trong tập huấn luyện (Training Data).

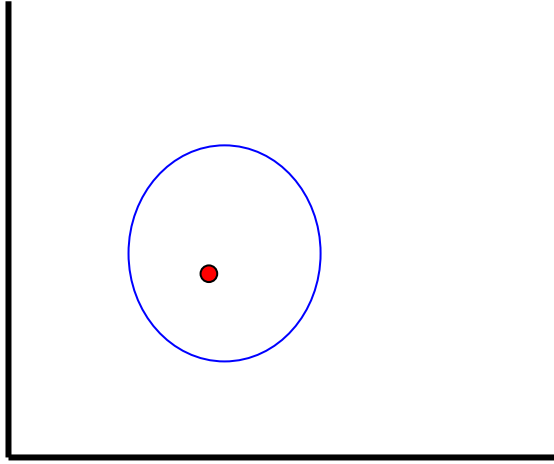
Một đối tượng được phân lớp dựa vào K láng giềng của nó. K là số nguyên dương được xác định trước khi thực hiện thuật toán. Người ta thường dùng khoảng cách Euclidean để tính khoảng cách giữa các đối tượng.

3.2.2. Thuật toán K-NN

1. Xác định giá trị tham số K (số láng giềng gần nhất)
2. Tính khoảng cách giữa đối tượng cần phân lớp (Query Point) với tất cả các đối tượng trong training data (thường sử dụng khoảng cách Euclidean)
3. Sắp xếp khoảng cách theo thứ tự tăng dần và xác định K láng giềng gần nhất với Query Point
4. Lấy tất cả các lớp của K láng giềng gần nhất đã xác định
5. Dựa vào phần lớn lớp của láng giềng gần nhất để xác định lớp cho Query Point

Trong hình dưới, training Data được mô tả bởi dấu (+) và dấu (-), đối tượng cần được xác định lớp cho nó (Query point) là hình tròn đỏ. Nhiệm vụ là ước lượng (hay dự đoán) lớp

của Query point dựa vào việc lựa chọn số láng giềng gần nhất với nó. Nói cách khác là cần biết liệu Query Point sẽ được phân vào lớp (+) hay lớp (-)



Hình 1.5: Hình K láng giềng gần nhất

3.2.3. Hàm tính khoảng cách

Đối với một đối tượng mới cần phân lớp, thuật toán KNN gán phân lớp của đối tượng như một hoặc nhiều đối tượng tương tự nó nhất. Nhưng làm thế nào để định nghĩa được độ tương tự?

Phân tích dữ liệu xác định thước đo khoảng cách để đo độ tương tự. Một thước đo khoảng cách hoặc một hàm khoảng cách d mang giá trị thực, sao cho với bất kỳ tọa độ x , y và z nào thì đảm bảo các tính chất.

1. $d(x, y) \geq 0$, and $d(x, y) = 0$ if and only if $x = y$
2. $d(x, y) = d(y, x)$
3. $d(x, z) \leq d(x, y) + d(y, z)$

Tính chất 1: đảm bảo khoảng cách giữa 2 tọa độ là không âm, và khoảng cách = 0 khi các tọa độ là như nhau.

Tính chất 2: chỉ giao hoán. Khoảng cách từ tọa độ x đến y bằng với khoảng cách từ y đến x

Tính chất 3: là bất đẳng thức tam giác: cho thêm 1 tọa độ thứ 3 thì không bao giờ có thể rút ngắn được khoảng cách giữa 2 tọa độ khác

Hàm khoảng cách phổ biến nhất là hàm khoảng cách Euclidean, đây là cách tính khoảng cách thông thường trong thực tế

$$d_{Euclidean}(x, y) = \sqrt{\sum_i (x_i - y_i)^2} \quad (1.17)$$

Với $x = x_1, x_2, \dots, x_m$ và $y = y_1, y_2, \dots, y_m$. Đại diện cho m giá trị thuộc tính của 2 đối tượng x và y (bản ghi).

Đối với các giá trị liên tục có thể sử dụng chuẩn hóa Min – Max hoặc chuẩn hóa Z-Score

- Chuẩn hóa Min – Max:

$$X^* = \frac{X - \min(X)}{\text{range}(X)} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (1.18)$$

Trong đó:

X : là giá trị thuộc tính của đối tượng

$\min(X)$: là giá trị nhỏ nhất trong miền giá trị của thuộc tính

$\max(X)$: là giá trị lớn nhất trong miền giá trị của thuộc tính

- Chuẩn hóa Z-Score:

$$X^* = \frac{X - \text{mean}(X)}{SD(X)} \quad (1.19)$$

Trong đó:

X : là giá trị thuộc tính của đối tượng

$\text{mean}(X)$ là giá trị trung bình trong miền giá trị của thuộc tính;

$SD(X)$ là độ lệch chuẩn của thuộc tính

Đối với các biến là danh sách thì đo khoảng cách bằng Euclide là không thích hợp, ta có thể định nghĩa một hàm khác, sử dụng để so sánh giá trị thuộc tính thứ i của 2 bản ghi như sau:

$$\text{Different}(x_i, y_i) = \begin{cases} 0 & \text{if } x_i = y_i \\ 1 & \text{otherwise} \end{cases} \quad (1.20)$$

Trong đó: x_i, y_i là thuộc tính thứ i loại danh sách của đối tượng x và y

Sau đó có thể thay thế hàm $\text{Different}(x_i, y_i)$ cho thuộc tính thứ i trong thước đo khoảng cách Euclide nói trên.

3.2.4. Chọn K

Trong thực tế, không có cách chọn nào được cho là giải pháp tốt nhất. Nếu chọn K nhỏ thì khi phân loại hay dự đoán dễ bị ảnh hưởng bởi các giá trị ngoại lai (nhiều). Thuật toán bị tình trạng quá khớp dữ liệu (overfitting) mất đi khả năng khái quát chung của dữ liệu. Còn khi chọn K quá lớn thì các đặc tính riêng biệt (các láng giềng giống nó nhất) từ tập huấn luyện sẽ bị bỏ qua (làm mịn dữ liệu).

3.3. Thuật toán học máy không có giám sát: One-class SVM

3.3.1. Tổng quan về One-class SVM

One-class SVM được đề xuất bởi Schölkopf để ước lượng sự hỗ trợ của một phân bố chiều cao (a high-dimensional distribution). Cho một tập dữ liệu không được đánh nhãn bất cứ thông tin gì, One-class SVM xây dựng một hàm quyết định, theo đó sẽ lấy các giá trị +1

tròn một vùng nhỏ bắt được hầu hết các điểm dữ liệu, và lấy giá trị -1 trong trường hợp còn lại.

3.3.2. Thuật toán One-class SVM

Bài toán One-class SVM: Cho một tập huấn luyện không có bất kỳ thông tin nào về nhãn phân loại như sau:

$$x_i \in R^n, i = 1, 2, \dots, l \quad (1.21)$$

Để phân chia tập dữ liệu từ tập gốc, One-class SVM cần giải phương trình (Schölkopf – 2001) sau:

$$\min_{w, \xi, \rho} \frac{1}{2} w^T w - \rho + \frac{1}{vl} \sum_{i=1}^l \xi_i \quad (1.22)$$

Với ràng buộc:

$$w^T \Phi(x_i) \geq \rho - \xi_i$$

$$\xi_i \geq 0, i = 1, \dots, l$$

Trong đó Φ là hàm ánh xạ, giá trị w và ρ của (1.23) có dạng hàm quyết định tuyến tính như sau:

$$f(x) = \text{sgn}((w^T \phi(x)) - \eta\rho) \quad (1.23)$$

η là một ngưỡng được sử dụng trong hàm quyết định (1.24) để tổng quát hóa các dữ liệu lạ thường. Vì các lý do tính toán, thay vì giải (1.23) trực tiếp, One-class SVM sẽ giải đồng thời 2 bài toán sau:

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha \quad (1.24)$$

Với ràng buộc:

$$0 \leq \alpha_i \leq \frac{1}{vl}, i = 1, \dots, l$$

$$e^T \alpha = 1$$

Trong đó $Q_{ij} = K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$. Lời giải giá trị α của (1.25) sẽ được sử dụng để tính các giá trị w và ρ của (33). Để giúp ngăn chặn số hạng sau dấu chấm trong tính toán trong không gian đặc trưng nhiều chiều, SVM sử dụng hàm nhân $K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$ là hàm Radial Basic Function (RBF), là loại hàm nhân hợp lý nhất hay sử dụng trong SVM.

$$K(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right) \quad (1.25)$$

CHƯƠNG 2. PHÁT HIỆN XÂM NHẬP TRÁI PHÉP VÀ CÁCH TIẾP CẬN BẰNG PHƯƠNG PHÁP HỌC MÁY

1. Khái niệm về xâm nhập trái phép

1.1. Xâm nhập trái phép

Xâm nhập trái phép mạng máy tính là hành vi đột nhập vào mạng (tấn công mạng) để truy cập, thao tác hoặc lạm dụng một số tài sản có giá trị trên mạng.

Phát hiện xâm nhập là tập hợp các kỹ thuật và phương pháp được sử dụng trong quá trình theo dõi các sự kiện bất thường đáng nghi ngờ xảy ra trên một hệ thống máy tính hoặc mạng, từ đó phân tích tìm ra các dấu hiệu sự cố có thể xảy ra.

1.2. Hệ thống IDS

1.2.1. Khái niệm

IDS là từ viết tắt tiếng Anh của Intrusion Detection System hay còn gọi là hệ thống phát hiện các truy nhập trái phép. Theo định nghĩa của Wiki, một hệ thống phát hiện xâm nhập IDS là một thiết bị phần cứng hoặc phần mềm theo dõi hệ thống mạng, có chức năng giám sát lưu thông mạng, tự động theo dõi các sự kiện xảy ra trên một hệ thống mạng máy tính, phân tích để phát hiện ra các vấn đề liên quan đến an ninh, bảo mật và đưa ra cảnh báo. IDS có nhiệm vụ rà quét các gói tin trên mạng, phát hiện các truy nhập trái phép, các dấu hiệu tấn công vào hệ thống từ đó cảnh báo cho người quản trị hay bộ phận điều khiển biết về nguy cơ xảy ra tấn công trước khi nó xảy ra.

1.2.2. Thành phần của hệ thống IDS

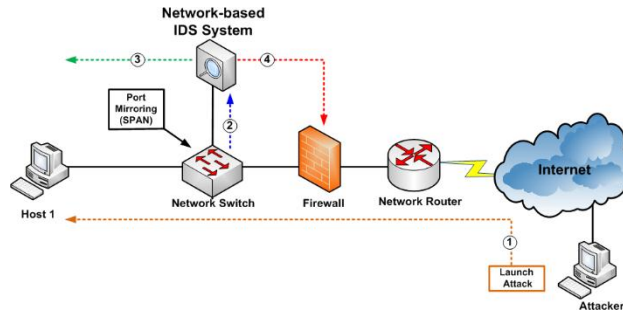
Kiến trúc của hệ thống IDS bao gồm các thành phần chính: Thành phần thu thập gói tin (information collection), thành phần phân tích gói tin (Detection), thành phần phản hồi (response) nếu gói tin đó được phát hiện là một tấn công của tin tặc. Trong 3 thành phần này thì thành phần phân tích gói tin là một thành phần quan trọng nhất và ở thành phần này bộ cảm biến đóng vai trò quyết định nên chúng ta đi sâu vào phân tích bộ cảm biến để hiểu rõ hơn kiến trúc của hệ thống phát hiện xâm nhập là như thế nào.

2. Phân loại phát hiện xâm nhập trái phép theo nguồn dữ liệu

Dựa trên nguồn dữ liệu có hai loại hệ thống phát hiện xâm nhập: Network Based IDS và Host Based IDS.

2.1. Phát hiện xâm nhập trái phép trên mạng (Network-based)

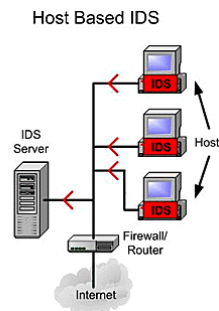
Network-based Intrusion Detection System (Hệ thống phát hiện truy nhập cho mạng) là một giải pháp độc lập để xác định các truy nhập trái phép bằng cách kiểm tra các luồng thông tin trên mạng và giám sát nhiều máy trạm.



Hình 2.1: Hệ thống Network-based IDS (NIDS)

2.2. Phát hiện xâm nhập trái phép trên máy chủ (Host-based)

Trong hệ thống HIDS (Hệ thống phát hiện truy nhập dựa trên máy trạm), các Sensor thường thường là một phần mềm trên máy trạm (Software agent), nó giám sát tất cả các hoạt động của máy trạm mà nó nằm trên đó.



Hình 2.2: Hệ thống Host-based IDS (HIDS)

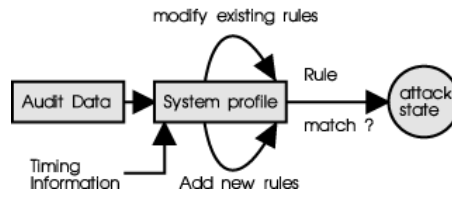
3. Phân loại phát hiện xâm nhập trái phép theo kỹ thuật phân tích dữ liệu

Phát hiện xâm nhập trái phép dựa vào kỹ thuật phân tích dữ liệu có hai phương pháp chính là phát hiện dựa trên sự lạm dụng (Misuse Detection) và phát hiện sự không bình thường (Anomaly Detection).

3.1. Misuse Detection

Phương pháp này phân tích các hoạt động của hệ thống, tìm kiếm các sự kiện giống với các mẫu tấn công đã biết trước. Thông thường hệ thống sẽ lưu trữ trong cơ sở dữ liệu những gói tin có liên quan đến kiểu tấn công từ trước dưới dạng so sánh được, trong quá trình xử lý sự kiện sẽ được so sánh Với các thông tin trong cơ sở dữ liệu nếu giống hệ thống sẽ đưa ra cảnh báo hoặc ngăn chặn. Các mẫu tấn công biết trước này gọi là các dấu hiệu tấn công. Do vậy phương pháp này còn được gọi là phương pháp dò dấu hiệu (Signature Detection).

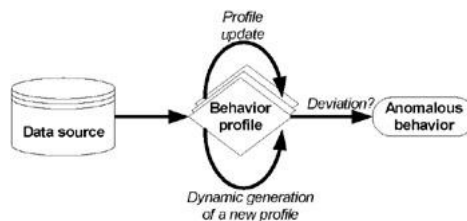
A typical misuse detection system

**Hình 2.3: Hệ thống Misuse Detection**

Kiểu phát hiện tấn công bằng dấu hiệu có ưu điểm là phát hiện các cuộc tấn công nhanh và chính xác, không đưa ra các cảnh báo Sai lầm giảm khả năng hoạt động của mạng và giúp người quản trị xác định các lỗ hổng bảo mật trong hệ thống của mình. Tuy nhiên, phương pháp này có nhược điểm là không phát hiện được các cuộc tấn công không có trong mẫu, các kiểu tấn công mới.

3.2. Anomaly Detection

Đây là kỹ thuật dò thông minh bằng cách nhận dạng các hành động không bình thường của mạng. Quan niệm của phương pháp này về các cuộc tấn công khác so với các hoạt động thông thường. Ban đầu, chúng lưu trữ các mô tả sơ lược về các hoạt động bình thường của hệ thống. Các cuộc tấn công sẽ có những hành động khác so với trạng thái bình thường do đó có thể nhận dạng được chúng.

**Hình 2.4: Hệ thống Anomaly Detection**

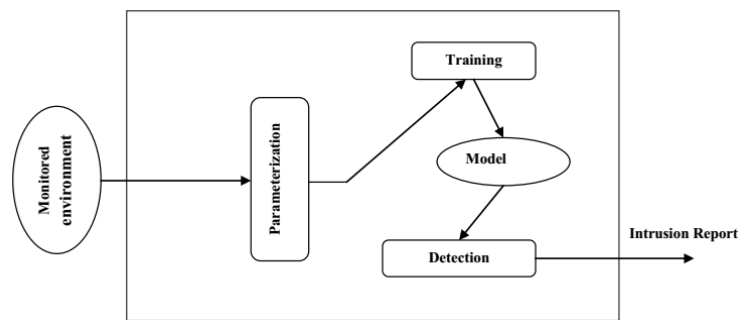
Phương pháp dò sự không bình thường của hệ thống rất hữu hiệu trong việc phát hiện các cuộc tấn công kiểu từ chối dịch vụ. Ưu điểm của phương pháp này là có thể phát hiện ra các kiểu tấn công mới, cung cấp các thông tin hữu ích bổ sung cho phương pháp dò sự lạm dụng, tuy nhiên chúng có nhược điểm là thường tạo ra một số lượng tương đối lớn các cảnh báo sai lầm giảm hiệu suất hoạt động của mạng. Tuy nhiên phương pháp này sẽ là hướng được nghiên cứu nhiều hơn, hoàn thiện các nhược điểm, đưa ra ít cảnh báo sai để hệ thống chạy chuẩn xác hơn.

4. Phát hiện xâm nhập trái phép tiếp cận bằng phương pháp học máy (Machine Learning Approach)

Hệ thống phát hiện xâm nhập dựa trên bất thường có 2 ưu điểm chính so với hệ thống dựa trên dấu hiệu. Thứ nhất, nó có khả năng phát hiện các cuộc tấn công chưa nhận diện trước bởi vì nó có thể mô hình hóa hoạt động bình thường của hệ thống và từ đó phát hiện ra độ lệch của các hành vi bất thường so với mô hình đó. Ưu điểm thứ 2 là nó có khả năng tùy biến các hồ sơ hành vi bình thường cho tất cả các hệ thống, các ứng dụng và các mạng. Do

đó, nó giúp làm tăng mức độ phức tạp đối với kẻ tấn công khi cố gắng thực hiện dò tìm cách tấn công mà không dễ bị phát hiện. Hệ thống phát hiện xâm nhập dựa trên bất thường hiện nay đang trở thành xu hướng phát triển của các hệ thống phát hiện xâm nhập; và để góp phần cải thiện hiệu quả của hệ thống hiện nay có rất nhiều kỹ thuật được áp dụng trong đó có kỹ thuật học máy.

Kiến trúc của một hệ thống phát hiện xâm nhập dựa trên bất thường được minh họa như sau: gồm các khối cơ bản là khối tham số hóa; khối huấn luyện và khối phát hiện. Khối tham số hóa bao gồm công việc thu thập các dữ liệu nguyên gốc từ một hệ thống giám sát mạng. Khối huấn luyện thực hiện tìm kiếm mô hình cho hệ thống sử dụng các phương pháp thủ công hoặc tự động. Khối phát hiện thực hiện so sánh hệ thống được sinh ra từ khối huấn luyện với phần dữ liệu được tham số hóa đưa vào. Một tiêu chí ngưỡng được lựa chọn để xác định dữ liệu đưa vào là bất thường hay không. Các phương pháp học máy có thể giúp xây dựng các mô hình một cách tự động dựa trên dữ liệu huấn luyện. Hiện có một số hướng áp dụng học máy trong phát hiện xâm nhập trái phép dựa trên bất thường, đó là: phương pháp học có giám sát và phương pháp học không có giám sát.



Hình 2.5: Kiến trúc của hệ thống phát hiện xâm nhập dựa trên bất thường

4.1. Phát hiện xâm nhập trái phép dựa vào học máy có giám sát

Phương pháp học có giám sát (hay còn gọi là phương pháp phân loại) yêu cầu một tập dữ liệu huấn luyện được đánh nhãn chứa cả dữ liệu thông thường và dữ liệu bất thường để từ đó xây dựng một mô hình dự đoán. Về mặt lý thuyết, phương pháp học có giám sát có hiệu năng phát hiện tốt hơn phương pháp học không giám sát. Tuy nhiên, phương pháp này tồn tại một số vấn đề ảnh hưởng tới tính chính xác. Đầu tiên, đó là kích thước của tập dữ liệu huấn luyện, nó quá bé để bao trùm hết tất cả các trường hợp. Ngoài ra, việc đánh nhãn một cách chính xác cũng là một thử thách và tập dữ liệu huấn luyện cũng tồn tại nhiều nên gây ra tình trạng cảnh báo sai với tần suất cao. Một số phương pháp học có giám sát thường được sử dụng như: Mạng Neural có giám sát, SVM, K-NN, mạng Bayes và Cây quyết định.

4.2. Phát hiện xâm nhập trái phép dựa vào học máy không có giám sát

Phương pháp học không có giám sát không yêu cầu cần phải có một bộ dữ liệu huấn luyện. Phương pháp này dựa trên 2 giả thiết. Thứ nhất, nó giả thiết phần lớn kết nối mạng là

các luồng dữ liệu bình thường và chỉ có một số ít luồng dữ liệu là bất thường. Thứ hai, nó biết trước các luồng dữ liệu bất thường là khác nhau về mặt thống kê so với các luồng dữ liệu bình thường. Theo 2 giải thiết trên, các nhóm dữ liệu của các trường hợp xuất hiện với tần suất thường xuyên được coi như là các luồng dữ liệu bình thường, trong khi đó các trường hợp xuất hiện với tần suất không bình thường khác với đa số các trường hợp khác thì được coi là bất thường. Một số phương pháp học không có giám sát là: K-means, C-Means, One-class SVM, kỹ thuật Clustering.

5. Mô tả bài toán đề xuất trong luận văn

Trong phạm vi luận văn này, tôi đề xuất sử dụng phương pháp học máy có giám sát (K-NN và SVM) và phương pháp học máy không giám sát (One-class SVM) để thực hiện kiểm thử việc phát hiện xâm nhập trái phép trong mạng dựa trên bất thường đối với dữ liệu mạng dạng Netflow.

5.1. Lựa chọn luồng dữ liệu Net-Flow

Netflow là một giao thức do hãng Cisco phát triển vào những năm 1996, được phát triển thành một công nghệ giám sát lưu lượng mạng. Netflow cho phép thực hiện giám sát, phân tích, tính toán lưu lượng gói. Một trong các ưu điểm của Netflow so với các giao thức khác là nó cho phép định danh và phân loại những loại tấn công như DoS, DDoS, Virus, Worm, ... theo thời gian thực dựa vào những sự hành vi thay đổi bất thường trong mạng, đặc biệt trong mạng có lưu lượng lớn.

Trên thế giới hiện nay tồn tại một số bộ dữ liệu nổi tiếng như DARPA, KDD-99, ISCX,... Tuy vậy, các bộ dữ liệu này tồn tại ở dạng tcpdump, không phải ở dạng Netflow nên không ứng dụng được trong nghiên cứu về IDS trên Netflow. Các bộ dữ liệu ở dạng Netflow rất ít, nếu có thì hoặc không đầy đủ hoặc chưa hoàn chỉnh. Để có bộ dữ liệu kiểm thử trong luận văn, ở phần sau, tôi sẽ trình bày phương pháp xây dựng bộ dữ liệu dạng Netflow.

5.2. Phát hiện bất thường bằng K-NN, SVM và One-class SVM

Với các thuật toán học máy KNN, SVM, One-class SVM đã tìm hiểu ở chương I, tôi đề xuất sử dụng các phương pháp thử nghiệm như sau:

- Thuật toán K-NN: lựa chọn $K = 1, 4$;
- Thuật toán SVM:
 - Hàm nhân tuyến tính (Linear kernel)

Hàm linear kernel có dạng:

$$K_{linear}(x_1, x_2) = x_1^T x_2 + c, \text{ chọn } c = 0 \quad (2.1)$$

- Hàm nhân đa thức (Polynomial kernel)

Hàm polynomial kernel có dạng:

$$K_{poly}(x_1, x_2) = (\alpha x_1^T x_2 + c)^d, \text{ chọn } \alpha = 0.25; c = 0; d = 3 \quad (2.2)$$

- Hàm nhân RBF (RBF kernel)

Hàm RBF kernel có dạng:

$$K_{RBF}(x_1, x_2) = e^{-\gamma \|x_1 - x_2\|^2}, \text{ chọn } \gamma = 0.25 \quad (2.3)$$

- Hàm nhân đường xích-ma (sigmoid kernel)

Hàm sigmoid kernel có dạng:

$$K_{sigmoid}(x_1, x_2) = \tanh(\alpha x_1^T x_2 + c) \quad (2.4)$$

$$\text{chọn } \alpha = 0.25; c = 0$$

- Thuật toán One-class SVM: sử dụng hàm nhân RBF

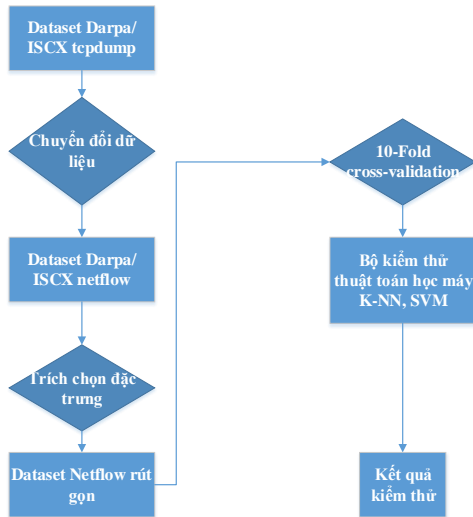
Hàm RBF kernel có dạng:

$$K_{RBF}(x_1, x_2) = e^{-\gamma \|x_1 - x_2\|^2}, \text{ chọn } \gamma = 0.25 \quad (2.5)$$

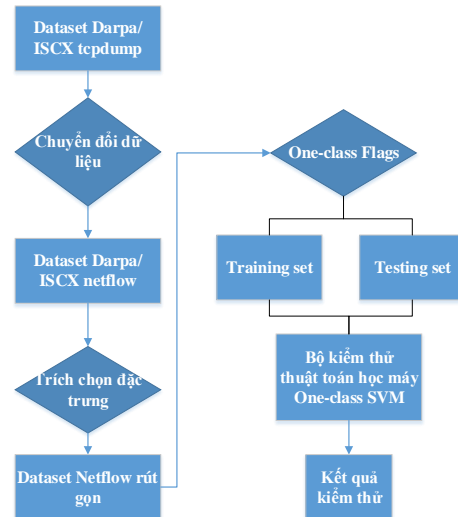
CHƯƠNG 3. ỨNG DỤNG PHƯƠNG PHÁP HỌC MÁY KNN, SVM VÀ ONE-CLASS SVM ĐỂ PHÁT HIỆN BẤT THƯỜNG

1. Mô hình thử nghiệm

Mô hình thử nghiệm gồm có:



Hình 3.1: Mô hình kiểm thử với thuật toán học máy có giám sát



Hình 3.2: Mô hình kiểm thử với thuật toán học máy không có giám sát

1.1. Giới thiệu bộ dữ liệu thử nghiệm của DARPA và ISCX Research Center

1.1.1. Bộ dữ liệu DARPA Tcpdump

Bộ dữ liệu DARPA hình thành do Cục dự án nghiên cứu cao cấp Bộ quốc phòng Mỹ (Defense Advanced Research Project Agency) tài trợ để xây dựng cơ sở dữ liệu mã xâm nhập trái phép tại Phòng thí nghiệm Lincoln, Đại học MIT. Để xây dựng tập dữ liệu này, các nhà khoa học đã lấy dữ liệu của một mạng quân sự Mỹ khi hoạt động bình thường làm dữ liệu bình thường; sau đó đưa thêm các dữ liệu xâm nhập trái phép vào trong tập dữ liệu đó. Cách làm trên cho phép biết được chắc chắn đâu là dữ liệu bình thường, đâu là dữ liệu xâm nhập trái phép.

Toàn bộ dữ liệu có kích thước khoảng 10Gb, trong đó gồm 54 loại xâm nhập được phân làm 4 nhóm: R2L (Remote to Local – là nhóm các xâm nhập cho phép kẻ tấn công từ xa lấy được quyền của người dung máy chủ), U2R (User to Root – là nhóm các xâm nhập cho phép người dùng bình thường trên máy chủ có thể đoạt quyền quản trị root), DoS (Denial of Service – là nhóm tấn công từ chối dịch vụ, phá hoạt tính sẵn sàng của hệ thống), Probe (là nhóm tấn công do thám, ảnh hưởng đến tính bảo mật của hệ thống, đồng thời cung cấp các thông tin cần thiết để tiến hành các bước tấn công tiếp theo).

1.1.2. Bộ dữ liệu ISCX Tcpdump

Information Security Centre of Excellence (ISCX) là một trung tâm nghiên cứu về an toàn thông tin của trường đại học New Brunswick (UNB) – Canada. ISCX đã xây dựng một mô hình mạng, mô phỏng các cuộc tấn công trong mạng dựa trên các giao thức HTTP, SMTP, SSH, IMAP, POP3 và FTP. Những luồng dữ liệu thông thường và bất thường được bắt giữ và được đánh dấu. Bộ dữ liệu UNB ISCX 2012 IDS bao gồm dữ liệu thu thập trong vòng 7 ngày, bao gồm cả dữ liệu thông thường và bất thường. Bộ dữ liệu ISCX cũng ở dạng Tcpdump.

1.2. Chuyển đổi dữ liệu từ Tcpdump sang Netflow

Để phục vụ cho đề tài luận văn này, tôi đã thực hiện xây dựng Bộ dữ liệu dạng Netflow trên cơ sở bộ dữ liệu Tcpdump để dùng trong IDS (Kết quả việc chuyển đổi dữ liệu từ dạng Tcpdump sang Netflow đã được tôi trình bày chi tiết trong bài báo khoa học “Bộ dữ liệu dạng Netflow dùng trong phát hiện xâm nhập trái phép và ứng dụng” được đăng trên Tạp chí Khoa học công nghệ thông tin và truyền thông, số 1, tháng 6 năm 2016 của Học viện Công nghệ Bưu chính viễn thông).

Đầu ra của việc chuyển đổi dữ liệu là 02 bộ dữ liệu dạng Netflow là DARPA Netflow và ISCX Netflow

2. Cài đặt thử nghiệm

2.1. Tập dữ liệu thử nghiệm DARPA Netflow

Như đã đề cập mục chuyển đổi dữ liệu nêu trên, sau khi chuyển đổi bộ dữ liệu DARPA Tcpdump, tôi đã thu thập và phân tách được 4 bộ dữ liệu netflow tương ứng với 4 máy chủ victim. Trong phạm vi luận văn này, tôi dự kiến sử dụng tập dữ liệu thử nghiệm là tập dữ liệu netflow của máy chủ Pascal (172.16.112.50), được trình bày chi tiết như sau:

Bảng 3.1: Các thông số cơ bản của bộ dữ liệu netflow máy chủ Pascal

Mô tả	Giá trị
Số lượng flow đến máy chủ Pascal	170.153
Số lượng flow tấn công vào máy chủ Pascal	29.416
Số lượng flow bình thường vào máy chủ Pascal	140.737

2.2. Tập dữ liệu thử nghiệm ISCX Netflow

Như đã đề cập mục chuyển đổi dữ liệu nêu trên, sau khi chuyển đổi bộ dữ liệu ISCX Tcpdump, tôi đã thu thập được các bộ dữ liệu netflow tương ứng với từng ngày thu thập của ISCX. Trong phạm vi luận văn này, tôi dự kiến sử dụng tập dữ liệu thử nghiệm là tập dữ liệu netflow của ngày 14/6/2010 có ghi nhận xâm nhập trái phép loại HTTP Denial of Service, được trình bày chi tiết như sau:

Bảng 3.2: Các thông số cơ bản của bộ dữ liệu ISCX Netflow

Mô tả	Giá trị
Tổng số lượng flow	18.682
Số lượng flow tấn công	1.000
Số lượng flow bình thường	17.682

2.3. Trích chọn đặc trưng

Bộ dữ liệu netflow gồm rất nhiều trường dữ liệu khác nhau. Tuy nhiên, để ứng dụng trong phát hiện xâm nhập trái phép, tôi lựa chọn sử dụng các đặc trưng như sau:

Bảng 3.3: Các đặc trưng lựa chọn trong phát hiện xâm nhập trái phép

Tên của đặc trưng	Mô tả
Protocol	Giao thức của kết nối
Packets	Số lượng gói tin (packet) trong một flow
Octets	Số lượng bytes trong một flow
Flags	Số dạng hexa biểu thị cờ của flow

Các đặc trưng được trích chọn nêu trên đều ở dạng số (numeric) nên rất thuận lợi cho việc thử nghiệm phát hiện xâm nhập trái phép bằng phương pháp học máy, mô phỏng trên phần mềm Weka. Đặc trưng Flags được gán nhãn là Normal hoặc Abnormal biểu thị Flow là bình thường hoặc bất thường được sử dụng làm nhãn phân loại.

Thông tin về bộ dữ liệu DARPA Netflow được dùng để thử nghiệm như sau: gồm 38.278 flows (dòng dữ liệu)

Bảng 3.4: Thông tin bộ dữ liệu thử nghiệm DARPA Netflow máy chủ Pascal

Thuộc tính	Giá trị
Protocol	1, 6, 17
Octets	46 – 8.279.218
Packets	1 – 179.983
Flags	Normal (9.856), Abnormal (29.422)

Thông tin về bộ dữ liệu ISCX Netflow được dùng để thử nghiệm như sau: gồm 18.682 flows (dòng dữ liệu)

Bảng 3.5: Thông tin bộ dữ liệu thử nghiệm ISCX Netflow ngày 14/6

Thuộc tính	Giá trị
Protocol	6
Octets	46 – 1.045.728
Packets	1 – 22.528
Flags	Normal (17.682), Abnormal (1.000)

2.4. Cài đặt

Trước khi thực hiện thử nghiệm với các thuật toán học máy K-NN, SVM và One-class SVM, dữ liệu cần thực hiện chuẩn hóa để nâng cao tính chính xác cho các thuật toán học máy. Đối với các thuật toán học máy có giám sát (K-NN và SVM), dữ liệu được chuẩn hóa sử dụng bộ lọc Discretize của Weka. Đối với thuật toán học máy không có giám sát (One-class SVM), dữ liệu được chuẩn hóa sử dụng bộ lọc Normalize của Weka.

- Đối với cài đặt thử nghiệm các thuật toán học máy có giám sát, cách thức thử nghiệm và đánh giá như sau:

Sử dụng phương pháp đánh giá 10-fold cross-validation của Weka đối với bộ dữ liệu dataset đầy đủ. Với phương pháp này, bộ dữ liệu dataset đầy đủ sẽ được chia một cách ngẫu nhiên thành 10 tập con. Với bộ 10 tập con, 1 tập con sẽ được sử dụng cho mục đích kiểm thử, 9 tập con khác được sử dụng cho mục đích dữ liệu huấn luyện. Phương pháp 10-fold cross-validation của Weka sẽ thực hiện lặp đi lặp lại 10 lần với tập dữ liệu, mỗi lần với một tập con làm tập kiểm thử. Kết quả của 10 lần thực hiện sẽ được tính giá trị trung bình để xác định hiệu năng tổng thể của từng giải thuật học máy.

- Đối với cài đặt thử nghiệm các thuật toán học máy không có giám sát, cách thức thử nghiệm và đánh giá như sau:

Thuật toán được sử dụng để thử nghiệm là One-class SVM. Do vậy nhãn phân loại của bộ dữ liệu chỉ là 01 nhãn duy nhất. Bộ dữ liệu được sử dụng trong quá trình Training là bộ dữ liệu Dataset gồm các Flow bất thường. Bộ dữ liệu được sử dụng trong quá trình kiểm thử Testing là bộ dữ liệu Dataset đầy đủ được gán nhãn phân loại toàn bộ là bất thường.

3. Kết quả và đánh giá

3.1. Tiêu chí đánh giá hệ thống IDS

Các thông số đánh giá độ chính xác của mô hình bao gồm:

- Độ chính xác (Accuracy – AC) là số lượng mẫu dự đoán đúng:
- True Positive (TP): tỉ lệ phát hiện đúng các mẫu tấn công (các nhãn positive được gán nhãn positive)
- False Positive (FP): tỉ lệ phát hiện sai các mẫu không phải là tấn công nhưng lại gán nhãn là tấn công (các nhãn negative nhưng bị gán nhãn nhầm là positive)
- True Negative (TN): tỉ lệ phát hiện đúng các mẫu không phải là tấn công (các nhãn negative được gán nhãn đúng là negative)
- False Negative (FN): tỉ lệ phát hiện sai các mẫu là tấn công nhưng lại gán nhãn là bình thường (các nhãn positive nhưng bị gán nhãn nhầm là negative)
- Precision (P):

- Đánh giá các hệ thống IPS người ta chủ yếu dựa vào 2 thông số False Positive (cảnh báo sai), True Positive (cảnh báo đúng), dùng tỷ lệ của 2 yếu tố này chúng ta có thể xây dựng nên đường cong ROC (Receiver Operating Characteristic Curve).

3.2. Kết quả thử nghiệm

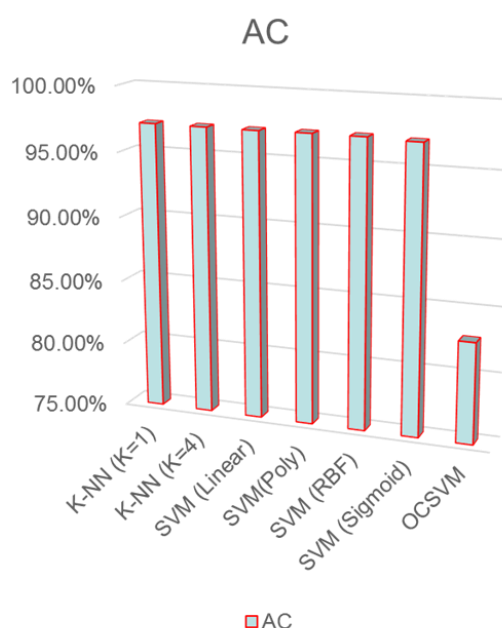
Chi tiết xem tại Mục 3, Mục 4 - Phụ lục 01 của Bản luận văn.

3.3. Đánh giá

Dựa trên kết quả thực nghiệm, ta thu được kết quả như sau:

- Đối với bộ dữ liệu DARPA Netflow

	Thuật toán có giám sát						Thuật toán không giám sát
	K-NN K=1	K-NN K=4	SVM Linear	SVM Poly	SVM RBF	SVM Sigmoid	SVM One-class
AC	97.1664 %	97.146 %	97.1282 %	97.1714 %	97.1664 %	97.0391 %	82.8072 %
TP	0.997	0.997	0.997	0.997	0.997	0.997	
FP	0.104	0.105	0.105	0.104	0.104	0.108	
P	0.966	0.966	0.966	0.966	0.966	0.965	

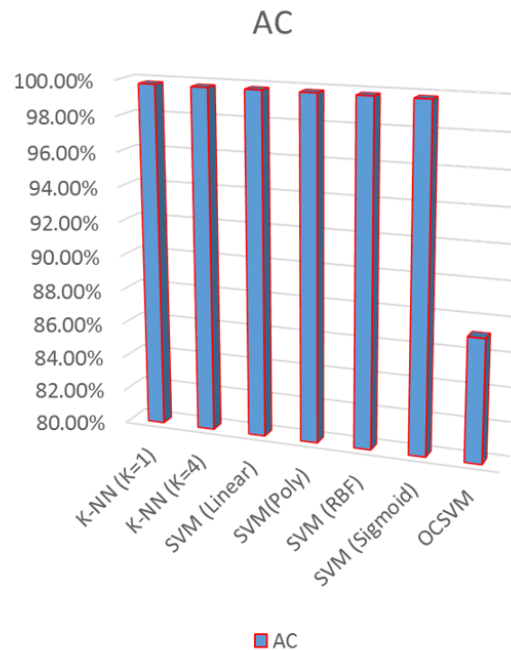


Hình 3.3: So sánh độ chính xác (AC) của các thuật toán học máy với bộ dữ liệu DARPA Netflow

- Đối với bộ dữ liệu ISCX Netflow

	Thuật toán có giám sát						Thuật toán không giám sát
	K-NN K=1	K-NN K=4	SVM Linear	SVM Poly	SVM RBF	SVM Sigmoid	SVM One-class
AC	99.7752 %	99.7538 %	99.7752 %	99.7805 %	99.7805 %	99.7805 %	87.1641 %

TP	0.999	1	0.999	0.999	0.999	0.999	
FP	0.002	0.003	0.002	0.002	0.002	0.002	
P	0.961	0.956	0.961	0.962	0.962	0.962	



Hình 3.4: So sánh độ chính xác (AC) của các thuật toán học máy với bộ dữ liệu ISCX Netflow

Nhận xét:

- Qua thử nghiệm với hai bộ dữ liệu DARPA Netflow và ISCX Netflow, các thuật toán học máy có giám sát có độ chính xác cao hơn thuật toán học máy không có giám sát;
- Đối với các thuật toán học máy có giám sát, do bộ dữ liệu huấn luyện và kiểm thử đã được đánh nhãn đầy đủ độ chính xác khi phân loại cao;
- Đối với thuật toán học không giám sát (One-class SVM): do bộ dữ liệu chỉ đánh một nhãn phân loại duy nhất nên độ chính xác khi phân loại là không cao;
- Các đặc trưng được sử dụng để thử nghiệm đối với mỗi flow là ít (4 đặc trưng). Điều này cũng góp phần làm tăng độ chính xác của các thuật toán trong quá trình kiểm thử.

KẾT LUẬN

1. Kết quả đạt được.

Luận văn đã trình bày được cách thức ứng dụng các phương pháp học máy để phát hiện xâm nhập trái phép trong mạng dựa trên bất thường. Dựa trên cơ sở lý thuyết của các thuật toán học máy SVM, K-NN, One-class SVM cùng các lý thuyết về phát hiện xâm nhập trái phép dựa trên bất thường, luận văn đã thực hiện đề xuất mô hình thử nghiệm một số thuật toán học máy để phát hiện bất thường trên bộ dữ liệu dạng Netflow. Nội dung chi tiết các phần mà luận văn đã thực hiện được:

- Nghiên cứu một cách tổng quan về phương pháp học máy; nghiên cứu chi tiết một số thuật toán học máy như SVM, K-NN, One-class SVM;
- Nghiên cứu về hệ thống phát hiện xâm nhập trái phép, một số hệ thống phát hiện xâm nhập trái phép, đặc biệt là hệ thống phát hiện xâm nhập trái phép dựa trên bất thường. Từ đó, đề xuất cách thức phát hiện xâm nhập trái phép dựa trên bất thường bằng phương pháp tiếp cận học máy;
- Đã xây dựng được một bộ dữ liệu thử nghiệm dạng Netflow bằng cách chuyển đổi từ các bộ dữ liệu nổi tiếng DARPA, ISCX ở dạng Tcpdump; trên cơ sở bộ dữ liệu thử nghiệm dạng Netflow đó đã thực hiện thử nghiệm phát hiện bất thường bằng công cụ WEKA sử dụng các thuật toán học máy SVM, K-NN, One-class SVM.

2. Hướng nghiên cứu trong tương lai.

Luận văn mới dừng lại ở việc thử nghiệm các thuật toán học máy trên các bộ dữ liệu thử nghiệm Netflow. Trong tương lai, tôi đề xuất hướng nghiên cứu tiếp theo:

- Bộ dữ liệu dạng Netflow là bộ dữ liệu mới, có nhiều ưu điểm trong việc nghiên cứu về IDS. Do đó, trong tương lai tôi sẽ tiếp tục nghiên cứu hoàn thiện các bộ dữ liệu dạng Netflow;
- Nghiên cứu, cải tiến các phương pháp học máy nhằm nâng cao hiệu quả phát hiện xâm nhập trái phép dựa trên bất thường trên các dữ liệu dạng Netflow;
- Tích hợp vào hệ thống bảo mật của một cơ quan, tổ chức để giúp tăng cường an toàn thông tin cho hệ thống mạng.