PHÁT HIỆN XÂM NHẬP MẠNG SỬ DỤNG KỸ THUẬT HỌC MÁY

Vũ Văn Cảnh*, Hoàng Tuấn Hảo, Nguyễn Văn Quân

Tóm tắt: *Cùng với sự phát triển của mạng máy tính, vấn đề an ninh mạng đang đối mặt với những thách thức lớn, các hệ thống mạng đang trở thành các mục tiêu tấn công phá hoại, xâm nhập trái phép và đánh cắp thông tin của các Hacker. Hầu hết các kỹ thuật phát hiện xâm nhập truyền thống có tỷ lệ phát hiện chính xác thấp và tỷ lệ phát hiện nhầm cao. Các nghiên cứu dựa trên kỹ thuật học máy trong phát hiện xâm nhập đã cho thấy hiệu quả trong việc phát hiện các tấn công mới với tỷ lệ phát hiện cao, tỷ lệ phát hiện nhầm thấp với chi phí tính toán hợp lý. Trong bài báo này, chúng tôi nghiên cứu một số kỹ thuật học máy trong phát hiện xâm nhập mạng. Các thí nghiệm đã được tiến hành trên bộ dữ liệu KDD99 tại phòng thí nghiệm An ninh mạng - Học viện Kỹ thuật quân sự.*

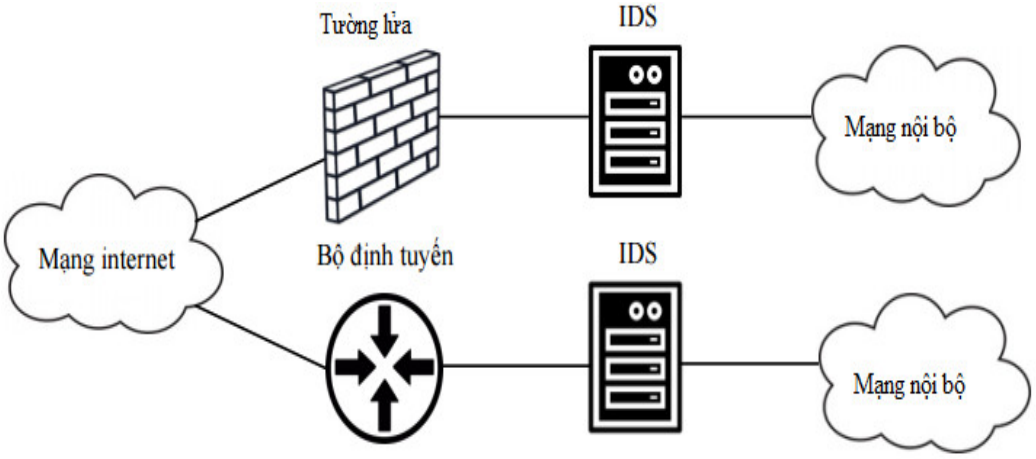
Từ khóa: Học máy, Xâm nhập mạng, Phát hiện xâm nhập, Phân cụm.

1. GIỚI THIỆU

Trong cuộc sống hiện đại, Internet là một trong những yếu tố quan trọng thúc đẩy sự phát triển của các cơ quan, tổ chức. Tuy nhiên, có khá nhiều rủi ro khi sử dụng Internet xuất phát từ các cuộc tấn công mạng. Vì vậy, các hệ thống phát hiện xâm nhập (Intrusion Detection System - IDS) khác nhau đã được thiết kế và xây dựng nhằm ngăn chặn các cuộc tấn công này. Mục tiêu của IDS là cung cấp một hàng rào bảo vệ, giúp các hệ thống mạng có khả năng phát hiện các cuộc tấn công từ bên ngoài. Việc phát hiện xâm nhập dựa trên giả thiết là hành vi của kẻ xâm nhập khác với người sử dụng hợp lệ [12]. Hình 1 dưới đây mô tả các vị trí điển hình của IDS trong hệ thống giám sát an ninh mạng. Trong đó, các dữ liệu vào ra giữa Internet và mạng nội bộ được các IDS bắt, xử lý và phân lớp để xác định đó là một truy cập bình thường hoặc một cuộc tấn công; Từ đó, có các cảnh báo, hành động phù hợp.

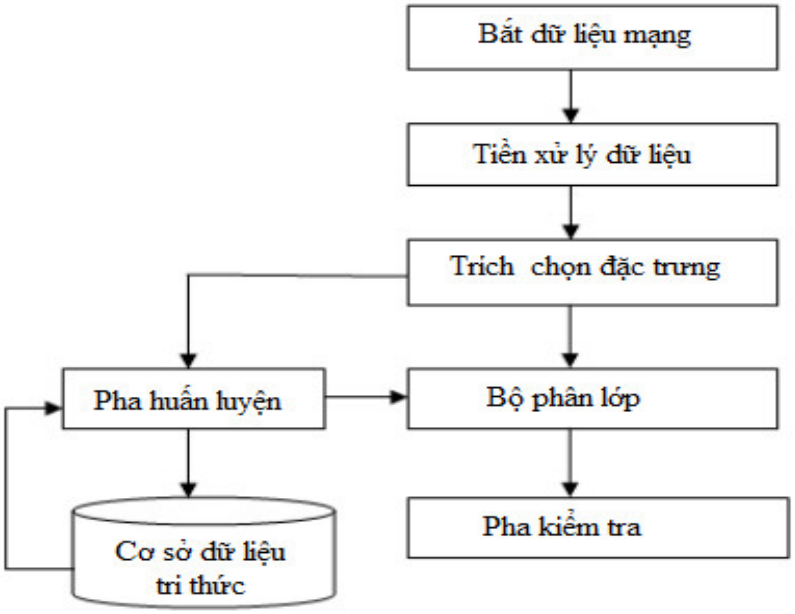
IDS được chia thành hai loại: IDS dựa trên dấu hiệu (misuse-based) và IDS dựa trên sự bất thường (anomaly-based) [2]. Việc phân lớp căn cứ vào cách tiếp cận phát hiện xâm nhập. IDS dựa trên dấu hiệu sử dụng mẫu của các cuộc tấn công đã biết hoặc điểm yếu của hệ thống để xác định xâm nhập, tương tự như các phần mềm chống virus sử dụng mẫu để phát hiện virus. Yếu điểm của kỹ thuật này là không thể phát hiện các mẫu tấn công mới, nên nó cần phải cập nhật liên tục các dấu hiệu tấn công để nhận dạng các cuộc tấn công mới.

IDS dựa trên sự bất thường cố gắng xác định độ lệch so với các mẫu sử dụng



***Hình 1**. Vị trí của IDS trong hệ thống giám sát an ninh mạng.*

thông thường đã được thiết lập trước để đánh dấu các xâm nhập. Vì vậy, các IDS dựa trên sự bất thường cần quen với các mẫu sử dụng thông thường thông qua việc học. Các kỹ thuật học máy khác nhau đã được sử dụng rộng rãi để phục vụ cho mục đích này. Hình 2 mô tả kiến trúc của một IDS sử dụng kỹ thuật học máy [7]. Trong đó, dữ liệu bắt được sau khi qua các công đoạn tiền xử lý, chọn lựa thuộc tính sẽ được phân lớp bởi các bộ phân lớp (classifier) đã được huấn luyện. Việc huấn luyện các bộ phân lớp được thực hiện qua pha huấn luyện và kiểm tra với tập dữ liệu huấn luyện đã lưu trữ.



***Hình 2**. Kiến trúc của một IDS.*

Bài báo được viết với cấu trúc như sau: sau phần 1 giới thiệu, phần 2 trình bày kiến thức nền tảng về tấn công đột nhập mạng, các kỹ thuật xâm nhập và kỹ thuật học máy. Một số kỹ thuật học máy ứng dụng trong phát hiện tấn công xâm nhập sẽ được trình bày trong phần 3. Phần 4 trình bày các thử nghiệm và kết quả đối với các kỹ thuật học máy đề xuất.

2. KIẾN THỨC NỀN TẢNG

2.1. Tấn công đột nhập mạng

Tấn công, đột nhập mạng là hành vi tấn công xâm nhập trái phép nhằm lạm dụng các tài nguyên trên mạng, việc lạm dụng có thể dẫn đến hậu quả có thể khiến cho tài nguyên mạng trở nên không đáng tin cậy hoặc không sử dụng được. Hầu hết các cuộc tấn công xâm nhập mạng máy tính vượt qua các lớp bảo mật của hệ thống theo những phương thức cụ thể nhằm phá vỡ các thuộc tính bảo mật của thông tin và hệ thống. Ví dụ một số cuộc tấn công nhằm đọc, đánh cắp các thông tin nhưng không thay đổi thành phần nào trong hệ thống. Một số cuộc tấn công lại tắt hoặc làm ngừng hoạt động thành phần nào đó trong hệ thống. Hoặc những cuộc tấn công khác lại có khả năng chiếm toàn quyền điều khiển hoặc phá huỷ hệ thống. Chung quy lại, chúng thường gây nên tổn thương đến các thuộc tính bảo mật thông tin và hệ thống: tính bí mật, tính toàn vẹn và tính khả dụng.

2.2. Các kỹ thuật phát hiện xâm nhập

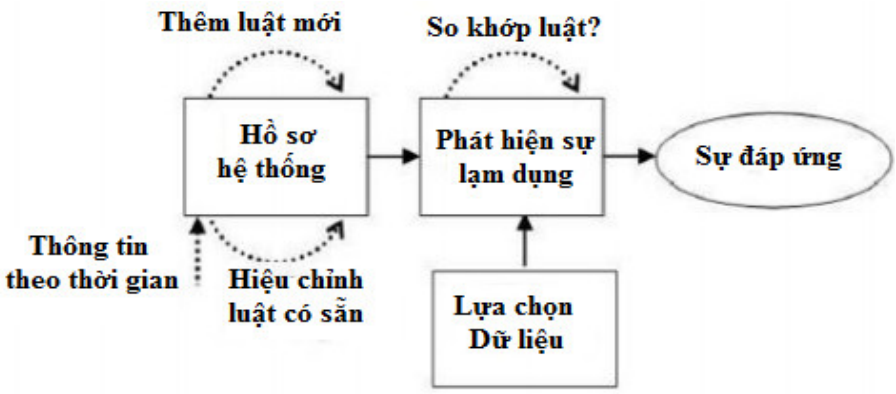
Hệ thống phát hiện xâm nhập (Intrusion Detection System - IDS) [10] là hệ thống có khả năng phân biệt hành vi người dùng bình thường và bất thường, ngoài ra, còn có chức năng giám sát, phân tích lưu lượng mạng, các hoạt động khả nghi và cảnh báo cho hệ thống, nhà quản trị.

2.2.1. Kỹ thuật phát hiện dựa trên phương pháp phát hiện sự lạm dụng

Những nghiên cứu về phát hiện xâm nhập dựa trên phương pháp phát hiện sự lạm dụng bắt đầu vào năm 1980 với báo cáo của Anderson [1]. Trong đó, hành vi xâm nhập được phát hiện bằng cách so sánh những hành vi được giám sát với các hành vi tấn công mẫu đã biết. Do đó, phương pháp này chỉ có hiệu quả trong việc phát hiện các dạng tấn công, đột nhập đã biết.

Mô hình phát hiện lạm dụng như minh họa trên hình 3 bao gồm bốn thành phần: thu thập dữ liệu, hồ sơ hệ thống, thành phần phát hiện sự lạm dụng, thành phần phản hồi. Dữ liệu được thu thập từ một hoặc nhiều nguồn, bao gồm báo cáo kiểm tra, lưu lượng mạng, dấu vết các lời gọi hệ thống, v.v... Dữ liệu thu thập được chuyển sang một định dạng mà các thành phần khác của hệ thống có thể xử lý được.

Hồ sơ hệ thống thường là một tập luật (rules), được sử dụng để mô tả các hành vi bình thường và bất thường.

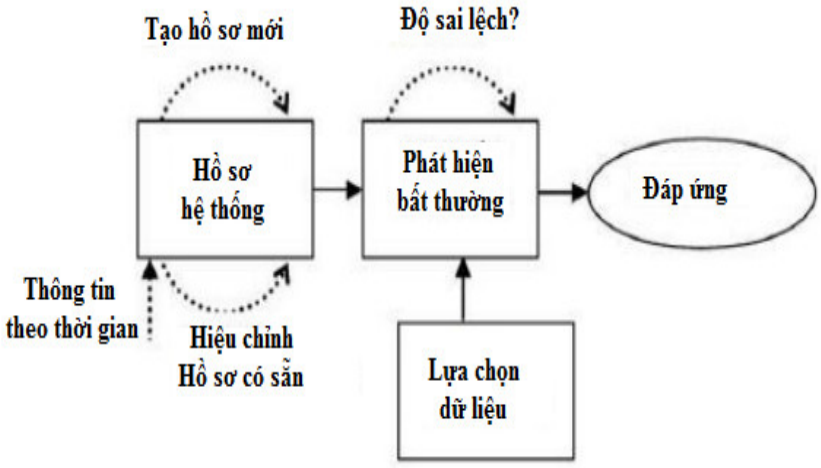


***Hình 3**. Mô hình phát hiện sự lạm dụng.*

Phương pháp phát hiện dựa trên sự lạm dụng có bốn kỹ thuật thường được sử dụng, kỹ thuật đối sánh mẫu, kỹ thuật dựa trên tập luật, kỹ thuật dựa trên trạng thái, và kỹ thuật dựa trên khai phá dữ liệu.

2.2.2. Kỹ thuật dựa trên phương pháp phát hiện sự bất thường

Khác với phát hiện dựa trên sự lạm dụng, phương pháp phát hiện dựa trên sự bất thường [1] là dựa vào việc thiết lập hồ sơ hoạt động bình thường cho hệ thống. Phương pháp này dựa trên giả định các hành vi tấn công, xâm nhập có quan hệ mật thiết với các hành vi bất thường. Các nghiên cứu phát hiện bất thường bắt đầu bằng cách định nghĩa những hành động như thế nào được coi là bình thường, và sau đó xác định những hoạt động nào là xâm nhập và phương pháp phân biệt từng hành động xâm nhập cụ thể.



***Hình 4**. Mô hình phát hiện sự bất thường.*

Mô hình phát hiện bất thường, như minh họa trên hình 4 bao gồm bốn thành phần: Thu thập dữ liệu, hồ sơ hệ thống bình thường, phát hiện bất thường và thành phần phản hồi. Các hành động sử dụng hệ thống bình thường hay lưu lượng dữ liệu được thu thập và lưu lại bởi thành phần thu thập dữ liệu. Các kỹ thuật mô hình cụ thể được sử dụng để tạo ra hồ sơ hệ thống bình thường. Thành phần phát hiện bất thường quyết định một hành vi được giám sát là bất thường thông qua mức sai lệch của hành vi đó với các hành vi bình thường trong tập hồ sơ. Cuối cùng, các thành phần phản ứng báo cáo sự xâm nhập được phát hiện. Ưu điểm chính của phương pháp dựa trên phát hiện bất thường là khả năng phát hiện các cuộc tấn công mới do nó không đòi hỏi có hiểu biết về các dạng tấn công này. Tuy nhiên, phương pháp này còn tồn tại một số hạn chế là tỷ lệ phát hiện sai thường khá cao do phương pháp này dựa trên giả định tấn công, xâm nhập đồng nghĩa với các bất thường. Trên thực tế, nhiều hành vi bất thường nhưng không phải là hành vi tấn công. Hơn nữa, phương pháp này cũng gặp phải khó khăn trong việc thu thập dữ liệu để xây dựng hồ sơ các hành vi bình thường. Chẳng hạn, hồ sơ hành vi bình thường của người dùng được xây dựng dựa trên dữ liệu thu thập được trong một khoảng thời gian hoạt động bình thường, những hoạt động xâm nhập không bị phát hiện trong thời gian này có thể được coi là hành vi bình thường. Điều này dẫn đến giảm tỷ lệ phát hiện đúng. Một vấn đề khác là kỹ thuật phát hiện bất thường khó có thể phát hiện các cuộc tấn công tàng hình, là một kiểu tấn công mà các hành vi tấn công ẩn trong số lượng lớn các hành vi bình thường.

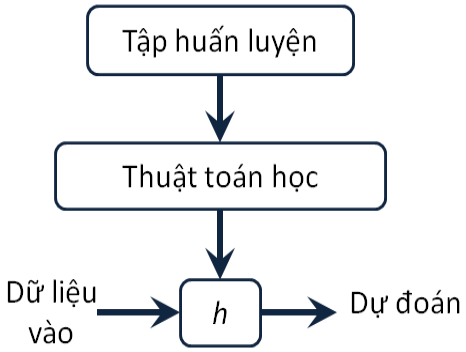
Phương pháp phát hiện dựa trên sự bất thường được chia thành các kỹ thuật chính như sau: kỹ thuật mô hình thống kê mở rộng, kỹ thuật dựa trên mô hình luật, kỹ thuật dựa trên mô hình sinh học và kỹ thuật dựa trên mô hình học.

2.3. Kỹ thuật học máy

Học máy (ML – Machine Learnning) [9] là kỹ thuật thiết kế và phát triển các thuật toán cho phép máy tính đánh giá hành vi dựa trên dữ liệu thực nghiệm, chẳng hạn như dữ liệu cảm biến hoặc cơ sở dữ liệu. Một chương trình học có thể tận dụng các mẫu (dữ liệu) để nắm bắt các đặc điểm quan tâm, dữ liệu có thể được xem như là ví dụ minh họa mối quan hệ giữa các biến quan sát được. Trọng tâm chính của nghiên cứu học máy là tự động học cách nhận ra các mẫu phức tạp và đưa ra quyết định thông minh dựa trên dữ liệu. Học máy có thể được chia thành các nhánh như sau: học có giám sát, học nửa giám sát và học không giám sát.

2.3.1. Kỹ thuật học có giám sát

Học có giám sát (Supervised learning) [9] là quá trình học với tập dữ liệu huấn luyện ban đầu hoàn toàn được gán nhãn từ trước. Học có giám sát sử dụng cho lớp bài toán phân lớp và phân loại. Với cách học này, kinh nghiệm được cho một cách tường minh dưới dạng đầu vào và đầu ra của hàm đích. Hình 5 mô tả kỹ thuật học có giám sát.



***Hình 5**. Mô hình học có giám sát.*

Một số kỹ thuật học có giám sát thường được quan tâm là máy hỗ trợ vector, cây quyết định, mạng thần kinh nhân tạo, lập trình di truyền …

2.3.2. Kỹ thuật học nửa giám sát

Kỹ thuật học nửa giám sát [9] là kỹ thuật học sử dụng cả dữ liệu đã gán nhãn và chưa gán nhãn để huấn luyện - điển hình là một lượng nhỏ dữ liệu có gán nhãn cùng với lượng lớn dữ liệu chưa gán nhãn. Nhiều nhà nghiên cứu nhận thấy dữ liệu không gán nhãn, khi được sử dụng kết hợp với một lượng nhỏ dữ liệu có gán nhãn, có thể cải thiện đáng kể độ chính xác. Trong kỹ thuật học có giám sát, để gán nhãn dữ liệu cho bài toán học máy thường đòi hỏi một chuyên viên có kỹ năng để phân loại bằng tay các mẫu huấn luyện. Trong khi đó, chi phí gán nhãn bằng tay cao, không khả thi. Với phương pháp kết hợp cả mẫu dữ liệu được gán nhãn và chưa gán nhãn sẽ đạt được hiệu quả cao hơn.

2.3.3. Kỹ thuật học không giám sát

Trong kỹ thuật học không giám sát [9], tập dữ liệu được cho dưới dạng $D = \{(x_1, x_2, \dots, x_N)\}$ với $(x_1, x_2, \dots, x_N)$ vector đặc trưng của mẫu huấn luyện. Nhiệm vụ của thuật toán là phải phân chia tập dữ liệu D thành các nhóm con, mỗi nhóm chứa các vector đầu vào có đặc trưng giống nhau.

Như vậy, việc học không giám sát, số lớp phân loại chưa biết trước, và tùy theo tiêu chuẩn đánh giá độ *tương tự* giữa các mẫu mà ta có thể có các lớp phân loại khác nhau.

3. ỨNG DỤNG KỸ THUẬT HỌC MÁY TRONG PHÁT HIỆN XÂM NHẬP MẠNG

3.1. Kỹ thuật học máy trong phát hiện xâm nhập

Học máy là kỹ thuật mạnh mẽ được một số nhà nghiên cứu ứng dụng vào giải quyết bài toán phát hiện xâm nhập mạng. Năm 1990 Fox và các cộng sự [6] lần đầu tiên cố gắng mô hình hóa hệ thống và các hành vi người dùng bằng mạng thần kinh nhân tạo. Đề xuất của họ sử dụng kỹ thuật học không giám sát để phát hiện cấu trúc cơ bản của dữ liệu mà không cần mẫu hành vi bất thường có sẵn. Năm 1994, Frank [5] sử dụng trí tuệ nhân tạo cho phát hiện xâm nhập theo hướng phân loại hành vi xâm nhập và giảm dữ liệu.

Một đề xuất dựa trên mạng lan truyền ngược để giám sát các chương trình đang chạy của Ghost [15] và các cộng sự dựa trên kỹ thuật học giám sát đã được đề xuất. Các tác giả đã sử dụng dữ liệu đầu vào được tạo ngẫu nhiên cho các hành vi bất thường, và cho rằng hiệu quả phát hiện của kỹ thuật này phụ thuộc vào trọng số khởi tạo đầu vào huấn luyện.

Một số nghiên cứu dựa trên thuật toán di truyền cũng được đề xuất, năm 1993 tác giả Me [8] sử dụng thuật toán di truyền cho phát hiện lạm dụng. Đề xuất này đã cải thiện tỷ lệ cảnh báo nhầm hiệu quả; tuy nhiên phương pháp này chưa xác định chính xác từng loại tấn công cụ thể.

3.2. Thuật toán quy nạp cây ID3

Thuật toán quy nạp cây ID3 [9] được Quinlan đề xuất cuối thập niên 1970s với ưu điểm là lựa chọn các thuộc tính tốt nhất để triển khai cây tại mỗi bước bằng cách sử dụng độ lợi (Gain) thông tin để đo tính hiệu quả của các thuộc tính phân lớp. Trong quá trình xây dựng cây quyết định theo thuật toán ID3 tại mỗi bước phát triển cây, thuộc tính được chọn để triển khai là thuộc tính có độ lợi lớn nhất.

Xét trường hợp đơn giản nhất cho bộ dữ liệu huấn luyện trên bài toán phát hiện xâm nhập, ta chỉ quan tâm đến địa chỉ IP nguồn, IP đích, cổng nguồn, cổng đích để xác định mẫu đó có phải tấn công hay không như biểu diễn trong bảng 1.

Bảng 1. *Tập dữ liệu huấn luyện cho bài toán phát hiện xâm nhập.*

IP nguồn	IP đích	Cổng nguồn	Cổng đích	Xâm nhập
123.202.72.109	225.142.187.12	001360	000080	False
123.202.72.109	225.142.187.12	001360	000025	False
225.142.147.75	225.142.187.12	001360	000080	True
233.167.15.65	150.216.191.119	001360	000080	True

233.167.15.65	125.250.187.19	001425	000080	True
233.167.15.65	125.250.187.19	001425	000025	False
225.142.147.75	125.250.187.19	001425	000025	True
123.202.72.109	150.216.191.119	001360	000080	False
123.202.72.109	125.250.187.19	001425	000080	True
233.167.15.65	150.216.191.119	001425	000080	True
123.202.72.109	150.216.191.119	001425	000025	True
225.142.147.75	150.216.191.119	001360	000025	True
225.142.147.75	225.142.187.12	001425	000080	True
233.167.15.65	150.216.191.119	001360	000025	False

Mỗi mẫu trong tập dữ liệu này được phân loại là "True" (xâm nhập) hoặc "False" (không phải xâm nhập), giá trị phân loại này được gọi là thuộc tính đích. Quá trình huấn luyện của ID3 sẽ xây dựng một cây quyết định có khả năng phân loại chính xác mẫu trong tập dữ liệu huấn luyện với kỳ vọng cho kết quả chuẩn đoán chính xác ở đầu ra.

ID3 xây dựng cây quyết định theo phương pháp từ trên xuống, tại mỗi nút ID3 sẽ chọn một thuộc tính để kiểm tra và phân vùng tập hợp các mẫu bằng cây con đệ quy cho mỗi vùng. Thuật toán lặp lại cho đến khi mọi thành viên của phân vùng đều nằm trong cùng một lớp, lớp đó trở thành nút lá của cây. Hiệu quả của thuật toán phụ thuộc rất nhiều vào tiêu chuẩn chọn giá trị gốc của cây.

ID3_algorithm(T_{Set}, Class_Labels, Attri){

If Tất_cả_các_mẫu của T_{Set} thuộc cùng Class_C

Return Nút Root được gắn với Class_C

If Tập thuộc tính Attri là rỗng

Return Nút Root được gắn nhãn lớp ≡ Majority_Class_Label(T_{Set})

A ← Thuộc tính⊂ Attri có khả năng phân loại "tốt nhất" đối với T_{Set}

Thuộc tính kiểm tra cho nút Root ← A

For each Giá trị có thể v của thuộc tính A

Bổ sung một nhánh cây mới dưới nút Root, tương ứng với: "*Giá trị của A là v*"

Xác định T_{Setv} = {mẫu x | x ⊆ T_{Set}, xA=v}

If (T_{Setv} là rỗng)

Tạo một nút lá với nhãn lớp ≡ Majority_Class_Label(T_{Set})

Gắn nút lá này vào nhánh cây mới vừa tạo

Else Gắn vào nhánh cây mới vừa tạo một cây con được tạo ra bởi ID3_algorithm(T_{Setv}, Class_Labels, { Attri A})

Return Root

}

Việc lựa chọn thuộc tính A có khả năng phân loại "tốt nhất" đối với tập dữ liệu huấn luyện T_{Set} được thực hiện theo công thức:

$$Gain(T_{Set}, A) = Entropy(T)Set - \sum v \in Value \frac{T_{Setv}}{T_{Set}} \text{Entropy}(T_{Set})$$

Với Values(A) là tập hợp có thể có các giá trị thuộc tính A và T_{Setv} là tập con của T_{Set} chứa các mẫu có thuộc tính A mang giá trị v; Độ thuần nhất cho tập dữ liệu Entropy(T_{Set}) xác định theo công thức sau:

$$\text{Entropy}(T_{Set}) = -\frac{A_{\text{True}}}{\text{Attri}} * \log_2\left(\frac{A_{\text{True}}}{\text{Attri}}\right) - \frac{A_{\text{False}}}{\text{Attri}} * \log_2\left(\frac{A_{\text{False}}}{\text{Attri}}\right)$$

Với tập huấn luyện được cho trong bảng 1 gồm 02 thuộc tính "*True*" và "*False*", do đó tỷ lệ các mẫu của mỗi thuộc tính được xác định là A_{True} và A_{False}.

Áp dụng cho dữ liệu huấn luyện trong bảng 1 ta có thể xây dựng cây quyết định theo thuật toán ID3 như sau:

Bước 1. Tính Entropy của tập dữ liệu

$$\text{Entropy}(T_{Set}) = -(9/14)\log_2(9/14) - (5/14)\log_2(5/14) = 0.940$$

Bước 2. Tình Gain cho từng thuộc tính để tìm thuộc tính làm gốc

$$\text{Gain}(T_{Set}, \text{IP}_{\text{Nguồn}}) = \text{Entropy}(T_{Set}) - (5/14)\text{Entropy}(S_{123.202.72.109}) - (4/14)\text{Entropy}(S_{225.142.147.75}) - (5/14)\text{Entropy}(S_{233.167.15.65}) = 0.299$$

Với:

$$\text{Entropy}(S_{123.202.72.109}) = -(2/5) * \log_2(2/5) - (3/5)\log_2(3/5) = 0.970$$

$$\text{Entropy}(S_{225.142.147.75}) = -(4/4)\log_2(4/4) = 0$$

$$\text{Entropy}(S_{233.167.15.65}) = -(3/5)\log_2(3/5) - (2/5)\log_2(2/5) = 0.794$$

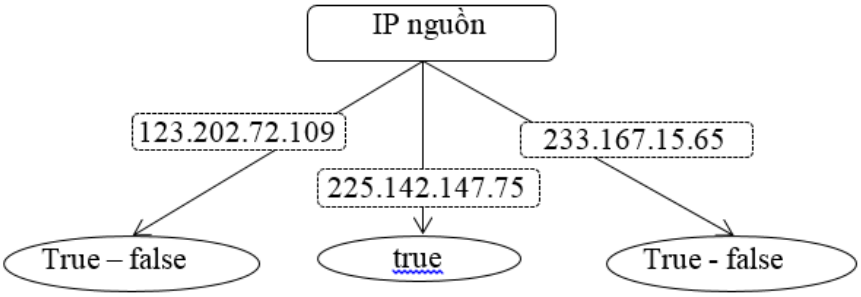
Tương tự với đích:

$$\text{Gain}(T_{Set}, \text{IP}_{\text{Đích}}) = \text{Entropy}(T_{Set}) - (4/14)\text{Entropy}(S_{225.142.187.12}) - (6/14)\text{Entropy}(S_{150.216.191.119}) - (4/14)\text{Entropy}(S_{125.250.187.19}) = 0.029$$

$$\text{Gain}(T_{Set}, \text{Port}_{\text{Đích}}) = \text{Entropy}(T_{Set}) - (6/14)\text{Entropy}(S_{000025}) - (8/14)\text{Entropy}(S_{00008}) = 0.145$$

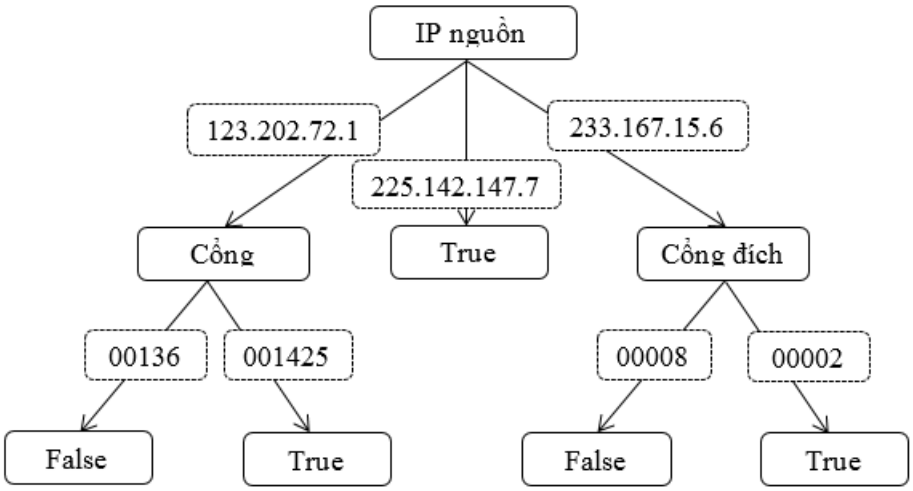
$$\text{Gain}(T_{\text{Set}}, \text{Port}_{\text{Nguồn}}) = \text{Entropy}(T_{\text{Set}}) - (7/14)\text{Entropy}(S_{00136}) - (7/14)\text{Entropy}(S_{001425}) = 0.15$$

Dễ thấy $\text{Gain}(T_{\text{Set}}, \text{IP}_{\text{Nguồn}})$ là lớn nhất. Vậy lấy thuộc tính $\text{IP}_{\text{Nguồn}}$ làm nút gốc. Và ta có thể xây dựng được cây ban đầu như hình 6.



***Hình 6**. Cây quyết định sau khi xác định được nút gốc (Root).*

Tiếp tục xét các nhánh con dưới nút gốc cho đến khi tất cả các Entropy đều thuần nhất (không thể xây dựng được cây nữa), thì ta có cây quyết định xây dựng bằng giải thuật ID3 như hình 7:



***Hình 7**. Cây quyết định được xây dựng theo thuật toán ID3.*

3.3. Thuật toán phân cụm dữ liệu mờ

Phân cụm dữ liệu mờ là phương pháp phân cụm dữ liệu cho phép mỗi điểm dữ liệu thuộc về hai hoặc nhiều cụm thông qua bậc thành viên. Ruspini [11] giới thiệu khái quát khái niệm phân hoạch mờ để mô tả cấu trúc cụm tập dữ liệu và đề xuất một thuật toán để tính toán tối ưu phân hoạch mờ. Dunn [4] mở rộng phương pháp phân cụm và đã phát triển thuật toán phân cụm mờ.

Ý tưởng của thuật toán là xây dựng một phương pháp phân cụm mờ dựa trên tối thiểu hóa hàm mục tiêu. Bezdek [3] cải tiến và tổng quá hóa hàm mục tiêu mờ bằng cách đưa ra trọng số mũ để xây dựng thuật toán phân cụm mờ FuzzyCMeans (FCM), và được chứng minh độ hội tụ của các thuật toán là cực tiểu cục bộ.

Thuật toán FCM thực hiện phân hoạch một tập *n* vector đối tượng dữ liệu $\mathbf{X}=\mathbf{x}_1,x_2,...,x_n \in R^d$ thành c nhóm mờ dựa trên tính toán tối thiểu hóa hàm mục tiêu để đo chất lượng của phân hoạch và tìm trọng tâm cụm trong mỗi nhóm, sao cho chi phí hàm đo độ tương tự là nhỏ nhất. Một phân hoạch mờ n vector điểm dữ liệu $\mathbf{X}=\mathbf{x}_1,x_2,...,x_n \in R^d$ là đặc trưng đầu vào được biểu diễn bởi ma trận $\mathbf{U}=[u_{ik}]$ sao cho điểm dữ liệu đã cho chỉ có thể thuộc về một số nhóm với bậc được xác định bởi mức độ giữa [0,1]. Như vậy, ma trận *U* được sử dụng để mô tả cấu trúc cụm của X bằng cách giải thích u_{ik} như bậc thành viên x_kvới cụm i.

Cho $\mathbf{u}=(u_1,u_2,...,\mathbf{u}_c)$ là phân hoạch mờ C, ta có:

$$U_{c\times n}=\begin{bmatrix} u_{11} & \cdots & u_{1n} \\ \vdots & \ddots & \vdots \\ u_{c1} & \cdots & u_{cn} \end{bmatrix}$$

Khi đó để tính toán hàm mục tiêu mờ với tham số mờ m, trọng tâm của cụm mờ thứ i là $v_i \in R^d$được xác định như sau:

$$J_m(\mathrm{U,v})=\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m(d_{ik})^2,\ 1\leq m\leq+\infty$$

Khoảng cách giữa mẫu dữ liệu x_kvới trọng tâm cụm thứ $(\mathbf{i},\mathbf{v}_i)$ được xác định theo phương pháp Euclide$d_{ik}=\mathrm{d}(x_k-v_i)=\|x_k-v_i\|=\left[\sum_{j=1}^{d}(x_{kj}-v_{ij})^2\right]^{1/2}$, bậc của mẫu dữ liệu x_kvới cụm thứ i là $u_{ik}\in[0,1]$. Ma trận biểu diễn các giá trị tâm của cụm được phân hoạch $\mathbf{V}=[v_{ij}]=[v_1,...,\mathbf{v}_c]\in R^{d\times c}$. Để đơn giản, ta coi mảng đối tượng dữ liệu $x_1,...,\mathbf{x}_n$là các cột trong ma trận đối tượng dữ liệu $\mathbf{X}=[x_{jk}]=[x_1,...,x_n]\in R^{d\times c}$. Ma trận phân hoạch U được sử dụng để mô tả cấu trúc cụm trong dữ liệu $\{x_1,...,x_n\}$.

Hàm mục tiêu đạt giá trị nhỏ nhất khi phân hoạch u_{ik} và phân cụm v_i thỏa mãn:

$$u_{ik}=\begin{cases} \dfrac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{\frac{2}{m-1}}},I_k=\emptyset \\ \qquad\qquad\qquad\qquad ,1\leqslant i\leqslant c,1\leqslant k\leqslant n \\ \begin{cases} 0,i\notin I_k \\ \sum_{i\in I_k}u_{ik}=1,i\in I_k \end{cases},I_k\neq\emptyset \end{cases}$$

và $v_i = \frac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m}, 1 \le i \le c$

Thuật toán:

Input: Số cụm *c* và tham số *m* cho hàm mục tiêu J, với sai số ε

Output: c cụm dữ liệu sao cho hàm mục tiêu đạt giá trị cực tiểu

Algorithm: Fuzzy C – Mean (FCM)

Begin

Bước 1. Khởi tạo

Nhập tham số $c(1<c<n), m(1<m<\infty)$,

Khởi tạo ma trận $V=[v_{ij}], V^{(0)} \in R^{d \times c}, j=0$

Bước 2. Tính ma trận phân hoạch U và cập nhật lại trọng tâm cụm V

2.1. j=j+1

2.2. Tính ma trận phân hoạch mờ $U^{(j)}$

2.3. Cập nhật các trọng tâm cụm $V^{(j)} = \left[v_1^{(j)}, v_2^{(j)}, ..., v_c^{(j)}\right]$ và $U^{(j)}$

Bước 3: Kiểm tra điều kiện dừng. Nếu $\max\left\{\left\|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right\|\right\} \le \varepsilon$ chuyển sang bước 4, ngược lại quay lại bước 2.

Bước 4. Đưa ra các cụm kết quả.

End

4. THỰC NGHIỆM VÀ KẾT QUẢ

Để đánh giá kết quả của các thuật toán học máy được giới thiệu trong phần 3. Nhóm đã tiến hành cài đặt thử nghiệm trên 10% của bộ dữ liệu huấn luyện và kiểm tra KDD'99 đối với thuật toán học có giám sát (ID3) và thuật toán học không giám sát (FCM) tại phòng thí nghiệm An ninh mạng – Học viện Kỹ thuật quân sự. Với các kết quả nhận được, để đánh giá độ tin cậy của phương pháp học, độ tin cậy (Accuracy) được tính toán như sau:

$$\text{Accuracy} = \frac{TN}{TN + FP} * 100$$

Trong đó: - TN: Số bản ghi được phân loại đúng

- FP: Số bản ghi bị phân loại nhầm

Mỗi thuật toán với các tham số thiết lập đầu vào được thực hiện 20 lần, kết quả thống kê là trung bình của cả 20 lần thực hiện.

***Bảng 2**. Kết quả thực nghiệm sử dụng thuật toán ID3 cho phân loại tấn công.*

Số bản ghi DL huấn luyện	**Số bản ghi DL kiểm tra**	**Thuật toán ID3**	
		Phân loại đúng (%)	Độ tin cậy (%)
800	2000	92,90	91.35
8000	4000	99,98	99.97
	8000	99,93	99,92
10000	8000	98,72	98,33

Kết quả bảng 2 cho thấy với trường hợp dữ liệu huấn luyện 800 bản ghi và dữ liệu kiểm tra là 2000 bản ghi; việc huấn luyện trên bộ dữ liệu với số lượng mẫu quá ít so với số mẫu kiểm tra đã ảnh hưởng lớn đến độ chính xác cũng như độ tin cậy của thuật toán.

Trong trường hợp huấn luyện trên bộ mẫu dữ liệu lớn và tiến hành kiểm tra trên bộ dữ liệu có số lượng mẫu nhỏ hơn sẽ nhận được độ chính xác và độ tin cậy cao. Khi huấn luyện với bộ dữ liệu có 8000 bản ghi và kiểm tra trên bộ dữ liệu 4000 bản ghi kết quả nhận dạng chính xác đạt 99.98% với độ tin cậy đạt 99.97%.

Tuy nhiên, khi huấn luyện và kiểm tra trên các bộ dữ liệu lớn hơn thì hiệu quả phân loại và độ tin cậy của thuật toán ID3 sẽ giảm dần, đặc biệt đối với trường hợp số mẫu của bộ dữ liệu huấn luyện lớn hơn không đáng kể so với số mẫu của bộ dữ liệu kiểm tra.

Đối với kỹ thuật học không giám sát FCM, trong quá trình thực nghiệm đã tiến hành 3 thiết lập tham số khác nhau với số cụm lần lượt là 2, 3 và 4; tham số mờ được lựa chọn là 2. Kết quả được thể hiện trong bảng 3. Trong quá trình thử nghiệm cho thấy khi thay đổi tham số mờ m thì hiệu quả và độ tin cậy của kỹ thuật thay đổi không đáng kể.

***Bảng 3**. Kết quả thực nghiệm sử dụng thuật toán FCM cho phân loại tấn công.*

Số bản ghi DL huấn luyện	Số bản ghi DL kiểm tra	**Thuật toán FCM**					
		c =2, m=2		**c=3, m=2**		**c = 4, m=2**	
		Phân loại đúng (%)	Độ tin cậy (%)	Phân loại đúng (%)	Độ tin cậy (%)	Phân loại đúng (%)	Độ tin cậy (%)
800	2000	50,35	100	49,85	100	94,4	93.43
8000	4000	49,75	100	49,75	100	96,3	96,2
	8000	48,63	100	49,7625	100	97,06	96.847
10000	8000	48,63	100	49,0625	100	94,91	93,893

Với trường hợp số cụm được thiết lập là 2 hoặc 3 thì hiệu quả phân loại rất thấp, chỉ đạt cao nhất là 50.35% , tuy nhiên độ tin cậy của thuật toán đạt 100%.

Trong trường hợp thiết lập tham số phân cụm là 4, hiệu quả phân loại tăng lên rất nhanh, điều này cho thấy số cụm được thiết lập càng lớn thì kết quả phân loại càng cao. Tuy nhiên, độ tin cậy của thuật toán sẽ giảm dần.

Mặt khác, tương tự với thuật toán học có giám sát (ID3), với số mẫu huấn luyện nhỏ hơn số mẫu kiểm tra sẽ cho kết quả phân loại và độ tin cậy thấp hơn khá nhiều so với các thí nghiệm đối với số mẫu huấn luyện lớn hơn khá nhiều so với số mẫu được kiểm tra.

Từ kết quả bảng 2 và bảng 3 cho thấy, quá trình học có giám sát đạt hiệu quả phân loại đúng và độ tin cậy cao hơn nhiều so với học không giám sát. Tuy nhiên, quá trình học có giám sát yêu cầu các mẫu dữ liệu huấn luyện đã được gán nhãn, mà chi phí để xây dựng bộ dữ liệu được gán nhãn là khá cao, trong khi huấn luyện đối với kỹ thuật học không giám sát thì không cần bộ dữ liệu huấn luyện được gán nhãn, trong trường hợp số cụm được thiết lập cao thì hiệu quả của kỹ thuật học không giám sát cũng tương đương với kỹ thuật học có giám sát.

5. KẾT LUẬN

Bài báo đã trình bày một số nội dung nghiên cứu về các kỹ thuật học máy và ứng dụng trong lĩnh vực phát hiện tấn công xâm nhập mạng. Các kết quả nghiên cứu áp dụng trên thuật toán ID3 và FCM đã cho thấy hiệu quả của học máy trong phân loại tấn công. Tuy nhiên, trong thuật toán FCM chưa có một quy tắc cụ thể để lựa chọn tham số m sao cho hiệu quả phân loại tối ưu. Do đó, trong thời gian tới, nhóm nghiên cứu sẽ tiếp tục định hướng nghiên cứu cho hệ thống tự đáp ứng tham số m để đạt được hiệu quả tối ưu của thuật toán.

TÀI LIỆU THAM KHẢO

[1]. Anderson, James P. "*Computer Security Threat Monitoring and Surveillance*", 15 April 1980

[2]. Bhat A. H., Patra S., Jena D. "*Machine learning approach for intrusion detection on cloud virtual machines*". International Journal of Application or Innovation in Engineering & Management (IJAIEM), 2013, 2(6) 56-66.

[3] J. C. Bezdek, "*Pattern Recognition with Fuzzy Objective Function Algorithms*", Plenum Press, New York, (1981).

[4]. Dunn JC. "*A fuzzy relative to the ISODATA process and its use in detecting compact well-separated clusters*". J Cybernet 1974, 3:310–313.

[5]. Frank, J. "*Artificial intelligence and intrusion detection: Current and future directions*". In Proceedings of the National 17th Computer Security Conference (1994).

[6]. Fox, K. L., Henning, R. R., Reed, J. H., and Simonian, R. "*A neural network approach towards intrusion detection*". In Proceedings of the 13th National Computer Security Conference, 125–134 (1990).

[7]. Gaidhane R., Vaidya C., Raghuwanshi M. "*Survey: Learning Techniques for Intrusion Detection System (IDS)*", International Journal of Advance Foundation and Research in Computer (IJAFRC), 2014, 1(2) 21-28

[8]. Me, Ludovic. "*Security Audit Trail Analysis Using Genetic Algorithms.*" Proceedings of the Twerfh International Conference on Computer Safety, Reliability, and Security, Poznan, Poland, 1993

[9]. Mitchell, Tom M. "*Machine Learning*". McGraw-Hill, 1997.

[10]. Stephen Northcutt, Judy Novak, "*Network Intrusion Detection*", Third Edition, New Riders Publishing, United States of America, 2004.

[11]. Ruspini, E.H. "*A new approach new clustering*". Information and control, 15 (1969), 22-32.

[12]. Devarakonda, N., S. Pamidi, et al. "*Intrusion Detection System using Bayesian Network and Hidden Markov Model*". Procedia Technology, 2012, 4(0) 506-514.

[13]. B.Ben Sujitha, R.Roja Ramani, Parameswari, "*Intrusion Detection System using Fuzzy Genetic Approach*", International Journal of Advanced Research in Computer and Communication Engineering, Vol.1, Issue 10, December 2012

[14]. KDD 99 Task. Avaiable at: http://kdd.ics.uci.edu/databases/kddcup99/task.html.

[15] A. K. Ghost, J. Wanken, F. Charron (September 27, 1997), "*Detecting Anomalous and Unknown Intrusions Against Programs in Real-Time*". DARPA SBIR FOCI Tutorial 2007 134 Phase I Final Report. Reliable Software Technologies.

ABSTRACT

INTRUSION DETECTION USING MACHINE LEARNING TECHNIQUES

In recent years, the growth of computer network is entailed many challenges to cyber security. Network is becoming the target of attacks, unauthorized intrusions and information stealed. Most of traditional intrusion detection techniques are known with relatively low true positive rate and high false alarm rate. Research on intrusion detection using machine learning techniques have proved effectively in detecting new attacks with high detection rate and low false alarm rate in reasonable computational cost. In this paper, we study some machine learning techniques (FCM, IC3) in network intrusion detection. Some experiments were conducted on KDD99 datasets at Laboratory of Network Security – Le Quy Don Technical University.

Keywords: Machine Learning, Network Intrusion, Intrusion Detection, Cluster.

Nhận bài ngày 03 tháng 3 năm 2017
Hoàn thiện ngày 04 tháng 4 năm 2017
Chấp nhận đăng ngày 01 tháng 5 năm 2017

Địa chỉ: Học viện Kỹ thuật quân sự
* Email: canhvuvan@yahoo.com.