PHÁT HIỆN XÂM NHẬP TRÁI PHÉP MẠNG KHÔNG DÂY SỬ DỤNG TÁC TỬ VỚI CƠ CHẾ HỌC MÁY

Intrusion Detection for Wireless Networks using Agents with Learning Mechanism

Từ Minh Phương, Nguyễn Minh Tuấn

Tóm tắt

Do đặc thù của mạng không dây, giải pháp phát hiện xâm nhập mạng trái phép cần có sự thay đổi và cải tiến so với mạng có dây thông thường. Bài báo trình bầy giải pháp phát hiện xâm nhập trái phép mạng không dây sử dụng tác tử với cơ chế học máy và phát hiện tình huống bất thường. Hai phương pháp phát hiện bất thường được lựa chọn là phương pháp dựa trên phân tích tương quan giữa các thuộc tính và phương pháp phát hiện mẫu mới bằng cách thực hiện các ánh xạ với hàm nhân (kernel). Hiệu quả của hai phương pháp được thử nghiệm và so sánh với nhau trên dữ liệu mô phỏng hoạt động của mạng không dây ad-hoc.

Từ khoá: *mạng không dây, phát hiện xâm nhập trái phép, học máy, phát hiện bất thường*

Abstract

Due to the differences between wireless and traditional networks, intrusion detection techniques developed for the later should be modified to be effective when applied to the former. This paper presents a solution to detecting intrusions in mobile wireless networks, which is based on intelligent agents with anomaly detection mechanism. We study two anomaly detection methods: the first method detects anomalies by analyzing cross-feature correlations, the second method uses kernel trick and margin maximization to detect novelty from log data. We present empirical results, which allow assessing and comparing the two methods when applied to simulated ad-hoc networks.

Keywords: *wireless network, intrusion detection, machine learning, anomaly detection*

1. ĐẶT VẤN ĐỀ

Mạng không dây ad-hoc được tạo thành từ các thiết bị không dây di động. Khác với mạng không dây thông thường, mạng ad-hoc không có hạ tầng mạng cố định, ví dụ không có các bộ định tuyến được quy định từ trước [6]. Các nút mạng đều tham gia vào việc định tuyến và chuyển tiếp các gói tin sao cho những nút nằm ngoài phạm vi liên lạc vô tuyến trực tiếp có thể liên lạc với nhau thông qua nút khác. Một trong những ưu điểm cơ bản của mạng ad-hoc là sự linh hoạt, khả năng thay đổi nhanh chóng thành phần và tô pô mạng. Mạng ad-hoc có thể dễ dàng hình thành và hoạt động trong những điều kiện khác nhau như chiến tranh, thiên tai, khi có hội nghị, mạng giữa các thiết bị không dây trong gia đình.v.v.

Bên cạnh sự tiện lợi mà mạng không dây đem lại, những đặc điểm của loại mạng này cũng đặt ra nhiều vấn đề về đảm bảo an ninh mạng. Đối với mạng có dây truyền thống, việc hạn chế xâm nhập trái phép mạng có thể thực hiện tập trung, chẳng hạn nhờ bức tường lửa. Trong khi đó, do tính mở và thiếu cơ chế quản lý tập trung, việc tấn công mạng ad-hoc có thể diễn ra ngay từ bên trong, từ mọi nút của

mạng, do vậy không thể xây dựng sẵn các tuyến an ninh cố định cho loại mạng này. Các yếu tố khác cũng gây khó khăn cho việc đảm bảo an ninh mạng bao gồm nguồn điện hạn chế của thiết bị mạng, tô pô mạng động và giao thức mạng đòi hỏi các nút hợp tác trên cơ sở tin tưởng lẫn nhau.

Do những lý do nêu trên, việc đảm bảo an ninh cho mạng không dây ad-hoc cần có những thay đổi thích hợp so với mạng thông thường. Phương pháp đảm bảo an ninh thông dụng nhất là *ngăn ngừa* xâm nhập trái phép bằng cách sử dụng cơ chế xác thực và mã hoá thông tin. Tuy nhiên, phương pháp này có thể bị vô hiệu hoá nếu các nút mạng đã bị tấn công và xâm chiếm (có thể từ trước khi tham gia vào mạng), ví dụ bị nhiễm vi rút, bị dò mật khẩu, bị các lỗi tràn bộ đệm.v.v.. Các nút này đều có khoá hợp lệ để tham gia vào mạng và tiến hành tấn công từ bên trong. Để đảm bảo an ninh cho mạng, do vậy, ngoài việc ngăn ngừa xâm nhập cần có thêm cơ chế *phát hiện* và *xử lý* các xâm nhập trái phép khi các xâm nhập này đã hoặc đang xảy ra.

Bài báo này trình bầy một số giải pháp phát hiện xâm nhập trái phép cho mạng ad-hoc. Việc phát hiện xâm nhập trái phép được thực hiện phân tán bởi các tác tử hoạt động trên mỗi nút mạng. Mỗi tác tử được trang bị cơ chế phát hiện dấu hiệu bất thường từ dữ liệu kiểm tra của nút. Hai phương pháp phát hiện bất thường được trình bầy trong bài báo là phương pháp phân tích tương quan giữa các thuộc tính [3,11] và phương pháp phát hiện mẫu mới sử dụng các *hàm nhân* (*kernel function*) [1]. Phương pháp thứ hai được đề xuất cho một ứng dụng khác, và đây là lần đầu tiên được sử dụng cho phát hiện xâm nhập trái phép mạng không dây. Việc ứng dụng phương pháp này trong phát hiện xâm nhập mạng ad-hoc và so sánh với phương pháp thứ nhất cho phép lựa chọn phương pháp có nhiều ưu điểm hơn. Đây cũng là đóng góp quan trọng nhất của bài báo này.

Phần còn lại của bài báo được bố cục như sau. Phần 2 phân tích sự cần thiết phải xây dựng cơ chế phát hiện xâm nhập trái phép dưới dạng tác tử và mô tả kiến trúc h hiện xâm nhập. Phần 3 trình bày hai p pháp phát hiện bất thường sử dụng sử d thuật học máy. Phần 4 mô tả các thực n và kết quả. Phần 5 là kết luận của bài bá

2. TÁC TỬ PHÁT HIỆN XÂM N MẠNG TRÁI PHÉP

2.1. Tại sao sử dụng tác tử

Như đã nhắc tới trong phần mở đầu với mạng truyền thống, mạng ad-hoc có hạ tầng mạng cố định, mỗi nút mạn tham gia vào việc định tuyến và chuyể gói tin một cách tự nguyện. Do đặc điển trong mạng ad-hoc không tồn tại cơ chế lý tập trung, mỗi nút chỉ có thông tin c về trạng thái mạng. Đặc thù của mạng a cũng khiến cho các hành động tấn công có thể xảy ra từ bên ngoài lẫn bên trong r bất cứ nút nào cũng có thể bị xâm nhập nút không thể tin cậy hoàn toàn thông t nút khác cung cấp. Việc phát hiện xâm trái phép, do vậy, phải được thực hiện tán. Mỗi nút cần tự chủ trong việc phát xâm nhập và cơ chế phát hiện xâm nhập từng nút phải dựa trên dữ liệu cục bộ m đó có.

Trong nhiều trường hợp, các hành vi thường của mạng rất dễ bị nhầm lẫn với vi tấn công mạng. Hệ thống phát hiện nhập khi đó có thể đưa ra những cảnh bác Việc kết hợp quyết định của nhiều nút k có thể tăng độ chính xác cảnh báo. Do các nút cần có cơ chế trao đổi quyết trước khi đưa ra cảnh báo.

Tác tử (mềm) là các chương trình máy có khả năng hoạt động độc lập và đưa ra q định một cách tự chủ. Ngoài ra, tác tử có năng trao đổi thông tin, cộng tác hoặc c tranh với tác tử khác. Các đặc điểm này tác tử phù hợp với những yêu cầu về hệ hiện xâm nhập trái phép cho mạng ad- được phân tích ở trên và sẽ được sử dụng kiến trúc cơ sở của hệ phát hiện xâm nhập phép.

2.2. Kiến trúc hệ tác tử phát hiện xâm nhập

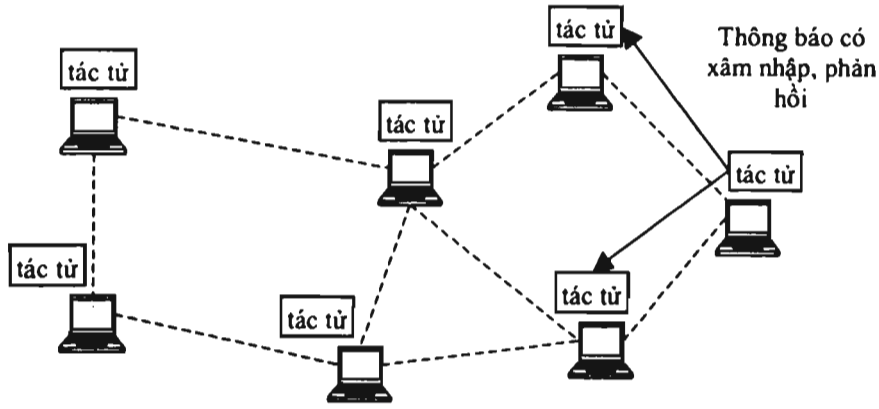
Trên hình 1 là kiến trúc hệ thống tác tử phát hiện xâm nhập trái phép [11]. Mỗi nút mạng có một tác tử của riêng mình. Tác tử theo dõi dữ liệu kiểm tra mà nút có và phát hiện các dấu hiệu bất thường bằng một trong các thuật toán mô tả ở phần sau. Khi phát hiện tình huống khả nghi, tác tử sẽ gửi thông báo tới các tác tử láng giềng. Tác tử láng giềng ở đây là tác tử của các nút nằm trong vùng liên lạc vô tuyến trực tiếp với nút hiện tại. Phản hồi của tác tử láng giềng cho phép tăng hoặc giảm mức độ cảnh báo. Chẳng hạn trong trường hợp nhiều nút cùng có phản hồi về xâm nhập mạng, tính chính xác của cảnh báo sẽ cao hơn.

Trong phạm vi nghiên cứu này, chúng tôi mới hạn chế trong việc nghiên cứu ảnh hưởng của cơ chế phát hiện xâm nhập trên từng tác tử. Ảnh hưởng của việc tương tác giữa các tác tử tới hiệu quả toàn hệ thống sẽ được đề cập trong những nghiên cứu sau.

3. PHÁT HIỆN BẤT THƯỜNG TRONG HOẠT ĐỘNG CỦA MẠNG AD-HOC

Phần này trình bầy cơ chế phát hiện xâm nhập trái phép của từng tác tử dựa trên thông tin cục bộ mà tác tử có.

Nguyên tắc chung của hệ thống phát hiện xâm nhập mạng trái phép là phân tích các hành vi của người dùng hay chương trình thể hiện qua dữ liệu kiểm tra (audit data), trên cơ sở đó phát hiện các hành vi có liên quan tới xâm nhập trái phép. Có hai cách chính để phát hiện xâm nhập là *phát hiện dùng sai* (misuse detection) và *phát hiện bất thường* (*anomaly detection*).



Hình 1. Kiến trúc hệ thống phát hiện xâm nhập trái phép sử dụng tác tử. Các đường không liền nét nối các nút có liên lạc vô tuyến trực tiếp

Phát hiện dùng sai là phương pháp lưu lại *mẫu đặc trưng* của các kiểu xâm nhập mạng đã biết và so sánh mẫu với hoạt động hiện thời của mạng. Mẫu đặc trưng có thể là chuỗi các hành vi liên quan tới các tấn công hoặc hậu quả của các đợt tấn công mạng. Ví dụ, chuỗi các thao tác đăng nhập sai mật khẩu liên tiếp thường là mẫu hành vi của kiểu xâm nhập sử dụng kỹ thuật dò mật khẩu. Ưu điểm của phương pháp phát hiện dùng sai là đơn giản, nhanh, và cho phép phát hiện khá chính xác các hình thức tấn công đã biết. Tuy nhiên phương pháp này không cho phép phát hiện các kiểu tấn công mới.

Phương pháp phát hiện bất thường phát hiện và cảnh báo về các hiện tượng không bình thường trong hoạt động của người dùng, chương trình hoặc hệ thống mạng. Hiện tượng không bình thường được phát hiện dựa trên việc so sánh với hoạt động bình thường đã

được ghi nhận trước đó, ví dụ chuỗi những lệnh mà một người dùng cụ thể thường sử dụng trong các phiên làm việc, hay tần suất thay đổi thông tin định tuyến. Do không cần biết trước thông tin về các dạng tấn công cụ thể, phương pháp này có khả năng phát hiện những kiểu tấn công mới. Nhược điểm của phương pháp này là khó xác định chính xác kiểu tấn công và có tỷ lệ cảnh báo nhầm cao.

Do ưu điểm của phương pháp phát hiện bất thường trong việc phát hiện các dạng tấn công mới, phần còn lại của bài báo sẽ chỉ tập trung nghiên cứu phương pháp phát hiện xâm nhập này. Cụ thể, chúng tôi sẽ trình bầy và so sánh bằng thực nghiệm hai phương pháp phát hiện bất thường là phương pháp phân tích tương quan giữa các thuộc tính [3] và phương pháp phát hiện mẫu mới (novelty detection) sử dụng hàm nhân [1]. Phân tích thương quan giữa các thuộc tính là phương pháp được đề xuất riêng cho dữ liệu kiểm tra của mạng không dây ad-hoc trong khi khi phương pháp phát hiện mẫu mới là phương pháp tổng quát được đề xuất để phát hiện bất thường cho các loại dữ liệu khác nhau và chưa được thử nghiệm cho bài toán phát hiện xâm nhập mạng.

3.1. Phân tích tương quan giữa các thuộc tính

Phương pháp phân tích tương quan giữa các thuộc tính dựa trên giả thiết: trong điều kiện bình thường, các thông số hay thuộc tính khác nhau của mạng không dây ad-hoc tương quan với nhau. Chẳng hạn, do khoảng cách và tốc độ di chuyển nút mạng có ảnh hưởng đến chất lượng truyền tin, trong điều kiện bình thường, tỷ lệ gói tin bị mất tỷ lệ thuận với mức độ thay đổi tô pô mạng. Các tương quan như vậy cho phép dự đoán giá trị của từng thông số dựa trên giá trị các thông số khác. Giá trị dự đoán sau đó được so sánh với giá trị quan sát được. Nếu giá trị dự đoán của thông số khác với giá trị thực quan sát được thì đây có thể là dấu hiệu bất thường liên quan tới xâm nhập mạng.

Giả sử dữ liệu kiểm tra bao gồm n thuộc tính $\{f_1, f_2, \ldots f_n\}$. Giá trị mỗi thuộc tính f_i được dự đoán từ các thuộc tính còn lại $\ldots f_{i-1}, \ldots f_{i+1}, \ldots f_n\}$ bằng cách sử dụng phu pháp học máy và phân loại tự động. Nếu thuộc tính có giá trị rời rạc, mỗi giá trị c được coi như một nhãn phân loại, nếu thuộc tính liên tục, f_i sẽ được rời rạc ho trở thành thuộc tính rời rạc (có thể trực tiế đoán giá trị các thuộc tính liên tục, khi đ có bài toán *regression* thay vì *classificat* tuy nhiên phương pháp dự đoán giá trị rời cho kết quả tốt hơn và sẽ được sử dụn đây). Bộ phân loại C_i là một ánh xạ

$$C_i: \{f_1, \ldots f_{i-1}, \ldots f_{i+1}, \ldots f_n\} \rightarrow f_i$$

Bộ phân loại C_i được huấn luyện trên liệu kiểm tra của mạng. Đây là dữ liệu k tra được ghi lại từ trước trong các tình hu không có xâm nhập trái phép.

Việc phát hiện bất thường được thực l như sau. Trước hết, sử dụng bộ phân loại để dự đoán giá trị f_i từ giá trị các thuộc còn lại. Sau đó, so sánh kết quả tính được giá trị thực tế. Nếu tổng số thuộc tính có trị dự đoán trùng với giá trị thực nhỏ hơn ngưỡng θ cho trước, tức là nếu

$$\sum_{i=1}^{n} \delta(C_i(x), f_i(x)) < \theta, \quad \text{trong}$$

$\delta(a,b) = 1$ nếu $a = b$ và $\delta(a,b) = 0$ nếu a :

thì đưa ra cảnh báo về khả năng mạn xâm nhập trái phép.

Một vấn đề có ảnh hưởng lớn tới kết dự đoán là lựa chọn phương pháp học máy phân loại phù hợp. Trong các thực ngh trình bầy ở đây, chúng tôi lựa chọn phương pháp học máy là *cây quyết định* C [10] và *support vector machine* (*SVM*) Cây quyết định là phương pháp học máy biểu và tương đối đơn giản trong khi SVM phương pháp ra đời sau và cho kết quả p loại tốt hơn các phương pháp khác trong nh ứng dụng khác nhau.

Cây quyết định C4.5. C4.5 là phư pháp biểu diễn hàm phân loại dưới dạng quyết định. Quá trình huấn luyện là quá tr xây dựng cây quyết định cho phép phân tốt nhất các ví dụ huấn luyện. Tại mỗi nút

cây, C4.5 phân chia các ví dụ huấn luyện thành hai nút con bằng cách so sánh giá trị một thuộc tính. Thuộc tính được chọn là thuộc tính cho phép tạo ra hai nút con có entropy thông tin nhỏ nhất.

Support vector machine. SVM xây dựng hàm phân loại dưới dạng một siêu phẳng trong không gian d chiều với d là kích thước của vectơ dữ liệu ($d = n$-1 trong bài toán đang xét). Giả sử cho m ví dụ huấn luyện $(\mathbf{x}_i, y_i) \in \Re^d \times \{\pm 1\}$, $i = 1,\ldots,m$, tương ứng với m điểm trong không gian Đề Các d chiều. SVM sử dụng hàm phân loại dưới dạng siêu phẳng

$$(\mathbf{w}.\mathbf{x}) + b = 0, \mathbf{w} \in \Re^d, b \in \Re \quad (1)$$

Các trường hợp nằm bên dưới siêu phẳng này sẽ có phân loại -1, các trường hợp nằm trên có phân loại +1. Nói cách khác, hàm phân loại được sử dụng là hàm

$$f(\mathbf{x}) = \text{sgn}((\mathbf{w}.\mathbf{x}) + b) \quad (2)$$

Gọi tổng khoảng cách từ siêu phẳng cho bởi (1) tới các ví dụ huấn luyện với nhãn +1(-1) gần nhất là "*lề*" (*margin*). SVM lựa chọn hàm phân loại tương ứng với siêu phẳng có lề cực đại. Sử dụng lý thuyết học máy có thể chứng minh siêu phẳng với lề cực đại cho kết quả phân loại tốt nhất với các trường hợp phân loại mới.

Sử dụng các biến đổi của hình học giải tích và phương pháp Lagrange, siêu phẳng tối ưu được xây dựng bằng cách giải bài toán tối ưu hoá sau:

Tìm số nhân Lagrange α_i cho phép cực đại hoá hàm

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i.\mathbf{x}_j) \quad (3)$$

thoả mãn điều kiện $\alpha_i \geq 0$, $i = 1,\ldots,m$ và

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \quad (4)$$

Giá trị $\mathbf{w}$ và b sau đó được tính theo các biểu thức sau

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \quad (5)$$

$$\alpha_i [y_i((\mathbf{w}.\mathbf{x}) + b) - 1] = 0 \quad (6)$$

Việc xây dựng hàm phân loại như trên chỉ có thể thực hiện trong trường hợp tồn tại hàm tuyến tính (siêu phẳng) phân chia ví dụ với nhãn khác nhau. Tuy nhiên, trên thực tế, đa số hàm phân loại là phi tuyến. Để giải quyết vấn đề này, thay vì tìm siêu phẳng (1) trong không gian gốc, trước hết dữ liệu được ánh xạ sang không gian (nhiều chiều hơn) H bởi phép ánh xạ

$$\Phi : \Re^d \to \mathrm{H}$$
$$\mathbf{x}_i \to \Phi(\mathbf{x}_i) \quad (7)$$

Do biểu thức (3) chứa tích vô hướng $\mathbf{x}_i.\mathbf{x}_j$, thay vì tính ánh xạ $\Phi(\mathbf{x}_i)$ của $\mathbf{x}i$ trong H, ta có thể tìm hàm K sao cho $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j)$. Hàm K như vậy được gọi là "*hàm nhân*" (*kernel function*). Các hàm nhân đóng vai trò quan trọng vì sử dụng hàm này cho phép tránh việc tính toán trực tiếp trong không gian H. Biểu thức (3) khi đó được viết lại như sau:

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \quad (8)$$

Quá trình huấn luyện bộ phân loại SVM được thực hiện bằng cách giải bài toán tối ưu hoá (8) với các ràng buộc (4).

Do SVM chỉ có thể phân chia các trường hợp thành hai lớp, đối với bài toán phân loại nhiều lớp như trong trường hợp các thuộc tính, nhiều bộ phân loại SVM sẽ được sử dụng cho một thuộc tính, mỗi bộ phân loại cho phép dự đoán thuộc tính có một giá trị cụ thể nào đó hay có các giá trị khác.

3.2. Phát hiện mẫu mới

Phát hiện mẫu mới (novelty detection) là bài toán xác định dữ liệu hay tín hiệu khác với dữ liệu bình thường mà hệ thống đã gặp trong quá trình huấn luyện. Đây là lớp bài toán có nhiều ứng dụng trong thực tế. Chẳng hạn phương pháp phát hiện mẫu mới có thể sử dụng trong chẩn đoán lỗi bằng cách phát hiện hoạt động khác với bình thường của động cơ hay các thiết bị khác. Do giá trị thực tế của bài toán này, nhiều phương pháp phát hiện mẫu

mới đã được đề xuất và thử nghiệm. Tổng quan về các phương pháp này có thể xem tại [4,5].

Trong bài báo này, chúng tôi lựa chọn thử nghiệm phương pháp phát hiện mẫu mới sử dụng hàm nhân kết hợp với quy hoạch tuyến tính [1] cho bài toán phát hiện xâm nhập mạng không dây. Phương pháp này được lựa chọn do có một số ưu điểm: quá trình huấn luyện đòi hỏi tính toán ít, chỉ phụ thuộc và một tham số duy nhất.

Cho dữ liệu huấn luyện $\mathbf{x}_i \in \Re^d$, $i = 1,\ldots,m$. Ý tưởng cơ bản của phương pháp là xây dựng một hàm phân loại nhị phân sao cho hàm này nhận giá trị dương trong vùng phân bố của dữ liệu huấn luyện và nhận giá trị âm ở ngoài vùng này. Thông thường, hàm phân loại như vậy không phải là hàm tuyến tính. Để sử dụng hàm tuyến tính, dữ liệu được ánh xạ sang không gian H tương tự như trong kỹ thuật SVM trình bầy ở trên $\mathbf{x}_i \rightarrow \Phi(\mathbf{x}_i)$. Để tránh phải tính toán với ánh xạ của $\mathbf{x}_i$, ta có thể sử dụng hàm nhân $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j)$ tương tự như phần trên.

Trong không gian gốc $\Re^d$ hàm phân loại cần tìm có dạng một mặt bao bọc lấy các điểm tương ứng với dữ liệu huấn luyện. Trong không gian ánh xạ H, hàm phân loại sẽ trở thành siêu phẳng $f(\mathbf{z}) = \Sigma_i\ \alpha_i\ K(\mathbf{z},\mathbf{x}_i) + b$. Phương pháp phát hiện mẫu mới trình bầy ở đây lựa chọn siêu phẳng gần các điểm dữ liệu nhất sao cho toàn bộ các điểm dữ liệu nằm về một phía hoặc nằm ngay trên bề mặt siêu phẳng. Siêu phẳng tối ưu này có thể tìm bằng cách cực tiểu hoá hàm

$$W(\alpha,b) = \sum_{i=1}^{m}\left(\sum_{j=1}^{m}\alpha_i K(\mathbf{x}_i,\mathbf{x}_j) + b\right)$$

với ràng buộc $\sum_{j=1}^{m}\alpha_i K(\mathbf{x}_i,\mathbf{x}_j) + b \geq 0$

và $\sum_{i=1}^{m}\alpha_i = 1, \alpha_i \geq 0$ (11)

Đây là bài toán quy hoạch tuyến tính truyền thống và có thể giải bằng phương pháp đơn hình hoặc các phương pháp quy hoạch tuyến tính khác. Ưu điểm cơ bản của phương pháp này là quá trình huấn luyện bằng cách giải bài toán quy hoạch tuyến tính đòi hỏi t toán ít hơn nhiều so với giải bài toán hoạch bậc hai (8) trình bầy ở phần trước.

Sau khi huấn luyện, ví dụ mới $\mathbf{x}$ sẽ đ phân loại bởi hàm

$$f(\mathbf{x}) = \text{sgn}(\sum_{i=1}^{m}\alpha_i K(\mathbf{x}_i,x) + b) \quad (1$$

Nếu $f(\mathbf{x}) < 0$, $\mathbf{x}$ sẽ được coi là mẫu mới là dấu hiệu xâm nhập mạng trong bài t đang xét.

Trong các thực nghiệm trình bầy ở đ chúng tôi sử dụng hàm nhân *RBF*

$$K(\mathbf{x}_i,\mathbf{x}_j) = e^{-(\|\mathbf{x}_i-\mathbf{x}_j\|)^2/2\sigma^2} \quad (1$$

Đây là hàm nhân cho phép tạo ra hàm phân loại bao kín các ví dụ huấn luyện tr không gian gốc. Thuật toán huấn luyện khi có một tham số duy nhất là σ. Giá trị tham này sẽ được chọn bằng thực nghiệm.

4. THỬ NGHIỆM

Trong phần này,chúng tôi trình bầy quả thử nghiệm và so sánh hai phương p phát hiện xâm nhập trình bầy ở phần trên c dữ liệu mô phỏng.

3.1. Dữ liệu mô phỏng

Dữ liệu thực nghiệm được sinh ra từ công cụ mô phỏng mạng ns-2 [14]. ns-2 công cụ mô phỏng mạng được sử dụng rộ rãi trong các nghiên cứu cả về mạng có d cũng như không dây.

Dữ liệu huấn luyện được tạo ra trong t gian 20000 giây. Hai kịch bản chứa các t huống tấn công được tạo ra trong thời g 7000 giây. Giao thức định tuyến là *Ad-h On-demand Distance Vector* (*AODV*) G thức tầng vận chuyển được chọn UDP/CBR(*constant bit rate*) và TCP, các c nút-nguồn được trải ngẫu nhiên trên toàn diện tích vật lý của mạng. Kích thước gói t được cố định là 512B. Tô pô mạng được t ra bằng cách sử dụng mô hình chuyển độ *random waypoint* trên một vùng hình vuô kích thước 670 x 670 m với 20 nút mạng. C nút di chuyển giữa các vị trí với tốc độ l

họn ngẫu nhiên trong khoảng 0-10m/giây. hời gian tạm dừng giữa các lần di chuyển là 0 giây. Đây là các tham số thường gặp trong ạng ad-hoc và là tham số mặc định trong một số kịch bản mẫu đi kèm các môđun của ns-2. Chu kỳ lấy mẫu và ghi lại dữ liệu kiểm ra là 5 giây.

Có thể giả định nhiều kiểu tấn công khác nhau. Các dạng tấn công thông thường là: 1) làm sai lệch thông tin định tuyến với các hậu quả như tạo ra các hố đen trong mạng hay tạo ra chu trình trong các bảng định tuyến; 2) thay đổi thông tin lưu lượng với các hậu quả như làm mất gói tin, tấn công từ chối dịch vụ.

Trong thực nghiệm của chúng tôi, tình huống tấn công giả định được sinh ra bằng cách sử dụng lớp `ErrorModel` của ns-2 với giả thiết có tất cả 4 nút mạng bị xâm nhập và là nguyên nhân gây mất gói tin. Tỷ lệ gói tin bị mất dao động từ 0.0 (0%) đến 1.0 (100%).

3.2. Lựa chọn thuộc tính

Các thuộc tính được lựa chọn như khuyến cáo trong [11] và đặc trưng cho ba thông số chính của mạng là lưu lượng, thông tin định tuyến và tô pô mạng. Tên viết tắt và ý nghĩa các thuộc tính sử dụng trong thực nghiệm được cho.

Trong các thuộc tính liệt kê trong bảng 1, thuộc tính đầu tiên đặc trưng cho lưu lượng, hai thuộc tính tiếp theo đặc trưng cho thông tin định tuyến, các thuộc tính còn lại đặc trưng cho chuyển động của nút hay tô pô mạng.

Bảng 1. Các thuộc tính sử dụng trong thử nghiệm

Tên thuộc tính	Ý nghĩa
PSTC	% thay đổi lưu lượng
PCR	% thay đổi trong bảng định tuyến
PCH	% thay đổi số lượng *hop*
VELOCITY	tốc độ di chuyển của nút
RDC	độ thay đổi khoảng cách tương đối
DISTANCE	khoảng cách tới vị trí lấy mẫu lần cuối

Để sử dụng làm dữ liệu phân loại và kiểm tra, các thuộc tính được rời rạc hoá. Giá trị của PSTC, PCR, PCH, RDC được chia thành sáu khoảng bằng nhau, mỗi khoảng tương ứng với một nhãn. Giá trị VELOCITY và DISTANCE được rời rạc hoá thành 10 giá trị ứng với 10 khoảng bằng nhau.

3.3. Thuật toán phân loại

Thuật toán học cây quyết định được lấy từ bộ công cụ Weka [12] với thông số mặc định. Phương pháp SVM sử dụng phiên bản SVM light của Joachim [13]. Các hàm nhân được thử nghiệm là hàm đa thức

$$K(x_i, x_j) = (\mathbf{x}_i.\mathbf{x}_j)^d \ , d > 1$$

(trong thử nghiệm chúng tôi sử dụng hàm nhân đa thức với d = 1, 2, 3), và hàm RBF (hay Gausse)

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-(|\mathbf{x}_i - \mathbf{x}_j|)^2 / 2\sigma^2}$$

Kết quả thử nghiệm cho thấy hàm nhân đa thức cho kết quả phát hiện xâm nhập kém hơn hàm nhân RBF và do vậy sẽ không được trình bầy ở đây.

Đối với phương pháp phân tích tương quan thuộc tính, ngưỡng phát hiện cảnh báo θ được thay đổi trong khoảng 0.8-1 với khoảng cách là 0.5. Tỷ lệ cảnh báo chính xác và cảnh báo sai được ghi lại cho mỗi giá trị của θ .

Đối với phương pháp phát hiện mẫu mới trình bầy trong phần 2.2, bài toán quy hoạch tuyến tính được giải bằng phương pháp simplex. Hàm nhân sử dụng là hàm RBF với các giá trị σ khác nhau.

Kết quả phát hiện xâm nhập được đánh giá theo hai tiêu chí: *tỷ lệ phát hiện xâm nhập* và *độ chính xác*. Hai tiêu chí này được định nghĩa như sau.

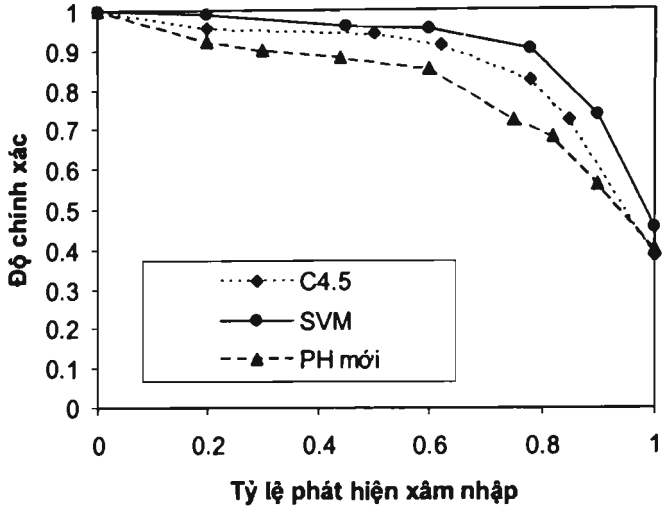
$$\text{Tỷ lệ phát hiện xâm nhập} = \frac{\text{Số lượng xâm nhập phát hiện đúng}}{\text{Số lượng xâm nhập thực tế}}$$

$$\text{Độ chính xác} = \frac{\text{Số lượng xâm nhập phát hiện đúng}}{\text{Số lượng xâm nhập phát hiện được}}$$

Trong định nghĩa trên, số lượng xâm nhập được tính bằng tổng số lượng mẫu lấy trong thời gian xảy ra xâm nhập chứ không phải số lượng phiên xâm nhập. Hai tiêu chí trên có giá trị dao động từ 0 tới 1.

Các giá trị ngưỡng θ và tham số hàm nhân σ khác nhau cho kết quả với tỷ lệ phát hiện xâm nhập và độ chính xác khác nhau. Trên hình 2 là biểu đồ giá trị hai tiêu chí đánh giá cho phương pháp phân tính tương quan thuộc tính sử dụng cây quyết định C4.5, sử SVM và phương pháp phát hiện mới trình trong phần 3.2. Trục hoành biểu thị tỷ lệ hiện xâm nhập, trục tung là độ chính xác giá trị này càng cao, tức là các điểm của c nằm càng gần góc trên bên phải thì càng t

Đồ thì trên hình 2 cho thấy với gi tham số được lựa chọn phù hợp, các phi pháp được thử nghiệm đều cho kết quả hiện xâm nhập trái phép tương đối tốt.



Hình 2. Biểu đồ giá trị recall và precision cho phương pháp phân tích tương quan giữa c thuộc tính sử dụng cây quyết định C4.5, SVM và phương pháp phát hiện mới sử dụng hàm n và quy hoạch tuyến tính

Trong hai phương pháp trình bầy ở phần trên, phương pháp phân tích tương quan thuộc tính cho kết quả tốt hơn phương pháp phát hiện mới. Có thể giải thích kết quả tốt hơn của phương pháp phân tích thuộc tính là do phương pháp này đã tính tới đặc thù của mạng ad-hoc, trong đó một số thông số hoạt động của mạng có tương quan với nhau trong điều kiện bình thường. Cũng như trong nhiều ứng dụng khác, bộ phân loại SVM cho kết quả tốt hơn cây quyết định. Tuy nhiên, để đạt được kết quả tốt với SVM, vấn đề quan trọng là lựa chọn hàm nhân phù hợp. Hàm nhân có thể lựa chọn qua thực nghiệm, chẳng hạn bằng phương pháp kiểm tra chéo (cross-validation). Tương tự như vậy, giá trị các tham số θ và σ sử dụng ở trên cũng được lựa chọn cho trường hợp cụ thể bằng cách kiểm tra chéo trên dữ liệu mẫu. Giá trị được chọn là giá trị kết quả kiểm tra chéo tốt nhất.

Vấn đề lựa chọn hàm nhân và tham số hàm nhân cũng có giá trị quan trọng đối phương pháp phát hiện mới. Như các tác của phương pháp này đã nhận xét, giá trị th số σ chỉ có thể lựa chọn chính xác tr trường hợp có đủ dữ liệu huấn luyện và k tra, hoặc có thêm thông tin về bài toán phép lựa chọn giá trị σ phù hợp.

5. KẾT LUẬN

Bài báo đã phân tính sự quan trọng việc phát hiện xâm nhập trái phép trong m không dây ad-hoc và các đặc điểm của d

mạng này có liên quan tới vấn đề phát hiện xâm nhập. Các phân tích cho thấy hệ thống phát hiện xâm nhập trái phép cho mạng ad-hoc nên được xây dựng phân tán dưới dạng các tác tử với cơ chế ra quyết định dựa trên thông tin cục bộ đồng thời với khả năng trao đổi kết quả với tác tử khác. Hai phương pháp phát hiện xâm nhập trái phép cho tác tử đã được trình bầy và so sánh bằng thực nghiệm. Cả hai phương pháp đều dựa trên nguyên tắc phát hiện tình trạng bất thường trong hệ thống. Phương pháp đầu tiên phát hiện bất thường khi tương quan giữa các thông số mạng bị phá vỡ, phương pháp thứ hai là phương pháp phát hiện mẫu mới cho phép phát hiện các tình huống khác với tình huống đã gặp. Dữ liệu thử nghiệm được sinh ra từ công cụ mô phỏng ns-2 với các tình huống tấn công giả định. Kết quả thử nghiệm cho thấy khả năng phát hiện xâm nhập của cả hai phương pháp là tương đối tốt, trong đó phương pháp thứ nhất có nhiều ưu điểm hơn. Trong nghiên cứu này, chúng tôi chưa thử nghiệm được kết quả phát hiện xâm nhập cho trường hợp các tác tử có thêm khả năng cộng tác trao đổi thông tin với nhau. Phương thức trao đổi cho tác tử, cách ra quyết định khi có thêm thông tin từ tác tử khác và kết quả thực nghiệm sẽ là nội dung của các nghiên cứu tiếp theo.

LỜI CẢM ƠN

Nghiên cứu được thực hiện với sự hỗ trợ kinh phí của đề tài NCCB do Bộ Khoa học và Công nghệ tài trợ.

Tài liệu tham khảo

1. C. Campbell, K.P. Bennett. A linear programming approach to noveltydetection. *Advances in NIPS, Vol. 14*, MIT Press, Cambridge, MA, 2001.
2. C. Cortes and V. Vapnik. Support vector networks. *Machine learning*. 20, pp:273-297, 1995.
3. Y. Huang, W. Fan, W. Lee, and P. Yu. Cross-feature analysis for detecting ad-hoc routing anomalies. *In Proceedings of the 23rd International Conference on Distributed Computing Systems*, Providence, RI, May 2003.
4. M. Markou, S. Singh. Novelty detection: a review-part 1: statistical approaches. *Signal processing*. 83, pp: 2481-2497, Elsevier 2003.
5. M. Markou, S. Singh. Novelty detection: a review-part 2: neural network based approaches. *Signal processing*. 83, pp: 2499-2521, Elsevier 2003.
6. C. E. Perkins. Ad hoc networking: An introduction. *In C. E. Perkins, editor, Ad Hoc Networking*. Addison-Wesley, 2000.
7. A. Scholkopf, R. Williamson, A. Smola, J.S. Taylor, J. Platt, Support vector method for noveltydetection, *in: S.A. Solla, T.K. Leen, K.R. MSuller (Eds.), Neural Information Processing Systems*, Elsevier, New York, 2000, pp. 582–588.
8. I. Steinwart, D. Hush, C. Scovel. A classification framework for anomaly detection. *Journal of machine learning research*, 6: 211-232, 2005.
9. Nguyễn Minh Tuấn. Phát hiện xâm nhập trái phép mạng không dây sử dụng kỹ thuật trí tuệ nhân tạo. *Luận văn cao học*. Học viện Bưu chính viễn thông, 2005.
10. J. R. Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann, San Mateo, CA, 1993.
11. Y. Zhang, W. Lee, Y. Huang. Intrusion Detection Techniques for Mobile Wireless Networks. *Wireless networks*, 9, pp: 545-556, Kluwer academics, 2003.
12. I. Witten, E. Frank. *Data mining*. Morgan Kaufman, 2000.
13. SVM Light website: http://svmlight.joachims.org/
14. ns-2 website: http://www.isi.edu/nsnam/ns/

Về các tác giả

Tiến sĩ **Từ Minh Phương** tốt nghiệp đại học tại trường Bách khoa Taskent năm 1993, bảo vệ tiến sỹ tại Viện hàn lâm khoa học Uzbekistant, Taskent, năm 1995.

Từ năm 2000 đến nay công tác tại khoa CNTT, Học viện Công nghệ Bưu chính Viễn thông. Hiện là

Q. Trưởng khoa, khoa CNTT Học viện Công nghệ Bưu chính Viễn thông.

Hướng nghiên cứu: học máy, hệ tác tử, logic mờ, tin sinh học

E-mail: phuongtm@fpt.com.vn

Thạc sĩ **Nguyễn Minh Tuấn** tốt nghiệp đại học Giao thông vận tải chuyên ngành kỹ thuật viễn thông năm 1997, nhận học vị thạc sỹ khoa học chuyên ngành điện tử viễn thông tại Học viện Công nghệ bưu chính viễn thông năm 2005.

Hiện đang công tác tại Trung tâm tin học, Công t Viễn thông, Bưu điện tỉnh Hà tây.

Hướng nghiên cứu: bảo mật cho mạng máy tính.

E-mail: minhtuantu@hn.vnn.vn