Một phương pháp trích rút đặc trưng tổng quát hóa trong phát hiện thâm nhập mạng dựa trên bất thường

Nguyễn Xuân Nam, Nguyễn Đại Thọ*

Trường Đại học Công nghệ
Đại học Quốc gia Hà Nội
* UMI UMMISCO 209 (IRD/UPMC), Hà Nội
nguyendaitho@vnu.edu.vn

Tóm tắt— Bài báo này đề xuất một phương pháp trích rút đặc trưng trong phân lớp dữ liệu tổng quát hơn so với một hệ thống phát hiện thâm nhập mạng dựa trên bất thường có tên là McPAD đã được đề xuất trước đây. Vấn đề đặt ra là số đặc trưng được trích rút tăng lên có thể làm tăng độ phức tạp và giảm độ chính xác của phương pháp phân lớp tổng quát hóa. Vì vậy, phương pháp lựa chọn đặc trưng Chi-Square được sử dụng để lọc ra chỉ những đặc trưng tốt nhất. Chúng tôi đã làm nhiều thí nghiệm với bộ dữ liệu là các tấn công HTTP thật để đánh giá hiệu năng của phương pháp trích rút đặc trưng tổng quát hóa kết hợp lọc Chi-Square. Kết quả cho thấy bộ phân lớp của chúng tôi có thể nhanh chóng phát hiện các gói tin tấn công với tỷ lệ dương tính thật rất cao trong khi duy trì tỷ lệ dương tính giả ở mức rất thấp. Ngoài ra, kết quả cũng cho thấy hiệu năng bộ phân lớp của chúng tôi vượt trội hơn so với McPAD.

I. GIỚI THIỆU

A. *Các hệ thống phát hiện thâm nhập*

Một trong những hiểm họa lớn nhất đối với an toàn an ninh các mạng truyền thông và cơ sở hạ tầng thông tin là các tấn công thâm nhập trái phép. Các hệ thống phát hiện thâm nhập mạng thường sử dụng một trong hai phương pháp sau đây để phân tích các gói tin phát hiện ra các hành vi thâm nhập: Phát hiện dựa trên chữ ký và phát hiện dựa trên bất thường. Phương pháp dựa trên chữ ký phát hiện các gói tin độc hại dựa theo các mẫu có sẵn trong cơ sở dữ liệu. Phương pháp dựa trên bất thường thu thập dữ liệu về các hành vi hợp lệ, bình thường để xây dựng mô hình phản ánh các hành vi này, trên cơ sở đó bất kỳ hành vi nào sai lệch so với mô hình sẽ bị coi là tấn công.

Các hệ thống phát hiện thâm nhập mạng trong thực tế thường sử dụng phương pháp dựa trên chữ ký vì nó cho phép phát hiện một cách hiệu quả các biểu hiện tấn công đã biết trong khi duy trì tỷ lệ dương tính giả (phát hiện sai) ở mức thấp. Tuy nhiên, chúng không thể phát hiện ra các hành vi thâm nhập mới chưa từng gặp, vì chưa có mẫu của các loại tấn công đó.

Ngược lại, các hệ thống phát hiện thâm nhập dựa trên bất thường có khả năng phát hiện được nhiều loại tấn công, kể cả các hành vi tấn công mới hoàn toàn. Nhược điểm của chúng là thường mắc tỷ lệ dương tính giả khá lớn. Vì vậy mục đích của các nghiên cứu phát hiện thâm nhập dựa trên bất thường là duy trì tỷ lệ dương tính thật (phát hiện đúng) ở mức cao trong khi kiềm chế tỷ lệ dương tính giả ở mức càng thấp càng tốt.

Để giải quyết bài toán đã nêu, các tác giả của hệ thống phát hiện thâm nhập dựa trên bất thường PAYL [13, 14] đề xuất một phương pháp trích rút đặc trưng từ phần nội dung (payload) của các gói tin. Các thông tin đặc trưng sẽ được sử dụng để đánh giá một gói tin có phải hợp lệ hay không. PAYL coi mỗi phần nội dung gói tin như một chuỗi các byte, không cần quan tâm đến ngữ nghĩa ở tầng ứng dụng. Để trích rút các đặc trưng từ một chuỗi byte, mỗi một trong số 256^n loại n-gram (n byte liên tiếp nhau) khác nhau được đếm số lần xuất hiện trong chuỗi, số lần xuất hiện của mỗi loại n-gram được lưu trong một chiều của véc tơ đặc trưng có 256^n chiều của chuỗi byte. Các tác giả đã làm thí nghiệm với n = 2, và bộ phân loại của họ khá là chính xác, nhưng tỷ lệ dương tính giả vẫn rất cao. Vấn đề của đề xuất này là với giá trị n nhỏ ($n \leq 2$), véc tơ đặc trưng không đủ thông tin để đại diện cho một phần nội dung gói tin, trong khi các giá trị cao hơn của n có thể làm số chiều tăng lên rất lớn, làm tăng một cách đáng kể độ phức tạp cũng như thời gian xử lý [4].

Để khắc phục hạn chế của PAYL, công trình [11] đề xuất một phiên bản cải tiến của phương pháp trích rút n-gram, sử dụng các 2_v-gram (2 byte cách nhau v vị trí) thay cho các 2-gram. Với mỗi phần nội dung gói tin, K véc tơ đặc trưng được tính, mỗi véc tơ tương ứng với một giá trị của v, v biến thiên từ 0 đến K - 1. Sau đó, K véc tơ này được dùng để huấn luyện K bộ phân lớp, mỗi bộ phân lớp được huấn luyện bằng một véc tơ tương ứng cho ra kết quả là xác suất đánh giá một gói tin là bình thường. Đầu ra của K bộ phân lớp được kết hợp bằng một luật như lấy max, min, hay trung bình để cho ra kết quả xác suất cuối cùng, dựa vào kết quả này ta có thể phân định gói tin đang kiểm tra là bình thường

hay tấn công. Toàn bộ hệ thống phát hiện thâm nhập dựa trên bất thường sử dụng kết hợp nhiều bộ phân lớp được các tác giả đặt tên là McPAD. Với việc sử dụng các 2_v-gram thay vì chỉ các 2-gram, McPAD đã cố gắng thu thập thêm nhiều thông tin đặc trưng hơn từ phần nội dung gói tin. Tuy nhiên, do mỗi véc tơ 2_v-gram chỉ được dùng để huấn luyện một bộ phân lớp nên mỗi bộ phân lớp không chứa đủ nhiều thông tin phản ánh cấu trúc phần nội dung gói tin. Hiệu quả phân loại của từng bộ phân lớp vốn đã thấp khiến cho ngay cả khi kết hợp nhiều bộ phân lớp với nhau, hiệu quả chung vẫn chỉ dừng ở mức thấp.

B. Đóng góp của chúng tôi

Do vấn đề phân lớp các gói tin khá tương đồng với vấn đề nhận biết ngôn ngữ tự nhiên, trong đó các byte trong phần nội dung gói tin tương ứng với các ký tự trong một văn bản, nên chúng tôi đã tiến hành một thí nghiệm nhỏ để xác định xem số lần xuất hiện của các cặp ký tự cách nhau một khoảng nhất định có tuân theo một quy luật nào hay không.

Ba văn bản tiếng Anh thuần túy đủ dài được chọn ngẫu nhiên trên mạng làm đầu vào của phép thống kê là *VB1*[1], *VB2*[2], và *VB3*[3]. Các bảng 1, 2, và 3 dưới đây thống kê các cặp ký tự cách nhau *v* vị trí xuất hiện nhiều nhất trong mỗi văn bản, trong đó *v* biến thiên từ 0 đến 7.

Bảng 1: Top 5 cặp ký tự xuất hiện nhiều nhất trong văn bản VB1

v	Top 5 cặp ký tự và tỷ lệ phần trăm xuất hiện nhiều nhất				
0	***er*** **1,31%**	***th*** **1,15%**	*in* 1,10%	***he*** **1,06%**	*re* 1,05%
1	***et*** **1,02%**	***te*** **0,89%**	*ae* 0,70%	*es* 0,67%	*ai* 0,67%
2	***ee*** **0,89%**	***tn*** **0,85%**	*rt* 0,67%	*ei* 0,67%	*eo* 0,57%
3	***ee*** **0,69%**	*re* 0,63%	*en* 0,62%	*et* 0,60%	*eo* 0,54%
4	***ee*** **0,80%**	*en* 0,56%	*ei* 0,55%	*ne* 0,55%	***et*** **0,55%**
5	***ee*** **0,94%**	***et*** **0,51%**	*he* 0,48%	*eo* 0,47%	*ar* 0,47%
6	***ee*** **0,64%**	***et*** **0,51%**	*en* 0,50%	*ei* 0,48%	*ae* 0,47%
7	***ee*** **0,55%**	*eo* 0,54%	***te*** **0,5%**	*ea* 0,49%	*en* 0,44%

Bảng 2: Top 5 cặp ký tự xuất hiện nhiều nhát trong văn bản VB2

v	Top 5 cặp ký tự và tỷ lệ phần trăm xuất hiện nhiều nhất				
0	***he*** **2.00%**	***th*** **2.09%**	*in* 1,33%	***er*** **1,22%**	*an* 1,00%
1	***te*** **1,67%**	*ad* 0,96%	***et*** **0,87%**	*ie* 0,70%	*ig* 0,63%
2	*ea* 0,87%	*eo* 0,77%	***ee*** **0,76%**	***tn*** **0,62%**	*ei* 0,61%
3	***eo*** **0,71%**	*te* 0,68%	*at* 0,65%	***et*** **0,65%**	***ee*** **0,62%**
4	*ee* 0,85%	*et* 0,74%	*te* 0,58%	*ta* 0,52%	*en* 0,49%
5	***ee*** **0,85%**	*ae* 0,64%	***et*** **0,63%**	*te* 0,61%	*oe* 0,53%
6	***ee*** **0,84%**	*te* 0,69%	***et*** **0,63%**	*he* 0,58%	*eo* 0,51%
7	***ee*** **0,76%**	***te*** **0,74%**	*et* 0,60%	*ae* 0,55%	*tt* 0,52%

Bảng 3: Top 5 cặp ký tự xuất hiện nhiều nhát trong văn bản VB3

v	Top 5 ký tự và tỷ lệ phần trăm xuất hiện nhiều nhất				
0	***er*** **1,46%**	***th*** **1,22%**	***he*** **1,19%**	*on* 1,09%	*re* 1,08%
1	***te*** **0,98%**	*to* 0,91%	***et*** **0,81%**	*ei* 0,72%	*ie* 0,72%

[1] https://tools.ietf.org/html/rfc3987

[2] https://www.gutenberg.org/files/11/11.txt

[3] https://tools.ietf.org/html/rfc4006

2	***ei*** 0,97%	***ee*** 0,93%	*et* 0,77%	***tn*** 0,67%	*er* 0,59%
3	*er* 0,76%	***ee*** **0,74%**	*re* 0,73%	*rt* 0,62%	*se* 0,59%
4	***ee*** **1,12%**	*tt* 0,56%	*in* 0,54%	***et*** **0,53%**	*te* 0,52%
5	***ee*** **0,74%**	*te* 0,68%	*tr* 0,63%	***et*** **0,59%**	*ei* 0,56%
6	***ee*** **0,69%**	*en* 0,65%	*ei* 0,60%	***et*** **0,60%**	*ro* 0,52%
7	*et* 0,90%	***ee*** **0,87%**	*rn* 0,58%	*ti* 0,58%	***te*** **0,57%**

Trong các bảng trên, tần suất xuất hiện của một cặp ký tự tính theo tỷ lệ phần trăm trên tổng số các cặp ký tự cách nhau tương ứng *v* vị trí trong toàn bộ văn bản (bằng *n*-*v*-1 nếu văn bản chứa *n* ký tự) được chỉ ra ngay trong ô của cặp ký tự tương ứng. Không chỉ các chữ cái tiếng Anh mà bất kỳ ký tự đọc được nào trong mỗi văn bản cũng được đưa vào các thống kê.

Có thể nhận thấy tương ứng với *v* = 0 các cặp ký tự *er*, *th*, và *he* luôn nằm trong nhóm các cặp ký tự xuất hiện nhiều nhất trong tất cả các văn bản. Khi v = 1, các cặp *et* và *te* xuất hiện nhiều nhất. In đậm là các cặp ký tự luôn nằm trong top 5 cặp ký tự xuất hiện nhiều nhất trong tất cả các văn bản. Ở khoảng cách rất xa (*v* = 7), các cặp ký tự như *ee* và *te* vẫn luôn nằm trong nhóm đứng đầu về số lần xuất hiện trong cả ba văn bản. Điều này nói nên rằng, dù dãn cách giữa hai ký tự trong cùng một cặp dài hay ngắn, số lần xuất hiện của các cặp ký tự vẫn có giá trị, và ta có thể sử dụng chúng là các đặc trưng cho các văn bản tiếng Anh bình thường. Do đó, tần số xuất hiện của các cặp byte cách nhau *v* vị trí phần nội dung gói tin có thể được dùng làm đặc trưng cho mỗi gói tin. Chúng tôi tin rằng nên quan tâm đến một khoảng giá trị của khoảng cách cùng một lúc thay vì mỗi lần chỉ quan tâm đến một giá trị khoảng cách cố định. Vì lý do đó, chúng tôi sử dụng tất cả các véc tơ 2_0-gram, 2_1-gram,…, 2_{N-1}-gram trong một bộ phân lớp duy nhất (*N* là một giá trị định trước, trong kết quả thống kê ở trên *N* = 8). Hệ quả là véc tơ đặc trưng thu được có số chiều là 256^2N. Số lượng chiều lớn có thể dẫn đến việc tăng độ phức tạp và thời gian xử lý [4], nên ta cần áp dụng một giải thuật thích hợp để giảm số chiều xuống mức hợp lý.

Trong McPAD, các tác giả sử dụng một thuật toán phân cụm đặc trưng được đề xuất bởi Dhillon et al [3] để giảm số chiều cho véc tơ đặc trưng. Các đặc trưng được phân phối vào các cụm sao cho số cụm là nhỏ nhất trong khi lượng mất mát thông tin là có thể chấp nhận được. Sau đó, với mỗi cụm, giá trị trung bình của các đặc trưng trong mỗi cụm được tính và được dùng như là đặc trưng của véc tơ đặc trưng mới.

Chúng tôi đã thử áp dụng phương pháp giảm số chiều của McPAD, nhưng kết quả không như mong đợi. Khi tỷ lệ dương tính giả ở mức chấp nhận được là 10^{-2}, tỷ lệ dương tính thật chỉ đạt tới 0,5. Lý do cho kết quả tồi này là véc tơ đặc trưng của chúng tôi có số chiều lớn hơn gấp *N* lần (256^2N đặc trưng) so với các véc tơ đặc trưng của McPAD (256^2 chiều), do đó sẽ có nhiều hơn cả về số đặc trưng có giá trị, lẫn các đặc trưng dư thừa. Chia đặc trưng thành các cụm và sử dụng gía trị trung bình của các cụm làm đặc trưng mới làm mất mát lượng thông tin chứa trong các đặc trưng tốt, do đó cần loại bỏ các đặc trưng dư thừa và chỉ giữ lại các đặc trưng tốt. Để làm được điều đó, chúng tôi sử dụng thuật toán giảm số chiều Chi-Square [9], chọn một tập con các đặc trưng tốt nhất, và sử dụng chúng trong véc tơ đặc trưng. Bộ phân lớp sử dụng phương pháp của chúng tôi đạt được độ chính xác rất cao nhưng chạy trong một khoản thời gian rất ngắn.

Phần tiếp theo của bài báo được tổ chức như sau. Mục II giới thiệu những kiến thức cơ sở cần thiết để hiểu và nắm bắt được phương pháp do chúng tôi đề xuất. Chi tiết về phương pháp này được mô tả trong mục III. Các kết quả thực nghiệm được trình bày ở mục IV. Cuối cùng là kết luận và phương hướng phát triển tiếp theo.

II. KIẾN THỨC CƠ SỞ

A. Phân lớp một lớp

Phát hiện dựa trên bất thường có thể coi là một bài toán phân lớp với hai nhãn mục tiêu là bình thường và tấn công. Do trong thực tế đa số các hành vi truy nhập vào mạng là bình thường, rất khó và tốn kém để thu thập được một bộ dữ liệu được gán nhãn sẵn phản ánh các hoạt động thật vừa bao gồm các gói tin bình thường vừa chứa các gói tin tấn công, gần đây một số phương pháp học máy không gán nhãn hay còn gọi là phân lớp một lớp đã được đề xuất. Các phương pháp phân lớp một lớp chỉ sử dụng dữ liệu mạng bình thường không bị tấn công để huấn luyện tạo mô hình phản ánh các hành vi truy nhập mạng bình thường, trên cơ sở đó phát hiện ra các hành vi bất thường là các tấn công thâm nhập mạng. Trong các phương pháp phân lớp một lớp, phương pháp phân lớp một lớp SVM (Support Vector Machine) được lựa chọn sử dụng trong hệ thống phát hiện thâm nhập dựa trên bất thường McPAD vì chứng tỏ đạt hiệu quả cao trong phân loại văn bản [6], một vấn đề tương đồng với phát hiện bất thường dựa trên các thống kê *n*-gram [11].

B. Thuật toán Chi-Square để lọc thông tin đặc trưng

Để hạn chế số chiều của véc tơ đặc trưng, tránh làm gia tăng độ phức tạp tính toán và thời gian thực hiện quá mức chấp nhận được, một tập đặc trưng quá lớn cần được lọc để

giữ lại một tập con chỉ bao gồm các đặc trưng thỏa mãn một số tiêu chí cho trước. Các đặc trưng được chọn giữ nguyên ngữ nghĩa ban đầu. Việc rút gọn tập đặc trưng tương đương với hướng quá trình học và xử lý dữ liệu vào những thông tin có ý nghĩa nhất đối với mục tiêu phân lớp, giúp phân biệt dữ liệu thuộc các lớp khác nhau một cách hiệu quả nhất.

Tùy vào việc sử dụng dữ liệu có hay không được gán nhãn, quá trình lọc thông tin đặc trưng có thể được giám sát hay không được giám sát. Các phương pháp được giám sát xác định mối tương quan giữa các đặc trưng và nhãn lớp dựa trên các độ đo khoảng cách, độ phụ thuộc thông tin, và tính nhất [2]. Ngược lại, các phương pháp không giám sát lọc thông tin chỉ dựa trên mối tương quan giữa các đặc trưng với nhau, mà không sử dụng gì đến các thông tin về nhãn. Các nghiên cứu tổng quan và chuyên sâu hơn về lý thuyết lọc thông tin đặc trưng được đề cập đến trong các bài báo [1,5, 7, 15]

Thủ tục lọc thông tin đặc trưng sẽ được chúng tôi áp dụng ở giai đoạn học máy khi các dữ liệu đầu vào đều được gán nhãn. Vì vậy, trong khuôn khổ công trình này, chúng tôi chỉ quan tâm đến các phương pháp lọc thông tin có giám sát.

Như chúng tôi đã đề cập trước đó, phát hiện xâm nhập dựa trên bất thường sử dụng đặc trưng là số lần xuất hiện của *n*-gram cũng tương đương như bài toán phân loại văn bản. Vì vậy, các phương pháp chọn đặc trưng tốt cho bài toán phân loại văn bản cũng sẽ tốt cho bài toán phát hiện bất thường.

Một số thuật toán chọn đặc trưng tiêu biểu thường được sử dụng trong phân loại văn bản có thể kể ra bao gồm "Information Gain" (IG) [15], "Chi-Square" (CHI) [8, 15], "Document Frequency" (DF) [15], và "Mutual Information" (MI) [10]. Trong [15], thí nghiệm đã cho thấy các thuật toán như IG hay CHI là những giải thuật cho hiệu quả cao nhất. Vì CHI dễ cài đặt hơn cả, chúng tôi sử dụng thuật toán này để giảm số đặc trưng của véc tơ đặc trưng.

Trong thống kê, phương pháp Chi-square được dùng để kiểm tra độ liên quan giữa 2 sự kiện A và B. Trong việc giảm đặc trưng của bài toán này, 2 sự kiện là sự xuất hiện của các đặc trưng và của nhãn lớp. Các đặc trưng sẽ được xếp hạng theo công thức sau:

$$X^2(\mathbb{D}, t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}}$$

Trong đó, $e_t = 1$ khi mà phần nội dung gói tin D chứa đặc trưng *t*, ngược lại $e_t = 0$; $e_c = 1$ khi *D* thuộc lớp *c*, ngược lại $e_c = 0$, *N* là tần số quan sát được của *D* và *E* là tần số kỳ vọng. Ví dụ, E_{11} là tần số kỳ vọng của sự kiện đặc trưng *t* xuất hiện trong phần nội dung gói tin thuộc lớp *c*. Tương tự, N_{11} là tần số quan sát được của sự kiện đặc trưng *t* xuất hiện trong phần nội dung gói tin thuộc lớp *c*. Ta giả thiết sự xuất hiện của một đặc trưng và sự xuất hiện của một nhãn lớp là độc lập với nhau.

III. BỘ PHÂN LỚP TỔNG QUÁT

A. *Trích rút đặc trưng*

Để trích rút các đặc trưng của dữ liệu, chúng tôi tập trung vào phần nội dung của gói tin. Cụ thể hơn, tần số xuất hiện của các giá trị 2 byte sẽ được đếm, và khoảng cách giữa các cặp byte thay đổi từ 0 đến *N*-1 thay vì dừng ở một giá trị cố định như McPAD. Ví dụ, với cặp *AB* (byte đầu tiên là *A*, byte tiếp sau là *B*) chúng tôi quan tâm đến *AB, A*B, A**B,…, A*(*)$^{N-1}$*B*. Tần suất xuất hiện của chúng được lưu vào các véc tơ 2_v-gram tương ứng, với *v* chạy từ 0 đến *N*-1, sau đó các véc tơ lại nối lại với nhau tạo thành một véc tơ đặc trưng có số chiều tổng cộng là 256^2N.

Ví dụ, nếu mỗi phần tử đơn vị có 3 g*iá trị khác nhau A, B,* và *C,* và chúng ta có payload *AABCCABAA*, thì số lần xuất hiện của các 2_0-gram sẽ như sau:

AA	*BB*	*CC*	*AB*	*BA*	*BC*	*CB*	*AC*	*CA*
2	0	1	2	1	1	0	0	1

Véc tơ số lần xuất hiện của các 2_1-gram:

*A*A*	*B*B*	*C*C*	*A*B*	*B*A*	*B*C*	*C*B*	*A*C*	*C*A*
1	0	0	1	1	1	1	1	1

Véc tơ đặc trưng cuối cùng thu được là kết quả nối 2 véc tơ trên:

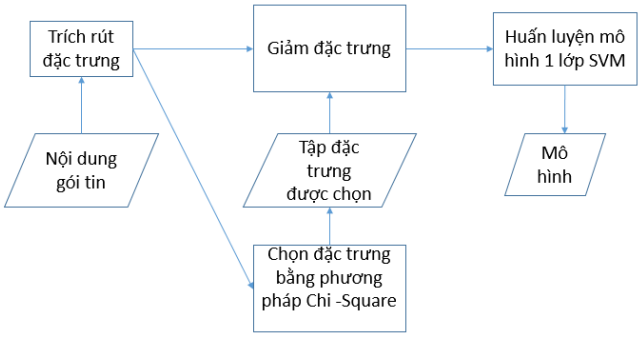
AA	*BB*	*CC*	*AB*	...	*B*C*	*C*B*	*A*C*	*C*A*
2	0	1	2	…	1	1	1	1

Phương pháp tổng quát hóa của chúng tôi tận dụng tối đa lượng thông tin về tần suất xuất hiện của các cặp byte. Tuy nhiên, vì số lượng đặc trưng tăng nên số lượng đặc trưng dư thừa, không có tác dụng thật sự trong việc phân biệt các gói tin hợp lệ với các gói tin tấn công cũng tăng. Vì thế, chúng tôi tính giá trị χ^2của mỗi đặc trưng, sau đó chỉ giữ lại *K* đặc trưng có giá trị χ^2 để đưa vào véc tơ đặc trưng sau cùng (*K* là một tham số cố định được xác định từ trước).

B. *Cơ chế hoạt động của bộ phân lớp*

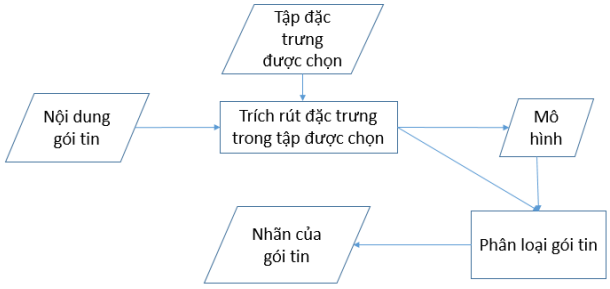
Bộ phân lớp của chúng tôi hoạt động theo 2 giai đoạn: giai đoạn huấn luyện và giai đoạn kiểm tra.

Giai đoạn huấn luyện (hình 1) có đầu vào là phần nội dung của các gói tin bình thường. Bộ phân lớp trích rút và rút gọn các véc tơ đặc trưng theo phương pháp đã nêu ở trên. Sau đó, các véc tơ đặc trưng được dùng để huấn luyện bộ phân loại một lớp SVM tạo ra một hình về các hành vi truy nhập mạng hợp lệ. Mô hình này sẽ được dùng để phân lớp các gói tin ở giai đoạn kiểm tra.



Hình 1. Sơ đồ quá trình huấn luyện

Giai đoạn kiểm tra (hình 2) có đầu vào là phần nội dung của gói tin cần kiểm tra để phát hiện thâm nhập. Sau quá trình trích rút đặc trưng ộ phân loại sử dụng model nhận được ở quá trình huấn luyện để kiểm tra các payload và xác định chúng là bình thường hay tấn công.



Hình 2. Sơ đồ quá trình kiểm tra

IV. THỰC NGHIỆM

A. *Cách thực hiện*

Chúng tôi mở rộng mã nguồn mở của McPAD[4] để tạo ra bộ phân loại mới. Mô đun đầu tiên mà chúng tôi viết được dùng để trích rút đặc trưng từ dữ liệu theo phương pháp đã được mô tả ở mục III.A. Mô đun tiếp theo là chọn đặc trưng Chi-Square, được dùng để chọn ra các đặc trưng tốt nhất và loại đi các đặc trưng dư thừa. Trong thí nghiệm, giá trị của N là 10 (v chạy từ 0 đến N-1), vì số lượng các bộ phân loại được dùng trong thí nghiệm của McPAD là 10. Ngoài ra, trong McPAD, số chiều được rút gọn từ 256^2 về 160, do đó chúng tôi để K (số chiều sau khi rút gọn) là 160.

B. *Các độ đo*

Có hai độ đo được sử dụng để đánh giá khả năng của bộ phân loại là đường cong ROC (Receiver Operating Characteristic) và diện tích dưới đường cong này AUC (Area Under the Curve). Đường cong ROC là một đồ thị biểu diễn sự tương quan giữa tỷ lệ dương tính thật và tỷ lệ dương tính giả, còn diện tích dưới đường cong cho thấy độ chính xác của bộ phân loại trong việc phát hiện hành vi độc hại.

C. *Dữ liệu*

Trong thí nghiệm, chúng tôi sử dụng hai bộ dữ liệu. Dữ liệu bình thường chứa các hành vi bình thường, được trích rút từ bộ dữ liệu DARPA'99[5] (bộ dữ liệu này phản ánh hoạt động truy cập mạng trong nhiều tuần ở một trường đại học, chúng tôi chỉ lấy ra dữ liệu của tuần đầu tiên). Bộ dữ liệu sau trích rút được dùng để huấn luyện mô hình và kiểm tra tỷ lệ dương tính giả. Bộ dữ liệu còn lại là các hành vi độc hại, được dùng để kiểm tra khả năng phát hiện độc hại của bộ phân loại. Bộ dữ liệu này được chia làm 3 phần:

- Tấn công tổng hợp: Bao gồm 66 loại tấn công HTTP[6], trong đó có 11 loại kiểu shellcode và mang theo mã thực thi trong phần nội dung gói tin, các loại tấn công còn lại bao gồm kiểu tấn công từ chối dịch vụ, tấn công rò rỉ thông tin,...
- Tấn công shellcode: Bao gồm 11 loại tấn công shellcode lấy từ phần trên.
- Tấn công CLET: Chứa 96 tấn công đa hình được sinh từ bộ sinh đa hình CLET [12].

D. *Kết quả thực nghiệm*

1. *Độ chính xác*

Bảng 4 cho thấy giá trị AUC của bộ phân loại của chúng tôi. Giá trị này rất gần với 1, tức là rất tốt, và có thể sử dụng trong thực tế.

Bảng 4: Giá trị AUC của các bộ phân loại

Loại tấn công	**AUC**
Tấn công tổng hợp	0,97227
Tấn công Shellcode	0,98656
CLET	0,96737

2. *So sánh với PAYL và McPAD*

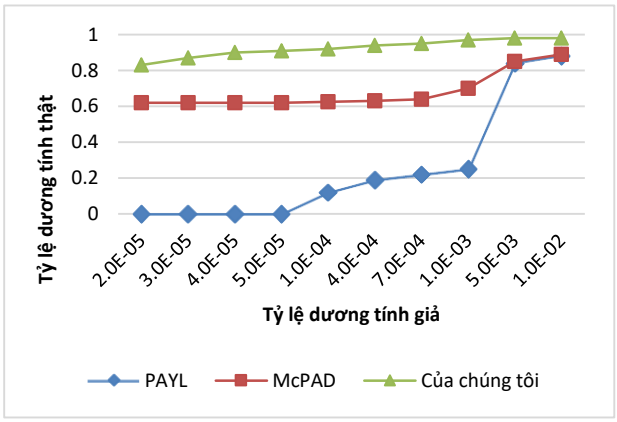
Trong phần này, chúng tôi so sánh khả năng phân lớp của bộ phân lớp của chúng tôi với PAYL và McPAD đối với 3 loại tấn công nêu trên.

[4] http://roberto.perdisci.googlepages.com/mcpad

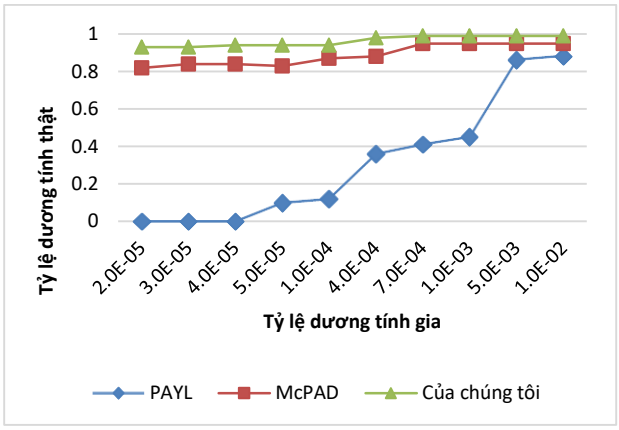
[5] https://www.ll.mit.edu/ideval/data/1999/training/week1/index.html

[6] http://roberto.perdisci.com/publications/publication-files/HTTP_generic-attacks.zip?attredirects=0

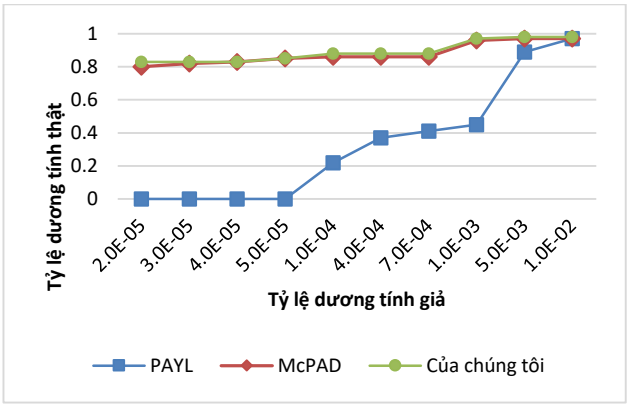
.



Đồ thị 1. Tấn công tổng hợp



Đồ thị 2. Tấn công shellcode



Đồ thị 3. Tấn công CLET

Các đồ thị 1, 2, 3 cho thấy với kiểu các tấn công tổng hợp, shellcode và CLET, PAYL đều có tỷ lệ dương tính thật thấp khi mà tỷ lệ dương tính giả ở mức rất thấp (bé hơn 10^{-3}). Với kiểu tấn công tổng hợp, McPAD đạt được tỷ lệ dương tính thật khá cao ở mức dương tính giả như trên, nhưng bộ phân loại của chúng tôi vượt trội so với 2 bộ đó trên cả miền giá trị của tỷ lệ dương tính giả. Ví dụ ở mức dương tính giả là 2×10^{-5}, giá trị dương tính thật của McPAD là 0,6, ít hơn nhiều so với của chúng tôi, 0,8. Tiếp đó, ở mức dương tính giả 10^{-2}, tỷ lệ dương tính thật của chúng tôi là gần 1, nhưng của McPAD chỉ là 0,9. Với kiểu tấn công shellcode cà CLET, bộ phân loại của chúng tôi chỉ tốt hơn một chút so với McPAD, vì thực ra McPAD vốn đã phát hiện rất tốt hai loại tấn công này.

Bảng 5: Thời gian xử lý trung bình mỗi gói tin

Bộ phân loại	Thời gian trung bình(ms)
PAYL	0,039
McPAD	13,39
Của chúng tôi	2,06

Bảng 5 so sánh thời gian trung bình xử lý một gói tin của các bộ phân loại trong quá trình kiểm tra. Mặc dù PAYL chạy rất nhanh, độ chính xác của nó chưa cao như ta có thể thấy ở thí nghiệm trên. Điều thú vị là bộ phân loại của chúng tôi chạy nhanh gấp 6 lần McPAD nhưng chính xác hơn nhiều, vì McPAD chạy nhiều bộ phân lớp, trong khi của chúng tôi chỉ có một.

IV. KẾT LUẬN

Chúng tôi đã đề xuất một phương pháp trích rút đặc trưng mới, mở rộng của các phương pháp trích rút 2-gram và 2_v-gram trong các hệ thống phát hiện thâm nhập mạng dựa trên bất thường PAYL và McPAD, sử dụng kết hợp với phương pháp chọn đặc trưng Chi-Square. Phương pháp của chúng tôi quan tâm đến các giá trị khác nhau của khoảng cách giữa hai byte, do đó nhận được nhiều thông tin hơn từ tần số xuất hiện của các cặp byte không liền nhau. Sau đó, chúng tôi áp dụng phương pháp này để xây dựng bộ phân lớp một lớp SVM để phát hiện một cách hiệu quả các tấn công vào hệ thống máy tính. Kết quả thí nghiệm cho thấy bộ phân loại này có thể phát hiện rất nhanh nhiều loại tấn công HTTP, luôn đạt được tỷ lệ dương tính thật cao khi mà tỷ lệ dương tính giả luôn ở mức thấp. Ngoài ra, kết quả thực nghiệm còn cho thấy bộ phân loại của chúng tôi có hiệu quả vượt trội so với các bộ phân loại PAYL và McPAD. Do đó chúng tôi tin rằng bộ phân loại này hoàn toàn có thể được áp dụng trong thực tế.

Phương hướng tiếp theo của bài báo là nghiên cứu đề xuất một giải pháp rút gọn các đặc trưng phù hợp hơn với các cấu trúc đặc trưng dạng 2_v-gram thay vì áp dụng một phương pháp có tính phổ quát như Chi-Square.

TÀI LIỆU THAM KHẢO

[1] A. Blum and L. Pat, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, pp. 245-271, 1997.

[2] Dash and Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 1-4, pp. 131-156, 1997.

[3] S. Dhillon, S. Mallela and R. Kumar, "A divisive information-theoretic feature clustering algorithm for text classification," *Journal of Machine Learning Research*, p. 1265–1287, 2003.

[4] R. O. Duda, P. E. Hart and D. G. Stork, "Pattern Classification," *Wiley*, 2000.

[5] D. Koller and S. Sahami, "Toward Optimal Feature Selection," *In Proceedings of the Thirteenth International Conference on Machine Learning (ICML)*, pp. 284-292, 1996.

[6] E. Leopold and j. Kindermann, "Text categorization with support vector machines," *How to represent texts in input space? Machine Learning*, p. 423–444, 2002.

[7] H. Liu and H. Motoda, "Computational Methods of Feature Selection," *Chapman and Hall/CRC Press*, p. 2007.

[8] H. Ogura, H. Amano and K. Kondo, "Feature selection with a measure of deviations from Poisson in text categorization," *Expert Systems with Applications*, p. pages 6826–6832, 2009.

[9] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philosophical Magazine*, p. 157–175, 1900.

[10] H. Peng, F. Long and C. Ding, "Feature selection based on mutual information: criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1226–1238, 2005.

[11] R. Perdisci, D. Ariu, P. Fogla, G. Giacinto and W. Lee, "McPAD : A Multiple Classifier System for Accurate Payload-based Anomaly Detection," *Computer Networks, Special Issue on Traffic Classification and Its Applications to Modern Networks*, pp. 864-881, 2009.

[12] J. Tax, "One-Class Classification, Concept Learning in the Absence of Counter Examples," *PhD thesis, Delft University of Technology*, 2001.

[13] K. Wang and S. Stolfo, "Anomalous payload-based network intrusion detection," *In Recent Advances in Intrusion Detection (RAID)*, 2004.

[14] K. Wang and S. Stolfo, "Anomalous payload-based worm detection and signature generation," *In Recent Advances in Intrusion Detection (RAID)*, 2005.

[15] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," *Proceedings of the 14th International Conference on Machine Learning (ICML '97)*, p. 412–420 , 1997.