PHÂN LOẠI LUỒNG DỮ LIỆU MẠNG SỬ DỤNG MẠNG NƠ-RON

NETWORK TRAFFIC CLASSIFICATION USING NEURAL NETWORK

Trần Văn Líc, Phan Trần Đăng Khoa

Trường Đại học Bách khoa – Đại học Đà Nẵng; tvlic@dut.udn.vn, ptdkhoa@dut.udn.vn

Tóm tắt - Với sự phát triển của hạ tầng mạng Internet, trong những năm gần đây tầm quan trọng của việc phân loại các luồng dữ liệu mạng nhằm nâng cao về chất lượng, bảo mật cho hệ thống mạng ngày càng thu hút được sự quan tâm nghiên cứu. Trong đó, phương pháp phân loại luồng dữ liệu dựa trên các mô hình học máy cũng đang được nghiên cứu và đã đạt được những kết quả đáng chú ý. Trong bài báo này, nhóm tác giả sử dụng mạng nơ-ron để phát triển một mô hình có thể đạt được độ chính xác cao trong việc phân loại luồng dữ liệu mạng. Các phương pháp xử lý dữ liệu cũng được áp dụng để tối ưu thời gian thực hiện và tài nguyên cho hệ thống, đồng thời nâng cao tỉ lệ phân loại đúng cho các nhóm có tần số xuất hiện thấp trong cơ sở dữ liệu. Kết quả thực nghiệm trên cơ sở dữ liệu mở đã cho thấy, mô hình đề xuất có độ ổn định theo thời gian và tỉ lệ phân loại cho các nhóm thiểu số tốt hơn so với các mô hình khác.

Từ khóa - luồng dữ liệu; phân loại; mạng nơ-ron

Abstract - With the rapid development of the Internet infrastructure, in recent years, Internet traffic classification has been intensively researched in order to improve the quality and security of the network. In particular, methods of traffic classification based on machine learning models are being studied and have achieved remarkable results. In this paper, we use neural networks to develop a model that can achieve high accuracy in classifying network traffic flows. Data processing methods are also applied to improve the classification ability for minority groups. Experimental results have shown that the proposed model has better stability and classification rate for minority groups than other models.

Key words - traffic flow; classification; neural network

1. Giới thiệu

Phân loại luồng dữ liệu mạng (Network traffic classification) là việc nhận dạng các loại ứng dụng và giao thức mạng khác nhau tồn tại trong hệ thống mạng. Với chức năng giám sát, khám phá, điều khiển và tối ưu hệ thống mạng, mục tiêu chung của phân loại luồng dữ liệu mạng là cải thiện hiệu năng hoạt động mạng. Khi các gói tin được phân loại sẽ giúp cho bộ định tuyến (router) tính toán chính sách (policy) dịch vụ thích hợp. Điều này cũng cho phép chúng ta dự đoán tốt hơn về luồng dữ liệu mạng, phát hiện và ngăn chặn các luồng dữ liệu mạng bất thường nhằm tăng bảo mật dữ liệu cá nhân. Ngoài ra, dựa trên sự phân loại này, các chính sách dịch vụ có thể được áp dụng như với VoIP (Voice over Internet Protocol), dịch vụ giải trí trực truyến sẽ được cam kết về chất lượng [1]. Tuy nhiên, với sự phát triển liên tục và đa dạng của các ứng dụng, số lượng host và khối lượng luồng dữ liệu trên mạng Internet đã tạo nên thách thức lớn cho các phương pháp phân loại luồng dữ liệu mạng ứng với từng ứng dụng và mức độ phát triển này dự đoán vẫn tiếp tục tăng trong tương lai.

Mặc dù một số phương pháp phân loại luồng dữ liệu truyền thống đang được áp dụng phổ biến hiện nay như phương pháp phân loại dựa trên số định danh của cổng (port number) và phương pháp phân tích gói dữ liệu (deep packet inspection), nhưng vẫn tồn tại một số vấn đề chưa được giải quyết [2]. Đầu tiên, phần lớn bởi vì luồng dữ liệu mạng không dễ dàng phân loại dựa vào chuẩn IANA (International Assigned Number Authority) theo danh sách các cổng ứng dụng, các ứng dụng khẩn cấp và proxy thường tránh sử dụng các cổng chuẩn. Thứ hai, các cổng ứng dụng và ký hiệu giao thức có thể không đủ để xác định các ứng dụng thực tế. Về nguyên tắc, không có ràng buộc rõ ràng giữa các ứng dụng và giao thức cơ bản. Ví dụ, các ứng dụng như MSN Messenger, BitTorrent và Gnutella có thể sử dụng giao thức HTTP (HyperText Transfer Protocol) cổng 80, trong khi Skype có thể hoạt động ở cả cổng 80 và 443. Thứ ba, việc mã hóa và đóng gói luồng dữ liệu ngày càng tăng như SOCKS proxy hay VPN (Virtual Private Network) làm thay đổi mô hình trong giao thức gốc, trong khi mã hóa gói làm cho việc kiểm tra, đào sâu vào dữ liệu không sử dụng được. Do đó, phương pháp phân loại dựa trên số định danh của cổng chỉ hiệu quả cho các ứng dụng và dịch vụ sử dụng các cổng cố định; còn phương pháp phân tích gói dữ liệu đòi hỏi tài nguyên và thời gian lớn để phân tích dữ liệu của gói tin lớn. Nhìn chung, cả 2 hướng tiếp cận trên đều có những hạn chế nhất định về độ chính xác trong việc phân loại và tài nguyên sử dụng.

Trong những năm gần đây, việc giải quyết vấn đề phân loại luồng dữ liệu mạng sử dụng các mô hình học máy thu hút được sự quan tâm nghiên cứu [2-9]. Dựa trên các thuộc tính của gói tin như tần suất byte (byte frequency), kích thước gói tin (packet size), khe thời gian đến giữa các gói tin (packet inter-arrival time), v.v… và kết hợp với các mô hình học máy như cây quyết định (decision tree), bộ phân loại Naïve Bayes, mạng nơ-ron, các phương pháp này có ưu điểm là độ chính xác cao và xử lý nhanh hơn so với các phương pháp phân loại đã nêu trên vì không đào sâu tới phần nội dung (content) của gói dữ liệu mà chỉ sử dụng các header của gói dữ liệu để phân tích [5]. Các phương pháp này sử dụng các công cụ phân loại thống kê để xây dựng các mô hình phân loại dựa trên các cơ sở dữ liệu huấn luyện đã được gắn nhãn. Các mô hình này có thể cho ra kết quả là nhóm đối tượng hoặc là phân bố xác suất của các nhóm đối với từng mẫu. Khác với các phương pháp truyền thống, các phương pháp học máy sử dụng đặc trưng đầu vào là thành phần siêu dữ liệu của dữ liệu (payload metadata) nên thường gặp phải vấn đề quá khớp (overfitting), tương ứng với tỷ lệ phân loại đúng cao (99%-100%) đối với quá trình huấn luyện, tuy nhiên không ổn định khi áp dụng kết quả mô hình cho cơ sở dữ liệu được thu thập từ các mạng khác hoặc từ cùng một mạng nhưng tại các thời điểm khác nhau [5].

Trong nghiên cứu [5], nhóm tác giả đã sử dụng phương pháp học máy có giám sát với mạng nơ-ron để xây dựng mô hình phân loại luồng dữ liệu có độ chính xác cao. Nghiên cứu đã đánh giá độ ổn định của mô hình đối với các mạng khác

nhau và tại các thời điểm khác nhau. Tuy nhiên, kết quả nghiên cứu cũng cho thấy tỷ lệ phân loại đúng rất thấp đối với các nhóm có tần suất xuất hiện thấp (được gọi là nhóm thiểu số) trong cơ sở dữ liệu huấn luyện. Trong một nghiên cứu khác cùng hướng, nhóm tác giả đã sử dụng mạng nơ-ron để phân loại luồng dữ liệu giao thức TCP (Transmission Control Protocol) với các giao thức khác dựa vào các giá trị thống kê về thông tin và thuộc tính ở lớp IP (Internet Protocol) [6]. Trong nghiên cứu [4], nhóm tác giả đã khai thác mô hình mạng nơ-ron Bayes và đạt được độ chính xác 99,3% và giảm xuống còn 95,3% khi kiểm thử với nguồn dữ liệu khác.

Kỹ thuật học sâu (Deep learning) đã được áp dụng để phân loại luồng dữ liệu mạng và đã có một vài nghiên cứu trong những năm gần đây. Wang Z. đã sử dụng 1000 bytes đầu tiên của mỗi luồng dữ liệu TCP làm dữ liệu đầu vào. Kết quả huấn luyện đã chỉ ra các bytes quan trọng cho việc phân loại. Tỷ lệ phân loại đúng là 55% khi lấy ngưỡng là 90% [7].

Nhóm tác giả trong nghiên cứu [8] đã áp dụng và so sánh 5 phương pháp học máy khác nhau để phân loại luồng dữ liệu IP ở thời gian thực. Nghiên cứu cho ra được kết quả phân loại với độ chính xác 91,875%, và kết quả này thấp hơn các nghiên cứu ban đầu do nhóm tác giả tập trung phát triển thuật toán hoạt động trong thời gian thực.

Qua các phân tích nêu trên, có thể thấy rằng, các phương pháp phân loại luồng dữ liệu dựa trên học máy, đặc biệt là mạng nơ-ron, có tỷ lệ nhận dạng đúng cao. Tuy nhiên, vấn đề cần được giải quyết là tránh việc quá khớp và tăng tỷ lệ phân loại đúng đối với các nhóm thiểu số. Ngoài ra, khả năng thực thi mô hình trong thời gian thực cũng là một vấn đề cần được nghiên cứu.

Trong bài báo này, nhóm tác giả trình bày mô hình phân loại luồng dữ liệu mạng dựa trên mạng nơ-ron. So với các nghiên cứu khác, nghiên cứu này có những đóng góp chính như sau:

+ Chọn lọc ra bộ đặc trưng với số chiều không gian ít hơn và thích ứng với mô hình mạng nơ-ron, tuy nhiên vẫn duy trì được độ chính xác và độ ổn định của mô hình.

+ Nâng cao khả năng phân loại của mô hình đối với các nhóm thiểu số dựa trên một số kỹ thuật xử lý dữ liệu.

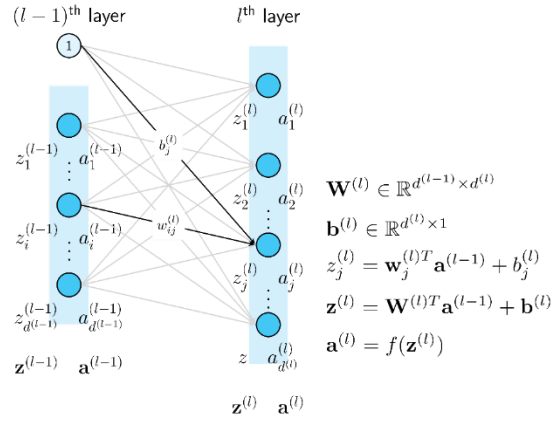
2. Mô hình phân loại luồng dữ liệu mạng

2.1. Mô hình mạng nơ-ron

Mạng nơ-ron bao gồm 3 lớp cơ bản: Lớp đầu vào (Input layer), lớp ẩn (Hidden layer) và lớp đầu ra (Output layer). Mạng nơ-ron có thể có một hoặc nhiều lớp ẩn. Mỗi lớp được cấu tạo từ một hoặc nhiều nơ-ron, và mỗi nơ-ron ở lớp trước được kết nối với tất cả các nơ-ron ở lớp kế tiếp như Hình 1.

Đầu vào của các lớp ẩn được ký hiệu bởi $\boldsymbol{z}$, đầu ra của mỗi nơ-ron thường được ký hiệu là $\boldsymbol{a}$. Đầu ra của nơ-ron thứ i trong lớp thứ l được ký hiệu là $\boldsymbol{a_i^{(l)}}$. Vector biểu diễn lớp đầu ra của lớp thứ l được ký hiệu là $\boldsymbol{a^{(l)} \in R^{d^{(l)}}}$.

Có L ma trận trọng số cho một mạng nơ-ron có L lớp. Các ma trận này được ký hiệu là $\boldsymbol{W^{(l)} \in R^{d^{(l-1)} x\, d^{(l)}}}$, $l = 1,2,\dots,L$, trong đó $\boldsymbol{W^{(l)}}$ thể hiện các kết nối từ lớp thứ $l-1$ tới layer thứ l. Các độ lệch của lớp thứ l được ký hiệu là $\boldsymbol{b^{(l)} \in R^{d^{(l)}}}$. Các trọng số này được ký hiệu như trên Hình 1. Tập hợp các trọng số và độ lệch lần lượt được ký hiệu là $\boldsymbol{W}$ và $\boldsymbol{b}$.



***Hình 1**. Mô hình mạng nơ-ron và các ký hiệu sử dụng trong mạng nơ-ron [10]*

Mỗi lớp đầu ra của một nơ-ron được tính dựa vào công thức:

$$\mathbf{a}_i^{(l)} = \boldsymbol{f}\left(\mathbf{w}_i^{(l)T}\mathbf{a}^{(l-1)} + \boldsymbol{b}_i^{(l)}\right) \quad (1)$$

trong đó $f(\cdot)$ là hàm kích hoạt phi tuyến.

Gọi $\hat{\boldsymbol{y}} = \boldsymbol{a}^{(L)}$ là đầu ra dự đoán của mạng nơ-ron, tương ứng với đầu ra của các nơ-ron thuộc lớp đầu ra (lớp thứ L). Đối với các bài toán phân loại, đầu ra dự đoán $\hat{y}$ được chuyển đổi sang dạng xác suất, trong đó $\hat{y}_i$ là xác suất mẫu thuộc về nhóm i. Việc biến đổi này cần phải đảm bảo các xác suất đầu ra là dương và tổng chúng bằng 1. Hàm biến đổi được sử dụng trong nghiên cứu này là hàm softmax với biểu thức như sau:

$$a_i^{(L)} = \frac{\exp\left(z_i^{(L)}\right)}{\sum_{j=1}^{C}\exp\left(z_j^{(L)}\right)} \quad (2)$$

trong đó, $z_i^L = \mathbf{w}_i^{(L)^T}\mathbf{a}^{(L-1)}$ là đầu vào của nơ-ron lớp đầu ra; C – số đầu ra.

Việc huấn luyện mạng nơ-ron tương ứng với việc tối ưu hàm mất mát theo các trọng số $\boldsymbol{W}$ và độ lệch $\boldsymbol{b}$. Gọi $\hat{\boldsymbol{y}} = \boldsymbol{a}^{(L)}$ là đầu ra dự đoán của mạng nơ-ron, tương ứng với đầu ra của các nơ-ron thuộc lớp đầu ra (lớp thứ L). Hàm mất mát trong bài toán phân loại là cross-entropy và được biểu diễn bằng biểu thức:

$$\mathcal{L}(\mathbf{W},\mathbf{b},\mathbf{X},\mathbf{Y}) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log^{(\hat{y}_i)}_{+(1-y_i)\log(1-\hat{y}_i)}\right] + \lambda\Phi(\boldsymbol{W}) \quad (3)$$

trong đó, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ là tập hợp các biến đầu vào và $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N\}$ là tập hợp các nhãn tương ứng; hàm $\Phi(\boldsymbol{W})$ là thành phần ổn định hóa (regularization).

Phương pháp Gradient Descent được áp dụng để tối ưu hàm mất mát theo các các trọng số $\boldsymbol{W}$ và độ lệch $\boldsymbol{b}$.

$$\mathbf{w}_{t+1}^l = \mathbf{w}_t^l - \eta\frac{\partial\mathcal{L}}{\partial\mathbf{w}^{(l)}} \quad (4)$$

$$\mathbf{b}_{t+1}^l = \mathbf{b}_t^l - \eta\frac{\partial\mathcal{L}}{\partial\mathbf{b}^{(l)}} \quad (5)$$

trong đó, η là hệ số học (learning rate).

Đối với phương pháp này, ta cần tính gradient của hàm mất mát theo từng trọng số và độ lệch, tức $\frac{\partial\mathcal{L}}{\partial\mathbf{w}^{(l)}}$ và $\frac{\partial\mathcal{L}}{\partial\mathbf{b}^{(l)}}$. Một phương pháp phổ biến để tính các gradient này là thuật toán lan truyền ngược (back-propagation) cho phép tính

gradient ngược từ lớp cuối đến lớp đầu.

2.2. Phân tích cơ sở dữ liệu

Một phần quan trọng đối với các mô hình học máy là cơ sở dữ liệu. Việc phân tích cơ sở dữ liệu cho phép làm rõ các tính chất đặc thù của từng dữ liệu và từ đó đưa ra các phương pháp, kỹ thuật xử lý dữ liệu, cũng như mô hình phù hợp để nâng cao khả năng phân loại. Hiện nay, để phục vụ cho việc nghiên cứu các mô hình phân loại luồng dữ liệu mạng, một số cơ sở dữ liệu mở với số lượng mẫu lớn (lên đến vài trăm ngàn mẫu) đã được xây dựng và sử dụng rộng rãi để làm cơ sở phát triển và so sánh các mô hình. Việc phân tích các cơ sở dữ liệu này sẽ cho thấy các đặc thù và các vấn đề gặp phải khi triển khai mô hình.

Cơ sở dữ liệu được sử dụng cho mô hình mạng nơ-ron bao gồm các mẫu (example), và mỗi mẫu chứa các đặc trưng dùng để phân loại và nhãn của mẫu. Các cơ sở dữ liệu dùng để phân loại luồng dữ liệu mạng được thu thập tại các cơ sở mạng thường gặp phải vấn đề mất cân bằng (unbalanced) về số lượng mẫu giữa các nhóm, tức là một số ít nhóm chiếm số lượng mẫu đa số trong cơ sở dữ liệu (được gọi là các nhóm đa số), còn lại số lượng rất ít mẫu thuộc về phần lớn các nhóm còn lại (được gọi là các nhóm thiểu số). Sự chênh lệch này trong các cơ sở dữ liệu hiện nay là rất lớn, nhóm đa số có thể có số lượng mẫu lên đến vài trăm ngàn, trong khi nhóm thiểu số chỉ có vài mẫu. Hiện tượng này xảy ra do thói quen sử dụng của người dùng tại các cơ sở mạng, và có thể thấy rằng, đa số các cơ sở dữ liệu mở để phục vụ cho việc nghiên cứu các thuật toán phân loại luồng dữ liệu đều bị mất cân bằng [11]. Do đó, việc thu thập được một cơ sở dữ liệu với số lượng mẫu lớn, đồng thời đảm bảo sự cân bằng giữa các nhóm là rất khó khăn.

Việc mất cân bằng trong cơ sở dữ liệu dễ dẫn đến việc chênh lệch trong khả năng phân loại đối với nhóm đa số và nhóm thiểu số. Các mô hình học máy sẽ có xu hướng phân loại các nhóm đa số để tối thiểu hóa hàm mất mát mà không chú trọng đến việc phân loại các nhóm thiểu số. Các kết quả phân tích trong bài báo này cũng cho thấy, các mẫu của nhóm đa số vẫn đóng góp vào sai số phân loại, trong khi các mẫu của các nhóm thiểu số gần như không được phân loại. Do đó, việc tập trung nâng cao khả năng phân loại đối với các nhóm thiểu số là cần thiết.

Trong các cơ sở dữ liệu mở để phục vụ cho nghiên cứu các mô hình phân loại luồng dữ liệu, các đặc trưng cơ bản thường xuất hiện là server port và client port (ví dụ, port 80). Các đặc trưng này mang tính chất định danh, tức giá trị của chúng không mang ý nghĩa đo lường. Rõ ràng rằng, các đặc trưng định danh không phù hợp để đưa trực tiếp vào mô hình do không thể hiện được sự đối sánh định lượng tương đối giữa các giá trị định danh. Ví dụ, so với port 1280, port 80 không phải gần hơn port 20.

Các cơ sở dữ liệu thường thu thập được số lượng lớn các đặc trưng (lên đến vài trăm). Tuy nhiên, mối quan hệ tương quan giữa các đặc trưng này thường không được xem xét. Việc đưa các đặc trưng có mối quan hệ tương quan dẫn tới sự thừa thông tin, tăng độ phức tạp của mô hình. Ngoài ra, một lượng thừa các đặc trưng có thể làm giảm độ ổn định của mô hình.

Từ các phân tích trên, nhóm tác giả đề xuất một số kỹ thuật xử lý cơ sở dữ liệu nhằm khắc phục các vấn đề nêu trên, bao gồm phân bố cơ sở dữ liệu, biến đổi các đặc trưng định danh, chuẩn hóa dữ liệu, chọn lọc các đặc trưng. Kết quả thực nghiệm trên một cơ sở dữ liệu mở cho thấy hiệu quả của các kỹ thuật này.

2.3. Xử lý cơ sở dữ liệu

2.3.1. Biến đổi đặc trưng định danh

Như đã trình bày ở trên, ta cần phải biến đổi các đặc trưng định danh, cụ thể là server port và client port, sang giá trị định lượng để có thể sử dụng làm đầu vào của mô hình. Việc biến đổi cần đảm bảo tầm quan trọng tương đối giữa các giá trị định danh. Qua khảo sát, nhóm tác giả thấy rằng tần số xuất hiện của đặc trưng trong cơ sở dữ liệu là phù hợp để làm cơ sở cho biến đổi đặc trưng định danh. Giá trị định lượng của đặc trưng định danh được xác định bằng biểu thức sau:

$$R_i = \frac{n_i}{\sum_{i=1}^{N} n_i} \quad (6)$$

trong đó, n_i – số lượng mẫu chứa đặc trưng định danh i.

Việc biến đổi này đảm bảo các Port có tần số xuất hiện lớn sẽ có giá trị định lượng lớn tương ứng, và các giá trị này được chuẩn hóa trong khoảng [0 1].

2.3.2. Chuẩn hóa dữ liệu đầu vào

Cơ sở dữ liệu bao gồm các đặc trưng có giá trị nằm ở các tỷ lệ (scale) khác nhau. Để đảm bảo công bằng giữa các đặc trưng, ta cần phải chuẩn hóa các giá trị của cơ sở dữ liệu đối với từng đặc trưng. Việc chuẩn hóa nhằm đảm bảo các giá trị của mỗi đặc trưng có giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1. Biểu thức để chuẩn hóa cơ sở dữ liệu như sau:

$$x_i^j = \frac{x_i^j - \mu^j}{\sigma^j} \quad (7)$$

trong đó, x_i^j – giá trị thứ i của đối với đặc trưng thứ j; μ^j và σ^j lần lượt là giá trị trung bình và độ lệch chuẩn của các giá đối với đặc trưng thứ j.

2.3.3. Phân chia cơ sở dữ liệu

Do cơ sở dữ liệu không cân bằng nên việc phân chia cơ sở dữ liệu thành các tập huấn luyện (train set) và kiểm tra (test set) cần đảm bảo sự có mặt của các mẫu thuộc nhóm thiểu số. Nếu các nhóm thiểu số không xuất hiện trong các tập kiểm tra thì mô hình sẽ không học được các đặc trưng để phân loại các nhóm này. Để giải quyết vấn đề này, tỉ lệ số mẫu của các nhóm trong các tập huấn luyện và kiểm thử được xác định bởi tỉ lệ của chúng trong toàn bộ cơ sở dữ liệu. Điều này đảm bảo số lượng mẫu của tất cả các nhóm trong tập huấn luyện và tập kiểm thử đều có mặt với số lượng phù hợp.

2.3.4. Chọn lọc bộ đặc trưng

Việc chọn lọc đặc trưng cho phép lấy ra các đặc trưng có ảnh hưởng nhất, từ đó làm giảm số chiều của không gian đặc trưng và sự thừa thông tin. Các đặc trưng hữu ích nhất giúp cho mô hình học được huấn luyện hiệu quả hơn và có độ ổn định cao hơn.

Trong nghiên cứu [4], đại lượng đo lường Symmetrical Uncertainty được sử dụng để xác định độ tương quan giữa các đặc trưng. Từ đó, lựa chọn ra bộ đặc trưng có ảnh hưởng nhất. Tuy nhiên, phương pháp không dựa trên mô hình học được sử dụng mà chỉ dựa vào mối quan hệ tương

quan giữa các đặc trưng. Trong khi đó, quá trình huấn luyện một mô hình học máy là một quá trình tối ưu hóa hàm mất mát, và mức độ ảnh hưởng của từng đặc trưng đối với mô hình là rất khó dự đoán trước. Chính vì vậy, trong nghiên cứu này, chúng tôi đề xuất thực hiện phương pháp chọn lọc đặc trưng tuần tự (Sequential Feature Selection) nhằm thích nghi với mô hình được sử dụng.

Phương pháp chọn lọc đặc trưng tuần tự lựa chọn tập các đặc trưng sao cho dự đoán tốt nhất cơ sở dữ liệu thông qua tuần tự đánh giá ảnh hưởng của từng đặc trưng đến khi không thể cải thiện thêm kết quả dự đoán. Phương thức chọn lọc đặc trưng bắt đầu bằng tập đặc trưng rỗng và lần lượt đưa các đặc trưng chưa được lựa chọn vào. Đối với mỗi đặc trưng, ta huấn luyện và đánh giá mô hình mạng nơ-ron theo phương pháp k-fold cross-validation, tức là chia tập huấn luyện thành k tập con và lần lượt lấy 1 tập con làm tập kiểm tra và $(k-1)$ tập con còn lại làm tập huấn luyện. Việc này đảm bảo đặc trưng được đánh giá trên toàn bộ dữ liệu. Tiêu chí đánh giá là tỷ lệ phân loại sai trên tập kiểm thử. Quy trình này được thực hiện cho đến khi tìm được tập đặc trưng tối ưu hóa tiêu chí đánh giá.

Phương pháp chọn lọc đặc trưng tuần tự cho phép lựa chọn các đặc trưng có ảnh hưởng đối với mô hình học cụ thể. Tuy nhiên, phương pháp này đòi hỏi khối lượng tính toán lớn vì cần phải tuần tự đánh giá ảnh hưởng của từng đặc trưng đối với mô hình. Do đó, đối với các cơ sở dữ liệu có số lượng đặc trưng lớn thì ta cần phải chọn lọc sơ bộ các đặc trưng tốt trước khi thực hiện phương pháp chọn lọc đặc trưng tuần tự. Để thực hiện bước này, chúng tôi áp dụng thuật toán Neighborhood Component Analysis (NCA). Thuật toán NCA thực hiện phân nhóm dữ liệu dựa trên tiêu chí khoảng cách để xác định các điểm dữ liệu lân cận [12].

2.3.5. Gia tăng dữ liệu

Đối với các cơ sở dữ liệu không cân bằng, các mô hình học máy sẽ có xu hướng phân loại các nhóm đa số để tối thiểu hóa hàm mất mát mà không chú trọng đến việc phân loại các nhóm thiểu số. Để khắc phục vấn đề này, ta cần phải tăng trọng số các mẫu của các nhóm thiểu số. Như vậy, trong quá trình tối ưu hàm mất mát của mạng nơ-ron, mô hình buộc phải điều chỉnh để tăng khả năng phân loại cho các nhóm thiểu số. Trong nghiên cứu này, chúng tôi tăng trọng số của các nhóm thiểu số bằng cách sao chép các mẫu của các nhóm này kết hợp cùng với việc thêm nhiễu, từ đó làm tăng đáng kể số lượng mẫu của các nhóm thiểu số. Kết quả thực nghiệm cho thấy, kỹ thuật gia tăng dữ liệu này góp phần cải thiện đáng kể tỷ lệ phân loại đúng đối với các nhóm thiểu số.

3. Kết quả thí nghiệm

3.1. Cơ sở dữ liệu

Cơ sở dữ liệu được sử dụng trong nghiên cứu này được lấy từ nguồn [11]. Đây là cơ sở dữ liệu được nhiều nghiên cứu liên quan sử dụng, do đó việc sử dụng cơ sở dữ liệu này tạo điều kiện thuận lợi cho việc đánh giá và so sánh với các nghiên cứu khác.

Cơ sở dữ liệu được thu thập từ máy chủ tại trung tâm nghiên cứu của Đại học Cambridge với khoảng 1000 người dùng kết nối với Internet thông qua đường kết nối Full-duplex Gigabit Ethernet. Tổng số mẫu của cơ sở dữ liệu là 190748 mẫu. Dữ liệu này được ghi lại với sự hỗ trợ của phần mềm Weka và lưu trữ vào ổ đĩa với độ phân giải lớn hơn 35 nano giây. Thời gian lấy mẫu là 24 giờ và dữ liệu được phân chia thành 10 tập con tương ứng với các thời điểm khác nhau trong ngày. Ngoài ra, cơ sở dữ liệu còn cung cấp 1 tập dữ liệu được thu thập trên cùng máy chủ sau 12 tháng để đánh giá độ ổn định của mô hình.

Cơ sở dữ liệu cung cấp 248 đặc trưng đối với mỗi luồng dữ liệu [11]. Các đặc trưng bao gồm các giá trị thống kê liên quan đến gói dữ liệu như chiều dài gói dữ liệu (packet length), khe thời gian giữa các gói dữ liệu (inter-packet timings) và các thông tin khác được suy ra từ giao thức lớp vận chuyển (transport protocol) TCP như số lượng SYN và ACK,... Nhiều thông tin thống kê về gói tin được trực tiếp suy ra bởi bộ đếm gói tin và kích thước của header. Các đặc trưng liên quan đến băng thông sử dụng (effective bandwidth utilization) được tính dựa trên entropy nhằm đánh giá đặc tính thông tin của luồng dữ liệu. Ngoài ra, các thông tin liên quan đến thời gian đến giữa các gói tin (2 chiều) được thể hiện thông qua 10 thành phần tần số của biến đổi Fourier.

Các luồng dữ liệu được phân loại thành 12 nhóm và đươc gắn nhãn. Việc gắn nhãn cho các luồng dữ liệu được thực hiện bằng tay thông qua giám sát các nội dung của các luồng dữ liệu cũng như các thông tin biết trước về hệ thống. Danh sách 12 nhóm và số lượng mẫu của từng nhóm được thể hiện ở Bảng 1.

Bảng 1 cho thấy, nhóm WWW (World Wide Web) chiếm số lượng mẫu đa số (84,9%), trong khi một số nhóm thiểu số như GAMES và INTERACTIVE lần lượt chỉ có 4 và 33 mẫu. Do đó, việc huấn luyện mô hình để phân loại được các nhóm thiểu số là khó khăn.

3.2. Điều kiện thiết lập và tiến hành thí nghiệm

Việc xử lý cơ sở dữ liệu và thực thi mô hình mạng nơ-ron được thực hiện trong môi trường Matlab. Thí nghiệm được thực hiện trên máy tính có cấu hình như sau: CPU Xeon E5-2630 - 2.40 GHz, RAM: 32 GB, GPU Nvidia Titan V 12Gb.

Theo định lý xấp xỉ tổng quát (Universal Approximation Theorem) [12], đối với một hàm số liên tục bất kỳ $f(x)$ và một số $\varepsilon > 0$, luôn luôn tồn tại một mạng nơ-ron một lớp ẩn với đầu ra có dạng $g(x)$ (với số nơ-ron lớp ẩn đủ lớn và hàm kích hoạt phù hợp) sao cho với mọi x, $|f(x) - g(x)| < \varepsilon$. Nói một cách khác, mạng nơ-ron một lớp ẩn có khả năng xấp xỉ hầu hết các hàm liên tục. Thông qua thực nghiệm và cùng với mục tiêu giảm thời gian thực thi của mô hình, chúng tôi nhận thấy rằng, mô hình mạng nơ-ron một lớp ẩn là phù hợp đối với cơ sở dữ liệu được lựa chọn. Do đó, trong phần thí nghiệm nhóm tác giả chỉ xem xét và đánh giá các kiến trúc mạng nơ-ron một lớp ẩn. Các siêu tham số cần điều chỉnh để đạt được mô hình mạng nơ-ron tối ưu bao gồm: số nơ-ron lớp ẩn, số đặc trưng đầu vào, loại hàm kích hoạt.

Để tránh vấn đề quá khớp, nhóm tác giả áp dụng kỹ thuật dừng học sớm (Early Stopping) và sử dụng thành phần ổn định hóa (l_2-norm regularization) với hệ số $\lambda = 10^{-3}$.

Tiêu chí đánh giá được lựa chọn là tỷ lệ phân loại đúng được tính bởi tỷ số giữa số mẫu được phân loại đúng trên tổng số mẫu. Tỷ lệ này được đánh giá không chỉ cho toàn bộ tập dữ liệu mà còn cho từng nhóm.

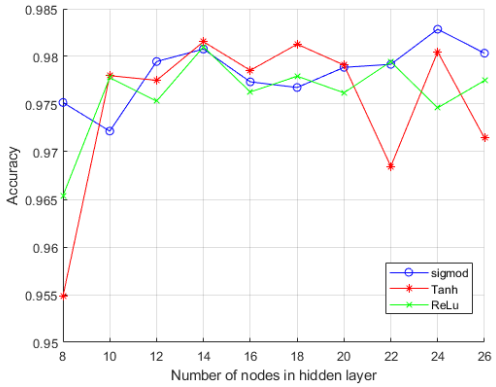
Bảng 1. *Danh sách các nhóm và số lượng mẫu của từng nhóm*

STT	Nhóm	Tần số xuất hiện	Tỷ lệ phần trăm (%)
1	MAIL	15789	8,277
2	FTP-CONTROL	2835	1,486
3	FTP-PASV	917	0,481
4	ATTACK	851	0,446
5	P2P	932	0,489
6	DATABASE	817	0,428
7	FTP-DATA	4729	2,479
8	MULTIMEDIA	543	0,285
9	SERVICES	1330	0,697
10	INTERACTIVE	33	0,017
11	GAMES	4	0,002
12	WWW	161968	84,912

3.3. Kết quả thí nghiệm và đánh giá

3.3.1. Khảo sát và đánh giá các kiến trúc mạng nơ-ron

Để đánh giá các kiến trúc mạng khác nhau, ta thay đổi số lượng nơ-ron lớp ẩn và loại hàm kích hoạt. nhóm tác giả xem xét 3 loại hàm kích hoạt được sử dụng phổ biến là Sigmoid, Tanh và ReLu. Hình 2 thể hiện tỉ lệ phân loại đúng đối với từng hàm kích hoạt khi thay đổi số nơ-ron lớp ẩn. Kết quả cho thấy, hàm kích hoạt Sigmoid cho kết quả phân loại tốt nhất. Do vậy, nhóm tác giả sử dụng hàm này cho các thí nghiệm tiếp theo để đánh giá mô hình.



Hình 2. *Sự phụ thuộc của kết quả phân loại vào số lượng nơ-ron lớp ẩn và hàm kích hoạt*

3.3.2. Khảo sát và đánh giá ảnh hưởng của số lượng đặc trưng

Trong phần này, tập các đặc trưng được chọn lọc sơ bộ dựa trên thuật toán Neighborhood Component Analysis (NCA). Tiếp theo, tập các đặc trưng được tiếp tục chọn lọc bằng phương thức chọn lọc đặc trưng tuần tự để chọn ra nhóm 10, 20, 30, 40 và 50 đặc trưng có ảnh hưởng nhất, được đặt tên lần lượt là Top 10, Top 20, Top 30, Top 40, Top 50. Nhóm tác giả cũng so sánh phương pháp chọn lọc đặc trưng này với phương pháp giảm số chiều không gian PCA (Principal Component Analysis). Biết rằng, PCA là thuật toán biến đổi không gian đặc trưng dựa trên phương sai của các giá trị đặc trưng, từ đó lựa chọn không gian với số chiều ít hơn để thể hiện dữ liệu [14]. Ngoài ra, mô hình mạng nơ-ron với toàn bộ số lượng đặc trưng (248) cũng được đưa vào so sánh. Bảng 2 thể hiện kết quả phân loại đối với các số lượng đặc trưng đầu vào khác nhau đối với tập huấn luyện và tập kiểm tra, và đối với tập kiểm tra được thu thập sau 12 tháng nhằm đánh giá độ ổn định của mô hình. Đối với mỗi trường hợp, nhóm tác giả lựa chọn số lượng nơ-ron lớp ẩn cho kết quả phân loại tốt nhất.

Bảng 2 cho thấy, sử dụng toàn bộ đặc trưng cho kết quả phân loại tốt nhất trên cả 2 tập huấn luyện và kiểm tra. Tuy nhiên đối với tập dữ liệu sau 12 tháng thì kết quả giảm đáng kể. Điều này có thể được giải thích là do sự dư thừa thông tin có thể dẫn đến việc mô hình học những đặc trưng không hữu ích nên kết quả phân loại giảm khi áp dụng cho dữ liệu (sau 12 tháng) có phân bố khác với tập huấn luyện. Bộ 70 đặc trưng cho kết quả ổn định nhất với độ chênh lệch sau 12 tháng vào khoảng 2,5%. Kiến trúc mạng đối với bộ đặc trưng này cũng là nhỏ nhất (12 nơ-ron lớp ẩn).

Bảng 2. *Bảng so sánh ảnh hưởng của số lượng đặc trưng*

Số lượng đặc trưng	Tỉ lệ phân loại đúng đối với tập huấn luyện (%)	Tỉ lệ phân loại đúng đối với tập kiểm tra (%)	Số nơ-ron lớp ẩn	Tỉ lệ phân loại đúng sau 12 tháng (%)
10	94,52	88,37	18	83,87
20	98,79	95,32	24	89,10
30	99,07	97,43	16	84,05
40	99,38	97,38	26	86,68
50	99,06	97,17	12	92,09
60	99,34	97,1	14	92,13
70	99,44	97,29	20	94,77
80	99,54	97,19	22	91,39
90	99,65	98,27	20	79,47
PCA	98,43	96,4	18	88,27
Tất cả (248)	99,73	98,29	24	79,46

3.3.3. So sánh với các mô hình khác

Nhóm tác giả tiến hành so sánh mô hình đề xuất với các mô hình cây quyết định (DCM) trong các nghiên cứu khác [2, 3]. Bảng 3 thể hiện kết quả phân loại của các mô hình đối với các tập dữ liệu huấn luyện, kiểm tra và sau 12 tháng. Kết quả cho thấy, mô hình đề xuất có độ ổn định cao hơn khi kiểm tra trên tập dữ liệu được thu thập sau 12 tháng.

Bảng 3. *So sánh các mô hình phân loại luồng dữ liệu*

Số lượng đặc trưng	Tỉ lệ phân loại đúng đối với tập huấn luyện (%)	Tỉ lệ phân loại đúng đối với tập kiểm tra (%)	Tỉ lệ phân loại đúng sau 12 tháng (%)
DCM	99,94	97,50	79,38
Mô hình đề xuất	99,44	97,29	94,77

Phân tích khả năng phân loại cụ thể đối với từng nhóm cho thấy, các mô hình DCM không phân loại được các nhóm thiểu số như INTERACTIVE (nhóm 10) và GAMES (nhóm 11) với tỉ lệ phân loại đúng lần lượt là 10,71% và 0,39% (Bảng 4). Biết rằng, số lượng mẫu của 2 nhóm này lần lượt là 33 và 4, được phân chia cho các tập huấn luyện và kiểm thử theo số lượng 23-10 và 3-1. Do áp dụng kỹ thuật gia tăng dữ liệu trong quá trình huấn luyện mô hình

mạng nơ-ron nên khả năng phân loại cho các nhóm thiểu số được cải thiện đáng kể, với tỉ lệ 50% cho nhóm INTERACTIVE và 100% cho nhóm GAMES.

***Bảng 4**. Tỉ lệ phân loại đúng đối với từng nhóm của mô hình cây quyết định (DCM) và mô hình đề xuất*

STT	Nhóm	DCM	Mô hình đề xuất
1	MAIL	97,73	96,72
2	FTP-CONTROL	100	84,25
3	FTP-PASV	85,71	94,54
4	ATTACK	19	54,90
5	P2P	61,71	58,57
6	DATABASE	100	90,20
7	FTP-DATA	100	90,83
8	MULTIMEDIA	96,36	80,98
9	SERVICES	96,41	91,97
10	INTERACTIVE	10,71	50,00
11	GAMES	0,39	100
12	WWW	99,65	98,87

4. Kết luận

Trong bài báo này, nhóm tác giả đã trình bày mô hình phân loại luồng dữ liệu dựa trên mạng nơ-ron. Để nâng cao tỷ lệ nhận dạng và tốc độ thực thi của mô hình, nhóm tác giả đã sử dụng các phương pháp chọn lọc đặc trưng để giảm số chiều không gian đặc trưng thích ứng với mô hình. Ngoài ra, kỹ thuật gia tăng dữ liệu cũng được áp dụng để nâng cao tỷ lệ phân loại đối với các nhóm có số lượng mẫu rất ít. Trong phần thí nghiệm, nhóm tác giả đã đánh giá ảnh hưởng của các yếu tố khác nhau (số nơ-ron lớp ẩn, hàm kích hoạt, số lượng đặc trưng, kỹ thuật gia tăng dữ liệu) vào hiệu năng phân loại của mô hình. Mô hình nghiên cứu cũng được đối sánh với các mô hình khác. Mô hình đề xuất cũng được đánh giá với các mô hình của các nghiên cứu khác. Kết quả cho thấy mô hình đề xuất có độ ổn định tốt hơn khi thực hiện kiểm tra với dữ liệu được thu thập sau 12 tháng. Ngoài ra, kết quả phân loại đối với các nhóm thiểu số của mô hình đề xuất cũng được cải thiện đáng kể so với các mô hình khác nhờ vào các các bước xử lý dữ liệu.

Lời cảm ơn: Chúng tôi gửi lời cảm ơn tới hãng NVIDIA đã gửi tặng GPU Titan V cho nhóm nghiên cứu CIVIC để phục vụ nghiên cứu này. Bài báo này được tài trợ bởi Trường Đại học Bách khoa – ĐHĐN với đề tài có mã số: T2019-02-26.

TÀI LIỆU THAM KHẢO

[1] Shaikh, Z. A., and D. Harkut. "An overview of network traffic classification methods". *Int. J. Recent Innovation Trends Comput. Commun* 3.2 (2015): 482-488.

[2] Li, Wei, et al. "Efficient application identification and the temporal and spatial stability of classification schema". *Computer Networks* 53.6 (2009): 790-809.

[3] Li, Wei, and Andrew W. Moore. "A machine learning approach for efficient traffic classification". *2007 15th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems*. IEEE, 2007.

[4] Auld, Tom, Andrew W. Moore, and Stephen F. Gull. "Bayesian neural networks for internet traffic classification". *IEEE Transactions on neural networks* 18.1 (2007): 223-239.

[5] Michael, Ang Kun Joo, et al. *Network traffic classification via neural networks*. No. UCAM-CL-TR-912. University of Cambridge, Computer Laboratory, 2017.

[6] Trivedi, Chintan, et al. *Classification of Internet traffic using artificial neural networks*. North Carolina State University. Center for Advanced Computing and Communication, 2002.

[7] Wang, Zhanyi. "The applications of deep learning on traffic identification". *BlackHat USA* 24 (2015).

[8] Singh, Kuldeep, S. Agrawal, and B. S. Sohi. "A near real-time IP traffic classification using machine learning". *International Journal of Intelligent Systems and Applications* 5.3 (2013): 83.

[9] Smit, Daniel, et al. "Looking deeper: Using deep learning to identify internet communications traffic". *2017 Australasian Conference of Undergraduate Research (ACUR)*. 2017.

[10] Vũ Hữu Tiệp, *Machine Learning cơ bản*, Nhà xuất bản khoa học và kỹ thuật, 2018.

[11] Andrew Moore, Denis Zuev and Michael Crogan, *Discriminators for use in flow-based Classification*, Department of Computer Science, University of London, ISSN 1470-5559

[12] Qin, Chen, et al. "Unsupervised neighborhood component analysis for clustering". *Neurocomputing* 168 (2015): 609-617.

[13] Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". *Neural networks* 2.5 (1989): 359-366.

[14] Jolliffe, Ian. *Principal component analysis*. Springer Berlin Heidelberg, 2011.

(BBT nhận bài: 23/4/2019, hoàn tất thủ tục phản biện: 13/5/2019)