TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 19, Số 11 (2022): 1878-1887
Website: https://journal.hcmue.edu.vn
ISSN: 2734-9918

HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 19, No. 11 (2022): 1878-1887
https://doi.org/10.54607/hcmue.js.19.11.3625(2022)

Bài báo nghiên cứu

DÒ TÌM BẤT THƯỜNG THIẾT BỊ ĐỊNH TUYẾN BẰNG KĨ THUẬT PHÂN LỚP

Nguyễn Quốc Huy
Trường Đại học Sài Gòn, Việt Nam
Tác giả liên hệ: Nguyễn Quốc Huy – Email: nqhuy@sgu.edu.vn
Ngày nhận bài: 11-10-2022; ngày nhận bài sửa: 18-11-2022; ngày duyệt đăng: 21-11-2022

TÓM TẮT

Phát hiện sớm tín hiệu bất thường của bộ định tuyến giúp dự đoán lỗi và có phương án thay thế kịp thời. Dữ liệu bất thường được phân tích thông qua dữ liệu cập nhật hoạt động của thiết bị. Bài báo đề xuất cách thức mới để phát hiện dữ liệu bất thường thông qua các kĩ thuật phân lớp dữ liệu. Dữ liệu BGL được sử dụng lại của tổ chức Usenix được gán nhãn theo kinh nghiệm của nhiều chuyên gia. Quá trình thực hiện bao gồm giai đoạn lựa chọn đặc trưng, huấn luyện mô hình, và kiểm thử. Kết quả khả quan khi các dự đoán lỗi hệ thống của các bộ định tuyến được phát hiện nhanh chóng và chính xác, và quan trọng là đã xác thực các đặc trưng được đặt giả thiết là quan trọng qua quá trình quan sát.

Từ khóa: phát hiện bất thường; kĩ thuật phân lớp; rút trích đặc trưng; phân loại dòng nhật kí; thiết bị định tuyến

1. Giới thiệu

Thiết bị định tuyến là một trong những thiết bị quan trọng trong các cơ quan, tổ chức có hạ tầng lớn về công nghệ thông tin. Các sự cố liên quan đến thiết bị định tuyến cần được dự đoán sớm và có phương án phòng ngừa, thay thế. Vì chi phí các loại thiết bị này không những mắc mà còn rất khó trong việc đặt mua. Chính vì vậy, việc đảm bảo các thiết bị này hoạt động ổn định và phát hiện, bảo vệ chúng khi có sự cố xảy ra là vô cùng quan trọng. Nhưng thiết bị định tuyến nổi tiếng là khó hiểu hoặc khó chẩn đoán, do sự không đồng nhất và bản chất các thiết bị là hộp đen. Cách phổ biến để hiểu rõ hơn về hệ thống bộ định tuyến và phát hiện các hành vi bất thường là kiểm tra các tập tin nhật kí của bộ định tuyến. Tuy nhiên, dữ liệu nhật kí rất khó kiểm tra vì dữ liệu vừa lớn, vừa không có cấu trúc, và đến từ nhiều nhà cung cấp khác nhau. Bên cạnh đó, độ tương quan giữa dữ liệu và việc dò tìm bất thường lại không cao (Yadav, 2020).

Hiện nay, đã có nhiều giải pháp để theo dõi hoạt động các thiết bị mạng, các cách tiếp cận thông thường để hiểu nhật kí hệ thống tập trung vào tìm kiếm từ khóa (chẳng hạn như

Cite this article as: Nguyen Quoc Huy (2022). Anomaly detection of router devices by classification techniques. *Ho Chi Minh City University of Education Journal of Science, 19*(11), 1878-1887.

lỗi và ngoại lệ) của các nhật kí có thể được liên kết với những hỏng hóc. Một cách tiếp cận như vậy là tốn thời gian và dễ xảy ra lỗi. Bài báo cũng cung cấp một phương pháp dự đoán hiệu quả thông qua dữ liệu từ các dòng nhật kí của thiết bị.

Các hoạt động bất thường của thiết bị có thể xuất phát từ lỗi định tuyến trong mạng, lỗi về phần cứng thiết bị, các cuộc tấn công DDOS hoặc lỗi tiến trình xử lí bên trong thiết bị. Các hoạt động bất thường này của thiết bị định tuyến thường được gửi đến nhật kí hệ thống máy chủ dưới dạng các dòng nhật kí nếu như thiết bị được cấu hình. Các tập dòng kí bất thường này thường đi kèm với nhiều dòng nhật kí bình thường khác nhằm mang lại nhiều thông tin hơn cho người quản trị. Đối với các hệ thống nhỏ, người quản trị mạng có thể định nghĩa các luật ràng buộc hoặc kiểm tra các dòng nhật kí của hệ thống để xác định các bất thường dựa trên các kiến thức của họ về hệ thống đang vận hành. Các từ khóa cũng có thể được sử dụng để việc tìm kiếm các dòng nhật kí bất thường được nhanh hơn. Như đã đề cập ở trên, kích thước vô cùng lớn của dữ liệu nhật kí được tạo ra (lên đến hàng triệu dòng nhật kí) bởi hàng trăm thiết bị khiến việc phân tích thủ công trở nên bất khả thi.

Do đó, một phương pháp phân tích các bản tin log tự động để xác định bất thường của thiết bị định tuyến như Cisco, Huawei, Dlink, và Juniper dựa trên các kĩ thuật phân lớp là vô cùng cần thiết. Mặc dù hiện nay, đã có các nghiên cứu về phân tích, phân loại các dòng nhật kí của nhiều thiết bị khác nhau. Phần thực nghiệm được thực hiện trên tập tin nhật kí bất thường BGL của tổ chức Usenix (Usenix, 2022) cho thiết bị định tuyến với các kĩ thuật phân lớp phổ biến. Thực nghiệm sẽ làm rõ về các giả thiết đặc trưng quan trọng qua quan sát khi dự đoán thủ công: Các từ quan trọng và Độ dài dòng nhật kí. Thông tin nhật kí của hệ thống sẽ được phân thành hai lớp riêng biệt: các dòng nhật kí bất thường và các dòng nhật kí bình thường.

2. Đối tượng và phương pháp nghiên cứu

Do dòng nhật kí chứa nhiều thông tin không cần thiết cho việc huấn luyện, nên những thông tin này cần được loại bỏ. Sau khi loại bỏ những thông tin cần thiết, thì những thông tin còn lại cần được biểu diễn theo một cấu trúc thích hợp để dễ dàng cho việc huấn luyện mô hình.

2.1. Tiền xử lí

Thiết bị định tuyến thường xuyên tạo ra các dòng nhật kí để ghi nhận lại trạng thái của thiết bị và thông tin thời gian hoạt động. Mỗi dòng bao gồm nội dung xác định điều gì đã được ghi nhận và nhãn thời gian. Những thiết bị này được cấu hình để gửi dữ liệu nhật kí về nhật kí hệ thống máy chủ, máy chủ này luôn luôn nhận những dữ liệu nhật kí từ cổng đã được cấu hình trước. Mặc dù định dạng của các dòng nhật kí không có chiều dài cố định, nhưng chúng có các đặc điểm chung. Những thông tin hữu ích này có thể được sử dụng để phân loại hoặc dùng cho những mục đích khác, do đó dữ liệu nhật kí được thu thập đầu tiên và lưu trên nhật kí hệ thống máy chủ dưới dạng dòng kí tự (tham khảo Hình 1).

```
14-07-2022 06:46:27 AM HNI-HED-MX5-01 last message
repeated 11730 times
14-07-2022 06:52:50 AM HCM-Q12-MX5 last message
repeated 397 times
14-07-2022 06:54:52 AM HCM-Q12-MX5 last message
repeated 64 times
14-07-2022 06:55:07 AM HNI-BDH-MX5 inetd[1379]:
/usr/sbin/sshd[37926]: exited, status 255
14-07-2022 07:04:06 AM QNH-HED-MX5-02 inetd[1616]:
/usr/sbin/sshd[12522]: exited, status 255
14-07-2022 07:07:18 AM HCM-Q12-MX5 last message
repeated 284 times
14-07-2022 07:10:02 AM TLC-MAN-AC-MX5-PE inetd[1652]:
/usr/sbin/sshd[96416]: exited, status 255
14-07-2022 07:19:10 AM TBH-HED-MX5 inetd[1627]:
/usr/sbin/sshd[12936]: exited, status 255
14-07-2022 07:21:32 AM VTU-HED-MX5 inetd[1606]:
/usr/sbin/sshd[52044]: exited, status 255
14-07-2022 07:28:07 AM TLC-MAN-AC-MX5-PE xntpd[1790]:
kernel time sync enabled 2001
14-07-2022 07:35:44 AM HCM-Q12-MX5 last message
repeated 17 times
```

Hình 1. *Nội dung tập tin nhật kí thiết bị định tuyến*

Như ta thấy ở ví dụ trên, các dòng nhật kí được bắt đầu bởi nhãn thời gian, tiếp đó là tên của thiết bị định tuyến với cùng định dạng và các đặc điểm chung khác: các tập tin nhật được viết bằng tiếng Anh, chúng được tạo thành bởi các số, chữ cái thường, chữ cái hoa và những kí tự đặc biệt khác.

Các dữ liệu nhật kí thô này cần được tiền xử lí trước khi sử dụng ở các bước tiếp theo. Các công việc tiền xử lí dữ liệu nhật kí được liệt kê cụ thể như sau:

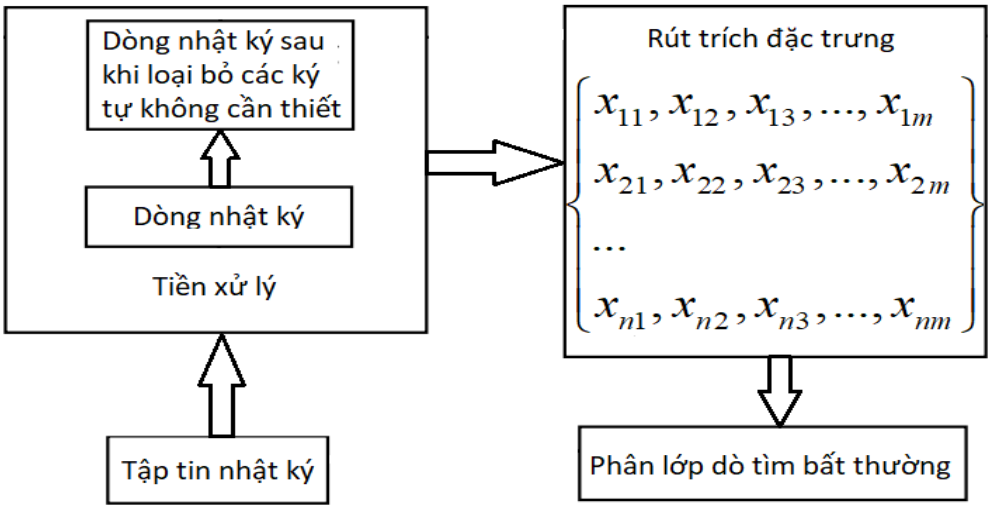
- Loại bỏ nhãn thời gian: loại bỏ toàn bộ nhãn thời gian trước mỗi bản tin (bao gồm cả ngày tháng và thời gian).
- Loại bỏ tên thiết bị: loại bỏ tên thiết bị trong bản tin.
- Loại bỏ các chữ số, kí tự đặc biệt: loại bỏ bất cứ kí tự đặc biệt nào, bao gồm dấu chấm câu và số.
- Thay thế các khoảng trắng liên tục bằng một khoảng trắng đơn: thay thế các khoảng trắng liên tục bằng một khoảng trắng đơn.
- Kí tự chữ thường: thay thế các chữ cái hoa bằng chữ cái thường tương ứng.

Mặc dù, nhãn thời gian trong các dòng nhật kí là một thông tin quan trọng cho người quản trị mạng (thể hiện thời gian mà sự kiện trong dòng nhật kí xảy ra), thông tin này vẫn được loại bỏ ra khỏi dòng nhật kí để thuận tiện cho việc xử lí. Sau khi phân loại vào các lớp khác nhau, các dòng nhật đầy đủ (bao gồm nhãn thời gian) sẽ được gửi tới người giám sát hệ thống.

2.2. Biểu diễn và xử lí dữ liệu

Bài báo này dựa trên một mô hình phát hiện bất thường có kiến trúc dựa trên LogEvent2Vec (He & Zhu, 2017) có một số điều chỉnh, bao gồm cả việc giảm đáng kể độ dài chuỗi tin và xác định tầm ảnh hưởng của thông số này trong quy trình phát hiện bất thường.

Sau đó phương pháp kiểm tra chéo được sử dụng phù hợp với các bộ dữ liệu không có độ cân bằng cao.



Hình 2. *Mô hình dò tìm bất thường từ tập tin nhật kí*

Quá trình thực nghiệm trên năm kĩ thuật phân lớp và chọn ra bộ phân lớp phù hợp nhất. Quá trình phát hiện bất thường được trình bày trong bài báo này thực hiện ba bước: phân tích cú pháp nhật kí, trích xuất tính năng mẫu nhật kí và phát hiện các trình tự bất thường (Hình 2). Xem bản tin nhật kí như là ngôn ngữ tự nhiên, mô hình này xử lí các mẫu nhật kí dưới dạng các từ và chuỗi của các mẫu nhật kí dưới dạng câu. Sau khi chia bộ nhật kí các mẫu thành chuỗi, mỗi mẫu trong chuỗi được biểu diễn dưới dạng vectơ (Zhao, 2018). Vectơ này được đưa ra làm đầu vào cho thuật toán phát hiện bất thường, cố gắng dự đoán liệu một điều bất thường có xảy ra hay không trong trình tự này.

Mô hình Bag-of-Words (BoW) và độ dài của dòng nhật kí được sử dụng để tính toán thông số này. Từ điển được sử dụng trong mô hình BoW được xây dựng dựa trên các dòng nhật kí bình thường đã được tiền xử lí trước đó. BoW là một mô hình đơn giản để tìm được số lượng từ của dòng nhật kí xuất hiện trong từ điển. Dựa vào BoW, mỗi dòng nhật kí được biểu diễn bằng một vector thưa có chiều dài bằng với chiều dài của từ điển. Mỗi phần tử của biểu thị số lần xuất hiện của từ đó trong dòng nhật kí. Các từ xuất hiện trong dòng nhật kí nhưng không được bao gồm trong từ điển sẽ không được tính khi thực hiện mô hình BoW. Dự liệu nhật kí sẽ trở thành ma trận như Hình 2.

Term frequency – inverse document frequency (TF-IDF) là một trong những thuật toán phổ biến trong khai thác dữ liệu văn bản. Phương pháp này dùng để tính toán trọng số của mỗi từ trong đoạn văn bản. Trong bài báo này, frequency thể hiện tần số xuất hiện của từ trong dòng nhật kí, inverse document frequency được sử dụng để tính toán mức độ quan trọng của từ đó trong tập tin nhật kí (Sokolova & Lapalme, 2009). Một vài sự cải tiến của thuật toán TF-IDF được đề cập trong (Guo & Yang, 2016). Công thức tính giá trị TF-IDF:

$$\omega_{ij} = tf_{ij} * \log \frac{n}{df_j} \quad (1)$$

Với ω_{ij} là trọng số TF-IDF của từ *j* trong dòng nhật kí *i*, *n* là kích thước tập tin nhật kí, tf_{ij} là tần suất xuất hiện của từ *j* trong dòng nhật kí *i*, df_j là số lượng dòng nhật kí chứa từ *j*.

Tổng giá trị TF-IDF của các từ trong dòng nhật kí là thành phần thứ ba của vector đặc trưng. Công thức dưới đây được dùng để tính tổng giá trị TF-IDF trong dòng nhật kí:

$$W_i = \sum_{j=1}^{h} \omega_{ij} \quad (2)$$

Với W_i là tổng giá trị TF-IDF của dòng nhật kí *i*, *h* là số lượng từ trong dòng nhật kí.

2.3. Các kĩ thuật dò tìm bất thường

Bài toán phát hiện dòng nhật kí bất thường được xem như bài toán phân lớp nhị phân. Với việc xử lí và biểu diễn dữ liệu phù hợp, dữ liệu đầu vào có thể áp dụng được trong bất cứ kĩ thuật phân lớp nhị phân nào. Bài báo thực nghiệm trên các kĩ thuật phân lớp tiêu biểu như: cây quyết định (DT), rừng ngẫu nhiên (RF), AdaBoost (AB), mạng nơron (MLP), và máy vectơ hỗ trợ (SVM).

Dữ liệu đầu vào của bộ phân loại nhị phân có 2 thành phần: một vectơ biểu diễn thông tin dòng nhật kí và thông tin đã được gán nhãn về việc liệu dòng nhật kí có bất thường hay không.

3. Kết quả và thảo luận

3.1. Tập dữ liệu

Bài báo sử dụng bộ dữ liệu Blue Gene / L (BGL) được thực hiện bởi Usenix. Nhật kí từ máy chủ này này đã được thu thập qua 215 ngày. Tập tin nhật kí có 4,747,963 dòng nhật kí, dung lượng 708 MB. Đây là loại dữ liệu mất cân bằng cao khi có 348.460 dòng nhật kí bất thường, chiếm 7,3% tổng số tập dữ liệu. Việc gán nhãn trong bộ dữ liệu BGL thông qua phương pháp lọc nhật kí bán tự động kết hợp với các thao tác thủ công của các người quản trị hệ thống. Các bất thường phổ biến nhất, chiếm 44% tổng số trường hợp (152.734 dòng nhật kí), là cảnh báo KERNDTLB mô tả việc ngắt lỗi TLB dữ liệu. 48% tiếp theo của tất cả các bất thường (168.011 dòng nhật kí) là loại KERNSTOR, APPSEV, KERNMNTF và ERNTERM. Còn lại 8% dòng nhật kí bất thường mô tả 36 loại cảnh báo hiếm khi xảy ra.

3.2. Môi trường thực nghiệm

Trình phân tích cú pháp Drain (IBM. Drain3, 2022) có thể cấu hình với một số tham số ảnh hưởng đến quá trình phân tích cú pháp nhật kí. Hai tham số quan trọng nhất là độ sâu tối đa của cây phân tích cú pháp và tham số độ tương tự, chính là ngưỡng tương tự của nhật kí so với độ tương tự mẫu phải lớn hơn nhật kí được gán cho nhóm mẫu cụ thể. Trong thực nghiệm, độ sâu cây tối đa là 3 và ngưỡng tương tự là 0,3, theo các đề xuất từ. Tham số xác định số lượng nút con tối đa trong cây phân tích cú pháp là 100. Đây là các thông số mặc định vì khá ổn định khi làm thực nghiệm.

Bảng 1. *Ảnh hưởng về độ dài dòng nhật kí*

STT	Độ dài dòng nhật kí	Số dòng nhật kí	
		Tập huấn luyện (70%)	Tập kiểm thử (30%)
1	10	332,357	142,439
2	20	166,178	71,220
3	50	66,579	28,533
4	100	33,178	14,219

Độ dài của dòng nhật kí cũng là là một thông tin quan trọng. Trong quá trình tìm hiểu các dòng nhật kí cho thấy rằng các dòng có độ dài đặc biệt nào đó có khả năng cao là bất thường hơn các dòng có độ dài khác. Gọi S_i là độ dài của dòng nhật kí và sử dụng khoảng trắng trong dòng nhật kí *i* sau khi đã được tiền xử lí để tính toán S_i.

Với *m* là độ dài của từ điển, mỗi hàng của ma trận thể hiện một dòng nhật kí trong dữ liệu nhật kí. Số lượng từ trong dòng nhật kí khác với từ điển được là độ dài của dòng nhật kí tính theo công thức :

$$L_i = S_i - \sum_{j=1}^{m} x_{ij} \tag{3}$$

Trong giai đoạn huấn luyện, bộ dữ liệu huấn luyện được phân tích theo Drain, tạo ra một tập các thông tin nhật kí có thể vectơ hóa. Tuy nhiên, bộ dữ liệu thử thì không nên làm như vậy vì kết quả sẽ bị sai lệch. Thay vào đó, mục tiêu của Drain là phân tích cú pháp các nhật kí đến và gán chúng vào các nhóm mẫu phù hợp nếu độ đo lớn hơn ngưỡng vì mô hình rút trích đặc trưng chỉ có thể biểu diễn các mẫu nhật kí dưới dạng có thể vectơ hóa trên tập dữ liệu huấn luyện. Quá trình phân tích cú pháp của tập dữ liệu huấn luyện đã tạo một từ điển chứa trung bình 1557 mẫu nhật kí. Trong giai đoạn kiểm thử mô hình, trình phân tích cú pháp đã xác định có 265 nhật kí không khớp với bất kì mẫu nào. Các mẫu nhật kí còn lại trong tập dữ liệu kiểm thử được gán vào các nhóm phù hợp có trong tập dữ liệu huấn luyện. Sau quá trình phân tích cú pháp nhật kí, các mẫu nhật kí được chia thành các chuỗi nhật kí không chồng chéo với bốn độ dài là 10, 20, 50 và 100 (xem Bảng 1). Việc chia dữ liệu thành hai tập huấn luyện và kiểm thử theo đúng quá trình kiểm tra chéo K-fold.

Ngoài Drain, môi trường thực nghiệm còn có sử dụng các thư viện trong ngôn ngữ lập trình Python như Re (xử lí biểu thức chính quy), NumPy, Pandas, fastText (thực hiện mô hình rút trích đặc trưng), Scikit-Learn (dùng các kĩ thuật phân lớp).

3.3. Phương pháp đánh giá

Dùng phương pháp ma trận lỗi (confusion matrix), nhưng trước tiên chúng ta thống nhất các thuật ngữ cơ bản sau:

- True positive (TP) — số dòng nhật kí lỗi được dự đoán chính xác.
- False positive (FP) —số dòng nhật kí bình thường được phân loại là bất thường (dự đoán sai).
- True negative (TN) —số dòng nhật kí bình thường được dự đoán chính xác.

- False negative (FN) —số dòng nhật kí lỗi được phân loại là bình thường.

Hiệu suất của các mô hình thử nghiệm được phân tích theo các độ đo:

$$Precision = \frac{Anomalies\ detected}{Anomalies\ reported}$$

$$Recall = \frac{Anomalies\ detected}{All\ anomalies}$$

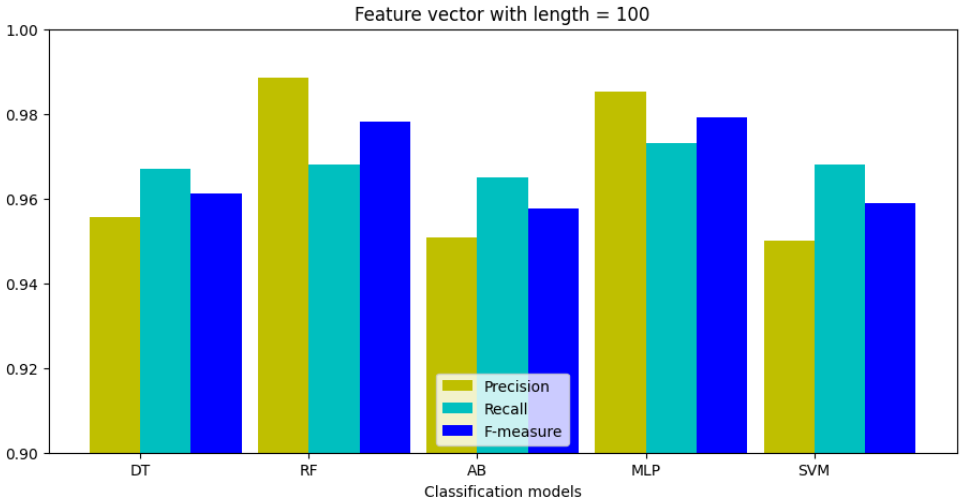
$$F\_measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

- Precision, cho biết tỉ lệ dự đoán đúng thực sự trong tất cả các dòng là lỗi theo mô hình dự đoán.
- Recall, cho biết tỉ lệ dự đoán đúng thực sự trong tất cả các dòng là lỗi theo thực tế.
- F-measure, trung bình điều hòa giữa 2 độ đo Precision và Recall.

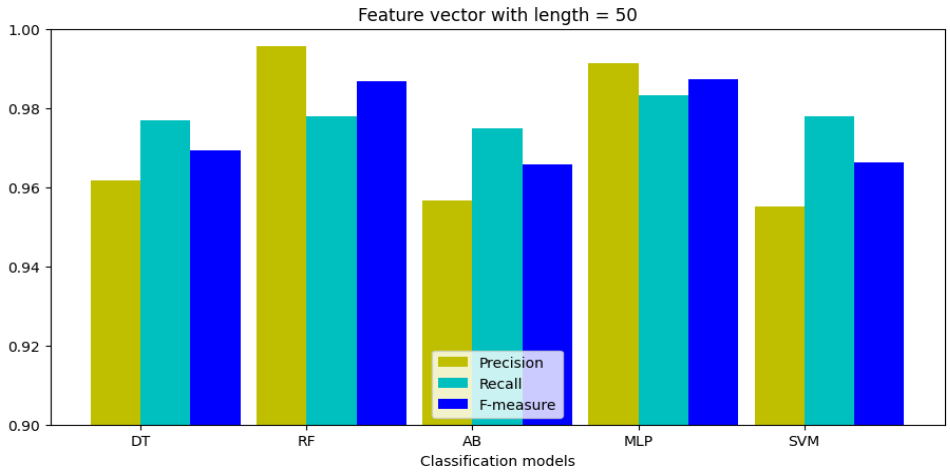
3.4. Kết quả thực nghiệm

Quá trình thực nghiệm đã chạy thử nhiều lần trên năm kĩ thuật phân lớp DT, RF, AB, MLP, và SVM (Ertam & Kaya, 2018) (với hàm hạt nhận RBF) của bộ trên bốn độ dài của vectơ đặc trưng biến đổi từ các dòng nhật kí. Hình 3 (trên đặc trưng có độ dài là 100), Hình 4 (trên đặc trưng có độ dài là 50), Hình 5 (trên đặc trưng có độ dài là 20), Hình 6 (trên đặc trưng có độ dài là 10) cung cấp tóm tắt các chỉ số đánh giá (Precision, Recall, và F-measure) là kết quả của các thử nghiệm trên tập dữ liệu BGL. Mỗi điểm số liệu đánh giá là một trung bình 10 lần thực nghiệm của các giá trị thu được trong quá trình quá trình đánh giá chéo.

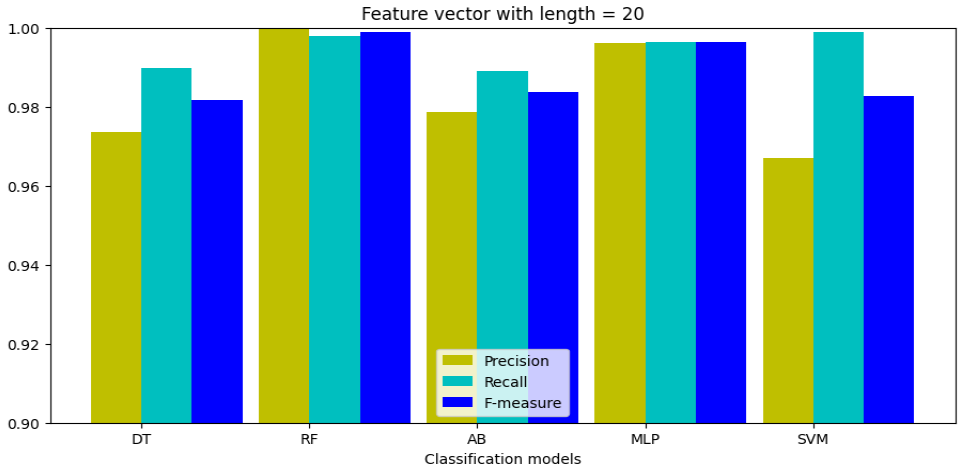
Rõ ràng, bộ phân loại MLP và RF đã đạt được hiệu suất tổng thể tốt nhất, với F-measure trên 98%. Các bộ phân lớp khác đạt điểm thấp hơn một chút, nhưng vẫn có các thước đo đánh giá rất tốt. Tất cả chúng đều đạt điểm F1 ở trên 95,5%. Tuy nhiên, thuật toán RF là thuật toán chính xác nhất (99,84% kết quả chính xác trong số tổng các dòng được dự đoán có lỗi).



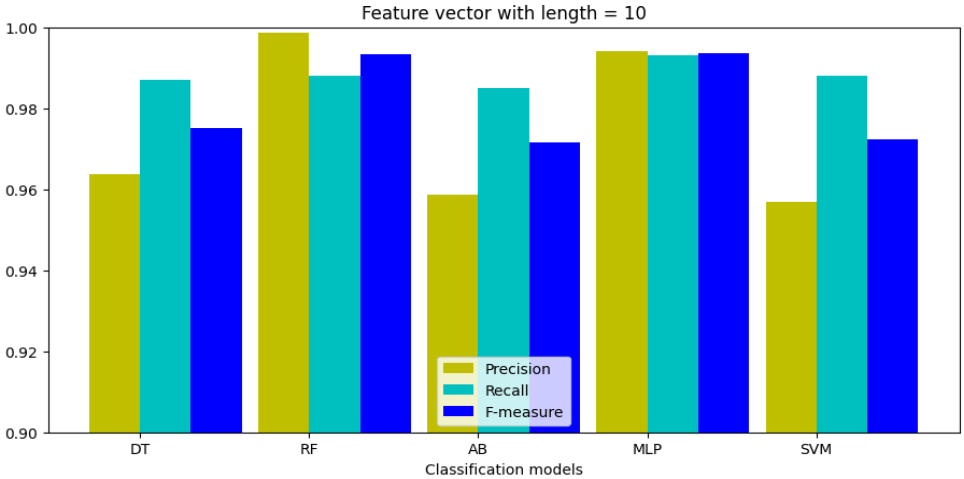
Hình 3. *Các mô hình với vectơ đặc trưng có độ dài 100*



Hình 4. *Các mô hình với vectơ đặc trưng có độ dài 50*

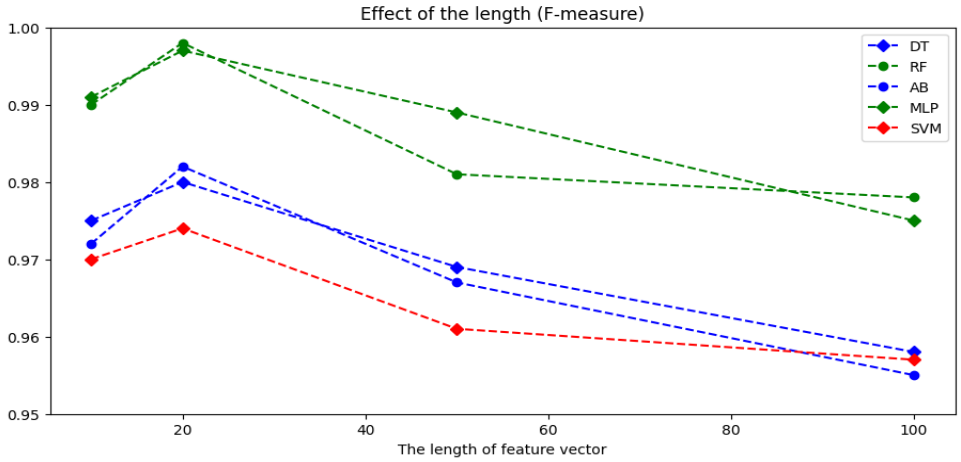


Hình 5. *Các mô hình với vectơ đặc trưng có độ dài 20*



Hình 6. *Các mô hình với vectơ đặc trưng có độ dài 10*

Hình 7 mô tả đặc trưng về độ dài của vectơ biểu diễn dòng nhật kí. Thông qua các độ đo Precision, Recall, F-measure trên các đăng trưng độ dài 10, 20, 50, 100. Nói chung, độ dài của vectơ biểu diễn dòng nhật kí có ảnh hưởng đến các độ đo, dài quá cũng không tốt và ngắn quá cũng không cho kết quả tối ưu. Qua yếu tố này, các kĩ thuật phân lớp nhị phân cần xem xét đến độ dài. Có những đặc trưng về lỗi nhiều khi ít quan trọng hơn đặc trưng độ dài của vectơ biểu diễn dòng nhật kí Các tính năng cá nhân chịu trách nhiệm đối với các dị thường ít quan trọng hơn trong các chuỗi dài hơn, do kết quả của sự biến đổi của vectơ chưa thật chính xác thông qua cách tính trung bình. Cũng thật may mắn trong năm kĩ thuật áp dụng thực nghiệm, độ dài của vectơ biểu diễn dòng nhật kí có giá trị lân cận 20 cho ra kết quả tối ưu hơn cả. Một số kĩ thuật khác cho ra kết quả không hội tự như mong muốn.



Hình 7. *Ảnh hưởng của đặc trưng chiều dài*

4. Kết luận

Mặc dù hiện nay đã có các nghiên cứu về phân tích, phân loại các dòng nhật kí của nhiều thiết bị khác nhau. Phần thực nghiệm trên tập tin nhật kí bất thường BGL của tổ chức Usenix cho thiết bị định tuyến với các kĩ thuật phân lớp phổ biến đã cho thấy giả thiết đặc trưng quan trọng qua quan sát khi dự đoán thủ công là đúng với thực nghiệm: Các từ quan trọng và Độ dài dòng nhật kí. Kết quả nghiên cứu chỉ ra một phương pháp để các nhà quản trị mạng có thể tạo ra ứng dụng để thao tác tự động trên dữ liệu nhật kí lớn và phức tạp.

Để kết quả được thuyết phục hơn, công việc trong tương lai về chủ đề này có thể kết hợp với các nhà quản trị hệ thống mạng để nhận các phản hồi thực tế nhằm so sánh với độ chính xác của mô hình. Ngoài ra, công việc tiếp theo có thể huấn luyện mô hình học trên dữ liệu không có bất thường nhằm huấn luyện các đặc trưng của một hệ thống bình thường và ổn định. Qua đó dễ dàng xác định những tín hiệu bất thường nằm ngoài đặc trưng mô hình và phát đi những cảnh báo nhẹ hơn.

❖ ***Tuyên bố về quyền lợi:*** *Tác giả xác nhận hoàn toàn không có xung đột về quyền lợi.*

TÀI LIỆU THAM KHẢO

Ertam, F., & Kaya, M. (2018). Classification of Firewall Log Files with Multiclass Support Vector Machine. In: A, Varol, M. Karabatak and C. Varol (editors). International Symposium on Digital Forensic and Security, 22-25, *Antalya, Turkey*. IEEE. Piscataway, New Jersey, 1-4.

Guo, A., & Yang, T. (2016). Research and improvement of feature words weight based on TFIDF algorithm. *In: Information Technology, Networking, Electronic and Automation Control Conference*, 20-22 May 2016, Chongqing, China. IEEE. 415-419.

He, P., Zhu, J., Zheng, Z.,& Lyu, M. R (2017). Drain: An Online Log Parsing Approach with Fixed Depth Tree. In *Proceedings of the IEEE International Conference on Web Services* (ICWS), Honolulu, HI, USA, 25-30 June 2017.

IBM. Drain3. Retrieved from https://github.com/IBM/Drain3 (accessed on 10 January 2022)

Sokolova, M., & Lapalm, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management, 45*, 427-437.

Usenix. The HPC4 Data. Retrieved from https://www.usenix.org/cfdr-data#hpc4 (accessed on 20 February 2022).

Yadav, R. B., Kumar, P. S., & Dhavale, S. V. (2020). *A Survey on Log Anomaly Detection using Deep Learning*. In *Proceedings of the 8th International Conference on Reliability, Infocom Technologies and Optimization* (Trends and Future Directions) (ICRITO). Noida, India, 4-5 June 2020, 1215-1220.

Ying, S., Wang, B., Wang, L., Li, Q., Zhao, Y., Shang, J., … Geng, J. (2021). An Improved KNN-Based Efficient Log Anomaly Detection Method with Automatically Labeled Samples. *ACM Trans. Knowl. Discov, 15*(3), 1-22.

Zhao, X., Wang, H. Xiao, & Chi, X. (2018). Improvement of the Log Pattern Extracting Algorithm Using Text Similarity. In*: International Parallel and Distributed Processing Symposium Workshops*, 21-25, May 2018, Vancouver, BC, Canada. IEEE. Los Alamitos, California, 507-514.

ANOMALY DETECTION OF ROUTER DEVICES BY CLASSIFICATION TECHNIQUES

Nguyen Quoc Huy

Saigon University, Vietnam

Corresponding author: Nguyen Quoc Huy – Email: nqhuy@sgu.edu.vn

Received: October 11, 2022; Revised: November 18, 2022; Accepted: November 21, 2022

ABSTRACT

Detecting early the anomaly signal of routers helps to predict errors and to prepare suitable solutions. Anomaly signals are analysed from the data log of devices. In this study, we have proposed an approach to detect anomaly signals from log files of routers using classification techniques. The log files BGL from the Usenix organization are collected and labelled based on the experience of many experts. Feature extraction is performed before training and testing the model. The results are efficient in almost realistic environments and especially confirm the assumption of the important features via our observation process.

Keywords: anomaly detection; classification techniques; feature extraction; log file classification; router devices