PRIVACY PRESERVING NAIVE BAYES CLASSIFIER FOR HORIZONTALLY PARTITIONED DATA

Nguyen Van Chung[1*], Nguyen Van Tao[2]
[1]*Vinh Phuc Technology - Economic College*
[2]*TNU - University of Information and Communication Technology*

ARTICLE INFO

Received: 12/10/2023
Revised: 06/11/2023
Published: 06/11/2023

KEYWORDS

Partitioned data
Horizontally partitioned
Privacy
Accuracy
Semi-honest model

ABSTRACT

The data mining process can reveal sensitive information about individuals or organizations thereby violating their privacy. The main purpose of the field of privacy preserving data mining is to develop various techniques to find valuable knowledge or information while still keeping sensitive data and private information for the owners. Up to now, there have been many solutions proposed, however these solutions either have low efficiency or do not ensure privacy. This article builds a privacy preserving Naive Bayes classifier solution in a multi-member classifier for horizontally partitioned data scenario based on the application of the secure sum protocol. The proposed protocol is assessed as good privacy, accuracy and efficiency in comparison to contemporary solutions. To confirm the effectiveness of the proposed solution, in the experimental part, the author used the python programming language to visualize the results. The author specifically, build a privacy preserving Naive Bayes classifier solution for the spam message detection model. Experimental results show that the proposed solution has good applicability in practice.

PHÂN LỚP NAIVE BAYES ĐẢM BẢO TÍNH RIÊNG TƯ CHO MÔ HÌNH DỮ LIỆU PHÂN TÁN NGANG

Nguyễn Văn Chung[1*], Nguyễn Văn Tảo[2]
[1]*Trường Cao đẳng Kinh tế - Kỹ thuật Vĩnh Phúc*
[2]*Trường Đại học Công nghệ thông tin và Truyền thông - ĐH Thái Nguyên*

THÔNG TIN BÀI BÁO

Ngày nhận bài: 12/10/2023
Ngày hoàn thiện: 06/11/2023
Ngày đăng: 06/11/2023

TỪ KHÓA

Phân lớp dữ liệu
Phân mảnh ngang
Tính riêng tư
Tính chính xác
Mô hình bán trung thực

TÓM TẮT

Quá trình khai phá dữ liệu có thể tiết lộ thông tin nhạy cảm về các cá nhân hoặc tổ chức vì thế xâm phạm quyền riêng tư của họ. Mục đích chính của lĩnh vực khai phá dữ liệu đảm bảo tính riêng tư là xây dựng các kỹ thuật khác nhau nhằm tìm kiếm ra các tri thức hoặc thông tin có giá trị trong khi dữ liệu và thông tin nhạy cảm vẫn được giữ riêng bởi các chủ sở hữu. Tính đến nay đã có nhiều giải pháp được đề xuất, tuy nhiên các giải pháp này có hiệu năng thấp hoặc chưa đảm bảo được tính riêng tư. Bài báo này xây dựng giải pháp phân lớp dữ liệu Naive Bayes đảm bảo tính riêng tư trong kịch bản dữ liệu phân tán ngang nhiều thành viên trên cơ sở ứng dụng giao thức tính Tổng bảo mật. Giao thức đề xuất được đánh giá tính riêng tư, tính chính xác và hiệu quả tốt so với các giải pháp hiện tại. Để khẳng định tính hiệu quả của giải pháp đề xuất, phần thực nghiệm tác giả sử dụng ngôn ngữ lập trình python để trực quan hóa kết quả. Cụ thể là xây dựng giải pháp phân lớp dữ liệu Naive Bayes đảm bảo tính riêng tư cho mô hình phát hiện tin nhắn rác, kết quả thực nghiệm cho thấy giải pháp được đề xuất có khả năng ứng dụng tốt vào thực tế.

DOI: https://doi.org/10.34238/tnu-jst.8980

[*] Corresponding author. *Email: nguyenvanchung.vtec@gmail.com*

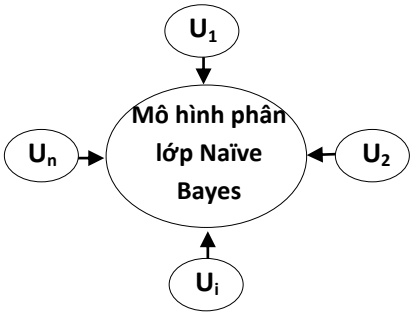
1. Giới thiệu

Gần đây, sự phát triển mạnh mẽ của các hệ thống thông tin cũng như các ứng dụng máy tính đã tạo ra một lượng dữ liệu lớn. Dựa trên dữ liệu này, các tổ chức hoặc cá nhân có thể xây dựng nên các mô hình học máy thông minh hỗ trợ giải quyết các vấn đề phức tạp đang gặp phải. Tuy nhiên, trong nhiều trường hợp, dữ liệu chứa các thông tin riêng tư hoặc nhạy cảm (ví dụ: thu nhập của khách hàng, quan điểm chính trị, bệnh tình của bệnh nhân…) khiến cho quá trình tạo ra mô hình học máy đối mặt với không ít khó khăn. Do đó, lĩnh vực học máy đảm bảo tính riêng tư đã ra đời từ đầu những năm 2000 và thu hút được nhiều sự quan tâm của cộng đồng nghiên cứu [1] - [4]. Mục tiêu của lĩnh vực nghiên cứu này là cho phép xây dựng các mô hình học máy trong khi những thông tin riêng tư, nhạy cảm tồn tại trong dữ liệu vẫn được bảo vệ an toàn theo một trong hai mô hình bán trung thực và độc hại (semi-honest, malicious models) [5]. Chính vì vậy, ba đặc trưng quan trọng của một giải pháp học máy đảm bảo tính riêng tư là: tính chính xác, tính riêng tư, và tính hiệu quả.

Tính đến nay, các giải pháp học máy đảm bảo tính riêng tư đã được đề xuất cho cả mô hình dữ liệu tập trung và phân tán (phân tán ngang, phân tán dọc, và phân tán tùy ý). Bên cạnh đó, để xây dựng nên một kỹ thuật học máy đảm bảo tính riêng tư, các nhà nghiên cứu tiếp cận theo ba hướng: ngẫu nhiên hóa, mật mã học, và lai hóa. Mỗi hướng tiếp cận ở đây đều có những ưu điểm và nhược điểm, cụ thể như sau:

Với các giải pháp theo hướng tiếp cận ngẫu nhiên hóa, chúng có thể áp dụng cho tất cả các mô hình dữ liệu và có hiệu quả cao. Tuy nhiên, do dữ liệu gốc đã được biến đổi nên những giải pháp này phải đánh đổi giữa tính riêng tư và độ chính xác của mô hình học máy. Tiếp theo là các giải pháp dựa trên các kỹ thuật mật mã, cụ thể là các giao thức tính toán bảo mật nhiều thành viên, sự riêng tư của dữ liệu có thể được bảo vệ nhờ các hệ mã hóa an toàn, nhưng chi phí tính toán của những giải pháp này tương đối lớn. Bên cạnh đó các giải pháp lai hóa cả hai hướng tiếp cận trên, một tham số thường được sử dụng để cân bằng giữa tính riêng tư và tính chính xác [6].

Trong nội dung này, bài báo xây dựng giải pháp phân lớp dữ liệu Naive Bayes có đảm bảo tính riêng tư trong kịch bản dữ liệu phân tán ngang, trong đó mỗi bên sở hữu dữ liệu U_i $(i = \overline{1,n})$ nắm giữ một số bản ghi như được mô tả trong Bảng 1 và mô hình phân lớp được mô tả trong Hình 1.



Hình 1. *Ví dụ về mô hình dữ liệu phân tán ngang*

Bảng 1. *Ví dụ về bảng dữ liệu phân tán ngang*

	Thời hạn	Lịch sử tín dụng	Mục đích vay	Giới tính	Tuổi	Tình trạng nhà ở	Nghề nghiệp	Hạng tín dụng
U_1	6	Đang vay của NH khác	Mua thiết bị âm thanh/ hình ảnh	Nữ	18-35	Nhà riêng	Hành chính	"Tốt"
	24	Đang vay của NH khác	Kinh doanh	Nam	18-35	Nhà riêng	Hành chính	"Tốt"
	42	Không có	Kinh doanh	Nam	18-35	Nhà riêng	Hành chính	"Tốt"
	18	Đúng hạn tới hiện tại	Mua thiết bị âm thanh/ hình ảnh	Nam	18-35	Nhà riêng	Hành chính	"Xấu"

	Thời hạn	Lịch sử tín dụng	Mục đích vay	Giới tính	Tuổi	Tình trạng nhà ở	Nghề nghiệp	Hạng tín dụng
	6	Đang vay của NH khác	Mua xe mới	Nữ	36-45	Nhà riêng	Phổ thông	"Tốt"
	18	Nợ quá hạn	Sửa chữa nhà cửa	Nam	36-45	Nhà riêng	Phổ thông	"Tốt"
	18	Đúng hạn tới hiện tại	Mua nội thất	Nữ	18-35	Nhà thuê	Hành chính	"Tốt"
...	24	Đúng hạn tới hiện tại	Mua thiết bị âm thanh/ hình ảnh	Nam	18-35	Nhà riêng	Hành chính	"Xấu"
	4	Đúng hạn tới hiện tại	Mua nội thất	Nữ	18-35	Nhà thuê	Phổ thông	"Tốt"
	24	Đúng hạn tới hiện tại	Mua xe mới	Nam	18-35	Nhà riêng	Hành chính	"Tốt"
	36	Đang vay của NH khác	Mua thiết bị âm thanh/ hình ảnh	Nam	36-45	Nhà riêng	Hành chính	"Tốt"
	12	Đang vay của NH khác	Mua nội thất	Nam	46-60	Nhà riêng	Hành chính	"Xấu"
U_n	10	Đúng hạn tới hiện tại	Đào tạo lại	Nữ	36-45	Nhà riêng	Hành chính	"Tốt"
	24	Đúng hạn tới hiện tại	Mua nội thất	Nữ	18-35	Nhà thuê	Phổ thông	"Xấu"
	12	Đang vay của NH khác	Mua nội thất	Nữ	36-45	Nhà thuê	Hành chính	"Tốt"

Hiện tại, đã có nhiều nghiên cứu liên quan tới giải pháp phân lớp dữ liệu Naive Bayes đảm bảo tính riêng tư trong kịch bản dữ liệu phân tán ngang được công bố bởi các nhà khoa học trên thế giới.

Năm 2003, Kantarcıoglu và cộng sự đã đề xuất bộ phân lớp Naïve Bayes có đảm bảo tính riêng tư lần đầu tiên [7] dựa trên giao thức tổng bảo mật đơn giản [8], [9]. Tuy nhiên, do giao thức này không đảm bảo an toàn trong trường hợp mỗi cặp thành viên thứ $(i+1)$ và $(i-1)$ thông đồng nên giải pháp đề xuất của Kantarcıoglu và cộng sự không thể bảo vệ được dữ liệu riêng tư của mỗi thành viên.

Sau đó, dựa trên hệ mã hóa đồng cấu ElGamal, Yang và cộng sự [10] cũng đề xuất phương thức phân lớp Naïve Bayes đảm bảo tính riêng tư cho mô hình dữ liệu phân tán đầy đủ - một trường hợp đặc biệt của mô hình phân tán ngang. Do sự an toàn của hệ mật ElGamal nên giải pháp của Yang có thể bảo vệ dữ liêu riêng tư của mỗi thành viên. Tuy nhiên, bên xây dựng mô hình phân lớp và các thành viên cần phải thực hiện giao thức khai phá tần suất nhiều lần khiến cho bộ phân lớp Naïve Bayes trong [10] có hiệu quả thấp.

Dựa trên giao thức tổng hợp dữ liệu đảm bảo tính riêng tư với công cụ làm nhiễu dữ liệu gốc, Huai và cộng sự đã mô tả giải pháp phân lớp dữ liệu Naïve Bayes đảm bảo tính riêng tư trong nghiên cứu [11]. Bên cạnh đó, giải pháp này còn sử dụng một thực thể tin cậy để sinh các tham số bí mật cần thiết. Kết quả, giải pháp của Huai và cộng sự phải đánh đổi giữa sự riêng tư của dữ liệu người dùng với mức độ chính xác của mô hình phân lớp.

Gần đây, P. Li cùng cộng sự [12] và T. Li cùng cộng sự [13] đã đề xuất hai giải pháp phân lớp Naïve Bayes đảm bảo tính riêng tư dựa trên các hệ mã hóa công khai DD-PKE và Paillier cùng với kỹ thuật làm nhiễu [6]. Hai giải pháp đề xuất này vừa có chi phí tính toán lớn, vừa phải đánh đổi giữa tính riêng tư của dữ liệu người dùng với mức độ chính xác của mô hình phân lớp.

Như vậy, các giải pháp phân lớp Naïve Bayes đảm bảo tính riêng tư cho mô hình dữ liệu phân tán ngang ở trên còn tồn tại nhiều hạn chế. Tiếp theo, bài báo đề xuất một giải pháp phân lớp Naïve Bayes đảm bảo tính riêng tư tổng quát cho phép các nhà phát triển ứng dụng có thể tùy chọn các tham số phù hợp với các yêu cầu thực tế về đảm bảo tính riêng tư và hiệu quả cho từng bài toán thự tế.

2. Phương pháp nghiên cứu

2.1. Bài toán phân lớp Naïve Bayes trong mô hình dữ liệu phân tán ngang có ràng buộc tính riêng tư

Trong bài toán phân lớp Naïve Bayes trong mô hình dữ liệu phân tán ngang, giả sử có l bên sở hữu dữ liệu $\{U_1, \dots, U_l\}$, trong đó:

- Mỗi U_i sở hữu một số bản ghi đã được gán nhãn có dạng $(u_{(1)}, \dots, u_{(m)}, \tau)$ với $(u_{(1)}, \dots, u_{(m)})$ là giá trị đặc trưng của m thuộc tính độc lập $\{A_1, \dots, A_m\}$ và τ là một nhãn thuộc tập các nhãn $\{L_1, \dots, L_r\}$.
- Mỗi thuộc tính độc lập A_j $(j = \overline{1,m})$ bao gồm r_j giá trị đặc trưng $\{a_j^1, \dots, a_j^{r_j}\}$. Ví dụ: thuộc tính '*Tuổi*' bao gồm ba loại giá trị '*18-40*', '*41-60*', '*> 60*'.

Để dự đoán nhãn của bản ghi mới $\{x_1, \dots, x_m\}$ là nhãn L_y nào trong số các nhãn $\{L_1, \dots, L_r\}$ sử dụng bộ phân lớp Naïve Bayes được xây dựng dựa trên toàn bộ bản ghi (ký hiệu là n) của l bên sở hữu dữ liệu ở trên, l bên sở hữu dữ liệu cùng nhau hợp tác để thực hiện tính toán theo công thức sau đây:

$$y = \underset{k=1,..,r}{\arg max} \left(p[k] \prod_{j=1}^{m} p[j,k]\right) \quad (1)$$

trong đó:

- $p[k]$ biểu diễn xác suất của nhãn L_k và $p[j,k]$ biểu diễn xác suất có điều kiện của từng giá trị trong thuộc tính x_j $(j = \overline{1,m})$ biết nhãn L_k $(k = \overline{1,r})$.
- y là chỉ số của nhãn L_k được mô hình phân lớp Naive Bayes dự đoán cho bản ghi mới, được hiểu là một trong các giá trị từ $1, \dots, r$ mà giá trị của hàm $p[k] \prod_{j=1}^{m} p[j,k]$ đạt giá trị lớn nhất.

Để thuận tiện khi làm việc với các giá trị xác suất nhỏ, công thức trên được biến đổi thành:

$$y = \underset{k=1,..,r}{\arg max} \left(\log p[k] + \sum_{j=1}^{m} \log p[j,k]\right) \quad (2)$$

Nếu ký hiệu $n[k]$ là số bản ghi có nhãn L_k $(k = \overline{1,r})$ và $n[j,k]$ là số bản ghi có thuộc tính thứ j là x_j $(j = \overline{1,m})$ đồng thời có nhãn là L_k $(k = \overline{1,r})$, ta có:

$$p[k] = \frac{n[k]}{n} \text{ và } p[j,k] = \frac{n[j,k]}{n[k]} \quad (3)$$

và: $$y = \underset{k=1,..,r}{\arg max} \left(\log n[k] - \log n + \sum_{j=1}^{m} (\log n[j,k] - \log n[k])\right) \quad (4)$$

Nói cách khác, để dự đoán nhãn cho bản ghi mới $\{x_1, \dots, x_m\}$ dựa trên mô hình phân lớp Naïve Bayes, các bên cần tính các giá trị $n[k]$ $(k = \overline{1,r})$ biểu diễn số bản ghi có nhãn tương ứng lần lượt là $L_1, \dots, L_r$ và các giá trị $n[j,k]$ biểu diễn số bản ghi có đặc trưng của thuộc tính thứ j là x_j và mang nhãn L_k $(j = \overline{1,m}, k = \overline{1,r})$. Trong ngữ cảnh mỗi bên sở hữu dữ liệu không tiết lộ dữ liệu mình nắm giữ, các bên cần phải cùng nhau tính toán các giá trị $n[k]$ và $n[j,k]$ trong khi không bên nào tiết lộ dữ liệu riêng tư của mình.

Như vậy, về tổng quát, để xây dựng mô hình phân lớp Naïve Bayes đảm bảo tính riêng tư trong mô hình phân tán ngang nhiều thành viên, các bên U_i $(i = 1, \dots, l)$ cần hợp tác cùng nhau tính toán bảo mật các giá trị sau:

- $n[k]$ $(k = \overline{1,r})$: số bản ghi có nhãn tương ứng lần lượt là $L_1, \dots, L_r$.
- $n[j^s,k]$ $(j = \overline{1,m}, s = \overline{1,r_j}, k = \overline{1,r})$: số bản ghi có đặc trưng của thuộc tính thứ j là a_j^s và mang nhãn L_k.

Về phía địch thủ, tương tự như trong mô hình tính toán nhiều thành viên tổng quát được trình bày trong [5], địch thủ được giả sử rằng kiểm soát một số thành viên nguy hại tham gia vào mô hình. Mục đích tấn công của địch thủ chủ yếu là khai thác dữ liệu riêng tư của các bên trung thực còn lại.

2.2. Giao thức phân lớp Naive Bayes đảm bảo tính riêng tư

Trong phần này, bài báo xây dựng giải pháp phân lớp dữ liệu Naive Bayes đảm bảo tính riêng tư theo mô hình bán trung trực (semi-honest model) phổ biến. Mỗi thành viên tham gia vào mô hình được giả sử là bán trung thực, nói một cách tường minh thì các thành viên này thực hiện theo hướng dẫn của giao thức nhưng khi kết thúc thì một số thành viên nguy hại có thể thông đồng để khai thác thông tin riêng tư của những thành viên trung thực.

Như đã đề cập ở trên, để xây dựng được mô hình phân lớp dữ liệu Naive Bayes có đảm bảo tính riêng tư, các bên cùng nhau tính toán các giá trị cần thiết bằng cách triển khai giao thức tính tổng bảo mật tổng quát (GSSP) được đề xuất trong [14]. Giao thức được trình bày như sau:

Giao thức phân lớp Naive Bayes đảm bảo tính riêng tư cho mô hình dữ liệu phân tán ngang:

- **Đầu vào:** n bản ghi dữ liệu được phân tán tại l bên sở hữu dữ liệu
- **Đầu ra**: Các giá trị $n[k]$, $n[j^s, k]$ với ($j = \overline{1,m}, s = \overline{1,r_j}, k = \overline{1,r}$)
- **Bước 1:** Mỗi bên sở hữu dữ liệu $U_i \in \{U_1, \dots, U_l\}$ chuẩn bị các giá trị đầu vào riêng tư từ tập bản ghi của mình.

  For i=1, .., l
  - Tính $n_i[k]$: số bản ghi U_i sở hữu mà có nhãn là $L_k (k = \overline{1,r})$.
  - Tính $n_i[j^s, k]$: số bản ghi U_i sở hữu mà có đặc trưng của thuộc tính thứ j là a_j^s và mang nhãn L_k ($j = \overline{1,m}, s = \overline{1,r_j}, k = \overline{1,r}$).
- **Bước 2:** Các bên cùng thực thi giao thức GSSP

  For i=1, .., l

  Thực thi giao thức GSSP
  - Tính các giá trị $n[k] = \sum_{i=1}^{l} n_i[k]$,
  - Tính $n[j^s, k] = \sum_{i=1}^{l} n_i[j^s, k]$ ($j = \overline{1,m}, s = \overline{1,r_j}, k = \overline{1,r}$).
- **Bước 3:**

  Bộ phân lớp Naïve Bayes được xây dựng dựa trên các giá trị xác suất: $p[k] = \frac{n[k]}{n}$, $p[j^s, k] = \frac{n[j^s,k]}{n[k]} (j = \overline{1,m}, s = \overline{1,r_j}, k = \overline{1,r})$.

3. Đánh giá giao thức đề xuất

3.1. Tính chính xác

Giao thức phân lớp Naive Bayes đảm bảo tính riêng tư được trình bày ở trên được cấu thành từ các giao thức GSSP. Mặt khác, tính đúng đắn của kết quả đầu ra của giao thức GSSP đã được chỉ ra trong [14], do đó các giá trị xác suất được sử dụng để xây dựng bộ phân lớp Naïve Bayes đề xuất được bảo toàn tính chính xác so với phương pháp phân lớp Naïve Bayes nguyên thủy.

3.2. Tính riêng tư

Mỗi giá trị tần suất được tính bởi một giao thức GSSP nên khả năng bảo vệ dữ liệu riêng tư của giao thức phân lớp Naive Bayes đề xuất tương đương với giao thức GSSP.

3.3. Tính hiệu quả

Để chỉ ra tính hiệu quả của giải pháp phân lớp Naïve Bayes đề xuất, bài báo ứng dụng giải pháp này cho bài toán thực tế xây dựng mô hình phát hiện tin nhắn rác (spam) với ràng buộc đảm bảo tính riêng tư của dữ liệu tin nhắn của những người cung cấp. Ngữ cảnh bài toán này được phát biểu như sau:

- Giả sử có 10 nhà mạng $\{U_1, \dots, U_{10}\}$, trong đó mỗi nhà mạng sẵn sàng sử dụng tập 500 tin nhắn ngắn bao gồm cả tin nhắn thông thường và tin nhắn rác. Cụ thể, mỗi nhà mạng dành 350 tin nhắn cho quá trình huấn luyện mô hình và 150 tin nhắn cho quá trình kiểm thử.

- Các nhà mạng này mong muốn phát triển công cụ phát hiện tin nhắn rác bằng mô hình phân lớp Naïve Bayes được xây dựng dựa trên tập tất cả các tin nhắn trên, trong khi mỗi nhà mạng không cung cấp tập tin nhắn của khách hàng của họ.

3.3.1. Mô tả dữ liệu thực nghiệm

Bộ dữ liệu thực nghiệm là 5000 tin nhắn tiếng Anh được lấy ra từ tập tin nhắn công khai trên Kaggle [15]. Mỗi tin nhắn trong bộ dữ liệu bao gồm hai thành phần chính: nội dung tin nhắn và nhãn tin nhắn (thông thường hoặc rác).

Bảng 2. *Thông tin bộ dữ liệu thực nghiệm*

Phân loại	Số lượng
Tổng số tin nhắn sử dụng	5000
Số tin nhắn thông thường	4311
Số tin nhắn rác	689

Như đã đề cập ở trên, 5000 tin nhắn này giả sử được lưu trữ phân tán tại cơ sở dữ liệu riêng của 10 nhà mạng $\{U_1, \dots, U_{10}\}$.

3.3.2. Tiền xử lý dữ liệu

Các nhà mạng xử lý mỗi tin nhắn trong tập dữ liệu của họ qua các bước sau:

Bước 1. Chuyển tất cả các ký tự sang dạng thường.

Bước 2. Loại bỏ dấu chấm câu, các từ phổ biến và các từ hiếm gặp.

Ngoài ra, khi làm việc với bộ phân lớp Naïve Bayes, mỗi tin nhắn thường được chuyển đổi sang một véc tơ đặc trưng Bag-of-Words (BOW). Tuy nhiên, trong ngữ cảnh ràng buộc về tính riêng tư, mỗi nhà mạng không tiết lộ tin nhắn khách hàng của họ. Điều này dẫn đến không thể xây dựng được bộ từ điển từ tập tất cả các tin nhắn phục vụ cho việc chuyển đổi dữ liệu cho mỗi tin nhắn đầu vào. Vì thế, trong phần này bài báo giả định 10 nhà mạng thống nhất sử dụng một kho từ phổ biến được thu thập từ các bài viết công khai trên trang Wikipedia có kích thước khoảng 6 Gigabytes.

Dựa trên bộ từ điển được trích xuất ra từ kho này, mỗi tin nhắn của các nhà mạng tiếp tục được chuyển đổi thành một véc tơ nhị phân đặc trưng 1500 chiều và kèm theo một giá trị nhãn "thông thường hoặc rác". Chú ý, số chiều của mỗi véc tơ đặc trưng phụ thuộc vào số từ trong bộ từ điển được lựa chọn sử dụng.

Như vậy, dựa vào phân tích ở mục 2.1, các nhà mạng cần cùng nhau thực thi 6002 lần giao thức GSSP để tính toán 6002 giá trị tần suất cần thiết để huấn luyện mô hình phân lớp Naïve Bayes.

3.3.3. Thiết kế và kịch bản thực nghiệm

Trong thí nghiệm này, các tính toán được cài đặt bằng ngôn ngữ lập trình Python sử dụng thư viện mật mã [16] tích hợp với môi trường phát triển Anaconda. Các chương trình thực nghiệm được thực thi trên một máy tính xách tay với hệ điều hành Windows 10, bộ vi xử lý Intel core i5-8250U xung nhịp 1,6GHz và 8GB bộ nhớ trong RAM.

Ngoài ra, để đảm bảo an toàn cho thông điệp truyền thông, mỗi cặp thành viên sử dụng công cụ mã hóa AES-256 và mã xác thực thông điệp HMAC-SHA256 được tích hợp trong bộ giao thức mật mã SSL/TLS.

Kịch bản thực nghiệm được diễn ra như sau:

- Đầu tiên, mỗi nhà mạng $U_1, \dots, U_{10}$ thực hiện pha 1 của giao thức được trình bày tại mục 2.2 với các giá trị t lần lượt là $2, 4, 6, 8$.

- Tiếp theo, mỗi nhà mạng $U_2, \dots, U_{10}$ thực hiện bước 1 của pha 2 của giao thức được trình bày tại mục 2.2. Chú ý, chúng ta không thể biết được mỗi nhà mạng $U_2, \dots, U_{10}$ nhận được chính xác bao nhiêu thông điệp từ các nhà mạng còn lại nên khối lệnh này sẽ thực thi với số lượng thông điệp nhận được trong trường hợp tốt nhất là 0 và xấu nhất là 9.

- Cuối cùng, nhà mạng U_1 thực hiện bước 2 của pha 2 của giao thức được trình bày tại mục 2.2.

3.3.4. Kết quả thực nghiệm

a. Thời gian thực thi

Trong 10 nhà mạng tham gia tính toán của giao thức GSSP, những nhà mạng $\{U_2, \dots, U_{10}\}$ thực hiện các tính toán tương tự nhau, chỉ riêng nhà mạng U_1 thực hiện khác ở pha thứ 2.

Bên cạnh đó, đối với giao thức GSSP, tham số t ở pha 1 được thử nghiệm với nhiều giá trị khác nhau, lần lượt là $2, 4, 6, 8$. Tại bước 1 trong pha 2 của giao thức GSSP, trường hợp xấu nhất là mỗi nhà mạng (ngoại trừ U_1) nhận được dữ liệu từ 9 nhà mạng còn lại, và ngược lại trường hợp tốt nhất là một nhà mạng không nhận được chia sẻ dữ liệu từ bất kỳ nhà mạng nào.

Các kết quả thực nghiệm được trình bày trong Bảng 3.

Bảng 3. *Kết quả thực nghiệm chương trình huấn luyện mô hình Naïve Bayes có đảm bảo tính riêng tư trên bộ dữ liệu tin nhắn*

Tham số t	Thời gian tính toán của U_1 (giây)	Thời gian tính toán của U_i (i=2, …, 10) (giây)	
		Trường hợp xấu nhất	Trường hợp tốt nhất
2	1,679	1,757	0,491
4	2,030	2,054	0,932
6	2,296	2,296	1,118
8	2,748	2,615	1,433

Chúng ta có thể thấy rằng, thời gian thực thi của hai nhóm nhà mạng là U_1 và $U_i (i = 2, \dots, 10)$ tương đối nhỏ. Cụ thể, trong trường hợp tham số $t = 2$, tổng thời gian tính toán của nhà mạng U_1 chưa tới 2 giây, còn thời gian tính toán của từng nhà mạng $U_i (i = 2, \dots, 10)$ trong trường hợp xấu nhất và tốt nhất tương ứng là $1{,}757$ giây và $0{,}491$ giây. Thậm chí ngay cả khi tham số t lên tới giá trị 8 thì thời gian tính toán của nhà mạng U_1 và thời gian này của từng nhà mạng U_i $(i = 2, \dots, 10)$ trong trường hợp xấu nhất chưa tới 3 giây.

Một điểm đáng chú ý nữa là thời gian thực hiện tính toán của mỗi nhà mạng tăng tuyến tính theo giá trị theo số t.

b. Độ chính xác phân lớp

Khi thử nghiệm mô hình phân lớp tìm được trên bộ dữ liệu kiểm thử (1500 tin nhắn), bài báo thống kê các chỉ số độ chính xác thông thường (accuracy), độ chính xác cân bằng (balanced_accuracy), và chỉ số F1_score lần lượt tương ứng là 0,97, 0,92, và 0,93. Kết quả trên hoàn toàn tương đương khi thực thi kỹ thuật phân lớp Naïve Bayes truyền thống với dữ liệu đầu vào là bộ dữ liệu tin nhắn được tiền xử lý tương tự như ở mục 3.3.2. Điều này hoàn toàn lô gic với những phân tích đánh giá lý thuyết ở mục 3.1.

4. Kết luận

Trong ngữ cảnh dữ liệu phân mảnh ngang, bài báo xây dựng giải pháp phân lớp dữ liệu Naïve Bayes đảm bảo tính riêng tư cho ứng dụng cụ thể là phân loại thư rác đảm bảo tính riêng tư áp dụng giao thức GSSP, với kịch bản có 10 nhà mạng cùng tham gia khai phá để phát triển công cụ phát hiện tin nhắn rác. Qua thực nghiệm với 5000 tin nhắn tiếng Anh được lấy từ các nguồn được trình bày trong mục 3.3.1 cho thấy giao thức GSSP thể hiện được tính hiệu quả khi được triển khai trên bộ dữ liệu và có thể khẳng định rằng giao thức này có khả năng ứng dụng rộng rãi trong các bài toán thực tế để xây dựng mô hình phân lớp Naïve Bayes có ràng buộc đảm bảo tính riêng tư.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, “Practical Secure Aggregation for Privacy-Preserving Machine Learning,” in *CCS’17*, ACM, 2017, pp. 1175–1191, doi: 10.1145/3133956.3133982.

[2] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, "Secure, privacy-preserving and federated machine learning in medical imaging," *Nat. Mach. Intell.*, vol. 2, no. 6, pp. 305–311, Jun. 2020, doi: 10.1038/s42256-020-0186-1.

[3] X. Zhou, K. Xu, N. Wang, J. Jiao, N. Dong, M. Han, and H. Xu, "A Secure and Privacy-Preserving Machine Learning Model Sharing Scheme for Edge-Enabled IoT," *IEEE Access*, vol. 9, pp. 17256–17265, 2021, doi: 10.1109/ACCESS.2021.3051945.

[4] E. Zorarpacı and S. A. Özel, "Privacy preserving classification over differentially private data," *WIREs Data Min. Knowl. Discov.*, vol. 11, no. 3, pp. 1–20, May 2021, doi: 10.1002/widm.1399.

[5] O. Goldreich, "Foundations of Cryptography," in *volume II, Basic Applications*, Cambridge University Press, 1998, p. 108.

[6] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, pp. 211–407, 2014.

[7] M. Kantarcioglu, J. Vaidya, and C. Clifton, "Privacy Preserving Naive Bayes Classifier for Horizontally Partitioned Data," *IEEE ICDM Workshop Priv. Preserv. Data Min.*, 2003, pp. 3–9.

[8] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu, "Tools for Privacy Preserving Distributed Data Mining," *ACM SIGKDD Explor. Newsl.*, vol. 4, no. 2, pp. 28–34, 2002, doi: 10.1145/772862.772867.

[9] B. Schneier, *Applied Cryptography*, 2nd ed. John Wiley & Sons, 1996.

[10] Z. Yang, S. Zhong, and R. N. Wright, "Privacy-Preserving Classification of Customer Data without Loss of Accuracy," in *Proceedings of the 2005 SIAM International Conference on Data Mining*, SIAM, 2005, pp. 92–102, doi: 10.1137/1.9781611972757.9.

[11] M. Huai, L. Huang, W. Yang, L. Li, and M. Qi, "Privacy-preserving Naive Bayes classification," in *International conference on knowledge science, engineering and management*, Springer, Cham, 2015, pp. 627–638, doi: 10.1007/978-3-319-25159-2_57.

[12] P. Li, T. Li, H. Ye, J. Li, X. Chen, and Y. Xiang, "Privacy-preserving machine learning with multiple data providers," *Future Gener. Comput. Syst.*, vol. 87, pp. 341–350, Oct. 2018, doi: 10.1016/j.future.2018.04.076.

[13] T. Li, J. Li, Z. Liu, P. Li, and C. Jia, "Differentially private Naive Bayes learning over multiple data sources," *Inf. Sci.*, vol. 444, pp. 89–104, May 2018, doi: 10.1016/j.ins.2018.02.056.

[14] V. C. Nguyen, "A general secure sum protocol," *J. Sci. Tech.*, vol. 11, no. 1, Jun. 2022, doi: 10.56651/lqdtu.jst.v11.n01.362.ict.

[15] Kaggle, "SMS Spam Collection Dataset," 2017. [Online]. Available: https://www.kaggle.com/uciml/sms-spam-collection-dataset. [Accessed Jun. 13, 2022].

[16] PyCryptodome, "Welcome to PyCryptodome's documentation — PyCryptodome 3.15.0 documentation," 2021. [Online]. Available: https://pycryptodome.readthedocs.io/en/latest/index.html. [Accessed Jul. 10, 2022].