ẨN TẬP PHỔ BIẾN DỰA TRÊN PHƯƠNG PHÁP QUY HOẠCH TUYẾN TÍNH NGUYÊN KẾT HỢP VỚI BIÊN DƯƠNG LÝ TƯỞNG

Nguyễn Thị Thu Tâm, Đinh Nguyễn Trọng Nghĩa*
Trường Đại học Công nghiệp Thực phẩm TP.HCM
*Email: *nghiadnt@hufi.edu.vn*
Ngày nhận bài: 07/4/2022; Ngày chấp nhận đăng: 08/6/2022

TÓM TẮT

Nghiên cứu này đề xuất một phương pháp để ẩn các tập phổ biến nhạy cảm trong cơ sở dữ liệu giao tác. Phương pháp này dựa trên thông tin từ biên dương lý tưởng và đề xuất một hệ phương trình quy hoạch tuyến tính nguyên. Thực hiện giải phương trình này sẽ xác định được các giao tác cần phải hiệu chỉnh để ẩn hoàn toàn các tập phổ biến nhạy cảm. Trong trường hợp phương trình vô nghiệm, một số hệ số được thêm vào để nới lỏng các ràng buộc của bài toán. Thực nghiệm đánh giá phương pháp trên một số tập dữ liệu nổi tiếng cho thấy phương pháp này có độ chính xác cao hơn phương pháp sử dụng quy hoạch tuyến tính nguyên truyền thống.

Từ khóa: Khai thác dữ liệu đảm bảo tính riêng tư, ẩn tập phổ biến, quy hoạch tuyến tính nguyên, biên dương lý tưởng.

1. MỞ ĐẦU

Khai thác dữ liệu đảm bảo tính riêng tư [1] là một lĩnh vực nghiên cứu mới kết hợp các phương pháp khai thác dữ liệu và riêng tư dữ liệu để thực hiện khai thác dữ liệu trong khi vẫn đảm bảo tính riêng tư. Sự gia tăng về lượng dữ liệu tạo nên động lực cho các ứng dụng khai thác dữ liệu, nhưng có thể ảnh hưởng tiêu cực đến sự riêng tư của các thông tin lưu trong dữ liệu. Cụ thể hơn, ta biết rằng các dữ liệu liên quan con người có thể làm lộ danh tính của họ, các dữ liệu liên quan kinh tế có thể chứa các dữ liệu mật của công ty như các kế hoạch kinh doanh, kế hoạch sáp nhập, mua lại công ty khác, v.v. Mục tiêu chính của khai thác dữ liệu đảm bảo tính riêng tư là thực hiện khai thác dữ liệu và đảm bảo không làm lộ thông tin nhạy cảm. Một số phương pháp đơn giản để bảo đảm tính riêng tư cổ điển như xóa danh tính hay các thuộc tính nhạy cảm là chưa đủ trong việc đạt được hiệu quả mong đợi.

Các kỹ thuật khai thác dữ liệu đảm bảo tính riêng tư có thể chia thành các phương pháp riêng tư đầu vào và riêng tư đầu ra. Các kỹ thuật riêng tư đầu vào nghiên cứu các kỹ thuật để bảo vệ dữ liệu nhạy cảm bằng cách biến đổi dữ liệu ban đầu để cho không có thông tin nhạy cảm lộ ra tường minh khi khai thác dữ liệu. Các phương pháp riêng tư đầu ra che chắn hoặc gây xáo trộn các mẫu kết quả, để cho các tri thức nhạy cảm bị cắt bỏ từ dữ liệu gốc. Trong số các kỹ thuật riêng tư đầu ra, các kỹ thuật ẩn tri thức như kỹ thuật hiệu chỉnh dữ liệu được sử dụng để biến đổi dữ liệu đầu vào theo cách mà tri thức chỉ định bị xóa hay ẩn đi từ dữ liệu. Phụ thuộc vào loại tri thức cần ẩn, có các dạng thuật toán khác nhau được sử dụng. Trong ngữ cảnh khai thác tập phổ biến, kỹ thuật được sử dụng trong bài báo là thuật toán ẩn tập phổ biến. Đó là kỹ thuật ẩn các tập có số lần xuất hiện của chúng trong cơ sở dữ liệu ít nhất lớn hơn một ngưỡng nào đó (gọi là minsup). Ngoài ra các tập này được đánh dấu là nhạy cảm và phải được ẩn đi từ cơ sở dữ liệu ban đầu.

Atallah và cộng sự đã đề xuất bài toán ẩn tập phổ biến, trong đó chủ yếu nghiên cứu các kỹ thuật hiệu chỉnh cơ sở dữ liệu ban đầu [2]. Kỹ thuật hiệu chỉnh này cần thỏa 3 điều kiện: (1) các tập phổ biến nhạy cảm không thể bị khai thác từ cơ sở dữ liệu đã hiệu chỉnh với cùng ngưỡng hỗ trợ minsup dùng để khai thác các tập trong cơ sở dữ liệu gốc, (2) chất lượng dữ liệu hiệu chỉnh đạt mức tối đa và (3) các tập phổ biến không nhạy cảm vẫn đạt trạng thái phổ biến trong cơ sở dữ liệu hiệu chỉnh.

Nhiều cách giải quyết gần đúng đã được đưa ra cho vấn đề này. Từ các kỹ thuật Heuristic [3], kỹ thuật khai thác lý thuyết biên [4, 5] đến các kỹ thuật dựa trên Deep Learning [6], để đạt được việc ẩn trong khi tác dụng phụ là nhỏ nhất. Các kỹ thuật quy hoạch tuyến tính nguyên (Linear-programming - ILP) [7-9] xây dựng vấn đề như là bài toán quy hoạch tuyến tính trong đó lời giải của nó chỉ ra các giao tác nào hay các hạng mục nào nên hiệu chỉnh. Mặc dù các kỹ thuật heuristic dễ hiểu và khả năng mở rộng tốt hơn các kỹ thuật dựa trên ILP, các kết quả của nó ít tin cậy hơn do có những tác dụng phụ. Mặc khác, các kỹ thuật dựa trên ILP khó hơn do phải giải hệ các phương trình quy hoạch tuyến tính. Tuy nhiên, các kỹ thuật này lại đáng tin cậy và tìm được các lời giải tốt hơn. Mặc khác, có một số khó khăn vốn có của bài toán và không có gần đúng nào đảm bảo giải pháp tối ưu tồn tại cho các kỹ thuật nêu trên. Ví dụ, trong một cơ sở dữ liệu giao tác với các tập phổ biến *abc* và *bc*, nếu *bc* là nhạy cảm, mọi cố gắng ẩn nó đều dẫn đến ẩn cả tập *abc*. Chính vì vậy không có giải pháp lý tưởng tồn tại trong bài toán này.

Nhìn chung, các phương pháp ẩn tập phổ biến có thể được chia thành 3 nhóm chính: phương pháp heuristic, phương pháp chính xác và phương pháp tiến hóa. Phương pháp heuristic có ưu điểm là tính toán hiệu quả và thời gian tính toán khá nhanh. Tuy nhiên, phương pháp này có một khuyết điểm lớn là có khả năng làm xuất hiện các hiệu ứng phụ không mong muốn. Phương pháp dựa trên thuật toán tiến hóa là phương pháp phổ biến nhất hiện nay. Phương pháp này đưa các hiệu ứng phụ vào hàm thích nghi của thuật toán và thực hiện tối ưu hóa dựa trên một thuật toán tiến hóa (GA, PSO, …). Các thuật toán tiến hóa về cơ bản là tìm kiếm lời giải gần đúng trong một không gian tìm kiếm. Việc tìm ra lời giải tối ưu khá nhanh nhưng cũng có khả năng lời giải này không thể ẩn được hoàn toàn các tập phổ biến. Phương pháp chính xác là một phương pháp dựa trên một hệ phương trình điều kiện (ví dụ hệ ILP). Như tên gọi, phương pháp này thông thường cho kết quả ẩn được hoàn toàn các tập phổ biến. Tuy nhiên, tài nguyên sử dụng cho phương pháp này khá lớn do phải giải các hệ phương trình điều kiện. Ngoài ra, do đưa vào hệ phương trình quá nhiều ràng buộc, một số phương pháp dựa trên ILP được đưa ra nhưng lại không có nghiệm khi giải hệ ILP.

Trong nhóm phương pháp sử dụng ILP phải kể đến Menon và cộng sự [9]. Đây là nhóm đầu tiên đưa ra một công thức quy hoạch tuyến tính nguyên cho bài toán ẩn tập phổ biến. Nghiệm của ILP chỉ ra các giao tác nào cần phải hiệu chỉnh để ẩn các tập nhạy cảm. Quá trình hiệu chỉnh được xem xét độc lập và không phụ thuộc nghiệm của bài toán quy hoạch. Công thức của phương pháp này bỏ qua hoàn toàn sự ảnh hưởng của quá trình hiệu chỉnh lên các tập phổ biến không nhạy cảm. Sau đó, một số công trình cũng đề xuất cải tiến phương pháp của Menon nhưng chưa thu được kết quả khả quan. Kagklis và cộng sự cải tiến công thức Menon chuyển tập *S* thành *Min(S)* để giảm bớt số lượng phương trình ràng buộc [7]. Tuy nhiên, công trình của Kagklis chỉ cải tiến được về thời gian tính toán mà không cho kết quả nào tốt hơn. Sau đó, một số công trình khác đưa thêm thông tin về biên dương lý tưởng vào điều kiện ràng buộc của công thức ILP. Tuy nhiên, sự xuất hiện thêm các điều kiện ràng buộc dẫn đến phương trình không có lời giải [14]. Stavropoulos và cộng sự đã bổ sung thêm một số số hạng nới lỏng ràng buộc [14]. Tuy nhiên, kiểm tra lại trên một số tập dữ liệu cho thấy công thức của Stavropoulos vẫn xuất hiện trường hợp vô nghiệm.

Với bài toán ẩn tập phổ biến, một đóng góp quan trọng có thể là một cách tiếp cận trong đó cho tập hợp các tập phổ biến và một tập hợp các tập nhạy cảm, nó tạo nên biên lý tưởng giữa hai tập hợp. Trong công trình này, mục tiêu là sử dụng kết hợp kỹ thuật biên với phương

pháp quy hoạch tuyến tính nguyên áp dụng để ẩn các tập nhạy cảm. Lời giải của phương trình xác định tập các giao tác cần phải hiệu chỉnh để cho sự ẩn có thể đạt được với độ chính xác cao nhất. Nếu không tồn tại lời giải, ta nới lỏng các ràng buộc cần thỏa mãn, để cho cơ sở dữ liệu sau hiệu chỉnh vẫn đảm bảo tính riêng tư với ảnh hưởng ít nhất đến chất lượng dữ liệu.

2. PHÁT BIỂU BÀI TOÁN VÀ ĐỀ XUẤT THUẬT TOÁN

2.1. Phát biểu bài toán

Cho $I = \{i_1, i_2, \dots, i_r\}$ là một tập hữu hạn các hạng mục (item). Cơ sở dữ liệu D là một tập hợp các giao tác $D = \{T_1, T_2, \dots, T_n\}$, trong đó mỗi giao tác $T_q \in D$, với T_q là một tập con I, và T_q có một định danh duy nhất q được gọi là định danh giao tác (Transaction IDentifier – TID). Ngưỡng hỗ trợ tối thiểu δ, được giả định là do người dùng hoặc chuyên gia đặt theo cách thủ công (theo tỷ lệ phần trăm).

Bảng 1. Một ví dụ về Cơ sở dữ liệu giao tác D_0

Tid	Itemset
1	B, D
2	B, C, D, E, F
3	A, B, C, D, F
4	A, B, C, E, F
5	A, B, C, D
6	B, C, D

Định nghĩa 1. Bài toán khai thác tập phổ biến là bài toán thực hiện trích xuất tất cả các tập hạng mục (itemset) có độ hỗ trợ (support count) không nhỏ hơn độ hỗ trợ tối thiểu (min support).

Định nghĩa 2. (Bài toán ẩn tập phổ biến) Cho trước một cơ sở dữ liệu giao tác D có tập các hạng mục $I = \{i_1, i_2, \dots, i_r\}$, một độ hỗ trợ tối thiểu σ và một tập các tập phổ biến nhạy cảm S. Thực hiện biến đổi D thành D' sao cho $sup_{D'}(X) < \sigma$ với mọi $X \in S$. Trong đó $sup_{D'}(X)$ là độ hỗ trợ của tập X khi tính trên cơ sở dữ liệu mới D'.

Điều này có nghĩa là cơ sở dữ liệu ban đầu phải được hiệu chỉnh sao cho không thể khai thác được các tập phổ biến nằm trong tập hợp các tập nhạy cảm bằng bất cứ thuật toán khai thác tập phổ biến nào. Vấn đề đặt ra cho bài toán là phải xác định được các giao tác nào cần hiệu chỉnh để ẩn hoàn toàn các tập nhạy cảm đồng thời lượng mất mát thông tin được giữ thấp nhất. Lượng mất mát thông tin có thể đo đạc thông qua số các tập phổ biến không nhạy cảm bị mất đi trong quá trình hiệu chỉnh dữ liệu.

2.2. Đề xuất thuật toán

2.2.1. Định nghĩa biên

Định nghĩa 3. (Định nghĩa biên) Biên dương $Bd^+(F_D^\sigma)$ là tập hợp tất cả các tập phổ biến tối đại (các tập phổ biến nhưng không có tập nào cha của chúng là phổ biến). Biên âm $Bd^-(F_D^\sigma)$ là tập hợp tất cả các tập không phổ biến tối tiểu (các tập không phổ biến nhưng không có tập con nào của chúng là không phổ biến).

$$Bd^+(F_D^\sigma) = \{X \in \mathcal{F}_D^\sigma \mid \forall Y, X \subset Y \Rightarrow Y \notin \mathcal{F}_D^\sigma\}, \text{ và}$$
$$Bd^-(F_D^\sigma) = \{X \notin \mathcal{F}_D^\sigma \mid \forall Y, Y \subset X \Rightarrow Y \in \mathcal{F}_D^\sigma\} \quad (1)$$

Ví dụ với cơ sở dữ liệu giao tác ví dụ trong Bảng 1 và độ hỗ trợ $\sigma = 3$, biên dương và biên âm là

$$Bd^+(F_D^\sigma) = \{\{B, C, F\}, \{B, C, D\}, \{A, B, C\}\}$$
$$Bd^-(F_D^\sigma) = \{\{E\}, \{A, D\}, \{A, F\}, \{D, F\}\}$$

2.2.2. Siêu đồ thị và đường ngang tối tiểu

Định nghĩa 4. (Siêu đồ thị) Một siêu đồ thị $H = (V, E)$ là một tập hữu hạn các siêu cạnh $E = \{E_1, E_2, \dots, E_m\}$, mỗi siêu cạnh là một tập hữu hạn của các đỉnh $V = \{v_1, v_2, \dots, v_n\}$ sao cho $E_i \neq \emptyset, i = 1..m$ và $\bigcup_{i=1}^{m} E_i = V$.

Định nghĩa 5. (Phần bù) Phần bù của siêu đồ thị H ký hiệu là $H^c = (V, E^c)$ trong đó $E^c = \{V \backslash e \mid e \in E\}$

Định nghĩa 6. (Siêu đồ thị đơn) Siêu đồ thị H được gọi là siêu đồ thị đơn nếu với mỗi cặp $E_i, E_j \in E, E_j \subseteq E_i \Rightarrow j = i$.

Định nghĩa 7. (Đường ngang của siêu đồ thị) Đường ngang (transversal) của siêu đồ thị là một tập $T \subseteq V$ sao cho $T \cap E_i \neq \emptyset \; \forall E_i \in E$. Điều đó có nghĩa là đường ngang này sẽ cắt ngang mọi siêu cạnh của siêu đồ thị.

Định nghĩa 8. (Đường ngang tối tiểu) Một đường ngang của siêu đồ thị được gọi là tối tiểu nếu không có một tập con nào của nó là đường ngang của siêu đồ thị đó.

Định nghĩa 9. (Siêu đồ thị ngang) Tập hợp tất cả các đường ngang tối tiểu của siêu đồ thị H được gọi là siêu đồ thị ngang (transversal hypergraph) của H, ký hiệu *Tr(H)*.

2.2.3. Áp dụng siêu đồ thị ngang

Gọi R là hàm chuyển tập hợp các tập hạng mục thành một siêu đồ thị, trong đó mỗi tập hạng mục tương ứng với một siêu cạnh và tập các hạng mục tương ứng với tập các đỉnh. Đặt

$$H^+ = R\{Bd^+(F)\}$$
$$H^- = R\{Bd^-(F)\} \quad (2)$$

Với H^+ và H^- là các siêu đồ thị tương ứng với biên dương và biên âm của tập dữ liệu ban đầu.

Theo Stavropoulos và cs, biên âm của một cơ sở giao tác có thể thu được bằng cách lấy Siêu đồ thị ngang của phần bù biên dương của chính cơ sở dữ liệu đó [14]. Điều đó có nghĩa là

$$H^- = Tr[(H^+)^c] \quad (3)$$

Áp dụng tính chất $Tr[Tr(H)] = H$ và $(H^c)^c = H$, sẽ thu được hệ thức tương ứng cho biên dương

$$H^+ = \left(Tr(H^-)\right)^c \quad (4)$$

Bằng cách biểu diễn các tập hạng mục của cơ sở dữ liệu giao tác dưới dạng các siêu cạnh của siêu đồ thị, có thể có được biên âm hoặc biên dương nếu có được thông tin của biên còn lại sử dụng một số tính chất đặc biệt của siêu đồ thị.

2.2.4. Tính các biên lý tưởng

Định nghĩa 10. (Toán tử *Min*) Toán tử *Min* trên một tập S, ký hiệu là *Min(S)*, là toán tử chỉ giữ lại các tập tối tiểu trong S, loại bỏ các tập không tối tiểu. Một tập được gọi là tối tiểu trong S nếu không có tập con nào của nó tồn tại trong S.

Thuật toán 1. Thuật toán để tính toán biên dương lý tưởng của cơ sở dữ liệu hiệu chỉnh

Input: Tập nhạy cảm *S*, Biên âm $Bd^-(F)$

Đặt $H^- = R(Bd^-(F))$ và $H_S = R(S)$;

Tính siêu đồ thị ngang $\widehat{H^-} = Min(H^- \cup H_S)$;

Gọi $THG(\widehat{H^-})$ để tính $Tr(\widehat{H^-})$;

Tính $\widehat{H^+} = \left(Tr(\widehat{H^-})\right)^c$

Đặt $Bd^+(F) = R^{-1}(\widehat{H^+})$

Output: $Bd^+(F)$

Ví dụ với cơ sở dữ liệu mẫu ở Bảng 1. Giả sử tập nhạy cảm được chọn là $S = \{\{A, B, C\}, \{C, D\}, \{B, C, F\}\}$. Biên âm lý tưởng có thể được xác định như sau

$$\widehat{H^-} = Min(H^- \cup H_S) = \{\{E\}, \{A, D\}, \{A, F\}, \{D, F\}, \{A, B, C\}, \{C, D\}, \{B, C, F\}\}$$

Dễ dàng thu được siêu đồ thị ngang

$$Tr(\widehat{H^-}) = \{\{A, C, D, E\}, \{A, C, E, F\}, \{C, D, E, F\}, \{A, D, E, F\}, \{A, B, D, E\}, \{B, D, E, F\}\}$$

Do đó, biên dương lý tưởng là

$$\widehat{H^+} = \left(Tr(\widehat{H^-})\right)^c = \{\{B, F\}, \{B, D\}, \{A, B\}, \{B, C\}, \{C, F\}, \{A, C\}\}$$

2.2.5. Công thức ILP đề xuất để xác định giao tác cần hiệu chỉnh

$$\min \sum_{\forall i: T_i \in D} x_i + \sum_{\forall j: X_j \in Bd^+(\hat{F})} s_j, \tag{5}$$

$$s.t. \sum_{\forall i: T_i \in D} a_{ij} x_i \geq sup_D(X_j) - \sigma + 1, \forall X_j \in Min(S), \tag{6}$$

$$\sum_{\forall i: T_i \in D} a_{ij} x_i - s_j . sup_D(X_j) \leq sup_D(X_j) - \sigma, \forall X_j \in Bd^+(\hat{F}), \tag{7}$$

$$x_i \in \{0,1\}, \forall i: T_i \in D \tag{8}$$

$$s_j \in \{0,1\},, \forall j: X_j \in Bd^+(\hat{F}) \tag{9}$$

Trong đó, các biến x_i tương ứng với các giao tác trong cơ sở dữ liệu ban đầu. Nếu nghiệm cho $x_i = 1$, có nghĩa là giao tác thứ *i* cần được hiệu chỉnh. Các biến s_j tương ứng với tập X_j trong biên dương lý tưởng. Đây là các biến hỗ trợ nới lỏng ràng buộc. Số hạng $a_{ij} = 1$ nếu giao tác thứ *i* có hỗ trợ cho tập X_j, ngược lại $a_{ij} = 0$.

Trong công thức trên, hàm mục tiêu (5) gồm 2 phần: phần thứ nhất đòi hỏi số lượng giao tác cần hiệu chỉnh phải nhỏ nhất, phần thứ hai đặt một số biến nới lỏng s_j để đảm bảo công thức ILP có nghiệm. Các ràng buộc (6) chỉ ra rằng cần phải hiệu chỉnh ít nhất $n_j = sup_D(X_j) - \sigma + 1$ giao tác trong các giao tác hỗ trợ cho một tập phổ biến nhạy cảm X_j để tất cả mọi tập nhạy cảm trong S có thể ẩn. Các ràng buộc (7) chỉ ra rằng các tập hạng mục trong $Bd^+(\hat{F})$ nên được giữ lại. Tuy nhiên, nếu điều đó làm cho công thức ILP trở nên vô nghiệm, bằng cách cho $s_j = 1$, số hạng $s_j.sup_D(X_j)$ sẽ vô hiệu hóa ràng buộc để công thức có nghiệm chấp nhận việc phát sinh một hiệu ứng phụ. Các ràng buộc (8) và (9) đảm bảo rằng các biến của bài toán chỉ nhận các giá trị 0 và 1 theo đúng như định nghĩa. Nghiệm của công thức ILP này là các x_i tương ứng với các giao tác nào cần hiệu chỉnh để ẩn các tập nhạy cảm.

Ví dụ với cơ sở dữ liệu mẫu ở Bảng 1. Công thức ILP thu được là

$$\begin{aligned}
&\min x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + (s_1 + s_2 + s_3 + s_4 + s_5 + s_6),\\
&s.t. \quad x_3 + x_4 + x_5 \geq 1\\
&\quad x_2 + x_3 + x_5 + x_6 \geq 2\\
&\quad x_2 + x_3 + x_4 \geq 1\\
&\quad x_2 + x_3 + x_4 - 3s_1 \leq 0\\
&\quad x_1 + x_2 + x_3 + x_5 + x_6 - 5s_2 \leq 2\\
&\quad x_3 + x_4 + x_5 - 3s_3 \leq 0\\
&\quad x_2 + x_3 + x_4 + x_5 + x_6 - 5s_4 \leq 2\\
&\quad x_2 + x_3 + x_4 - 3s_5 \leq 0\\
&\quad x_3 + x_4 + x_5 - 3s_6 \leq 0\\
&\quad x_i \in \{0,1\}, \forall i: 1..6;\ s_j \in \{0,1\}, \forall j: 1..6
\end{aligned}$$

Nghiệm thu được từ công thức ILP trên là $(x_1, \dots, x_6, s_1, \dots, s_6) = (0,0,1,0,0,1,1,0,1,0,1,1)$. Điều này cho thấy cần thực hiện hiệu chỉnh các giao tác số 3 và 6 để ẩn hoàn toàn các tập nhạy cảm. Thực hiện hiệu chỉnh cơ sở dữ liệu giao tác D thu được cơ sở dữ liệu hiệu chỉnh $D' = \{\{B,D\},\{B,C,D,E,F\},\{A,B,D,F\},\{A,B,C,E,F\},\{A,B,C,D\},\{B,D\}\}$. Biên dương tương ứng với cơ sở dữ liệu này là $Bd^+ = \{\{B,F\},\{B,D\},\{A,B\},\{B,C\}\}$. So sánh với biên dương lý tưởng thì độ mất mát dữ liệu là 33%.

3. KẾT QUẢ THỰC NGHIỆM

Các thực nghiệm đều được tiến hành trong cùng một nền tảng, code bằng ngôn ngữ lập trình python, và thực hiện trên cùng một máy tính PC với CPU AMD Ryzen5 2600 Six-Core Processor 3.40 GHz và RAM 16GB chạy trên hệ điều hành Windows 10 (64-bit). Kết quả phương pháp được so sánh với công thức ILP thuần túy không sử dụng đến thông tin của biên dương được đề xuất bởi Menon [9]. Phần bên dưới gọi công thức của Menon là ILP1 và công thức cải tiến là ILP4.

Như đã trình bày ở mục 1, công thức ILP của Menon [9] là công thức đầu tiên và công thức đang có kết quả ổn định nhất trong nhóm các phương pháp chính xác. Các cố gắng cải tiến công thức của Menon đều chưa cho kết quả tốt hơn về việc ẩn tập phổ biến hoặc chưa ổn định (có nhiều trường hợp công thức không có lời giải). Ngoài ra, việc so sánh một phương pháp thuộc nhóm phương pháp chính xác với phương pháp thuộc nhóm heuristic hay phương pháp tiến hóa là không có ý nghĩa. Thời gian tính toán phương pháp chính xác luôn cao hơn với độ chính xác trong việc ẩn tập phổ biến luôn tốt hơn các nhóm còn lại. Do đó, bài báo này

chọn so sánh với phương pháp được đề xuất bởi Menon [9] và được cải tiến bởi Kagklis [7] với toán tử *Min(S)* dù kết quả được công bố khá lâu.

3.1. Các độ đo đánh giá hiệu quả phương pháp ẩn tập phổ biến

- *Độ mất mát thông tin dựa trên biên dương*: Đây là độ đo chính để đánh giá hiệu quả phương pháp ẩn tập phổ biến theo hướng tiếp cận chính xác. Trong hướng tiếp cận này nếu hệ các ràng buộc thỏa mãn và có nghiệm thì chắc chắn phương pháp sẽ ẩn được tất cả các tập phổ biến nhạy cảm. Chính vì vậy, hiệu quả phương pháp được đánh giá bằng các hiệu ứng phụ. Độ mất mát thông tin dựa trên biên dương được xác định bởi tỷ lệ số lượng tập phổ biến trong biên dương lý tưởng bị mất đi do quá trình hiệu chỉnh dữ liệu và số lượng tập phổ biến trong biên dương lý tưởng ban đầu

$$IL_{Bd}\left(Bd^{+}(\hat{F}), Bd^{+}(F')\right) = \frac{|X \in Bd^{+}(\hat{F}) : X \notin Bd^{+}(F')|}{|Bd^{+}(\hat{F})|} \quad (10)$$

Tỷ số này càng nhỏ, độ mất mát thông tin của biên dương lý tưởng càng ít, phương pháp càng hiệu quả.

- *Độ mất mát thông tin dựa trên tập phổ biến*: là tỷ số giữa tổng độ lệch tuyệt đối các hỗ trợ của các tập phổ biến lý tưởng trong cơ sở dữ liệu ban đầu và cơ sỡ dữ liệu hiệu chỉnh

$$IL(D, D') = \frac{\sum_{X \in \hat{F}} |sup_D(X) - sup_{D'}(X')|}{\sum_{X \in \hat{F}} sup_D(X)} \quad (11)$$

Tỷ số này càng nhỏ, độ mất mát thông tin của tập phổ biến lý tưởng càng ít, phương pháp càng hiệu quả.

- *Hiệu ứng phụ*: được xác định bằng sự khác biệt giữa số lượng tập phổ biến trong cơ sở dữ liệu lý tưởng và số lượng tập phổ biến trong cơ sở dữ liệu đã hiệu chỉnh.

$$SE(\hat{F}, F) = \frac{|\hat{F}| - |F|}{|\hat{F}|} \quad (11)$$

Tỷ số này càng nhỏ cho thấy tập hiệu chỉnh càng gần với tập lý tưởng, phương pháp càng hiệu quả.

3.2. Mô tả các tập dữ liệu sử dụng trong thực nghiệm

Bảng 2. Đặc tính các tập dữ liệu

Tập dữ liệu	Số giao tác	Số hạng mục	Chiều dài trung bình	Ngưỡng hỗ trợ
chess	3196	75	37	2557
mushroom	8124	119	23	1625
BMS1	59601	497	2.5	51
BMS2	77512	3340	5.6	39

Tập dữ liệu Chess được sinh ra và mô tả bởi Alen Shapiro và được công bố công khai thông qua UCI (Machine Learning Repository) [12]. Định dạng cho các thể hiện trong tập dữ liệu là một chuỗi gồm 37 giá trị thuộc tính.

Tập dữ liệu Mushroom [12] được rút ra từ "Hướng dẫn lĩnh vực xã hội Audubon" về các loại nấm Bắc Mỹ của tác giả G. H. Lincoff, NewYork: Alfred A. Knopf, và được công bố công khai thông qua UCI (Machine Learning Repository) [12]. Đây là tập dữ liệu của các mẫu giả định mô tả về 23 loài "nấm lá tia" trong họ Agricus và Lepiota.

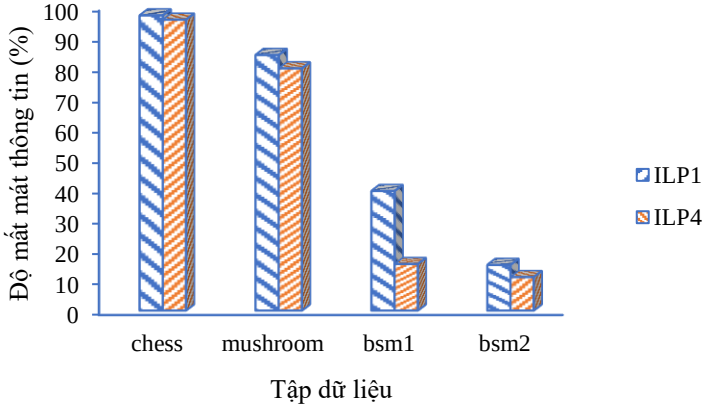
Hai tập Bms-1 và Bms-2 là hai tập dữ liệu được sử dụng cho KDD Cup 2000 [13] và chứa những dòng dữ liệu click và mua từ trang web của nhà bán lẻ legwear và legcare.

3.3. Đánh giá thực nghiệm

** So sánh độ mất mát thông tin dựa trên biên dương*

Bảng 3. So sánh độ mất mát thông tin dựa trên biên dương

Tập dữ liệu	Độ hỗ trợ	Số tập nhạy cảm (tỷ lệ so với tập phổ biến)	Độ mất mát thông tin dựa trên biên dương (%)	
			ILP1	ILP4
chess	2557	10 (0,12%)	93	92
		20 (0,25%)	100	97
		50 (0,6%)	98	98
mushroom	1625	10 (0,02%)	74	66
		20 (0,04%)	87	85
		50 (0,1%)	91	88
bms1	51	10 (0,17%)	19	4
		20 (0,34%)	38	13
		50 (0,85%)	61	29
bms2	39	10 (0,01%)	9	8
		20 (0,02%)	17	13
		50 (0,05%)	19	12



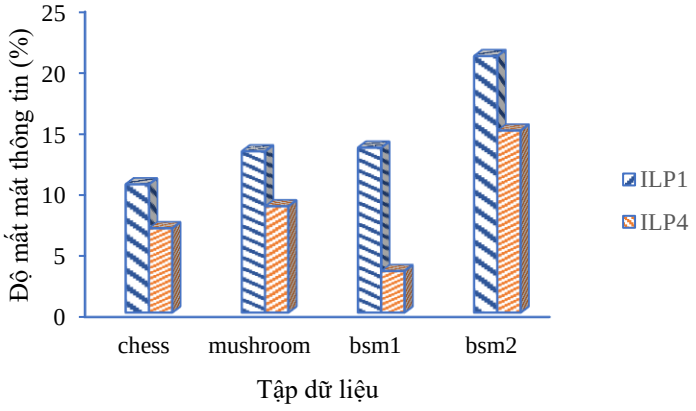
Hình 1. Biểu đồ thể hiện giá trị trung bình của độ mất mát thông tin dựa trên biên dương của một số tập dữ liệu

Hình 1 và Bảng 3 cho ta thấy trong hầu hết các trường hợp, công thức ILP4 đều cho kết quả tốt hơn công thức ILP1 ngoại trừ tập dữ liệu chess với số lượng tập phổ biến nhạy cảm lớn thì sự chênh lệch không đáng kể.

***** *So sánh độ mất mát thông tin dựa trên tập phổ biến*

Bảng 4. So sánh độ mất mát thông tin dựa trên biên dương

Tập dữ liệu	Độ hỗ trợ	Số tập nhạy cảm (tỷ lệ so với tập phổ biến)	Độ mất mát thông tin dựa trên tập phổ biến (%)	
			ILP1	ILP4
chess	2557	10 (0,12%)	5,8	4,3
		20 (0,25%)	9,7	7,5
		50 (0,6%)	15,9	8,9
mushroom	1625	10 (0,02%)	9,6	6,2
		20 (0,04%)	10,2	7,3
		50 (0,1%)	19,7	12,7
bms1	51	10 (0,17%)	3,4	1,0
		20 (0,34%)	11,8	2,4
		50 (0,85%)	25,3	6,7
bms2	39	10 (0,01%)	12,3	12,1
		20 (0,02%)	16,2	15,7
		50 (0,05%)	34,6	17,0



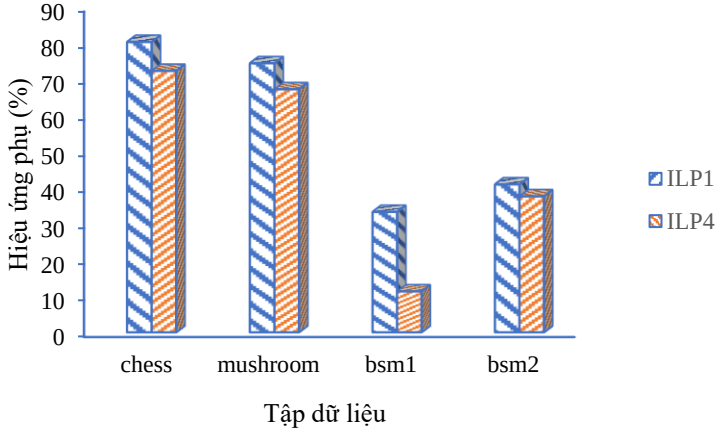
Hình 2. Biểu đồ thể hiện giá trị trung bình của độ mất mát thông tin dựa trên các tập phổ biến của một số tập dữ liệu

Bảng 5 và Hình 2 cho thấy đối với độ đo độ mất mát thông tin dựa trên các tập phổ biến, công thức ILP4 cho kết quả tốt hơn hẳn công thức ILP1. Giữa hai công thức có độ chênh lệnh khá lớn.

** So sánh hiệu ứng phụ*

Bảng 5. So sánh hiệu ứng phụ

Tập dữ liệu	Độ hỗ trợ	Số tập nhạy cảm (tỷ lệ so với tập phổ biến)	Hiệu ứng phụ (%)	
			ILP1	ILP4
chess	2557	10 (0,12%)	61,6	52,1
		20 (0,25%)	83,9	78,2
		50 (0,6%)	95,4	86,6
mushroom	1625	10 (0,02%)	68,7	62,1
		20 (0,04%)	69,2	62,5
		50 (0,1%)	85,4	77,0
bms1	51	10 (0,17%)	14,4	2,7
		20 (0,34%)	33,0	9,5
		50 (0,85%)	52,6	21,7
bms2	39	10 (0,01%)	33,6	33,5
		20 (0,02%)	39,1	37,9
		50 (0,05%)	50,1	41,3



Hình 3. Biểu đồ thể hiện giá trị trung bình của hiệu ứng phụ của một số tập dữ liệu

Bảng 5 và Hình 3 cho thấy với độ đo hiệu ứng phụ, ILP4 cũng tỏ ra nhỉnh hơn so với ILP1 đặc biệt với tập bms1. Tuy nhiên với bms2 thì hiệu ứng phụ của hai công thức có chênh lệch ít không đáng kể.

Các thực nghiệm cho thấy một cải tiến rõ rệt trong các kỹ thuật sử dụng các phương trình ràng buộc vào bài toán ẩn tập phổ biến. So với phương pháp ILP truyền thống, phương pháp đề xuất trong công trình này đưa thêm thông tin biên dương lý tưởng vào các hệ phương trình ràng buộc. Việc thông tin về biên dương xuất hiện yêu cầu nghiệm của hệ ILP mới phải giữ lại được các tập trong biên dương lý tưởng khi ẩn tập phổ biến. Điều này dẫn đến các hiệu ứng phụ của phương pháp giảm rõ rệt. Ngoài ra, công thức ILP4 còn đưa vào các hệ số để nới lỏng ràng buộc khi cần. Một số công trình cố gắng đưa vào quá nhiều ràng buộc gây ảnh hưởng đến tính ổn định bài toán. Công thức ILP4 do có các hệ số giúp nới lỏng ràng buộc nên sẽ có

nghiệm trong mọi trường hợp. Trường hợp xấu nhất là tất cả các phương trình ràng buộc liên quan đến biên dương lý tưởng đều bị nới lỏng, công thức ILP4 sẽ quay đúng như công thức ILP1 của Menon [9] với cải tiến Min(S) của Kagklis [7].

4. KẾT LUẬN

Công trình này đã cải tiến một công thức ẩn tập phổ biến nhạy cảm trong một cơ sở dữ liệu giao tác. Bài toán ẩn tập phổ biến nhạy cảm được mô hình hóa bằng công thức ILP1 của Menon nhưng chưa xét đến các ràng buộc liên quan đến biên dương lý tưởng. Công thức ILP cải tiến trong công trình này ngoài việc thêm vào các ràng buộc cho biên dương còn thêm vào các biến để nới lỏng các ràng buộc trong trường hợp bài toán vô nghiệm. Thực nghiệm cũng cho thấy công thức ILP cải tiến mặc dù đã nới lỏng các ràng buộc nhưng vẫn cho các kết quả khả quan hơn công thức ILP ban đầu thông qua tính toán các độ đo như độ mất mát thông tin và hiệu ứng phụ. Phương pháp này cũng có hạn chế cơ bản là thời gian tính toán khá lâu do việc phình không gian nghiệm sử dụng cho các biến nới lỏng. Do đó, hướng phát triển trong thời gian tới của phương pháp là tìm cách biểu diễn các phương trình ràng buộc tốt hơn giúp rút gọn thời gian tính toán.

Lời cảm ơn: Nghiên cứu này do trường Đại học Công nghiệp Thực phẩm TP.Hồ Chí Minh bảo trợ và cấp kinh phí theo Hợp đồng số 05/HĐ-DCT ký ngày 5/01/2021.

TÀI LIỆU THAM KHẢO

1. Inda K., Amra K., Lejla T. - Data mining privacy preserving: Research agenda, WIREs Data Mining Knowl Discov **11** (2021) e1392.
2. Atallah M., Bertino E., Elmagarmid A., Ibrahim M., Verykios V. - Disclosure limitation of sensitive rules, Proceedings of the knowledge and data engineering exchange **99** (1999) 45-52.
3. Moustakides G., Verykios V. - A maxmin approach for hiding frequent itemsets, Data Knowl Eng **65** (1) (2008) 75-89.
4. Sun X., Yu P. - A border–based approach for hiding sensitive frequent itemsets, Proceedings of 5th IEEE internationa conference on data mining (2005) 426-433.
5. Sun X., Yu P. - Hiding sensitive frequent itemsets by a border-based approach, J. Comput. Sci. Eng. **1** (1) (2007) 74-94.
6. Usman A., Jerry C., Wei L., Gautam S., and Youcef D. - A Deep Q-Learning Sanitization Approach for Privacy Preserving Data Mining, Proceedings of the 2021 International Conference on Distributed Computing and Networking **21** (2021) 43-48.
7. Kagklis V., Verykios V., Tzimas G., Tsakalidis A. - An integer linear programming scheme to sanitize sensitive frequent itemsets, Proceedings of 2014 IEEE international Conference on Tools with AI **14** (2014) 771-775.
8. Gkoulalas-Divanis A., Verykios V. - Hiding sensitive knowledge without side effects, Knowl Info Syst **20** (3) (2009) 263-299.
9. Menon S., Sarkar S., Mukherjee S. - Maximizing accuracy of shared databases when concealing sensitive patterns, Info Syst Res **16** (3) (2005) 256-270.
10. Bayardo R. - Efficiently mining long patterns from databases, Proceedings of the 1998 ACM SIGMOD international conference on management of data (1998) 85-93.

11. Jimmy M., Tai W., Gautam S., Unil Y., Shahab T., Jerry C., Wei L. - An evolutionary computation-based privacy-preserving data mining model under a multithreshold constraint, Trans Emerging Tel Tech. **32** (2021) e4209.
12. Dua D., Graff C. – UCI Machine Learning Repository, University of California, School of Information and Computer Science (2019). http://archive.ics.uci.edu/ml.
13. Kohavi R., Brodley C., Frasca B., Mason L., Zheng Z. - KDD-Cup 2000 organizers' report: peeling the onion, SIGKDD explorations (2000)

    http://www.ecn.purdue.edu/KDDCUP.
14. Stavropoulos C., Vassilios S., Kagklis V. - A transversal hypergraph approach for the frequent itemset hiding problem, Knowledge and Information Systems **47** (2016) 625- 645.

ABSTRACT

HIDING FREQUENT ITEMSETS BASED ON INTEGER LINEAR PROGRAMMING METHOD COMBINED WITH IDEAL POSITIVE BORDER

Nguyen Thi Thu Tam, Dinh Nguyen Trong Nghia*
Ho Chi Minh City University of Food Industry
*Email: *nghiadnt@hufi.edu.vn*

This study proposes a method to hide sensitive frequent itemsets in transaction databases. This proposed method is based on using information from the ideal positive border to build integer linear programming equations. The solution of this equation determines the transactions that need to be sanitized to completely hide the sensitive frequent itemsets. In case the equation has no solution, some coefficients are added to loosen the constraints of the problem. Experimental evaluation of this method on some well-known data sets shows that this method has higher accuracy than the method using traditional integer linear programming.

Keywords: Privacy-preserving aata mining, hiding frequent itemsets, integer linear programming, ideal positive border.