Kỷ yếu Hội nghị Quốc gia lần thứ X về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR), Đà Nẵng, ngày 17-18/08/2017
DOI: 10.15625/vap.2017.00034

HHUAL: THUẬT TOÁN ẨN LUẬT KẾT HỢP HỮU ÍCH CAO NHẠY CẢM VÀ MỘT ĐỀ XUẤT CẢI TIẾN DỰA TRÊN GIÀN GIAO

Huỳnh Triệu Vỹ[1], Lê Quốc Hải[2], Trương Ngọc Châu[3]

[1]Trường ĐH Phạm Văn Đồng;

[2]Trường CĐSP Quảng Trị

[3]Trường ĐH Bách khoa Đà Nẵng

htrvy@yahoo.com, hailq79@gmail.com, truongngocchau@yahoo.com

TÓM TẮT: *Ẩn luật kết hợp hữu ích cao nhạy cảm là một bài toán quan trọng nhằm bảo vệ tính riêng tư của tri thức được khai phá từ dữ liệu khi chúng được chia sẻ ra bên ngoài. Mục tiêu của bài toán này là tác động lên dữ liệu để tất cả các luật kết hợp hữu ích cao nhạy cảm không thể khai thác được từ dữ liệu khi chúng được chia sẻ sao cho hiệu ứng phụ sinh ra là thấp nhất. Bài toán ẩn luật kết hợp hữu ích cao không giống như ẩn luật kết hợp phổ biến, bởi vì khai phá luật kết hợp phổ biến dựa trên ràng buộc về độ hỗ trợ và độ tin cậy, còn đối với khai phá luật kết hợp hữu ích cao dựa trên ràng buộc về hữu ích và độ tin cậy hữu ích. Trong bài báo này chúng tôi đề xuất thuật toán ẩn luật kết hợp hữu ích cao có tên là HHUAL, dựa trên thuật toán này chúng tôi cũng đề xuất một hướng tiếp cận mới cho việc lựa chọn mục ứng viên và giao tác để sửa đổi sao cho có thể ẩn luật nhưng gây ra hiệu ứng phụ là thấp nhất dựa trên Giàn giao. Đóng góp mới của chúng tôi là: (1) Đây là lần đầu tiên bài toán ẩn luật kết hợp hữu ích cao nhạy cảm được giới thiệu; (2) Phương pháp xác định mục ứng viên và giao tác để sửa dựa trên Giàn giao của các tập mục có trọng số hữu ích cao.*

Từ khóa: *Tập mục hữu ích cao, Luật kết hợp hữu ích, Luật kết hợp hữu ích nhạy cảm, Giàn giao, Độ tin cậy hữu ích.*

I. GIỚI THIỆU

Khai phá luật kết hợp từ cơ sở dữ liệu (CSDL) là một bài toán quan trọng trong khai phá dữ liệu. Mục đích của khai phá luật kết hợp là tìm ra những luật có dạng $X \rightarrow Y$ từ tập các tập mục thường xuyên XY. Hai đặc trưng quan trọng của một luật là độ hỗ trợ và độ tin cậy. Những luật phổ biến là những luật có độ hỗ trợ và độ tin cậy lớn hơn ngưỡng hỗ trợ tối thiểu và độ tin cậy tối thiểu [1]. Ý nghĩa của luật kết hợp là xác định được mối quan hệ giữa các tập mục dựa trên sự xuất hiện hoặc số lượng của chúng trong CSDL. Hạn chế của luật kết hợp thông thường là không xác định được mối quan hệ giữa các tập mục khi gắn với độ hữu ích của chúng. Để khắc phục hạn chế này, năm 2015, Jayakrushna Sahoo, Ashok Kumar Das, A. Goswami [10] đã đề xuất thuật toán để khai phá luật kết hợp hữu ích cao nhằm khai thác mối quan hệ giữa các tập mục trong CSDL dựa trên độ hữu ích của chúng. Hai đặc trưng của luật kết hợp hữu ích cao là giá trị hữu ích và độ tin cậy hữu ích.

Bài toán khai phá luật kết hợp hữu ích cao được thực hiện qua 2 bước chính: (1) tìm tất cả các tập mục hữu ích cao (các tập mục có giá trị hữu ích lớn hơn ngưỡng hữu ích tối thiểu) [2, 6, 9], (2) hai là từ tập mục hữu ích cao tìm các luật có độ tin cậy hữu ích lớn hơn độ tin cậy hữu ích tối thiểu. Dựa trên nền tảng này, năm 2017 Thang Mai, Bay Vo, Loan T. T. Nguyen [15] đã đưa ra thuật toán khai phá luật kết hợp hữu ích cao dựa trên Giàn và cũng đã chứng minh được thuật toán này hiệu quả hơn thuật toán trong [10].

Hiện nay, hợp tác thương mại toàn cầu và kinh doanh xuyên quốc gia là xu thế chung của thế giới. Các doanh nghiệp muốn phát triển thì không thể hoạt động độc lập mà có sự liên kết, hợp tác chặt chẽ với các đối tác khác. Chia sẻ dữ liệu là một yêu cầu quan trọng để thúc đẩy quan hệ hợp tác giữa các đối tác vì mục đích đôi bên cùng có lợi. Tuy nhiên, trong CSDL được chia sẻ có thể ẩn chứa các thông tin mà nếu các đối tác khai thác được các thông tin này sẽ gây bất lợi cho bên chia sẻ, các thông tin này được gọi là thông tin nhạy cảm. Những luật kết hợp được sử dụng để suy luận ra các thông tin nhạy cảm được gọi là luật kết hợp nhạy cảm. Để đảm bảo các đối tác sẽ không khai thác được luật kết hợp nhạy cảm thì dữ liệu phải được sửa đổi trước khi chia sẻ để các luật này sẽ không được khai thác bởi đối tác. Việc sửa đổi dữ liệu phải đảm bảo rằng các luật kết hợp nhạy cảm sẽ được ẩn nhưng gây ra hiệu ứng phụ trên dữ liệu sửa đổi là thấp nhất. Để giải quyết điều này, năm 1999 M. Atallah [3] lần đầu tiên đề xuất mô hình ẩn luật kết hợp nhạy cảm để bảo vệ thông tin riêng tư trong khai phá dữ liệu. Từ nền tảng này nhiều thuật toán ẩn tập mục nhạy cảm hiệu quả hơn cũng được đề xuất [5, 7, 8].

Các kỹ thuật được áp dụng trong ẩn luật kết hợp phổ biến không thể áp dụng trực tiếp vào trong ẩn tập mục hữu ích cao nhạy cảm hay ẩn luật kết hợp hữu ích cao nhạy cảm vì tiêu chí đánh giá là khác nhau. Năm 2010, Jieh-Shan Yeh, Po-Chiang Hsu [12] đã đề xuất hai thuật toán HHUIF và MSICF ẩn tập mục hữu ích cao nhạy cảm. Từ hai thuật toán tiền đề này có nhiều tác giả đã đưa ra các thuật toán ẩn tập mục hữu ích cao nhạy cảm hiệu quả hơn [4, 14]. Vấn đề ẩn luật kết hợp hữu ích cao nhạy cảm là bài toán cần được nghiên cứu và giải quyết nhưng cho đến nay chưa có kết quả được công bố. Việc ẩn luật kết hợp hữu ích cao nhạy cảm có thể được thực hiện nhờ vào việc ẩn tập mục hữu ích cao nhạy cảm sinh ra nó. Tuy nhiên, cách làm này kéo theo tất cả các luật được sinh ra từ tập mục đều bị ẩn và do đó làm tăng hiệu ứng phụ. Trong bài báo này, chúng tôi lần đầu tiên đề xuất phương pháp ẩn luật kết hợp hữu ích cao nhạy cảm, thuật toán được chúng tôi đề xuất có tên HHUAL (Hiding High Utility Association Rules) ẩn luật kết hợp hữu ích cao nhạy cảm dựa trên việc làm giảm độ tin cậy hữu ích của luật và một đề xuất cải tiến của nó dựa trên Giàn

giao các tập mục có trọng số hữu ích cao. Nội dung tiếp theo của bài báo được tổ chức như sau: Phần II trình bày các vấn đề liên quan đến bài toán ẩn luật, phần III trình bày thuật toán HHUAL và kết quả thực nghiệm, Phần IV thảo luận và đề xuất cải tiến thuật toán HHUAL, phần V kết luận.

II. CÁC VẤN ĐỀ LIÊN QUAN

Định nghĩa 1 (Cơ sở dữ liệu giao tác và giá trị hữu ích của tập mục): Cho $I=\{i_1, i_2,..., i_m\}$ là một tập các mục. $D=\{T_1,T_2,...,T_m\}$ là cơ sở dữ liệu giao tác, ở đây, mỗi $T_c \in D$ là tập con của I. Mỗi mục $i \in T_c$ có một giá trị dương, ký hiệu là $q(i,Tc)$ được gọi là giá trị hữu ích nội của i (tương ứng với số lượng của i trong mỗi T_c). Mỗi mục $i \in I$ có một giá trị hữu ích ngoại, ký hiệu là $p(i)$ (tương ứng với giá trị hữu ích của mục i).

Hữu ích của mục $i \in T_c$, được định nghĩa là $u(i,T_c)=q(i,T_c)\times p(i)$. Hữu ích của tập mục X trong giao tác T_c, được định nghĩa $u(X,T_c)=\sum_{i\in X\wedge X\subseteq T_c} u(i,T_c)$. Hữu ích của tập mục X trong cơ sở dữ liệu D, được định nghĩa $u(X)=\sum_{T_c\in D\wedge X\subseteq T_c} u(X,T_c)$.

Bảng 1. Cơ sở dữ liệu giao tác D

Tid	Transaction	Tid	Transaction	Tid	Transaction
T1	A(4), C(1), E(6), F(2)	**T4**	D(1), E(2), F(6)	**T7**	D(1), E(1), F(4), G(1), H(1)
T2	D(1), E(4), F(5)	**T5**	A(3), C(1), E(1)	**T8**	D(7), E(3)
T3	B(4), D(1), E(5), F(1)	**T6**	B(1), F(2), H(1)	**T9**	G(10)

Bảng 2. Bảng giá trị hữu ích

Item	A	B	C	D	E	F	G	H
Utility	3	4	5	2	1	1	1	2

Định nghĩa 2 (Giá trị hữu ích giao tác và trọng số hữu ích giao tác): Giá trị hữu ích giao tác của một giao tác T_c là tổng giá trị hữu ích của các mục trong giao tác T_c, được định nghĩa là: $TU(T_c)=\sum_{x\in T_c} u(x,T_c)$.

Trọng số hữu ích giao tác của một tập mục X là tổng giá trị hữu ích của các giao tác chứa tập X, được định nghĩa là: $TWU(X)=\sum_{T_c\in g(X)} TU(T_c)$.

Bảng 3. Tập $HTWU_D$ được rút trích từ CSDL cho ở bảng 1 & 2 với min_util=30

Item	TWU	Item	TWU	Item	TWU	Item	TWU
A	40	E	112	CE	40	BF	32
B	32	F	88	DE	72	ACE	40
C	40	AC	40	DF	55	DEF	55
D	72	AE	40	EF	80		

Tính chất 1: Trọng số hữu ích của một tập mục X luôn luôn lớn hơn hoặc bằng giá trị hữu ích của chính nó, tức là: $TWU(X)\geq u(X)$.

Tính chất 2: Cho X là một tập mục, nếu $TWU(X)< min\_util$ thì tập mục X và tất cả tập cha của X không phải là tập mục hữu ích cao.

Tính chất 3: Cho $X\subseteq Y\forall X,Y\in I$ thì $TWU(X)\geq TWU(Y)$, như vậy, trọng số hữu ích giao tác của tập mục thỏa mãn tính chất phản đơn điệu.

Bảng 4. Tập H_D được rút trích từ CSDL cho ở bảng 1 & 2 với min_util=30

Item	Utility	Item	Utility	Item	Utility	Item	Utility
AC	31	EF	36	ACE	38	DEF	36
DE	37						

Định nghĩa 3(Tập mục có trọng số hữu ích cao): Cho *min_util* là ngưỡng hữu ích tối thiểu. Tập mục X được gọi là tập mục có trọng số hữu ích cao nếu $TWU(X)\geq \min\_util$. Cho $HTWU_D$ là một tập của các tập mục có trọng số hữu ích cao thì $HTWU_D=\{X \mid X\subseteq I, TWU(X)\geq \min\_util\}$.

Định nghĩa 4 (Tập mục hữu ích cao): Cho *min_util* là ngưỡng hữu ích tối thiểu. Tập mục *X* được gọi là tập mục hữu ích cao nếu $u(X) \geq \min\_util$, ngược lại *X* là tập mục hữu ích thấp. Cho H_D là một tập của các tập mục hữu ích cao thì $H_D = \{X \mid X \subseteq I, u(X) \geq \min\_util\}$.

Tính chất 4: Tập của tập mục có trọng số hữu ích cao là tập ứng viên để tìm tập mục hữu ích cao, tức $H_D \subseteq HTWU_D$.

Định nghĩa 5 (Giá trị hữu ích cục bộ của một mục trong tập mục trên một giao tác): Giá trị hữu ích cục bộ của một mục x_i trong tập mục *X* trên giao tác T_c là giá trị hữu ích của mục x_i trong các giao tác có chức *X*, được định nghĩa là: $luv(x_i, X, T_c) = u(x_i, T_c) \mid x_i \in X \subseteq T_c \wedge T_c \in D$.

Định nghĩa 6 (Giá trị hữu ích cục bộ của một tập mục trong tập cha của nó trên một giao tác): Giá trị hữu ích cục bộ của một tập mục *X* trong tập *Y* trên giao tác T_c, sao cho $X \subseteq Y \wedge Y \subseteq T_c$ là tổng các giá trị hữu ích cục bộ của mỗi mục $x_i \subseteq X$ trong giao T_c, được định nghĩa là: $luv(X, Y, T_c) = \sum_{x_i \in X \wedge X \subseteq Y \wedge Y \subseteq T_c} luv(x_i, X, T_c)$.

Định nghĩa 7 (Giá trị hữu ích cục bộ của một tập mục trong tập cha của nó trên CSDL giao tác): Giá trị hữu ích cục bộ của một tập mục *X* trong tập *Y* trên CSDL giao tác *D*, sao cho $X \subseteq Y \wedge Y \subseteq T_c \wedge T_c \in D$ là tổng giá trị hữu ích cục bộ của mỗi mục $x_i \subseteq X$ trong tập mục *Y* trong các giao tác chứa *Y*, được định nghĩa là: $luv(X, Y) = \sum_{T_c \in D \wedge X \subseteq Y \wedge Y \subseteq T_c} \sum_{x_i \in X} luv(x_i, X, T_c)$

Định nghĩa 8 (Luật kết hợp hữu ích cao) [13]: Một luật kết hợp $R: X \rightarrow Y$, ở đây $X, Y \subseteq I \wedge Y, X \neq \phi \wedge X \cap Y = \phi$ được gọi là một luật kết hợp hữu ích cao nếu như độ tin cậy hữu ích của luật *R* không nhỏ hơn độ tin cậy hữu ích tối thiểu (*min_uconf*) do người dùng đưa ra. Độ tin cậy hữu ích tối thiểu của luật *R* được định nghĩa là: $uconf(R) = \frac{luv(X, XY)}{u(X)}$.

Định nghĩa 9 (Khai phá luật kết hợp hữu ích cao) [13]: Khai phá luật kết hợp hữu ích cao là một quá trình đi tìm tất cả các luật kết hợp được suy diễn từ các tập mục hữu ích cao thỏa mãn độ tin cậy hữu ích của luật không nhỏ hơn độ tin cậy hữu ích tối thiểu.

Cho HR_D là một tập chứa các luật kết hợp hữu ích cao thì: $HR_D = \{R: X \rightarrow Y \mid uconf(R) \geq \min\_uconf\}$.

Bảng 5. Tập HR_D được khai thác từ tập H_D trong bảng 4 với *min_uconf* = 80%

Rules	Uconf(%)	Rules	Uconf(%)	Rules	Uconf(%)	Rules	Uconf(%)
A→C	100	F→E	90	A→CE	100	AE→C	100
C→A	100	E→F	81	F→DE	80	DF→E	100
D→E	100	AC→E	100	C→AE	100	CE→A	100

III. THUẬT TOÁN ẨN LUẬT KẾT HỢP HỮU ÍCH CAO NHẠY CẢM HHUAL

1. Ẩn luật kết hợp hữu ích cao nhạy cảm

Định nghĩa 10 (Tập luật kết hợp hữu ích cao nhạy cảm): Cho $R_S \subseteq HR_D$ là một luật kết hợp hữu ích cao nhạy cảm thì tập luật kết hợp hữu ích cao nhạy cảm được định nghĩa: $HRS_D = \{R_S \mid R_S \subseteq HR_D\}$. Như vậy, $HRS_D \subseteq HR_D$.

Định nghĩa 11 (Ẩn luật kết hợp hữu ích cao nhạy cảm): Ẩn luật kết hợp hữu ích cao nhạy cảm là quá trình chuyển đổi cơ sở dữ liệu giao tác *D* thành *D'* mà đảm bảo rằng các luật *R* vẫn được khai thác từ *D'* ngoài trừ R_S, tức $HR_{D'} = HR_D \setminus HRS_D$.

Định nghĩa 12 (Bảng chỉ mục nhạy cảm): Để tăng tốc độ tính toán độ tin cậy hữu ích của luật nhạy cảm $R_S: X \rightarrow Y$ và xác định giao tác mục tiêu trong quá trình ẩn luật R_S, cần xây dựng một bảng chỉ mục chứa các ID của các giao tác chứa XY của các luật $R_S: X \rightarrow Y \in HRS_D$. Bảng chỉ mục này được định nghĩa như sau: $iTable \leftarrow \{T_c \mid XY \subseteq T_c \wedge T_c \in D\}$.

Ví dụ: Cho $HRS_D = \{A \to C, A \to CE, BE \to D\}$ thì *iTable* được xác định từ cơ sở dữ liệu ở bảng 1 như trong bảng 6.

Bảng 6. Bảng iTable

Itemset	Transaction IDs
AC	T1,T5
ACE	T1,T5
BED	T3

2. Mục tiêu của ẩn luật kết hợp hữu ích cao ngạy cảm: Để đảm bảo quá trình ẩn luật gây ra ít hiệu ứng phụ nhất, cụ thể hạn chế tối thiểu các vấn đề sau:

- Sinh luật ma: Luật không khai thác được từ *D* nhưng khai thác được từ *D'*.

- Mất luật: Ẩn đi các luật không phải là luật nhạy cảm.

- Ẩn sai luật: Không ẩn được luật nhạy cảm.

3. Phương pháp ẩn luật kết hợp hữu ích cao nhạy cảm: Để ẩn một luật nhạy cảm R_S:X→Y, có hai cách tiếp cận:

1) Giảm độ hữu ích của tập mục *XY* xuống dưới độ hữu ích tối thiểu để $XY \notin H_D$, để thực hiện điều này cần sửa số lượng của *B* hoặc *E* hoặc cả hai trong các giao tác chứa *XY*, ví dụ ẩn R_S:*B→E*, nếu sửa *B* trong giao tác T_3 từ 4 xuống 3, khi đó *u(BE)=17<min_util* nên *B→E* không thể được sinh ra, tuy nhiên các tập mục *B, BE, BF, BDE,BFE* cũng bị ẩn theo và tất nhiên các luật được sinh ra từ các tập này cũng không được sinh ra. Như vậy, hiệu ứng phụ do phương án này gây ra là rất lớn.

2) Giảm độ tin cậy hữu ích của luật R_S xuống dưới độ tin cậy hữu ích tối thiểu, tức $uconf(R_S : X \to Y) = \dfrac{luv(X, XY)}{u(X)} < \min\_uconf$. Để thực hiện điều này có hai phương án:

+ Phương án 1: Tăng mẫu số (tức tăng *u(X)*) lên bằng cách chèn thêm một giao tác chỉ chứa mục *X* với số lượng của mỗi mục trong *X* sao cho $\dfrac{luv(X, XY)}{u(X)} < \min\_uconf$.

Nhận xét: Đối với phương án này có ưu điểm là ẩn được tất cả các luật nhạy cảm, nhưng có nhược điểm làm tăng kích thước của cơ sở dữ liệu và có thể sinh ra luật ma.

+ Phương án 2: Giảm tử số (tức giảm *luv(X,XY)*) sao cho $\dfrac{luv(X, XY)}{u(X)} < \min\_uconf$. Để giảm *luv(X,XY)* có hai cách:

- Một là sửa mục trong tập mục *X* ở các giao tác chứa *X* nhằm mục đích giảm *luv(X,XY)*.

Nhận xét: Khi *luv(X,XY)* giảm kéo theo *u(X)* cũng giảm làm cho thuật toán hội tụ chậm, ngoài ra trong trường hợp độ tin cậy hữu ích của luật là 100% sẽ không hội tụ.

- Hai là giảm số giao tác chứa *XY* tức xóa mục trong *Y* để cho *luv(X,XY)* giảm xuống nhưng *u(X)* không giảm theo.

Nhận xét: Ưu điểm của cách này là ẩn được tất cả các luật nhạy cảm, nhưng có thể ẩn nhầm nhiều luật không nhạy cảm nếu như việc chọn giao tác để xóa mục và mục cần xóa không phù hợp. Bởi vì, khi xóa mục $x_i \in Y \wedge XY \in T_{victim}$, ở đây T_{victim} là giao tác được chọn để xóa mục x_i, khi đó *luv(X,XY)* sẽ giảm xuống còn $luv(X, XY) - \sum_{x_i \in X \subseteq T_{victim} \wedge T_{victim} \in D} u(x_i, XY, T_{victim})$ và các tập mục $Z \supseteq XY \wedge Z \in T_{victim}$ cũng bị ảnh hưởng và giá trị hữu ích của chúng giảm xuống còn $u(Z) - u(Z, T_{victim})$. Như vậy, phương án tối ưu để lựu chọn giao tác mục tiêu và mục mục tiêu như sau:

- Chọn giao tác mục tiêu: Với luật nhạy cảm $R_S : X \to Y$, chọn giao tác có $luv(X, XY, T_c)$ nhỏ nhất trong số các giao tác thuộc tập giao tác chứa *XY* có $\dfrac{luv(X, XY, T_c)}{u(X)} > uconf(R_S : X \to Y) - \min\_uconf$, nếu không tồn tại giao tác này ta chọn giao tác có $luv(X, XY, T_c)$ lớn nhất trong các giao tác thuộc tập giao tác chứa *XY*.

- Chọn mục mục tiêu: Chọn mục có giá trị hữu ích nhỏ nhất trong tập mục Y của giao tác mục tiêu.

3. Thuật toán HHUAL

Mô tả thuật toán:

- Input: *D*: Cơ sở dữ liệu giáo tác gốc, HRS_D : Tập luật kết hợp hữu ích cao nhạy cảm, *min_uconf*: Độ tin cậy hữu ích tối thiểu.

- Output: *D'*: Cơ sở dữ liệu sau khi được sửa để ẩn luật kết hợp hữu ích cao nhạy cảm.

Bước 1: Từ cơ sở dữ liệu *D* và tập luật kết hợp hữu ích cao nhạy cảm, xây dựng bảng *iTable* (nêu trong định nghĩa 12) nhằm hạn chế truy xuất đến cơ sở dữ liệu *D* trong quá trình ẩn luật.

Bước 2: Duyệt qua từng $R_S \in HRS_D$, với mỗi R_S thực hiện các bước sau:

Bước 2.1: Tính độ sai khác *dr* giữa *uconf*(R_S) và *min_uconf* ($dr = uconf(R_S) - \min\_uconf$); Tạo ra một cơ sở dữ liệu phụ D_{R_S} gồm các giao tác chứa *XY* của luật $R_S : X \to Y \in HRS_D$ cần ẩn nhằm mục đích hạn chế truy xuất cơ sở dữ liệu *D*.

Bước 2.2: Nếu $dr < 0$ thực hiện bước 3, ngược lại thực hiên sửa cơ sở dữ liệu để ẩn luật, quá trình sửa CSDL này thực hiện qua các bước:

Bước 2.2.1: Xác định giao tác mục tiêu T_{victim}. Trên D_{R_S} tìm tập giao tác ứng viên $S_{T_{victim}}$ là các giao tác có $\frac{luv(X, XY, T_c)}{u(X)} > dr$. Nếu $S_{T_{victim}} \neq \phi$ thì $T_{victim} = \min\{luv(X, XY, T_c) \mid XY \subset T_c \wedge T_c \in S_{T_{victem}}\}$, ngược lại nếu $S_{T_{victim}} = \phi$, $T_{victim} = \max\{luv(X, XY, T_c) \mid XY \subset T_c \in D_{R_S}\}$.

Bước 2.2.2: Xác định mục mục tiêu từ giao tác mục tiêu đã tìm được ở bước 2.2.1. Tìm mục $x_i \in Y$ có $u(x_i, T_{victim})$ nhỏ nhất.

Bước 2.2.3: Xóa mục mục tiêu đã tìm được ở bước 2.2.2 ra khỏi giao tác mục tiêu; tính lại *dr*; cập nhật lại *iTable* và D_{R_S} .

Bước 3: Gỡ bỏ R_S ra khỏ HRS_D, quay lại bước 2.

Mục tiêu của việc lựa chọn giao tác mục tiêu là giảm tối đa giao tác bị sửa đổi khi ẩn luật và mục tiêu của lựa chọn mục mục tiêu trong giao tác mục tiêu để sửa là chọn mục có giá trị hữu ích thấp nhất để giảm tối đa tác động đến các tập mục chứa mục tiêu. Chính vì thế, trong bước 2.2.1 nếu tồn tại $S_{T_{victim}}$ thì số giao tác cần sửa để ẩn luật là một giao tác nên ta chọn giao tác có $luv(X, XY, T_c)$ nhỏ nhất, trường hợp không tồn tại $S_{T_{victim}}$ thì số giao tác cần sửa để ẩn luật là nhiều hơn một giao tác nên ta chọn giao tác có $luv(X, XY, T_c)$ lớn nhất để giảm số lượng giao tác bị tác động.

Chạy thuật toán với CSDL trong bảng 1, với luật cần ẩn là E→F: Tập giao tác mục tiêu $S_{T_{victim}} = \{T_1, T_2, T_3, T_4, T_7\}$, suy ra T_{victim}=T_7 và mục mục tiêu cần xóa là F. Kết quả luật E→F được ẩn và các luật bị ẩn nhầm là F→DE và DFE→E.

4. Kết quả thực nghiệm

- Mô tả các cơ sở dữ liệu và hệ thống chạy thực nghiệm: Cơ sở dữ liệu Foodmart [13] gồm 4141 giao tác, 1559 mục, giao tác dài nhất có 11 mục, số mục trung bình trong mỗi giao tác là 4 mục. Hệ thống máy tính có cấu hình: CPU Core I5 2.4GHz, RAM 4Gb, Windows 10.

Ngưỡng tối thiểu được được thiết lập: min_utility = 8000, min_uconf = 60%.

- Kết quả chạy thực nghiệm:

Tổng số tập mục hữu ích cao được sinh ra: 769. Tổng số luật kết hợp hữu ích cao được sinh ra: 414709.

Tập luật nhạy cảm được lựa chọn ngẫu nhiên từ tập các luật kết hợp hữu ích cao được sinh ra từ CSDL gồm 3 tập: tập có 3 luật, tập có 4 luật và tập có 5 luật. Kết quả ẩn luật như sau:

Bảng 7. Kết quả sau khi ẩn luật với HHUAL

Databases	Số luật cần ẩn	% luật ẩn sai	% luật ma	% luật ẩn nhầm
Foodmart	3	0%	0%	7,5%
	4	0%	0%	9,1%
	5	0%	0%	15,4%

IV. THẢO LUẬN

1. Hạn chế của thuật toán HHUAL

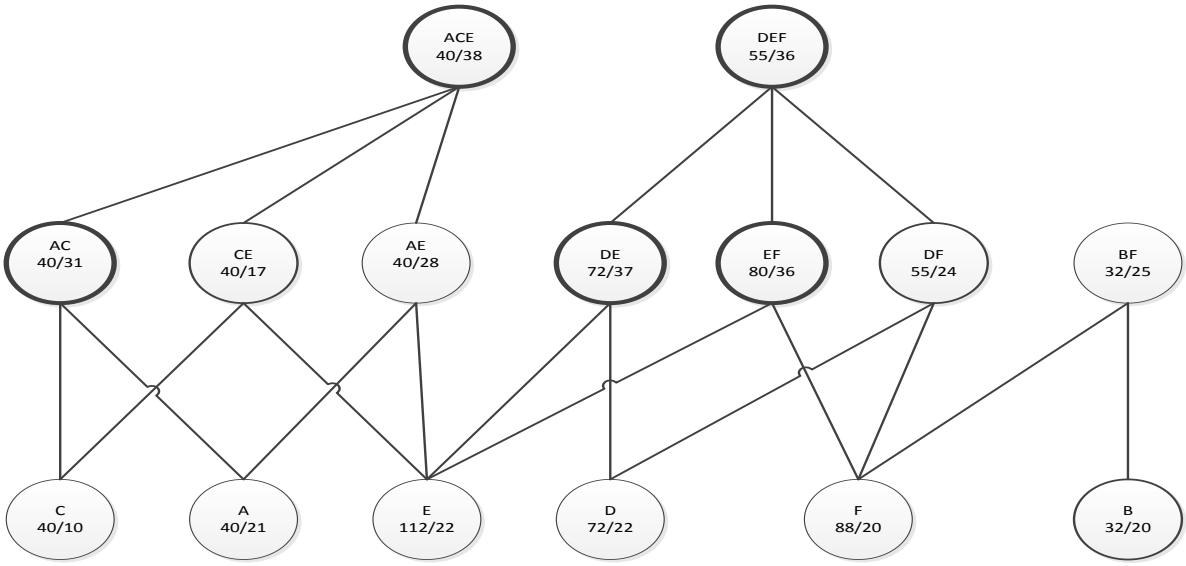
Kết quả thực nghiệm cho thấy tỉ lệ các hiệu ứng phụ gây ra gồm các luật bị ẩn nhầm là tương đối cao. Mặc dù phương pháp chọn giao tác mục tiêu và mục mục tiêu của HHUAL dựa theo tiêu chí hạn chế tối đa số giao tác và số mục bị can thiệp để sửa dữ liệu nhằm mục đích hạn chế hiệu ứng phụ của quá trình này gây ra. Tuy nhiên, phương pháp này không xét đến yếu tố tác động lên tất cả các tập trên toàn bộ CSDL, nên trong một số trường hợp số giao tác và số mục bị tác động ít lại gây ra hiệu ứng phụ nhiều hơn.

2. Đề xuất hướng tiếp cận ẩn luật kết hợp hữu ích cao nhạy cảm dựa trên Giàn giao

Dựa trên lý thuyết giàn [11], Hai Quoc Le [8] đã xây dựng được giàn giao của tập mục phổ biến và đưa bài toán ẩn luật kết hợp phổ biến về bài toán bảo vệ giàn giao để đảm bảo trong quá trình sửa đổi dữ liệu để ẩn luật, hạn chế tối thiểu gây ảnh hưởng đến các nút trên giàn giao làm thay đổi cấu trúc của giàn. Các nút dễ bị tác động mạnh làm thay đổi giàn đó chính là các nút có độ hỗ trợ thấp ở phía trên của giàn. Tác giả cũng đã chứng minh được các nút có độ hỗ trợ thấp nhất ở trên giàn chính là các nút thuộc tập sinh của tập mục phổ biến. Như vậy, để bảo vệ giàn cần bảo vệ các nút thuộc tập sinh của tập mục phổ biến. Dựa trên tính chất này tác giả đã đề xuất thuật toán hueristic có tên là *HCSRIL* để ẩn luật kết hợp phổ biến nhạy cảm. Kết quả thực nghiệm cho thấy *HCSRIL* làm việc tối ưu hơn nhiều thuật toán trước đó.

Vì tập phổ biến thỏa mãn tính chất Apriori nên ta dễ dàng xây dựng được giàn từ tập phổ biến, còn đối với tập mục hữu ích cao thì không thể xây dựng được giàn vì nó không thỏa mãn tính chất Apriori. Tuy nhiên, ta có thể xây dựng được giàn của tập mục có trọng số hữu ích cao vì nó thỏa mãn tính chất phản đơn điệu (Tính chất 3), hình 1 là giàn giao của tập mục có trọng số hữu ích cao được xây dựng từ bảng 3.

Chúng ta thấy rằng tập mục hữu ích cao luôn luôn là tập mục có trọng số hữu ích cao, nhưng ngược lại thì không đúng. Như vậy, giàn của tập mục có trọng số hữu ích cao luôn chứa tập mục hữu ích cao (Tính chất 4). Do đó, để bảo vệ các tập mục hữu ích cao trong quá trình ẩn luật thì cần phải bảo vệ các tập mục có trọng số hữu ích thấp để hạn chế việc làm thay đổi giàn. Như vậy, đối với bài toán ẩn luật kết hợp hữu ích cao nhạy cảm chúng ta có thể áp dụng kỹ thuật ẩn luật ở trong [8].



Hình 1. Giàn giao của tập mục có trọng số hữu ích cao trong bảng 3 (nút viền đậm là nút của tập mục hữu ích cao, bên trong mỗi nút là giá trị của TWU/Utility của tập mục)

Dựa vào lý thuyết giàn trong [11], chúng tôi trình bày lại một số vấn đề về lý thuyết giàn và từ đó xây dựng giàn giao của tập mục trọng trọng số hữu ích cao như sau:

Cho *(L; ≤)* là một tập hợp sắp thứ tự. Ta nói (L; ≤) là một giàn nếu tồn tại chặn trên tối thiểu và chận dưới tối đại với mọi a, b ∈ L, tức tồn tại $\sup(a,b) = a \vee b$ và $\inf(a,b) = a \wedge b$.

Cho $(A;o)$ là một đại số với toán tử hai ngôi o. Đại số $(A;o)$ là nửa giàn nếu o là:

(1) Lũy đẳng: $a \wedge a = a$ và $a \vee a = a$.

(2) Giao hoán: $a \wedge b = b \wedge a$ và $a \vee b = b \vee a$.

(3) Kết hợp: $a \wedge (b \wedge c) = (a \wedge b) \wedge c$ và $a \vee (b \vee c) = (a \vee b) \vee c$.

Một đại số $(L;\wedge,\vee)$ là một giàn nếu *L* là một tập hợp khác rỗng, $(L;\wedge)$ và $(L;\vee)$ là nửa giàn và thỏa mãn luật hấp thụ: $a\vee(a\wedge b)=a$ và $a\wedge(a\vee b)=a$.

Cho *U* là một tập hợp sắp thứ tự hữu hạn khác rỗng. Ta kí hiệu *Poset(U)* là họ toàn thể các tập con của *U* với thứ tự bộ phận là phép bao hàm $\subseteq$. $(Poset(U);\subseteq)$ là một giàn với $\sup(X,Y)=X\cup Y$ và $\inf(X,Y)=X\cap Y$. Nếu $L\subseteq U$ và $(L;\subseteq)$ là một giàn thỏa mãn tính chất $\sup(X,Y)=X\cup Y$ và $\inf(X,Y)=X\cap Y$ với mọi *X, Y* thì $(L;\subseteq)$ được gọi là một giàn tập hợp.

Cho $L=(L;\subseteq)$, nếu L thỏa mãn $\inf(X,Y)=X\cap Y$ với mọi X, Y thì *L* được gọi là một giàn giao. Ta có, giao của các phần tử trong giàn $(L;\subseteq)$ thuộc *L*. Hay nói cách khác giàn giao $(L;\subseteq)$ đóng với phép giao.

Tập các tập mục có trọng số hữu ích cao tạo nên giàn giao: Cho $HTWU_D$ là một tập của các tập mục có trọng số hữu ích cao được khai thác từ cơ sở dữ liệu *D*. Theo tính chất phản đơn điệu *TWU*, nếu $X,Y\in HTWU_D$ thì $X\cap Y\in HTWU_D$. Có thể suy ra được $HTWU_D$ là một giàn giao. Tập sinh của $HTWU_D$ ký hiệu là $Gen(HTWU_D)$ là tập nhỏ nhất trong các tập mục của $HTWU_D$ sao cho mỗi tập mục của $HTWU_D$ được sinh ra bởi giao của một số tập mục trong $Gen(HTWU_D)$. Nói cách khác:

$$HTWU_D = \{X \mid X=\bigcap_{k\in N^*} Y^k, Y^k\in Gen(HTWU_D)\}$$

Tập $Gen(HTWU_D)$ có thể được tính như sau:

$Gen(HTWU_D)=\{X\in HTWU_D \mid d(X)\le 1\}$, ở đây $d(X)=|\{Y\in HTWU_D \mid X\subset Y\}|$

Với mỗi $HTWU_D$, tập chứa tất cả tập mục cực đại của $HTWU_D$ được gọi là $Coatom(HTWU_D)$

$Coatom(HTWU_D)=MAX(Gen(HTWU_D))$

Mỗi tập mục $X\in HTWU_D, Gen(X)=\{X\setminus I_k, I_k\in X, k=1...|X|\}$ là một tập sinh của $Poset(X)\setminus\{X\}$.

Theo tính chất phản đơn điệu *TWU*, những tập mục được chứa trong tập $Gen(HTWU_D)$ có trọng số hữu ích nhỏ nhất trong $HTWU_D$, vì thế các tập mục này là các tập mục dễ bị tổn thương nhất khi giảm giá trị hữu ích. Hơn thế nữa nếu mỗi tập mục của $Gen(HTWU_D)$ là tập có trọng số hữu ích cao thì tất cả các tập mục của $HTWU_D$ là có trọng số hữu ích cao. Do vậy, việc cần duy trì các nút thuộc $Gen(HTWU_D)$ trên giàn trong quá trình đi ẩn luật sẽ giúp giảm tối đa luật bị mất. Vậy có thể đề xuất một thuật toán ẩn luật kết hợp hữu ích cao nhạy cảm giảm tối đa hiệu ứng phụ bằng cách sử dụng Heuristic bảo vệ $Gen(HTWU_D)$ để giàn giao các tập mục có trọng số hữu ích cao bị ảnh hưởng ít nhất trong quá trình ẩn luật.

V. KẾT LUẬN

Ẩn luật kết hợp hữu ích cao nhạy cảm là một bài toán quan trọng nhằm bảo vệ tính riêng tư của tri thức được khai phá từ dữ liệu khi chúng được chia sẻ ra bên ngoài. Trong bài báo này lần đầu tiên chúng tôi đề xuất hướng giải quyết bài toán ẩn luật kết hợp hữu ích cao nhạy cảm. Trên cơ sở đó chúng tôi đã đề xuất thuật toán HHUAL để ẩn luật kết hợp hữu ích cao nhạy. Qua thực nghiệm cho thấy thuật toán HHUAL là hiệu quả. Bên cạnh đó, chúng tôi cũng đã đề xuất một hướng tiếp cận mới dựa trên Giàn giao của tập mục có trọng số hữu ích cao để xác định giao tác mục tiêu và mục mục tiêu đảm bảo hạn chế tối đa hiệu ứng phụ khi tiến hành sửa dữ liệu để ẩn luật.

TÀI LIỆU THAM KHẢO

[1] R. Agrawal, T. Imielinski, A. Swami, "Mining association rules between sets of items in large databases", Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 1993, pp. 207-216.

[2] Erwin A., Gopalan R., Achutan, "Efficient Mining of High Utility Itemsets from Large Datasets", T. Washio: PAKDD 2008, LNAI 5012, 2008, pp. 554-561.

[3] M. Atallah, E. Bertino, A. Elmagamind, M. Ibrahim, and V. S. Verykios "Disclosure limitation of sensitive rules," .In Proc. of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop(KDEX 1999), 1999, pp. 45- 52.

[4]Chun-Wei Lin, Tzung-Pei Hong, Jia-WeiWong, Guo-Cheng Lan, Wen-Yang Lin3, A GA-Based Approach to Hide Sensitive High Utility Itemsets, Hindawi Publishing Corporatione Scientific World Journal Volume 2014.

[5] C. Dong and L. Chen, "A fast secure dot product protocol with application to privacy preserving association rule mining", In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer International Publishing, 2014, pp. 606-617.

[6] Fournier-Viger, P., Wu, C.-W., Zida, S., Tseng, V. S, "FHM: Faster high-utility itemset mining using estimated utility co-occurrence pruning". In: Proc. 21st Intern. Symp. on Methodologies for Intell. Syst., 2014, pp. 83-92.

[7] S-L Wang, T Hong, Y-C Tsai, and H-Y Kao, "Hiding Sensitive Association Rules on Stars" 2010 IEEE International Conference on Granular Computing, 2010, pp 505-508.

[8] Hai Quoc Le, Somjit Arch-int, and Ngamnij Arch-int, "Association Rule Hiding Based on Intersection Lattice", Hindawi Publishing Corporation, Mathematical Problems in Engineering Volume 2013, pp. 776-784.

[9] Hong Yao, Howard J. Hamilton, "Mining Itemset Utilities from Transaction Databases, Journal Data & Knowledge Engineering", Volume 59 Issue 3, December 2006, pp 603-626.

[10] Jayakrushna Sahoo, Ashok Kumar Das, A. Goswami, "An efficient approach for mining association rules from high utility itemsets", Expert Systems with Applications, 2015.

[11] George Grätzer, Lattice Theory: Foundation, 2010 Mathematics Subject Classification, Springer Basel AG, 2011, pp. 1-29.

[12] Jieh-Shan Yeh, Po-Chiang Hsu, "HHUIF and MSICF: Novel algorithms for privacy preserving utility mining", Expert Systems with Applications, Vol: 37, 2010, pp: 4779-4786.

[13] Jayakrushna Sahoo, Ashok Kumar Das, A. Goswami, "An efficient approach for mining association rules from high utility itemsets", Expert Systems with Applications, 2015.

[14] Rajalakshmi Selvaraj, Venu Madhav Kuthadi, A modified hiding high utility item first algorithm (HHUIF) with item selector (MHIS) for hiding sensitive itemsets, International Journal of Innovative Computing, Information and Control ICIC International Volume 9, 2013, pp. 4851-4862.

[15] Thang Mai, Bay Vo, Loan T. T. Nguyen, A lattice-based approach for mining high utility association rules, Information Sciences Volume 399, 2017, pp 81-97.

HHUAL: ALGORITHM HIDING SENSITIVE HIGH UTILITY ASSOCIATION RULE AND PROPOSE AN ENHANCE APPROACH BASED ON INTERSECTION LATTICE

Huynh Trieu Vy, Le Quoc Hai, Truong Ngoc Chau

ABSTRACT: *Hiding sensitive high utility association rule (HSHUAR) is an important problem in privacy preserving data mining. This problem aims at protecting sensitive high utility association rules discovered from databases from being revealed when sharing them outside the parties. The target is to distort data in such a way that the sensitive rules are hidden while the side effects caused by data distortion are minimal. HSHUAR is different from hiding sensitive popular association rule (HSPAR) because HSPAR based on support and confidence constraint while HSHUAR based on the utility and utility confidence constraint. In this paper, we propose an algorithm, entitled HHUAL, for HSHUAR. By analyzing this algorithm, we propose an enhance approach which can be applied to reduce the side effects of hiding process. The contributions of this paper are: (1) This is the first time an algorithm for HSHUAR problem is proposed; (2) Intersection lattice of high weight utility itemsets is applied in order to determine victim items and transactions for data distortion process.*