RESEARCH ARTICLE

KHẢO SÁT PHƯƠNG PHÁP ẨN LUẬT KẾT HỢP TRONG DỮ LIỆU GIAO DỊCH

Trần Minh Thái, Trần Anh Duy, Lê Thị Minh Nguyện

Khoa Công nghệ thông tin, Trường Đại học Ngoại ngữ -Tin học TP.HCM

minhthai@huflit.edu.vn, duy.ta@huflit.edu.vn, nguyenltm@huflit.edu.vn

TÓM TẮT— Khai thác dữ liệu bảo toàn tính riêng tư (Privacy-Preserving Data Mining - PPDM) là một lĩnh vực nghiên cứu tương đối mới trong cộng đồng khai thác dữ liệu và đã tồn tại khoảng hơn một thập kỷ. PPDM nghiên cứu các hiệu ứng phụ của phương pháp khai thác dữ liệu có nguồn gốc từ sự xâm nhập vào sự riêng tư của các cá nhân và tổ chức. Một số phương pháp tiếp cận để giải quyết vấn đề này đã được nghiên cứu và áp dụng. Các phương pháp được đề xuất có thể được phân loại theo hai hướng nghiên cứu chính đó là ẩn dữ liệu và ẩn tri thức. Ẩn dữ liệu là hướng nghiên cứu về tính riêng tư trong các dữ liệu thô hay thông tin, có thể được đảm bảo trong quá trình khai thác dữ liệu. Các phương pháp của nhóm này tác động vào bản thân dữ liệu nhằm mục đích làm ẩn các thông tin nhạy cảm bằng các phương pháp khác nhau. Ẩn tri thức liên quan đến các phương pháp nhằm bảo vệ các kết quả khai thác dữ liệu nhạy cảm chứ không phải chính dữ liệu thô. Đây là hướng ứng dụng chính của các công cụ và thuật toán khai thác dữ liệu. Trong đó, ẩn luật kết hợp là một hướng nghiên cứu trong nhóm ẩn tri thức. Trong bài báo này, chúng tôi tập trung vào việc trình bày bài toán liên quan đến ẩn luật kết hợp. Bên cạnh đó, chúng tôi khảo sát các kỹ thuật ẩn luật kết hợp và so sánh các phương pháp đã được đề xuất nhằm làm rõ sự thay đổi hướng tiếp cận của các phương pháp ẩn luật. Cuối cùng, các phương pháp thực nghiệm cùng với các độ đo được sử dụng để so sánh hiệu quả của các thuật toán cũng được trình bày cụ thể trong bài báo.

Từ khóa— Ẩn luật kết hợp; bảo toàn tính riêng tư; khai thác dữ liệu; ẩn luật nhạy cảm.

I. GIỚI THIỆU

Hiện nay, trong bối cảnh số lượng thông tin được trao đổi giữa các công ty, cơ quan chính phủ và các tổ chức được gia tăng rất nhanh chóng. Hơn nữa, cùng với sự phát triển của công nghệ khai thác thông tin, các mối quan hệ tiềm ẩn giữa các đối tượng bên trong dữ liệu có thể được khám phá ra bằng cách suy đoán, nhằm mục đích đưa ra quyết định hoặc khám phá thông tin cá nhân của người dùng. Do vậy, một vấn đề lớn phát sinh là các tri thức được khai thác bằng kĩ thuật khai thác dữ liệu có thể ngầm chứa các bí mật, thông tin riêng tư hoặc thông tin nhạy cảm (ví dụ như số chứng minh nhân dân, địa chỉ nhà, thông tin tài khoản ngân hàng, v.v.). Vấn đề này trở nên đặc biệt quan trọng khi các tổ chức tiến hành công khai các thông tin. Trong trường hợp này, sử dụng các kĩ thuật khai thác dữ liệu có thể dẫn đến các nguy cơ về riêng tư hay dữ liệu bị lạm dụng. Vấn đề tương tự có thể xảy ra khi chia sẻ dữ liệu giữa các tổ chức với nhau. Dữ liệu có thể bị phân tích bởi đối tác hoặc đối thủ cạnh tranh để tìm kiếm các thông tin nhạy cảm hay thông tin chiến lược, mà có thể ảnh hưởng đến lợi nhuận của công ty hoặc các nguy cơ bảo mật. Trong bối cảnh như vậy, sự cần thiết có một lĩnh vực nghiên cứu để vừa có thể khai thác dữ liệu vừa đảm bảo những tri thức nhạy cảm trong dữ liệu không bị khai thác. Những lý do đó đã thúc đẩy lĩnh vực khai thác dữ liệu đảm bảo sự riêng tư ra đời và đang được phát triển mạnh mẽ trong những năm gần đây. Từ khi công trình tiên phong của Agrawal và Srikant [1] và của Y. Lindell và Pinkas [2] vào năm 2000, một số phương pháp đã được đề xuất nhằm mục đích đảm bảo tính riêng tư trong khai thác dữ liệu. Dựa vào phương pháp tiếp cận được đề xuất, chúng có thể được chia thành hai hướng nghiên cứu chính là ẩn dữ liệu và ẩn tri thức.

Phương pháp ẩn dữ liệu nhằm sửa đổi dữ liệu thô nhạy cảm thông qua các kỹ thuật ngẫu nhiên [1], [3], [4] hoặc sửa đổi các thông tin khả định danh (quasi-identifier) bằng cách sử dụng các kỹ thuật nặc danh để làm mờ đi chủ sở hữu bản ghi [5], [6] và không phụ thuộc vào loại phân tích. Các thuộc tính khả - định danh là các thuộc tính không thể tự có khả năng xác định thông tin chủ sở hữu bản ghi, nhưng khi chúng được kết hợp với nhau có thể xác định các thực thể như tuổi tác và zip code [6], [7].

Phương pháp ẩn tri thức tập trung vào việc bảo vệ các kết quả khai thác dữ liệu nhạy cảm [8]. Các mối đe dọa sự riêng tư gây ra bởi các kết quả khai thác dữ liệu đã được giới thiệu đầu tiên bởi O'Leary [9], [10]. Hướng tiếp cận PPDM có thể được áp dụng trong các tác vụ khai thác dữ liệu khác nhau chẳng hạn như khai thác luật kết hợp, gom cụm và phân lớp. Khai thác luật kết hợp bảo toàn tính riêng tư liên quan đến việc thanh lọc dữ liệu mà có thể dẫn đến tiết lộ tri thức riêng tư và bí mật [8]. Phương pháp này được gọi là ẩn luật kết hợp hoặc thanh lọc dữ liệu.

Ẩn luật kết hợp là một trong những lĩnh vực nghiên cứu chính trong PPDM được đề xuất lần đầu tiên bởi Atallah và cộng sự [11]. Quá trình ẩn luật kết hợp là thanh lọc các giao dịch để giảm độ tin cậy hoặc độ hỗ trợ của các mẫu nhạy cảm dưới một ngưỡng xác định trước. Quá trình này tạo ra một số hiệu ứng phụ trên dữ liệu đã thanh lọc như là một số các mẫu không nhạy cảm bị mất hay các mẫu mới có thể được sinh ra. Một giải pháp thanh lọc mà ẩn đi tất cả các tri thức nhạy cảm và cũng không tạo ra các hiệu ứng phụ được biết đến như một "giải pháp tối ưu". Tuy nhiên, vấn đề để tìm kiếm một quá trình thanh lọc dữ liệu tối ưu là một vấn đề NP-hard [11].

Nội dung bài báo sẽ tập trung vào khảo sát các phương pháp ẩn tri thức trong khai thác dữ liệu đảm bảo tính riêng tư của tập phổ biến và ẩn luật kết hợp nhằm ẩn các luật kết hợp nhạy cảm. Nội dung của bài báo gồm 5 phần. Trong đó, phần I trình bày giới thiệu bài toán; phần định nghĩa bài toán thể hiện trong mục II; phần III trình bày các công trình nghiên cứu liên quan; mô tả các độ đo đánh giá trong mục IV; và cuối cùng phần V là phần kết luận.

II. ĐỊNH NGHĨA BÀI TOÁN

Khai thác luật kết hợp là một trong những kỹ thuật khai thác dữ liệu quan trọng nhất, được giới thiệu lần đầu bởi Agrawal và cộng sự [12].

Cho I = {i_1, i_2, i_3, ..., i_m} là một tập của các item và D là một cơ sở dữ liệu (CSDL) bao gồm nhiều giao dịch, D = (t_1, t_2, ..., t_n). Mỗi giao dịch t_i là một tập con của I ($t_i \subseteq I$). Tập các luật kết hợp được rút ra từ D là R. Mỗi luật kết hợp được biểu diễn theo dạng: A → B. Trong đó, A là tiền đề hoặc vế trái của luật và B là kết quả hoặc vế phải của luật, sao cho A, B ⊂ I và A ∩ B = ∅. Hai tiêu chí được xem xét trong việc khai thác luật kết hợp bao gồm: Thứ nhất là độ hỗ trợ của luật cho biết tần suất của một luật trong dữ liệu và được tính bằng công thức: $Sup(A \to B) = \frac{|A \cup B|}{|D|}$ (trong đó, $Sup(A \to B)$ là độ hỗ trợ của luật kết hợp: $A \to B$, $|A \cup B|$ là số giao tác chứa tất cả các item trong cả hai tập A và B, |D| là tổng số giao tác trong dữ liệu). Thứ hai là độ tin cậy luật cho biết độ mạnh của luật trong dữ liệu và được tính bằng công thức: $Conf(A \to B) = \frac{|A \cup B|}{|A|}$ (trong đó, $Conf(A \to B)$ là độ tin cậy của luật kết hợp: $A \to B$, $|A \cup B|$ là số giao tác chứa tất cả các item trong cả hai tập A và B, $|A|$ là số giao tác chứa tất cả các item của tập A).

Đối với mỗi luật kết hợp, một ngưỡng hỗ trợ tối thiểu (*Minimum Support Threshold - MST*) và một ngưỡng tin cậy tối thiểu (*Minimum Confidence Threshold - MCT*) được xác định trước bởi người dùng. Một luật kết hợp thỏa mãn khi độ hỗ trợ của nó lớn hơn hoặc bằng MST và độ tin cậy của nó cũng lớn hơn hoặc bằng MCT. Khai thác luật kết hợp thường bao gồm hai giai đoạn: Giai đoạn 1 tìm tập các item phổ biến được khai thác với ngưỡng MST và giai đoạn 2 là luật kết hợp mạnh được sinh ra từ các tập phổ biến thu được trong giai đoạn 1 với ngưỡng MCT.

Dựa trên tính chất khai thác luật kết hợp, một luật nhạy cảm tiết lộ sự riêng tư khi độ hỗ trợ của nó lớn hơn hay bằng MST hoặc độ tin cậy của nó cao hơn hay bằng MCT. Do đó, để ẩn một luật nhạy cảm, cần giảm độ hỗ trợ hay độ tin cậy của nó dưới ngưỡng tối thiểu để luật không thể bị phát hiện từ CSDL đã được thanh lọc. Như vậy, ẩn luật kết hợp có thể được phát biểu: Cho một CSDL giao dịch, tập các mẫu có ý nghĩa được khai thác từ CSDL ban đầu và một tập con các mẫu nhạy cảm trong các mẫu được khai thác. Chúng ta muốn chuyển đổi CSDL thành một CSDL đã được thanh lọc sao cho tất cả các mẫu nhạy cảm được ẩn, trong khi các mẫu không nhạy cảm vẫn có thể được khai thác bình thường.

Trong quá trình ẩn luật kết hợp [13], ngưỡng hỗ trợ và tin cậy được xem là mức nhạy cảm. Nếu độ hỗ trợ hoặc độ tin cậy của một luật mạnh và phổ biến là trên một mức nhạy cảm nhất định, quá trình ẩn nên được áp dụng để độ phổ biến hoặc độ mạnh của luật bị giảm. Quá trình này bao gồm bốn bước: rút trích mẫu, đặc tả, thanh lọc và đánh giá.

Bước 1 Rút trích mẫu: một tập các itemset phổ biến hay các luật kết hợp được khai thác từ CSDL ban đầu thông qua một thuật toán khai thác luật kết hợp.

Bước 2 Đặc tả: một số mẫu hay item mà vi phạm sự riêng tư được xác định bởi người sử dụng được gọi là mẫu nhạy cảm.

Bước 3 Thanh lọc: CSDL được thanh lọc bằng cách sử dụng một thuật toán thanh lọc để ẩn các mẫu nhạy cảm. Áp dụng một thuật toán tối ưu làm giảm các hiệu ứng phụ trên CSDL đã thanh lọc. Điều này phụ thuộc chủ yếu vào loại mẫu. Một tập phổ biến không thể được ẩn bằng cách sử dụng một thuật toán ẩn luật trong khi một luật kết hợp có thể được ẩn bằng cách sử dụng một thuật toán ẩn itemset để giảm độ hỗ trợ của itemset hoặc bằng cách sử dụng một thuật toán ẩn luật để giảm độ tin cậy của luật.

Bước 4 Đánh giá hiệu ứng phụ của quá trình thanh lọc: được đo đối với các mẫu nhạy cảm và không nhạy cảm mà đã được xác định tại bước 2. Với mục đích này, việc khai thác luật kết hợp với ngưỡng tối thiểu cho trước được áp dụng trên CSDL thanh lọc để xác nhận mức độ hữu dụng và bảo đảm của CSDL thanh lọc.

Khi mục tiêu của nhà quản trị CSDL hoặc chủ sở hữu dữ liệu được đáp ứng, CSDL thanh lọc được chia sẻ. Nếu không, quá trình thanh lọc lại được thực hiện một lần nữa bằng cách sử dụng các thông số khác nhau hoặc sử dụng thuật toán khác. Các phương pháp ẩn luật kết hợp nhằm mục đích lọc sạch CSDL ban đầu sao cho ít nhất một trong các mục tiêu sau được đáp ứng: (1) Không luật nhạy cảm nào được chỉ định bởi người sở hữu trong CSDL ban đầu với ngưỡng hỗ trợ và tin cậy được chỉ định trước có thể bị tiết lộ ra trong CSDL đã được thanh lọc khi dữ liệu này được khai thác với cùng hay ngưỡng cao hơn; (2) Tất cả các luật không nhạy cảm đã được rút

trích trong dữ liệu ban đầu với ngưỡng hỗ trợ và tin cậy chỉ định trước, có thể được khai thác lại trong dữ liệu thanh lọc với cùng hay ngưỡng cao hơn; và (3) Không luật nào không thuộc các luật kết hợp được khai thác trong dữ liệu ban đầu với ngưỡng độ tin cậy và độ hỗ trợ chỉ định trước có thể xuất hiện trong dữ liệu thanh lọc khi dữ liệu này được khai thác với cùng hay ngưỡng cao hơn.

Dựa vào ba mục tiêu này, quá trình thanh lọc của một thuật toán ẩn được xem là trọn vẹn khi mà ít gây ảnh hưởng nhất đến các CSDL ban đầu, giữ lại được các mẫu không nhạy cảm và ẩn được tất cả các luật kết hợp nhạy cảm. Một giải pháp giải quyết được tất cả ba mục tiêu trên (không có "hiệu ứng phụ") được gọi là lý tưởng hay tối ưu. Trường hợp không xử lý hoàn toàn các mục tiêu này nhưng khả thi được gọi là gần đúng.

Như vậy, các phương pháp ẩn luật kết hợp chủ yếu khác nhau về khả năng mà chúng có thể đáp ứng các mục tiêu nói trên (đặc biệt là thứ hai và thứ ba). Đối với mục tiêu thứ nhất, nó là điều kiện quyết định tính khả thi của một giải pháp ẩn, tức là một giải pháp ẩn hiệu quả phải ẩn hết tất cả các luật kết hợp nhạy cảm trong CSDL. Điều này có nghĩa rằng mục tiêu đầu tiên luôn luôn có thể đạt được bất chấp các yếu tố khác. Một cách đơn giản nhất thì một giải pháp ẩn có khả thi trong một CSDL là chọn một item từ các itemset sinh ra của từng luật nhạy cảm và xóa nó ra khỏi tất cả các giao dịch của dữ liệu.

Với một CSDL D gồm các giao dịch, và một ngưỡng MST và MCT tạo bởi chủ của dữ liệu. Sau khi thực hiện khai thác luật kết hợp trong D với ngưỡng MST và MCT, tạo ra một tập các luật kết hợp R, với một tập con R_S của R chứa các luật được xem là nhạy cảm theo quan điểm của chủ dữ liệu ($Rs \subset R$).

Cho tập các luật kết hợp nhạy cảm R_S, mục tiêu của các phương pháp ẩn luật kết hợp là tạo ra một dữ liệu thanh lọc D' từ D, để bảo vệ các luật kết hợp nhạy cảm R_S khỏi bị lộ, trong khi giảm thiểu ảnh hưởng đến các luật không nhạy cảm hiện có trong R. Quá trình ẩn có thể ảnh hưởng đến các luật không nhạy cảm trong D hoặc các luật tiền mạnh trong D. Các luật tiền mạnh là những luật với độ hỗ trợ không nhỏ hơn MST và độ tin cậy nhỏ hơn MCT. Một luật tiền mạnh có thể trở nên mạnh khi độ tin cậy của nó lớn hơn hoặc bằng MCT. Một luật không nhạy cảm trong D có thể chấm dứt mạnh khi độ hỗ trợ của nó giảm xuống dưới MST hay độ tin cậy của nó giảm xuống dưới MCT trong D' do việc loại bỏ item. Bảng 1 trình bày tóm tắt các hiệu ứng phụ trong quá trình ẩn luật kết hợp.

Bảng 1. *Các hiệu ứng phụ trong quá trình ẩn luật kết hợp*

Trước quá trình ẩn	Sau quá trình ẩn	Kết quả
Supp(r) ≥ MST và *Conf*(r) ≥ MCT và r ∈ R_S	*Supp*(r) ≥ MST và *Conf*(r) ≥ MCT	**HF** *(Hiding Failure)*
Supp(r) ≥ MST và *Conf*(r) ≥ MCT và r ∈ (R - R_S)	*Supp*(r) < MST hay *Conf*(r) < MCT	**LR** *(Lost Rules)*
Supp(r) < MST hay *Conf*(r) < MCT và r ∉ R	*Supp*(r) ≥ MST và *Conf*(r) ≥ MCT	**GR** *(Ghost Rules)*

III. CÁC CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN

Vấn đề ẩn luật kết hợp được đề xuất đầu tiên bởi Atallah và cộng sự [11]. Nhóm tác giả sử dụng phương pháp biến dạng để giảm độ hỗ trợ của các itemset phổ biến. Tiếp theo đó, Oliveira và cộng sự [14] đưa ra một cách tiếp cận ẩn nhiều luật. Các ảnh hưởng lên các mẫu không nhạy cảm được xem xét trong cách tiếp cận này. Wu và cộng sự [15] đã đề xuất một phương pháp nhằm tránh tất cả các "hiệu ứng phụ" trong quá trình ẩn luật thay vì ẩn tất cả các luật nhạy cảm. Bảng 2 mô tả các phương pháp tiếp cận tiêu biểu trong việc ẩn luật kết hợp được đề xuất.

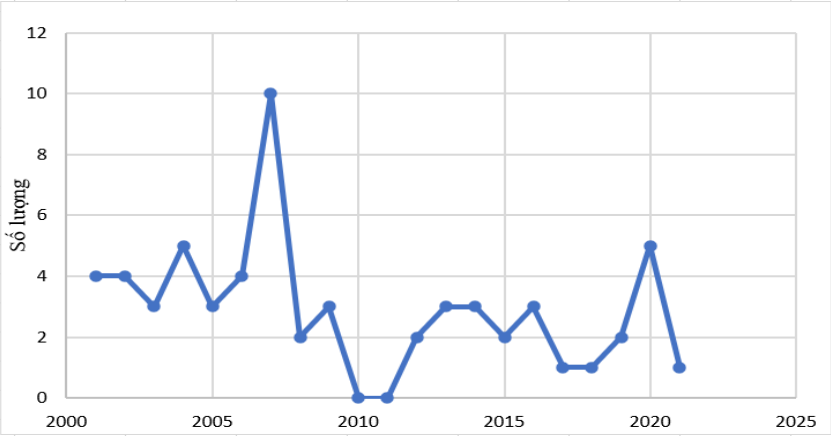
Bảng 2. *Các phương pháp tiếp cận tiêu biểu trong ẩn luật kết hợp*

Năm	Tác giả	Phương pháp tiếp cận
2001	Dasseni và cộng sự [16]	Đề xuất ba thuật toán để ẩn luật nhạy cảm. Hai thuật toán đầu giảm độ tin cậy của luật bằng cách tăng hỗ trợ ở vế trái của luật và giảm tương ứng độ hỗ trợ ở vế phải, thuật toán thứ ba giảm hỗ độ trợ ở tập phổ biến của luật.
	Saygin và cộng sự [17]	Đề xuất thuật toán giảm độ tin cậy (CR), và thuật toán ẩn tập phổ biến (GIH). Tương tự như ba thuật toán của Dasseni và cộng sự [16], nhưng có sự khác biệt trong việc thay thế các item ẩn bằng cách đánh dấu thay vì loại bỏ các item này đi.
2002	Oliveira và Zaiane [14]	Đề xuất bốn thuật toán ẩn itemset, gồm: Maximum Frequency Item Algorithm (MaxFIA), Minimum Frequency Item Algorithm (MinFIA), Item Grouping Algorithm (IGA), và Naïve. Những thuật toán này xét tác động của việc sửa đổi giao dịch và item trên CSDL được thanh lọc bằng cách tính toán sự xung đột của nó.
2003	Oliveira và Zaiane [18], [19]	Trong [18] đề xuất hai thuật toán: Thuật toán Random Algorithm (RA) và Thuật toán Round Robin Algorithm (RRA), để ẩn các luật nhạy cảm bằng cách giảm các tập phổ biến. Hai thuật toán này xét tác động của việc thay đổi giao dịch đối với các luật nhạy cảm. Thuật toán Sliding Window size (SWA) [19] đề xuất ẩn các itemset nhạy cảm trong một lần quét trên tập dữ liệu. Trước tiên, thuật toán sao chép các giao dịch không nhạy cảm vào CSDL đã được thanh lọc và sau đó sử dụng cơ chế lập chỉ mục để tăng tốc quá trình ẩn. Đa số các thuật toán khác chỉ có một ngưỡng công khai duy nhất

Năm	Tác giả	Phương pháp tiếp cận
		được gán cho tất cả các luật nhạy cảm. Trong khi đó, mỗi ngưỡng công khai trong SWA được gán cho từng luật nhạy cảm. Tập các quyền khai thác được tham chiếu đến tập hợp các ánh xạ của luật nhạy cảm vào ngưỡng công khai tương ứng.
2004	Pontikakis và cộng sự [20]	Trong [20] đề xuất thuật toán biến dạng dựa trên độ ưu tiên Priority-based Distortion Algorithm (PDA) và thuật toán biến dạng có sắp xếp dựa trên trọng số Weight-based Sorting Distortion Algorithm (WSDA) thực hiện ẩn các luật nhạy cảm bằng phương pháp heuristic trong giai đoạn chọn item ở PDA và trong giai đoạn chọn giao dịch ở WSDA. Đây là hai thuật toán đầu tiên gán trọng số cho các giao dịch. Thuật toán Blocking (BA) [20] tạo ra các luật không tồn tại trong tập dữ liệu gốc bằng cách thêm ẩn số (đánh dấu) vào giao dịch.
2005	Menon và cộng sự [21]	Việc ẩn itemset phổ biến được xây dựng dưới dạng Constraint Satisfaction Problem (CSP). Đề xuất thuật toán Blanket và Intelligence giải quyết CSP bằng cách sử dụng lập trình số nguyên để giảm thiểu số lượng giao dịch được thanh lọc, thuật toán này sử dụng phương pháp heuristics để tìm ra các item cần xử lý.
	Sun và Yu [22]	Đề xuất Border-Based Approach (BBA) lấy cảm hứng từ lý thuyết biên của các tập phổ biến [23] để duy trì chất lượng biên của các tập phổ biến không nhạy cảm trong dàn tập phổ biến.
2006	Divanis và Verykios [24]	Đưa ra khái niệm về khoảng cách giữa CSDL gốc và CSDL đã được thanh lọc trong thuật toán nội biên (Inline). Thuật toán này dựa vào quá trình sửa đổi đường biên để xác định số lượng item ít nhất để thanh lọc thay vì xét số lượng giao dịch được thanh lọc tối thiểu. Nó giải quyết CSP bằng cách sử dụng Binary Integer Programming (BIP).
	Moustakides và Verykios [25]	Đề xuất Max-Min1 và Max-Min2 nhằm kiểm soát tác động của việc thanh lọc đối với các tập itemset dễ bị tấn công nhiều trong quá trình ẩn thay vì tất cả các itemset trên đường biên như trong [24]
2007	Amiri [26]	Đề xuất ba phương pháp heuristics: Aggregate, Disaggregate và Hybrid vượt trội hơn SWA vì phương pháp này cung cấp dữ liệu hữu ích cao hơn và độ biến dạng thấp hơn.
	Li và Yeh [27]	Đề xuất thuật toán Maximum Item Conflict First (MICF) làm tốt hơn IGA về việc giảm số lượng item bị xóa và khắc phục sự chồng chéo giữa các nhóm.
	Wang và cộng sự [28]	Mở rộng các thuật toán ISL và DSR [29] bằng kỹ thuật biến dạng. Thuật toán Decrease Confidence by Decrease Support (DCDS) và thuật toán Decrease Confidence by Increase Support (DCIS) tiếp tục được Wang và cộng sự [30] đề xuất để tự động ẩn các luật mà không cần tiền khai thác và chọn luật ẩn.
	Verykios và cộng sự [31]	Cải tiến thuật toán BA bằng cách áp dụng phương pháp heuristic lựa chọn giao dịch đã được sử dụng trong WSDA [20].
	Wang và cộng sự [30]	Trình bày phương pháp hiệu ứng phụ giới hạn để phân loại tất cả các sửa đổi hợp lệ liên quan đến các luật nhạy cảm, các luật không nhạy cảm và các luật giả có thể bị ảnh hưởng khi được sửa đổi. Sau này, phương pháp heuristic được sử dụng để sửa đổi các giao dịch nhằm tăng số lượng luật nhạy cảm ẩn, đồng thời giảm số lượng các item được sửa đổi [32]
2008	Wang và cộng sự [33]	Đề xuất thuật toán Decrease Support and Confidence (DSC) để ẩn luật kết hợp dự đoán.
	Menon và Sarkar [34]	Mở rộng thuật toán [21] để giảm thiểu cả số lượng giao dịch được thanh lọc và số lượng itemset không nhạy cảm bị mất.
2009	Divanis và Verykios [32]	Bổ sung phần CSDL mở rộng vào CSDL gốc thay vì sửa đổi các giao dịch hiện có. Phần CSDL mở rộng chứa một tập những giao dịch làm giảm bớt tầm quan trọng của các mẫu nhạy cảm ở mức độ mà nó không thú vị theo quan điểm của các thuật toán khai thác dữ liệu, đồng thời ảnh hưởng tối thiểu đến tầm quan trọng của các itemset không nhạy cảm. Đề xuất một thuật toán lai ghép giữa CSP, BIP và sửa đổi đường biên để ẩn các itemset nhạy cảm.
	Wang [35]	Cải tiến [33] và giới thiệu thuật toán Maintenance of Sanitizing Informative association rules (MSI) để bảo vệ thông tin nhạy cảm khi CSDL được cập nhật thường xuyên. Tập dữ liệu mới được bổ sung vào được MSI thanh lọc riêng và sau đó được kết hợp với CSDL gốc.
	Divanis và Verykios [36]	Cải tiến cách tiếp cận Inline bằng một quy trình hai giai đoạn. Quá trình thanh lọc kết thúc trong giai đoạn đầu, nếu luật nhạy cảm được ẩn mà không gây ra hiệu ứng phụ. Nếu không, bản đối ngẫu của thuật toán Inline được thực hiện trong giai đoạn thứ hai để loại bỏ các bất đẳng thức CSP không khả thi, cho đến khi chỉ còn CSP khả thi, và sau đó xử lý CSP để có được tập dữ liệu đã được thanh lọc.

Năm	Tác giả	Phương pháp tiếp cận
2012	Grätzer [37]	Lần đầu tiên đưa ra thuật toán Ẩn luật dựa trên dàn giao (Intersection Lattice-based Association Rule Hiding - ILARH) để lựa chọn item ẩn.
2013	Hai và cộng sự [38]	Trình bày ẩn luật kết hợp dựa trên dàn (Association Rule Hiding based on Intersection Lattice - ARHIL) và Heuristic để giảm độ tin cậy và hỗ trợ dựa trên dàn (Heuristic for Confidence and Support Reduction based on Intersection Lattice - HCSRIL) để ẩn các luật.
	Hong và cộng sự [39]	Áp dụng khái niệm tần suất tài liệu nghịch đảo (TFIDF), và đưa ra tần suất CSDL nghịch đảo (SIF-IDF) cho các item nhạy cảm để gán trọng số cho mỗi giao dịch.
2014	Lin và cộng sự [40] [41]	Sử dụng thuật toán di truyền (GA) để lựa chọn giao dịch ẩn. Thuật toán Compact Prelarge GA-based algorithm to Delete Transactions (cpGA2DT) [40] xóa các giao dịch được chỉ định, trong khi thuật toán được đề xuất trong [41] tạo và chèn các giao dịch mới vào CSDL.
	Cheng và cộng sự [42]	Đề xuất thuật toán ẩn luật dựa trên cơ sở tối ưu hóa đa mục tiêu (Evolutionary Multi-objective Optimization-base Rule Hiding - EMO-RH). Kiến trúc của thuật toán này dựa trên nền tảng PISA [43].
2015	Lin và cộng sự [44]	Giới thiệu hai thuật toán ẩn tập phổ biến, đó là thuật toán Simple Genetic Algorithm to Delete Transactions (sGA2DT) và Pre-large Genetic Algorithm to Delete Transactions (pGA2DT) sử dụng di truyền để chọn giao dịch và sau đó xóa giao dịch khỏi CSDL ban đầu.
2016	Lin và cộng sự [45]	Hạn chế của các thuật toán dựa trên GA là một số tham số phải được chỉ định bởi người dùng, chẳng hạn như kích thước nhiễm sắc thể, tỷ lệ đột biến và tỷ lệ lai ghép. Bên cạnh đó, các thuật toán này yêu cầu xác định thủ công số lượng giao dịch để xóa. Để giải quyết những vấn đề này, nhóm tác giả đề xuất thuật toán Particle Swarm Optimization based algorithm to Delete Transactions (PSO2DT) có thể xác định số lượng giao dịch tối đa có thể bị xóa, cũng như ít tham số hơn.
	Afshari và cộng sự [46]	Đề xuất thuật toán Cuckoo Optimization Algorithm for Association Rules Hiding (COA4ARH) để ẩn luật nhạy cảm bằng thuật toán Cuckoo [47]
	Cheng và cộng [48]	Đề xuất thuật toán sắp xếp theo mức độ liên quan, xây dựng phương pháp heuristic để xác định các giao dịch thanh lọc. Để giảm tỷ lệ biến dạng, thuật toán tính toán số lượng giao dịch tối thiểu cần phải sửa đổi để ẩn luật nhạy cảm.
2017	Telikani và Shahbahrami [49]	Đề xuất thuật toán Decrease the Confidence of Rule (DCR) để cải thiện giải pháp MaxMin [25] sử dụng hai phương pháp heuristics để ẩn luật. Trong thuật toán này, kết hợp phương pháp tiếp cận MaxMin và phương pháp heuristic được xây dựng để chọn các item, trong khi đó đối với những giao dịch nhạy cảm chọn giải pháp heuristic.
2018	Talebi và Dehkordi [50]	Lấy cảm hứng từ vật lý, tính bầy đàn và sự tiến hóa trong thuật toán tối ưu hóa metaheuristic [51], thuật toán tối ưu hóa trường điện từ (Electromagnetic Field Optimization Algorithm - EFO4ARH). Thuật toán sử dụng kỹ thuật làm nhiễu dữ liệu để ẩn các luật, đồng thời làm giảm "hiệu ứng phụ" và bảo toàn chất lượng dữ liệu tốt hơn.
2019	Bac Le và cộng sự [52]	Đưa ra giải pháp xác định các giao dịch quan trọng dựa trên số lượng tập phổ biến tối đa không nhạy cảm nhưng có chứa ít nhất một luật nhạy cảm. Chúng có thể bị ảnh hưởng trực tiếp bởi các giao dịch đã sửa đổi, sau đó tính số lượng giao dịch nhỏ nhất để sửa đổi trước nhằm giảm thiểu thiệt hại cho CSDL.
	Shaoxin Li và cộng sự [53]	Những phương pháp được đề xuất trước đây đều gây ra nhiều hiệu ứng phụ do thực hiện thay đổi trên CSDL. Để giảm bớt vấn đề này, nhóm tác giả áp dụng khai thác tập hữu ích cao, đề xuất thuật toán mới dựa trên lập trình tuyến tính số nguyên (Integer Linear Programming - ILP) thu được tỷ lệ hiệu ứng phụ thấp hơn và không lộ thông tin nhạy cảm trong CSDL đã được thanh lọc.
2020	Akbar Telikani và cộng sự [54]	Đề xuất thuật toán ẩn luật mới dựa trên cách tiếp cận thuộc địa đàn ong nhân tạo nhị phân (Artificial Bee Colony - ABC) có khả năng thăm dò tốt. Cải tiến thuật toán ABC thành thuật toán Improved Binary ABC (IBABC) để tăng khả năng khai thác bằng cách thiết kế một cơ chế tạo vùng lân cận mới để cân bằng giữa thăm dò và khai thác. Đồng thời, phương pháp tiếp cận IBABC kết hợp với thuật toán ẩn luật gọi là ABC4ARH để chọn các giao dịch nhạy cảm cần sửa đổi.
	S. Jangra và D. Toshniwal [55]	Các phương pháp như di truyền (GA), tối ưu hóa bầy đàn (PSO) và tối ưu hóa đàn kiến (ACO) thực hiện ẩn các mẫu nhạy cảm bằng cách xóa các giao dịch nhạy cảm dẫn đến mất dữ liệu là thách thức rất lớn đối với các thuật toán trên đồng thời hiệu suất các thuật toán tiến hóa càng bị suy giảm khi áp dụng trên các tập dữ liệu dày. Do đó [55]

Năm	Tác giả	Phương pháp tiếp cận
		lấy cảm hứng từ PSO đề xuất thuật toán Victim Item Deletion based PSO (VIDPSO) để thanh lọc các tập dữ liệu đặc.
2021	Bac Le và cộng sự [56]	Khi khai phá tập dữ liệu lớn thì các giải pháp trước đây hầu như ít phù hợp nên thuật toán EFODBV4ARH, áp dụng cấu trúc dữ liệu vector bit động kết hợp với phương pháp tối ưu hóa trường điện từ hiệu quả hơn thuật toán trường điện từ EFO4ARH [50].



Hình 1. *Số lượng thuật toán ẩn luật kết hợp được công bố từ năm 2001 đến 2021.*

Hình 1 thể hiện thống kê số lượng thuật toán ẩn luật kết hợp được đề xuất từ năm 2001 đến năm 2021. Kỹ thuật chặn và biến dạng dữ liệu đã được sử dụng vào năm 2001 để sửa đổi các giao dịch nhạy cảm. Vào năm 2005, trọng tâm của các thuật toán là duy trì tính hữu ích và độ chính xác của CSDL thanh lọc bằng lý thuyết biên, vì thế các phương pháp tiếp cận chính xác và đường biên đã xuất hiện vào năm 2005. Đồng thời với việc loại bỏ kỹ thuật chặn vào năm 2007, kỹ thuật xóa giao dịch đã được giới thiệu bởi Amiri (2007). Kỹ thuật chèn giao dịch sử dụng vào năm 2009 với mục đích làm giảm tầm quan trọng của các itemset nhạy cảm. Vào năm 2012, lý thuyết dàn giao đã được áp dụng vào trong xử lý. Tiếp theo, thuật giải di truyền (GA) đầu tiên được áp dụng để chọn các giao dịch vào năm 2014. Từ đó phương pháp tiếp cận tiến hóa đã được tập trung đề xuất vào thời bấy giờ. Đến năm 2018, phương pháp điện từ trường lấy cảm hứng từ vật lý các thuật toán tối ưu metaheuristic được công bố [50] và phương pháp này vẫn đang được cải tiến [56]. Dựa trên các phương pháp GA, lấy cảm hứng từ tự nhiên như phương pháp bầy đàn tiếp tục được nghiên cứu vào năm 2020. Hình 2 thể hiện trục thời gian của các cách tiếp cận chính được đề xuất và nghiên cứu liên quan đến ẩn luật kết hợp.

Hình 2. *Phương pháp tiếp cận chính của các thuật toán ẩn luật kết hợp.*

Trong những năm gần đây, các thuật toán meta-heuristic đã được sử dụng để khai thác luật kết hợp đảm bảo sự riêng tư, chẳng hạn như "thuật toán tối ưu hóa Cuckoo". Thuật toán Cuckoo được giới thiệu lần đầu tiên vào năm 2009 bởi Yang và Deb [47]. Gần đây nhất, Mahtab Hossein Afshar và cộng sự [46] đã phát triển "thuật toán tối ưu hóa Cuckoo" cho vấn đề ẩn luật kết hợp. Tuy nhiên, các thuật toán này vẫn còn một số "hiệu ứng phụ", đặc biệt là về việc mất luật còn rất cao.

Trong phần II tiếp theo dùng để trình bày quy cách viết bài báo, phần III đưa ra một số thông tin khác.

IV. TIÊU CHUẨN ĐÁNH GIÁ

Điều quan trọng của việc ẩn luật kết hợp là đánh giá hiệu ứng phụ và hiệu quả CSDL được tạo ra bởi quá trình thanh lọc. Như vậy, cần phải xác định một tập các độ đo cho mục đích này. Đã có rất nhiều công trình đề xuất những độ đo khác nhau. Các độ đo được chia thành bốn loại: (1) Độ đo dựa trên thay đổi tập dữ liệu thô; (2) Độ đo dựa trên mức độ bảo toàn CSDL; (3) Độ đo dựa trên những hiệu ứng lề; và (4) Độ đo dựa trên hiệu suất của thuật toán.

A. ĐỘ ĐO DỰA TRÊN SỰ THAY ĐỔI TẬP DỮ LIỆU THÔ

Nội dung chính của hầu hết các thuật toán khai thác luật kết hợp bảo toàn tính riêng tư là biến đổi CSDL ban đầu thành CSDL thanh lọc sao cho người khai thác không thể phát hiện các thông tin nhạy cảm. Do đó, chất lượng của CSDL thanh lọc là yếu tố cần thiết phải xem xét để đánh giá hiệu quả thuật toán. Một trong những chiến lược để kiểm tra chất lượng của CSDL thanh lọc là kiểm tra số lượng biến đổi đã thực hiện để tạo ra CSDL thanh lọc. Độ đo dựa trên thay đổi tập dữ liệu thô được chia ra hai cấp độ: cấp độ giao dịch và cấp độ item.

Đối với cấp độ giao dịch, hiệu quả của thuật toán được đánh giá dựa trên số lượng các giao dịch bị thay đổi để tạo ra CSDL mới.

Ở cấp độ item [19] [14], hiệu quả của thuật toán được đánh giá dựa trên sự khác biệt giữa tập dữ liệu ban đầu và tập dữ liệu thanh lọc (Công thức 1):

$$Diss\ (D, D') = \frac{\sum_{i=1}^{n} |f_D(i) - f_{D'}(i)|}{\sum_{i=1}^{n} f_D(i)} \quad (1)$$

Trong đó n là số lượng các item trong tập dữ liệu, $f_D(i)$ là tần suất của item *i* trong tập dữ liệu ban đầu, và $f_{D'}(i)$ là tần suất của item *i* trong tập dữ liệu thanh lọc.

B. ĐỘ ĐO DỰA TRÊN MỨC ĐỘ BẢO TOÀN CSDL

CSDL sau khi đã thanh lọc phải đảm bảo được tính riêng tư của dữ liệu. Với tập các dữ liệu nhạy cảm của người dùng cho trước, CSDL ban đầu phải được biến đổi để ẩn các dữ liệu nhạy cảm. Cụ thể là khi dữ liệu đã được thanh lọc, người dùng không thể phát hiện ra những dữ liệu nhạy cảm này khi áp dụng các phương pháp khai thác dữ liệu. Độ đo Hiding Failure (HF) [18] [19] [14] được đề xuất để đo hiệu quả của việc ẩn các luật nhạy cảm. Độ đo HF cho biết số lượng các luật nhạy cảm mà thuật toán thanh lọc không thể ẩn và vẫn đang được khai thác từ CSDL đã thanh lọc. HF được tính theo công thức (2):

$$HF = \frac{|R_s(D')|}{|R_s(D)|} \quad (2)$$

Trong đó, $R_s(D')$ là số lượng luật nhạy cảm tìm thấy trong CSDL thanh lọc D' và $R_s(D)$ là số lượng luật nhạy cảm trong CSDL ban đầu D. Khi quá trình thanh lọc kết thúc, tất cả các luật nhạy cảm được ẩn thì khi đó HF bằng 0. Hầu hết các thuật toán hiện có đều hướng đến việc ẩn tất cả các luật nhạy cảm. Tuy nhiên, việc ẩn đi các luật nhạy cảm có thể dẫn đến việc mất mát thông tin khi thuật toán thực hiện thao tác xóa các item khỏi CSDL ban đầu. Do đó, nhiều thuật toán hiện nay được nghiên cứu để cho phép thực hiện ẩn một số luật nhạy cảm có độ quan trọng nhất định và cho phép phát hiện một số luật nhạy cảm khác để đảm bảo tính cân bằng cho CSDL thanh lọc.

C. ĐỘ ĐO DỰA TRÊN HIỆU ỨNG LỀ

Để biến đổi CSDL ban đầu thành CSDL thanh lọc, một số thuật toán sử dụng phương pháp xóa đi một hoặc nhiều item có trong các giao dịch ban đầu nhằm giảm tần suất xuất hiện của các mẫu nhạy cảm. Tuy nhiên, thao tác xóa cũng sẽ giảm tần suất xuất hiện của một số mẫu không nhạy cảm khác, dẫn đến trường hợp mất đi thông tin mà người dùng muốn chia sẻ. Do đó, độ đo Lost Rules (LR) [18] [19] [14] được sử dụng để đánh giá mức độ mất mát thông tin của CSDL thanh lọc. Độ đo LR cho biết số lượng các luật không nhạy cảm bị mất do hoạt động thanh lọc và sẽ không còn được khai thác từ tập dữ liệu đã thanh lọc. LR được tính theo công thức (3):

$$LR = \frac{|{\sim}R_s(D)| - |{\sim}R_s(D')|}{|{\sim}R_s(D)|} \quad (3)$$

Trong đó $|{\sim}R_s(D)|$ là số lượng các luật không nhạy cảm trong tập dữ liệu ban đầu D và $|{\sim}R_s(D')|$ là số lượng các luật không nhạy cảm trong tập dữ liệu thanh lọc D'.

Bên cạnh thao tác biến đổi xóa, một số thuật toán sử dụng thao tác thay đổi các item có trong giao dịch ban đầu thành các item khác nhằm giảm tần suất xuất hiện của các mẫu nhạy cảm. Tuy nhiên, thao tác này có thể tạo ra các luật giả không có trong CSDL ban đầu, dẫn đến trường hợp thông tin được chia sẻ có thể bị sai lệch. Độ đo Ghost Rules (GR) cho biết số lượng các luật giả không có trong CSDL gốc ban đầu, được tạo ra do hoạt động thanh lọc và được khai thác từ CSDL thanh lọc. GR được tính theo công thức (4):

$$GR = \frac{|R'| - |R \cap R'|}{|R'|} \quad (4)$$

Trong đó, $|R'|$ là số lượng luật khai thác từ D' và $|R|$ là số lượng luật khai thác từ D.

D. ĐỘ ĐO DỰA TRÊN HIỆU SUẤT CỦA THUẬT TOÁN

Một số tiêu chí khác được dùng để đánh giá bao gồm: (1) **Số vòng lặp**: một trong những tiêu chí đánh giá quan trọng nhất trong các thuật toán là số lần lặp cần thiết để đạt được giải pháp tối ưu; (2) **Thời gian khai thác**:

được đo ở cả hai giai đoạn chuyển đổi dữ liệu ban đầu sang dữ liệu thanh lọc và thời gian khai thác dữ liệu thanh lọc để rút trích ra các mẫu nhạy cảm [16] [17] [19]; (3) **Tài nguyên**: không gian vùng nhớ được sử dụng trong suốt quá trình thực thi thuật toán; (4) **Thiết bị giao tiếp**: trong trường hợp thuật toán được cài đặt trên các hệ thống phân tán, cần có sự đánh giá về quá trình giao tiếp giữa các thiết bị trong hệ thống để đảm bảo thuật toán được thực hiện đúng và hiệu quả; và (5) **Khả năng mở rộng**: với sự phát triển của khoa học và công nghệ, dữ liệu được khai thác sẽ được mở rộng theo thời gian [19]. Do đó, cần phải đánh giá khả năng mở rộng của các thuật toán để có thể xử lý được khối lượng dữ liệu ngày càng tăng.

V. KẾT LUẬN

Bài báo trình bày một khảo sát về các phương pháp thực hiện và phương pháp đánh giá thuật toán ẩn luật kết hợp thông qua các độ đo trong bài toán khai thác dữ liệu bảo toàn tính riêng tư. Kể từ khi được giới thiệu lần đầu vào năm 2000, ẩn luật kết hợp đã được mở rộng nghiên cứu trong cộng đồng nghiên cứu khai thác dữ liệu, dẫn đến nhiều công trình nghiên cứu đáng kể trong những năm qua.

Thông qua các công trình nghiên cứu gần đây thì có thể thấy đa phần các công trình được đề xuất để ẩn luật kết hợp đều tiếp cận dựa trên heuristic. Yếu tố căn bản khiến heuristic hấp dẫn các nhà nghiên cứu là tính hiệu quả về mặt tính toán và khả năng sử dụng bộ nhớ; cho phép mở rộng thuật toán trong trường hợp các tập dữ liệu trở nên lớn hơn, từ đó nhanh chóng cho ra một lời giải tối ưu hoặc là giải pháp gần đúng. Tuy nhiên, phần lớn các thuật toán heuristic hoạt động bằng cách lấy quyết định cục bộ tốt nhất mà không đi đến các giải pháp ẩn tối ưu toàn cục. Do vậy, cách tiếp cận meta-heuristic đang là xu hướng: trước tiên các thông tin thỏa mãn các yêu cầu ràng buộc được tính toán, sau đó dựa trên các thông tin này để thực hiện heuristic.

VI. TÀI LIỆU THAM KHẢO

[1] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *ACM SIGMOD International Conference on Management of Data*, 2000.

[2] Y. Lindell and B. Pinkas, "Privacy preserving data mining," *Journal of Cryptology,* vol. 15, no. 3, p. 36–54, 2000.

[3] A. Evfimievski, R. Srikant, R. Agrawal and J. Gehrke, "Privacy preserving mining of association rules," *Information Systems,* vol. 29, no. 4, p. 343–364, 2004.

[4] J.-L. Lin and Y.-W. Cheng, "Privacy preserving itemset mining through noisy items," *Expert Systems with Applications,* vol. 36, p. 5711–5717, 2009.

[5] P. Samarati, "Protecting respondents' identities in microdata release.," *IEEE Transactions on Knowledge and Data Engineering,* vol. 13, no. 6, p. 1010–1027, 2001.

[6] S. Hajian, J. Domingo-Ferrer and O. Farr`as, "Generalization-based privacy preservation and discrimination prevention in data publishing and mining," *Data Mining and Knowledge Discovery,* vol. 28, p. 1158–1188, 2014.

[7] B. C. Fung, K. Wang, R. Chen and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Computing Surveys,* vol. 42, no. 4, p. 141–172, 2010.

[8] A. Gkoulalas-Divanis and V. S. Verykios, "Association rule hiding for data mining," *Springer Science & Business Media,* 2010.

[9] D. O'Leary, G. Piatetsky-Shapiro and W. J. Frawley, "Knowledge Discovery as a Threat to Database Security," *Knowledge discovery in databases. Menlo Park: AAAI/MIT Press,* p. 507–516, 1991.

[10] D. E. O'Leary, S. a. K. W. Bonorris, Y.-T. Khaw, H.-Y. Lee and W. Ziarko, "Some privacy issues in knowledge discovery: The OECD personal privacy guidelines," *IEEE Expert,* vol. 10, no. 2, pp. 48--59, 1995.

[11] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim and V. Verykios, "Disclosure limitation of sensitive rules.," in *The IEEE knowledge and data engineering exchange workshop*, 1999.

[12] R. T. I. Agrawal and A. Swami, "Mining association rules between sets of items in large databases," in *The ACM SIGMOD conference on management of data*, 1993.

[13] A. Telikani and A. Shahbahrami, "Data sanitization in association rule mining: An analytical review," *Expert Systems with Applications,* vol. 96, pp. 406-426, 2018.

[14] Stanley R. M. Oliveira and Osmar R. Zaıane, "Privacy Preserving Frequent Itemset Mining," in *Proceedings of the IEEE international conference on privacy, security and data mining (pp. 43-54)*, 2002.

[15] Y.-H. Wu, C.-M. Chiang and A. L. Chen, "Hiding Sensitive Association Rules with Limited Side Effects," *IEEE transactions on knowledge and data engineering,* vol. 19, no. 1, pp. 29-42, 2007.

[16] Elena Dasseni, Vassilios S. Verykios, Ahmed K. Elmagarmid3 and Elisa Bertino, "Hiding Association Rules by Using Confidence and Support," in *Proceedings of the 4th international workshop on information hiding (pp.369-383)*, 2001.

[17] Yucel Saygın, Vassilios S. Verykios and Chris Clifton, "Using Unknowns to Prevent Discovery of Association Rules," *ACM SIGMOD,* vol. 30, no. 4, p. pp 45–54, 2001.

[18] Stanley R. M. Oliveira and Osmar R. Zaıane, "Algorithms for Balancing Privacy and Knowledge Discovery in Association Rule Mining," in *Proceedings of the international database engineering and application symposium (pp. 54-63)*, 2003.

[19] Stanley R. M. Oliveira and Osmar R. Zaıane, "Protecting Sensitive Knowledge By Data Sanitization," in *Proceedings of the IEEE international conference on data mining (pp. 211-218)*, 2003.

[20] Emmanuel D. Pontikakis, Achilleas A. Tsitsonis and Vassilios S. Verykios, "An experimental study of distortion based techniques for association rule hiding," in *Proceedings of the ACM workshop on privacy in the electronic society*, 2004.

[21] S. Menon, S. Sarkar and S. Mukherjee, "Maximizing Accuracy of Shared Databases when Concealing Sensitive Patterns," *Information Systems Research,* pp. 256-270, 2005.

[22] X. Sun and P. S. Yu, "A border-based approach for hiding sensitive frequent itemsets," in *Proceedings of the Fifth IEEE International Conference on Data Mining*, 2005.

[23] Heikki Mannila and Hannu Toivonen, "Levelwise Search and Borders of Theories in Knowledge Discovery," *Data Mining and Knowledge Discovery,* 1997.

[24] Aris Gkoulalas-Divanis and Vassilios S. Verykios, "An integer programming approach for frequent itemset hiding," in *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management*, 2006.

[25] George V. Moustakides and Vassilios S. Verykios, "A Max-Min Approach for Hiding Frequent Itemsets," in *Sixth IEEE International Conference on Data Mining - Workshops*, 2006.

[26] Ali Amiri, "Dare to share: Protecting sensitive knowledge with data sanitization," *Decision Support Systems,* pp. 181-191, 2007.

[27] Yu-Chiang Li and Jieh-Shan Yeh, "MICF: An effective sanitization algorithm for hiding sensitive patterns on data mining," *Advanced Engineering Informatics,* vol. 21, no. 3, pp. 269-280, 2007.

[28] Shyue-LiangWang, Ayat Jafari and Bhavesh Parikh, "Hiding informative association rule sets," *Expert Systems with Applications,* vol. 33, no. 2, pp. pp 316-323, 2007.

[29] Shyue-LiangWang and A. Jafari, "Using unknowns for hiding sensitive predictive association rules," in *Proceedings of the IEEE International Conference on Information Reuse and Integration*, 2005.

[30] Shyue-Liang Wang, Dipen Patel, Ayat Jafari and Tzung-Pei Hong, "Hiding collaborative recommendation association rules," *Applied Intelligence,* vol. 27, no. 1, pp. 67-77, 2007.

[31] V. S. Verykios, E. D. Pontikakis, Y. Theodoridis and L. Chang, "Efficient algorithms for distortion and blocking techniques in association rule hiding," *Distributed and Parallel Databases,* vol. 22, p. 85–104, 2007.

[32] A. Gkoulalas-Divanis and V. S. Verykios, "Exact knowledge hiding through database extension," *IEEE Transactions on Knowledge and Data Engineering,* vol. 21, no. 5, pp. 699-713, 2009.

[33] Shyue-Liang Wang, Rajeev Maskey, Ayat Jafari and Tzung-Pei Hong, "Efficient sanitization of informative association rules," *Expert Systems with Applications,* vol. 35, no. (1-2), pp. 442-450, 2008.

[34] S. Menon and S. Sarkar, "Minimizing information loss and preserving privacy. Manage Science," *Manage Science,* vol. 53, pp. 101-116, 2008.

[35] Shyue-Liang Wang, "Maintenance of sanitizing informative association rules," *Expert Systems with Applications,* vol. 36, no. 2, pp. 4006-4012, 2009.

[36] Aris Gkoulalas-Divanis and Vassilios S. Verykios , "Hiding sensitive knowledge without side effects," *Knowledge and Information Systems ,* p. 263–299, 2009.

[37] G. Grätzer, Lattice Theory: Foundation, Springer link, 2011.

[38] Hai Quoc Le, Somjit Arch-int and Ngamnij Arch, "Association Rule Hiding Based on Intersection Lattice," in *Proceedings of the 4th International Conference on computer technology and development*, 2013.

[39] Tzung-Pei Hong, Chun-Wei Lin, Kuo-Tung Yang and Shyu, "Using TF-IDF to hide sensitive itemsets," *Applied Intelligence,* vol. 38, no. 4, pp. 502-510, 2013.

[40] Chun-Wei Lin, Binbin Zhang, Kuo-Tung Yang and Tzung-Pei Hong, "Efficiently Hiding Sensitive Itemsets with Transaction Deletion Based on Genetic Algorithms," *Scientific World,* 2014.

[41] Chun-Wei Lin, Tzung-Pei Hong, Jia-Wei Wong, Guo-Cheng Lan and Wen-Yang Lin, "A GA-Based Approach to Hide Sensitive High Utility Itemsets," *Scientific World,* 2014.

[42] Peng Cheng, Jeng-Shyang Pan and Chun-Wei Lin, "Privacy Preserving Association Rule Mining Using Binary Encoded NSGA-II," in *Proceedings of the 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2014.

[43] Stefan Bleuler, Marco Laumanns, Lothar Thiele and Eckart Zi, "PISA — A Platform and Programming Language Independent Interface for Search Algorithms," in *Proceedings of the International Conference on Evolutionary Multi-Criterion Optimization*, 2014.

[44] Chun-Wei Lin, Tzung-Pei Hong, Kuo-Tung Yang and Shyu, "The GA-based algorithms for optimizing hiding sensitive itemsets through transaction deletion," *Applied Intelligence,* vol. 42, p. 210–230, 2015.

[45] Jerry Chun-WeiLin, Qiankun Liu, Philippe Fournier-Vigerb, Tzung-PeiHong, Miroslav Voznak and Justin Zhan, "A sanitization approach for hiding sensitive itemsets based on particle swarm optimization," *Engineering Applications of Artificial Intelligence,* vol. 53, pp. 1-18, 2016.

[46] M. H. Afshari, M. N. Dehkordi and M. Akbari, "Association rule hiding using cuckoo optimization algorithm," *Expert Systems with Applications,* vol. 64, pp. 340-351, 2016.

[47] X.-S. Yang and S. Deb, "Cuckoo search via Lévy flights," in *Nature & biologically inspired computing*, 2009.

[48] Peng Cheng, John F. Roddick, Shu-Chuan Chu and Chun-Wei Lin , "Privacy preservation through a greedy, distortion-based rule-hiding method," *Applied Intelligence,* p. 295–306, 2016.

[49] Akbar Telikani and Asadollah Shahbahrami, "Optimizing association rule hiding using combination of border and heuristic approaches," *Applied Intelligence ,* vol. 47, p. 544–557, 2017.

[50] Behnam Talebi and Mohammad Naderi Dehkordi, "Sensitive Association Rules Hiding Using Electromagnetic Field Optimization Algorithm," *Expert Systems with Applications,* vol. 114, pp. 155-172, 2018.

[51] H. Abedinpourshotorban, S. M. Shamsuddin, Z. Beheshti and D. N. Jawawi, "Electromagnetic field optimization: A physics-inspired metaheuristic optimization algorithm," *Swarm and Evolutionary Computation,* vol. 26, pp. 8-22, 2016.

[52] Bac Le, Lien Kieu and Dat Tran, "Distortion-based heuristic method for sensitive association rule hiding," *Journal of Computer Science and Cybernetics,* vol. 35, p. 337–354, 2019.

[53] Shaoxin Li, Nankun Mu, Junqing Le and Xiaofeng Liao, "A novel algorithm for privacy preserving utility mining based on integer linear programming," *Engineering Applications of Artificial Intelligence,* pp. 300-312, 2019.

[54] A. Telikani, A. H. Gandomi, A. Shahbahrami and M. N. Dehkordi, "Privacy-preserving in association rule mining using an improved discrete binary artificial bee colony," *Expert Systems With Applications,* vol. 144, 2020.

[55] S. Jangra and D. Toshniwal, "Victim item deletion based PSO inspired sensitive pattern hiding algorithm for dense datasets," *Information Processing and Managemen,* vol. 57, no. 5, 2020.

[56] Bac Le, Dong Phuong Le and Minh Thai - Tran, "Hiding sensitive association rules using the optimal electromagnetic optimization method and a dynamic bit vector data structure," *Expert Systems With Applications ,* vol. 176, 2021.

A SURVEY OF HIDING ASSOCIATION RULE METHODS IN TRANSACTION DATASETS

Tran Minh Thai, Tran Anh Duy, Le Thi Minh Nguyen

ABSTRACT—Privacy-Preserving Data Mining (PPDM) is a new area of research in the data mining community and has been focused on for over a decade. PPDM studies the side effects of data mining methods that stem from intrusions into the privacy of individuals and organizations. Several approaches to solving this problem have been studied and applied. The proposed methods can be classified according to two main research directions: data hiding and knowledge hiding. Data hiding is a research direction on the privacy of raw data or information, which can be guaranteed during data mining. The methods of this group work on the data itself to hide sensitive information by different methods. Knowledge hiding refers to protecting the results of mining sensitive data instead of the raw data itself. It is the main application direction of data mining tools and algorithms. In which, association rule hidden is a research direction in knowledge hidden group. In this paper, we focus on presenting the problem related to hidden association rules. Besides, we investigate the association rule hiding techniques and compare the proposed methods to clarify the change of approach of the hiding rule methods. Finally, the experimental methods performed with the measures used to compare the efficiency of the algorithms are also presented in the paper.

TS. Trần Minh Thái tốt nghiệp cử nhân ngành Công nghệ Phần mềm vào năm 2001 và thạc sỹ Tin học vào năm 2006 tại trường Đại học Khoa học Tự nhiên TP. Hồ Chí Minh, nhận bằng tiến sỹ Khoa học Máy tính vào năm 2017 do Đại học Quốc gia TP. Hồ Chí Minh cấp. Anh ta từng là giảng viên và quản lý khoa Công nghệ Thông tin trường Cao đẳng Công nghệ Thông tin TP. Hồ Chí Minh từ năm 2002 đến 2015. Từ năm 2015 đến hiện tại, anh ta là giảng viên và là trưởng bộ môn Hệ thống Thông tin thuộc khoa Công nghệ Thông tin trường Đại học Ngoại ngữ Tin học TP. Hồ Chí Minh. Lĩnh vực nghiên cứu chính của anh ta liên quan đến vấn đề khai thác dữ liệu, ẩn dữ liệu, xử lý dữ liệu lớn và nhận dạng.

ThS. Lê Thị Minh Nguyện tốt nghiệp thạc sĩ Khoa học máy tính năm 2007 tại trường Đại Học Công Nghệ Thông Tin Tp.HCM. Từng là giảng viên tại trường Cao đẳng Công nghệ Thông Tin từ 2003-2015. Từ năm 2015 đến nay là giảng viên thuộc khoa Công nghệ Thông tin trường Đại học Ngoại ngữ Tin học Tp.HCM. Lĩnh vực nghiên cứu quan tâm là Khai thác dữ liệu.

ThS. Trần Anh Duy Nhận học vị thạc sĩ Khoa học máy tính trường Đại học Khoa Học Tự Nhiên năm 2017. Hiện là giảng viên khoa Công Nghệ Thông Tin trường Đại Học Ngoại Ngữ Tin Học thành phố Hồ Chí Minh. Lĩnh vực nghiên cứu đang quan tâm là: Khai thác dữ liệu.