Hệ thống chuyên gia có ứng dụng 96 (2018) 406–426

Danh sách nội dung có sẵn tại ScienceDirect

Hệ thống chuyên gia có ứng dụng

trang chủ tạp chí: www.elsevier.com/locate/eswa

Ôn tập

Làm sạch dữ liệu trong khai thác luật kết hợp: Đánh giá phân tích

Akbar Telikani, Asadollah Shahbahrami∗

Khoa Kỹ thuật Máy tính, Khoa Kỹ thuật, Đại học Guilan, Rasht, Iran

bài báo thông tin

Lịch sử bài viết: Nhận ngày 20 tháng 3 năm 2017 Sửa đổi ngày 21 tháng 10 năm 2017 Được chấp nhận ngày 22 tháng 10 năm 2017 Có sẵn trực tuyến ngày 24 tháng 10 năm 2017

Từ khóa: Bảo đảm quyền riêng tư trong khai thác dữ liệu Khai thác quy tắc hiệp hội Ẩn quy tắc hiệp hội

Vệ sinh dữ liệu

trừu tượng

Ẩn quy tắc kết hợp là quá trình chuyển đổi cơ sở dữ liệu giao dịch thành phiên bản được làm sạch để bảo vệ kiến thức và mẫu nhạy cảm.Thách thức là giảm thiểu các tác dụng phụ trên cơ sở dữ liệu đã được vệ sinh.Nhiều thuật toán khử trùng khác nhau đã được đề xuất để đạt được mục đích này.Bài viết này trình bày phân tích và phân loại có cấu trúc về những thách thức và hướng đi hiện có đối với các thuật toán dọn dẹp tiên tiến, trong đó nêu bật các đặc điểm của chúng.Năm mươi bốn thuật toán khoa học, chủ yếu trải dài trong giai đoạn 2001–2017, đã được phân tích và nghiên cứu theo bốn khía cạnh, bao gồm chiến lược ẩn giấu, kỹ thuật dọn dẹp, phương pháp dọn dẹp và phương pháp lựa chọn.Về kết quả và phát hiện, đánh giá này cho thấy (i) so với các khía cạnh khác của thuật toán dọn dẹp, phương pháp lựa chọn giao dịch và vật phẩm ảnh hưởng đáng kể hơn đến tính tối ưu của quá trình ẩn, (ii) kỹ thuật chặn làm tăng rủi ro tiết lộ trong khi kỹ thuật bóp méo tốt hơn trong lĩnh vực bảo vệ tri thức và kỹ thuật chèn/xóa giao dịch là một hướng mới, (iii) các thuật toán dựa trên heuristic đã thu hút được nhiều sự chú ý hơn các thuật toán khác, đặc biệt là trong bối cảnh ẩn các luật kết hợp, (iv) một thuật toán mới xu hướng là sử dụng mô hình tiến hóa để ẩn tri thức thường được tích hợp với kỹ thuật chèn/xóa giao dịch và (V) ẩn các quy tắc kết hợp đưa ra nhiều thách thức hơn là ẩn các tập phổ biến về mặt xác định chiến lược và công thức của phương pháp tuyển chọn.Nghiên cứu này nhằm mục đích giúp các nhà nghiên cứu và quản trị viên cơ sở dữ liệu tìm thấy những phát triển gần đây trong việc ẩn luật kết hợp.

1. Giới thiệu

Chia sẻ dữ liệu có thể dẫn đến rò rỉ thông tin nhạy cảm liên quan đến lợi thế cạnh tranh của doanh nghiệp hoặc quyền riêng tư của cá nhân (Verykios et al., 2004a).Các chính sách và hướng dẫn trong việc xuất bản dữ liệu không thể bảo vệ thông tin này vì chúng dựa vào việc xuất bản các loại dữ liệu cụ thể và dựa trên các thỏa thuận về việc sử dụng dữ liệu đã xuất bản (Fung và cộng sự, 2010).Do đó, trước khi phát hành dữ liệu, thông tin nhạy cảm sẽ được bảo vệ thông qua sửa đổi cơ sở dữ liệu (Divanis & Verykios, 2009b).Các phương pháp bảo vệ có thể được phân thành hai loại chính, (i) ẩn dữ liệu và (ii) ẩn tri thức.Các phương pháp ẩn dữ liệu sửa đổi dữ liệu thô nhạy cảm bằng cách sử dụng các kỹ thuật ngẫu nhiên hóa (Agrawal & Srikant, 20 0 0; Evfimievski và cộng sự, 2004; Lin & Liu, 2007; Rizvi & Haritsa, 2002; Lin & Cheng, 2009) hoặc sửa đổi các định danh gần như bằng cách sử dụng kỹ thuật ẩn danh để che giấu chủ sở hữu hồ sơ (Samarati, 2001; Sweeney, 2002; Hajian và cộng sự, 2014), bất kể loại phân tích được thực hiện bởi bên thứ ba (Prakash & Singaravel, 2015).Các thuộc tính gần như định danh là những thuộc tính không thể chỉ xác định chủ sở hữu bản ghi nhưng nếu chúng được kết hợp với nhau, có thể xác định rõ ràng thực thể như độ tuổi và mã zip (Fung và cộng sự, 2010; Hajian và cộng sự, 2014).Trong tài liệu, những phương pháp này được gọi là "Xuất bản dữ liệu bảo vệ quyền riêng tư (PPDP)" (Fung và cộng sự, 2010).

Các phương pháp ẩn giấu tri thức tập trung vào việc bảo vệ kết quả khai thác dữ liệu nhạy cảm (Divanis & Verykios, 2010).Danh mục này được ký hiệu là Khai thác dữ liệu bảo vệ quyền riêng tư (PPDM).Các mối đe dọa về quyền riêng tư do kết quả khai thác dữ liệu được đưa ra lần đầu tiên bởi O' Leary (1991, 1995).Sau đó, Clifton và Marks (1996) trình bày một số chiến lược che giấu dữ liệu để cấm suy luận và khám phá những kiến thức nhạy cảm.PPDM có thể được áp dụng trong các tác vụ khai thác dữ liệu khác nhau như khai thác quy tắc kết hợp, phân cụm và phân loại.Khai thác quy tắc kết hợp bảo vệ quyền riêng tư liên quan đến việc làm sạch dữ liệu dẫn đến tiết lộ kiến thức bí mật và riêng tư (Divanis & Verykios, 2010).Nó được gọi là ẩn quy tắc kết hợp/khử trùng dữ liệu.Trong phân cụm bảo vệ quyền riêng tư, trung tâm cụm được thay đổi bằng cách bóp méo các thuộc tính bí mật.Oliveira và Zaiane (2004) đã làm xáo trộn các thuộc tính số bí mật bằng cách sử dụng chuyển đổi dữ liệu hình học để đáp ứng việc bảo vệ quyền riêng tư trong phân tích phân cụm, đặc biệt là phân cụm dựa trên phân vùng và phân cấp.Oliveira và Za-iane (2010) cũng áp dụng phép chuyển đổi dựa trên vòng quay để bảo vệ tính riêng tư của thông tin một cách độc lập với bất kỳ cụm nào.

∗ Tác giả tương ứng.Địa chỉ email: shahbahrami@guilan.ac.ir (A. Shahbahrami).

https://doi.org/10.1016/j.eswa.2017.10.048 0957-4174/

làm phiền thuật toán. Sự riêng tư bảo quản phân loại cách tiếp cận

hạ thấp hiệu quả của các bộ phân loại sao cho các bộ phân loại không tiết lộ bất kỳ kiến thức nhạy cảm nào.Một số kỹ thuật PPDM được sử dụng trong các ứng dụng cây quyết định và quy tắc phân loại đã được thảo luận trong Chang và Moskowitz (1998) và Moskowitz và Chang (20 0 0).Tính toán an toàn của nhiều bên (Yao, 1982) là kỹ thuật dựa trên mật mã phổ biến nhất để bảo vệ quyền riêng tư trong khai thác dữ liệu phân tán (Lindell & Pinkas, 2009).Clifton và cộng sự.(2002) đã trình bày một số kỹ thuật như tổng an toàn, liên kết tập hợp an toàn, kích thước an toàn của giao tập hợp và tích vô hướng hữu ích cho nhiều nhiệm vụ khai thác dữ liệu.Trong những năm gần đây, thuật ngữ "PPDP" đôi khi đã phát triển để bao hàm nhiều vấn đề nghiên cứu về PPDM, mặc dù chúng không hoàn toàn giống nhau.

Việc ẩn luật kết hợp là một trong những lĩnh vực nghiên cứu chính trong PPDM được đề xuất lần đầu tiên bởi Atallah et al.(1999).Để minh họa sự cần thiết phải duy trì các quy tắc kết hợp nhạy cảm, một hồ sơ được công bố công khai về các giao dịch của một hiệu sách tiết lộ rằng những người đã mua một cuốn sách có tựa đề "Romeo và Juliet" vào tháng trước cũng đã mua một cuốn sách có tựa đề "Free Alaska".Alice gặp Bob đang đọc "Romeo và Juliet" và được biết rằng anh ấy đã mua nó vào tháng trước.Vì vậy, cô ấy đưa ra một suy luận vi phạm quyền riêng tư của Bob liên quan đến các thuyết phục chính trị của anh ấy (Cao và cộng sự, 2010).Các ví dụ thúc đẩy tương tự cho việc ẩn quy tắc kết hợp được thảo luận trong Clifton và Marks (1996), Divanis và Verykios, (2010), Oliveira và Zaiane (2002), và Sun và Yu (2007).Quá trình ẩn quy tắc kết hợp sẽ lọc sạch các giao dịch để giảm độ tin cậy/hỗ trợ của các mẫu nhạy cảm xuống dưới ngưỡng được xác định trước (Divanis & Verykios, 2010).Quá trình này tạo ra một số tác dụng phụ trên cơ sở dữ liệu đã được làm sạch để một số mẫu không nhạy cảm bị mất hoặc các mẫu mới có thể được đưa vào.Một giải pháp vệ sinh che giấu mọi kiến thức nhạy cảm và cũng không gây ra tác dụng phụ được gọi là "giải pháp tối ưu".Điều quan trọng cần lưu ý là vấn đề tìm kiếm phương pháp làm sạch dữ liệu tối ưu là một vấn đề NP-khó (Atallah và cộng sự, 1999).Trong hai thập kỷ qua, một số thuật toán đã được đề xuất để tìm ra giải pháp che giấu thông tin nhạy cảm.

Ẩn quy tắc kết hợp là chủ đề của một số bài khảo sát và đánh giá, cũng như sách, trong đó mục tiêu là thu thập và phân loại các thuật toán ẩn quy tắc kết hợp.Verykios và Divanis (2008) và Verykios (2013) đã tổng quan ngắn gọn về việc ẩn luật kết hợp và trình bày phân loại của các mẫu thuật toán quan trọng.Divanis và Verykios (2010) đã cung cấp một mô tả sâu rộng về các đặc điểm chính của thuật toán dọn dẹp.Tất cả các khảo sát hiện tại đều tập trung vào việc cung cấp phân loại các thuật toán ẩn quy tắc kết hợp dựa trên cách tiếp cận cơ bản được áp dụng bởi mỗi thuật toán.Họ đã phân loại các thuật toán được đề xuất từ năm 2001 đến năm 2009 thành ba loại dựa trên cách tiếp cận, bao gồm heuristic, border và formal.Ngoài ra, theo hiểu biết tốt nhất của chúng tôi, chưa có công trình nghiên cứu nào được thực hiện để phân tích các yếu tố góp phần tạo nên tiện ích của quá trình dọn dẹp, trong khi nghiên cứu về tác động và tầm quan trọng của các yếu tố này trong quá trình dọn dẹp dữ liệu là một vấn đề rất quan trọng đối với cả hai. nhà nghiên cứu và quản trị viên cơ sở dữ liệu.

Trong khảo sát này, chúng tôi cố gắng cung cấp một đánh giá mang tính phân tích về các hướng chính trong việc ẩn giấu luật kết hợp và trình bày những hiểu biết sâu sắc của chúng tôi về chủ đề này.Không giống như các khảo sát khác mô tả các thuật toán theo cách phân loại dựa trên cách tiếp cận, nghiên cứu của chúng tôi không có ý định cung cấp mô tả chi tiết về các thuật toán ẩn quy tắc kết hợp vì đã tồn tại một số khảo sát phù hợp.Mục tiêu chính là giới thiệu các khái niệm và xu hướng mới về các quan điểm khác nhau của quá trình vệ sinh dữ liệu.

Bài viết có ba mục tiêu: thứ nhất là xem xét các thuật toán dọn dẹp cập nhật để cung cấp điểm tham chiếu cho các nhà nghiên cứu và quản trị viên cơ sở dữ liệu.Thứ hai là phân tích các khía cạnh khác nhau của việc ẩn luật kết hợp với thảo luận rộng rãi về các tính năng, ưu điểm và nhược điểm nổi bật của chúng.

Mục tiêu thứ ba là trình bày các hướng và xu hướng ẩn quy tắc kết hợp bằng cách giới thiệu đánh giá thống kê về việc áp dụng các khía cạnh làm sạch trong các thuật toán đã xác định.

Phần còn lại của bài báo được tổ chức như sau: Phần 2 trình bày cách trình bày vấn đề và các ký hiệu được sử dụng trong nghiên cứu này.Phần 3 mô tả quy trình ẩn quy tắc kết hợp và Phần 4 giới thiệu các tác dụng phụ của nó bằng một ví dụ minh họa.Phần 5 mô tả các thuật toán dọn dẹp dữ liệu được thu thập từ năm 2001 đến năm 2017 theo thứ tự thời gian.Phần 6 phân tích các hướng hiện tại trong việc dọn dẹp dữ liệu để che giấu kiến thức nhạy cảm.Trong Phần 7, phần thảo luận và phân tích thống kê về các thuật toán được thu thập sẽ được cung cấp, đồng thời nêu bật những lợi ích và hạn chế chính.Phần 8 trình bày các thước đo và bộ dữ liệu tiêu chuẩn để đánh giá hiệu suất của các thuật toán dọn dẹp.Cuối cùng, Phần 9 đưa ra kết luận và hướng nghiên cứu tiếp theo.

2. Tuyên bố vấn đề

Khai thác luật kết hợp là một trong những kỹ thuật khai thác dữ liệu quan trọng nhất được giới thiệu lần đầu tiên bởi Agrawal et al. (1993).Các thuật toán khai thác luật kết hợp được phân thành hai loại: cấp độ và tăng trưởng mẫu.Các thuật toán Eclat (Zaki, 20 0 0) và Apriori (Agrawal & Srikant, 1994) đã được thiết kế để khai thác các luật kết hợp theo cách khôn ngoan hơn.Apriori sử dụng tìm kiếm theo chiều rộng để tính mức độ hỗ trợ của các tập mục trong khi Eclat sử dụng tìm kiếm theo chiều sâu bằng cách sử dụng tập hợp giao nhau.Giống như thuật toán Eclat, thuật toán tăng trưởng FP (Han và cộng sự, 20 0 0) thực hiện tìm kiếm theo chiều sâu mà không cần tạo ứng cử viên bằng cách áp dụng phương pháp "phân chia và chinh phục" (Sohrabi & Roshani, 2017).Thay vì tính độ hỗ trợ của tập ứng cử viên bằng cách sử dụng cách tiếp cận dựa trên giao lộ, nó sử dụng kỹ thuật cây mẫu thường xuyên (Han và cộng sự, 20 0 0) để lưu trữ tất cả các giao dịch của cơ sở dữ liệu trong cấu trúc dựa trên trie (Hạn và cộng sự. , 2004).Các khái niệm cơ bản của khai phá luật kết hợp được định nghĩa như sau:
Cho I = {i1 , i2 ,...,in } là tập n mục riêng biệt trong cơ sở dữ liệu (D), D = (t1 , t2 ,..., tm) trên I là tập hữu hạn các giao dịch.Mỗi giao dịch t là một tập các mục trong I, sao cho t ⊆ I. Một luật kết hợp được biểu diễn dưới dạng X ⇒Y, sao cho X ⊆ I, Y ⊆ I và X∩Y = ∅.Để khám phá một luật kết hợp, trước tiên, tập mục tạo ra nó được trích xuất dựa trên tiêu chí hỗ trợ, ký hiệu là α.Khi đó, quy tắc được rút ra từ tập mục được tạo dựa trên tiêu chí độ tin cậy, ký hiệu là β.Độ hỗ trợ của quy tắc X ⇒Y là tỷ lệ phần trăm giao dịch chứa cả X và Y, được tính theo phương trình dưới đây

$$\alpha (X \Rightarrow Y) = |X \cup Y |t\hat{o}i \quad (1)$$

Trong phương trình này, m là số lượng giao dịch trong D và |X∪Y|là số lượng giao dịch bao gồm cả tập mục X và Y. Nếu độ hỗ trợ của quy tắc vượt quá ngưỡng hỗ trợ tối thiểu, được ký hiệu là αmin, thì tập mục tạo ra nó được gọi là tập mục thường xuyên (Zeng và cộng sự, 2015).Độ tin cậy của quy tắc là tỷ lệ các giao dịch chứa X cũng chứa Y độ tin cậy được tính theo phương trình dưới đây.

$$\beta (X \Rightarrow Y) = |X \cup Y ||X | \quad (2)$$

Ở đâu |X|là số lượng giao dịch bao gồm tập mục X. Một quy tắc mạnh khi độ tin cậy của nó vượt quá ngưỡng tin cậy tối thiểu, do βmin tặng (Hai và cộng sự, 2013a).
Dựa trên thuộc tính khai thác quy tắc kết hợp, một quy tắc nhạy cảm sẽ tiết lộ tính riêng tư khi độ hỗ trợ của nó lớn hơn αmin hoặc độ tin cậy của nó cao hơn βmin.Do đó, để ẩn một quy tắc nhạy cảm, cần phải giảm độ hỗ trợ hoặc độ tin cậy của nó xuống dưới ngưỡng tối thiểu để quy tắc đó không thể bị phát hiện từ cơ sở dữ liệu đã được làm sạch.Tóm lại, việc ẩn luật kết hợp có thể được thực hiện

Hình 1. Khung chung cho quá trình ẩn luật kết hợp.

được nêu như sau: cho một cơ sở dữ liệu giao dịch, một tập hợp các mẫu có ý nghĩa được khai thác từ cơ sở dữ liệu gốc và một tập hợp con các mẫu nhạy cảm có trong các mẫu được khai thác.Chúng tôi muốn chuyển đổi cơ sở dữ liệu thành cơ sở dữ liệu sạch sẽ theo cách mà tất cả các mẫu nhạy cảm đều bị ẩn, trong khi các mẫu không nhạy cảm vẫn có thể được khai thác.Trên thực tế, việc ẩn luật kết hợp hoạt động theo cách khác với việc khai thác luật kết hợp.Mục tiêu của việc ẩn quy tắc kết hợp là làm giảm tầm quan trọng của các mẫu nhạy cảm đến mức chúng trở nên không còn thú vị theo quan điểm của các thuật toán khai thác dữ liệu, trong khi mục tiêu của việc khai thác quy tắc kết hợp là trích xuất các mẫu chưa biết và thú vị từ cơ sở dữ liệu giao dịch.

3. Quá trình ẩn luật kết hợp

Trong quá trình ẩn luật kết hợp, ngưỡng hỗ trợ và tin cậy được coi là mức độ nhạy cảm.Nếu độ hỗ trợ/độ tin cậy của một quy tắc mạnh và thường xuyên vượt quá mức độ nhạy cảm nhất định thì quy trình ẩn nên được áp dụng theo cách làm giảm tần suất hoặc độ mạnh của quy tắc.Quá trình này bao gồm bốn bước bao gồm trích xuất mẫu, đặc tả, làm sạch và đánh giá, được mô tả trong Hình 1.

Bước 1: Trích xuất mẫu, một tập mục phổ biến hoặc quy tắc kết hợp được khai thác từ cơ sở dữ liệu gốc bằng cách sử dụng thuật toán khai thác quy tắc kết hợp.

Bước 2: Thông số kỹ thuật, một số mẫu hoặc mục vi phạm quyền riêng tư được người dùng chỉ định là nhạy cảm.Bước này khác với khuyến nghị hợp tác (Lin và cộng sự, 2002; O'Mahony và cộng sự, 2004) và quy tắc kết hợp dự đoán (Li et al., 2001).Các quy tắc kết hợp dự đoán nhạy cảm và quy tắc đề xuất cộng tác là các quy tắc chứa các mục nhạy cảm tương ứng ở phía bên trái và phía bên phải của quy tắc.Đối với các quy tắc như vậy, các mục nhạy cảm được xác định mà không cần khai thác trước và lựa chọn các quy tắc ẩn.Do đó, quá trình ẩn được tích hợp vào quá trình tiền xử lý để tìm ra các quy tắc ẩn này (Wang và cộng sự, 2007a, 2007b, 2008; Wang & Jafari, 2005; Wang, 2009).

Bước 3: Dọn dẹp, trong bước này, cơ sở dữ liệu được dọn dẹp bằng thuật toán dọn dẹp để ẩn các mẫu nhạy cảm.Áp dụng thuật toán tối ưu giúp giảm tác dụng phụ trên cơ sở dữ liệu đã được vệ sinh.Điều này phụ thuộc chủ yếu vào loại mẫu.Không thể ẩn tập mục thường xuyên bằng thuật toán ẩn quy tắc trong khi quy tắc kết hợp có thể được ẩn bằng thuật toán ẩn tập mục bằng cách giảm độ hỗ trợ của nó hoặc bằng cách sử dụng thuật toán ẩn quy tắc bằng cách giảm độ tin cậy của nó.Do đó, bước này cố gắng tìm hiểu các điều kiện mà thuật toán dọn dẹp dữ liệu hữu ích nhất cho việc ẩn giấu kiến thức.

Bước 4: Đánh giá, tác dụng phụ của quá trình khử trùng được đo lường theo các mẫu nhạy cảm và không nhạy cảm được chỉ định ở bước thứ hai.Với mục đích này, việc khai thác luật kết hợp với các ngưỡng tối thiểu trước đó được áp dụng trên cơ sở dữ liệu đã được vệ sinh để xác nhận tiện ích và mức độ bảo vệ của cơ sở dữ liệu đã được vệ sinh.Khi các mục tiêu của quản trị viên cơ sở dữ liệu hoặc chủ sở hữu dữ liệu được hoàn thành, cơ sở dữ liệu đã được vệ sinh sẽ được phát hành cho người khác;nếu không, quá trình khử trùng sẽ được thực hiện lại bằng các tham số khác hoặc sử dụng thuật toán khác.

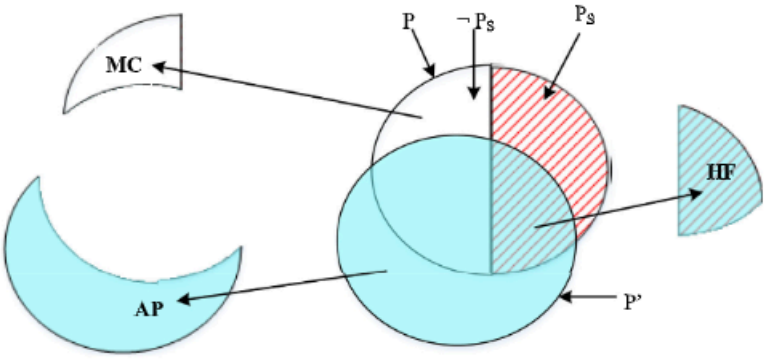
4. Tác dụng phụ của việc ẩn luật kết hợp

Rõ ràng, quy trình ẩn quy tắc kết hợp không thể kiểm soát được các tác dụng phụ của việc ẩn kiến thức nhạy cảm và tiện ích dữ liệu của dữ liệu đã được lọc sẽ bị giảm.Lỗi ẩn (HF), Chi phí bỏ lỡ (MC) và Mẫu giả tạo (AP) là ba tác dụng phụ được tạo ra trên cơ sở dữ liệu đã được vệ sinh.Hình 2 minh họa tác động của quá trình dọn dẹp dữ liệu trên cơ sở dữ liệu đã phát hành.

Tập P là các mẫu có ý nghĩa trong D, các tập PS và ¥PS lần lượt là các mẫu nhạy cảm và không nhạy cảm trong D.Tập P′ là các mẫu được phát hiện từ cơ sở dữ liệu đã được làm sạch (D′).Thuật toán dọn dẹp tối ưu sẽ khám phá tất cả các mẫu trong âPS từ D′ , điều này có nghĩa là P′ = âPS .Nhưng tập P′ có thể không
chứa tất cả các mẫu không nhạy cảm cũng như có thể chứa các mẫu nhạy cảm hoặc các mẫu được tạo mới.Trong Hình 2, MC có nghĩa là một số mẫu không nhạy cảm bị ẩn khỏi cơ sở dữ liệu đã phát hành (mất quy tắc/chi phí bỏ lỡ) trong khi HF có nghĩa là một số mẫu nhạy cảm được phát hiện từ cơ sở dữ liệu đã được làm sạch.Trong AP, một số mẫu nhân tạo được tạo ra trong D′ là kết quả của quá trình ẩn (quy tắc ma/quy tắc mới) (Oliveira & Zaiane, 2006; Verykios et al., 2004b).Tóm lại, thuật toán dọn dẹp có ba mục tiêu.Đầu tiên là ẩn tất cả các mẫu được chỉ định trong PS (HF = ∅).Thứ hai là duy trì tất cả các mẫu trong иPS (MC = ∅) và thứ ba là không nên tạo ra các mẫu giả (AP = ∅).HF phát sinh khi thuật toán dọn dẹp dữ liệu không thực hiện đủ các sửa đổi để ẩn tất cả các mẫu nhạy cảm.Một quy trình khử trùng có ngưỡng hỗ trợ tối thiểu thấp không thể tránh khỏi thất bại che giấu (Li & Chang, 2007).MC xảy ra khi một số giao dịch hỗ trợ toàn bộ hoặc một phần các mẫu không nhạy cảm bị sửa đổi hoặc xóa;do đó, độ hỗ trợ hoặc độ tin cậy của các mẫu này có thể bị giảm và không thể trích xuất được từ cơ sở dữ liệu đã được làm sạch.Điều đáng lưu ý là có sự thỏa hiệp giữa thất bại trong việc che giấu và chi phí bỏ lỡ.Chúng ta càng che giấu những khuôn mẫu nhạy cảm thì chúng ta càng bỏ sót những khuôn mẫu chính đáng (Oliveira & Zaiane, 2002).AP xảy ra khi độ tin cậy của một quy tắc không mạnh hoặc số lượng hỗ trợ của một tập mục không thường xuyên tăng lên (Oliveira & Zaiane, 2006).

Sau đây, các tác dụng phụ của quá trình khử trùng được minh họa bằng một ví dụ.Bảng 1(a) hiển thị một cơ sở dữ liệu nhất định.Xét αmin = 3 và βmin = 75%, các tập phổ biến và luật kết hợp mạnh lần lượt được liệt kê trong Bảng 1(b) và Bảng 1(c).

Giả sử rằng quy tắc {a} ⇒{b, c} là nhạy cảm.Vì độ tin cậy của nó là 80% (45) và βmin = 75% nên cần phải sửa đổi một lần



Hình 2. Mối liên hệ giữa các tác dụng phụ do thuật toán dọn dẹp dữ liệu tạo ra.

Bảng 1 Một ví dụ về quá trình khai thác luật kết hợp.

(a) Dữ liệu giao dịch		(b) Tập mục thường xuyên	(c) Luật kết hợp mạnh			
NHẬN DẠNG	Mặt hàng	Tập mục thường xuyên/α	AR	β (%)	AR	β (%)
T1	{A B C D}	{a}/5	{a} ⇒{b}	100	{a, d} ⇒{c}	75
T2	{b, d, e}	{b}/7	{a} ⇒{c}	80	{c, d} ⇒{a}	75
T3	{a, b, c}	{c}/5	{c} ⇒{a}	80	{b, c} ⇒{d}	75
T4	{a,b,d,e}	{d}/7	{a} ⇒{d}	80	{c, d} ⇒{b}	75
T5	{A B C D}	{e 4	{c} ⇒{b}	80	{e} ⇒{b, d}	100
T6	{b, d, e}	{a, b}/5	{b} ⇒{d}	85	{b, e} ⇒{d}	100
T7	{a,b,c,d,e}	{a, c}/4	{d} ⇒{b}	85	{d, e} ⇒{b}	100
Q8	{đĩa CD}	{a, d}/4	{e} ⇒{b}	100	{a, c} ⇒{b, d}	75
		{b, c}/4	{c} ⇒{d}	80	{a, d} ⇒{b, c}	75
		{b, d}/6	{e} ⇒{d}	100	{b, c} ⇒{a, d}	75
		{b, e}/4	{a} ⇒{b, c}	80	{c, d} ⇒{a, b}	75
		{c, d}/4	{c} ⇒{a, b}	80	{a, b, c} ⇒{d}	75
		{d, e}/4	{a, b} ⇒{c}	80	{a, b, d} ⇒{c}	75
		{a,b,c}/4	{a, c} ⇒{b}	100	{b, c, d} ⇒{a}	100
		{a,b,d}/4	{b, c} ⇒{a}	100		
		{a,c,d}/3	{a} ⇒{b, d}	80		
		{b, c, d}/3	{a, b} ⇒{d}	80		
		{b, d, e}/4	{a, d} ⇒{b}	100		
		{a,b,c,d}/3	{a, c} ⇒{d}	75		

Bảng 2 Tác động của các sửa đổi khác nhau đối với việc ẩn {a} ⇒{b, c}.

sửa đổi	#MC (bộ vật phẩm)	#MC (quy tắc)	#AP
{b}—→T1	2	14	1
{b}—→T3	0	3	1
{b}—→T5	2	14	1
{b}—→T7	2	14	0
{c}—→T1	3	15	0
{c}—→T3	0	3	0
{c}—→T5	3	15	0
{c}—→T7	3	15	0

để ẩn quy tắc này, trong đó mục {b} hoặc {c} bị xóa khỏi một trong các giao dịch hỗ trợ (T1, T3, T5 hoặc T7).Do đó, độ tin cậy giảm xuống còn 60% (35) sau lần sửa đổi này.Bảng 2 cho thấy các tác dụng phụ do áp dụng tất cả tám sửa đổi trong Bảng 1(a).Việc xóa mục {c} khỏi giao dịch T3 tạo ra ít tác dụng phụ nhất và chỉ mất ba quy tắc không nhạy cảm.Do đó, việc phát triển một thuật toán thực hiện những sửa đổi như vậy có thể làm giảm số lượng tác dụng phụ.

5. Thuật toán dọn dẹp dữ liệu

Cấu trúc của thuật toán dọn dẹp bao gồm ba phần, bao gồm chiến lược ẩn, kỹ thuật dọn dẹp và chọn lọc.

phương pháp tạo/tạo.Trọng tâm của chiến lược ẩn là làm thế nào để che giấu các mẫu nhạy cảm, trong khi trọng tâm của việc dọn dẹp kỹ thuật là làm thế nào để vệ sinh các giao dịch để thực hiện chiến lược ẩn.Chiến lược ẩn có được thông qua việc sửa đổi hoặc xóa các giao dịch nhạy cảm hiện có cũng như thông qua việc chèn các giao dịch mới vào cơ sở dữ liệu.Những sửa đổi này tạo ra các tác dụng phụ trên cơ sở dữ liệu đã được làm sạch cần được giảm thiểu.Do đó, việc xác định phương pháp lựa chọn có thể giảm thiểu tác dụng phụ bằng cách chọn một nhóm giao dịch và hạng mục thích hợp là vấn đề then chốt.Thuật toán dọn dẹp chấp nhận tập dữ liệu gốc, ngưỡng tối thiểu, mẫu nhạy cảm cũng như mẫu không nhạy cảm làm đầu vào.Sau đó, nó thực hiện các bước dọn dẹp dựa trên chiến lược ẩn, kỹ thuật dọn dẹp và các phương pháp lựa chọn được xác định cho việc dọn dẹp để tạo ra cơ sở dữ liệu được dọn dẹp làm đầu ra.

Chúng tôi đã xác định hai tiêu chí để chọn các bài báo đã xuất bản trong phạm vi ẩn quy tắc kết hợp, (1) cơ sở dữ liệu “Chỉ số trích dẫn khoa học”: Web of Science bao gồm các tạp chí có tác động cao nhất trong khoa học, kỹ thuật và nhân văn;do đó, chúng tôi đã sử dụng cơ sở dữ liệu này để xác thực các tạp chí.(2) Trích dẫn: tiêu chí này được sử dụng cho các tài liệu đã được trích dẫn trong các bài báo được xác nhận theo tiêu chí thứ nhất.Chúng tôi đã sử dụng tiêu chí này cho tài liệu không phải tạp chí, chẳng hạn như sách, hội nghị, hội thảo, v.v., cũng như cho các bài báo cung cấp giải pháp vệ sinh dữ liệu nhưng không có trong danh sách chính.Từ 109 tài liệu được chọn cho nghiên cứu này, 42 bài báo đã trình bày các thuật toán dọn dẹp để ẩn giấu tri thức.Số lượng các thuật toán này là 54 được sử dụng để phân tích của chúng tôi.Các thuật toán này được trình bày trong Bảng 3 với các chữ viết tắt của chúng.Sau đây, chúng tôi cung cấp

Bảng 3 Thuật toán dọn dẹp dữ liệu để ẩn quy tắc kết hợp.

Thẩm quyền giải quyết	Thuật toán
Saygin và cộng sự, 2001, 2002	Giảm độ tin cậy (CR), CR2, Tạo ẩn tập mục (GIH)
Dasseni và cộng sự, 2001;Verykios và cộng sự, 2004b	1.a, 1.b, 2.a
Oliveira và Zaiane, 2002	Thuật toán mục tần suất tối đa (MaxFIA), Thuật toán mục tần suất tối thiểu (MinFIA), Mục Thuật toán nhóm (IGA), Naïve
Oliveira và Zaiane, 2003a, 2006	Thuật toán kích thước cửa sổ trượt (SWA)
Oliveira và Zaiane, 2003b	Thuật toán ngẫu nhiên (RA), Thuật toán Round Robin (RRA)
Pontikakis và cộng sự, 2004a;Verykios và cộng sự, 2007	Thuật toán biến dạng dựa trên mức độ ưu tiên (PDA), Thuật toán biến dạng sắp xếp dựa trên trọng số (WSDA)
Verykios và cộng sự, 2004b	2.b
Pontikakis và cộng sự, 2004b;Verykios và cộng sự, 2007	Thuật toán chặn (BA)
Menon và cộng sự, 2005	Chăn, Trí Tuệ
Wang và Jafari, 2005;Wang và cộng sự, 2007a	Tăng hỗ trợ bên tay trái (ISL), Giảm hỗ trợ bên tay phải (DSR)
Tôn và Yu, 2005, 2007	Phương pháp tiếp cận dựa trên biên giới (BBA)
Divanis và Verykios, 2006	Nội tuyến
Moustakides & Verykios, 2006, 2008	Max-Min1, Max-Min2
Amiri, 2007	Tổng hợp, tách rời, lai
Lý và Chang, 2007	Xung đột mục tối đa đầu tiên (MICF)
Wang và cộng sự, 2007b	Giảm niềm tin bằng cách giảm hỗ trợ (DCDS), giảm niềm tin bằng cách tăng hỗ trợ (DCIS)
Wu và cộng sự, 2007	–
Wang và cộng sự, 2008	Giảm hỗ trợ và niềm tin (DSC)
Menon và Sarkar, 2008	–
Divanis và Verykios, 2009a	Hỗn hợp
Vương, 2009	Duy trì quy tắc kết hợp thông tin (MSI)
Divanis và Verykios, 2009b	–
Hải và Somjit, 2012	Ẩn quy tắc kết hợp dựa trên mạng giao nhau (ILARH)
Hải và cộng sự, 2012	Dựa trên mạng khoảng cách và giao điểm (DIL)
Hải và cộng sự, 2013a	Ẩn quy tắc kết hợp dựa trên lưới giao nhau (ARHIL)
Hải và cộng sự, 2013b	Heuristic để giảm độ tin cậy và hỗ trợ dựa trên mạng giao nhau (HCSRIL)
Hồng và cộng sự, 2013	Các mục nhạy cảm Tần số cơ sở dữ liệu nghịch đảo tần số (SIF-IDF)
Lin và cộng sự, 2014a	Thuật toán dựa trên GA thu nhỏ trước để xóa giao dịch (cpGA2DT)
Lin và cộng sự, 2014b	–
Lin và cộng sự, 2015	Thuật toán di truyền đơn giản để xóa giao dịch (sGA2DT), Thuật toán di truyền tiền lớn để xóa giao dịch (pGA2DT)
Lin và cộng sự, 2016	Thuật toán dựa trên Tối ưu hóa nhóm hạt để xóa giao dịch (PSO2DT)
Cheng và cộng sự, 2014, 2016a	Ẩn quy tắc cơ sở tối ưu hóa đa mục tiêu tiến hóa (EMO-RH)
Afshari và cộng sự, 2016	Thuật toán tối ưu hóa Cuckoo để ẩn quy tắc kết hợp (COA4ARH)
Cheng và cộng sự, 2016b	Sắp xếp mức độ liên quan
Telikani và Shahbahrami, 2017	Giảm độ tin cậy của các quy tắc (DCR)

đánh giá lịch sử của 54 thuật toán được thực hiện dựa trên việc ẩn quy tắc kết hợp cũng như trình bày một số mã giả đang phát triển cho hầu hết các thuật toán vệ sinh dữ liệu phổ biến.

2001. Dasseni và cộng sự.(2001) đã trình bày ba thuật toán là 1.a, 1.b và 2.a để ẩn các quy tắc nhạy cảm.Hai thuật toán đầu tiên làm giảm độ tin cậy của một quy tắc bằng cách tăng độ hỗ trợ của quy tắc tiền đề và bằng cách giảm độ hỗ trợ của quy tắc hệ quả tương ứng, trong khi thuật toán thứ ba giảm sự hỗ trợ của việc tạo tập mục của quy tắc.Saygin và cộng sự.(2001) đề xuất các thuật toán Giảm độ tin cậy (CR), CR2 và Tạo ẩn tập mục (GIH) được thực hiện tương tự như ba thuật toán trước đó;sự khác biệt là các thuật toán đề xuất thay thế các mục bằng ẩn số (dấu chấm hỏi) thay vì loại bỏ các mục.Mã giả của thuật toán 2.a được mô tả trong Hình 3(a).Thuật toán 2.a ước tính số lượng giao dịch tối thiểu cần được sửa đổi giữa N_iter_conf và N_iter_supp.Trong đó N_iter_conf và N_iter_supp là số lần lặp để ẩn một quy tắc nhạy cảm bằng các chiến lược giảm độ tin cậy và giảm hỗ trợ tương ứng, đồng thời thuật toán 2.a xem xét số lần lặp tối thiểu trong số đó để giảm thiểu các giao dịch được vệ sinh.

Algorithm: 2.a

Input: D, a set Ps of rules to hide, |D|, α_{min}, β_{min}

Output: D'

1. While sensitive rules not hidden {
2. Find corresponding transactions
3. Compute length of sensitive transactions
4. Sort sensitive transactions in ascending order of length
5. $N\_iter\_conf = \left\lceil |D| * \left(\frac{supp\ (X \Rightarrow Y)}{\beta_{min}} - supp(X)\right)\right\rceil$
6. $N\_iter\_supp = \left\lceil |D| * \left(\frac{supp\ (X \Rightarrow Y)}{\alpha_{min}}\right)\right\rceil$
7. $N\_iteration = min\ (N\_iter\_conf, N\_iter\_supp)$
8. While $N\_iteration > 0$ {
9. Select item with minimum impact on ($|X \Rightarrow Y|$-1)-itemsets
10. Remove victim item from a transaction
11. Supp (X⇒Y) = supp (X⇒Y)-1
12. Conf (X⇒Y) = supp (X⇒Y)/supp (X)
13. Remove sanitized transactions from sensitive list}
14. Remove hidden rule from sensitive rules}

Hình 3. (a): Mã giả của Thuật toán 2.a.(b): Mã giả của thuật toán Kích thước cửa sổ trượt (SWA).

2002. Oliveira và Zaiane (2002) đã đề xuất bốn thuật toán ẩn tập mục, đó là Thuật toán mục tần số tối đa (MaxFIA), Thuật toán mục tần suất tối thiểu (MinFIA), Thuật toán nhóm mục (IGA) và Naïve.Lần đầu tiên, họ xem xét tác động của việc sửa đổi giao dịch và mục trên cơ sở dữ liệu đã được làm sạch bằng cách tính toán xung đột của chúng.Mã giả của thuật toán IGA được hiển thị trong Hình 4.

2003. Thuật toán kích thước cửa sổ trượt (SWA) được đề xuất để ẩn các tập mục nhạy cảm trong một lần quét trên toàn bộ tập dữ liệu (Oliveira & Zaiane, 2003a).Bản phác thảo của SWA được đưa ra trong Hình 3(b).Trước tiên, thuật toán sao chép các giao dịch không nhạy cảm vào cơ sở dữ liệu đã được vệ sinh và sau đó sử dụng cơ chế lập chỉ mục để tăng tốc quá trình ẩn.Không giống như các thuật toán khác có ngưỡng tiết lộ duy nhất cho tất cả các quy tắc nhạy cảm, SWA

Algorithm: SWA
Input: D, mining permissions (M_P), Sliding window (K)
Output: D'

```
For each K transactions in D{
  For each transaction t ∈ K{
    For each sensitive itemset ∈ MP{
      If transaction t is not sensitive
        Transaction t is copied into D'
      Else{
        Transaction t is added to inverted index list
        Update support of each item in t}}
    If transaction t is sensitive{
      Sort its items in descending order of frequency
      For each sensitive itemset ∈ MP
      Select item with highest frequency as victim}}}
For each sensitive itemset ∈ MP{
  N_iteration=| sensitive transactions |*(1-α_min)
  Sort the transactions in ascending order of size}
For each sensitive itemset ∈ MP{
  While N_iteration>0
    Remove victim item from sensitive transactions}
```

Hình 3. Tiếp tục

Algorithm: IGA
Input: D, a set P_S of itemsets to hide, α_{min}
Output: D'

```
For each transaction t in D{
  Update support of each item in t
  Sort the items in t in alphabetic order
  For each sensitive itemset{
    if t is correspond to sensitive itemset
      transaction t is added to inverted index list}}
For each sensitive itemset{
  sort sensitive transactions in descending order of conflict degree
  N_iteration=| sensitive transactions |*(1-α_min)}
Group sensitive rules in a set of groups
Assign labels to each group and select item with lower support as victim item
Order the groups in by size in term of number of sensitive itemsets in group
Compare groups pairwise and start with the largest
For each sensitive itemset{
  Sort Transactions in descending order of conflict degree
  While N_iteration>0
    Remove victim item from sensitive transactions}
```

Hình 4. Mã giả của thuật toán Thuật toán nhóm mục (IGA).

có ngưỡng tiết lộ được chỉ định cho từng liên kết nhạy cảm

luật lệ.Tập hợp các quyền khai thác (MP) được gọi là tập hợp các ánh xạ của quy tắc kết hợp nhạy cảm vào ngưỡng tiết lộ tương ứng của nó.Oliveira và Zaiane (2003b) đã đề xuất hai thuật toán, Thuật toán ngẫu nhiên (RA) và Thuật toán Round Robin (RRA), để ẩn các quy tắc nhạy cảm bằng cách giảm các tập mục tạo ra chúng.Các thuật toán này xem xét tác động của việc thay đổi giao dịch đối với các quy tắc nhạy cảm.

2004. Verykios và al. (2004b) mở rộng các công việc của Dasseni và cộng sự.(2001) cũng như đề xuất thuật toán 2.b để ẩn tập mục tạo ra các quy tắc nhạy cảm.Thuật toán bóp méo dựa trên mức độ ưu tiên (PDA) và Thuật toán biến dạng sắp xếp dựa trên trọng số (WSDA) (Pontikakis và cộng sự, 2004a) đã được trình bày để che giấu các quy tắc nhạy cảm bằng cách xây dựng một phương pháp phỏng đoán trong giai đoạn lựa chọn vật phẩm của PDA và trong lựa chọn giao dịch giai đoạn của WSDA.Những thuật toán này là nỗ lực đầu tiên nhằm xác định trọng số cho các giao dịch.Pontikakis và cộng sự.(2004b) đã đề xuất Thuật toán chặn (BA) nhằm mục đích tạo ra các quy tắc không tồn tại trong tập dữ liệu gốc bằng cách thêm các ẩn số vào các giao dịch.

2005. Việc ẩn tập mục thường xuyên được Menon et al.(2005).Họ đề xuất

Algorithm: Max-Min2
Input: D, a set P_S of itemsets to hide, α_{min}, the positive border
Output: D'

```
While |PS|≠ ∅ {
  Sort PS in increasing order of support
    Select itemset with lowest support
    Lsensitive ← Find corresponding transactions for sensitive itemset
    Build vi-list representation
    While α(itemset)≥ α_min {
      If max-min is not attained by a vi-list
        Determine an itemset with minimum impact as max-min itemset
      Lmax-min← Find corresponding transactions for max-min itemset
      Sanitization_list ← Compute Lsensitive - Lmax-min
      If the Sanitization_list is not empty then
        Remove victim item from transaction t in the Sanitization_list
      Else remove item from transaction t with minimum impact on max-min itemsets
        α(itemset) = α(itemset)-1}
    Remove hidden itemset from PS}
```

Hình 5. Mã giả cho thuật toán Max-Min2.

các thuật toán Chăn và Thông minh giải quyết CSP bằng cách sử dụng lập trình số nguyên để giảm thiểu số lượng giao dịch được chọn lọc, trong khi các thuật toán này sử dụng phương pháp phỏng đoán để tìm các mục nạn nhân.Wang và Jafari (2005) đã kết hợp các ẩn số để che giấu các quy tắc kết hợp dự đoán và trình bày các thuật toán Tăng cường hỗ trợ bên tay trái (ISL) và Giảm hỗ trợ bên tay phải (DSR).Sun và Yu (2005) đề xuất Phương pháp tiếp cận dựa trên biên giới (BBA) lấy cảm hứng từ lý thuyết biên giới của các tập mục phổ biến (Mannila & Toivonen, 1997) để duy trì chất lượng đường viền của các tập mục không nhạy cảm trong mạng tập mục.

2006. Divanis và Verykios (2006) đã đưa ra khái niệm về khoảng cách giữa cơ sở dữ liệu gốc và cơ sở dữ liệu đã được làm sạch trong thuật toán Nội tuyến.Thuật toán này dựa vào quy trình sửa đổi biên giới để xác định số lượng mặt hàng cần vệ sinh ít nhất, thay vì xem xét số lượng giao dịch được vệ sinh tối thiểu.Nó giải quyết CSP bằng cách sử dụng Lập trình số nguyên nhị phân (BIP).Moustakides và Verykios (2008) đã đề xuất hai thuật toán dựa trên đường viền, đó là Max-Min1 và Max-Min2, kiểm soát tác động của việc dọn dẹp đối với các tập mục dễ bị tổn thương hơn trong quá trình ẩn, thay vì tất cả các tập mục trên ranh giới.Hình 5 là bản phác thảo của thuật toán Max-Min2.

20 07. Amiri (20 07) đã đề xuất ba phương pháp phỏng đoán, đó là Tổng hợp, Phân tách và Kết hợp, hoạt động tốt hơn SWA bằng cách cung cấp tiện ích dữ liệu cao hơn và độ biến dạng thấp hơn nhưng phải trả giá bằng tốc độ tính toán tăng lên.Thuật toán Xung đột mục tối đa đầu tiên (MICF) được đề xuất để vượt trội hơn IGA về mặt giảm số lượng mục bị xóa và khắc phục sự chồng chéo giữa các nhóm (Li & Chang, 2007).Vương và cộng sự.(2007a) đã mở rộng thuật toán ISL và DSR (Wang & Jafari, 2005) bằng kỹ thuật biến dạng.Các thuật toán Giảm độ tin cậy bằng cách giảm hỗ trợ (DCDS) và Giảm độ tin cậy bằng tăng hỗ trợ (DCIS) đã được đề xuất để tự động ẩn các quy tắc kết hợp đề xuất hợp tác mà không cần khai thác trước và chọn các quy tắc ẩn (Wang và cộng sự, 2007b).Verykios và cộng sự.(2007) đã cải tiến thuật toán BA (Pontikakis và cộng sự, 2004b) bằng cách áp dụng phương pháp phỏng đoán lựa chọn giao dịch được sử dụng trong WSDA (Pontikakis và cộng sự, 2004a).Wu và cộng sự.(2007), đã trình bày một phương pháp tác dụng phụ hạn chế nhằm phân loại tất cả các sửa đổi hợp lệ liên quan đến các quy tắc nhạy cảm, các quy tắc không nhạy cảm và các quy tắc giả mạo có thể bị ảnh hưởng bởi các sửa đổi.Sau đó, các phương pháp heuristic được sử dụng để sửa đổi các giao dịch nhằm tăng số lượng quy tắc nhạy cảm ẩn, đồng thời giảm số lượng mục được sửa đổi (Divanis & Verykios, 2009a).

2008. Giảm bớt Ủng hộ Và Sự tự tin (DSC) thuật toán đã được đề xuất để che giấu các quy tắc kết hợp dự đoán (Wang và cộng sự, 2008).Menon và Sarkar (2008) đã mở rộng thuật toán được trình bày trong Menon et al.(2005) để giảm thiểu cả số lượng giao dịch được kiểm soát và số lượng tập mục không nhạy cảm bị mất.

2009. Divanis và Verykios (2009a) đã thêm phần mở rộng cơ sở dữ liệu vào cơ sở dữ liệu gốc thay vì sửa đổi các giao dịch hiện có.Phần mở rộng chứa một tập hợp các giao dịch làm giảm tầm quan trọng của các mẫu nhạy cảm đến mức chúng trở nên không còn thú vị theo quan điểm của thuật toán khai thác dữ liệu, đồng thời ảnh hưởng tối thiểu đến tầm quan trọng của các tập mục không nhạy cảm.Họ đã đề xuất một thuật toán lai kết hợp CSP, BIP và sửa đổi đường viền để ẩn các tập mục nhạy cảm.Wang (2009) đã cải tiến thuật toán DSC (Wang và cộng sự, 2008) và giới thiệu thuật toán Bảo trì quy tắc kết hợp thông tin (MSI) để bảo vệ thông tin nhạy cảm khi cơ sở dữ liệu được cập nhật thường xuyên.Tập dữ liệu mới được thêm vào sẽ được MSI làm sạch riêng biệt và sau đó được kết hợp với cơ sở dữ liệu gốc.Divanis và Verykios (2009b) đã cải tiến cách tiếp cận Nội tuyến (Divanis & Verykios, 2006) bằng quy trình hai giai đoạn.Quá trình khử trùng sẽ kết thúc ở giai đoạn đầu tiên nếu kiến thức nhạy cảm được che giấu mà không gây ra tác dụng phụ.Mặt khác, bản sao kép của thuật toán Nội tuyến được thực hiện trong giai đoạn thứ hai để thuật toán ẩn loại bỏ có chọn lọc những bất bình đẳng khỏi CSP không khả thi, cho đến khi CSP trở nên khả thi và sau đó CSP được giải quyết để đạt được tập dữ liệu đã được làm sạch.

2012. Lý thuyết mạng giao nhau (Grätzer, 2010) lần đầu tiên được nghiên cứu trong thuật toán Ẩn quy tắc kết hợp dựa trên mạng giao nhau (ILARH) (Hai & Somjit, 2012) để lựa chọn mục.Thuật toán dựa trên khoảng cách và giao điểm mạng (DIL) được đề xuất bởi Hai et al. (2012) đo lường tác động của quá trình che giấu đối với các quy tắc không nhạy cảm bằng cách gán trọng số cho mỗi giao dịch.Hơn nữa, nó xem xét khoảng cách từ các quy tắc nhạy cảm đến tập các tập mục tối đa và quy tắc không nhạy cảm gần nhất để chọn các mục nạn nhân.

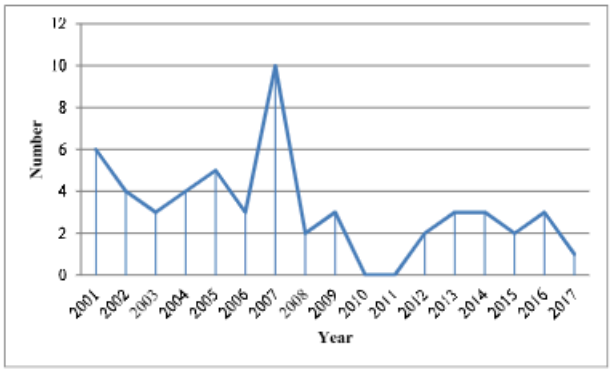
2013. Hải và cộng sự.đã trình bày Ẩn quy tắc kết hợp dựa trên mạng giao nhau (ARHIL) (Hai và cộng sự, 2013a) và Heuristic cho sự tự tin và giảm hỗ trợ dựa trên mạng giao nhau (HCSRIL) (Hai và cộng sự, 2013b) để ẩn các quy tắc nhạy cảm.ARHIL tận dụng tối đa các ưu điểm của thuật toán ILARH (Hai & Somjit, 2012), DIL (Hai và cộng sự, 2012) và HCSRIL (Hai và cộng sự, 2013b).Nó sử dụng các đặc điểm của mạng giao nhau của các tập phổ biến để chọn các mục nạn nhân, đồng thời xác định các giao dịch dựa trên trọng số của chúng lấy cảm hứng từ thuật toán DIL (Hai và cộng sự, 2012).Bằng cách áp dụng khái niệm Tần số nghịch đảo tần số thuật ngữ (TF-IDF), Hong et al. (2013) đã giới thiệu thuật toán Tần số cơ sở dữ liệu nghịch đảo tần số (SIF-IDF) của các hạng mục nhạy cảm để

chỉ định giá trị trọng số cho mỗi giao dịch.2014. Việc sử dụng thuật toán di truyền (GA) để giao dịch

lựa chọn trong bối cảnh ẩn tập mục lần đầu tiên được đề xuất bởi Lin et al.(2014a, 2014b).Lin và cộng sự.(2014a, 2014b) Thuật toán xóa giao dịch dựa trên GA nhỏ gọn (cpGA2DT) (Lin và cộng sự, 2014a) xóa các giao dịch được chỉ định, trong khi thuật toán được đề xuất trong Lin và cộng sự.(2014b) tạo và chèn các giao dịch mới vào cơ sở dữ liệu.Cheng và cộng sự.(2014) đã đề xuất thuật toán Ẩn quy tắc cơ sở tối ưu hóa đa mục tiêu tiến hóa (EMO-RH).Kiến trúc của thuật toán này dựa trên nền tảng PISA (Bleuler và cộng sự, 2003).Trong phần biến thể của nền tảng, sơ đồ mã hóa dành riêng cho vấn đề và toán tử biến thể hiệu quả được đưa ra.Phần chọn lọc của PISA được triển khai bằng thuật toán NSGA II (Deb và cộng sự, 2002).

2015. Lin và cộng sự.(2015) đã giới thiệu hai thuật toán ẩn tập mục, đó là Thuật toán di truyền đơn giản để xóa giao dịch (sGA2DT) và Thuật toán di truyền tiền lớn để xóa giao dịch (pGA2DT) sử dụng GA để chọn giao dịch và sau đó xóa giao dịch khỏi cơ sở dữ liệu gốc.

2016. Hạn chế của thuật toán dựa trên GA là người dùng phải chỉ định một số tham số, bất kể việc tìm giá trị (tỷ lệ) thích hợp cho tham số, chẳng hạn như kích thước nhiễm sắc thể, tỷ lệ đột biến và tỷ lệ chéo.Ngoài ra, các thuật toán này



Hình 6. Số lượng thuật toán dọn dẹp dữ liệu từ năm 2001 đến 2017.

yêu cầu chỉ định thủ công số lượng giao dịch để xóa

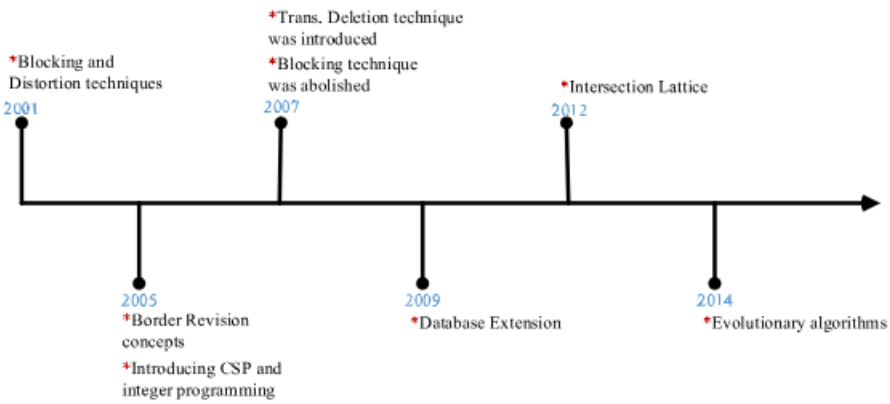
sự.Để giải quyết những vấn đề này, thuật toán xóa giao dịch dựa trên Particle Swarm Optimization (PSO2DT) (Lin và cộng sự, 2016) đã được trình bày có thể xác định số lượng giao dịch tối đa có thể bị xóa cũng như cần ít tham số hơn. bộ.Afshari và cộng sự.(2016) đã đề xuất Thuật toán tối ưu hóa Cuckoo cho việc ẩn các quy tắc kết hợp (COA4ARH) để ẩn các quy tắc kết hợp nhạy cảm bằng cách áp dụng Thuật toán Cuckoo (COA) (Yang & Deb, 2009).Họ đã xác định một hoạt động tiền xử lý với hai giai đoạn khi bắt đầu thuật toán đề xuất.Hoạt động này làm giảm đáng kể cả số lần lặp và thời gian truy cập vào giải pháp tối ưu.Thuật toán sắp xếp mức độ liên quan được đề xuất bởi Cheng et al.(2016b) xây dựng phương pháp phỏng đoán để xác định các giao dịch nhằm vệ sinh.Để giảm tỷ lệ biến dạng, thuật toán tính toán số lượng giao dịch tối thiểu cần được sửa đổi để che giấu quy tắc nhạy cảm.

2017. Thuật toán Giảm độ tin cậy của quy tắc (DCR) đã được đề xuất trong Telikani và Shahbahrami (2017) nhằm cải thiện giải pháp MaxMin (Moustakides & Verykios, 20 06, 20 08) bằng cách sử dụng hai phương pháp phỏng đoán để ẩn các quy tắc kết hợp.Trong thuật toán này, sự kết hợp giữa MaxMin và phương pháp heuristic được xây dựng để chọn các mục nạn nhân, trong khi các giao dịch nhạy cảm được chọn bằng giải pháp heuristic.

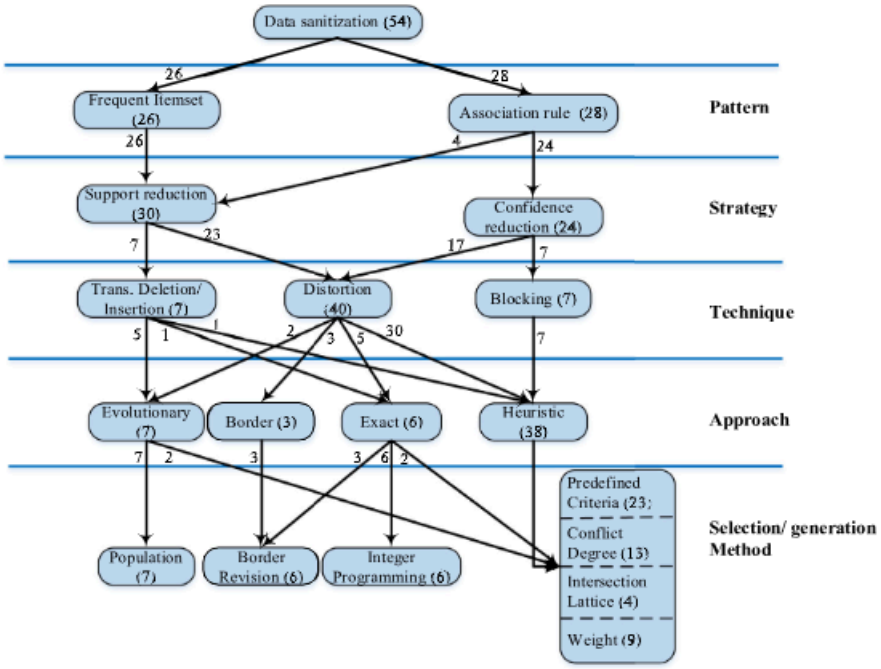
Hình 6 mô tả số lượng thuật toán vệ sinh từ năm 2001 đến năm 2017. Như được hiển thị, số lượng thuật toán cao nhất đã được đề xuất vào năm 2007 (10 trên 54, 19%);Ngoài ra, không có thuật toán nào được trình bày trong năm 2010 và 2011. Kể từ thời điểm đó, việc ẩn luật kết hợp đã thu hút nhiều sự quan tâm hơn trong những năm gần đây (2012–2017).

Hình 7 cho thấy các điểm chính của nghiên cứu về quá trình vệ sinh dữ liệu

thôi.Như có thể thấy trong hình, các kỹ thuật chặn và bóp méo đã được sử dụng vào năm 2001 để sửa đổi các giao dịch nhạy cảm.Năm 2005, trọng tâm của các thuật toán là duy trì tính tiện ích và độ chính xác của cơ sở dữ liệu đã được làm sạch để lý thuyết biên giới và CSP được xây dựng tương ứng cho các mục đích này.Kết quả của việc áp dụng các kỹ thuật này là các phương pháp tiếp cận chính xác và biên giới đã xuất hiện vào năm 2005. Đồng thời với việc bãi bỏ kỹ thuật chặn vào năm 2007, kỹ thuật xóa giao dịch đã được Amiri (2007) giới thiệu.Trên thực tế, các nghiên cứu được công bố năm 2007 đã tập trung vào các kỹ thuật dọn dẹp trong khi tập trung vào các phương pháp lựa chọn vào năm 2005. Kỹ thuật chèn giao dịch đã được sử dụng vào năm 2009 để làm giảm tầm quan trọng của các tập mục nhạy cảm.Vào năm 2012, lý thuyết mạng lưới giao nhau đã làm tăng động lực nghiên cứu sau hai năm nghiên cứu trì trệ từ năm 2010 đến năm 2011. Khung dựa trên GA lần đầu tiên được áp dụng để chọn các giao dịch vào năm 2014, và do đó phương pháp tiến hóa đã được giới thiệu.



Hình 7. Những điểm chính trong vòng đời của quá trình vệ sinh dữ liệu.



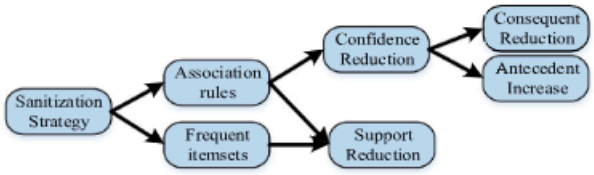
Hình 8. Phân loại các thuật toán dọn dẹp dữ liệu theo bốn khái niệm về chiến lược, kỹ thuật, cách tiếp cận và phương pháp.

6. Công nghệ vệ sinh dữ liệu tiên tiến

Chúng tôi nghiên cứu và phân tích tất cả 54 thuật toán và trình bày phân loại các hướng tiên tiến nhất dựa trên bốn khía cạnh của quy trình sàng lọc dữ liệu, bao gồm chiến lược, kỹ thuật, cách tiếp cận và phương pháp lựa chọn.Hình 8 cho thấy phân loại của phân tích này trình bày sự phân bố của các thuật toán liên quan đến từng danh mục.Tổng phân phối ở cấp độ cuối cùng khác với các cấp độ khác vì một số thuật toán là kết hợp và kết hợp các phương pháp lựa chọn khác nhau.Ví dụ: ba thuật toán dựa trên chính xác sử dụng phương pháp sửa đổi đường viền và lập trình số nguyên để chọn các sửa đổi tốt nhất.Đặc điểm của từng khía cạnh cũng như thảo luận về các giải pháp nhằm đẩy nhanh quá trình vệ sinh sẽ được trình bày trong các phần sau.

6.1.Ẩn chiến lược

Từ góc độ chiến lược ẩn, các thuật toán dọn dẹp dữ liệu được phân tích thành hai nhóm, (1) ẩn tập mục thường xuyên và (2) ẩn quy tắc.Danh mục đầu tiên ẩn các tập mục nhạy cảm bằng cách giảm độ hỗ trợ của chúng xuống dưới mức αmin.Loại thứ hai áp dụng chiến lược hỗ trợ hoặc giảm độ tin cậy để che giấu các quy tắc nhạy cảm.Trong chiến lược giảm độ hỗ trợ cho các luật kết hợp, việc hỗ trợ tạo tập mục của luật bị giảm.Trong chiến lược giảm độ tin cậy, sự hỗ trợ của hệ quả của quy tắc bị giảm hoặc tiền đề của quy tắc được tăng lên.Hình 9 thể hiện sự phân loại của các chiến lược được đề cập trong quá trình khử trùng.Các phần tiếp theo mô tả các cách khác nhau để làm giảm sự tin cậy và hỗ trợ với các tác dụng phụ khác nhau do mỗi chiến lược tạo ra.



Hình 9. Các chiến lược trong quy trình vệ sinh dữ liệu.

Bảng 4 Tác động của việc thêm {a} vào giao dịch T2.

sửa đổi	AP (bộ vật phẩm)	AP (quy tắc)
{a}−→T2	{a, e} {a, b, e} {a, d, e}	{e} ⇒{a}, {e} ⇒{a, b}, {a, e} ⇒{b}, {b, e} ⇒{a}, {e} ⇒{a, d}, {a, e} ⇒{d}, {d, e} ⇒{a}, {b} ⇒{a}, {d} ⇒{a}, {b, d} ⇒{a}

6.1.1.Giảm độ tin cậy Như thể hiện trong Hình 9, có thể xem xét hai chiến lược để giảm

độ tin cậy của quy tắc nhạy cảm X ⇒Y, bao gồm giảm hệ quả và tăng tiền đề.Trong quá trình giảm thiểu, số lượng hỗ trợ của tập mục Y sẽ giảm bằng cách loại bỏ các mục khỏi các giao dịch nhạy cảm.Chiến lược này chỉ tạo ra các tác dụng phụ MC và AP, được thể hiện bằng cách sử dụng ví dụ trong Bảng 2. Trong chiến lược tăng trước đó, số lượng hỗ trợ của X được tăng lên bằng cách thêm các mục vào các giao dịch không nhạy cảm hỗ trợ một phần X và hỗ trợ đầy đủ Y. Chiến lược này không chỉ đưa ra các tác dụng phụ MC và AP tương tự như chiến lược giảm thiểu hậu quả mà còn có thể không che giấu được tất cả các mô hình nhạy cảm.Ngoài ra, nó làm sạch nhiều giao dịch hơn chiến lược tăng trước khi ẩn một quy tắc cụ thể vì nó làm giảm cả tử số và mẫu số của phương trình độ tin cậy.(2) trong khi chiến lược rút gọn hệ quả chỉ rút gọn tử số.Ba ví dụ sau đây cho thấy tác dụng phụ do chiến lược tăng trước đó gây ra.Tất cả các ví dụ được thực hiện trên cơ sở dữ liệu giao dịch được trình bày trong Bảng 1(a).

Ví dụ 1. Ẩn lỗi.Giả sử rằng {d} ⇒{b} là một quy tắc nhạy cảm thì cần phải có hai giao dịch được làm sạch.Không thể thêm mục {d} vào bất kỳ giao dịch nào vì tất cả các giao dịch không nhạy cảm cho mục {d} cũng hỗ trợ mục {b}, do đó, không thể ẩn quy tắc {d} ⇒{b}.Điều này thường xảy ra khi số lượng hỗ trợ của Y rất cao và số lượng giao dịch cần thiết để ẩn quy tắc là không đủ, điều đó có nghĩa là δ < (m- α(Y)) trong đó δ là số lần lặp, được định nghĩa là số lượng giao dịch cần thiết để giảm độ tin cậy của quy tắc xuống dưới βmin .Giá trị δ cho chiến lược tăng tiền đề được tính bằng phương trình.(3).

$$\delta (X \Rightarrow Y) = |D| * (\alpha (X \Rightarrow Y) / \beta min - \alpha (X)) \quad (3)$$

Ví dụ 2. Các mẫu tạo tác.Giả sử rằng quy tắc {a} ⇒{b, d} là nhạy cảm, cần phải sửa đổi một lần để ẩn quy tắc, trong đó mục {a} được thêm vào giao dịch T2 hoặc giao dịch T6.Ví dụ: nếu chúng ta thêm mục {a} vào giao dịch T2;sửa đổi này giới thiệu ba tập mục mới và 10 quy tắc mới, như được trình bày trong Bảng 4.

Ví dụ 3. Bỏ lỡ chi phí.Chúng tôi xem xét Ví dụ 2 để chứng minh các quy tắc không nhạy cảm bị mất khỏi dữ liệu đã được lọc.Khi thêm mục {a} vào T2, các quy tắc {a} ⇒{c}, {a} ⇒{b, c}, {a, b} ⇒{c}, {a, d} ⇒{c} , {a, d} ⇒{b, c} và {a, b, d} ⇒{c} bị ẩn.Điều này là do thực tế là số lượng hỗ trợ tiền đề của các quy tắc này tăng lên và do đó độ tin cậy của các quy tắc này bị giảm.

1.b (Dasseni và cộng sự, 2001; Verykios và cộng sự, 2004b), DCDS (Wang và cộng sự, 2007b), DSR (Wang và cộng sự, 2007a, 2007b), ILARH (Hai & Somjit, 2012) , DIL (Hai và cộng sự, 2012), ARHIL (Hai và cộng sự, 2013a), HCSRIL (Hai và cộng sự, 2013b), PDA, WSDA (Pontikakis và cộng sự, 2004a; Verykios và cộng sự, 2007), CR (Saygin và cộng sự, 20 01, 20 02), COA4ARH (Afshari và cộng sự, 2016) và DCR (Telikani & Shahbahrami, 2017) sử dụng chiến lược giảm thiểu hậu quả để giảm độ tin cậy của các quy tắc nhạy cảm.1.a (Dasseni và cộng sự, 2001; Verykios và cộng sự, 2004b), DCIS (Wang và cộng sự, 2007b), ISL (Wang và cộng sự, 20 07a, 20 07b), BA (Pontikakis và cộng sự. , 2004b) và thuật toán CR (Saygin et al., 20 01, 20 02) áp dụng chiến lược tăng tiền đề để ẩn các quy tắc kết hợp nhạy cảm.Không giống như các thuật toán ẩn quy tắc kết hợp khác ẩn một tập hợp các quy tắc cụ thể, ISL, DSR (Wang và cộng sự, 2007a, 2007b), DSC (Wang và cộng sự, 2008) và MSI (Wang, 2009) ẩn các quy tắc có chứa một tập hợp các mục nhạy cảm ở phía bên trái của quy tắc.Mặt khác, DCIS và DCDS (Wang và cộng sự, 2007b) ẩn các quy tắc kết hợp rằng các mục nhạy cảm thuộc về phía bên phải của quy tắc.

Sự kết hợp của hai chiến lược trên đã được trình bày trong Verykios et al.(2007) và Wu và cộng sự.(2007) để giảm độ tin cậy của các quy tắc nhạy cảm bằng cách giảm số lượng hỗ trợ của Y và bằng cách tăng số lượng hỗ trợ của X. Ý tưởng chính đằng sau chiến lược này là cân bằng các tác dụng phụ về các quy tắc bị mất và các quy tắc ma được tạo ra. trong cơ sở dữ liệu đã được vệ sinh;bởi vì chỉ giảm số lượng hỗ trợ của quy tắc dẫn đến nhiều quy tắc bị mất và cũng chỉ tăng số lượng hỗ trợ của quy tắc tiền đề sẽ tạo ra nhiều quy tắc mới.Thuật toán BA mở rộng (Verykios et al., 2007) và thuật toán được đề xuất trong Wu et al.(2007) sử dụng chiến lược này để ẩn các luật kết hợp.

6.1.2.Giảm hỗ trợ Chiến lược giảm hỗ trợ không tạo ra bất kỳ

tập mục trong khi nó có thể tạo ra các quy tắc mới do tác dụng phụ vì độ mạnh của các quy tắc yếu có thể được tăng lên bằng cách giảm sự hỗ trợ của phía bên trái của chúng.Mặt khác, chiến lược này che giấu một số tập mục không nhạy cảm do bỏ lỡ chi phí.Hơn nữa, số lượng các quy tắc không nhạy cảm được ẩn bởi chiến lược giảm độ hỗ trợ nhiều hơn so với chiến lược giảm độ tin cậy vì tất cả các quy tắc kết hợp xuất phát từ các tập mục ẩn đều được ẩn sau quá trình chuẩn hóa.

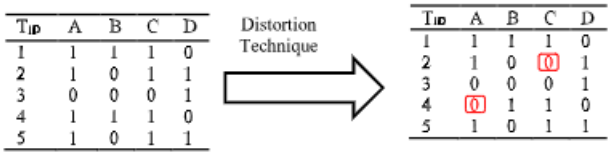
RA, RRA (Oliveira & Zaiane, 2003b), 2.b (Verykios et al., 2004b) và GIH (Saygin et al., 20 01, 20 02) ẩn các quy tắc nhạy cảm bằng cách giảm tập mục tạo ra của chúng.DSC (Wang và cộng sự, 2008) và MSI (Wang, 2009) giảm số lượng hỗ trợ của quy tắc trước đó nhằm giảm độ tin cậy của các quy tắc này.Divanis và Verykios (2009a) đã đề xuất một phương pháp ẩn tập mục nhằm chèn các giao dịch tổng hợp theo cách giảm số lượng hỗ trợ của các tập mục nhạy cảm và số lượng hỗ trợ của các tập mục không nhạy cảm được duy trì ở mức có thể.

6.2.Kỹ thuật vệ sinh

Trong quá trình dọn dẹp, các giao dịch hiện tại được sửa đổi hoặc xóa khỏi cơ sở dữ liệu hoặc các giao dịch mới được thêm vào cơ sở dữ liệu.Các kỹ thuật bóp méo và chặn sửa đổi các giao dịch hiện có bằng cách xóa/chèn các mục từ/vào giao dịch.Nghiên cứu này coi kỹ thuật chèn và xóa giao dịch là một khái niệm thống nhất được gọi là kỹ thuật xóa/chèn giao dịch để bao quát phạm vi thuật toán rộng hơn.Phần này thảo luận về quá trình vệ sinh theo các kỹ thuật vệ sinh khác nhau.

6.2.1.Kỹ thuật làm biến dạng Kỹ thuật này lần đầu tiên được đề xuất bởi Atallah et al.(1999) đến

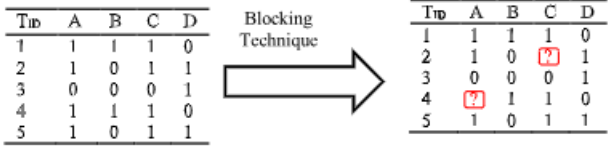
xóa các mục cụ thể khỏi các giao dịch.Cách triển khai kỹ thuật biến dạng trong cơ sở dữ liệu nhị phân và phân loại



Tid	A	B	C	D
1	1	1	1	0
2	1	0	1	1
3	0	0	0	1
4	1	1	1	0
5	1	0	1	1

Tid	A	B	C	D
1	1	1	1	0
2	1	0	0	1
3	0	0	0	1
4	0	1	1	0
5	1	0	1	1

Hình 10. Một ví dụ về kỹ thuật bóp méo.



Tid	A	B	C	D
1	1	1	1	0
2	1	0	1	1
3	0	0	0	1
4	1	1	1	0
5	1	0	1	1

Tid	A	B	C	D
1	1	1	1	0
2	1	0	?	1
3	0	0	0	1
4	?	1	1	0
5	1	0	1	1

Hình 11. Các giá trị không xác định trong kỹ thuật chặn.

khác.Trong cơ sở dữ liệu nhị phân, các thuật toán sửa đổi các giao dịch bằng cách chèn (thay thế 0 bằng 1) hoặc xóa (thay thế 1 bằng 0), trong khi ở cơ sở dữ liệu phân loại, các mục nạn nhân bị xóa khỏi các giao dịch nhạy cảm hoặc các mục được thêm vào giao dịch không nhạy cảmNgười ta đã chứng minh rằng kỹ thuật làm biến dạng để ẩn luật kết hợp là một bài toán NP-khó (Atallah và cộng sự, 1999).Trong cơ sở dữ liệu được làm sạch bằng kỹ thuật bóp méo, người nhận dữ liệu không thể chắc chắn về sự thay đổi của bất kỳ mục cụ thể nào trong cơ sở dữ liệu vì bất kỳ mục nào có thể đã được chèn hoặc xóa bởi quá trình làm sạch.Một ví dụ về kỹ thuật biến dạng trong cơ sở dữ liệu nhị phân được hiển thị trong Hình 10. Như có thể thấy trong hình này, một số số 1 đã được thay thế bằng số 0.

6.2.2.Kỹ thuật chặn Ý tưởng chính của các chặn kỹ thuật đã từng là Đầu tiên ủng hộ đặt ra TROIChang Và Moskowitz (1998) cái mà các nạn nhân các mục được thay thế bằng các giá trị không xác định (Saygin và cộng sự, 2001).Pontikakis và cộng sự.(2004b) gợi ý rằng cần phải tối đa hóa số lượng ẩn số để ngăn chặn đối thủ khôi phục các quy tắc nhạy cảm.Hình 11 cho thấy một ví dụ về kỹ thuật chặn để ẩn luật kết hợp.

Kỹ thuật chặn tạo ra sự khác biệt giữa các mục bị bóp méo và các mục không bị ảnh hưởng của cơ sở dữ liệu gốc vì chỉ các mục bị bóp méo mới được thay thế bằng các mục không xác định.Đây là một đặc tính hữu ích trong các ứng dụng quan trọng trong đời sống, trong đó sự phân biệt giữa "sai" và "không xác định" có thể rất quan trọng vì nó không thêm bất kỳ quy tắc sai nào vào cơ sở dữ liệu gốc.Về mặt tiêu cực, nó có thể làm mờ đi sự hỗ trợ và độ tin cậy của luật kết hợp do có sự kết hợp của các ẩn số (Cheng và cộng sự, 2016b).Thật vậy, rất khó để khai thác các quy tắc kết hợp không nhạy cảm quan trọng từ cơ sở dữ liệu được làm sạch bằng kỹ thuật chặn vì sự xuất hiện của một mục được biểu thị bằng '1' và các thuật toán khai thác quy tắc kết hợp sẽ tính giá trị này cho mỗi tập mục và bất kỳ giá trị nào ngoại trừ giá trị này có nghĩa là bằng không.Ngoài ra, vì các giao dịch đã được lọc là rõ ràng đối với người nhận dữ liệu nên kẻ tấn công có thể quan sát thấy rằng dữ liệu nhận được đã được lọc do sự tồn tại của kiến thức nhạy cảm.Do đó, rủi ro tiết lộ sẽ tăng lên, đặc biệt là trong tập dữ liệu nhị phân, trong đó mỗi giao dịch chứa các giá trị nhị phân (0 hoặc 1).CR, CR2, GIH (Saygin và cộng sự, 2001, 2002), BA (Pontikakis và cộng sự, 2004b; Verykios và cộng sự, 2007), và phiên bản mở rộng của ISL và DSR (Wang & Jafari, 2005) thuật toán thay thế các mục nhị phân bằng các giá trị chưa biết.

6.2.3.Xóa/chèn giao dịch Kỹ thuật xóa/chèn giao dịch đã nhận được tín hiệu

được quan tâm đặc biệt trong những năm gần đây.Nó nhằm mục đích xóa các giao dịch khỏi cơ sở dữ liệu hoặc chèn các giao dịch mới vào cơ sở dữ liệu gốc.Vấn đề xóa các giao dịch hiện có từ tập dữ liệu lần đầu tiên được giới thiệu trong thuật toán Tổng hợp (Amiri, 2007).Divanis và Verykios (2009a) đề xuất một cách tiếp cận chính xác dựa trên biên giới để các giao dịch không quan trọng được

được thêm vào cơ sở dữ liệu gốc.Hầu hết các thuật toán dựa trên tiến hóa đều sử dụng kỹ thuật chèn/xóa giao dịch.Lin và cộng sự.(2014b) đã đề xuất thuật toán dựa trên GA để chèn các giao dịch vào cơ sở dữ liệu.cpGA2DT (Lin và cộng sự, 2014a), sGA2DT, pGA2DT (Lin và cộng sự, 2015) và PSO2DT (Lin và cộng sự, 2016) loại bỏ các giao dịch khỏi cơ sở dữ liệu gốc để giảm sự hỗ trợ của các tập mục nhạy cảm.Kỹ thuật chèn/xóa giao dịch làm thay đổi số lượng giao dịch của cơ sở dữ liệu;do đó, rất khó để khai thác các mẫu có ngưỡng tối thiểu trước đó.

6.3.Phương pháp vệ sinh

Trọng tâm chính của các thuật toán dọn dẹp là tìm ra giải pháp tối ưu để che giấu tất cả kiến thức nhạy cảm với tác dụng phụ tối thiểu.Với mục đích này, bốn cách tiếp cận đã được trình bày, bao gồm phương pháp heuristic, biên giới, chính xác và tiến hóa.Những cách tiếp cận này được mô tả trong các phần sau.

6.3.1.Phương pháp tiếp cận heuristic Phương pháp này bao gồm các thuật toán hiệu quả và nhanh chóng giúp

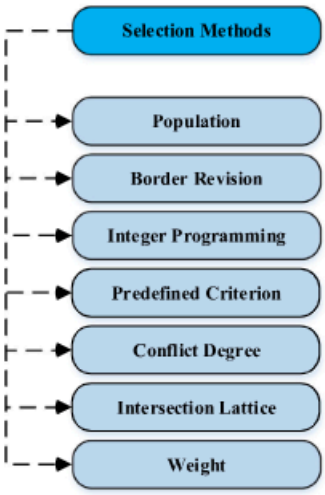
chọn một tập hợp các giao dịch bằng cách sử dụng các tiêu chí được xác định trước.Mặc dù cách tiếp cận heuristic đã nhận được rất nhiều sự chú ý của các nhà nghiên cứu trong những năm gần đây, nhưng nó không nhất thiết là tốt nhất trên toàn cầu và không đảm bảo tính tối ưu của giải pháp ẩn, tuy nhiên, nó thường tìm ra giải pháp gần giải pháp tốt nhất trong thời gian phản hồi nhanh hơn.Do thực tế là các thuật toán dựa trên heuristic luôn hướng đến việc đưa ra các quyết định tốt nhất cục bộ liên quan đến việc che giấu kiến thức nhạy cảm, nên chúng tạo ra nhiều tác dụng phụ không mong muốn hơn các thuật toán khác, đặc biệt là trong việc ẩn các tập mục (Divanis & Verykios, 2010).Một số thuật toán dựa trên heuristic thú vị nhất để che giấu kiến thức nhạy cảm đã được trình bày trong Divanis và Verykios (2010), Verykios (2013), Verykios và Divanis (2008).

6.3.2.Cách tiếp cận biên giới Biên giới (Mannila & Toivonen, 1997) nắm bắt các tập mục đó

của mạng tập mục phổ biến kiểm soát vị trí của đường biên ngăn cách các tập mục phổ biến với các tập mục phổ biến đối tác không thường xuyên của chúng (Verykios & Divanis, 2008).Việc dọn dẹp dữ liệu ảnh hưởng đến ranh giới của cơ sở dữ liệu đã được dọn dẹp;do đó, cách tiếp cận này xem xét quy tắc kết hợp ẩn thông qua việc sửa đổi các đường viền trong mạng của các tập mục phổ biến và không thường xuyên của cơ sở dữ liệu gốc.Nó được Sun và Yu (2007) giới thiệu lần đầu tiên để ẩn các tập phổ biến trong khi vẫn duy trì các tập mục không nhạy cảm với độ hỗ trợ thấp.Các thuật toán dựa trên biên giới làm sạch các giao dịch với tác động tối thiểu đến kết quả của cơ sở dữ liệu được phát hành (Moustakides & Verykios, 2008).Thuật toán BBA (Sun & Yu, 2005, 2007), Max-Min1 và Max-Min2 (Moustakides & Verykios, 20 06, 20 08) sử dụng lý thuyết đường viền để ẩn các tập phổ biến.

6.3.3.Cách tiếp cận chính xác Cách tiếp cận chính xác cố gắng che giấu các mô hình nhạy cảm bằng nguyên nhân-

giảm thiểu biến dạng cho cơ sở dữ liệu đã được làm sạch.Nó xem xét vấn đề ẩn tập mục thường xuyên như một CSP và xây dựng CSP như một chương trình số nguyên để giảm thiểu số lượng các giao dịch hoặc mục được làm sạch (Menon và cộng sự, 2005).Do sử dụng bộ giải lập trình số nguyên để giải bài toán tối ưu nên các thuật toán dựa trên chính xác rất phức tạp (Divanis & Verykios, 2010).Cách tiếp cận đường viền thường được coi là sự bổ sung cho cách tiếp cận ẩn chính xác vì hầu hết các thuật toán dựa trên chính xác đều sử dụng lý thuyết đường viền.Trong phương pháp này, trước tiên, lý thuyết đường biên được áp dụng để tính toán một phần nhỏ



Hình 12. Các phương pháp lựa chọn/tạo ra cho quá trình khử trùng.

các tập mục có thể đóng vai trò quan trọng trong việc duy trì chất lượng của giải pháp ẩn và sau đó sự bất bình đẳng được tạo ra bằng cách sử dụng các phương pháp chính xác để kiểm soát trạng thái của các tập mục được chọn ở đường biên (Verykios, 2013).Không có thuật toán dựa trên độ chính xác tuyệt đối vì các giải pháp heuristic hoặc biên thường được kết hợp để tìm ra giải pháp dọn dẹp.Chăn, Thông minh (Menon và cộng sự, 2005), Nội tuyến (Divanis & Verykios, 2006), và thuật toán được trình bày trong Divanis và Verykios (20 09a, 20 09b) và Menon và Sarkar (2008) đều dựa trên cơ sở chính xác.

6.3.4.Cách tiếp cận tiến hóa Mô hình của các thuật toán tiến hóa bao gồm ngẫu nhiên

các thuật toán tìm kiếm lấy cảm hứng từ quá trình phát triển của thuyết Darwin mới

dung dịch.Các thuật toán này hoạt động với một nhóm cá nhân.Mỗi cá nhân là một giải pháp ứng cử viên cho một vấn đề nhất định và được phát triển theo hướng giải pháp ngày càng tốt hơn cho vấn đề đó.Chất lượng của giải pháp ứng viên được đo bằng hàm thích hợp do người dùng xác định trước.Đây là một mô hình tìm kiếm rất chung chung và có thể được sử dụng để giải quyết nhiều loại vấn đề khác nhau (Maimon & Rokach, 2010).cpGA2DT (Lin và cộng sự, 2014a), thuật toán được đề xuất trong Lin và cộng sự.(2014b), sGA2DT, pGA2DT (Lin và cộng sự, 2015), PSO2DT (Lin và cộng sự, 2016), EMO-RH (Cheng và cộng sự, 2014, 2016a) và COA4ARH (Afshari và cộng sự, 2016) thuật toán sử dụng các thuật toán tiến hóa để ẩn giấu tri thức.

6.4.Phương pháp lựa chọn/tạo

Tìm các mục và giao dịch nạn nhân thích hợp để dọn dẹp cũng như tạo ra các giao dịch mới để thêm vào cơ sở dữ liệu là bước quan trọng nhất trong quy trình dọn dẹp.Nó đóng một vai trò quan trọng trong việc giảm tác dụng phụ.Hình 12 mô tả phân loại các phương pháp được xây dựng để chọn/tạo các mục và giao dịch.

6.4.1.Dân số Trong các bài toán NP-khó, các phương pháp tiếp cận dựa trên dân số là

được sử dụng rộng rãi để tìm giải pháp gần tốt nhất nhằm tối ưu hóa các vấn đề bằng cách đánh giá tất cả các giải pháp.Những cách tiếp cận này tạo điều kiện thuận lợi cho việc tìm kiếm giải pháp tốt bằng cách áp dụng các nguyên tắc tiến hóa tự nhiên.Các thuật toán di truyền (GA) (Holland, 1992) là cách tiếp cận dựa trên dân số cơ bản nhất (Lin và cộng sự, 2016).Trong GA, ý tưởng là mã hóa từng giải pháp dưới dạng nhiễm sắc thể.Quần thể ban đầu được tạo ngẫu nhiên và sau đó mỗi giải pháp được được sao chép bằng nhiều thao tác khác nhau như chọn lọc, đột biến và lai ghép.Cuối cùng, mức độ tốt của nhiễm sắc thể được đánh giá bằng cách sử dụng hàm thích ứng được thiết kế.Quá trình này được lặp lại cho đến khi thỏa mãn điều kiện dừng (Lin và cộng sự, 2016).Trong lĩnh vực ẩn quy tắc kết hợp, mỗi nhiễm sắc thể mã hóa một giải pháp bao gồm một tập hợp các giao dịch nhạy cảm và chức năng thích ứng được thiết kế dựa trên cả ba tác dụng phụ của quá trình ẩn, bao gồm quy tắc bị mất, quy tắc ma và thất bại trong việc ẩn.Lin và cộng sự.(2014a) đã sử dụng GA để chọn một tập hợp các giao dịch để xóa.Họ đã trình bày các thuật toán cpGA2DT (Lin và cộng sự, 2014a), sGA2DT và pGA2DT (Lin và cộng sự, 2015).Lin và cộng sự.(2014b) cũng sử dụng GA để tạo ra các giao dịch thích hợp để đưa vào cơ sở dữ liệu.

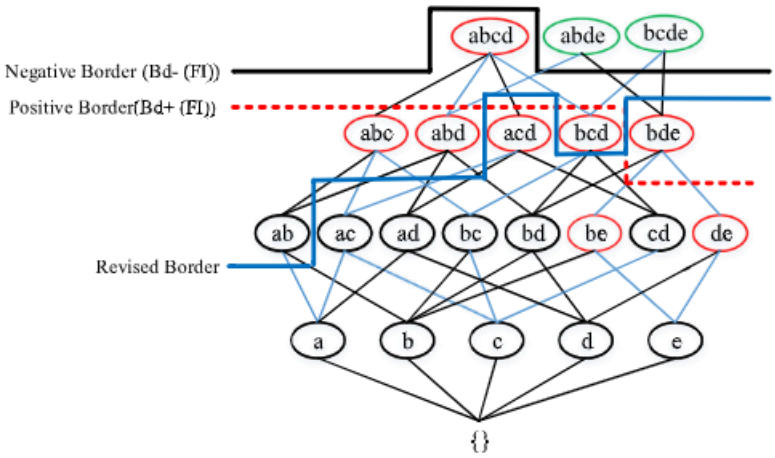
PSO (Kennedy & Eberhart, 1995) được lấy cảm hứng từ hành vi của các loài chim đồ xô đi tìm nguồn thức ăn tốt hơn.Trong PSO, các hạt đại diện cho lời giải của bài toán, trong đó mỗi hạt có vận tốc biểu thị hướng bay về phía các lời giải khác.Quy trình PSO trước tiên khởi tạo ngẫu nhiên các hạt và sau đó thực hiện quá trình tiến hóa lặp đi lặp lại.Trong mỗi lần lặp, mỗi hạt được cập nhật bằng cách sử dụng giá trị tốt nhất cá nhân (pbest) và giá trị tốt nhất toàn cầu (gbest) dựa trên hàm thích ứng được thiết kế để cập nhật các hạt cũ và tạo ra con cái của quần thể.Giá trị gbest là nghiệm tốt nhất trong số tất cả các giá trị pbest trong tổng thể.Các hạt và vận tốc tương ứng của chúng được đánh giá và cập nhật bằng hai giá trị tốt nhất này (Lin và cộng sự, 2016).Việc triển khai PSO dễ dàng hơn GA trong việc khám phá các giải pháp tối ưu vì PSO, không giống như GA, không có các hoạt động chéo và đột biến (Kuo và cộng sự, 2011).Tương tự như GA, nó sử dụng phương pháp tiến hóa ngẫu nhiên cho vấn đề vệ sinh.PSO rời rạc được áp dụng trong thuật toán PSO2DT (Lin và cộng sự, 2016) để tìm một tập hợp các giao dịch.Trong thuật toán này, các hạt và vận tốc của chúng được gán cho tập hợp các mã định danh giao dịch.

Các bài toán tối ưu hóa đa mục tiêu thường được giải quyết bằng thuật toán tiến hóa dựa trên quần thể (Bandaru và cộng sự, 2016).Tối ưu hóa đa mục tiêu tiến hóa đã được sử dụng trong thuật toán EMO-RH (Cheng và cộng sự, 2014, 2016a) bằng cách áp dụng sơ đồ mã hóa nhị phân để ẩn các quy tắc nhạy cảm.Trong thuật toán này, các tác dụng phụ được xây dựng dưới dạng mục tiêu tối ưu hóa nhằm tìm ra tập hợp con giao dịch phù hợp để sửa đổi.Mỗi bit trong nhiễm sắc thể chỉ tương ứng với một giao dịch hỗ trợ, do đó làm giảm kích thước của không gian tìm kiếm.Mỗi nhiễm sắc thể được chia thành k đoạn, trong đó k là số quy tắc nhạy cảm.Độ dài của đoạn thứ j là số lượng giao dịch hỗ trợ quy tắc nhạy cảm thứ j.Trong EMO-RH, phép lai chéo đồng nhất và đột biến bit độc lập được sử dụng trong quá trình tiến hóa.

Lấy cảm hứng từ loài chim cúc cu, Thuật toán chim cúc cu (COA) lần đầu tiên được phát triển bởi Yang và Deb (2009).Giống như các thuật toán tiến hóa khác, COA cũng bắt đầu công việc của mình từ một quần thể ban đầu ngẫu nhiên được gọi là "môi trường sống" và nó được hình thành bởi chim cu gáy.Thuật toán này đã được áp dụng trong thuật toán COA4ARH (Afshari và cộng sự, 2016) để chọn các giao dịch nhạy cảm dựa trên ba hàm thích hợp được xác định cho ba tác dụng phụ.COA4ARH cũng giới thiệu một hàm nhập cư để thoát khỏi bất kỳ mức tối ưu cục bộ nào.Mỗi giải pháp của tập hợp ban đầu được hiển thị bằng một chuỗi 0 và 1.Trong thuật toán này, giải pháp đầu tiên của tập hợp ban đầu là một chuỗi các giao dịch nhạy cảm được chọn bởi bước tiền xử lý và các giải pháp khác được tạo ra bằng cách định lượng ngẫu nhiên các giao dịch nhạy cảm đó đã được xử lý trong bước tiền xử lý trong khi các giao dịch khác không thay đổi từ giải pháp đầu tiên.

6.4.2.Sửa đổi biên giới Để giảm thiểu tác động của quá trình vệ sinh đối với

đối với các mẫu không nhạy cảm, cần kiểm soát tác động của việc sửa đổi đối với các mẫu này.Trên thực tế, phương pháp này duy trì



Hình 13. Lưới các tập mục phổ biến có đường viền ban đầu và đường viền được sửa đổi cho Bảng 1(b).

Bảng 5 Các tập mục phổ biến không nhạy cảm cần được trích xuất từ cơ sở dữ liệu đã được làm sạch.

Các tập mục thường xuyên được mong đợi trong D′
{a, c, d}, {b, d, e} {a, c},{a, d},{b, c},{b, d}, {b, e}, {c, d} ,{d, e} {a}, {b}, {c}, {d}, {e}

chất lượng tổng hợp của cơ sở dữ liệu kết quả trong quá trình ẩn giấu tri thức (Sun & Yu, 2007).Theo lý thuyết biên giới (Mannila & Toivonen, 1997), các phần tử trên biên của mạng các tập mục là ranh giới của các tập mục không phổ biến.Sau đây, phương pháp sửa đổi đường viền được mô tả rõ ràng bằng cách sử dụng các tập phổ biến của Bảng 1(b).Giả sử rằng các tập mục {a, b} và {b, c, d} là nhạy cảm.Bảng 5 trình bày các tập mục phổ biến dự kiến, tức là các tập mục phổ biến trong cơ sở dữ liệu đã được chọn lọc (D′).Theo thuộc tính Apriori (Agrawal & Srikant, 1994), khi một tập mục phổ biến nhạy cảm bị ẩn, các siêu tập hợp của nó cũng bị ẩn khỏi cơ sở dữ liệu đã được làm sạch.Do đó, các tập mục {a, b, c}, {a, b, d} và {a, b, c, d} bị ẩn khỏi D′ sau khi ẩn các tập mục {a, b} và {b, c, d }.

Hình 13 cho thấy mạng tập mục cho Bảng 1(b) với đường viền âm, đường viền dương và đường viền được sửa đổi.Đường viền âm (hoặc đường viền gốc) của Tập mục thường xuyên (FI), ký hiệu là Bd− (FI), là tập hợp tất cả các tập mục không thường xuyên từ D trong đó tất cả các tập con thích hợp đều xuất hiện trong FI.Đường viền dương của FI, ký hiệu là Bd+ (FI), là tập hợp tất cả các tập phổ biến tối đa xuất hiện trong FI.Đường viền được sửa đổi là đường biên lý tưởng sau khi ẩn các tập mục, cho phép chúng ta ẩn các tập mục nhạy cảm mà không thay đổi bất kỳ tập mục thường xuyên dự kiến nào thành không thường xuyên.Nếu tất cả các tập mục trên biên giới sửa đổi vẫn thường xuyên thì giải pháp vệ sinh là tối ưu (Sun & Yu, 2007).Do đó, phương pháp này tập trung vào việc duy trì chất lượng của đường viền được sửa đổi bằng cách tham lam lựa chọn các sửa đổi với tác dụng phụ tối thiểu.

Thuật toán BBA (Sun & Yu, 20 05, 20 07) chọn các sửa đổi có liên quan bằng cách đánh giá tác động của bất kỳ sửa đổi nào lên đường viền trong quá trình ẩn tập mục, sao cho các mục có tác động tối thiểu đến đường viền sửa đổi sẽ bị xóa.Cách tiếp cận Max-Min (Moustakides & Verykios, 20 06, 20 08) sử dụng lý thuyết đường viền để lựa chọn mục.Theo cách tiếp cận này, đối với mỗi mục của tập mục nhạy cảm, các tập mục không nhạy cảm có chứa mục đó sẽ được chỉ định.Trong số các tập mục được chỉ định, một tập mục có độ hỗ trợ tối thiểu được chọn làm tập mục ứng cử viên để được bảo vệ khỏi việc ẩn.Do đó, số lượng tập mục ứng viên bằng với số phần tử của tập mục nhạy cảm.Sau đó, trong số các tập mục ứng cử viên, một tập mục có độ hỗ trợ tối đa được chọn làm tập mục tối thiểu.Cuối cùng, mục thuộc tập mục maxmin được chọn làm mục nạn nhân.Dựa trên cách tiếp cận này, các thuật toán Max-Min1 và Max-Min2 đã được đề xuất để thực hiện các sửa đổi theo cách sao cho số lượng hỗ trợ của tập mục maxmin, nếu có thể, không bị sửa đổi (Moustakides & Verykios, 2008).Khi có nhiều hơn một tập mục tối đa, hai thuật toán được thực hiện theo hai cách khác nhau để chọn mục nạn nhân.Trong trường hợp này, Max-Min1 chọn ngẫu nhiên một trong số chúng làm tập mục tối đa, trong khi MaxMin2 chọn tập mục có hiệu ứng tối thiểu.Telikani và Shahbahrami (2017) đã phát triển phương pháp Max-Min để chọn các mục nạn nhân trong bối cảnh ẩn quy tắc để thuật toán DCR của họ kiểm soát tác động của việc dọn dẹp đối với các quy tắc kết hợp với độ tin cậy thấp.

Việc áp dụng phương pháp sửa đổi đường viền lần đầu tiên được giới thiệu trong Divanis và Verykios (2006) để duy trì độ chính xác của dữ liệu đã được làm sạch, tức là số lần sửa đổi mục thực tế được giảm thiểu.Phương pháp này cũng đã được Divanis và Verykios (2009a) áp dụng để tạo ra các giao dịch mới được thêm vào cơ sở dữ liệu gốc cũng như Divanis và Verykios (2009b) để tìm ra giải pháp chính xác để ẩn các tập phổ biến nhạy cảm mà không có tác dụng phụ.

6.4.3.Lập trình số nguyên Một CSP (Russell & Norvig, 2003) được xác định bởi một tập hợp các biến,

một miền hữu hạn và rời rạc cho mỗi biến và một tập hợp các ràng buộc, trong đó mỗi biến có một miền khác trống của các giá trị tiềm năng.Các ràng buộc liên quan đến một tập hợp con của các biến để chỉ định sự kết hợp các giá trị được phép mà các biến này có thể đạt được.Một phép gán không vi phạm tập hợp các ràng buộc được gọi là "nhất quán".Mục tiêu là thỏa mãn tất cả các ràng buộc nhằm tối đa hóa hoặc giảm thiểu hàm mục tiêu chịu một số ràng buộc (Divanis & Verykios, 2009a; Kumar, 1992).Kỹ thuật quy hoạch số nguyên có thể được áp dụng để tìm giải pháp chính xác bằng cách sử dụng quy hoạch tuyến tính hoặc phi tuyến tính (Luenberger, 1973).Vì tất cả các biến trong bài toán vệ sinh đều là nhị phân nên kỹ thuật BIP (Gueret và cộng sự, 2002) có thể hữu ích trong việc chuyển CSP thành bài toán tối ưu hóa.

CSP cố gắng thu hẹp khoảng cách giữa cơ sở dữ liệu gốc và phiên bản đã được làm sạch của nó, nói cách khác, mục tiêu là duy trì độ chính xác của dữ liệu.Trong định nghĩa cơ bản về độ chính xác, độ chính xác của một quan hệ là tỷ lệ các bộ dữ liệu chính xác trong quan hệ, trong đó một bộ dữ liệu được cho là chính xác khi và chỉ khi mọi giá trị thuộc tính trong bộ dữ liệu đều chính xác (Reddy & Wang, 1995).

Menon và cộng sự.(2005) định nghĩa độ chính xác là số lượng giao dịch không được vệ sinh.Họ tin rằng độ chính xác của dữ liệu và tiện ích dữ liệu của cơ sở dữ liệu đã được vệ sinh có thể được tối đa hóa bằng cách giảm thiểu số lượng giao dịch được vệ sinh.Về vấn đề này, Menon et al. (2005), Menon và Sarkar (2008).và Divanis và Verykios (20 06, 20 09a, 20 09b) coi quá trình ẩn náu là một bài toán CSP để tìm ra giải pháp vệ sinh tối ưu.

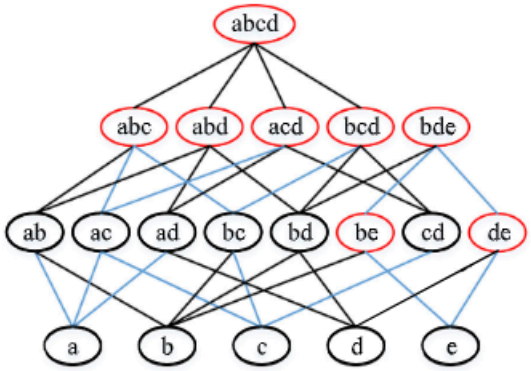
6.4.4.Tiêu chí được xác định trước Trong danh mục này, các mục nạn nhân được chọn dựa trên sự hỗ trợ

và các giao dịch được lựa chọn dựa trên độ dài.Có hai tiêu chí để lựa chọn dựa trên hỗ trợ: hỗ trợ tối thiểu và hỗ trợ tối đa.Nguyên nhân chính là do mặt hàng có độ hỗ trợ thấp nên có số lượng mẫu ít hơn;do đó, việc sửa đổi mục này sẽ gây ra ít tác động nhất đến các mẫu không nhạy cảm.Lý do đằng sau điều sau là các mẫu không nhạy cảm chứa mục có tần suất cao nhất có độ hỗ trợ cao và do đó các mẫu này bị ảnh hưởng tối thiểu bởi quá trình khử trùng.Thuật toán MinFIA (Oliveira & Zaiane, 2002) và DCR (Telikani & Shahbahrami, 2017) nhằm mục đích giảm tác động của từng sửa đổi bằng cách chọn các mục có độ hỗ trợ tối thiểu.Thuật toán Naïve (Oliveira & Zaiane, 2002) và Blanket (Menon et al., 2005) cũng tuân theo chiến lược này theo cách loại bỏ tất cả các mục của giao dịch nhạy cảm ngoại trừ mục có tần suất cao nhất.MaxFIA (Oliveira & Zaiane, 2002), SWA (Oliveira & Zaiane, 2003a), 2.b (Verykios et al., 2004b), GIH (Saygin et al., 2001, 2002), EMO-RH (Cheng et al. ., 2014; Cheng và cộng sự, 2016a), và Phân loại mức độ liên quan (Cheng và cộng sự, 2016b) chọn mục có tần suất cao nhất làm mục nạn nhân.

Chỉ có một tiêu chí để chọn các giao dịch trong đó các giao dịch có độ dài ngắn nhất được chọn để khử trùng.Giả định là những giao dịch này tạo ra ít mẫu thường xuyên hơn;kết quả là, ít quy tắc kết hợp được tạo ra hơn so với các giao dịch có độ dài dài nhất.Ví dụ, giao dịch {c, d} có độ dài ngắn nhất trong Bảng 1(a), ba tập phổ biến và một luật kết hợp được tạo ra từ giao dịch này.Mặt khác, giao dịch {a, b, c, d, e} có độ dài cao nhất và tất cả các mẫu đều được tạo từ giao dịch này.HCSRIL (Hai và cộng sự, 2013b), 1.b, 2.a (Dasseni và cộng sự, 2001; Verykios và cộng sự, 2004b), 2.b (Verykios và cộng sự, 2004b), DSC (Wang và cộng sự cộng sự, 2008), DCIS, DCDS (Wang và cộng sự, 2007b), ISL, DSR (Wang & Jafari, 2005; Wang và cộng sự, 2007a), CR, GIH (Saygin và cộng sự, 20 01, 20 02) và thuật toán SWA (Oliveira & Zaiane, 2003a) chọn các giao dịch có độ dài ngắn nhất.Thuật toán 1.a (Dasseni và cộng sự, 2001; Verykios và cộng sự, 2004b) và CR2 (Saygin và cộng sự, 2001, 2002) chọn các giao dịch không nhạy cảm hỗ trợ số lượng mục tối đa ở bên trái. -mặt trái của quy tắc nhạy cảm.

6.4.5.Mức độ xung đột Mức độ xung đột cho biết số lượng các mẫu được

bị ảnh hưởng nếu một mặt hàng hoặc giao dịch được vệ sinh.Nó có thể đo lường tác động của việc khử trùng đối với các mẫu không nhạy cảm hoặc trên các mẫu nhạy cảm.Phép đo đầu tiên trực tiếp kiểm soát tác động của việc dọn dẹp dữ liệu đối với các mẫu không nhạy cảm bằng cách chọn các sửa đổi có tác động tối thiểu.Phép đo thứ hai có hai mục tiêu: thứ nhất là đồng thời ẩn các mẫu nhạy cảm hơn và do đó những sửa đổi có tác động tối đa sẽ được chọn.Mục tiêu thứ hai là vệ sinh các giao dịch/vật phẩm với tác động tối thiểu.Giả định chính đằng sau mục tiêu này là mặt hàng hoặc giao dịch phụ thuộc vào ít mẫu nhạy cảm hơn cũng phụ thuộc vào ít mẫu không nhạy cảm hơn và do đó có thể tạo ra ít tác dụng phụ.Xem xét mức độ xung đột trên các mẫu không nhạy cảm, quá trình dọn dẹp được thực hiện với số lần lặp lớn hơn so với khi xem xét các mẫu nhạy cảm do số lượng mẫu không nhạy cảm.



Hình 14. Biểu đồ mạng giao nhau của các tập mục phổ biến trong Bảng 1(b).

nhận biến thường nhiều hơn số lượng các mẫu nhạy cảm.Trong trường hợp luật kết hợp, mức độ xung đột có thể được xem xét ở phía bên phải, bên trái hoặc cả hai phía.Về bên trái của các luật có liên quan đến chiến lược ẩn, ví dụ, khi chiến lược là sự rút gọn hệ quả, mức độ xung đột được xem xét ở phía bên phải của các luật kết hợp.

Trong giai đoạn lựa chọn giao dịch, các thuật toán MinFIA, MaxFIA, ngây thơ và RRA (Oliveira & Zaiane, 2003b) chọn các giao dịch có mức độ xung đột thấp nhất trên các mẫu nhạy cảm, trong khi MICF (Li & Chang, 2007), IGA (Oliveira & Thuật toán Zaiane, 2002) và RA (Oliveira & Zaiane, 2003b) chọn các giao dịch có mức độ xung đột cao nhất trên các mẫu nhạy cảm.Trong giai đoạn lựa chọn mục, các thuật toán MICF (Li & Chang, 2007), Intelligence (Menon và cộng sự, 2005), IGA (Oliveira & Zaiane, 2002) và SIF-IDF (Hong et al., 2013) chọn các mục với mức độ xung đột tối đa trên các mẫu nhạy cảm.Trong IGA, mục nạn nhân trong một quy tắc được cố định và bị xóa khỏi tất cả các giao dịch nhạy cảm liên quan đến tập mục nhạy cảm đó.Mặc dù điều này làm tăng hiệu quả của quá trình dọn dẹp nhưng nó có thể ảnh hưởng tối đa đến các tập mục không nhạy cảm liên quan đến mục nạn nhân.Không giống như các thuật toán nêu trên xem xét mức độ xung đột trên các mẫu nhạy cảm, thuật toán Phân chia (Amiri, 2007) xem xét mức độ xung đột của mục trên các tập mục không nhạy cảm và chọn mục có mức độ xung đột tối thiểu làm nạn nhân.Thuật toán 1.b (Dasseni và cộng sự, 2001; Verykios và cộng sự, 2004b) tính toán tác động của các mục lên tập mục (|Y| −1).Thuật toán 2.a (Dasseni và cộng sự, 2001; Verykios và cộng sự, 2004b) đo lường tác động của các mục trên các tập mục (|X∪Y| −1).Cả hai thuật toán 1.b và 2.a đều chọn một mục có tác động tối thiểu.COA4ARH (Afshari và cộng sự, 2016) chọn một mục nhạy cảm có tần suất xuất hiện cao nhất từ phía bên phải của các quy tắc nhạy cảm và tần suất thấp nhất trong các quy tắc không nhạy cảm.

6.4.6.Lưới giao nhau Lý thuyết mạng tinh thể (Grätzer, 2010) đã từng là Đầu tiên con nuôi qua Hải và cộng sự.(2012) để lựa chọn các đối tượng nạn nhân.Họ đã phân tích các đặc điểm của mạng lưới giao nhau của các tập mục phổ biến để giảm thiểu tác dụng phụ lên các tập mục phổ biến có độ hỗ trợ thấp.Các khái niệm cơ bản của việc lựa chọn dựa trên lý thuyết mạng được trình bày như sau.Đầu tiên, mạng giao nhau của tất cả các tập mục phổ biến được tạo ra và sau đó số siêu tập hợp của mỗi tập mục trong U được tính bằng hàm d (Z), trong đó Z là một tập mục phổ biến và U là tập hợp tất cả các tập mục phổ biến (Hai et cộng sự, 2013b).Tập mục W là tập siêu của tập mục Z khi Z⊆W.Hình 14 cho thấy biểu đồ mạng giao nhau của các tập phổ biến trong Bảng 1(b).

Theo tính chất Apriori, nếu Z và W ∈ U thì Z ∩W ∈ U. Có thể suy ra U là một mạng giao nhau.Bộ tạo (GS)

Bảng 6 Các tập phổ biến và GS (U).

bạn	Một	b	c	d	e	bụng	AC	quảng cáo	bc	bd	là	đĩa CD	của	abc	bụng	acd	bcd	bạn ơi	A B C D
S	3	4	3	3	2	2	2	2	2	3	1	2	1	1	1	1	1	0	0

của U, ký hiệu là GS (U), là tập mục nhỏ nhất của U sao cho mọi tập mục của U đều có thể được tạo ra bằng cách lấy giao của một số tập mục trong GS (U).Tập GS (U) có thể được tính bằng

$$GS(U) = \{Z \in U | d(Z) \leq 1\}, \text{ trong đó } d(Z) = |Z \in U | Z \subset W| \quad (4)$$

Số lượng siêu tập hợp, ký hiệu là s, được tính toán bởi hàm d(Z) và tập U được trình bày trong Bảng 6. Coatom (U) là tập mục tối đa của U, nói cách khác, tập mục có s = 0 là GS (U), Vì vậy,

$$GS(U) = \{be, de, abc, abd, acd, bcd, bde, abcd\}, \text{ và}$$
$$Coatom(U) = \max(GS(U)) = \{bde, abcd\}$$

Các tập mục chứa trong GS (U) có độ hỗ trợ thấp nhất trong U. Do đó, các tập mục này dễ bị giảm hỗ trợ của bất kỳ mục nào.Hơn nữa, nếu mọi tập mục của GS (U) đều phổ biến thì tất cả các tập mục của U cũng phổ biến.Lấy cảm hứng từ ý tưởng về mạng giao nhau, một số thuật toán đã được đề xuất để duy trì GS (U) trong quá trình ẩn nhằm hạn chế các luật bị mất.ILARH (Hai & Somjit, 2012), HCSRIL (Hai và cộng sự, 2013b), ARHIL (Hai và cộng sự, 2013a) và DIL (Hai và cộng sự, 2012) chỉ định các đối tượng nạn nhân dựa trên đặc điểm của lưới giao nhau của các tập phổ biến.

6.4.7.Trọng số Ưu tiên dựa trên trọng lượng là một phương pháp phỏng đoán để chọn giao dịch Và mặt hàng, Ở đâu các cân nặng của mỗi nhạy cảm giao dịch hoặc hạng mục được tính toán.Trong thuật toán WSDA (Pontikakis và cộng sự, 2004a; Verykios và cộng sự, 2007), trước tiên, trọng số được gán cho từng quy tắc nhạy cảm tùy theo mức độ gần với ngưỡng tin cậy tối thiểu.Sau đó, giá trị ưu tiên của mỗi giao dịch nhạy cảm được tính toán dựa trên trọng số và cuối cùng, các giao dịch nhạy cảm có mức độ ưu tiên thấp nhất sẽ được loại bỏ.

Verykios và cộng sự.(2007) đã áp dụng phương pháp lựa chọn này trong giai đoạn lựa chọn mục của thuật toán BA.Thuật toán DIL (Hai và cộng sự, 2012) và ARHIL (Hai và cộng sự, 2013a) ấn định trọng số cho mỗi giao dịch nhạy cảm dựa trên mức độ an toàn của nó, số lượng quy tắc nhạy cảm và số lượng quy tắc kết hợp không nhạy cảm có trong giao dịch đó.Không giống như các thuật toán dựa trên trọng số ở trên, DIL sẽ lọc các giao dịch có trọng số cao nhất.PDA (Pontikakis và cộng sự, 2004a) chọn một mục có tác động tối thiểu đến các quy tắc kết hợp nhạy cảm bằng cách chỉ định mức độ ưu tiên cho từng mục.

Hồng và cộng sự.(2013) đã cải thiện khái niệm TF-IDF (Salton và cộng sự, 1983) được sử dụng trong khai thác văn bản để ước tính mức độ giao dịch liên quan đến các tập mục nhạy cảm.Trong phương pháp này, mức độ tương quan giữa từng giao dịch nhạy cảm và các tập mục nhạy cảm được tính toán.Thuật toán SIF-IDF (Hong và cộng sự, 2013) ưu tiên các giao dịch hỗ trợ có trọng số sửa đổi cao nhất.Hạn chế lớn của thuật toán này là nó chỉ sử dụng thông tin trên các tập mục nhạy cảm có trong giao dịch, trong khi những tập mục không nhạy cảm không được xem xét mặc dù chúng có liên quan nhiều hơn đến các tác dụng phụ (Cheng và cộng sự, 2016a).Cheng và cộng sự.(2016b) đã xây dựng một phương pháp phỏng đoán để tính toán mức độ liên quan của từng giao dịch bằng cách xem xét xung đột giao dịch trên các tập mục tạo ra các quy tắc không nhạy cảm.Trong phương pháp này, các giao dịch có giá trị liên quan cao nhất sẽ được chọn để sàng lọc.Thật vậy, các giao dịch có mối quan hệ tối thiểu với các tập mục không nhạy cảm đã được loại bỏ.

Bảng 7 Tệp đảo ngược của Bảng 1(a).

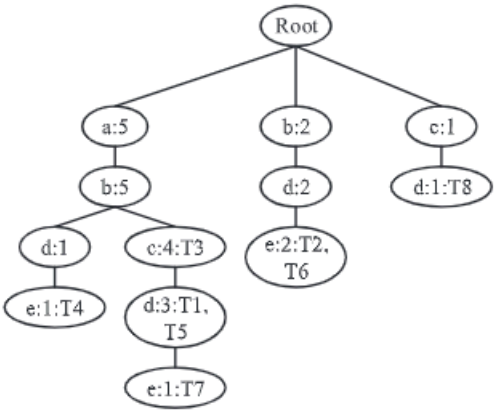
Mặt hàng	Tính thường xuyên		ID giao dịch
MỘT	5	→	T1, T3, T4, T5, T7
B	7	→	T1, T2, T3, T4, T5, T6, T7
C	5	→	T1, T3, T5, T7, T8
D	7	→	T1, T2, T4, T5, T6, T7, T8
E	4	→	T2, T4, T6, T7

6.5.Đẩy nhanh quá trình khử trùng

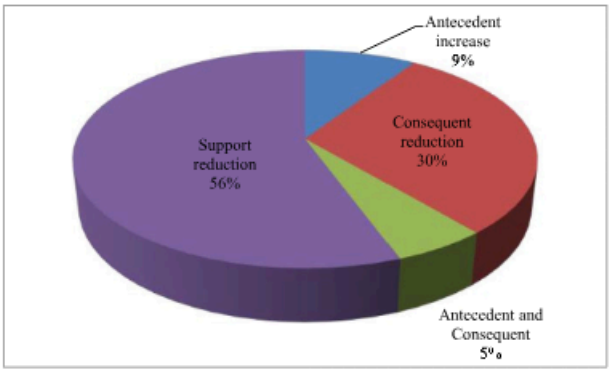
Mục tiêu thay thế của thuật toán dọn dẹp dữ liệu là giảm thời gian tính toán cần thiết cho quá trình ẩn.Có hai cách để tăng tốc quá trình vệ sinh;trong phương pháp đầu tiên, một giải pháp hiệu quả được xác định cho các giai đoạn giao dịch và lựa chọn mặt hàng.Hầu hết các thuật toán heuristic sử dụng cách này vì chúng gán giá trị cho các giao dịch dựa trên các heuristic khác nhau như độ dài, mức độ xung đột hoặc trọng lượng.Vì lý do này, trước tiên, họ sắp xếp các giao dịch theo thứ tự tăng dần/giảm dần về giá trị của các giao dịch trong O(S), trong đó S là số lượng giao dịch nhạy cảm, sau đó các giao dịch được chọn lần lượt từ đầu danh sách. .Như vậy, tốc độ lựa chọn của mỗi giao dịch được tăng từ O(S) lên O(1).Trong phương pháp thứ hai, một kỹ thuật độc lập được áp dụng để giảm việc quét cơ sở dữ liệu nhằm xác định các giao dịch nhạy cảm.Kỹ thuật này có thể cộng tác với quy trình khử trùng hoặc có thể được nhúng vào quy trình khử trùng.Sau đây, các kỹ thuật khác nhau được đề xuất cho cách thứ hai sẽ được thảo luận.

Oliveira và Zaiane (20 02, 20 03b) đã giới thiệu một công cụ truy xuất giao dịch dựa trên một tệp đảo ngược để truy xuất ID giao dịch từ cơ sở dữ liệu.Cơ sở dữ liệu giao dịch được lập chỉ mục thành một tệp đảo ngược, trong đó, đối với mỗi mục của cơ sở dữ liệu, có một danh sách ID giao dịch tương ứng liên quan đến mục đó.ID giao dịch được sắp xếp theo thứ tự tăng dần của ID giao dịch.Do đó, trong trường hợp xấu nhất, ID giao dịch được tìm thấy bằng cách sử dụng tìm kiếm nhị phân với thời gian truy cập là O (logN), trong đó N là số ID giao dịch trong danh sách.Từ vựng của tệp đảo ngược bao gồm tất cả các mục khác nhau trong cơ sở dữ liệu giao dịch được triển khai dựa trên bảng băm hoàn hảo (Dietzfelbinger et al., 1994).Bảng 7 hiển thị tệp đảo ngược cho Bảng 1(a).Các thuật toán RRA, RA (Oliveira & Zaiane, 2003b), MaxFIA, MinFIA, Naïve và IGA (Oliveira & Zaiane, 2002) sử dụng kỹ thuật này.

Cấu trúc dữ liệu cây đảo mẫu (PI-tree) (Wang, 2009; Wang và cộng sự, 2008) là phần mở rộng của kỹ thuật cây mẫu (Huang và cộng sự, 2002) để giảm số lần quét cơ sở dữ liệu.Mỗi nút trong cây PI chứa ba trường: tên mục, số lượng giao dịch chứa các mục trên đường dẫn từ nút gốc đến nút hiện tại và danh sách ID giao dịch chứa tất cả các mục trên đường dẫn từ nút gốc đến nút hiện tại. nút.Các giao dịch trước tiên được đọc từ cơ sở dữ liệu từng cái một.Mỗi giao dịch được sắp xếp theo tên mục và được chèn vào cây PI.Danh sách tần số được cập nhật tương ứng.Sau đó, danh sách tần số được sắp xếp theo số lượng hỗ trợ của các mục.Cuối cùng, cây PI được cơ cấu lại tương tự như bước đầu tiên.Hình 15 trình bày cây PI cho các giao dịch trong Bảng 1(a).DSC (Wang và cộng sự, 2008) và MSI (Wang, 2009) sử dụng cây PI để che giấu kiến thức nhạy cảm chỉ bằng một lần quét cơ sở dữ liệu.



Hình 15. Cây đào ngược mẫu cho tập dữ liệu trong Bảng 1(a).



Hình 16. Phân phối thuật toán ẩn luật kết hợp dựa trên chiến lược ẩn.

Để đẩy nhanh quá trình tiến hóa trong các thuật toán dựa trên tiến hóa, pGA2DT (Lin và cộng sự, 2015) và PSO2DT (Lin và cộng sự, 2016) áp dụng khái niệm tiền lớn (Hong và cộng sự, 2001) để giảm bớt thời gian thực hiện quét lại cơ sở dữ liệu gốc trong quá trình đánh giá.Sự tối ưu hóa này bao gồm việc duy trì một bộ đệm gồm các tập mục có kích thước lớn trong quá trình tiến hóa để tránh thực hiện nhiều lần quét cơ sở dữ liệu.Lin và cộng sự.(2014b) đã áp dụng thuộc tính đóng xuống (Liu và cộng sự, 2005) và khái niệm tiền lớn để giảm chi phí quét lại cơ sở dữ liệu.Phương pháp GA nhỏ gọn (Harik và cộng sự, 1999) đã được áp dụng trong cpGA2DT (Lin và cộng sự, 2014a) để chỉ tạo ra hai cá thể trên mỗi quần thể để cạnh tranh nhằm giảm mức sử dụng bộ nhớ.

MICF (Li & Chang, 2007) ban đầu tải tất cả các giao dịch nhạy cảm vào bộ nhớ chính.Do đó, các giao dịch được vệ sinh trong bộ nhớ chính thay vì trên đĩa.MICF áp dụng bảng tra cứu chỉ mục để sắp xếp hiệu quả các giao dịch nhạy cảm và giảm yêu cầu về dung lượng bộ nhớ.Mỗi giao dịch được liên kết với một cặp gồm hai giá trị trong bảng tra cứu: giá trị đầu tiên là giá trị chỉ mục trỏ đến giao dịch và giá trị thứ hai là giá trị mức độ xung đột của giao dịch.Khi kích thước của các giao dịch nhạy cảm vượt quá dung lượng bộ nhớ khả dụng hiện tại, MICF phải chịu rất nhiều sự hoán đổi trang giữa đĩa và bộ nhớ chính.Trong trường hợp này, kỹ thuật phân vùng được kết hợp với MICF ban đầu để xử lý cơ sở dữ liệu rất lớn.Trong quá trình dọn dẹp, chỉ các giao dịch nhạy cảm của phân vùng hiện tại mới được tải vào bộ nhớ chính.Sau đó, MICF ban đầu được áp dụng để vệ sinh phân vùng.Nếu số lượng hỗ trợ của các tập mục nhạy cảm không giảm xuống dưới ngưỡng quyền riêng tư thì thuật toán sẽ tải các giao dịch nhạy cảm của phân vùng tiếp theo.

Cấu trúc dữ liệu ba cây (Bodon, 2005) đã được Cheng và cộng sự áp dụng.(2014, 2016a) để lưu trữ số lượng hỗ trợ của các tập mục thường xuyên nhằm tăng tốc độ truy cập và cập nhật số lượng hỗ trợ của các tập mục này.Mảng hai chiều trước tiên được sử dụng để lưu trữ sự hỗ trợ của các quy tắc có hai mục và sau đó cây trie được áp dụng để dự trữ và truy xuất các tập mục tạo tương ứng của các quy tắc có nhiều hơn hai mục.Khi mảng được sử dụng, dữ liệu có thể được truy cập trực tiếp theo cách ánh xạ và do đó không cần truy xuất toàn bộ quy tắc đã đặt để tìm quy tắc cập nhật hỗ trợ cho mảng đó.

Wu và cộng sự.(2007) đã sử dụng hai kỹ thuật để tăng tốc quá trình dọn dẹp, (1) cơ sở dữ liệu gốc được biểu diễn dưới dạng vectơ bit trong đó mỗi mục riêng biệt được mã hóa dưới dạng một số nguyên tố duy nhất. (2) Chỉ mục quy tắc giao dịch được xây dựng bằng cách sử dụng khái niệm danh sách đảo ngược (Kowalski & Maybury, 2006) để tương quan với các bảng nhằm truy xuất hiệu quả.

Thuật toán BA mở rộng (Verykios et al., 2007) sử dụng một số cấu trúc dữ liệu để truy cập các quy tắc trong cơ sở dữ liệu.Nó tạo ra bốn bảng, bao gồm một bảng chỉ mục đảo ngược và ba bảng băm.(1) Bảng chỉ mục đảo ngược được tạo cho mọi mục của cơ sở dữ liệu sao cho thuật toán sắp xếp hợp nhất được sử dụng để tìm sự hỗ trợ của các tập mục lớn.(2) Đối với mỗi quy tắc, quy tắc cùng với số lượng giao dịch hỗ trợ vượt quá ngưỡng tin cậy sẽ được lưu trữ.(3) Các tập mục lớn và giá trị hỗ trợ của chúng được lưu trữ để phục hồi nhanh chóng và (4) các quy tắc không nhạy cảm với độ tin cậy thấp được lưu trữ cùng với số lượng giao dịch có dưới ngưỡng tối thiểu.

7. Thảo luận và phân tích

Bốn khái niệm liên quan đến việc dọn dẹp dữ liệu, bao gồm chiến lược ẩn, kỹ thuật dọn dẹp, phương pháp dọn dẹp và phương pháp lựa chọn/tạo được sử dụng bởi một tập hợp các thuật toán ẩn quy tắc kết hợp để che giấu kiến thức nhạy cảm.Phần này trình bày phân tích thống kê về việc áp dụng các hướng này trong tất cả 54 thuật toán data sanitization cũng như so sánh ưu điểm và nhược điểm của từng hướng.Đặc điểm của các hướng này đã được mô tả rộng rãi và độc lập với nhau trong Phần 6.1–6.4.

7.1.Phân tích bằng cách ẩn chiến lược

Bảng 8 tóm tắt các đặc điểm của chiến lược vệ sinh được thảo luận trong Phần 6.1.Chiến lược giảm hỗ trợ làm giảm chất lượng của kết quả khai phá luật kết hợp theo hướng làm mất đi các mẫu không nhạy cảm và tạo ra các quy tắc mới, không có tập mục mới.Trong vấn đề ẩn quy tắc, chiến lược giảm thiểu hệ quả sẽ loại bỏ ít giao dịch hơn so với chiến lược tăng trước đó;do đó, độ chính xác dữ liệu của nó cao.Bất lợi của chiến lược tăng trước đó là nó không phải lúc nào cũng che giấu được tất cả các quy tắc nhạy cảm.Nói chung, chiến lược tăng trước kém hơn chiến lược giảm sau.Việc áp dụng đồng thời các chiến lược giảm hệ quả và tăng tiền đề có thể dẫn đến ít quy tắc bị mất hơn và các quy tắc mới hơn so với khi hai chiến lược được áp dụng riêng biệt.

Hình 16 trình bày sự phân bố của thuật toán bằng cách ẩn chiến lược.Từ 54 thuật toán dọn dẹp dữ liệu, hầu hết trong số đó, 56% (30 trên 54), sử dụng chiến lược giảm hỗ trợ để ẩn các mẫu nhạy cảm.Trong số 30 thuật toán này, có 7 thuật toán dựa trên ẩn quy tắc và 23 thuật toán dựa trên ẩn tập mục.Mặt khác, 24 (44%) thuật toán áp dụng chiến lược giảm độ tin cậy để ẩn các quy tắc nhạy cảm.Hầu hết các thuật toán này làm giảm độ tin cậy bằng cách giảm số lượng hỗ trợ của hệ quả quy tắc, 16 trên 24, sau đó là mức tăng trước đó.

Bảng 8 So sánh chiến lược ẩn để ẩn luật kết hợp.

Phương pháp	Thuận lợi	Nhược điểm
Giảm hỗ trợ	Nó không tạo ra bất kỳ tập mục nhân tạo nào.	Tiện ích dữ liệu trong việc khai thác các quy tắc kết hợp từ cơ sở dữ liệu đã được làm sạch bị giảm đi.
Hậu quả giảm	Không có sự che giấu thất bại. Độ chính xác của dữ liệu cao hơn chiến lược tăng trước đó.	–
Tăng trước đó	–	Nó có thể thất bại trong việc che giấu tất cả các quy tắc nhạy cảm.Tỷ lệ vệ sinh cao.
Hậu quả là giảm với tăng trước đó	Số lượng quy tắc bị mất và quy tắc mới thấp.	–

Bảng 9 So sánh các kỹ thuật ẩn luật kết hợp.

Phương pháp	Thuận lợi	Nhược điểm
Méo mó	Rủi ro tiết lộ các mẫu nhạy cảm là thấp.	–
Chặn	Nó không thêm bất kỳ thông tin sai lệch nào vào dữ liệu.	Nó có thể làm mờ độ hỗ trợ và độ tin cậy của luật kết hợp.Nó không thể đảm bảo việc bảo vệ các mẫu nhạy cảm.
Xóa/chèn giao dịch	Không cần chọn mục nạn nhân	Nó thay đổi số lượng giao dịch.

Bảng 10 Phân phối các thuật toán theo kỹ thuật vệ sinh và năm xuất bản.

Sanitization technique	2001	2002	2003	2004	2005	2006	2007	2008	2009	2012	2013	2014	2015	2016	2017	Total
Distortion	3	4	3	3	3	3	8	2	2	2	3	1	0	2	1	**40**
Blocking	3	0	0	1	2	0	1	0	0	0	0	0	0	0	0	**7**
Transaction deletion/insertion	0	0	0	0	0	0	1	0	1	0	0	2	2	1	0	**7**
Total	**6**	**4**	**3**	**4**	**5**	**3**	**10**	**2**	**3**	**2**	**3**	**3**	**2**	**3**	**1**	**54**

chiến lược, 5 trên 24, và sự kết hợp giữa chiến lược giảm thiểu và tăng trước đó, 3 trên 24.

7.2.Phân tích bằng kỹ thuật vệ sinh

Bảng 9 trình bày sự so sánh giữa các phương pháp vệ sinh khác nhau

kỹ thuật.Trong kỹ thuật chặn, đối thủ có thể tiết lộ các quy tắc ẩn bằng cách xác định các tập mục được tạo có chứa dấu chấm hỏi và sau đó anh ta/cô ta có thể thay thế các dấu chấm hỏi bằng các mục thực tế.Mặt khác, do kỹ thuật bóp méo đảo ngược các mục của cơ sở dữ liệu nhị phân hoặc xóa/chèn các mục từ/vào cơ sở dữ liệu phân loại nên rủi ro tiết lộ các mẫu nhạy cảm sẽ giảm.Một nhược điểm quan trọng của kỹ thuật chèn/xóa giao dịch là nó thay đổi số lượng giao dịch của cơ sở dữ liệu sao cho tầm quan trọng của các mẫu có ý nghĩa và vô dụng có thể bị giảm bớt.

7.3.Phân tích bằng phương pháp vệ sinh

ảnh hưởng.

Bảng 10 thể hiện sự phân bố của các thuật toán dọn dẹp dựa trên kỹ thuật dọn dẹp và năm xuất bản của chúng từ 2001 đến 2017. Từ bảng này có thể thấy rằng kỹ thuật dọn dẹp được sử dụng phổ biến nhất là biến dạng, 74% (40 trên 54), trong khi chặn kỹ thuật không được sử dụng để sửa đổi cơ sở dữ liệu trong những năm gần đây (từ 2007 đến 2017).Mặt khác, kỹ thuật chèn/xóa giao dịch gần đây đã được áp dụng để vệ sinh cơ sở dữ liệu, đặc biệt là từ năm 2014 đến 2016.

Bảng 11 trình bày những lợi ích và hạn chế của các phương pháp vệ sinh.Vì các cách tiếp cận đường viền và chính xác che giấu một quy tắc nhạy cảm bằng cách ẩn tập mục tạo ra nó, nên chúng không đạt được kết quả tốt khi ẩn một tập hợp quy tắc so với các phương pháp heuristic và tiến hóa.Các

Sự khác biệt chính giữa cách tiếp cận tiến hóa và các cách tiếp cận khác là các thuật toán dựa trên tiến hóa hình thành chức năng thích hợp bằng cách xem xét tất cả ba tác dụng phụ của quá trình dọn dẹp để tìm ra các giao dịch tốt nhất cho việc dọn dẹp, trong khi các thuật toán khác xem xét chi phí bỏ lỡ để chọn một giao dịch hoặc vật phẩm.Ngoài ra, cách tiếp cận tiến hóa có thể không phải lúc nào cũng đáp ứng được điều kiện chính của quá trình vệ sinh.Thật vậy, các thuật toán này có thể thất bại trong việc che giấu tất cả các mẫu nhạy cảm nhưng chúng làm giảm số lượng quy tắc bị mất và quy tắc ma.Mặt khác, mục tiêu chính của ba phương pháp còn lại là che giấu tất cả các mẫu nhạy cảm và việc giảm MC và AP được coi là mục tiêu tiếp theo.Cách tiếp cận biên giới cố gắng duy trì tiện ích dữ liệu trong khi cách tiếp cận chính xác tập trung vào việc duy trì độ chính xác của dữ liệu.Mặc dù phương pháp heuristic không thể tìm ra giải pháp tối ưu để duy trì tiện ích của cơ sở dữ liệu đã được dọn dẹp nhưng nó thường có thể thực hiện quá trình dọn dẹp một cách hiệu quả.

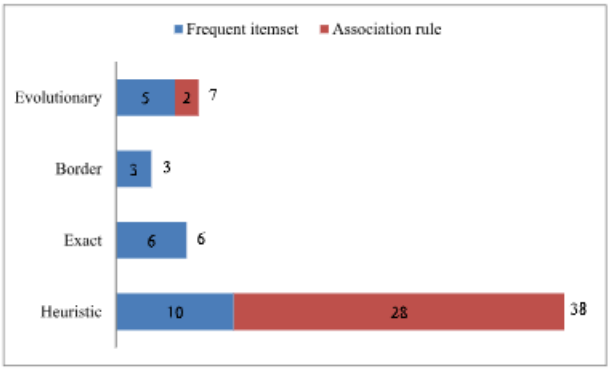
Hình 17 cho thấy sự phân bố của các thuật toán theo phương pháp làm sạch và loại mẫu.Phân tích này cho thấy rằng phần lớn các thuật toán, 70% (38 trên 54), dựa trên heuristic, tiếp theo là dựa trên tiến hóa, dựa trên chính xác và dựa trên đường viền bao gồm 7, 6 và 3 trên 54, tương ứng.Tuy nhiên, sự phân loại này không thể tuyệt đối vì hầu hết các thuật toán không heuristic là kết hợp và thường sử dụng lý thuyết heuristic hoặc lý thuyết biên để lựa chọn.Ví dụ: Chăn, Thông minh (Menon và cộng sự, 2005), Nội tuyến (Divanis & Verykios, 2006), thuật toán được đề xuất trong Menon và Sarkar (2008) và thuật toán Kết hợp (Divanis & Verykios, 2009a), được phân loại trong danh mục chính xác, tích hợp các giải pháp khác trong giai đoạn lựa chọn.Hai thuật toán đầu tiên sử dụng heuristic để chọn mục nạn nhân và ba thuật toán còn lại áp dụng lý thuyết biên để chọn giao dịch.Hai thuật toán dựa trên tiến hóa, EMO-RH (Cheng và cộng sự, 2014, 2016a) và COA4ARH (Afshari và cộng sự, 2016) sử dụng phương pháp phỏng đoán để chọn các mục nạn nhân.Tuy nhiên, chúng tôi đã tuân theo cách phân loại dựa trên cách tiếp cận trước

Bảng 11 So sánh các phương pháp dọn dẹp để ẩn luật kết hợp.

Phương pháp	Thuận lợi	Nhược điểm
Tự tìm tòi	Hiệu quả và độ ổn định cao.	Chất lượng dữ liệu kém hơn các phương pháp khác.
	Chất lượng kết quả của việc ẩn các luật kết hợp không chỉ là việc ẩn tập mục thông thường.	Nó không đảm bảo giải pháp tối ưu.
Ranh giới	Tác động của quá trình ẩn đến chất lượng kết quả bị giảm.	–
Chính xác	Độ chính xác của dữ liệu cao.	Độ phức tạp tính toán rất cao.Nó chỉ tập trung vào độ chính xác của dữ liệu.
tiến hóa	Tất cả ba tác dụng phụ đều được xem xét trong định nghĩa về chức năng thể lực.	Nó có thể thất bại trong việc che giấu tất cả các mẫu nhạy cảm.

Bảng 12 So sánh các phương pháp lựa chọn để ẩn luật kết hợp.

Phương pháp	Thuận lợi	Nhược điểm
Dân số	–	Nó thất bại trong việc che giấu một số mẫu.Nó chỉ có thể được áp dụng trong bước lựa chọn giao dịch.
Sửa đổi đường viền	Chất lượng kết quả cao.	–
Lập trình số nguyên	Độ chính xác của dữ liệu cao	Thời gian tính toán cao.Kích thước của không gian tìm kiếm cao.
Tiêu chí được xác định trước	Thời gian tính toán thấp.	Nó lựa chọn dựa trên giả thuyết.
Mức độ xung đột	Nó tính toán tác động của từng mặt hàng/giao dịch lên các mẫu.	–
Lưới giao nhau	–	–
Dựa trên trọng lượng	–	–



Hình 17. Phân phối các thuật toán ẩn dựa trên các phương pháp dọn dẹp và loại mẫu.

Bảng 13 Phân bổ các phương pháp lựa chọn được sử dụng trong quy trình vệ sinh dựa trên giao dịch và lựa chọn mặt hàng.

Phương pháp	Lựa chọn giao dịch	Lựa chọn mục
Dân số	5	0
Sửa đổi đường viền	4	3
Lập trình số nguyên	6	4
Tiêu chí được xác định trước	17	11
Mức độ xung đột	11	9
Lưới giao nhau	0	4
Dựa trên trọng lượng	6	1
Tổng cộng	49	32

gửi trong các cuộc khảo sát khác (Verykios & Divanis, 2008; Divanis & Verykios, 2010; Verykios, 2013).Như được hiển thị trong Hình 17, hầu hết các thuật toán dựa trên heuristic đã được thiết kế để ẩn các quy tắc kết hợp, 28 trên 38. Tất cả các thuật toán dựa trên đường viền và dựa trên chính xác đều ẩn các tập phổ biến.

7.4.Phân tích bằng phương pháp chọn lọc/tạo

Bảng 12 tóm tắt những ưu điểm và nhược điểm của các phương pháp chọn lọc/tạo ra.Do các phương pháp được xác định trước thực hiện giai đoạn lựa chọn thông qua một số giả định nên nó đòi hỏi ít thời gian tính toán hơn các phương pháp khác.Việc sửa đổi đường viền và lập trình số nguyên được sử dụng tương ứng trong các phương pháp tiếp cận đường viền và chính xác, do đó các phương pháp này mang các đặc điểm của chúng, bao gồm độ phức tạp cao, tiện ích dữ liệu tối đa và độ chính xác dữ liệu cao.

Bảng 13 cho thấy sự phân bố của các phương pháp lựa chọn/tạo trong thuật toán dọn dẹp theo tần suất sử dụng trong các giai đoạn lựa chọn.Từ 54 thuật toán, các phương pháp lựa chọn đã được sử dụng trong 49 thuật toán để chọn các giao dịch để khử trùng trong khi chúng được áp dụng trong 32 thuật toán để chọn các mục nạn nhân.Điều này cho thấy việc lựa chọn giao dịch nhận được nhiều sự quan tâm hơn so với việc lựa chọn mục nạn nhân vì việc sửa đổi giao dịch ảnh hưởng nhiều hơn đến tiện ích của cơ sở dữ liệu được làm sạch hơn là sửa đổi mục nạn nhân.Các tiêu chí được xác định trước là phương pháp phỏng đoán được sử dụng rộng rãi nhất trong các thuật toán dọn dẹp, trong bước lựa chọn giao dịch của 17 thuật toán và trong bước chọn mục nạn nhân của 11 thuật toán.Tiếp theo là mức độ xung đột, được sử dụng trong bước chọn giao dịch của 11 thuật toán và trong bước chọn mục nạn nhân của 9 thuật toán dọn dẹp.Các phương pháp dựa trên dân số chưa bao giờ được sử dụng trong giai đoạn lựa chọn mục nạn nhân vì hầu hết các thuật toán dựa trên tiến hóa đều sử dụng kỹ thuật chèn/xóa giao dịch, do đó, chúng không cần chọn mục nạn nhân.Phương pháp dựa trên mạng giao nhau chỉ được sử dụng trong giai đoạn lựa chọn mục nạn nhân.

Phân tích của chúng tôi cho thấy rằng việc lựa chọn mục có tần suất cao nhất và cả lựa chọn giao dịch có thời lượng ngắn nhất là những phương pháp phỏng đoán dựa trên được xác định trước được sử dụng nhiều nhất, lần lượt là 8 trên 11 và 16 trên 17 thuật toán.Điều này có thể là do các giao dịch/mục này tạo ra ít mẫu hơn, do đó cơ sở dữ liệu được phát hành có thể ít bị ảnh hưởng hơn bởi quá trình dọn dẹp.Do đó, những phương pháp phỏng đoán này đạt được sự cân bằng tốt giữa hiệu quả và tiện ích.

Bảng 14 cho thấy số lượng việc sử dụng phương pháp phỏng đoán dựa trên mức độ xung đột để lựa chọn.Rõ ràng là trong cả hai bước lựa chọn vật phẩm và giao dịch, mức độ xung đột trên các mẫu nhạy cảm là phương pháp được sử dụng rộng rãi nhất, tiếp theo là tính toán mức độ xung đột trên cả các mẫu không nhạy cảm tối đa và không nhạy cảm tối thiểu.Mặt khác, mức độ xung đột trên các mẫu không nhạy cảm chỉ được sử dụng bởi một thuật toán dọn dẹp làm tiêu chí lựa chọn.

Bảng 14 Sự phân bố của các phương pháp dựa trên Mức độ Xung đột (CD) trong các thuật toán dọn dẹp.

Bước chân	Phương pháp	Con số
Lựa chọn giao dịch	CD về các mẫu nhạy cảm	7
	CD trên các mẫu không nhạy cảm	0
	CD ở mức nhạy cảm tối đa và không nhạy cảm tối thiểu	4
	CD trên tất cả các mẫu	0
	Tổng cộng	11
Lựa chọn mục	CD về các mẫu nhạy cảm	4
	CD trên các mẫu không nhạy cảm	1
	CD ở mức nhạy cảm tối đa và không nhạy cảm tối thiểu	3
	CD trên tất cả các mẫu	2
	Tổng cộng	10

8. Tiêu chuẩn đánh giá

Trong việc ẩn quy tắc kết hợp, điều quan trọng là phải đánh giá tác dụng phụ và tác động cơ sở dữ liệu do quá trình dọn dẹp tạo ra.Do đó, cần phải xác định một loạt các biện pháp cho mục đích này.Các tác dụng phụ được đo lường dựa trên những thất bại trong việc che giấu, các quy tắc mới và các quy tắc bị mất.Hiệu ứng cơ sở dữ liệu được đo lường ở hai cấp độ: cấp độ giao dịch và cấp độ vật phẩm.Ở cấp độ giao dịch, tỷ lệ phần trăm của các giao dịch bị thay đổi được đo lường và ở cấp độ hạng mục, tỷ lệ phần trăm tần suất của các mục bị thay đổi được đo lường (Amiri, 2007; Wang và cộng sự, 2008).Những đánh giá này có thể được thực hiện bởi các quản trị viên cơ sở dữ liệu để xác định xem quy trình ẩn có thể đáp ứng mục tiêu của họ hay bởi các nhà nghiên cứu để ước tính tác động của thuật toán được thiết kế mới bằng cách sử dụng một số bộ dữ liệu giao dịch tiêu chuẩn so với các thuật toán khác.Các thước đo và tập dữ liệu giao dịch sẽ được thảo luận trong các phần sau.

8.1.Đo

Chi phí bỏ lỡ (MC), Lỗi ẩn (HF) và Mẫu giả (AP) được đo bằng các phương trình.(5), (6) và (7), tương ứng.MC được đo bằng tỷ lệ phần trăm các mẫu hợp pháp không được phát hiện từ cơ sở dữ liệu đã được làm sạch (D′).HF được đo bằng phần trăm các quy tắc nhạy cảm được phát hiện từ D′.AP được đo bằng phần trăm các quy tắc không có trong cơ sở dữ liệu gốc nhưng được phát hiện từ cơ sở dữ liệu đã được lọc sạch.

$$MC = \#€PS\ (I\frac{(D')}{\#€PS\ (D)} \quad (5)$$

$$HF = \#\#PS\ (D') \quad (6)$$

$$AP = |P| - |P \cap P'| \quad (7)$$

Trong đó #PS (D) và #PS (D′) lần lượt là số lượng mẫu không nhạy cảm hiện có trong cơ sở dữ liệu gốc và cơ sở dữ liệu đã được làm sạch, # PS (D′) là số lượng mẫu nhạy cảm được phát hiện từ cơ sở dữ liệu đã được làm sạch và # PS (D) là số lượng mẫu nhạy cảm hiện có trong cơ sở dữ liệu gốc.

Mục tiêu thứ hai của quá trình dọn dẹp là duy trì độ chính xác của dữ liệu nhằm giảm thiểu số lượng thay đổi trong cơ sở dữ liệu đã được dọn dẹp.Nếu độ chính xác của dữ liệu bị suy giảm đáng kể thì cơ sở dữ liệu đã được làm sạch sẽ trở nên vô dụng cho mục đích trích xuất kiến thức.Ở cấp độ giao dịch, độ chính xác có thể được đo lường bằng số lượng giao dịch không được kiểm soát (Menon và cộng sự, 2005).Để đánh giá độ chính xác của dữ liệu ở cấp độ mục, sự khác biệt giữa tập dữ liệu gốc và tập dữ liệu đã được làm sạch thường được sử dụng.Sự khác biệt được đo bằng phần trăm tần số của các mục của hai bộ dữ liệu trước và sau khi dọn dẹp, như được biểu thị trong phương trình dưới đây

$$\text{Bất đồng } (D, D') = \frac{\sum_{\text{tôi}=1} |fD(i) - fD'(i)|}{\sum_{ni=1} fD(i)} \quad \text{(số 8)}$$

Trong đó n là số mục trong tập dữ liệu, fD (i) là tần suất của mục i trong tập dữ liệu gốc và fD′ (i) là tần số của mục i trong tập dữ liệu đã được chọn lọc.Bertino và cộng sự.(2005) đã đề xuất một khung đánh giá để đo lường tính hiệu quả của các thuật toán ẩn quy tắc kết hợp cũng như các loại nhiệm vụ PPDM khác.Các thước đo cũng đã được nghiên cứu rộng rãi trong tài liệu (Menon & Sarkar, 2008; Verykios & Divanis, 2008).

8.2.Bộ dữ liệu

Nhiều bộ dữ liệu thực được coi là để đánh giá tính hiệu quả của thuật toán vệ sinh dữ liệu.Các bộ dữ liệu này được thu thập trong hội thảo đầu tiên về Triển khai khai thác tập mục thường xuyên (FIMI) và có sẵn thông qua kho FIMI.1 Chúng đã được mô tả chi tiết hơn trong Goethals và Zaki (2003) và Menon et al.(2005).Các bộ dữ liệu này thể hiện các đặc điểm khác nhau về số lượng giao dịch, số lượng mặt hàng và thời lượng giao dịch trung bình (Telikani & Shahbahrami, 2017).Những đặc điểm này được mô tả lần lượt ở cột thứ ba, thứ tư và thứ năm của Bảng 15.Trong số các bộ dữ liệu được mô tả, các cơ sở dữ liệu bms-pos, Bmsl, Bms2 (Zheng và cộng sự, 1999), Kosarak (Bodon, 2003) và Cơ sở dữ liệu bán lẻ (Brijs, 2003) rất thưa thớt và Tai nạn (Geurts và cộng sự, 2003).), Bộ dữ liệu Connect, Chess và Mushroom (Byardo, 1998) dày đặc hơn.Mật độ của tập dữ liệu, độ dài giao dịch trung bình chia cho số mục, ảnh hưởng đến hiệu quả của thuật toán ẩn quy tắc kết hợp.Như đã chỉ ra trong Menon et al. (2005), bộ dữ liệu rất dày đặc không đại diện cho dữ liệu giao dịch trong thực tế.Các bộ dữ liệu thưa thớt thường được quan sát phổ biến hơn trong kịch bản thế giới thực.Amiri (2007) đã chứng minh rằng mật độ của cơ sở dữ liệu dường như không có tác động đáng kể đến độ chính xác của dữ liệu của cơ sở dữ liệu đã được làm sạch, đặc biệt là ở cấp độ mục.Cột thứ sáu của Bảng 15 trình bày mật độ của các tập dữ liệu thực.Bộ dữ liệu Cờ vua có mật độ cao nhất (0,493), tiếp theo là bộ dữ liệu Kết nối, Nấm và Tai nạn với lần lượt là 0,33, 0,19 và 0,08.Các bộ dữ liệu khác có mật độ rất thấp, ví dụ: bộ dữ liệu Kosarak là bộ dữ liệu thưa thớt nhất (0,0 0 02).

Tập dữ liệu tổng hợp là một loại tập dữ liệu khác được tạo bởi Trình tạo dữ liệu tổng hợp của IBM.2 Cột thứ bảy

trong Bảng 15 trình bày số lần một tập dữ liệu đã được sử dụng để đánh giá các thuật toán dọn dẹp.Từ 42 bài báo, có 21 bài sử dụng bộ dữ liệu tổng hợp để đánh giá hiệu suất.Điều này là do các nhà nghiên cứu có thể tạo ra một tập dữ liệu với số lượng giao dịch và mặt hàng một cách tùy ý.Ngược lại,

1 http://fimi.cs.helsinki.fi 2 http://www.almaden.ibm.com/cs/quest/syndata.html

Bảng 15 Đặc điểm của cơ sở dữ liệu thực.

Số.	Tên	#Giao dịch	#Mặt hàng	Trung bìnhDịch.Chiều dài	Tỉ trọng	#Cách sử dụng
1	bms-pos	515.597	1657	7 giờ 50	0,0045 (Thưa thớt)	1
2	Bms1	59.602	497	2,5	0,005 (Thưa thớt)	15
3	Bms2	77.512	3340	5.6	0,0016 (Thưa thớt)	13
4	Giỏ	990.002	41.217	8.10	0,0 0 02 (Thưa thớt)	2
5	Bán lẻ	88.162	16.470	10h30	0,0 0 06 (Thưa thớt)	số 8
6	Tai nạn	340.183	468	33,80	0,08 (Dày đặc)	1
7	Kết nối	67.557	129	43:00	0,33 (Dày đặc)	1
số 8	Cờ vua	3196	75	37:00	0,493 (Dày đặc)	6
9	Nấm	8124	119	23:00	0,193 (Dày đặc)	14
10	Dữ liệu tổng hợp của IBM	–	–	–	–	21

Các bộ dữ liệu bms-pos, Connect và Tai nạn là ít phổ biến nhất trong các đánh giá.

9. Kết luận và hướng phát triển trong tương lai

Tính tối ưu của thuật toán sanitization trong việc giảm thiểu những tác động không mong muốn của quá trình ẩn là một vấn đề cấp thiết.Bằng chứng thu thập được cho bài viết này chỉ ra rằng các yếu tố khác nhau ảnh hưởng đến sự tối ưu.Trong nghiên cứu này, chúng tôi đã trình bày một đánh giá phân tích mới về 54 thuật toán được đề xuất để ẩn luật kết hợp, tập trung vào việc điều tra các khái niệm đóng góp chính trong quá trình làm sạch.Các kết quả được trình bày trong bài viết này có một số ý nghĩa quan trọng:

Về phương pháp tiếp cận, các phương pháp heuristic và tiến hóa được cập nhật, trong khi các phương pháp tiếp cận đường biên và chính xác chỉ được áp dụng từ năm 2005 đến 2009. Các phương pháp heuristic và tiến hóa có thể được áp dụng cả trong ẩn quy tắc và ẩn tập mục khu vực.Các phương pháp tiếp cận đường viền và chính xác chỉ thực hiện quy trình ẩn tập mục, trước đây tập trung vào việc giảm tác động của việc dọn dẹp đối với các tập mục không nhạy cảm để duy trì tiện ích của cơ sở dữ liệu đã được làm sạch.Mục đích thứ hai là giảm số lượng giao dịch hoặc vật phẩm được vệ sinh bằng quy trình ẩn.

Của chúng tôi nghiên cứu chứng minh cái đó các méo mó kỹ thuật

đã thu hút được sự chú ý lớn nhất kể từ khi xuất hiện bài toán ẩn luật kết hợp.Mặt khác, kỹ thuật chặn từ năm 2007 đã lỗi thời, trong khi kỹ thuật chèn/xóa giao dịch đã xuất hiện trong năm đó.

Trọng tâm chính của tất cả các thuật toán là xây dựng giải pháp lựa chọn các giao dịch và mục nạn nhân phù hợp nhằm tối ưu hóa quá trình ẩn giấu.Đây là động lực quan trọng để các nhà nghiên cứu thiết kế một thuật toán mới.Hầu hết các thuật toán đều sử dụng các phương pháp dựa trên heuristic nên 74% thuật toán sử dụng các phương pháp này.Việc áp dụng các phương pháp dựa trên quần thể đã thu hút được sự quan tâm đáng kể trong những năm gần đây, đặc biệt là trong lĩnh vực ẩn tập mục.

Mặc dù có sự khác biệt trong các phương pháp lựa chọn nhưng vẫn có sự đồng thuận lớn về việc sử dụng kỹ thuật dọn dẹp và chiến lược ẩn sao cho 74% thuật toán sử dụng kỹ thuật biến dạng và 65% thuật toán ẩn quy tắc áp dụng chiến lược giảm thiểu hệ quả để giảm độ tin cậy.

Phân tích của chúng tôi cho thấy rằng chi phí bỏ lỡ và các mẫu giả tạo phụ thuộc chủ yếu vào phương pháp lựa chọn và sau đó phụ thuộc vào chiến lược ẩn, trong khi thất bại trong việc ẩn chỉ phụ thuộc vào chiến lược ẩn khi độ tin cậy bị giảm bởi chiến lược tăng tiền đề.Quá trình dọn dẹp che giấu các tập mục bằng cách xóa mục không tạo ra bất kỳ tập mục nhân tạo nào trong khi nó có thể tạo ra các quy tắc nhân tạo.

Có một số hướng đầy hứa hẹn để nghiên cứu sâu hơn về ẩn luật kết hợp.Một trong số đó là việc mở rộng các giải pháp sửa đổi, chính xác và tiến hóa đường viền để che giấu quy tắc vấn đề, thay vì ẩn các tập mục tổng quát của sự kết hợp quy tắc.Vì mục đích của các thuật toán dựa trên tiến hóa là chọn các giao dịch nhạy cảm để xóa nên chúng có thể được kết hợp với giải pháp lựa chọn mục tối ưu để làm sai lệch các giao dịch thay vì xóa.Một số phương pháp tiếp cận dựa trên quần thể khác như đàn ong nhân tạo, tìm kiếm thức ăn của vi khuẩn và tối ưu hóa đàn kiến có thể được sử dụng để lựa chọn các giao dịch nhằm vệ sinh.Giảm thời gian tính toán của phương pháp tiếp cận chính xác, đặc biệt đối với cơ sở dữ liệu rất lớn, là một vấn đề khác có thể được giải quyết bằng cách sử dụng xử lý song song sao cho bài toán thỏa mãn ràng buộc được phân tách thành các thành phần khác nhau và mỗi thành phần được giải độc lập.Là cơ hội tốt cho các nhà nghiên cứu nhưng lại thiếu các nghiên cứu đánh giá tính hiệu quả của thuật toán ẩn giúp người quản trị cơ sở dữ liệu trong việc quyết định lựa chọn thuật toán ẩn tri thức.

Người giới thiệu

Afshari, MH, Dehkordi, MN, & Akbari, M. (2016).Ẩn luật kết hợp bằng thuật toán tối ưu hóa cuckoo.Hệ thống chuyên gia với các ứng dụng, 64, 340–351.Agrawal, R., & Srikant, R. (1994).các thuật toán nhanh chóng cho các luật kết hợp khai thác mỏ.

Trong Kỷ yếu hội nghị quốc tế lần thứ 20 về cơ sở dữ liệu rất lớn (trang 487–499).Agrawal, R., & Srikant, R. (20 0 0).Khai thác dữ liệu bảo vệ quyền riêng tư.Trong Kỷ yếu của hội nghị quốc tế ACM SIGMOD về quản lý dữ liệu (trang 439–450).Agrawal, R., Imielinski, T., & Swami, A. (1993).Quy tắc kết hợp khai thác giữa

tập hợp các mục trong cơ sở dữ liệu lớn.Trong Kỷ yếu của hội nghị ACM SIGMOD về quản lý dữ liệu (trang 207–216).

Amiri, A. (2007).Dám chia sẻ: Bảo vệ kiến thức nhạy cảm bằng cách vệ sinh dữ liệu.Hệ thống hỗ trợ quyết định, 43(1), 181–191.

Atallah, M., Bertino, E., Elmagarmid, A., Ibrahim, M., & Verykios, VS (1999).Hạn chế tiết lộ các quy định nhạy cảm.Trong Kỷ yếu hội thảo trao đổi kiến thức và kỹ thuật dữ liệu của IEEE (trang 45–52).

Bandaru, S., Ng, AH, & Deb, K. (2016).Các phương pháp khai thác dữ liệu để khám phá tri thức trong tối ưu hóa đa mục tiêu: Phần A-Khảo sát.Hệ chuyên gia với các ứng dụng, 70, 139–159.

Bayardo, R. (1998).Khai thác hiệu quả các mẫu dài từ cơ sở dữ liệu.Trong Kỷ yếu của hội nghị quốc tế ACM SIGMOD về quản lý dữ liệu.

Bertino, E., Fovino, IN, & Povenza, LP (2005).Một khung để đánh giá các thuật toán khai thác dữ liệu bảo vệ quyền riêng tư.Khai thác dữ liệu và khám phá tri thức, 11(2), 121–154.

Bleuler, S., Laumanns, M., Thiele, L., & Zitzler, E. (2003).PISA- một nền tảng và giao diện độc lập với ngôn ngữ lập trình cho các thuật toán tìm kiếm.Trong Kỷ yếu của hội nghị quốc tế về tối ưu hóa đa tiêu chí tiến hóa (trang 494–508).

Bodon, F. (2003).Triển khai APRIORI nhanh chóng.Trong Kỷ yếu hội thảo về triển khai khai phá tập mục thường xuyên.

Bodon, F. (2005).Triển khai APRIORI dựa trên trie để khai thác chuỗi mục thường xuyên.Trong Kỷ yếu hội thảo quốc tế lần thứ nhất về khai thác dữ liệu nguồn mở: Triển khai khai thác mẫu thường xuyên (trang 56–65).ACM.

Brijs, T. (2003).Bộ dữ liệu giỏ thị trường bán lẻ.Trong Kỷ yếu hội thảo về triển khai khai thác tập mục thường xuyên.

Cao, J., Karras, P., Raïssi, C., & Tan, K. (2010).Độ không chắc chắn P: ẩn danh giao dịch bằng chứng suy luận.Trong Kỷ yếu của cơ sở dữ liệu rất lớn Tài trợ (trang 1033–1044).

Chang, L., & Moskowitz, I. (1998).Cây quyết định và hạ cấp một cách chi tiết được áp dụng cho bài toán suy luận.Trong Kỷ yếu hội thảo về mô hình bảo mật mới (trang 82–89).

Cheng, P., Lee, I., Lin, CW, & Pan, JS (2016a).Ẩn luật kết hợp dựa trên tối ưu hóa đa mục tiêu tiến hóa.Phân tích dữ liệu thông minh, 20(3), 495–514.

Cheng, P., Pan, JS, & Lin, CW (2014).Khai thác quy tắc kết hợp bảo vệ quyền riêng tư bằng cách sử dụng NSGA-II được mã hóa nhị phân.Trong Kỷ yếu của hội nghị Châu Á Thái Bình Dương lần thứ 18 về khám phá kiến thức và khai thác dữ liệu (trang 87–99).

Cheng, P., Roddick, JF, Chu, SC, & Lin, CW (2016b).Bảo vệ quyền riêng tư thông qua phương pháp che giấu quy tắc dựa trên sự bóp méo tham lam.Trí tuệ Ứng dụng, 44(2), 295–306.

Clifton, C., & Marks, D. (1996).Ý nghĩa bảo mật và quyền riêng tư của khai thác dữ liệu.Trong Kỷ yếu hội thảo ACM về khai thác dữ liệu và khám phá tri thức (trang 15–19).

Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., & Zhu, MY (2002).Các công cụ bảo vệ quyền riêng tư khi khai thác dữ liệu phân tán.Bản tin Khám phá ACM SIGKDD, 4(2), 28–34.

Dasseni, E., Verykios, VS, Elmagarmid, AK, & Bertino, E. (2001).Ẩn các quy tắc kết hợp bằng cách sử dụng sự tin cậy và hỗ trợ.Trong Kỷ yếu của hội thảo quốc tế lần thứ 4 về che giấu thông tin (trang 369–383).

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002).Một thuật toán di truyền đa mục tiêu nhanh và ưu tú: NSGA-II.Giao dịch của IEEE về tính toán tiến hóa, 6(2), 182–197.

Dietzfelbinger, M., Karlin, AR, Mehlhorn, K., auf der Heide, FM, Rohnert, H., & Tarjan, RE (1994).Băm hoàn hảo động: Giới hạn trên và dưới.Tạp chí Máy tính SIAM, 23(4), 738–761.

Divanis, AG, & Verykios, V. (2006).Một phương pháp lập trình số nguyên để ẩn tập mục thường xuyên.Trong Kỷ yếu của hội nghị ACM lần thứ 15 về quản lý thông tin và kiến thức (trang 748–757).

Divanis, AG, & Verykios, V. (2009a).Kiến thức chính xác ẩn thông qua phần mở rộng cơ sở dữ liệu.Giao dịch của IEEE về Kỹ thuật Kiến thức và Dữ liệu, 21(5),

699–713.Divanis, AG, & Verykios, V. (2009b).Che giấu kiến thức nhạy cảm mà không có tác dụng phụ

fect.Hệ thống Thông tin và Tri thức, 20(3), 263–299.Divanis, AG, & Verykios, VS (2010).Ẩn luật kết hợp để khai thác dữ liệu.

Truyền thông Khoa học & Kinh doanh Springer.Evfimievski, A., Srikant, R., Agrawal, R., & Gehrke, J. (2004).Bảo vệ quyền riêng tư tối thiểu

việc áp dụng các luật kết hợp.Hệ thống Thông tin, 29(4), 343–364.Fung, BCM, Wang, K., Chen, R., & Yu, PS (2010).Xuất bản dữ liệu bảo vệ quyền riêng tư-

ing: Một cuộc khảo sát về những phát triển gần đây.Khảo sát máy tính ACM, 42(4), 141–153.Geurts, K., Wets, G., Brijs, T., & Vanhoof, K. (2003).Lập hồ sơ tai nạn tần số cao

vị trí sử dụng luật kết hợp.Tạp chí của Ban Nghiên cứu Giao thông Vận tải, 123–130.

Goethals, B., & Zaki, M. (2003).Những tiến bộ trong việc triển khai khai thác tập mục thường xuyên: Báo cáo về FIMI'03.Trong Kỷ yếu của hội thảo việc triển khai khai thác tập mục thường xuyên (trang 109–117).

Gratzer, G. (2010).Lý thuyết mạng: Nền tảng.Mùa xuân.Gueret, C., Prins, C., & Sevaux, M. (2002).Ứng dụng tối ưu hóa với Xpress-MP.

Tối ưu hóa Dash.Hải, LQ, & Somjit, A. (2012).Khung khái niệm về bảo vệ quyền riêng tư của

Khai thác luật kết hợp trong thương mại điện tửTrong Kỷ yếu của hội nghị IEEE lần thứ 7 về điện tử công nghiệp và ứng dụng (trang 1999–2003).

Hải, LQ, Somjit, A., & Ngamnij, A. (2012).Ẩn luật kết hợp dựa trên khoảng cách và mạng giao nhau.Trong Kỷ yếu của hội nghị quốc tế lần thứ 4 về công nghệ và phát triển máy tính (trang 227–231).

Hải, LQ, Somjit, A., & Ngamnij, A. (2013a).Ẩn luật kết hợp dựa trên mạng giao nhau.Các vấn đề toán học trong kỹ thuật.

Hải, LQ, Somjit, A., Huy, XN, & Ngamnij, A. (2013b).Quy tắc hiệp hội ẩn trong quản lý rủi ro để hợp tác chuỗi cung ứng bán lẻ.Máy tính trong Công nghiệp, 64, 776–784.

Hajian, S., Domingo-Ferrer, J., & Farràs, O. (2014).Bảo vệ quyền riêng tư và ngăn chặn phân biệt đối xử dựa trên khái quát hóa trong xuất bản và khai thác dữ liệu.Khai thác dữ liệu và khám phá tri thức, 28, 1158–1188.

Han, J., Pei, J., & Yin, Y. (20 0 0).Khai thác các mẫu phổ biến mà không cần tạo ứng cử viên.Trong Kỷ yếu của hội nghị quốc tế về quản lý dữ liệu (trang 1–12).

Han, J., Pei, J., Yin, Y., & Mao, R. (2004).Khai thác mẫu phổ biến mà không tạo ứng cử viên: Cách tiếp cận cây mẫu phổ biến.Khai thác dữ liệu và khám phá tri thức, 8(1), 53–87.

Harik, GR, Lobo, FG, & Goldberg, DE (1999).Thuật toán di truyền nhỏ gọn.Giao dịch của IEEE về tính toán tiến hóa, 3(4), 287–297.

Hà Lan, JH (1992).Thích ứng trong các hệ thống tự nhiên và nhân tạo.Nhà xuất bản MIT.Hồng, TP, Lin, CW, Yang, KT, & Wang, SL (2013).Sử dụng TF-IDF để ẩn thông tin nhạy cảm

itemset.Trí tuệ ứng dụng, 38(4), 502–510.Hồng, TP, Wang, CY, & Tao, YH (2001).Một thuật toán khai thác dữ liệu gia tăng mới

Rithm sử dụng tập mục trước lớn.Phân tích dữ liệu thông minh, 5, 111–129.Huang, H., Wu, X., & Relue, R. (2002).Phân tích liên kết với một lần quét

cơ sở dữ liệu.Trong Kỷ yếu của hội nghị quốc tế IEEE về khai thác dữ liệu

(trang 629–632).Kennedy, J., & Eberhart, R. (1995).Phương pháp tối ưu bầy đàn.Trong Kỷ yếu tổ tụng của

Hội nghị quốc tế của IEEE về mạng lưới thần kinh (trang 1942–1948).Kowalski, GJ, & Maybury, MT (2006).Hệ thống lưu trữ và truy xuất thông tin:

Lý thuyết và thực hiện.Mùa xuân.Kumar, V. (1992).Các thuật toán cho các vấn đề thỏa mãn ràng buộc: Một cuộc khảo sát.AI Mag-

tạp chí, 13(1).Kuo, RJ, Chao, CM, & Chiu, YT (2011).Ứng dụng tối ưu hóa bầy đàn hạt

về khai phá luật kết hợp.Máy tính mềm ứng dụng, 11(1), 326–336.Li, J., Shen, H., & Topor, R. (2001).Khai phá tập luật kết hợp nhỏ nhất cho

phỏng đoán.Trong Kỷ yếu của hội nghị quốc tế IEEE về khai thác dữ liệu

(trang 361–368).

Li, YC, & Chang, CC (2007).MICF: Một thuật toán dọn dẹp hiệu quả để ẩn các mẫu nhạy cảm khi khai thác dữ liệu.Tin học kỹ thuật nâng cao, 21, 269–280.

Lin, CW, Hong, TP, Wong, JW, Lan, GC, & Lin, WY (2014b).Cách tiếp cận dựa trên GA để ẩn các tập mục có tính tiện ích cao nhạy cảm.Tạp chí Khoa học Thế giới, 2014. doi:10.1155/2014/804629.

Lin, CW, Hong, TP, Yang, KT, & Wang, SL (2015).Các thuật toán dựa trên GA để tối ưu hóa việc ẩn các tập mục nhạy cảm thông qua việc xóa giao dịch.Trí tuệ ứng dụng, 42(2), 210–230.

Lin, CW, Liu, Q., Fournier-Viger, P., Hong, TP, Voznak, M., & Zhan, JA (2016).Một cách tiếp cận dọn dẹp để ẩn các tập mục nhạy cảm dựa trên tối ưu hóa bầy đàn hạt.Ứng dụng kỹ thuật của trí tuệ nhân tạo, 53, 1–18.

Lin, CW, Zhang, B., Yang, KT, & Hong, TP (2014a).Ẩn các tập mục nhạy cảm một cách hiệu quả bằng cách xóa giao dịch dựa trên thuật toán di truyền.Tạp chí Khoa học Thế giới, 2014. doi:10.1155/2014/398269.

Lin, JL, & Cheng, YW (2009).Bảo vệ quyền riêng tư khi khai thác tập mục thông qua các mục ồn ào.Hệ thống chuyên gia với các ứng dụng, 36, 5711–5717.

Lin, JL, & Liu, JYC (2007).Bảo mật quyền riêng tư trong việc khai thác tập mục thông qua các giao dịch giả mạo.Trong Kỷ yếu của hội nghị chuyên đề ACM thường niên lần thứ 22 về điện toán ứng dụng.

Lin, W., Alvarez, S., & Ruiz, C. (2002).Khai thác quy tắc kết hợp hỗ trợ thích ứng hiệu quả cho các hệ thống gợi ý.Khai thác dữ liệu và khám phá tri thức, 6, 83–105.

Lindell, Y., & Pinkas, B. (2009).Bảo mật tính toán đa bên để khai thác dữ liệu bảo đảm quyền riêng tư.Tạp chí Quyền riêng tư và Bảo mật, 1(1), 59–98.

Liu, Y., Liao, WK, & Choudhary, A. (2005).Thuật toán hai pha để phát hiện nhanh các tập mục có tiện ích cao.Trong Kỷ yếu của hội nghị Châu Á Thái Bình Dương về khám phá kiến thức và khai thác dữ liệu (trang 689–695).

Luenberger, D. (1973).Giới thiệu về lập trình tuyến tính và phi tuyến tính.Addis-

Công ty xuất bản on-Wesley.Maimon, O., & Rokach, L. (2010).Sổ tay khai phá dữ liệu và khám phá tri thức.

Mùa xuân.Mannila, H., & Toivonen, H. (1997).Tìm kiếm theo cấp độ và ranh giới của các lý thuyết trong khám phá tri thức.Khai thác dữ liệu và khám phá tri thức, 1(3), 241–258.Menon, S., & Sarkar, S. (2008).Giảm thiểu mất mát thông tin và bảo vệ sự riêng tư.

Khoa học Quản lý, 53, 101–116.Menon, S., Sarkar, S., & Mukherjee, S. (2005).Tối đa hóa độ chính xác của việc chia sẻ

cơ sở dữ liệu khi che giấu các mẫu nhạy cảm.Nghiên cứu Hệ thống Thông tin, 16(3), 256–270.

Moskowitz, I., & Chang, L. (20 0 0).Hệ thống lý thuyết quyết định về thông tin

hạ cấp.Trong Kỷ yếu hội nghị chung về khoa học thông tin.Moustakides, GV, & Verykios, VS (2006).Cách tiếp cận tối đa-tối thiểu để ẩn thường xuyên

tập mục quen thuộc.Trong Kỷ yếu của hội nghị quốc tế IEEE lần thứ 6 về khai thác dữ liệu (trang 502–506).

Moustakides, GV, & Verykios, VS (2008).Cách tiếp cận MaxMin để ẩn thường xuyên

itemset.Kỹ thuật Dữ liệu & Tri thức, 65, 75–89.O'Leary, D. (1991).Khám phá tri thức như một mối đe dọa đối với bảo mật cơ sở dữ liệu.TRONG

G. Piatetsky-Shapiro, & WJ Frawley (Biên tập), Khám phá tri thức trong cơ sở dữ liệu

(trang 507–516).Công viên Menlo: Nhà xuất bản AAAI/MIT.O'Leary, DE (1995).Một số vấn đề về quyền riêng tư trong khám phá tri thức: Nguyên tắc OECD

hướng dẫn về quyền riêng tư của sonal.Chuyên gia IEEE, 10(2), 48–52.O'Mahony, M., Hurley, N., Kushmerick, N., & Silvestre, G. (2004).Đề xuất hợp tác

Khuyến nghị: Một phân tích mạnh mẽ.Giao dịch ACM trên Công nghệ Internet, 4(4), 344–377.

Oliveira, SRM, & Zaiane, O. (2004).Sự xáo trộn dữ liệu bằng cách xoay vòng để phân cụm bảo vệ quyền riêng tư.Báo cáo kỹ thuật TR04-17.

Oliveira, SRM, & Zaiane, OR (2002).Quyền riêng tư bảo vệ việc khai thác tập mục thường xuyên.Trong Kỷ yếu của hội nghị quốc tế IEEE về quyền riêng tư, bảo mật và khai thác dữ liệu (trang 43–54).

Oliveira, SRM, & Zaiane, OR (2003a).Bảo vệ kiến thức nhạy cảm bằng cách vệ sinh dữ liệu.Trong Kỷ yếu của hội nghị quốc tế IEEE lần thứ 3 về khai thác dữ liệu (trang 211–218).

Oliveira, SRM, & Zaiane, OR (2003b).Các thuật toán cân bằng quyền riêng tư và khám phá tri thức trong khai phá luật kết hợp.Trong Kỷ yếu của hội nghị chuyên đề ứng dụng và kỹ thuật cơ sở dữ liệu quốc tế (trang 54–63).

Oliveira, SRM, & Zaiane, OR (2006).Một khuôn khổ thống nhất để bảo vệ các quy tắc kết hợp nhạy cảm trong cộng tác kinh doanh.Tạp chí Quốc tế về Kinh doanh Thông minh và Khai thác Dữ liệu, 1(3), 247–287.

Oliveira, SR, & Zaiane, OR (2010).Phân cụm bảo vệ quyền riêng tư bằng cách chuyển đổi dữ liệu.Tạp chí Quản lý thông tin và dữ liệu, 1(1), 37.

Pontikakis, ED, Tsitsonis, AA, & Verykios, VS (2004a).Một nghiên cứu thực nghiệm về các kỹ thuật dựa trên biến dạng để ẩn luật kết hợp.Trong Kỷ yếu của hội nghị lần thứ 18 về bảo mật cơ sở dữ liệu (trang 325–339).

Pontikakis, ED, Theodoridis, Y., Tsitsonis, AA, Chang, L., & Verykios, VS (2004b).Một phân tích định lượng và định tính về việc chặn trong việc ẩn luật kết hợp.Trong Kỷ yếu của hội thảo ACM về quyền riêng tư trong xã hội điện tử (trang 29–30).

Prakash, M., & Singaravel, G. (2015).Một cách tiếp cận để ngăn chặn vi phạm quyền riêng tư và rò rỉ thông tin trong khai thác dữ liệu nhạy cảm.Máy tính và Kỹ thuật Điện, 45, 134–140.

Reddy, MR, & Wang, RY (1995).Ước tính độ chính xác của dữ liệu trong môi trường cơ sở dữ liệu liên kết.Trong Kỷ yếu của hội nghị quốc tế lần thứ 6 về hệ thống thông tin và quản lý dữ liệu (trang 115–134).

Rizvi, SJ, & Haritsa, JR (2002).Duy trì sự riêng tư của dữ liệu trong khai phá luật kết hợp.Trong Kỷ yếu của hội nghị lần thứ 28 về cơ sở dữ liệu rất lớn (trang 682–693).

Russell, S., & Norvig, P. (2003).Trí tuệ nhân tạo: một cách tiếp cận hiện đại: 27. Prentice-Hall.

Salton, G., Fox, EA, & Wu, H. (1983).Truy xuất thông tin Boolean mở rộng.Thông tin của ACM, 26(11), 1022–1036.

Samarati, P. (2001).Bảo vệ danh tính của người trả lời trong việc phát hành dữ liệu vi mô.Giao dịch của IEEE về Kỹ thuật Kiến thức và Dữ liệu, 13(6), 1010–1027.

Saygin, Y., Verykios, VS, & Clifton, C. (2001).Sử dụng ẩn số để ngăn cản việc khám phá các luật kết hợp.Bản ghi ACM SIGMOD, 30(4), 45–54.

Saygin, Y., Verykios, VS, & Elmagarmid, AK (2002).Khai thác quy tắc kết hợp bảo vệ quyền riêng tư.Trong Kỷ yếu hội thảo quốc tế về các vấn đề nghiên cứu trong kỹ thuật dữ liệu: Kỹ thuật hệ thống thương mại điện tử/kinh doanh điện tử (trang 151–158).Sohrabi, MK, & Roshani, R. (2017).Khai thác tập mục thường xuyên bằng cách sử dụng học tập di động

máy tự động.Máy tính trong hành vi con người, 68, 244–253.Sun, X., & Yu, PS (2005).Một cách tiếp cận dựa trên đường viền để ẩn các tần số nhạy cảm

itemset.Trong Kỷ yếu của hội nghị quốc tế IEEE lần thứ 5 về khai thác dữ liệu

(trang 426–433).Sun, X., & Yu, PS (2007).Ẩn các tập mục thường xuyên nhạy cảm theo đường viền

tiếp cận.Khoa học và Kỹ thuật Máy tính, 1(1), 74–94.Sweeny, L. (2002).K-Anonymity: Một mô hình bảo vệ quyền riêng tư.Tạp chí quốc tế-

nal Hệ thống dựa trên tri thức mờ không chắc chắn, 10(5), 557–570.Telikani, A., & Shahbahrami, A. (2017).Tối ưu hóa quy tắc kết hợp ẩn chúng tôi-

sự kết hợp giữa các phương pháp biên giới và heuristic.Trí tuệ ứng dụng, 47, 544–557.

Verykios, VS (2013).Các phương pháp che giấu luật kết hợpKhai thác dữ liệu và khám phá kiến thức của WIRE, 3, 28–36.

Verykios, VS, & Divanis, AG (2008).Khảo sát các phương pháp ẩn luật kết hợp để bảo mật.Trong CC Aggarwal, & PS Yu (Eds.), Khai thác dữ liệu bảo vệ quyền riêng tư: mô hình và thuật toán (trang 267–289).New York: Springer.

Verykios, VS, Bertino, E., Fovino, IN, Parasiliti, L., Saygin, Y., & Theodor-idis, Y. (2004a).Công nghệ tiên tiến nhất trong việc bảo vệ quyền riêng tư trong khai thác dữ liệu.Bản ghi ACM SIGMOD, 33(1), 50–57.

Verykios, VS, Elmagarmid, AK, Bertino, E., Saygin, Y., & Dasseni, E. (2004b).Ẩn quy tắc kết hợp.Giao dịch của IEEE về Kỹ thuật Kiến thức và Dữ liệu, 16(4), 434–447.

Verykios, VS, Pontikakis, ED, Theodoridis, Y., & Chang, L. (2007).Các thuật toán hiệu quả cho kỹ thuật bóp méo và chặn trong ẩn luật kết hợp.Cơ sở dữ liệu song song và phân tán, 22(1), 85–104.

Vương, SL (2009).Duy trì vệ sinh các quy tắc kết hợp thông tin.Hệ thống chuyên gia với các ứng dụng, 36, 4006–4012.

Wang, SL, & Jafari, A. (2005).Sử dụng ẩn số để ẩn các quy tắc kết hợp dự đoán nhạy cảm.Trong Kỷ yếu của hội nghị quốc tế IEEE về tái sử dụng và tích hợp thông tin (trang 223–228).

Wang, SL, Maskey, R., Jafari, A., & Hong, TP (2008).Làm sạch hiệu quả các quy tắc kết hợp thông tin.Hệ thống chuyên gia với các ứng dụng, 35(1-2), 442–450.Wang, SL, Parikh, B., & Jafari, A. (2007a).Ẩn bộ quy tắc kết hợp thông tin.

Hệ thống chuyên gia với các ứng dụng, 33(2), 316–323.Wang, SL, Patel, D., Jafari, A., & Hong, TP (2007b).Ẩn đề xuất cộng tác

quy tắc kết hợp sửa đổi.Trí tuệ ứng dụng, 26(1), 66–77.Wu, YH, Chiang, CM, & Chen, AL (2007).Ẩn các quy tắc kết hợp nhạy cảm với

tác dụng phụ hạn chế.Giao dịch của IEEE về Kỹ thuật tri thức và dữ liệu, 19(1).Yang, XS, & Deb, S. (2009).Tìm kiếm chim cu qua các chuyến bay Lévy.Trong Kỷ yếu tố tụng của

Bản chất của IEEE và điện toán lấy cảm hứng từ sinh học (trang 210–214).Yao, AC (1982).Giao thức cho tính toán an toàn.Trong Kỷ yếu thường niên lần thứ 23

Hội nghị chuyên đề của IEEE về nền tảng khoa học máy tính (trang 160–164).Zaki, MJ (20 0 0).Các thuật toán có thể mở rộng để khai thác liên kết.Giao dịch IEEE trên

Kỹ thuật Kiến thức và Dữ liệu, 12(3), 372–390.Zeng, Y., Yin, S., Liu, J., & Zhang, M. (2015).Nghiên cứu cải tiến thuật toán tăng trưởng FP

nguyên tắc trong khai phá luật kết hợp.Lập trình khoa học.Zheng, Z., Kohavi, R., & Mason, L. (1999).Hiệu suất của hiệp hội trong thế giới thực

thuật toán quy luật.Trong Kỷ yếu của hội nghị quốc tế ACM SIGKDD lần thứ 7 về khai thác dữ liệu khám phá tri thức (trang 401–406).