Che giấu hiệu quả các tập vật phẩm bí mật có tính tiện ích cao với tác dụng phụ tối thiểu

Bài viết trên Tạp chí Trí tuệ nhân tạo thực nghiệm và lý thuyết · Tháng 5 năm 2017

DOI: 10.1080/0952813X.2017.1328462

Tạp chí Trí tuệ nhân tạo thực nghiệm và lý thuyết

ISSN: 0952-813X (Bản in) 1362-3079 (Trực tuyến) Trang chủ tạp chí: http://www.tandfonline.com/loi/teta20

Che giấu hiệu quả các tập vật phẩm bí mật có tính tiện ích cao với tác dụng phụ tối thiểu

Jerry Chun-Wei Lin, Tzung-Pei Hong, Philippe Fournier-Viger, Qiankun Liu, Jia-Wei Wong & Justin Zhan

Để trích dẫn bài viết này: Jerry Chun-Wei Lin, Tzung-Pei Hong, Philippe Fournier-Viger, Qiankun Liu, Jia-Wei Wong & Justin Zhan (2017): Che giấu hiệu quả các bộ vật phẩm bí mật có tiện ích cao với tác dụng phụ tối thiểu, Tạp chí của Trí tuệ nhân tạo thực nghiệm & lý thuyết, DOI: 10.1080/0952813X.2017.1328462

Để liên kết đến bài viết này: http://dx.doi.org/10.1080/0952813X.2017.1328462

Xuất bản trực tuyến: 15 tháng 5 năm 2017.

Taylor & Francis
Taylor & Francis Group

TẠP CHÍ TRÍ TUỆ NHÂN TẠO THỰC NGHIỆM & LÝ LUẬN, 2017
https://doi.org/10.1080/0952813X.2017.1328462

Che giấu hiệu quả các tập vật phẩm bí mật có tính tiện ích cao với tác dụng phụ tối thiểu

Jerry Chun-Wei Lin a,b , Tzung-Pei Hong c,d , Philippe Fournier-Viger B, Càn Khôn Lưu Jia-Wei Wong d và Justin Zhanf

Phòng thí nghiệm trọng điểm về khai thác và ứng dụng dữ liệu lớn của tỉnh Phúc Kiến, Đại học Công nghệ Phúc Kiến, Phúc Kiến, Trung Quốc; b Trường Khoa học và Công nghệ Máy tính, Viện Công nghệ Cáp Nhĩ Tân Trường sau đại học Thâm Quyến, Thâm Quyến, Trung Quốc; c Khoa Khoa học Máy tính và Kỹ thuật Thông tin, Đại học Quốc gia Cao Hùng, Cao Hùng, Đài Loan; d Khoa Khoa học và Kỹ thuật Máy tính, Đại học Quốc gia Sun Yat-sen, Cao Hùng, Đài Loan; e Trường Khoa học Tự nhiên và Nhân văn, Viện Công nghệ Cáp Nhĩ Tân Trường sau đại học Thâm Quyến, Thâm Quyến, Trung Quốc; f Khoa Khoa học Máy tính, Đại học Nevada, Las Vegas, NV, Hoa Kỳ

TÓM TẮT Khai thác dữ liệu bảo vệ quyền riêng tư (PPDM) là một vấn đề nghiên cứu mới nổi và đã trở nên quan trọng trong những thập kỷ qua. PPDM bao gồm việc ẩn thông tin nhạy cảm để đảm bảo rằng nó không thể bị phát hiện bởi các thuật toán khai thác dữ liệu. Một số thuật toán PPDM đã được phát triển. Hầu hết chúng được thiết kế để ẩn các tập phổ biến nhạy cảm hoặc các quy tắc kết hợp. Việc ẩn thông tin nhạy cảm trong cơ sở dữ liệu có thể có một số tác dụng phụ như ẩn thông tin không nhạy cảm khác và đưa ra thông tin dư thừa. Việc tìm tập hợp các tập mục hoặc giao dịch cần được làm sạch để giảm thiểu tác dụng phụ là một bài toán khó. Trong bài báo này, một thuật toán di truyền (GA) sử dụng tính năng xóa giao dịch được thiết kế để ẩn các tập mục hữu ích cao nhạy cảm cho PPUM. Chức năng thể dục linh hoạt với ba trọng số có thể điều chỉnh được sử dụng để đánh giá mức độ tốt của từng nhiễm sắc thể trong việc ẩn các tập mục có tính tiện ích cao nhạy cảm. Để tăng tốc quá trình tiến hóa, khái niệm tiền lớn được áp dụng trong thuật toán được thiết kế. Nó làm giảm số lần quét cơ sở dữ liệu cần thiết để xác minh tính tốt của nhiễm sắc thể được đánh giá. Các thử nghiệm đáng kể được tiến hành để so sánh hiệu suất của phương pháp GA được thiết kế (có/không có khái niệm tiền lớn), với cách tiếp cận dựa trên GA dựa vào việc chèn giao dịch và thuật toán không tiến hóa, về thời gian thực hiện, tác dụng phụ, tính toàn vẹn của cơ sở dữ liệu và tính toàn vẹn của tiện ích. Kết quả chứng minh rằng thuật toán đề xuất ẩn các tập mục có tính tiện ích cao nhạy cảm với ít tác dụng phụ hơn so với các nghiên cứu trước đây, trong khi vẫn duy trì tính toàn vẹn của cơ sở dữ liệu và tiện ích cao.

LỊCH SỬ BÀI VIẾT Nhận được ngày 7 tháng 11 năm 2015 Được chấp nhận ngày 29 tháng 4 năm 2017

TỪ KHÓA Thuật toán di truyền; xóa giao dịch; khai thác tiện ích cao; khai thác dữ liệu bảo vệ quyền riêng tư

1. Giới thiệu

Khai thác tập mục thường xuyên (FIM), khai thác quy tắc kết hợp (ARM) và khai thác mẫu tuần tự là các kỹ thuật cơ bản để khám phá tri thức trong cơ sở dữ liệu (Agrawal & Srikant, 1994b; Fournier-Viger, Lin, Kiran, Koh, & Thomas, 2017; Han , Pei, Yin, & Mao, 2004; Lin, Hong, & Lu, 2009). Mặc dù các kỹ thuật này rất hữu ích nhưng một vấn đề quan trọng là thông tin được phát hiện bằng kỹ thuật khai thác dữ liệu có thể tiết lộ thông tin bí mật hoặc riêng tư như số an sinh xã hội, số thẻ tín dụng và các vấn đề sức khỏe.

LIÊN HỆ Jerry Chun-Wei Lin jerrylin@ieee.org

Trong những thập kỷ gần đây, việc khai thác dữ liệu đảm bảo quyền riêng tư (PPDM) (Agrawal & Srikant, 2000; Atallah, Elmagarmid, Ibrahim, Bertino, & Verykios, 1999; Verykios et al., 2004) đã thu hút sự chú ý của nhiều nhà nghiên cứu và học viên vì nó cho phép vệ sinh cơ sở dữ liệu bằng cách ẩn thông tin bí mật và riêng tư, đồng thời đảm bảo rằng các thông tin quan trọng khác vẫn có thể được trích xuất để ra quyết định. Agrawal và Srikant (2000) đã chỉ ra rằng các hàm nhiễu loạn Gaussian và thống nhất ngẫu nhiên có thể được áp dụng để ẩn danh cơ sở dữ liệu. Evfimievski, Srikant, Agrawal và Gehrke (2002) đã trình bày một khuôn khổ để khai thác các quy tắc kết hợp trong các giao dịch bao gồm các mục phân loại trong đó dữ liệu được chọn ngẫu nhiên để bảo vệ quyền riêng tư của các giao dịch riêng lẻ. Clifton, Kantarcioglu, Vaidya, Lin và Zhu (2002) đã thiết kế một bộ công cụ với một số thành phần để giải quyết các vấn đề của PPDM. Lindell và Pinkas (2000) đã đưa ra khái niệm PPDM và sử dụng nguyên tắc ID3 trong PPDM. Wu, Chiang và Chen (2007) đã thiết kế một thuật toán để sửa đổi các giao dịch tương ứng nhằm làm giảm sự hỗ trợ và độ tin cậy của các quy tắc nhạy cảm đối với PPDM. Hong, Lin, Yang và Wang (2013) đã thiết kế một phương pháp tiếp cận tần số cơ sở dữ liệu nghịch đảo tần số của các mục nhạy cảm để vệ sinh cơ sở dữ liệu bằng cách giảm sự hỗ trợ của các tập mục nhạy cảm. Lin, Hong, Chang và Wang (2013) đã thiết kế một cách tiếp cận tham lam nhằm tăng kích thước cơ sở dữ liệu nhằm che giấu thông tin nhạy cảm.

Khai thác tập mục tiện ích cao (HUIM) (Lin, Hong, & Lu, 2011; Liu, Liao, & Choudhary, 2005; Yao, Hamilton, & Butz, 2004; Yao & Hamilton, 2006) là một chủ đề nghiên cứu mới nổi xem xét cả hai lợi nhuận đơn vị của các mặt hàng và số lượng mua của chúng trong các giao dịch để khai thác các tập mặt hàng có tiện ích cao (HUI). Do HUIM gặp phải các vấn đề bảo mật tương tự như ARM truyền thống nên việc khai thác tiện ích bảo vệ quyền riêng tư (PPUM) (Rajalaxmi & Nataraja, 2012; Yeh & Hsu, 2010; Yun & Kim, 2015) đã nổi lên như một biến thể quan trọng của PPDM. Trong PPUM, các mẫu tiện ích cao nhạy cảm bị ẩn bằng cách làm xáo trộn cơ sở dữ liệu gốc bằng cách loại bỏ các tập mục hoặc giảm tiện ích của các tập mục nhạy cảm trong cơ sở dữ liệu. Yeh và Hsu (2010) đã thiết kế các phương pháp tiếp cận HHUIF và MSICF tiên tiến nhất để che giấu các mẫu có tính tiện ích cao nhạy cảm. Rajalaxmi và Nataraja (2012) đã trình bày hai thuật toán dọn dẹp có tên MSMU và MCRSU để ẩn cả các tập mục hữu ích cao nhạy cảm và các tập phổ biến bằng cách sửa đổi cơ sở dữ liệu. Yun và Kim (2015) đã trình bày một cấu trúc cây hiệu quả để che giấu các mẫu hữu ích cao nhạy cảm.

Vì mục đích của PPDM và PPUM là ẩn thông tin nhạy cảm trong cơ sở dữ liệu đồng thời đảm bảo rằng các thông tin hoặc quy tắc quan trọng khác vẫn có thể bị tiết lộ, nên những vấn đề này có thể được coi là vấn đề tối ưu hóa trong đó phải tìm ra giải pháp tối ưu, đó là NP-hard (Agrawal & Srikant, 2000; Aggarwal, Pei, & Zhang, 2006). Thuật toán di truyền (GA) Holland (1992) là một mô hình tiến hóa có thể tìm ra các giải pháp khả thi gần như tối ưu trong một khoảng thời gian giới hạn. Trong bài viết này, một cách tiếp cận dựa trên GA được phát triển để làm sạch cơ sở dữ liệu nhằm ẩn các tập mục nhạy cảm có tính tiện ích cao thông qua việc xóa giao dịch. Những đóng góp chính của thuật toán được thiết kế có hai phần và được tóm tắt như sau.

(1) Đây là bài viết đầu tiên giải quyết vấn đề bảo vệ quyền riêng tư của HUIM bằng cách áp dụng cách tiếp cận dựa trên GA để tìm các giao dịch thích hợp cần xóa để ẩn các tập mục tiện ích cao nhạy cảm. Chức năng thích ứng linh hoạt được áp dụng trong quá trình tiến hóa để đánh giá mức độ tốt của từng nhiễm sắc thể được xử lý và giảm thiểu tác dụng phụ do quá trình vệ sinh gây ra.

(2) Hạn chế chính của tính toán tiến hóa là tốn thời gian. Để giải quyết vấn đề này, bài viết này đề xuất một khái niệm tiền lớn nâng cao, nhằm tránh thực hiện nhiều lần quét cơ sở dữ liệu trong quá trình tiến hóa nhằm phát hiện nhanh các tác dụng phụ của từng nhiễm sắc thể được xử lý. Quy trình này làm giảm đáng kể thời gian chạy của quá trình tiến hóa.

2. Công việc liên quan

Trong phần này, công việc liên quan về HUIM và PPDM sẽ được xem xét và thảo luận.

2.1. Khai thác tập mục tiện ích cao

HUIM (Chan, Yang, & Shen, 2003; Yao và cộng sự, 2004; Yao & Hamilton, 2006) là một chủ đề mới nổi và ngày càng trở nên quan trọng trong những năm gần đây, vì nó có thể tiết lộ nhiều thông tin có lợi và có ý nghĩa hơn ARM truyền thống. HUI có thể được các nhà quản lý và người ra quyết định sử dụng để đưa ra các quyết định kinh doanh sáng suốt và điều chỉnh các chiến lược bán hàng. Yao và cộng sự. (2004) và Yao và Hamilton (2006) đã phát triển các thuật toán để khai thác các tập mục có lợi nhuận bằng cách xem xét cả số lượng mua và lợi nhuận đơn vị của các mặt hàng để tiết lộ HUI mong muốn. Bởi vì thuộc tính đóng xuống theo trọng số giao dịch không có trong HUIM truyền thống, nên mô hình sử dụng theo trọng số giao dịch (Liu và cộng sự, 2005) với thuộc tính đóng theo trọng số giao dịch được thiết kế đã được đề xuất để tìm các tập mục sử dụng theo trọng số giao dịch cao (HTWUI), và tăng tốc độ khám phá HUI. Cách tiếp cận này thực hiện tìm kiếm theo cấp độ để tìm HUI.

Để tránh những hạn chế của cách tiếp cận theo cấp độ, Lin et al. đã thiết kế cấu trúc cây mô hình tiện ích cao (HUP) để khai thác HUI (Lin và cộng sự, 2011). Thuật toán cây HUP trước tiên phát hiện ra 1-HTWUI (HTWUI chứa một mục duy nhất) bằng mô hình TWU. Sau đó, 1-HTWUI được phát hiện sẽ được sử dụng để xây dựng cấu trúc cây HUP, cấu trúc này sau đó được sử dụng để khám phá tất cả các HUI. Một thuật toán dựa trên danh sách có tên HUI-Miner (Liu & Qu, 2012) cũng được thiết kế để khai thác trực tiếp HUI mà không cần tạo ứng viên. Một số phần mở rộng của HUIM lần lượt được trình bày và thảo luận (Lin, Gan, Fournier-Viger, Hong, & Tseng, 2015; Lin, Gan, Fournier-Viger, Hong, & Zhan, 2016; Lin, Fournier-Viger, & Gan , 2016; Liu, Wang, & Fung, 2016).

2.2. Kỹ thuật bảo vệ quyền riêng tư

Kỹ thuật khai thác dữ liệu (Agrawal & Srikant, 1994b; Han và cộng sự, 2004) được sử dụng để khám phá và phân tích mối quan hệ tiềm ẩn giữa các mặt hàng được mua. Thông tin riêng tư hoặc bí mật giữa các mục trong tập mục cũng có thể bị tiết lộ bởi các kỹ thuật khai thác dữ liệu, điều này có thể dẫn đến các mối đe dọa bảo mật và đánh cắp thông tin. Do đó, PPDM (Agrawal & Srikant, 2000; Aggarwal và cộng sự, 2006; Verykios và cộng sự, 2004) đã nổi lên như một chủ đề quan trọng trong những năm gần đây vì thông tin riêng tư hoặc bí mật có thể được ẩn đi, đồng thời đảm bảo rằng thông tin mong muốn về việc mua hàng của các tập mục vẫn có thể được khám phá. Khử trùng là một kỹ thuật PPDM, có thể che giấu thành công thông tin riêng tư hoặc bí mật bằng cách xóa hoặc chèn. Để so sánh hiệu suất của các thuật toán vệ sinh, ba tác dụng phụ có tên là thất bại ẩn, chi phí bị thiếu và chi phí nhân tạo đã được đề xuất (Oliveria & Zaiane, 2002; Wu và cộng sự, 2007). Atallah và cộng sự. (1999) đã phát triển một cơ chế bảo vệ để làm sạch cơ sở dữ liệu nhằm che giấu các quy tắc kết hợp nhạy cảm. Wu và cộng sự. (2007) đã thiết kế một số phương pháp heuristic để ẩn các quy tắc nhạy cảm trong khi giảm thiểu số lượng mục được sửa đổi. Giannotti, Lakshmanan, Monreale, Pedreschi và Wang (2013) đã nghiên cứu vấn đề thuê ngoài nhiệm vụ ARM trong khuôn khổ bảo vệ quyền riêng tư của công ty. Bonam, Reddy và Kalyani (2014) đã giải quyết vấn đề bảo vệ quyền riêng tư trong ARM bằng cách phát triển phương pháp tiếp cận dựa trên PSO bằng cách sử dụng biến dạng dữ liệu. Cheng, Lin và Pan (2015) đã áp dụng cơ chế tối ưu hóa đa mục tiêu (EMO) để xem xét nhiều yếu tố nhằm che giấu các tập mục nhạy cảm. Cheng, Roddick, Chu và Lin (2016) sau đó đã đề xuất một phương pháp xóa để giảm mức độ hỗ trợ hoặc độ tin cậy của các quy tắc nhạy cảm dưới ngưỡng quy định cho PPDM.

Bên cạnh PPDM, PPUM cũng đã trở thành một vấn đề quan trọng vì nó xem xét cả số lượng mua và lợi nhuận đơn vị của mặt hàng để che giấu các HUI nhạy cảm. Yeh và Hsu (2010) lần đầu tiên thiết kế hai thuật toán có tên HHUIF và MSICF để ẩn các tập mục hữu ích cao nhạy cảm. Lin, Wu và cộng sự. (2015) đã thiết kế ba phép đo tương tự được sử dụng làm tiêu chuẩn mới để đánh giá tác dụng phụ trong PPUM. Rajalaxmi và Nataraja (2012) đã trình bày thuật toán MSMU và MCRSU để ẩn cả tiện ích nhạy cảm và vật phẩm/bộ thường xuyên. Yun và Kim (2015) đã thiết kế cấu trúc cây và thuật toán FPUTT để tăng tốc thuật toán HHUIF và MSICF. Việc tìm các giao dịch hoặc mục thích hợp cần được sửa đổi để ẩn các HUI nhạy cảm trong PPUM là một vấn đề NP-khó và không hề đơn giản. Lin và cộng sự. (2014) lần đầu tiên trình bày cách tiếp cận dựa trên GA để ẩn các HUI nhạy cảm bằng cách chèn các giao dịch giả để tăng tổng tiện ích trong cơ sở dữ liệu. Tuy nhiên, cách tiếp cận này dẫn đến chi phí nhân tạo cao do quy mô giao dịch tăng lên và nhiều không phải HUI trở thành HUI sau quá trình vệ sinh.

Bảng 1. Cơ sở dữ liệu định lượng.

THỜI GIAN	Các mặt hàng và số lượng của chúng
1	Đ:6,
2	F:E:6
3	Đ:5,
4	B:5,
5	C:8,
6	Đ:4,
7	B:2, C:3,
8	A:7, B:3,
9	E:2 E:4
10	A:5, B:2,
11	E:C:1,
12	Đ:3, E:3

3. Sơ bộ và nêu vấn đề

Cho I = {i1 , i2 , . . . , im} là một tập hữu hạn gồm m mục riêng biệt. Cơ sở dữ liệu định lượng là một tập hợp các giao dịch D = {T1, T2, . . . , Tn}, trong đó mỗi giao dịch Tq ∈ D (1 ≤ q ≤ n) là tập con của I và có mã định danh duy nhất q, được gọi là TID của nó. Ngoài ra, mỗi mặt hàng ij trong giao dịch Tq đều có số lượng mua (hữu dụng nội tại) ký hiệu là q(ij , Tq). Bảng lợi nhuận ptable = {pr1 , pr2 , . . . , prm} cho biết giá trị lợi nhuận (đơn vị lợi nhuận) prj của từng mặt hàng ij . Một tập hợp k phần tử riêng biệt X = {i1 , i2 , . . . , ik } sao cho X ⊆ I được gọi là tập mục k, trong đó k là độ dài của tập mục. Tập mục X được gọi là có trong giao dịch Tq nếu X ⊆ Tq. Ngưỡng tiện ích tối thiểu (MUT) δ được người dùng đặt theo sở thích của mình.

Một ví dụ minh họa được trình bày trong Bảng 1. Nó sẽ được sử dụng làm ví dụ thực tế cho phần còn lại của bài viết này. Nó chứa 12 giao dịch và 5 mục riêng biệt, được biểu thị bằng các chữ cái từ (A) đến (F). Giá trị lợi nhuận (tiện ích bên ngoài) của từng mặt hàng được đặt là ptable = {A:7, B:15, C:10, D:6, E:2, F:1}. MUT được đặt thành (δ = 20%).

Định nghĩa 1: Tiện ích của mục ij trong giao dịch Tq được ký hiệu là u(ij , Tq) và được định nghĩa là:

$$u(ij , Tq) = q(ij , Tq) \times pr(ij). \quad (1)$$

Ví dụ: Tiện ích của mục (A) trong giao dịch T3 được tính như sau: u(A, T3) = 5 × 7 (= 35). Định nghĩa 2: Tiện ích của tập mục X trong giao dịch Tq được ký hiệu là u(X, Tq) và được định nghĩa là:

$$u(X, Tq) = \sum_{ij \subseteq X \wedge X \in T} qu(ij , Tq). \quad (2)$$

Ví dụ: Tiện ích của tập mục (AE) trong giao dịch T3 được tính như sau: u(AE, T3) = 35 + 2 (= 37)

Định nghĩa 3: Tiện ích của tập mục X trong cơ sở dữ liệu D được ký hiệu là u(X) và được định nghĩa là:

$$u(X) = \sum_{X \subseteq T \, q \wedge T \, q \in D} u(X, Tq). \quad (3)$$

Ví dụ, tiện ích của tập mục (AE) trong D được tính như sau: u(AE) = 37 + 30 + 53 + 45 + 27 (= 192). Định nghĩa 4: Tiện ích giao dịch của giao dịch Tq được ký hiệu là tu(Tq) và được định nghĩa là:

$$tu(Tq) = \sum_{X \in T} qu(X,Tq). \quad (4)$$

Ví dụ, tu(T1) được tính như sau: tu(T1) = 36 + 1 (= 37).

Định nghĩa 5: Việc sử dụng tập mục X theo trọng số giao dịch được ký hiệu là TWU(X) và được xác định BẰNG:

$$TWU(X) = \sum_{X \subseteq T \wedge T \in D} tu(Tq). \quad (5)$$

Ví dụ: TWU(AE) được tính như sau: TWU(AE) = 37 + 30 + 98 + 75 + 27 (= 267). Định nghĩa 6: Tổng tiện ích của cơ sở dữ liệu D được ký hiệu là TU và được định nghĩa là:

$$TU' = \sum_{T \in D} tu(Tq). \quad (6)$$

Ví dụ: tổng tiện ích của cơ sở dữ liệu D được tính như sau: TU = 37 + 12 + 37 + 77 + 82 + 30 + 72 + 98 + 8 + 75 + 11 + 27 (= 566).

Định nghĩa 7: Tập mục X là HUI trong cơ sở dữ liệu D nếu tiện ích của nó không nhỏ hơn tiện ích tối thiểu số được xác định là TU × δ. Do đó, tập hợp các HUI được định nghĩa là:

$$HUI \leftarrow \{X | u(X) \geq TU \times \delta\}. \quad (7)$$

Giả sử rằng MUT được đặt thành δ (= 20%). Trong ví dụ này, các HUI được phát hiện và tiện ích của chúng là {A:168, B:180, C:120, AB:159, AE:192, ABE:173}.

ĐịnCho HS = {s1 , s2 , . . . , sk } biểu thị tập hợp các HUI nhạy cảm bị ẩn trong cơ sở dữ liệu D. Ví dụ: giả sử rằng hai tập mục (B) và (ABE) được coi là các HUI nhạy cảm mà

cần phải được ẩn giấu. Do đó HS = {B,

Trong PPDM truyền thống (Wu và cộng sự, 2007), ba tác dụng phụ tiêu chuẩn được xem xét để đánh giá hiệu suất của phương pháp vệ sinh, đó là lỗi ẩn (HF), chi phí bị thiếu (MC) và chi phí nhân tạo (AC). HF là tập hợp các tập mục nhạy cảm mà quá trình dọn dẹp không thể che giấu được. Nếu số lượng tập mục trong HF lớn, điều đó có nghĩa là quá trình dọn dẹp không thể che giấu thành công thông tin nhạy cảm. Nếu số lượng tập mục trong HF bằng 0, điều đó cho biết rằng tất cả thông tin nhạy cảm đã bị ẩn. MC là tập hợp các tập mục không nhạy cảm trước quá trình dọn dẹp nhưng đã bị quá trình đó ẩn đi. Nếu số lượng tập mục trong MC cao, điều đó có nghĩa là nhiều tập mục không nhạy cảm nhưng được coi là quan trọng đã bị ẩn hoặc thiếu. Do đó, nó dẫn đến mất thông tin, từ đó có thể đưa ra quyết định sai lầm hoặc sử dụng các chiến lược bán hàng không hiệu quả. AC là tập hợp các tập mục không được phát hiện bởi quá trình khai thác dữ liệu trước khi dọn dẹp nhưng được phát hiện sau quá trình dọn dẹp. Số lượng tập mục trong AC cao, điều đó cho thấy nhiều tập mục dư thừa hoặc không cần thiết được phát hiện và coi là quan trọng khi áp dụng kỹ thuật khai thác dữ liệu vào cơ sở dữ liệu đã được làm sạch. Đối với PPUM (Yeh & Hsu, 2010), các tiêu chí tương tự như PPDM (Wu và cộng sự, 2007) được sử dụng và được xác định như sau.

Định nghĩa 9: Cho α (= HF) là các HUI nhạy cảm mà quá trình khử trùng không thể che giấu được, tức là số lượng HUI nhạy cảm vẫn xuất hiện trong cơ sở dữ liệu sau quá trình khử trùng. Về mặt hình thức, nó được định nghĩa là:

$$\alpha = HS \cap HUIs', \quad (8)$$

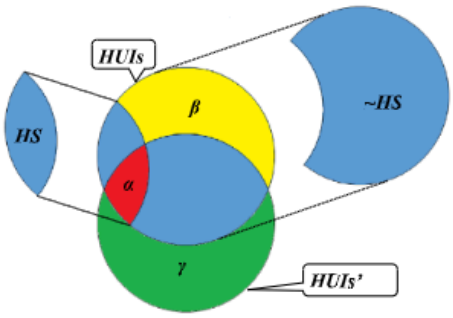
trong đó HS là tập hợp các HUI nhạy cảm trước quá trình vệ sinh và HUI′ là tập hợp các HUI sau quá trình vệ sinh.

Định nghĩa 10: Đặt β (= MC) là các HUI bị thiếu, tức là các HUI không nhạy cảm nhưng quá trình vệ sinh đã ẩn:

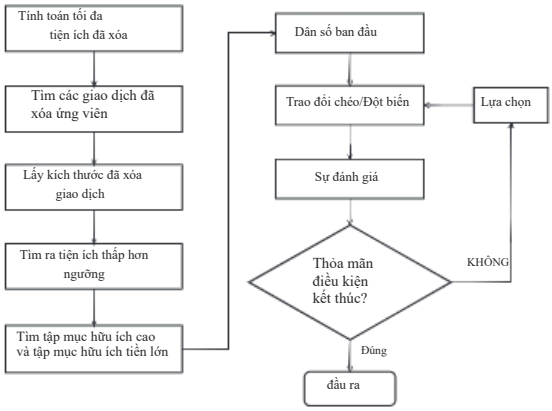
$$\beta = \sim HS - HUI'. \quad (9)$$

Định nghĩa 11: Đặt γ (= AC) là các tập mục không phải là HUI trước quá trình vệ sinh nhưng hiện là HUI, đó là điểm khác biệt giữa HUI′ và HUI:

$$\gamma = HUI' - HUI, \quad (10)$$



Hình 1. Mối quan hệ giữa ba tác dụng phụ và HUI được phát hiện trước và sau khi vệ sinh.



Hình 2. Lưu đồ của thuật toán PPUMGAT được thiết kế.

trong đó HUI′ là tập hợp các HUI thu được sau quá trình khử trùng. Mối quan hệ giữa ba tác dụng phụ này và các tập vật phẩm có tính tiện ích cao được phát hiện trước đó (HUI) và sau khi vệ sinh (HUI′) được minh họa trong Hình 1. Vấn đề: Với một tập hợp các HUI nhạy cảm, vấn đề PPUM (Yeh & Hsu, 2010) sử dụng GA (Holland, 1992) và xóa giao dịch là tìm một tập hợp giao dịch thích hợp để xóa như một giải pháp tối ưu để ẩn càng nhiều HUI nhạy cảm càng tốt, đồng thời giảm thiểu ba tác dụng phụ (HF, MC và AC).

4. Đề xuất thuật toán PPUM dựa trên GA

Tìm lời giải tối ưu cho bài toán PPUM là bài toán NP-khó khi xét HF, MC và AC. Các phương pháp tiếp cận truyền thống đã được thiết kế để che giấu các HUI nhạy cảm bằng cách giảm số lượng mua các tập mục và do đó làm giảm tiện ích của các tập mục nhạy cảm. Cách tiếp cận này có thể hiệu quả nhưng có thể tạo ra nhiều quy tắc hoặc thông tin sai lệch vì nó không có cơ chế lựa chọn giao dịch nào cần được sửa đổi để giảm thiểu tác dụng phụ. Do đó, bài viết này trình bày một thuật toán dựa trên GA (Holland, 1992) để làm sạch cơ sở dữ liệu nhằm ẩn các HUI nhạy cảm bằng cách tìm các giao dịch thích hợp để xóa. Thuật toán dựa trên GA cho PPUM sử dụng tính năng xóa giao dịch có tên PPUMGAT

do đó được thiết kế để ẩn các HUI nhạy cảm với tác dụng phụ tối thiểu. Mỗi nhiễm sắc thể bao gồm một số gen và mỗi gen đại diện cho một giao dịch sẽ bị xóa theo phương pháp được thiết kế. Sơ đồ của thuật toán đề xuất được hiển thị trong Hình 2.

4.1. Chức năng tập thể dục

Trong thuật toán PPUMGAT được thiết kế, hàm thích nghi linh hoạt được đưa ra để đánh giá mức độ tốt của từng nhiễm sắc thể được xử lý. Trong chức năng thể dục được phát triển, ba mức trọng lượng tương ứng được gán cho ba tác dụng phụ, tùy theo sở thích của người dùng. Số lượng trường hợp của từng tác dụng phụ (HF, MC và AC) được xem xét trong hàm thích ứng được thiết kế, được sử dụng trong quá trình tiến hóa.

Định nghĩa 12: Hàm thích nghi được đề xuất để đánh giá mức độ tốt của nhiễm sắc thể được xử lý trong PPUM được định nghĩa là:

$$\text{thể lực}(ci) = |a| \times w1 + |b| \times w2 + |c| \times w3 , \quad (11)$$

trong đó w1, w2 và w3 là ba trọng số có thể điều chỉnh được xác định theo sở thích của người dùng. Như đã nêu trong báo cáo vấn đề, mục đích của PPUM như được định nghĩa trong bài viết này là tìm các giao dịch thích hợp sẽ bị xóa để ẩn các HUI nhạy cảm, đồng thời giảm thiểu ba tác dụng phụ do quá trình tiến hóa.

4.2. Thuật toán PPUMGAT được đề xuất

Mã giả đầy đủ của thuật toán PPUMGAT được thiết kế được trình bày trong Thuật toán 1.

Thuật toán 1: Thuật toán PPUMGAT đề xuất

Đầu vào: D, cơ sở dữ liệu định lượng; ptable, bảng lợi nhuận, Su, ngưỡng hữu dụng trên (tối thiểu); M, số lượng nhiễm sắc thể trong quần thể; N, số lần lặp lại trong quá trình tiến hóa.
Đầu ra: D', một cơ sở dữ liệu đã được làm sạch.
1 cho mỗi si ∈ HS do
2 MDU += TWU(si) − TU × S u;
3 với mỗi Tq ∈ D thực hiện
4 phép tính tu(Tq);
5 với mỗi si ∈ HS làm
6 nếu si ∈ T q và tu(T q) < MDU thì
7 Candi_Delete ← Tq;
8 sắp xếp các giao dịch trong Candi_Delete theo thứ tự tăng dần của tu;
9 tập m = 0, tổng = 0;
10 cho mỗi Tq trong Candi_Delete thực hiện
11 nếu tổng < MDU thì
12 tổng+ = tu(Tq);
13 m = m + 1;
14 đặt kích thước của nhiễm sắc thể thành m;
15 nhiễm sắc thể M được tạo ngẫu nhiên làm quần thể ban đầu;
16 trong khi không đạt được tiêu chí chấm dứt
17 đối với mỗi nhiễm sắc thể ci trong số M nhiễm sắc thể trong quần thể thì
18 thực hiện hoạt động chéo;
19 thực hiện thao tác đột biến;
20 đánh giá mức độ phù hợp(ci);
21 chọn nhiễm sắc thể M/2 hàng đầu trong quần thể;
22 tạo ngẫu nhiên nhiễm sắc thể M/2 ở thế hệ tiếp theo;
23 thu được nhiễm sắc thể ci tối ưu với giá trị thích nghi tối thiểu từ M;
24 xóa T q của ci khỏi D thành D';
25 trả lại D';

Đầu tiên, Tiện ích bị xóa tối đa được tính toán (MDU), được định nghĩa là tổng chênh lệch giữa TWU của từng tập mục nhạy cảm và số lượng tiện ích tối thiểu (Dòng 1 đến 2).

Định nghĩa 13: Tiện ích bị xóa tối đa được ký hiệu là MDU và được định nghĩa là:

$$MDU = \sum_{\forall si \in HS} (TWU(si) - TU \times \delta). \quad (12)$$

Trong thuật toán được thiết kế, các giao dịch được chọn để xóa nhằm ẩn các HUI nhạy cảm. Trong quy trình xóa giao dịch này, mỗi giao dịch được đánh giá và dự kiến là giao dịch dự kiến để xóa trong Candi_Delete nếu nó chứa ít nhất một HUI nhạy cảm. Ngoài ra, giá trị tiện ích bị xóa tối đa được tính toán từ các HUI nhạy cảm được xác định trước. Để giảm thiểu tác dụng phụ MC, nên xóa các giao dịch có tiện ích giao dịch không nhỏ hơn MDU (Dòng 3–7). Quá trình này có thể giúp tìm ra giải pháp tối ưu về các giao dịch cần xóa để giảm thiểu tác dụng phụ AC.

Định nghĩa 14: Mỗi giao dịch chứa ít nhất một tập mục hữu ích cao nhạy cảm trong HS và có tiện ích giao dịch không nhỏ hơn MDU được dự đoán là giao dịch đã xóa ứng viên:

$$Candi\_Delete \leftarrow \{Tq | Tq \in D, \exists si \in HS, si \subseteq Tq \wedge tu(Tq) < MDU\}. \quad (13)$$

Các giao dịch trong Candid_Delete được sắp xếp theo thứ tự tăng dần của tu (Dòng 8). Các tiện ích giao dịch trong Candid_Delete sau đó được tính tổng cho đến khi giá trị lớn hơn MDU. Sau đó, số lượng giao dịch đã được tổng hợp sẽ được sử dụng làm độ dài nhiễm sắc thể cho quá trình tiến hóa (Dòng 9–14). Sau đó, một bộ nhiễm sắc thể được tạo ra làm quần thể ban đầu, trong đó các gen của mỗi nhiễm sắc thể được chọn ngẫu nhiên từ Candi_Delete (Dòng 15). Trong thuật toán PPUMGAT được đề xuất, một nhiễm sắc thể đại diện cho một giải pháp khả thi, đó là một tập hợp các giao dịch sẽ bị xóa để ẩn các HUI nhạy cảm. Mỗi gen của nhiễm sắc thể đại diện cho ID của một giao dịch trong Candi_Delete. Lưu ý rằng việc tìm ra giải pháp tối ưu cho PPUM từ các nhiễm sắc thể được tạo ra trong quần thể là một vấn đề NP-khó và gen nhiễm sắc thể được phép lấy giá trị null. Ngoài ra, một giao dịch có thể được chọn nhiều lần trong cùng một nhiễm sắc thể nhưng nó chỉ có thể bị xóa một lần trong quá trình tiến hóa.

Các hoạt động lai ghép và đột biến cũng được thực hiện để cập nhật các nhiễm sắc thể cho lần lặp tiếp theo của quá trình tiến hóa (Dòng 18 đến 19). Sau đó, mỗi nhiễm sắc thể được đánh giá bằng hàm thích ứng được thiết kế (Dòng 20). Một nửa số nhiễm sắc thể, có giá trị hàm thích hợp thấp nhất, được giữ lại cho thế hệ tiếp theo và nửa còn lại được tạo ngẫu nhiên (Dòng 21 đến 22). Quy trình này sau đó được thực hiện lặp đi lặp lại cho đến khi đáp ứng tiêu chí kết thúc (tức là số lần lặp tối đa nhất định trong GA) (Dòng 16–22). Sau đó, nhiễm sắc thể có giá trị hàm thích hợp thấp nhất (Dòng 23) được chiếu và ID giao dịch trong nhiễm sắc thể này được chọn làm giao dịch để xóa, do đó ẩn thành công các HUI nhạy cảm (Dòng 24). Cuối cùng, cơ sở dữ liệu được cập nhật và quá trình dọn dẹp hoàn tất (Dòng 24).

4.3. Khái niệm tiền lớn

Trong quá trình tiến hóa, việc đánh giá mức độ tốt của nhiễm sắc thể ở mỗi lần lặp rất tốn thời gian. Quá trình này yêu cầu phải quét liên tục cơ sở dữ liệu gốc để đánh giá ba tác dụng phụ cho từng nhiễm sắc thể để tính toán hàm thích ứng được thiết kế và đặc biệt là đánh giá tác dụng phụ AC. Để tăng tốc quá trình đánh giá này, khái niệm tiền lớn được áp dụng trong thuật toán được thiết kế. Nhờ khái niệm này, có thể tránh được việc quét nhiều lần cơ sở dữ liệu ở mỗi lần đánh giá. Khái niệm tiền lớn được Hong và Wang (2006) đề xuất để duy trì hiệu quả thông tin được phát hiện trong các tình huống động. Nó sử dụng hai ngưỡng gọi là ngưỡng hỗ trợ trên (Su) và ngưỡng hỗ trợ dưới (Sl) để duy trì bộ đệm gồm các tập mục hứa hẹn có mức hỗ trợ trong khoảng [Su, Sl], có xác suất cao là các tập mục phổ biến. Khi các giao dịch bị xóa, tác dụng phụ của AC

Bảng 2. Các giao dịch đã xóa ứng viên (Candi_Delete) và các tiện ích giao dịch của chúng.

THỜI GIAN	Các mặt hàng và số lượng của chúng	bạn
4	B:5,	77
7	B:2, C:3,	72
8	A:7, B:3,	98
10	A:5, B:2, E:5	75

có thể được tính toán trực tiếp bằng cách sử dụng bộ đệm lớn trước và do đó có thể tránh được việc quét nhiều cơ sở dữ liệu. Chiến lược này đơn giản nhưng có thể được sử dụng để duy trì thông tin cập nhật một cách hiệu quả.

Trong thuật toán PPUMGAT được thiết kế, tiện ích xóa tối đa (MDU) trước tiên được tính toán bằng cách sử dụng các HUI nhạy cảm. MDU có thể được coi là giới hạn an toàn được khái niệm tiền lớn sử dụng để tránh quét nhiều lần cơ sở dữ liệu ở mỗi lần lặp và tổng tiện ích (TU) trong cơ sở dữ liệu có thể được coi là kích thước cơ sở dữ liệu trong cơ sở dữ liệu gốc. Dựa trên khái niệm tiền lớn, chúng ta có thể tính toán giá trị SI để giữ tập hợp các tập mục tiện ích tiền lớn đầy hứa hẹn (PUI) trong bộ đệm để tránh thực hiện quét cơ sở dữ liệu.

Định nghĩa 15: Đặt Su là ngưỡng hỗ trợ trên do người dùng xác định và MDU là tiện ích bị xóa tối đa thu được từ các HUI nhạy cảm và TU là tổng tiện ích trong cơ sở dữ liệu gốc. Giá trị SI của khái niệm tiền lớn được xác định lại trong phương pháp đề xuất là:

$$SI = Su \times \left(1 - \frac{MDU}{TU}\right). \quad (14)$$

Nhờ khái niệm pre-large đã được sửa đổi, có thể tính toán các tác dụng phụ mà không cần quét lại cơ sở dữ liệu. Sau đó, các nhiễm sắc thể được tạo ra sẽ được đánh giá bằng hàm thích ứng được thiết kế. Thủ tục này được lặp lại cho đến khi đáp ứng tiêu chí kết thúc (tức là tối đa 100 lần lặp). Đối với mỗi thế hệ, một nửa số nhiễm sắc thể có giá trị thích nghi thấp nhất được chọn làm nhiễm sắc thể còn lại cho thế hệ tiếp theo. Nửa nhiễm sắc thể còn lại được tạo ra bằng cách sử dụng các giao dịch trong cơ sở dữ liệu dự kiến Candi_Delete. Cuối cùng, nhiễm sắc thể có giá trị thích hợp thấp nhất sẽ được giữ lại và ID giao dịch trong nhiễm sắc thể đó sẽ bị xóa khỏi cơ sở dữ liệu ban đầu để khử trùng. Sau đó, cơ sở dữ liệu đã được làm sạch cuối cùng sẽ được xuất ra dưới dạng cơ sở dữ liệu được cập nhật cuối cùng.

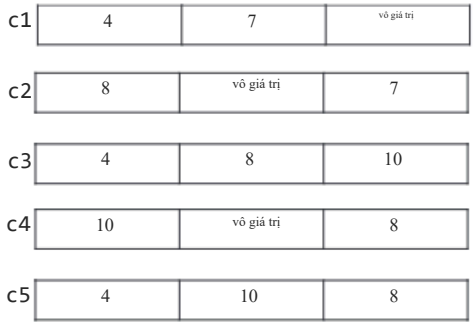
5. Ví dụ minh họa

Một ví dụ được đưa ra để minh họa từng bước thuật toán được đề xuất. Hãy xem xét cơ sở dữ liệu của Bảng 1 và giả sử rằng (B) và (ABE) là các tập mục có tính tiện ích cao nhạy cảm cần được ẩn đi. Do đó, MDU của Bảng 1 được tính như sau: MDU = (TWU(B) − 113) + (TWU(ABE) − 113)(= 269). Sau đó, các giao dịch dự kiến của Candi_Delete từ Bảng 1 được hiển thị trong Bảng 2.

Trong ví dụ đang chạy, MDU(= 269), và tu(T7) + tu(T10) + tu(T4)(= 224) < 269, và tu(T7) + tu(T10) + tu(T4) + tu (T8) (= 322) > 269. Như vậy, số lượng giao dịch tổng hợp là 3, được dùng làm độ dài nhiễm sắc thể (số lượng gen) cho quá trình tiến hóa.

Từ ví dụ trên, độ dài nhiễm sắc thể được đặt thành 3, điều này cho thấy rằng có thể xóa tối đa ba giao dịch để ẩn các HUI nhạy cảm. Số lượng nhiễm sắc thể trong quần thể được đặt thành 5. Do đó, các nhiễm sắc thể được tạo ra trong quần thể được hiển thị trong Hình 3. Đối với nhiễm sắc thể đầu tiên trong Hình 3, các giao dịch T4 và T7 là các giao dịch sẽ bị xóa bởi thuật toán đề xuất.

Giả sử rằng ngưỡng hỗ trợ trên được người dùng đặt thành 20% và tiện ích bị xóa tối đa được tính là MDU (= 113) bằng cách sử dụng các HUI nhạy cảm được xác định trước. Ngưỡng tiện ích thấp hơn là



Hình 3. Các nhiễm sắc thể được tạo ra trong quần thể.

được tính như sau:

$$SI = Su \times \left(1 - \frac{MDUTU}{\quad}\right) = 0,2 \times (1 - 269566) = 10,4\%.$$

Do đó, mỗi tập mục có TWU trong khoảng từ 566 × 0,2 (= 113,2) đến 566 × 0,104 (= 58,86) theo quy trình HUIM, sẽ được thêm vào bộ đệm của PUI. Trong ví dụ này, các HUI và PUI được phát hiện lần lượt là: HUI = {A:168, B:180, C:120, AB:159, AE:192, ABE:173} và PUI = {BE:89, CF: 93}. Dựa trên khái niệm tiền lớn được cải tiến này, có thể tránh được nhiều lần quét cơ sở dữ liệu để đánh giá các tác dụng phụ và đặc biệt là AC.

Hãy xem xét cơ sở dữ liệu của ví dụ đang chạy và giả sử rằng một trong các nhiễm sắc thể trong Hình 3 được cập nhật là (4, 7, null), điều này cho biết rằng các giao dịch T4 và T7 sẽ bị xóa bởi quá trình làm sạch. Nhờ khái niệm pre-large đã được sửa đổi, có thể tính toán các tác dụng phụ mà không cần quét lại cơ sở dữ liệu. Sau đó, các HUI và PUI được phát hiện sẽ được cập nhật tương ứng.

Trong ví dụ này, tổng tiện ích giao dịch được cập nhật thành (566 − 77 − 72) (= 412) và số tiện ích cao hơn được cập nhật thành (412 × 0,2) (= 82,4) Trong ví dụ này, HUI nhạy cảm (B) do đó bị ẩn vì tiện ích của nó được cập nhật lên (180 − 75 − 30) (= 75), thấp hơn 82,4. Tập mục (BE) trở thành HUI vì tiện ích của nó bây giờ lớn hơn (89 > 82.4). Sau đó, nhiễm sắc thể được đánh giá bằng hàm thích ứng được thiết kế.

Trong ví dụ này, tác động phụ HF được tính như sau: {B, ABE}−{B, ABE} = {ABE}, tức là (|α| = 1). Tác dụng phụ MC được tính như sau: {A, B, C, AB, AE, ABE} − {A, C, AB, AE, BE, CF, ABE} = {B}, và do đó (|β| = 1). Tác dụng phụ AC được tính như sau: {A, C, AB, AE, BE, CF, ABE} − {A, B, C, AB, AE, ABE} = {BE, CF}, và do đó (|γ | = 2). Lưu ý rằng đối với các HUI mới (BE và CF), không cần thiết phải quét lại cơ sở dữ liệu để tính toán tiện ích của nó vì tập mục này đã được lưu vào bộ đệm trong bộ PUI. Vì vậy, việc tính toán có thể tránh được. Giả sử rằng ba trọng số của hàm thích nghi lần lượt được đặt là 0,5, 0,25 và 0,25. Độ tốt của nhiễm sắc thể được tạo ra đầu tiên trong quần thể sau đó được đánh giá bằng hàm thích ứng được thiết kế như sau:

$$\text{thể lực}(4, 7, \text{null}) = 0,5 \times 1 + 0,25 \times 1 + 0,25 \times 2 = 1,25.$$

Các nhiễm sắc thể khác sau đó được đánh giá theo cách tương tự để tìm ra các giao dịch tối ưu cần xóa nhằm ẩn các HUI nhạy cảm. Sau đó, các quy trình trên được lặp lại cho đến khi tiêu chí kết thúc và cơ sở dữ liệu đã được làm sạch sẽ được xuất ra dưới dạng cơ sở dữ liệu cập nhật cuối cùng.

6. Kết quả thực nghiệm

Trong các thử nghiệm, ba thuật toán tiến hóa như PPUMGA+insert (Lin et al., 2014) (thuật toán vệ sinh tiến hóa sử dụng chèn giao dịch), PPUMGAT+ (thuật toán PPUMGAT với khái niệm tiền lớn) và PPUMGAT- (thuật toán PPUMGAT không có khái niệm tiền lớn)

Bảng 3. Đặc điểm của các bộ dữ liệu.

Tập dữ liệu	#|D|	#|Tôi|	DepL	MaxLen	Kiểu
Cờ	3196	75	35	35	dày đặc
nấm	8124	120	23	23	dày đặc
Tai nạn	340.183	468	33,8	42	dày đặc
Siêu thị ẩm thực	21.556	1559	4	11	thưa thớt
Bán lẻ	88.162	16.470	10	176	thưa thớt
T10I4D100K	100.000	870	10.1	29	thưa thớt

và thuật toán HHUIF không cải tiến (Yeh & Hsu, 2010) được so sánh về thời gian chạy, ba tác dụng phụ, tính toàn vẹn cơ sở dữ liệu và tính toàn vẹn tiện ích. Sự khác biệt giữa PPUMGAT+ và PPUMGAT- chỉ là việc áp dụng khái niệm tiền lớn. Khái niệm tiền lớn có thể giảm thời gian chạy của quá trình dọn dẹp, nhưng không ảnh hưởng đến các tác dụng phụ về tính toàn vẹn của cơ sở dữ liệu và tính toàn vẹn của tiện ích. Do đó, các kết quả về tác dụng phụ, tính toàn vẹn của cơ sở dữ liệu và tính toàn vẹn của tiện ích không được hiển thị cho PPUMGAT-. Năm bộ dữ liệu trong thế giới thực (Fournier-Viger và cộng sự, 2016) và cơ sở dữ liệu tổng hợp (Agrawal và Srikant, 1994a) do trình tạo cơ sở dữ liệu IBM tạo ra đã được sử dụng trong các thử nghiệm để cho thấy hiệu suất của các phương pháp được thiết kế. Đặc điểm của các bộ dữ liệu được sử dụng trong thử nghiệm được thể hiện trong Bảng 3. Các tham số được sử dụng trong Bảng 3 lần lượt là #|D|: tổng số giao dịch; #|I|: số mục riêng biệt; AvgLen: thời lượng giao dịch trung bình; MaxLen: thời lượng giao dịch tối đa; và Loại: loại tập dữ liệu. Trong các thử nghiệm, MUT và tỷ lệ phần trăm HUI nhạy cảm (SP) được thay đổi để đánh giá hiệu suất của các thuật toán được so sánh.

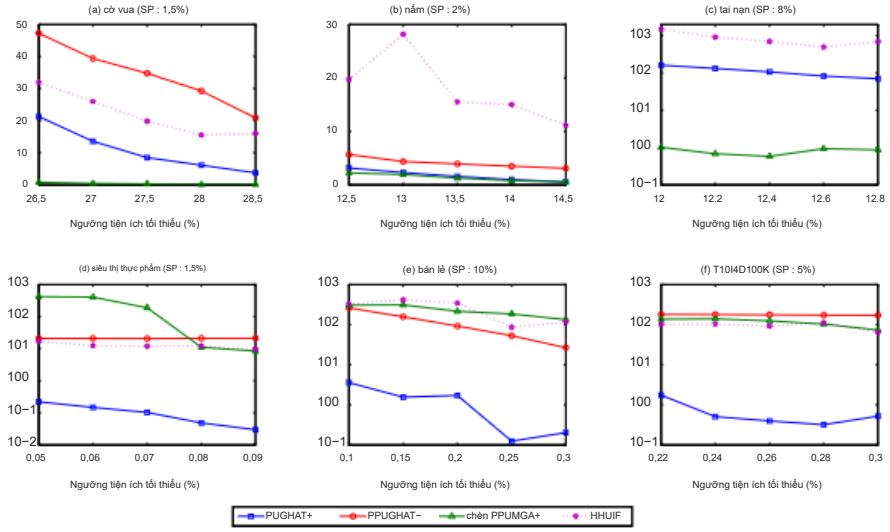
Đối với phương pháp được thiết kế, quy mô quần thể của GA được đặt thành 20 và số lần lặp tối đa được đặt thành 100. Phép lai điểm đơn được áp dụng và tỷ lệ đột biến được đặt thành 0,1. Đối với ba bộ dữ liệu dày đặc (cờ vua, nấm và tai nạn), trọng số của ba tác dụng phụ trong hàm thích nghi lần lượt được đặt thành 0,98, 0,01 và 0,01. Đối với ba tập dữ liệu thưa thớt (foodmart, bán lẻ và T10I4D100K), trọng số của ba tác dụng phụ được đặt thành 0,8, 0,1 và 0,1. Để đánh giá mức độ khác biệt đáng kể giữa các kết quả thu được từ các thuật toán tiến hóa được so sánh, phân tích ANOVA hai chiều đã được thực hiện. ANOVA là một phương pháp thường được sử dụng để phân tích thống kê. ANOVA hai chiều xác định mức độ phản hồi bị ảnh hưởng bởi hai yếu tố (trục x và y). Giá trị F là tỷ lệ giữa giá trị bình phương trung bình của nguồn biến thiên đó và bình phương trung bình dư. Nếu giá trị F lớn chứng tỏ sự khác biệt giữa các thuật toán so sánh là lớn. Nếu giá trị F gần bằng 1, điều đó cho thấy sự khác biệt giữa các thuật toán được so sánh là nhỏ. Giá trị P cho biết sự khác biệt giữa các thuật toán được so sánh có đáng kể hay không. Nếu giá trị P nhỏ hơn 0,05, điều đó cho thấy có sự khác biệt đáng kể giữa các thuật toán được so sánh. Nếu giá trị P lớn hơn 0,05 thì có nghĩa là không có sự khác biệt đáng kể giữa các thuật toán được so sánh. Vì HHUIF là thuật toán không tiến hóa nên không phù hợp để phân tích kết quả của thuật toán này bằng ANOVA hai chiều. Do thuật toán được thiết kế là thuật toán tiến hóa nên phân tích ANOVA hai chiều chỉ được sử dụng để so sánh các thuật toán tiến hóa.

6.1. Thời gian chạy

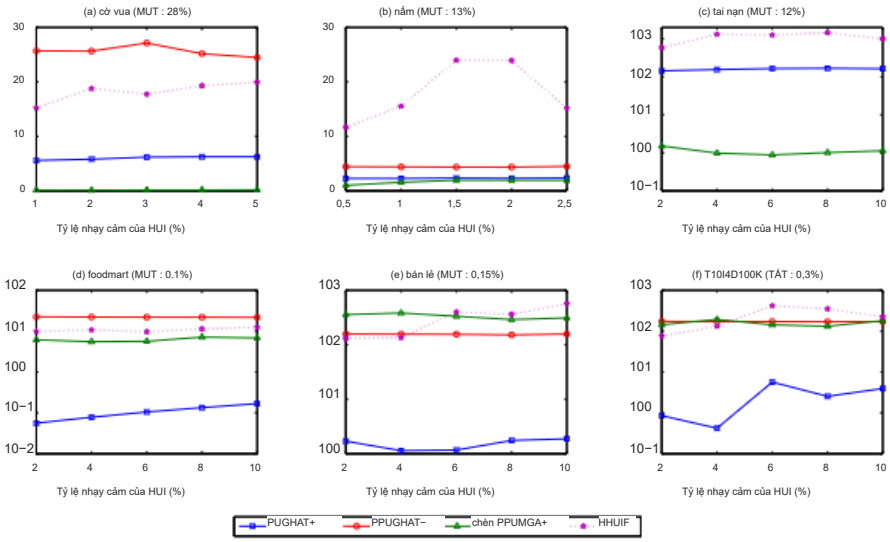
Tiểu mục này so sánh thời gian chạy của các thuật toán cho tất cả sáu bộ dữ liệu. Kết quả thời gian chạy của các thuật toán ghi các MUT khác nhau và một tỷ lệ cố định các HUI nhạy cảm được hiển thị trong Hình 4.

Từ Hình 4, có thể thấy rằng thuật toán PPUMGAT+ được đề xuất chạy nhanh hơn PPUMGAT-. Lý do là PPUMGAT+ có thể tránh thực hiện nhiều lần quét cơ sở dữ liệu không cần thiết để tính toán ba tác dụng phụ nhằm đánh giá nhiễm sắc thể trong quá trình tiến hóa. Nhưng thuật toán PPUMGAT+ được đề xuất yêu cầu nhiều thời gian chạy hơn PPUMGA+insert cho ba tập dữ liệu dày đặc. Lý do là thuật toán chèn PPUMGA+ cũng áp dụng cả khái niệm tiền lớn và cách tiếp cận GA để tìm ra giải pháp tối ưu nhằm thêm các giao dịch giả nhằm ẩn các HUI nhạy cảm.

Tiện ích tối đa để chèn vào PPUMGA+insert phụ thuộc vào tỷ lệ giữa tiện ích tập mục nhạy cảm tối đa và MUT. Đối với các tập dữ liệu dày đặc, tiện ích giao dịch trung bình cao hơn



Hình 4. Thời gian chạy ghi nhiều MUT khác nhau.



Hình 5. Thời gian chạy ghi nhiều SP khác nhau.

hơn đối với các tập dữ liệu thưa thớt. Thuật toán chèn PPUMGA+ chỉ thêm một số lượng nhỏ giao dịch để ẩn các HUI nhạy cảm. Do đó, thời gian chạy của thuật toán chèn PPUMGA+ nhỏ hơn thuật toán PPUMGAT+. Dựa trên phân tích ANOVA hai chiều, có sự khác biệt đáng kể giữa thời gian chạy của thuật toán chèn PPUMGAT+ và PPUMGAT+ ($F = 58,901$, $P < 0,001$, trong Hình 4(a); $F = 216,684$, $P < 0,01$, trong Hình 4(b); $F = 43,823$, $P = 0,003 < 0,05$, trong Hình 4(c)). Ở đó

cũng là sự khác biệt đáng kể giữa thời gian chạy của thuật toán PPUMGAT+ và PPUMGAT- cho bộ dữ liệu cờ vua (T = 7,376, P < 0,001, trong Hình 4(a)). Đối với tập dữ liệu tai nạn, kết quả được hiển thị trong Hình 4(c). Vì PPUMGAT- vượt quá 104 giây trên tập dữ liệu đó nên nó bị bỏ qua vì mục đích so sánh.

Đối với các tập dữ liệu thưa thớt, kết quả được hiển thị trong Hình 4(d)–(f). PPUMGA+ được thiết kế nhanh hơn tới hai bậc so với PPUMGA+insert. Dựa trên phân tích ANOVA hai chiều, có sự khác biệt đáng kể giữa thời gian chạy của thuật toán chèn PPUMGA+ và PPUMGA+ (F = 4,760, P = 0,043 < 0,05, trong Hình 4(d); F = 24,408, P < 0,001 , trong Hình 4(e); F = 178,373, P < 0,001, trong Hình 4(f)). Đối với tập dữ liệu foomart, không có sự khác biệt đáng kể giữa thuật toán chèn PPUMGAT- và PPUMGAT+ (P = 0,108). Tuy nhiên, có một sự khác biệt đáng kể giữa thuật toán chèn PPUMGAT- và PPUMGAT+ (P = 0,027 < 0,05). Đối với tất cả các thử nghiệm trong Hình 4, thời gian chạy của PPUMGAT+ giảm khi MUT tăng. Những kết quả này là hợp lý vì khi MUT tăng với SP cố định thì số lượng HUI nhạy cảm sẽ giảm. Do đó, sẽ mất ít thời gian hơn để tìm ra giải pháp vệ sinh tối ưu. Kết quả thử nghiệm với các SP khác nhau có MUT cố định cho sáu bộ dữ liệu được hiển thị trong Hình 5.

Có thể thấy trong Hình 5, thuật toán PPUMGAT+ được thiết kế chạy nhanh hơn thuật toán PPUMGAT- cho tất cả các tập dữ liệu và nhanh hơn PPUMGA+insert cho các tập dữ liệu thưa thớt (F = 3018.041, P < 0,001, trong Hình 5(d); F = 318,946, P < 0,001, trong Hình 5(e)); Đối với kết quả của Hình 4, thời gian chạy của PPUMGAT+ lớn hơn so với PPUMGA+insert đối với các tập dữ liệu dày đặc (F = 2423,296, P < 0,001, trong Hình 5(a); F = 194,929, P < 0,001, trong Hình 5(b); F = 1188,355, P < 0,001, trong Hình 5(c)). Ngoài ra, có sự khác biệt đáng kể giữa thời gian chạy của thuật toán PPUMGAT+ và PPUMGAT- (P < 0,001, trong Hình 5(a); P = 0,007 < 0,05, trong Hình 5(b)), điều này cho thấy rằng khái niệm này rất hữu ích để giảm thời gian chạy dựa trên phân tích ANOVA hai chiều. Nói chung, thời gian chạy tăng khi SP tăng. Điều này là hợp lý vì khi SP tăng lên, những thông tin nhạy cảm hơn cần được ẩn đi. Đối với tập dữ liệu tai nạn, thuật toán PPUMGAT- vượt quá 104 giây và do đó bị bỏ qua trong Hình 5(c). Từ các kết quả được hiển thị trong Hình 4 và 5, có thể kết luận rằng thuật toán PPUMGAT+ được đề xuất vượt trội hơn PPUMGAT- cả trong các MUT khác nhau và trong các SP khác nhau. Do đó, khái niệm tiền lớn được điều chỉnh để tăng tốc quá trình dọn dẹp nhằm che giấu các HUI nhạy cảm có thể chấp nhận được. Ngoài ra, thuật toán PPUMGAT+ được đề xuất cũng tốt hơn PPUMGA+insert đối với các tập dữ liệu thưa thớt, có các đặc điểm giống với các tập dữ liệu thực tế hơn.

6.2. tác dụng phụ

Trong phần này, các tỷ lệ tác dụng phụ về lỗi che giấu (HFr), chi phí bị thiếu (MC r) và chi phí nhân tạo (AC r) được sử dụng làm tiêu chí để đánh giá hiệu quả của phương pháp đề xuất. (HFr) là tỷ lệ HUI nhạy cảm mà quá trình khử trùng không thể che giấu được. (MCr) là tỷ lệ các HUI không nhạy cảm bị thiếu sau khi dọn dẹp và (ACr) là tỷ lệ các tập mục đã trở thành HUI do dọn dẹp. Ba tỷ lệ tương ứng với ba tác dụng phụ này được xác định chính thức là:

$$HFr = \frac{|HS' |}{|HS|}, \tag{15}$$

$$MC\ r = | \sim HS - \sim HS' | \ | \sim HS|, \tag{16}$$

$$AC\ r = | \sim HS' - \sim HS| \ | \sim HS|, \tag{17}$$

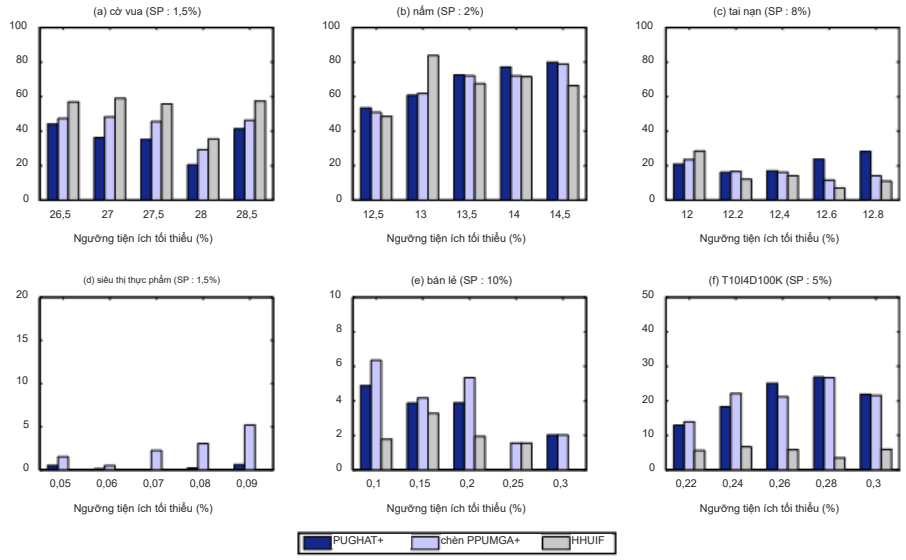
trong đó (HS) là tập hợp các HUI nhạy cảm trước khi khử trùng và HS′ là tập hợp các HUI nhạy cảm sau khi khử trùng.

Bảng 4. Kết quả HFr của HHUIF trên siêu thị thực phẩm đối với các SP khác nhau.

(MUT = 0,1%)					
SP(%)	2	4	6	8	10
HF r .%/	0	11.1	0	11.1	9.091

Bảng 5. Kết quả HFr của HHUIF trên siêu thị thực phẩm đối với các MUT khác nhau.

(SP = 1,5%)					
TÁT (%)	0,05	0,06	0,07	0,08	0,09
HFr (%)	20	15.385	20	0	20



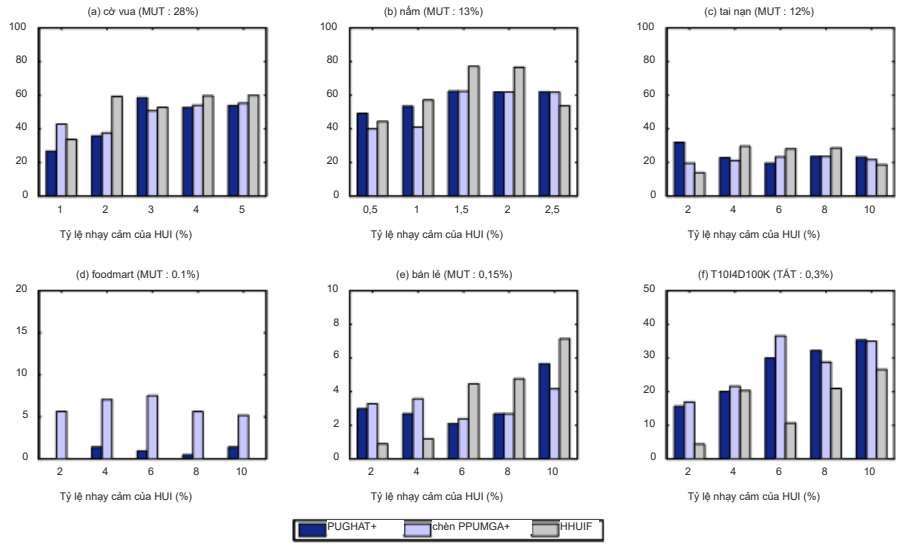
Hình 6. Thiếu chi phí cho các MUT khác nhau.

6.2.1. Ẩn thất bại Tất cả các thuật toán tiến hóa được so sánh đều ẩn thành công tất cả các tập mục HUI nhạy cảm, trong khi thuật toán HHUIF không tiến hóa không có trên tập dữ liệu foodmart. Do đó, kết quả chi tiết cho thuật toán HHUIF được trình bày trong Bảng 4 và 5.

Vì vậy, có thể thấy rằng HHUIF không thể đạt được mục đích của PPUM vì một số HUI nhạy cảm không bị quá trình vệ sinh của nó che giấu.

6.2.2. Thiếu chi phí Kết quả của MC r cho các MUT khác nhau khi SP cố định được hiển thị trong Hình 6.

Có thể thấy trong Hình 6 rằng thuật toán được đề xuất nhìn chung có thể đạt được kết quả tốt hơn về mặt (MC r) so với thuật toán chèn PPUMGA+. Mặc dù HHUIF trực tiếp xóa các mục (không phải giao dịch) để đạt được mục đích của PPUM, thuật toán đề xuất vẫn có thể thu được kết quả tốt hơn trong một số trường hợp theo (MC r). (MC r) cao đối với hầu hết các tập dữ liệu ngoại trừ trong Hình 6(c)–(e). Những kết quả này rất thuyết phục vì có mối quan hệ đánh đổi giữa tác dụng phụ của việc che giấu thất bại và chi phí bị thiếu. Khi các HUI nhạy cảm bị ẩn, nhiều thông tin hơn về các HUI không nhạy cảm có thể



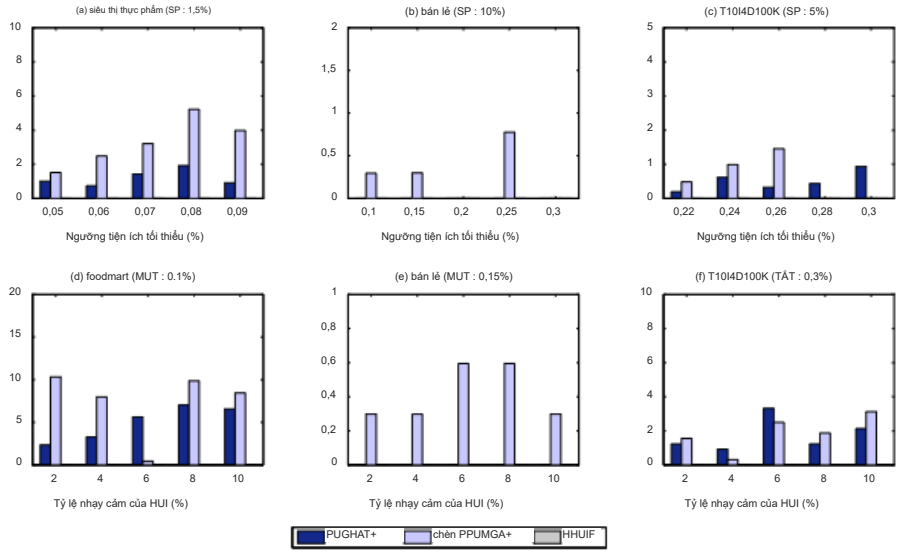
Hình 7. Thiếu chi phí cho các SP khác nhau.

cũng được ẩn đi cùng lúc. Dựa trên phân tích ANOVA hai chiều, có sự khác biệt đáng kể về chi phí còn thiếu đối với các bộ dữ liệu cờ vua, siêu thị thực phẩm và bán lẻ ($F = 22,956$, $P < 0,001$, trong Hình 6(a); $F = 8,994$, $P = 0,009 < 0,05$, trong Hình 6(d); $F = 8,230$, $P = 0,011 < 0,05$, trong Hình 6(e)). Tuy nhiên, không có sự khác biệt đáng kể về chi phí còn thiếu đối với các bộ dữ liệu nấm, tai nạn và T10I4D100K ($F = 2,451$, $P = 0,148 > 0,05$, trong Hình 6(b); $F = 1,888$, $P = 0,241 > 0,05$, trong Hình 6(c); $F = 0,00384$, $P = 0,996 > 0,05$, trong Hình 6(f)).

Từ các kết quả trình bày ở 6(a), (b) và (f), có thể thấy rằng tỷ lệ (MC r) có thể đạt giá trị cao tới 20%. Nguyên nhân là do việc phân bổ HUI cho các bộ dữ liệu này quá dày đặc. Khi các HUI nhạy cảm bị ẩn bởi quá trình khử trùng, các HUI không nhạy cảm hơn cũng bị ẩn do tác dụng phụ. Đối với các tập dữ liệu rất thưa thớt như siêu thị thực phẩm và bán lẻ, độ tương tự giữa các giao dịch là thấp. Kết quả là các HUI nhạy cảm có thể bị ẩn hoàn toàn ở mức thấp (MCr). Kết quả MC r cho các SP khác nhau và MUT cố định được thể hiện trong Hình 7.

Từ kết quả được trình bày trong Hình 7, có thể thấy rằng thuật toán PPUMGAT+ được đề xuất có thể đạt được kết quả tốt hơn thuật toán chèn PPUMGA+ và HHUIF. Khi SP tăng, thuật toán đề xuất và thuật toán chèn PPUMGA+ vẫn có thể ẩn hoàn toàn tất cả các HUI nhạy cảm và do đó tỷ lệ (HFr) bằng 0 đối với tất cả các tập dữ liệu. Tuy nhiên, HHUIF không thể ẩn tất cả các HUI nhạy cảm đối với tập dữ liệu foodmart. Lý do đã được đề cập trước đó. Thuật toán PPUMGAT+ được đề xuất đạt được kết quả tốt hơn về mặt (MCr) trong hầu hết các trường hợp, so với hai thuật toán còn lại. Tuy nhiên, dựa trên phân tích ANOVA hai chiều, không có sự khác biệt đáng kể về chi phí còn thiếu đối với tất cả các tập dữ liệu ($F = 0,471$, $P = 0,641 > 0,05$, trong Hình 7(a); $F = 2,744$, $P = 0,124 > 0,05$, trong Hình 7(b); $F = 0,772$, $P = 0,429 > 0,05$, trong Hình 7(c); $> 0,01$, trong Hình 7(f)) nhưng tập dữ liệu foodmart ($F = 135,375$, $P < 0,001$, trong Hình 7(d)). Tỷ lệ (MC r) tăng khi SP tăng do có mối quan hệ đánh đổi giữa việc che giấu thất bại và chi phí bị thiếu.

6.2.3. Chi phí nhân tạo Đối với tác dụng phụ của chi phí nhân tạo (ACr), các thuật toán so sánh tạo ra tỷ lệ gần như bằng 0 hoặc bằng 0 cho (AC r) cho ba bộ dữ liệu dày đặc (cờ vua, nấm và tai nạn). Vì vậy, kết quả của



Hình 8. Chi phí nhân tạo cho các SP và MUT khác nhau.

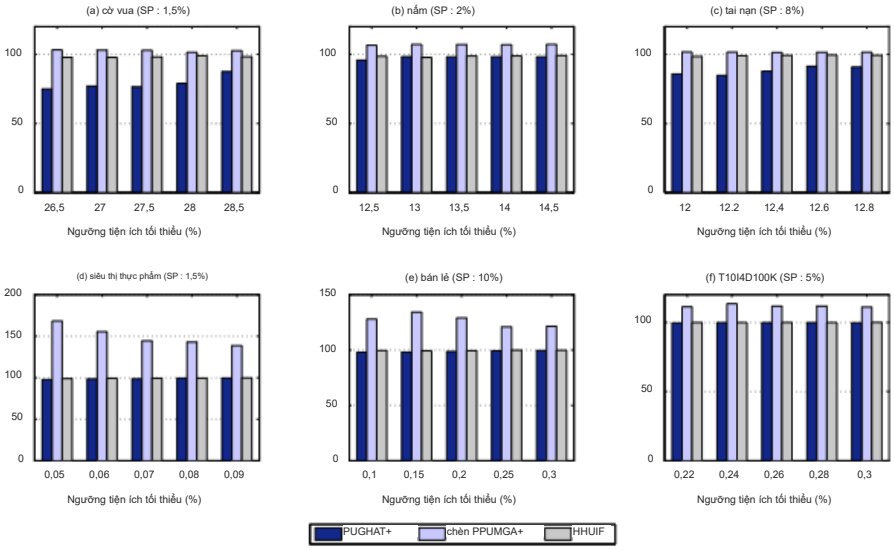
(AC r) cho các tập dữ liệu dày đặc không được đưa ra. Kết quả của (ACr) cho các MUT và SP khác nhau cho ba bộ dữ liệu thưa thớt được hiển thị trong Hình 8.

Từ kết quả trong Hình 8, rõ ràng thuật toán PPUMGAT+ được đề xuất có kết quả tốt hơn so với PPUMGA+insert xét về (ACr) trong hầu hết các trường hợp. Tuy nhiên, HHUIF không tạo ra bất kỳ chi phí nhân tạo nào trong mọi trường hợp. Lý do là ba bộ dữ liệu rất thưa thớt. Do đó, việc thực hiện xóa mục có thể không ảnh hưởng nhiều đến các tập mục không nhạy cảm khác. Mặc dù thuật toán PPUMGAT+ được đề xuất và thuật toán chèn PPUMGA+ được so sánh tạo ra các tác dụng phụ của (AC r) đối với các tập dữ liệu thưa thớt (foodmart, bán lẻ và T10I4D100K), như trong Hình 8(a) và (d), (b) và (e), và (c) và (f), tỷ lệ (ACr) thu được cho siêu thị thực phẩm, bán lẻ và T10I4D100K lần lượt nhỏ hơn 10, 1 và 4% đối với các thuật toán được so sánh và thuật toán PPUMGAT+ được đề xuất luôn vượt trội so với Thuật toán chèn PPUMGA+. Ngoài ra, dựa trên phân tích ANOVA hai chiều, có sự khác biệt đáng kể giữa PPUMGAT+ và PPUMGA+insert về chi phí nhân tạo cho tập dữ liệu foomart ($F = 16,979$, $P = 0,001 < 0,05$, trong Hình 8(a) ;$F = 113,797$, $P = 0,003 < 0,05$, trong Hình 8(d)) nhưng đối với bán lẻ ($F = 0,723$, $P = 0,072 > 0,05$, trong Hình 8(b); $F = 32,824$, $P < 0,001$, trong Hình 8 (e)) và T10I4D100K ($F = 0,0608$, $P = 0,941 > 0,05$, trong Hình 8(c); $F = 0,0848$, $P = 0,919 > 0,05$, trong Hình 8(f)). Nhìn chung, thuật toán đề xuất có hiệu suất tốt xét về ba tác dụng phụ và đặc biệt là về (HFr) và (AC r) cho tất cả các tập dữ liệu.

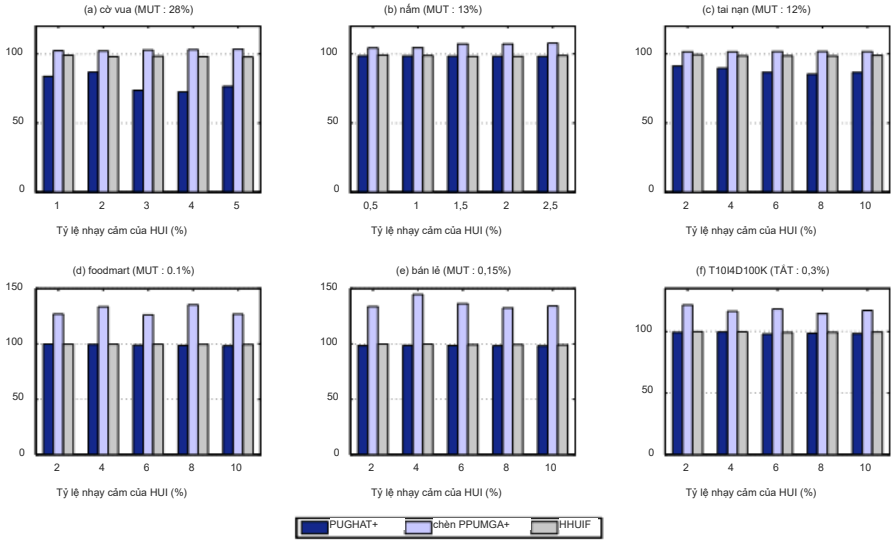
6.3. Tính toàn vẹn dữ liệu

Bên cạnh ba tác dụng phụ được sử dụng trong PPUM và PPDMA, một tiêu chí mới về tính toàn vẹn cơ sở dữ liệu (DI) cho thấy ảnh hưởng của việc xóa giao dịch trên cơ sở dữ liệu gốc, tiêu chí này có thể được sử dụng để xác minh sự giống nhau giữa cơ sở dữ liệu gốc. Nó được định nghĩa là:

$$DI = \frac{|D*|}{|D|}, \tag{18}$$



Hình 9. Tính toàn vẹn của cơ sở dữ liệu cho các MUT khác nhau.



Hình 10. Tính toàn vẹn của cơ sở dữ liệu cho các SP khác nhau.

ở đâu $|D*|$ là kích thước cơ sở dữ liệu sau khi dọn dẹp và $|D|$ là kích thước cơ sở dữ liệu ban đầu trước khi khử trùng. Kết quả của các thuật toán được so sánh cho các MUT khác nhau và các tỷ lệ phần trăm nhạy cảm khác nhau lần lượt được thể hiện trong Hình 9 và 10.

Trong những hình này, rõ ràng là thuật toán PPUMGAT+ duy trì tính toàn vẹn cơ sở dữ liệu cao (gần 100%) cho ba tập dữ liệu thưa thớt và cả tập dữ liệu nấm dày đặc. Dựa trên phân tích ANOVA hai chiều, có thể kết luận rằng thuật toán PPUMGAT+ vượt trội hơn thuật toán chèn PPUMGAT+ đối với các tập dữ liệu thưa thớt và đối với tập dữ liệu nấm ($F = 63,034$, $P < 0,001$, trong Hình 9(b) $F = 42,579$, $P < 0,001$, trong Hình 10(b); $F = 93,435$, $P < 0,001$, trong Hình 9(d); , $P < 0,001$, trong Hình 9(e); $F = 235,843$, $P < 0,001$, trong Hình 10(e); $F = 624,219$, $P < 0,001$, trong Hình 9(f); trong Hình 10(f)). Mặt khác, thuật toán chèn PPUMGA+ ảnh hưởng đến tính toàn vẹn của cơ sở dữ liệu lên tới 20% hoặc thậm chí 50%, đặc biệt đối với các tập dữ liệu bán lẻ và siêu thị thực phẩm.

Ngoài ra, có thể thấy rõ rằng HHUIF luôn cung cấp tính toàn vẹn cơ sở dữ liệu cao cho cả sáu bộ dữ liệu. Lý do là thuật toán HHUIF chỉ xóa các HUI nhạy cảm một phần để đạt được mục đích PPDM, trong khi thuật toán của chúng tôi xóa một tập hợp giao dịch với hiệu ứng kích thước tối thiểu. Đối với ba tập dữ liệu thưa thớt và tập dữ liệu nấm dày đặc, kết quả của thuật toán PPUMGAT+ được đề xuất và thuật toán HHUIF gần như giống nhau. Đối với các tập dữ liệu cờ vua và tai nạn, hiệu suất của thuật toán PPUMGAT+ kém hơn HHUIF và thuật toán chèn PPUMGA+ ($F = 15,777$, $P : 0,002 < 0,05$, trong Hình 9(a); $F = 66,185$, $P = 0,001 < 0,05$, trong Hình 9(c); $F = 50,399$, $P < 0,001$, trong Hình 10(a); $F = 101,505$, $P < 0,001$, trong Hình 10(c)), nhưng nó vẫn duy trì tính toàn vẹn của cơ sở dữ liệu gần 70 –80%. Nhìn chung, thuật toán được thiết kế vượt trội hơn thuật toán chèn PPUMGA+ về tính toàn vẹn của cơ sở dữ liệu đối với các tập dữ liệu thưa thớt.

6.4. Tính toàn vẹn của tiện ích

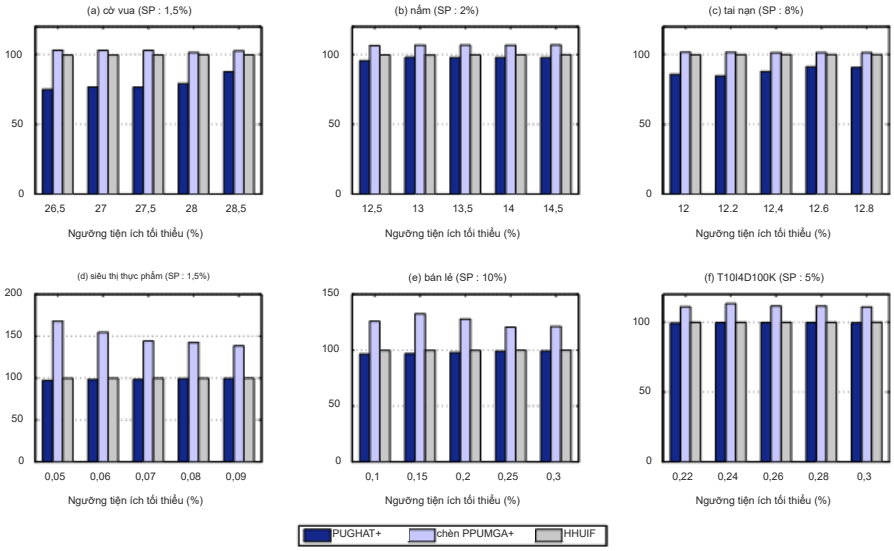
Bên cạnh tính toàn vẹn của cơ sở dữ liệu, tiện ích cũng cần được coi trong PPUM như một tiêu chí quan trọng để đánh giá hiệu suất của các thuật toán được so sánh. Do đó, khái niệm về tính toàn vẹn tiện ích (UI) được đề xuất để đánh giá sự khác biệt về tổng tiện ích trước và sau khi khử trùng.

$$UI = TU^{*}TU, \tag{19}$$

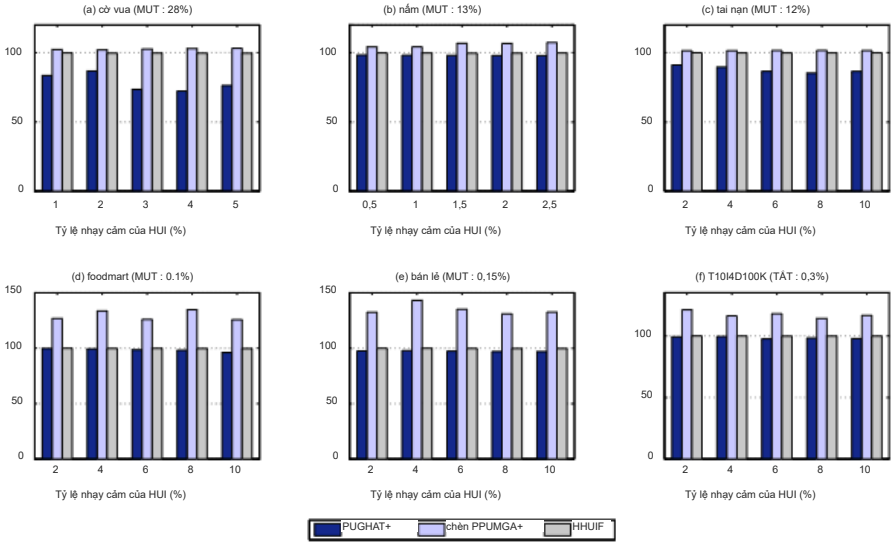
trong đó TU^{*} là tổng tiện ích của cơ sở dữ liệu sau khi áp dụng quy trình dọn dẹp và TU là tổng tiện ích của cơ sở dữ liệu trước quá trình dọn dẹp.

Kết quả của các thuật toán so sánh cho các MUT khác nhau và các tỷ lệ phần trăm nhạy cảm khác nhau được trình bày tương ứng trong Hình 11 và 12.

Rõ ràng là thuật toán chèn PPUMGA+ không thể đảm bảo tính toàn vẹn tiện ích cao cho các tập dữ liệu thưa thớt khi so sánh với thuật toán PPUMGA+ và thuật toán HHUIF được thiết kế. Có thể thấy rằng tổng tiện ích của thuật toán được thiết kế vẫn ổn định và gần 100% đối với hầu hết các tập dữ liệu và vượt trội hơn thuật toán chèn PPUMGA+ ngoại trừ các tập dữ liệu cờ vua và tai nạn dày đặc. Ngoài ra, có sự khác biệt đáng kể giữa PPUMGAT+ và PPUMGA+insert dựa trên phân tích ANOVA hai chiều cho cơ sở dữ liệu nấm dày đặc và ba cơ sở dữ liệu thưa thớt ($F = 55,073$, $P < 0,001$, trong Hình 11(b); $F = 41,557$, $P < 0,001$, trong Hình 12(b); $F = 5,834$, $P < 0,001$, trong Hình 11(d); $F = 169,826$, $P < 0,001$, trong Hình 11(e); trong Hình 11(f); $F = 151,971$, $P < 0,001$, trong Hình 12(d); $F = 189,492$, $P < 0,001$, trong Hình 12(e);)). Do thuật toán HHUIF chỉ bảo vệ quyền riêng tư bằng cách xóa các HUI nhạy cảm một phần nên tính toàn vẹn tiện ích của nó vẫn rất cao (gần 100%) đối với tất cả các cơ sở dữ liệu. Kết quả của thuật toán PPUMGAT+ được đề xuất gần như giống nhau đối với tập dữ liệu nấm dày đặc và ba tập dữ liệu thưa thớt. Tuy nhiên, đối với các tập dữ liệu rất dày đặc như cờ vua và tai nạn, thuật toán PPUMGA+ được thiết kế hoạt động kém hơn thuật toán chèn HHUIF và PPUMGA+. Điều này là hợp lý vì đối với các tập dữ liệu rất dày đặc, khi các HUI nhạy cảm bị xóa, nhiều HUI liên quan cũng bị xóa và tổng tiện ích trong tập dữ liệu đã được làm sạch sẽ giảm đi. Dựa trên phân tích ANOVA hai chiều, có sự khác biệt đáng kể giữa hai thuật toán so sánh dựa trên GA ($F = 15,704$, $P = 0,002 < 0,05$, trong Hình 11(a); $F = 70,094$, $P = 0,001 < 0,05$, trong Hình 11(c); $F = 50,306$, $P < 0,001$, trong Hình 12(a);



Hình 11. Tính toàn vẹn của tiện ích cho các MUT khác nhau.



Hình 12. Tính toàn vẹn của tiện ích cho các SP khác nhau.

7. Kết luận và thảo luận

Sử dụng các kỹ thuật khai thác dữ liệu, có thể dễ dàng phát hiện các mối quan hệ tiềm ẩn hoặc tiềm ẩn giữa các mục trong cơ sở dữ liệu. Do đó, thông tin riêng tư hoặc bí mật cũng có thể bị tiết lộ, điều này có thể gây ra các mối đe dọa về bảo mật và dẫn đến các vấn đề về quyền riêng tư. Vì vậy, PPDM đã nổi lên như một vấn đề quan trọng trong

những năm gần đây. Gần đây, PPUM cũng đã trở thành một vấn đề quan trọng vì nó giải quyết vấn đề bảo quản thông tin bí mật cho HUIM. Vì mục đích của PPDM hoặc PPUM là che giấu thông tin nhạy cảm trong khi vẫn đảm bảo rằng thông tin quan trọng không nhạy cảm vẫn có thể được phát hiện nên việc tìm ra giải pháp tối ưu cho vấn đề này là NP-hard.

Trong bài báo này, trước tiên chúng tôi đề xuất một phương pháp tối ưu hóa để ẩn các HUI nhạy cảm dựa trên GA và hoạt động xóa giao dịch. Một thuật toán có tên PPUMGAT đã được trình bày để tìm ra tập hợp giao dịch tối ưu cần xóa nhằm giảm thiểu ba tác dụng phụ dựa trên chức năng thể dục được thiết kế. Một chiến lược khái niệm tiền lớn cải tiến cũng đã được phát triển để đẩy nhanh quá trình tiến hóa. Dựa trên chiến lược khái niệm tiền lớn được thiết kế, có thể tránh được việc quét nhiều lần cơ sở dữ liệu và do đó thời gian chạy được giảm xuống. Các thử nghiệm được tiến hành đã chỉ ra rằng thuật toán PPUMGAT+ được đề xuất có thể ẩn thành công tất cả các HUI nhạy cảm, đồng thời duy trì tính toàn vẹn của cơ sở dữ liệu và tiện ích ở mức cao.

Bài viết này đã tập trung vào việc xóa giao dịch như một cơ chế để ẩn các tập mục hữu ích cao nhạy cảm. Là một cải tiến sẽ được xem xét trong công việc trong tương lai, mỗi mục trong HUI nhạy cảm trong các giao dịch có thể được mã hóa riêng lẻ dưới dạng gen cho quy trình khử trùng. Hơn nữa, thuộc tính đóng đi xuống theo trọng số giao dịch có thể được áp dụng thêm trong PPUM nếu một số tập mục nhạy cảm trong cơ sở dữ liệu có thể được sửa đổi để giảm tác dụng phụ. Vì PPUM là một vấn đề nghiên cứu mới nổi đã thu hút sự chú ý của nhiều nhà nghiên cứu trong những năm gần đây nên việc phát triển cơ chế ẩn danh hiệu quả cho PPUM cũng là một vấn đề quan trọng.

Tuyên bố tiết lộ

Không có xung đột lợi ích tiềm ẩn nào được các tác giả báo cáo.

Tài trợ

Công trình này được hỗ trợ một phần bởi quỹ mở của Phòng thí nghiệm ứng dụng và khai thác dữ liệu lớn trọng điểm tỉnh Phúc Kiến (Đại học Công nghệ Phúc Kiến; Quỹ khoa học tự nhiên quốc gia Trung Quốc (NSFC) theo [số cấp 61503092]; Dự án Tencent theo [số cấp CCF-Tencent IAGR20160115].

ORCID

Philippe Fournier-Viger http://orcid.org/0000-0002-7680-9899

Tài liệu tham khảo

Aggarwal, CC, Pei, J., & Zhang, B. (2006). Về bảo vệ quyền riêng tư chống lại việc khai thác dữ liệu đối nghịch. ACM SIGKDD
Hội nghị quốc tế về khám phá tri thức và khai thác dữ liệu, trang 510–516. Agrawal, R., & Srikant, R. (1994a). Trình tạo dữ liệu tổng hợp Quest. Lấy từ http://www.Almaden.ibm.com/cs/quest/
syndata.html Agrawal, R., & Srikant, R. (1994b). Thuật toán nhanh để khai thác luật kết hợp trong cơ sở dữ liệu lớn. quốc tế
Hội thảo về Cơ sở dữ liệu rất lớn, Nhà xuất bản Morgan Kaufmann, San Francisco, CA, Hoa Kỳ, trang 487–499.
Agrawal, R., & Srikant, R. (2000). Khai thác dữ liệu bảo vệ quyền riêng tư. Bản ghi ACM SIGMOD, 29, 439–450.
Atallah, M., Elmagarmid, A., Ibrahim, M., Bertino, E., & Verykios, V. (1999). Hạn chế tiết lộ của quy tắc nhạy cảm. các
Hội thảo về Trao đổi Kiến thức và Kỹ thuật Dữ liệu, Chicago, IL, trang 45–52. Bonam, J., Reddy, AR, & Kalyani, G. (2014). Bảo vệ quyền riêng tư trong khai thác quy tắc kết hợp bằng cách bóp méo dữ liệu bằng pso.
Những tiến bộ trong hệ thống thông minh và máy tính, 249, 551–558. Chan, R., Yang, Q., & Shen, YD (2003). Khai thác các tập mục có tính tiện ích cao. Hội nghị quốc tế của IEEE về khai thác dữ liệu,
Melbourne, FL, trang 19–
Chẽng, P., Lin, CW, & Pan, JS (2015). Sử dụng cường điệu để ẩn quy tắc kết hợp bằng cách thêm các mục. XIN VUI LÒNG Một, 10(6), 1–19.
Cheng, P., Roddick, JF, Chu, SC, & Lin, CW (2016). Bảo vệ quyền riêng tư thông qua việc che giấu quy tắc dựa trên sự bóp méo, tham lam
phương pháp. Trí tuệ ứng dụng, 44, 295–306. Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., & Zhu, MY (2002). Công cụ bảo vệ quyền riêng tư khi khai thác dữ liệu phân tán.
Bản tin Khám phá ACM SIGKDD, 4, 28–34.

Evfimievski, A., Srikant, R., Agrawal, R., & Gehrke, J. (2002). Bảo vệ quyền riêng tư khai thác các quy tắc kết hợp. ACM quốc tế

Hội thảo về Khám phá tri thức và Khai thác dữ liệu, trang 217–228. Fournier-Viger, P., Lin, JCW, Gomariz, A., Gueniche, T., Soltani, A., Đặng, Z., & Lam, HT (2016). Thư viện khai thác dữ liệu nguồn mở spmf phiên bản 2. Hội nghị chung châu Âu về học máy và khám phá kiến thức trong cơ sở dữ liệu, trang 36–40.

Fournier-Viger, P., Lin, JCW, Kiran, RU, Koh, YS, & Thomas, R. (2017). Khảo sát khai phá mẫu tuần tự dữ liệu

Khoa học và Nhận dạng Mẫu, 1, 54–77. Giannotti, F., Lakshmanan, LVS, Monreale, A., Pedreschi, D., & Wang, H. (2013). Khai thác hiệp hội bảo vệ quyền riêng tư

quy tắc từ cơ sở dữ liệu giao dịch thuê ngoài. Tạp chí Hệ thống IEEE, 7, 385–395. Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Khai thác các mẫu phổ biến mà không cần tạo ứng cử viên: Cây mẫu phổ biến

tiếp cận. Khai thác dữ liệu và khám phá tri thức, 8, 53–87.

Hà Lan, JH (1992). Thích ứng trong các hệ thống tự nhiên và nhân tạo. Cambridge, MA: Nhà xuất bản

Hồng, TP, & Wang, CY (2006). Duy trì các luật kết hợp bằng cách sử dụng các tập mục lớn. Cơ sở dữ liệu thông minh:

Công nghệ và Ứng dụng, trang 44–60.

Hồng, TP, Lin, CW, Yang, KT, & Wang, SL (2013). Sử dụng tf-idf để ẩn các tập mục nhạy cảm. Trí tuệ ứng dụng, 38,

502–510.

Lin, CW, Hong, TP, & Lu, WH (2009). Thuật toán tiền fufp để khai thác tăng dần. Hệ thống chuyên gia với các ứng dụng,

36, 9498–9505.

Lin, CW, Hong, TP, & Lu, WH (2011). Một cấu trúc cây hiệu quả để khai thác các tập mục hữu ích cao. Hệ thống chuyên gia với

Đơn đăng ký, 38, 7419–7424. Lin, CW, Hong, TP, Chang, CC, & Wang, SL (2013). Một cách tiếp cận dựa trên tham lam để ẩn các tập mục nhạy cảm bằng cách

chèn giao dịch. Tạp chí Ấn thông tin và Xử lý tín hiệu đa phương tiện, 4, 201–214. Lin, CW, Hong, TP, Wong, JW, Lan, GC & Lin, WY (2014). Cách tiếp cận dựa trên Ga để ẩn các tập mục hữu ích cao nhạy cảm.

Tạp chí Khoa học Thế giới, 12 tr. ID bài viết 804629. Lin, JCW, Gan, W., Fournier-Viger, P., Hong, TP, & Tseng, VS (2015). Các thuật toán hiệu quả để khai thác các tập mục có tiện ích cao

trong cơ sở dữ liệu không chắc chắn. Hệ thống dựa trên tri thức, 96, 171–187. Lin, JCW, Wu, TY, Fournier-Viger, P., Lin, G., Hong, TP, & Pan, JS (2015). Một cách tiếp cận vệ sinh để bảo vệ quyền riêng tư

khai thác tiện ích. Những tiến bộ trong hệ thống thông minh và máy tính, 388, 47–57. Lin, JCW, Fournier-Viger, P., & Gan, W. (2016). Fhn: Một thuật toán hiệu quả để khai thác các tập mục hữu ích cao với giá trị âm

lợi nhuận đơn vị. Hệ thống dựa trên tri thức, 111, 283–298. Lin, JCW, Gan, W., Fournier-Viger, P., Hong, TP, & Zhan, J. (2016). Khai thác hiệu quả các tập mục có tính tiện ích cao bằng cách sử dụng nhiều

ngưỡng tiện ích tối thiểu. Hệ thống dựa trên tri thức, 113, 100–115. Lindell, Y., & Pinkas, B. (2000). Khai thác dữ liệu bảo vệ quyền riêng tư. Hội nghị mật mã quốc tế thường niên về

Những tiến bộ trong Mật mã học, Santa Barbara, CA, Hoa Kỳ, trang 36–

Liu, J., Wang, K., & Fung, B. (2016). Khai thác các mẫu tiện ích cao trong một giai đoạn mà không tạo ra ứng viên. IEEE

Giao dịch về Kỹ thuật Tri thức và Dữ liệu, 28, 1245–1257. Liu, M., & Qu, J. (2012). Khai thác các tập mục hữu ích cao mà không cần tạo ứng cử viên. Hội nghị quốc tế ACM về

Quản lý thông tin và tri thức, trang 55–64.

Liu, Y., Liao, WK, & Choudhary, A. (2005). Thuật toán hai pha để phát hiện nhanh các tập mục có tiện ích cao. Ghi chú bài giảng

trong Khoa học Máy tính, 3518, 689–695.

Oliveria, SRM, & Zaiane, OR (2002). Quyền riêng tư bảo vệ việc khai thác tập mục thường xuyên. Hội nghị quốc tế về quyền riêng tư của

Khai thác dữ liệu và bảo mật, 14, 43–54. Rajalaxmi, RR, & Nataraja, AM (2012). Các phương pháp dọn dẹp hiệu quả để ẩn tiện ích nhạy cảm và các tập mục thường xuyên.

Phân tích dữ liệu thông minh, 16, 933–951. Verykios, VS, Bertino, E., Fovino, IN, Provenza, LP, Saygin, Y., & Theodoridis, Y. (2004). Sự riêng tư hiện đại

bảo tồn khai phá dữ liệu Bản ghi ACM SIGMOD, 33, 50–57.

Wu, YH, Chiang, CM, & Chen, ALP (2007). Ẩn các quy tắc kết hợp nhạy cảm với tác dụng phụ hạn chế. Giao dịch IEEE

về Kiến thức và Kỹ thuật Dữ liệu, 19, 29–42.

Yao, H., & Hamilton, HJ (2006). Khai thác tiện ích tập mục từ cơ sở dữ liệu giao dịch. Kỹ thuật Dữ liệu & Tri thức, 59,

603–626. Yao, H., Hamilton, HJ, & Butz, CJ (2004). Một cách tiếp cận cơ bản để khai thác các tiện ích tập mục từ cơ sở dữ liệu. SIAM

Hội nghị quốc tế về khai thác dữ liệu, Orlando, FL, trang 482–486.

Yeh, JS, & Hsu, PC (2010). Hhuif và msicf: các thuật toán mới để khai thác tiện ích bảo đảm quyền riêng tư. Hệ thống chuyên gia với Đơn đăng ký, 37, 4779–4786.

Yun, U., & Kim, J. (2015). Một thuật toán nhiễu loạn nhanh sử dụng cấu trúc cây để khai thác tiện ích bảo đảm quyền riêng tư. Chuyên gia Hệ thống có ứng dụng, 42, 1149–1165.