Máy tính & Bảo mật 132 (2023) 103360

Danh sách nội dung có sẵn tại ScienceDirect

Máy tính & Bảo mật

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

# Các thuật toán bảo vệ quyền riêng tư hiệu quả để ẩn các tập mục có tính tiện ích cao nhạy cảm

Mohamed Ashraf*, Sherine Rady, Tamer Abdelkader, Tarek F. Gharib

Khoa Hệ thống Thông tin, Khoa Khoa học Thông tin và Máy tính, Đại học Ain Shams, Cairo 11566, Ai Cập

bài báo thông tin

Lịch sử bài viết: Nhận được ngày 12 tháng 3 năm 2023 Sửa đổi ngày 14 tháng 5 năm 2023 Được chấp nhận ngày 20 tháng 6 năm 2023 Có sẵn trực tuyến ngày 21 tháng 6 năm 2023

Từ khóa: Bảo vệ quyền riêng tư Khai thác dữ liệu Khai thác tiện ích Mẫu nhạy cảm Số liệu bảo mật Số liệu tiện ích dữ liệu,

trừu tượng

Khai thác theo hướng tiện ích là một kỹ thuật khai thác dữ liệu mạnh mẽ nhằm mục đích trích xuất kiến thức quan trọng và có giá trị từ các loại bộ dữ liệu khác nhau. Tuy nhiên, việc phân tích các tập dữ liệu có thông tin nhạy cảm hoặc riêng tư có thể gây ra mối lo ngại về bảo mật và quyền riêng tư. Để cân bằng việc tối đa hóa tiện ích và bảo toàn quyền riêng tư, Khai thác tiện ích bảo vệ quyền riêng tư (PPUM) đã được trình bày. Mục tiêu chính của thuật toán PPUM là che giấu kiến thức nhạy cảm có thể được tìm thấy thông qua việc áp dụng thuật toán khai thác tiện ích cho dữ liệu nhạy cảm. Tuy nhiên, tài liệu PPUM hiện tại cho thấy rất ít thuật toán bảo vệ quyền riêng tư có thể mở rộng và đủ hiệu quả để xử lý các tập dữ liệu lớn và dày đặc. Dựa trên quan điểm này, bài viết này đề xuất ba thuật toán dựa trên heuristic, đó là Chọn tiện ích nhạy cảm với mục thực nhất trước tiên (SMRF), Chọn tiện ích nhạy cảm với mục ít thực nhất trước tiên (SLRF) và Chọn mục mong muốn nhất trước tiên. (SDIF), để che giấu một cách hiệu quả tất cả các Tập mục hữu ích có tính nhạy cảm cao (SHI) đồng thời giảm thiểu tác động bất lợi dự kiến đối với thông tin không nhạy cảm. Các thuật toán được đề xuất dựa trên một khái niệm mới, được gọi là Tiện ích nhạy cảm với vật phẩm thực (RISU), để chọn một cách hiệu quả một vật phẩm nạn nhân xác định cho mỗi SHI trong toàn bộ quá trình khử trùng. Hơn nữa, một kỹ thuật sắp xếp tập dữ liệu nhạy cảm mới được đề xuất để giảm thời gian cần thiết để tìm các giao dịch phù hợp để dọn dẹp. Thông qua các đánh giá thử nghiệm toàn diện với công nghệ tiên tiến, tính khả thi của ba thuật toán đề xuất đã được chứng thực. Những phát hiện thu được chứng minh rõ ràng tính hiệu quả của các thuật toán được đề xuất trong việc giảm thời gian khử trùng và các tác dụng phụ.



## 1. Giới thiệu

Với sự phát triển của công nghệ số trong lĩnh vực bán lẻ, thông tin về hành vi của khách hàng ngày càng trở nên có giá trị. Khả năng suy luận kiến thức từ dữ liệu của khách hàng có thể cho phép nhiều thực thể đưa ra các quyết định quan trọng và duy trì hoạt động kinh doanh hiệu quả. Theo hướng này, các công nghệ khai thác mẫu (Fournier-Viger và cộng sự, 2022) đã nổi lên như một phương pháp phân tích đáng tin cậy và mạnh mẽ trong kinh doanh và công nghiệp. Trong hơn một thập kỷ, khai thác theo định hướng tiện ích (Zhang và cộng sự, 2020) đã được nghiên cứu rộng rãi như một công nghệ khai thác mẫu để đưa ra các biến thể và dạng mẫu hữu ích mới. Thuật ngữ Khai thác tập mục tiện ích cao (HUIM) (Gan và cộng sự, 2019) chủ yếu đề cập đến nhiệm vụ khai thác dữ liệu xác định các tập mục có giá trị cao trong bộ dữ liệu giao dịch. So với Khai thác tập mục thường xuyên (FIM) (Agrawal và cộng sự, 1994), HUIM thường được coi là khả thi hơn đối với các ứng dụng trong thế giới thực vì nó xem xét các thuộc tính quan trọng như số lượng và lợi nhuận của các mục để khám phá các mô hình thú vị và có thể hành động hơn. . TRONG

Những năm gần đây, nhiều thuật toán khai thác tiện ích đã được giới thiệu để tính đến các tính năng dữ liệu khác nhau bên cạnh các yếu tố cụ thể và tình huống thực tế. Ví dụ: khai thác các mô hình tiện ích cao (HUP) (Liu và cộng sự, 2005; Qu và cộng sự, 2020; Tseng và cộng sự, 2012; Zida và cộng sự, 2015), HUP top-k (Ashraf và cộng sự, 2021), HUP đã đóng cửa (Lin và cộng sự, 2022), HUP sẵn có (Chen và cộng sự, 2022), HUP đa dạng (Verma và cộng sự, 2021) và HUP đa cấp (Tung và cộng sự, 2021). Khung khai thác tiện ích cũng đã được sử dụng trong các ứng dụng khác nhau như phân khúc khách hàng (Krishna và Ravi, 2021), phân tích nhật ký sự kiện (Fournier-Viger và cộng sự, 2020), phân tích dữ liệu sinh học (Segura-Delgado et al. ., 2022) và phân tích dữ liệu hoạt động taxi (Liu và Guo, 2021).

Có một số lợi ích hữu hình mà hoạt động khai thác theo định hướng tiện ích có thể mang lại. Tuy nhiên, những lo ngại về quyền riêng tư là một yếu tố thích hợp khác trong tài liệu khai thác tiện ích, điều này có khả năng làm giảm sự sẵn lòng của các công ty trong việc chia sẻ dữ liệu của họ với các đối tác vì mục đích nghiên cứu. Bất chấp tính thực tế và hữu ích của chúng, các kỹ thuật hướng đến tiện ích có thể dẫn đến các mối đe dọa lớn về bảo mật và quyền riêng tư nếu kiến thức thu được của chúng được cung cấp cho các bên không tin cậy (Tran và Hu, 2019). Ví dụ, trong các doanh nghiệp nghiên cứu thị trường, các công ty thường thu thập thông tin nhạy cảm về khách hàng.

* Tác giả tương ứng. Địa chỉ email: Mohamed.aahis@gmail.com (M. Ashraf).

https://doi.org/10.1016/j.cose.2023.103360 0167-4048/

khách hàng, bao gồm lịch sử mua hàng, lượt thích cá nhân và thông tin chi tiết về nhân khẩu học. Khai thác mẫu tiện ích có thể được áp dụng cho dữ liệu nhạy cảm như vậy để khám phá các mẫu và mối quan hệ ẩn, sau đó có thể được sử dụng để xây dựng hồ sơ khách hàng hoặc phát triển hệ thống đề xuất sản phẩm hiệu quả. Do đó, các công ty phải đánh giá mức độ nhạy cảm của dữ liệu họ đang chia sẻ và đảm bảo rằng các mẫu và mối quan hệ có tính tiện ích cao nhạy cảm bị ngăn chặn và không thể bị phát hiện ở ngưỡng quyền riêng tư cụ thể. Tuy nhiên, việc ẩn các mẫu nhạy cảm không phải là một nhiệm vụ tầm thường vì cấu trúc cơ sở dữ liệu ban đầu sẽ bị phá hủy nếu các mẫu nhạy cảm bị xóa một cách vô thức và không chính xác, điều này sẽ khiến cho tính tiện ích và độ tin cậy của dữ liệu giảm đáng kể.

Như một liều thuốc chữa bách bệnh cho những vấn đề nói trên, Khai thác dữ liệu bảo vệ quyền riêng tư (PPDM) (Kenthapadi và cộng sự, 2019; Mendes và Vilela, 2017) đã xuất hiện. Các thuật toán PPDM loại bỏ kiến thức nhạy cảm và bí mật khỏi các nguồn dữ liệu, một mặt, cải thiện khả năng chia sẻ và xuất bản dữ liệu, mặt khác, nâng cao việc sử dụng dữ liệu trong miền mục tiêu. Điều này cho phép các công ty và tổ chức kiểm soát thông tin có thể được lấy từ dữ liệu của họ và ngăn chặn kẻ thù khai thác thông tin thu được từ quá trình khai thác để cải thiện lợi nhuận hoặc đạt được lợi thế kinh doanh và tài chính. Khai thác tiện ích bảo vệ quyền riêng tư (PPUM) (Dinh và cộng sự, 2019; Gan và cộng sự, 2018) là một nhánh của PPDM với mục tiêu cuối cùng là che giấu tất cả các tập mục tiện ích cao nhạy cảm (SHI) đồng thời giảm tác động bất lợi của quá trình dọn dẹp trên các tập mục có tiện ích cao không nhạy cảm (NHI).

Cho đến nay, một số thuật toán (Ashraf và cộng sự, 2022; Jangra và Toshniwal, 2022; Lin và cộng sự, 2016; Liu và cộng sự, 2020b; Yeh và Hsu, 2010) đã được đề xuất để giải quyết vấn đề ẩn các mẫu tiện ích cao nhạy cảm trong bộ dữ liệu giao dịch. Tuy nhiên, người ta phân tích rằng hầu hết tất cả các thuật toán hiện có chỉ xem xét các tập dữ liệu có kích thước vừa và nhỏ, thời gian xử lý cũng như tác động tiêu cực của chúng đối với các mẫu không nhạy cảm có xu hướng cao đáng kể. Trong bài viết này, chúng tôi nghiên cứu và đóng góp chính xác cho tài liệu về PPUM bằng cách trình bày một khái niệm mới gọi là Tiện ích nhạy cảm đối tượng thực (RISU) để vệ sinh hiệu quả các tập dữ liệu giao dịch từ các mẫu tiện ích cao nhạy cảm. Đặc biệt, ba thuật toán dựa trên heuristic mới được đề xuất để khắc phục các vấn đề tồn tại của các thuật toán PPUM trước đây. Ngoài ra, một kỹ thuật sắp xếp mới được đề xuất để tăng tốc quá trình lọc và giảm sự mất mát về kiến thức không nhạy cảm. Các ý tưởng đề xuất đã được xác thực trên bốn bộ dữ liệu chuẩn khác nhau. Kết quả hoạt động được cho là khả hứa hẹn về việc giảm thời gian vệ sinh và tác dụng phụ so với các đối thủ hiện có.

Nội dung còn lại của bài viết này được sắp xếp như sau. Phần 2 xem xét các công trình liên quan về PPUM. Phần 3 mô tả những khái niệm sơ bộ và chính của khung HUIM và PPUM. Các thuật toán nhiều loạn được đề xuất được giới thiệu trong Phần 4. Phần 5 cung cấp một ví dụ minh họa cho các thuật toán được đề xuất. Phần 6 báo cáo kết quả đánh giá thực nghiệm. Cuối cùng, Phần 7 tóm tắt bài viết hiện tại.

## 2. Công trình liên quan

Trong thập kỷ qua, vấn đề bảo mật trong khai thác tiện ích đã được chú ý một cách khiêm tốn. Nói chung, các thuật toán PPUM hiện tại có thể được chia thành hai nhóm chính. Nhóm đầu tiên là các thuật toán vệ sinh dựa trên vật phẩm (Ashraf và cộng sự, 2022; Jangra và Toshniwal, 2022; Jisna và Salim, 2018; Lin và cộng sự, 2016; Liu và cộng sự, 2020a; 2020b; Selvaraj và Kuthadi, 2013; Yeh và Hsu, 2010; Yun và Kim, 2015), trong đó tiện ích của các mục thích hợp được sửa đổi trong từng tập mục có mức độ tiện ích cao nhạy cảm (SHI) cho đến khi chúng bị ẩn hoàn toàn. Và thứ hai là các thuật toán vệ sinh dựa trên giao dịch (Lin và cộng sự, 2014; 2017), trong đó

Các giao dịch được chèn hoặc xóa khỏi cơ sở dữ liệu gốc bằng cách áp dụng một số siêu chẩn đoán như thuật toán di truyền (Holland, 1992).

Bài toán PPUM được giới thiệu lần đầu tiên bởi Yeh và Hsu (2010), người đã trình bày hai thuật toán có tên HHUIF và MSCIF, đồng thời định nghĩa nhiều khái niệm cơ bản về PPUM. Trong cả hai thuật toán, các mẫu nhạy cảm được che giấu bằng cách giảm số lượng vật phẩm nạn nhân được chọn cho đến khi tiện ích của mẫu giảm xuống dưới ngưỡng tiện ích tối thiểu. Ngược lại, tiêu chí lựa chọn vật phẩm nạn nhân là khác nhau ở cả hai vì thuật toán HHUIF ưu tiên vật phẩm có tiện ích lớn nhất cho việc khử trùng, trong khi thuật toán MSCIF ưu tiên vật phẩm có tần suất xuất hiện tối đa trong các HUI nhạy cảm. Về kết quả, HHUIF được cho là yếu trong việc giảm tỷ lệ chênh lệch cơ sở dữ liệu trước và sau quá trình dọn dẹp, trong khi MSCIF được cho là yếu trong việc giảm số lượng HUI không nhạy cảm bị thiếu. Để nâng cao hiệu quả che giấu, Selvaraj và Kuthadi (2013) đã giới thiệu một phiên bản nâng cao của thuật toán HHUIF, được gọi là HHUIF*, áp dụng chiến lược chọn mục, được đặt tên là MHIS, để xử lý trường hợp khi có nhiều can- các vật phẩm nạn nhân có giá trị tiện ích tương tự. HHUIF* được cho là vượt trội so với HHUIF. Tuy nhiên, một nhược điểm lớn trong các thuật toán trước đó là chúng yêu cầu quét nhiều cơ sở dữ liệu để đạt được hoạt động nhiều loạn; như mong đợi, dẫn đến các vấn đề về hiệu suất và khả năng mở rộng. Được thúc đẩy bởi điều này, Yun và Kim (2015) đã đề xuất FPUTT mở rộng thuật toán UP-Growth+ (Tseng et al., 2012) để xác định các tập mục nhạy cảm và thu thập thông tin cần thiết cho quá trình ẩn. Mặc dù tốc độ tốt vì chỉ yêu cầu ba lần quét cơ sở dữ liệu, FPUTT tuân theo cùng một kỹ thuật ẩn của HHUIF và do đó nó chịu các tác dụng phụ tương tự.

Lin và cộng sự. (2016) đã đưa ra khái niệm về tiện ích tối đa và tối thiểu của các mục nhạy cảm bằng cách đề xuất hai thuật toán mới MSU-MAU và MSU-MIU để che giấu SHI với tốc độ cao hơn và giảm phản ứng của quá trình ẩn giấu. Cùng với đó, nhóm tác giả đề xuất 3 thước đo hiệu suất mới để đánh giá tốt hơn hiệu quả ẩn của thuật toán PPUM. Các thuật toán đề xuất đã được so sánh với HHUIF và MSCIF và cho thấy hiệu suất hàng đầu trong việc giảm thời gian thực hiện cũng như các tác động bất lợi. Để cải thiện hơn nữa tốc độ nhiễu loạn, Jisna và Salim (2018) đã đề xuất FPUFC áp dụng Bảng mục nhạy cảm (SIT) với cấu trúc nhiều tập hợp để lưu thông tin cần thiết cho quy trình thanh lọc. Kết quả đánh giá cho thấy FPUFC có hiệu năng thời gian chạy tốt hơn FPUTT.

Cuối cùng, tất cả các thuật toán được đề cập trước đó đều có một hạn chế lớn ở chỗ chúng không xem xét đến kiến thức không hạn chế. Việc xóa các tập mục nhạy cảm trong tập dữ liệu có thể dẫn đến mất các tập mục không nhạy cảm do các mục được chia sẻ giữa chúng, dẫn đến làm giảm chất lượng của tập dữ liệu bị nhiễu. Vì vậy, nhiều học giả bắt đầu chú ý nhiều hơn đến các tập mục không nhạy cảm. Lưu và cộng sự. (2020a) đã trình bày IM-SCIF là phiên bản cải tiến của MSCIF (Yeh và Hsu, 2010) với mức độ bảo mật cao hơn. IMSCIF lặp đi lặp lại việc chọn các mục nạn nhân bằng cách đánh giá số lượng xung đột của chúng giữa các bộ mục nhạy cảm. Đối với giao dịch nạn nhân, nó chọn giao dịch chứa số lượng tối thiểu các tập mục không bị hạn chế và giá trị tiện ích tối đa cho tập mục nhạy cảm đang được vệ sinh. IMSCIF cho thấy hiệu suất thời gian chạy kém nhưng kết quả đầy hứa hẹn về việc giảm chi phí còn thiếu so với các thuật toán HHUIF, MSCIF, MSU-MAU và MSU-MIU.

Li và cộng sự. (2019) đã nghiên cứu việc sử dụng lập trình tuyến tính số nguyên (ILP) để tìm ra giải pháp chính xác cho vấn đề khai thác tiện ích bảo đảm quyền riêng tư. Họ đã xác định lại phương pháp ẩn như một vấn đề thỏa mãn ràng buộc cùng với cơ chế nới lỏng để đảm bảo giảm thiểu các tác động tiêu cực do hoạt động che giấu tạo ra đến giới hạn tối đa. Tuy nhiên, một chính

thách thức đối với kỹ thuật được áp dụng của họ là khả năng mở rộng; bởi vì việc tìm ra giải pháp chính xác trong các bài toán NP-khó là một nhiệm vụ tính toán rất tốn kém. Nói cách khác, việc tìm ra lời giải chính xác cho một bài toán NP-khó thường đòi hỏi một lượng lớn tài nguyên tính toán, điều này có thể khiến việc giải những bài toán đó đối với các tập dữ liệu lớn trong một khoảng thời gian hợp lý là không thể thực hiện được. Do đó, hầu hết các thuật toán của khung PPUM đều sử dụng phương pháp phỏng đoán hoặc xấp xỉ để tìm ra giải pháp tốt trong một khoảng thời gian hợp lý, mặc dù chúng không thể đảm bảo tìm ra giải pháp chính xác.

Lưu và cộng sự. (2020b) đã trình bày ba thuật toán bóp méo dữ liệu là SMAU, SMIU và SMSE, thực hiện thao tác bóp méo bằng cách quét cơ sở dữ liệu hai lần. Ba thuật toán dựa trên cấu trúc dữ liệu phức tạp để lưu thông tin cần thiết về các tập mục nhạy cảm và không nhạy cảm. Giao dịch sở hữu số lượng mẫu không nhạy cảm ít nhất được chọn để thanh lọc trong ba thuật toán. Sự khác biệt cốt lõi giữa ba thuật toán nằm ở cách chúng chọn mục nạn nhân. Thuật toán SMAU ẩn SHI bằng cách giảm tiện ích của vật phẩm có giá trị tiện ích lớn nhất. Thuật toán SMIU ẩn SHI bằng cách giảm tiện ích của vật phẩm có giá trị hữu ích ít nhất. Thuật toán SMSE xác định mục nạn nhân sao cho nó tồn tại trong số lượng lớn nhất các tập mục nhạy cảm. Ba thuật toán cho thấy kết quả cạnh tranh về thời gian thực hiện và tác dụng phụ, mặc dù mức tiêu thụ bộ nhớ cao.

Gần đây, Jangra và Toshniwal (2022) đã đề xuất hai thuật toán mới là MinMax và Weighted để đạt được nhiều loạn nhanh đồng thời hạn chế tỷ lệ biến dạng dữ liệu. Mỗi thuật toán áp dụng một chiến lược lựa chọn mục nạn nhân khác nhau và dựa vào ba kỹ thuật sắp xếp tập dữ liệu, cụ thể là DoC_RoT, RoT_DoC và thời lượng giao dịch để xác định giao dịch của nạn nhân. Không giống như các nghiên cứu dọn dẹp dựa trên mục trước đây, một mục cụ thể được chọn trong mỗi tập mục nhạy cảm để trở thành mục nạn nhân của tập mục này trong toàn bộ quá trình dọn dẹp. Lý do đằng sau điều này là mục nạn nhân được chọn rất có thể sẽ gây ra ít tác động tiêu cực nhất đến các tập mục không nhạy cảm. Về mặt phát hiện, kỹ thuật sắp xếp kép, RoT_DoC, được cho là hiệu quả nhất đối với hai thuật toán. Tuy nhiên, chúng tôi cho rằng kỹ thuật sắp xếp kép mang lại chi phí tính toán tương đối cao, đặc biệt là trong các tập dữ liệu lớn và rất thưa thớt. Hơn nữa, trong cả hai thuật toán, các kỹ thuật lựa chọn mục nạn nhân được sử dụng chỉ xem xét các tập mục nhạy cảm và không nhạy cảm mà bỏ qua ảnh hưởng của các mục nạn nhân được chọn đối với toàn bộ giao dịch nhạy cảm, do đó có thể dẫn đến kết quả không thuận lợi.

Gần đây hơn, Ashraf et al. (2022) đã phát triển thuật toán SB2VF với nhận thức sâu sắc rằng có thể đạt được quá trình khử trùng hiệu quả cao bằng cách chọn các mục và giao dịch nạn nhân dựa trên các tiêu chí hiệu quả. Họ gợi ý rằng nên chọn hai mục làm mục nạn nhân ứng cử viên trong mỗi tập mục nhạy cảm, đó là mục trùng lặp trong hầu hết các tập mục nhạy cảm và mục trùng lặp trong hầu hết các tập mục không nhạy cảm. Hai mục này được sửa đổi có thể thay thế cho nhau trong khi che giấu các tập mục nhạy cảm của chúng sao cho mục có giá trị tiện ích thấp hơn trong giao dịch được xử lý sẽ đáng được sửa đổi hơn. Họ cũng đề xuất sắp xếp ba lần các giao dịch nhạy cảm để đảm bảo rằng các giao dịch có tác dụng phụ thấp hơn sẽ được xử lý trước. Bằng cách sử dụng ba bộ dữ liệu chuẩn, tính hiệu quả của ý tưởng của họ đã được chứng minh. Các tác giả tương tự cũng trình bày thuật toán HUP-Hiding (Ali và cộng sự, 2023), chọn một hoặc nhiều mục nạn nhân cho mỗi tập mục nhạy cảm dựa trên một tiêu chí cụ thể và ra lệnh cho các giao dịch nhạy cảm dựa trên tiện ích của các tập mục nhạy cảm. .

Tất cả các thuật toán được thảo luận ở trên đều áp dụng mô hình làm sạch dựa trên vật phẩm để ẩn SHI, đây là mô hình mà bài viết này quan tâm nhất. Chúng tôi tóm tắt những ưu điểm và nhược điểm cốt lõi của các thuật toán dọn dẹp dựa trên vật phẩm chính trong Bảng 1. Các nghiên cứu khác tuân theo mô hình dọn dẹp dựa trên giao dịch cho nhiệm vụ của PPUM. Lin và cộng sự. (2014) đã sử dụng thuật toán di truyền và khái niệm tiền lớn để làm sạch tập dữ liệu bằng cách chèn các giao dịch mới đồng thời giảm thời gian cần thiết để quét cơ sở dữ liệu. Các tác giả (Lin và cộng sự 2017) đã đề xuất một phương pháp tiếp cận dựa trên thuật toán di truyền, được gọi là PPUMGAT. Phương pháp đề xuất của họ sẽ tự động xóa các giao dịch nhạy cảm để che giấu SHI và sử dụng hàm thích ứng để đánh giá độ mịn của nhiễm sắc thể (giải pháp ứng cử viên).

Trong khi đó, một số thuật toán đã được đưa ra để ẩn cả tiện ích nhạy cảm và tập mục thường xuyên cùng một lúc. Trong (Rajalaxmi và Natarajan, 2012) hai thuật toán chung đã được giới thiệu, đó là MSMU và MCRSU. Trong thuật toán MSMU, quá trình ẩn được thực hiện bằng cách chọn giao dịch chứa số lượng tập mục không bị hạn chế ít nhất làm giao dịch nạn nhân và mục có hỗ trợ tối thiểu hoặc tiện ích tối đa làm mục nạn nhân. Thuật toán MCRSU áp dụng cách tiếp cận tương tự nhưng sử dụng tỷ lệ các tập mục không hạn chế bị ảnh hưởng bởi quá trình nhiễu loạn đối với việc lựa chọn mục nạn nhân. Người ta đã chứng minh rằng thuật toán MCRSU hiệu quả hơn MSMU trong việc giảm chi phí còn thiếu và duy trì tính tốt của tập dữ liệu đã được tinh lọc. Sau đó, Liu và cộng sự. (2018) đã phát triển thuật toán HUFI để bảo vệ tiện ích nhạy cảm và các mẫu thường xuyên. HUFI được phát triển áp dụng ý tưởng về ranh giới tối đa để kiểm soát cách vệ sinh.

Một số công trình tập trung vào việc áp dụng các kỹ thuật mới để tiến hành quá trình khử trùng. Bandil và cộng sự. (2018) đã áp dụng nguyên tắc bảo mật khác biệt bằng các phép biến đổi Laplace để che giấu các tập mục tiện ích nhạy cảm. Triệu và cộng sự. (2018) đã giới thiệu thuật toán HHUARL để che giấu các quy tắc kết hợp tiện ích cao nhạy cảm bằng cách sử dụng mạng lưới giao nhau của các tập mục (Grätzer, 2011) và khái niệm tiện ích theo trọng số giao dịch (Liu và cộng sự, 2005).

Bên cạnh các thuật toán đã đề cập trước đó, một số nghiên cứu đã điều tra vấn đề PPUM trong cơ sở dữ liệu tuần tự. Ví dụ, Lệ và cộng sự. (2018) đã đề xuất HUS-Hiding để giải quyết vấn đề che giấu các mẫu tiện ích nhạy cảm trong bộ dữ liệu trình tự định lượng. Trong (Huỳnh và cộng sự 2022), ba thuật toán song song đa lõi đã được giới thiệu để tăng tính tối ưu của quá trình ẩn.

Dựa trên khảo sát trước đây và so sánh chi tiết trong Bảng 1, rõ ràng là các nghiên cứu trước đây về khuôn khổ PPUM tiêu chuẩn đã ít chú ý đến việc duy trì chất lượng của tập dữ liệu giao dịch tổng hợp; vì nhiều tập mục không nhạy cảm bị mất trong quá trình nhiễu loạn và sự khác biệt giữa tập dữ liệu gốc và tập dữ liệu đã được tinh lọc có xu hướng rất lớn, tạo cơ hội để cải thiện khi giảm thời gian cần thiết để dọn dẹp, tối đa hóa tiện ích tập dữ liệu và giảm thiểu những tác động tiêu cực đi kèm. Để cải thiện các thuật toán trước đó, bài nghiên cứu này đề xuất ba thuật toán nỗ lực tốt nhất không chỉ nhanh mà còn hiệu quả trong việc giảm các hậu quả tiêu cực dự kiến đối với tập dữ liệu đã được làm sạch cuối cùng. Tóm lại, những đóng góp chính của công việc này được nêu bật như sau:

1. Ba thuật toán dọn dẹp dựa trên vật phẩm với ba kỹ thuật chọn mục nạn nhân khác nhau được thiết kế để duy trì hiệu suất dọn dẹp đồng thời tăng mức độ riêng tư.

2. Đề xuất về khái niệm Tiện ích nhạy cảm cho vật phẩm thực (RISU), có thể được tận dụng để chọn mục nạn nhân lý tưởng cho từng tập mục nhạy cảm trong giai đoạn đầu của quá trình dọn dẹp, dẫn đến việc dọn dẹp nhanh hơn.

3. Một kỹ thuật sắp xếp tập dữ liệu nhạy cảm mới được gọi là Sắp xếp có trọng số được giới thiệu. Kỹ thuật này được sử dụng cho quy trình lựa chọn giao dịch nạn nhân cho tất cả các thuật toán được đề xuất nhằm ưu tiên dọn dẹp các giao dịch có thể gây ra ít tác dụng phụ hơn.

4. Để xác minh tính hiệu quả của thuật toán được đề xuất, các so sánh rộng rãi được tiến hành trên bốn điểm chuẩn khác nhau

Bảng 1 So sánh giữa các thuật toán dọn dẹp dựa trên vật phẩm PPUM chính.

| Thuật toán | Chiến lược vệ sinh | Thuận lợi | Nhược điểm |
|---|---|---|---|
| HHUIF (Yeh và Hsu, 2010) | -Chọn vật phẩm có tiện ích cao nhất làm nạn nhân. | -Sự đơn giản của thiết kế và thực hiện. | -Yêu cầu quét cơ sở dữ liệu nhiều lần. -Mất thông tin không nhạy cảm đáng kể trong cơ sở dữ liệu đã được vệ sinh. |
| MSICF (Yeh và Hsu, 2010) | -Chọn mục có mức độ xung đột cao nhất trong số tất cả các tập mục nhạy cảm làm nạn nhân. -Chọn giao dịch có tiện ích cao nhất của mặt hàng nạn nhân. | -Sự đơn giản của thiết kế và thực hiện. | -Yêu cầu quét cơ sở dữ liệu nhiều lần. -Mất thông tin không nhạy cảm đáng kể trong cơ sở dữ liệu đã được vệ sinh. |
| FPUTT (Yun và Kim, 2015) | -Chọn vật phẩm có tiện ích cao nhất làm nạn nhân. | -Yêu cầu chỉ có ba lần quét cơ sở dữ liệu. | - Công tác vệ sinh còn chậm. -Mất thông tin không nhạy cảm đáng kể trong cơ sở dữ liệu đã được vệ sinh. |
| MSU-MAU (Lin và cộng sự, 2016) | -Chọn giao dịch có tiện ích cao nhất trong tập mục nhạy cảm. -Chọn vật phẩm có tiện ích cao nhất làm nạn nhân. | -Tránh quét cơ sở dữ liệu nhiều lần bằng cách lưu trữ thông tin nhạy cảm giao dịch của từng tập mục nhạy cảm trong bảng chỉ mục. | -Tổn thất các thông tin không nhạy cảm còn cao. |
| MSU-MIU (Lin và cộng sự, 2016) | -Chọn giao dịch có tiện ích cao nhất trong tập mục nhạy cảm. -Chọn vật phẩm có tiện ích thấp nhất làm nạn nhân. | -Tránh quét cơ sở dữ liệu nhiều lần bằng cách lưu trữ thông tin nhạy cảm giao dịch của từng tập mục nhạy cảm trong bảng chỉ mục. | -Tổn thất các thông tin không nhạy cảm còn cao. |
| SMAU (Liu và cộng sự, 2020b) | -Chọn giao dịch có số lượng mục không nhạy cảm ít nhất. -Chọn vật phẩm có tiện ích cao nhất làm nạn nhân. | -Yêu cầu chỉ có hai lần quét cơ sở dữ liệu. -Có thể giảm thiểu sự mất mát những thông tin không nhạy cảm. | - Gặp vấn đề về khả năng mở rộng do cấu trúc dữ liệu đắt tiền. |
| SMIU (Liu và cộng sự, 2020b) | -Chọn giao dịch có số lượng mục không nhạy cảm ít nhất. -Chọn vật phẩm có tiện ích thấp nhất làm nạn nhân. | -Yêu cầu chỉ có hai lần quét cơ sở dữ liệu. -Có thể giảm thiểu sự mất mát những thông tin không nhạy cảm. | - Gặp vấn đề về khả năng mở rộng do cấu trúc dữ liệu đắt tiền. |
| SMSE (Liu và cộng sự, 2020b) | -Chọn giao dịch có số lượng mục không nhạy cảm ít nhất. -Chọn mục có mức độ xung đột cao nhất trong số các tập mục nhạy cảm và mức độ xung đột thấp nhất trong số các tập mục không nhạy cảm làm nạn nhân. | -Yêu cầu chỉ có hai lần quét cơ sở dữ liệu. -Có thể giảm thiểu sự mất mát những thông tin không nhạy cảm. | - Gặp vấn đề về khả năng mở rộng do cấu trúc dữ liệu đắt tiền. |
| Có trọng số (Jangra và Toshni-wal, 2022) | -Mục có trọng lượng tính toán thấp nhất là mục nạn nhân lý tưởng của tập mục nhạy cảm. -Giao dịch được lựa chọn dựa trên tiêu chí sắp xếp được xác định trước. | -Không cần phải đánh giá các mặt hàng và giao dịch của nạn nhân nhiều lần | -Có thể sắp xếp kép cơ sở dữ liệu tốn kém. -Bỏ qua ảnh hưởng của các mục nạn nhân được chọn đối với toàn bộ các giao dịch nhạy cảm. |
| MinMax (Jangra và Toshni-wal, 2022) | -Mục có mức độ xung đột thấp nhất trong số các tập mục không nhạy cảm là mục nạn nhân lý tưởng của tập mục nhạy cảm. -Giao dịch được lựa chọn dựa trên tiêu chí sắp xếp được xác định trước. | -Không cần phải đánh giá các mặt hàng và giao dịch của nạn nhân nhiều lần | -Có thể sắp xếp kép cơ sở dữ liệu tốn kém. -Bỏ qua ảnh hưởng của các mục nạn nhân được chọn đối với toàn bộ các giao dịch nhạy cảm. |

Bảng 2 Một ví dụ về tập dữ liệu giao dịch.

| Thời gian | Giao dịch (mặt hàng, số lượng mua) |
|---|---|
| T1 | (A, 1), (B, 3), (E, 2) |
| T2 | (C, 4), (D, 5) |
| T3 | (E, 2), (F, 3) |
| T4 | (A, 5), (B, 4), (C, 2), (D, 2), (E, 4) |
| T5 | (B, 5), (C, 4), (E, 6), (F, 1) |
| T6 | (B, 1), (C, 3), (E, 6) |
| T7 | (A, 2), (E, 3), (F, 4) |

Bảng 3 Đơn vị lợi nhuận của các mặt hàng.

| Mục | Lợi nhuận đơn vị |
|---|---|
| MỘT | 6 |
| B | 4 |
| C | 2 |
| D | 5 |
| E | 3 |
| F | 9 |

bộ dữ liệu sử dụng năm thước đo hiệu suất. Kết quả thu được chỉ ra rằng các ý tưởng được áp dụng hoạt động rất tốt trong các tập dữ liệu lớn và dày đặc và có thể giảm tổn thất trong các tập mục không nhạy cảm tới 40% so với các thuật toán hiện có.

## 3. Sơ bộ

Giả sử D là tập dữ liệu giao dịch định lượng có dạng một tập hợp các giao dịch sao cho D = {T1 , T2 , ..., Tn } trong đó Ti (1 ≤ i ≤ n ) là giao dịch có mã định danh i. Ví dụ, Bảng 2 trình bày một ví dụ về tập dữ liệu giao dịch D có bảy giao dịch với số nhận dạng của chúng. Những giao dịch này có thể thể hiện hành vi mua hàng của khách hàng hoặc các hoạt động cụ thể của người dùng. Giả sử I = {i1 , i2 , ..., im } là tập hữu hạn gồm các mục riêng biệt, mỗi giao dịch Ti bao gồm m mục định lượng, mỗi mục ij ∈ I có giá trị tiện ích bên ngoài, eu(ij ), phản ánh việc bán nó doanh thu hoặc trọng số như trong Bảng 3 và có giá trị hữu dụng nội bộ iu(ij ), phản ánh số lượng hoặc tần suất bán của nó trong mỗi giao dịch Ti . Một tập mục X có thể được coi là một tập gồm k mục riêng biệt và có thể được gọi là tập mục k. Một giao dịch Ti có thể được coi là một giao dịch hỗ trợ cho X khi và chỉ khi X ⊆ Ti .

Để hiểu rõ hơn những thông tin sơ bộ ở trên, hãy kiểm tra tập dữ liệu mẫu D được đưa ra trong Bảng 2. Tổng cộng có bảy giao dịch, được biểu thị là T1 , T2 , T3 , T4 , T5 , T6 và T7 . Ngoài ra, có sáu mục trong D, được biểu thị là A, B, C, D, E và F. Tập hợp lợi nhuận đơn vị dương liên quan đến việc bán các mặt hàng này được cung cấp trong Bảng 3. Về nguyên tắc, mỗi giao dịch chỉ ra việc bán các mặt hàng cụ thể. Ví dụ: giao dịch T1 trong D ngụ ý rằng mặt hàng A, B và E được mua trong giao dịch này với số lượng lần lượt là 1, 3 và 2.

Định nghĩa 1. Tiện ích của mục x ∈ Ti được tính như sau, u(x, Ti ) = iu(x, Ti ) × eu(x ). Ví dụ, trong Bảng 2, u(A, T1 ) = iu(A, T1 ) × eu(A ) = 1 × 6 = 6.

Bảng 4 HUI được tạo khi minUtil = 70.

| Bộ vật phẩm | Tính thiết thực |
|---|---|
| F | 72 |
| A, B | 72 |
| A, E | 87 |
| LÀ | 102 |
| E, F | 105 |
| A, B, E | 90 |
| B, C, E | 106 |
| A, B, C, D, E | 72 |

Bảng 5 Các HUI nhạy cảm và các giao dịch của chúng.

| Bộ vật phẩm | ID giao dịch |
|---|---|
| A, B, E | T1, T4 |
| B, C, E | T4, T5, T6 |
| A, B, C, D, E | T4 |

Định nghĩa 2. Với tập mục X ⊆ Ti . u(X ) trong Ti có thể được định nghĩa như, u(X, Ti ) = ∑xj ∈X ∧X ⊆Ti u(xj , Ti ). Ví dụ: trong Bảng 2, u({A, B}, T1 ) = u(A, T1 ) + u(B, T1 ) = 6 + 12 = 18.

Định nghĩa 3. Tổng hữu dụng của tập mục X trong tập dữ liệu D là được định nghĩa là: u(X ) = ∑X⊆Ti ∧Ti ∈D u(X, Ti ). Ví dụ: trong Bảng 2, u({A, B} ) = u({A, B}, T1 ) + u({A, B}, T4 ) = 18 + 46 = 64.

Định nghĩa 4. Đối với mọi giao dịch Ti ⊂ D, tiện ích giao dịch của Ti có thể được tính như sau, TU (Ti ) = ∑x∈Ti u(x, Ti ). Ví dụ, trong Bảng 2, TU (T3 ) = u(E, T3 ) + u(F , T3 ) = 6 + 27 = 33.

Định nghĩa 5. Tập mục hữu ích cao (HUI) là tập mục có giá trị tiện ích lớn hơn hoặc bằng ngưỡng tiện ích tối thiểu do người dùng ưa thích (minUtil). Nhiệm vụ khai thác HUI từ bộ dữ liệu giao dịch là tìm tất cả các HUI thỏa mãn minUtil nhất định. Các tập mục hữu ích cao được tạo ra từ tập dữ liệu mẫu D khi minUtil = 70 được cung cấp trong Bảng 4.

Định nghĩa 6. Tập mục hữu ích cao nhạy cảm (SHI) là một mẫu rất có giá trị cần được giữ bí mật và không thể bị phát hiện bởi bất kỳ kỹ thuật khai thác dựa trên tiện ích nào vì nó có thể tiết lộ kiến thức bí mật hoặc thông tin riêng tư về chủ sở hữu dữ liệu. Ngược lại, một tập mục có tiện ích cao không nhạy cảm (NHI) phải được giữ ở mức có thể phát hiện được để đảm bảo độ tin cậy và tính hợp lệ của dữ liệu. Điều đáng nói là những gì được coi là thông tin nhạy cảm có thể thay đổi theo bối cảnh và phạm vi dữ liệu cũng như các bên liên quan tham gia vào quá trình phân tích, chẳng hạn như trong kỹ thuật tiếp thị, nếu một mẫu hữu ích cao tiết lộ thói quen hoặc hành vi mua hàng của khách hàng thuộc một giới tính hoặc chủng tộc nhất định, nó có thể được coi là một mô hình nhạy cảm vì nó có thể gây ra những lo ngại về đạo đức. Xem xét ví dụ đang chạy, các tập mục nhạy cảm được nêu trong Bảng 5.

Định nghĩa 7. Xét một tập mục hữu ích cao nhạy cảm SHIi , một mục nhạy cảm là và một giao dịch T j sao cho ∈ SHIi và SHIi ⊆ T j . được cho là mục nạn nhân của SHIi , chính thức hơn là Ivic (SHIi ), nếu tiện ích của nó cần được giảm xuống để ẩn tập mục nhạy cảm SHIi . Giả thuyết đằng sau việc chọn mục cụ thể trong số tất cả các mục trong SHIi là các tập mục có tiện ích cao (NHI) không nhạy cảm không tránh khỏi việc bị che, do đó việc chọn mục này dựa trên tiêu chí hiệu quả có thể làm giảm đáng kể tổn thất của chúng. Theo cách tương tự, T j được cho là giao dịch nạn nhân của SHIi, cụ thể là Tvic (SHIi), nếu nó được chọn để sửa đổi mục nạn nhân. Xem xét ví dụ đang chạy, theo Bảng 5, các mục nhạy cảm là A, B, C, D, E và các giao dịch nhạy cảm là T1 , T4 , T5 , T6 .

Định nghĩa 8. Lớp phủ nhạy cảm của một hạng mục xi , chính thức hơn là SC(xi ), đề cập đến số lượng SHI chứa hạng mục xi . Đang xem xét ví dụ đang chạy, bìa nhạy cảm của tất cả các mục nhạy cảm được đưa ra trong Bảng 6.

Bảng 6 Vỏ bọc nhạy cảm và không nhạy cảm của các mặt hàng.

| Mặt hàng nhạy cảm | MỘT | B | C | D | E |
|---|---|---|---|---|---|
| SC | 2 | 3 | 2 | 1 | 3 |
| NSC | 2 | 2 | 0 | 0 | 3 |

Bảng 7 Vỏ bọc nhạy cảm và không nhạy cảm của các giao dịch.

| Thời gian giao dịch nhạy cảm | T1T4 | T5 | T6 | |
|---|---|---|---|---|
| SC | 1 | 3 | 1 | 1 |
| NSC | 3 | 3 | 3 | 1 |

Định nghĩa 9. Phạm vi nhạy cảm của giao dịch T j , chính thức hơn là SC(T j ), đề cập đến số lượng SHI xuất hiện trong giao dịch T j . Xem xét ví dụ đang chạy, vỏ bọc nhạy cảm của tất cả các giao dịch nhạy cảm được đưa ra trong Bảng 7.

Định nghĩa 10. Lớp phủ không nhạy cảm của một hạng mục xi , chính thức hơn là NSC(xi ), đề cập đến số lượng NHI chứa hạng mục xi . Xem xét ví dụ đang thực hiện, bìa không nhạy cảm của tất cả các mục nhạy cảm được đưa ra trong Bảng 6.

Định nghĩa 11. Vỏ bọc không nhạy cảm của một giao dịch T j , chính thức hơn là NSC(T j ), đề cập đến số lượng NHI xuất hiện trong giao dịch T j . Xem xét ví dụ đang chạy, phạm vi không nhạy cảm của tất cả các giao dịch nhạy cảm được đưa ra trong Bảng 7.

Định nghĩa 12. Trọng số nhạy cảm của giao dịch T j , W t(T j ), có thể được tính bằng công thức sau:

$$W\,t(T_j) = \frac{SC(T_j)}{NSC(T_j) + 1} \quad (1)$$

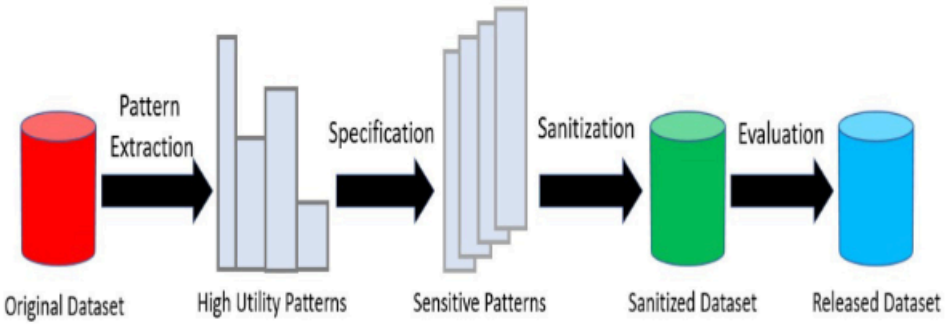
Bởi vì một giao dịch nhạy cảm có thể không có tập hợp mục không nhạy cảm nên chúng tôi đã thêm 1 vào mẫu số của công thức trên. Về bản chất, Vỏ bọc nhạy cảm (SC) cao của giao dịch biểu thị rằng việc xóa mục nạn nhân rất có thể sẽ ẩn các tập mục nhạy cảm hơn, trong khi Vỏ bọc không nhạy cảm (NSC) thấp biểu thị rằng việc sửa đổi giao dịch rất có thể sẽ tạo ra ít tác dụng phụ hơn. Do đó, các giao dịch có trọng số cao cần được làm sạch trước tiên.

Định nghĩa 13. Để giảm thiểu tình trạng mất thông tin do biến dạng giao dịch, lấy cảm hứng từ kỹ thuật sắp xếp kép trong (Jangra và Toshniwal, 2022), chúng tôi đề xuất kỹ thuật Sắp xếp có trọng số. Kỹ thuật này ngụ ý rằng các giao dịch nhạy cảm được sắp xếp theo thứ tự giảm dần trọng lượng của chúng. Nói cách khác, các giao dịch có trọng số cao hơn sẽ có mức độ ưu tiên cao hơn cho việc vệ sinh. Xem xét ví dụ đang chạy, thứ tự sắp xếp của các giao dịch nhạy cảm là T4 ≺ T6 ≺ T1 ≺ T5 .

Định nghĩa 14. Giả sử là một mục nhạy cảm, x, và tập hợp tất cả các giao dịch nhạy cảm trong tập dữ liệu mẫu, ST . Tiện ích nhạy cảm vật phẩm thực của x, RISU (x ), có thể được tính bằng công thức sau:

$$GẠO\,(x) = \sum_{x \in T_j \wedge T_j \in ST} u(x, T_j) \quad (2)$$

Tìm mục nạn nhân tối ưu trong tập mục nhạy cảm là một nhiệm vụ tế nhị và khá phức tạp. Điều này là do việc sửa đổi các mục trong các tập mục nhạy cảm có thể dẫn đến các tác dụng phụ ngoài ý muốn, đặc biệt nếu các mục được sửa đổi chồng chéo với nhiều giao dịch nhạy cảm và các tập mục không nhạy cảm. Đây là lúc khái niệm về Tiện ích nhạy cảm với tập mục thực sự (RISU) xuất hiện. RISU




Hình 1. Khung chung của quy trình khai thác tiện ích bảo vệ quyền riêng tư.

Bảng 8 Giá trị RISU của các hạng mục.

| Mặt hàng nhạy cảm | MỘT | B | C | D | E |
|---|---|---|---|---|---|
| TĂNG LÊN | 48 | 48 | 18 | 10 | 54 |

các giá trị có thể được coi là tham chiếu đến sự đóng góp của các hạng mục nhạy cảm vào tổng lợi ích của các giao dịch nhạy cảm của chúng. Do đó, các giá trị này có thể được sử dụng làm công cụ đánh giá để tìm mục nạn nhân thích hợp cho mỗi SHI, điều này có thể nâng cao hiệu suất tiềm năng của các thuật toán được đề xuất cho nhiệm vụ loại bỏ kiến thức nhạy cảm. Xem xét ví dụ đang chạy, các giá trị RISU của tất cả các hạng mục nhạy cảm được cung cấp trong Bảng 8.

Hình 1 thể hiện khuôn khổ chung cho quy trình PPUM. Nhìn chung, với tập dữ liệu giao dịch D, bảng có lợi nhuận và ngưỡng quyền riêng tư tối thiểu (minUtil). Nhiệm vụ của PPUM hay còn gọi là Ẩn mẫu tiện ích cao (HUPH) bao gồm bốn quy trình chính tiếp theo: (1) thuật toán điều khiển tiện ích được thực thi trên D để tạo ra tất cả các HUI ở một minUtil cụ thể (Trích xuất mẫu); (2) trong số các HUI được tạo ra, những HUI nhạy cảm phải được xác định theo một số yêu cầu kinh doanh, chính sách quyền riêng tư hoặc tùy chọn của chủ sở hữu dữ liệu (Thông số kỹ thuật); (3) một thuật toán dọn dẹp được thực hiện nhằm mục đích ẩn tất cả các tập mục nhạy cảm được xác định trước với ít tác dụng phụ nhất (Sanitization); (4) Các tác dụng phụ của quá trình vệ sinh được đánh giá liên quan đến các dạng nhạy cảm và không nhạy cảm được chỉ ra trong giai đoạn thứ hai (Đánh giá). Quá trình biến dạng của thuật toán PPUM áp dụng phương pháp dọn dẹp dựa trên vật phẩm liên quan đến việc chọn một số mục trong các mẫu nhạy cảm sẽ bị loại bỏ, cho dù bằng cách xóa chúng hay giảm tiện ích nội bộ của chúng trong một số giao dịch cho đến khi tiện ích hoặc độ tin cậy của mẫu giảm xuống dưới tối thiểu cho đến khi.

## 4. Các thuật toán đề xuất

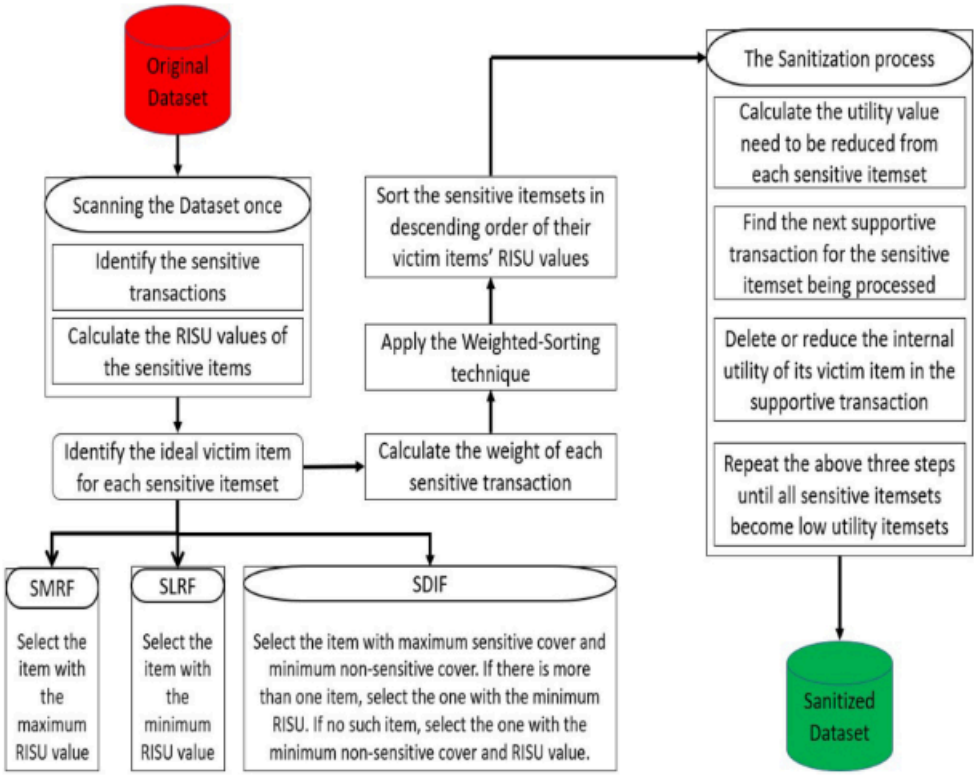
Khai thác tiện ích bảo vệ quyền riêng tư (PPUM) là một định nghĩa đưa ra lời hứa với chủ sở hữu dữ liệu rằng thông tin riêng tư và nhạy cảm của họ sẽ không bị tiết lộ, theo cách bất lợi hoặc theo cách khác, nếu họ cho phép chia sẻ hoặc xuất bản dữ liệu của mình để sử dụng trong bất kỳ hoạt động phân tích hướng đến tiện ích nào . Phù hợp với định nghĩa này, chúng tôi đề xuất ba thuật toán heuristic cho PPUM: (1) Chọn tiện ích nhạy cảm với mục thực nhất trước tiên (SMRF); (2) Chọn tiện ích nhạy cảm với mục Least Real đầu tiên (SLRF); và (3) Chọn mục mong muốn nhất trước tiên (SDIF).

Tóm lại, hiệu quả và hiệu quả của bất kỳ thuật toán PPUM nào đều phụ thuộc cơ bản vào ba yếu tố: (1) các mục nạn nhân được chọn; (2) các giao dịch nạn nhân được lựa chọn; (3) Thứ tự xử lý và ẩn của các tập mục nhạy cảm.

Để chọn giao dịch nạn nhân, các thuật toán được đề xuất sử dụng kỹ thuật sắp xếp tập dữ liệu, được đặt tên là Sắp xếp theo trọng số, sắp xếp các giao dịch nhạy cảm theo trọng số ước tính của chúng theo thứ tự giảm dần. Điều này là do các giao dịch có trọng số cao hơn rất có thể sẽ hỗ trợ số lượng tập mục nhạy cảm cao hơn và số lượng tập mục không nhạy cảm thấp hơn. Điều này chắc chắn đảm bảo rằng các mẫu không nhạy cảm sẽ có khả năng chống biến dạng tốt hơn và nhiều mẫu nhạy cảm sẽ bị ảnh hưởng hoặc thậm chí bị che giấu bằng cách sửa đổi giao dịch của nạn nhân. Để lựa chọn các mục nạn nhân, ba thuật toán được đề xuất sử dụng hiệu quả các giá trị RISU được tính toán của các mục nhạy cảm để chỉ định một mục nạn nhân xác định cho mỗi tập mục nhạy cảm. Do đó, điều này sẽ đẩy nhanh toàn bộ quá trình nhiễu loạn vì các thuật toán được đề xuất sẽ không cần phải chọn và đánh giá nhiều mục nạn nhân cho mỗi SHI trong quy trình dọn dẹp. Để tối ưu hóa quá trình ẩn, các tập mục nhạy cảm được xử lý theo thứ tự giảm dần của các giá trị RISU của các mục nạn nhân được chỉ định trước. Dựa trên ba tiêu chí khác nhau để chọn một mục nạn nhân cho mỗi tập mục nhạy cảm và một lần quét tập dữ liệu, chúng tôi đề xuất ba thuật toán SMRF, SLRF và SDIF. Hình 2 thể hiện cách làm việc tổng thể của các thuật toán được đề xuất. Giải thích về quy trình vệ sinh của ba thuật toán được đề xuất được đưa ra như sau:

### 4.1. Thuật toán SMRF

Mã giả của thuật toán SMRF có thể được xem trong Thuật toán 1. Đầu tiên, tập dữ liệu được quét một lần để xác định các giao dịch của nạn nhân và tính toán giá trị RISU của các mục nhạy cảm bằng cách sử dụng Eqn 2 (Dòng 1). Sau đó, đối với mỗi tập mục nhạy cảm, mục có giá trị RISU cao nhất sẽ được chọn làm mục nạn nhân của tập mục này (Dòng 2–4). Sau đó, đối với mỗi giao dịch nhạy cảm, trọng số nhạy cảm được tính bằng Công thức 1 (Dòng 5–7). Ở dòng 8, các giá trị trọng số ước tính sau đó được sử dụng để áp dụng kỹ thuật Sắp xếp theo Trọng số. Dòng 9 sắp xếp các tập mục nhạy cảm theo thứ tự không tăng của các giá trị RISU của các mục nạn nhân đã chọn của chúng. Sau đó, các mẫu nhạy cảm (SHI) sẽ được ẩn theo cách riêng lẻ (Dòng 10–31). Thuật toán lặp đi lặp lại ẩn từng SHI bằng cách tính toán chênh lệch bằng phương trình ở Dòng 11. Giá trị chênh lệch phản ánh tiện ích tối thiểu cần giảm khỏi các tập mục nhạy cảm để nó bị ẩn. Sau đó, thuật toán bắt đầu duyệt qua các giao dịch nhạy cảm đã được sắp xếp để xác định giao dịch nạn nhân và sửa đổi mục nạn nhân của mẫu đang được xử lý. Nếu một giao dịch hỗ trợ được tìm thấy và giá trị chênh lệch vẫn lớn hơn 0 thì mục nạn nhân trong giao dịch này sẽ bị xóa hoặc sửa đổi (Dòng 13–27). Trong trường hợp giá trị chênh lệch lớn hơn tiện ích của mục nạn nhân trong giao dịch nạn nhân, mục nạn nhân sẽ bị xóa khỏi giao dịch (Dòng 16–19). Nếu không đúng như vậy, số lượng mục nạn nhân sẽ giảm theo giá trị được tính ở dòng 21. Trong cả hai trường hợp trước, tiện ích của phần nhạy cảm còn lại




Hình 2. Cách thức hoạt động của các thuật toán đề xuất.

các mẫu có mục nạn nhân này và tồn tại trong cùng một giao dịch nạn nhân cũng bị giảm đi. Sau đó, các giao dịch nhạy cảm đã sửa đổi sẽ được chèn vào tập dữ liệu gốc khi tất cả các mẫu nhạy cảm được ẩn hoàn toàn và cuối cùng một tập dữ liệu đã được làm sạch hoàn toàn được tạo ra (Dòng 32–33).

### 4.2. Thuật toán SLRF

Thuật toán SLRF tuân theo phương pháp tương tự như thuật toán SMRF, ngoại trừ việc trong quá trình chọn mục, mục có giá trị RISU ít nhất trong mẫu nhạy cảm được chọn làm mục nạn nhân của mẫu này trong toàn bộ quá trình dọn dẹp. Thuật toán 2 hiển thị quy trình vệ sinh của SLRF được đề xuất. Đầu tiên, quá trình quét tập dữ liệu được thực hiện để xác định giao dịch nhạy cảm và ước tính giá trị RISU (Dòng 1). Sau đó, chúng tôi chỉ định mục nạn nhân cho từng tập mục nhạy cảm bằng cách chọn mục có giá trị RISU ít nhất trong số các mục nhạy cảm của mỗi tập mục (Dòng 2–4). Trong các dòng 5–8, trọng số nhạy cảm của mỗi giao dịch được ước tính và kỹ thuật Sắp xếp theo Trọng số sau đó được áp dụng cho các giao dịch nhạy cảm. Tiếp theo, tập mục nhạy cảm có mục nạn nhân có giá trị RISU cao nhất so với giá trị RISU của các mục nạn nhân được chọn trong các tập mục nhạy cảm khác, sẽ được chọn để khử trùng trước tiên. Sau đó, thuật toán tiến hành với kỹ thuật ẩn tương tự được trình bày trước đó trong thuật toán SMRF.

### 4.3. Thuật toán SDIF

Thuật toán 3 mô tả quy trình tổng thể của thuật toán SDIF. Không giống như hai thuật toán trước, chỉ xem xét các giá trị RISU để chọn mục nạn nhân cho từng tập mục nhạy cảm, thuật toán SDIF áp dụng các tiêu chí đánh giá nâng cao hơn. Trong dòng (2–9), quá trình lựa chọn mục nạn nhân sẽ xem xét ba vấn đề khác nhau.

các yếu tố khác; (1) bìa nhạy cảm của từng mặt hàng; (2) vỏ bọc không nhạy cảm của từng hạng mục; (3) giá trị RISU của từng hạng mục. Trong quá trình đánh giá hạng mục, trước tiên thuật toán sẽ cố gắng chọn hạng mục mong muốn nhất, đó là hạng mục có bìa nhạy cảm nhất và có bìa ít nhạy cảm nhất so với các hạng mục khác. Nếu không tìm thấy mục mong muốn nào, thuật toán sẽ chọn mục có vỏ bọc ít nhạy cảm nhất làm mục nạn nhân. Ở đây, các giá trị RISU được tận dụng làm công cụ ngắt kết nối cho cả hai trường hợp trước đó sao cho thuật toán luôn chọn mục có giá trị RISU ít nhất nếu có nhiều hơn một mục nạn nhân ứng cử viên cho mẫu nhạy cảm. Các giao dịch nhạy cảm sau đó được sắp xếp theo trọng số của chúng, trong khi các tập mục nhạy cảm được sắp xếp theo giá trị RISU của các mục nạn nhân (Dòng 10–14). Trong các dòng (15–36), quy trình dọn dẹp được áp dụng lặp đi lặp lại cho các mẫu nhạy cảm cho đến khi tất cả chúng đều bị ẩn.

## 5. Ví dụ minh họa

Trong phần này, chúng tôi cung cấp một ví dụ minh họa để làm sáng tỏ hơn quá trình dọn dẹp của ba thuật toán được đề xuất. Hãy xem xét tập dữ liệu D trong Bảng 2 và các giá trị lợi nhuận của các mặt hàng trong Bảng 3, các tập mục có tiện ích cao khi minUtil = 70 được mô tả trong Bảng 4 và các tập mục nhạy cảm với các giao dịch của chúng được nêu trong Bảng 5. Ba thuật toán bắt đầu bằng quét tập dữ liệu gốc một lần để xác định các giao dịch của nạn nhân và tính toán giá trị RISU của các mục nhạy cảm. Giá trị RISU của tất cả các mục nhạy cảm được cung cấp trong Bảng 8. Sau đó, tất cả các tập mục có tính tiện ích cao được quét một lần để ước tính mức độ nhạy cảm và không nhạy cảm của các mục như được trình bày trong Bảng 6. Ngoài ra, các giao dịch nhạy cảm cũng được quét một lần để ước tính mức độ nhạy cảm và không nhạy cảm của các giao dịch như trong Bảng 7. Sau đó, mỗi thuật toán thực hiện theo một cách khác nhau để chỉ định mục nạn nhân lý tưởng

Thuật toán 1: Thuật toán SMRF.

Đầu vào: D*, tập dữ liệu nhạy cảm; SHI, tập hợp các tập mục có tính tiện ích cao, nhạy cảm; NHI, tập hợp các tập mục có tính tiện ích cao không nhạy cảm; minUtil, ngưỡng tiện ích tối thiểu.

Đầu ra: D′ , một tập dữ liệu đã được làm sạch. 1 Quét tập dữ liệu một lần để xác định các giao dịch nhạy cảm và tính giá trị RISU của các mục nhạy cảm bằng phương trình. (2)

2 foreach If ∈ SHI của 3 Ivic (If ) ← xi sao cho xi ∈ If và ∀xj ∈ If RỦI RO (xi ) ≥ RỦI RO (xj )

4 end foreach 5 foreach giao dịch Ti ∈ D* do 6 Tính trọng số của Ti bằng phương trình. (1) 7 phần cuối bài giảng

8 Sắp xếp các giao dịch Ti ∈ D* theo thứ tự trọng lượng giảm dần (Sắp xếp có trọng số)

9 Sắp xếp các tập mục Si ∈ SHI theo thứ tự giảm dần của giá trị RISU của các mục nạn nhân đã được chọn Ivic (Si )

10 foreach Si trong các SHI được sắp xếp thực hiện 11 di ff = u(Si ) − minUtil + 1

```
12 foreach giao dịch Ti ∈ D* do
13     nếu di ff > 0 thì
14         nếu Si ⊆ Ti thì
15             Tvic (Tới ) = Nếu
16             nếu ff ≥ u(Ivic (Si ), Tvic (Si )) thì
17                 xóa Ivic khỏi Tvic
18                 di ff = di ff − u(Ivic (Si ), Tvic (Si ))
19                 Cập nhật S j ∈ SHI sao cho S j ⊆ Tvic và
                   Ivic (Si ) ∈ S j
20             kết thúc nếu
21             khác
22                 diu = ⌈di ff /i(Ivic (Si ))⌉
23                 iu(Ivic (Si ), Tvic (Si )) = iu(Ivic (Si ), Tvic (Si )) − diu
24                 Cập nhật S j ∈ SHI sao cho S j ⊆ Tvic và
                   Ivic (Si ) ∈ S j
25                 di ff = 0
26             kết thúc nếu
27         kết thúc nếu
28     khác
29         phá vỡ
30     kết thúc nếu
31 end foreach 33 end foreach
```

34 D′ ← D* 35 trả về tập dữ liệu đã được lọc sạch D′

Thuật toán 2: Thuật toán SLRF.

Đầu vào: D*, tập dữ liệu nhạy cảm; SHI, tập hợp các tập mục có tính tiện ích cao, nhạy cảm; NHI, tập hợp các tập mục có tính tiện ích cao không nhạy cảm; minUtil, ngưỡng tiện ích tối thiểu.

Đầu ra: D′ , một tập dữ liệu đã được làm sạch. 1 Quét tập dữ liệu một lần để xác định các giao dịch nhạy cảm và tính giá trị RISU của các mục nhạy cảm bằng phương trình. (2)

2 foreach If ∈ SHI của 3 Ivic (If ) ← xi sao cho xi ∈ If và ∀xj ∈ If RỦI RO (xi ) ≤ RỦI RO (xj )

4 end foreach 5 foreach giao dịch Ti ∈ D* do 6 Tính trọng số của Ti bằng phương trình. (1) 7 phần cuối bài giảng

8 Sắp xếp các giao dịch Ti ∈ D* theo thứ tự trọng lượng giảm dần (Sắp xếp có trọng số)

9 Sắp xếp các tập mục Si ∈ SHI theo thứ tự giảm dần của giá trị RISU của các mục nạn nhân đã được chọn Ivic (Si )

10 foreach Si trong các SHI được sắp xếp thực hiện 11 di ff = u(Si ) − minUtil + 1

```
12 foreach giao dịch Ti ∈ D* do
13     nếu di ff > 0 thì
14         nếu Si ⊆ Ti thì
15             Tvic (Tới ) = Nếu
16             nếu ff ≥ u(Ivic (Si ), Tvic (Si )) thì
17                 xóa Ivic khỏi Tvic
18                 di ff = di ff − u(Ivic (Si ), Tvic (Si ))
19                 Cập nhật S j ∈ SHI sao cho S j ⊆ Tvic và
                   Ivic (Si ) ∈ S j
20             kết thúc nếu
21             khác
22                 diu = ⌈di ff /i(Ivic (Si ))⌉
23                 iu(Ivic (Si ), Tvic (Si )) = iu(Ivic (Si ), Tvic (Si )) − diu
24                 Cập nhật S j ∈ SHI sao cho S j ⊆ Tvic và
                   Ivic (Si ) ∈ S j
25                 di ff = 0
26             kết thúc nếu
27         kết thúc nếu
28     khác
29         phá vỡ
30     kết thúc nếu
31 end foreach 33 end foreach
```

34 D′ ← D* 35 trả về tập dữ liệu đã được lọc sạch D′

cho từng tập mục nhạy cảm. Xem xét ví dụ hiện tại, các tập mục nhạy cảm là {A, B, E}, {B, C, E} và {A, B, C, D, E}. Đối với thuật toán SMRF, mục có giá trị RISU tối đa trong mỗi tập mục nhạy cảm sẽ được chọn làm mục nạn nhân trong toàn bộ quá trình sanitization. Đối với tập mục {A, B, E}, mục E được chọn để đại diện cho mục nạn nhân của tập mục này. Điều này là do RISU(A) = 48, RISU(B) = 48 và RISU(E) = 54. Tương tự, mục E được chọn trong {B, C, E} và {A, B, C, D, E} itemset. Đối với thuật toán SLRF, mục có giá trị RISU tối thiểu sẽ là mục nạn nhân. Đối với tập mục {A, B, E}, mục A và mục B có giá trị RISU nhỏ nhất, do đó bất kỳ mục nào trong số chúng đều có thể được chọn. Đối với các tập mục {B, C, E} và {A, B, C, D, E}, các mục C và D lần lượt được chọn làm mục nạn nhân cho cả hai tập mục. Điều này là do RISU(C) = 18 và RISU(D) = 10. Thuật toán SDIF xem xét Độ nhạy

Bao gồm SC và NSC không nhạy cảm cho các hạng mục trong quá trình đánh giá. Mục mong muốn nhất, có SC tối đa và NSC tối thiểu, được chọn trong mỗi tập mục. Đối với tập mục {A, B, E}, SC tối đa = 3 và NSC tối thiểu = 2. Do đó, mục B được chọn làm mục nạn nhân của tập mục trước đó. Đối với tập mục {B, C, E}, SC tối đa = 3 và NSC tối thiểu = 0. Do đó, không có mục mong muốn nào được tìm thấy trong tập mục này. Trong trường hợp như vậy, mục C được chọn làm mục nạn nhân vì nó có giá trị NSC tối thiểu. Đối với tập mục {A, B, C, D, E}, không tìm thấy mục mong muốn nào và có hai mục C và D có giá trị NSC tối thiểu. Trong trường hợp này, mục có giá trị RISU ít nhất được chọn, đó là mục D. Sau đó, sau khi xác định được mục nạn nhân, kỹ thuật Sắp xếp theo trọng số sẽ được áp dụng cho các giao dịch nhạy cảm. Vì các giao dịch nhạy cảm là , nên lệnh vệ sinh sẽ là

Thuật toán 3: Thuật toán SDIF.

Đầu vào: D*, tập dữ liệu nhạy cảm; SHI, tập hợp các tập mục có tính tiện ích cao, nhạy cảm; NHI, tập hợp các tập mục có tính tiện ích cao không nhạy cảm; minUtil, ngưỡng tiện ích tối thiểu.

Đầu ra: D′ , một tập dữ liệu đã được làm sạch. 1 Quét tập dữ liệu một lần để xác định các giao dịch nhạy cảm và tính giá trị RISU của các mục nhạy cảm bằng phương trình. (2)

2 foreach Si ∈ SHI thực hiện 3 CandIvic ← xi sao cho NSC(xi ) ≤ NSC(xj ) và SC(xi ) ≥ SC(xj ) ∀xj ∈ Si

4 nếu |CandIvic | ≥ 1 thì
5 Ivic (Si ) ← xi sao cho RISU (xi ) < RISU (xj ) ∀xj ∈ CandIvic
kết thúc nếu
6
7 khác
8 Ivic (Si ) ← xi sao cho NSC(xi ) < NSC(xj ) và RISU (xi ) < RISU (xj ) ∀xj ∈ Si

9 end if 10 end foreach 11 foreach giao dịch Ti ∈ D* do 12 Tính trọng số của Ti bằng phương trình. (1) 13 phần cuối bài giảng

14 Sắp xếp các giao dịch Ti ∈ D* theo thứ tự trọng lượng giảm dần (Sắp xếp có trọng số)

15 Sắp xếp các tập mục Si ∈ SHI theo thứ tự giảm dần của giá trị RISU của các mục nạn nhân đã được chọn Ivic (Si )

16 foreach Si trong các SHI được sắp xếp thực hiện 17 di ff = u(Si ) − minUtil + 1

18 foreach giao dịch Ti ∈ D* do
19 nếu di ff > 0 thì
20 nếu Si ⊆ Ti thì
21 Tvic (Tới ) = Nếu
22 nếu ff ≥ u(Ivic (Si ), Tvic (Si )) thì
23 xóa Ivic khỏi Tvic
24 di ff = di ff − u(Ivic (Si ), Tvic (Si ))
25 Cập nhật S j ∈ SHI sao cho S j ⊆ Tvic và Ivic (Si ) ∈ S j
kết thúc nếu
26
27 khác
28 diu = ⌈di ff /i(Ivic (Si ))⌉
29 iu(Ivic (Si ), Tvic (Si )) = iu(Ivic (Si ), Tvic (Si )) − diu
30 Cập nhật S j ∈ SHI sao cho S j ⊆ Tvic và Ivic (Si ) ∈ S j
31 di ff = 0
kết thúc nếu
32
33 kết thúc nếu
34 kết thúc nếu
35 khác
36 phá vỡ
37 kết thúc nếu

38 kết thúc foreach 39 kết thúc foreach 40 D′ ← D*; trả về tập dữ liệu đã được làm sạch D′

T4 ≺ T6 ≺ T1 ≺ T5 . Trong bước tiếp theo, chúng tôi sắp xếp các tập mục nhạy cảm dựa trên giá trị RISU của các mục nạn nhân theo thứ tự giảm dần. Theo đó, mỗi thuật toán hoạt động như tiếp theo.

Trong thuật toán SMRF, thứ tự xử lý các tập mục nhạy cảm là {B, C, E} - {A, B, C, DE} - {A, B, E}. Đối với tập mục {B, C, E}, giá trị sai phân (di ff ) cần giảm để ẩn tập mục này là di ff = 106 − 70 + 1 = 37. Đối với ba giao dịch T4 , T5 và T6 chứa tập mục đang được xử lý, giao dịch T4 đầu tiên được chọn làm giao dịch nạn nhân vì nó có trọng số cao hơn hai giao dịch còn lại. Vì tiện ích của mục nạn nhân E trong T4 là 12, nhỏ hơn giá trị của khác biệt nên mục E bị xóa khỏi giao dịch T4 và tiện ích của {B, C, E} giảm xuống còn 74. Bởi vì cả hai tập mục nhạy cảm {A, B, E} và {A, B, C, D, E} đều xuất hiện trong giao dịch hiện tại T4 và chứa mục E, tiện ích của chúng cũng giảm lần lượt xuống 32 và 0. Tiếp theo, {B, C, E} tiếp tục bị ẩn khi giao dịch T6 được chọn để sửa đổi. Do tiện ích của E trong T6 nhỏ hơn diff nên tiện ích bên trong của E giảm từ 6 xuống 4 và tiện ích của tập mục hiện tại giảm xuống 68, thấp hơn minUtil đã cho. Kết quả là tập mục {B, C, E} bị ẩn hoàn toàn và các tập mục nhạy cảm còn lại, {A, B, E} và {A, B, C, D, E}, không cần phải xử lý vì chúng cũng bị ẩn. Tuy nhiên, các tập mục không nhạy cảm, {A, E} và {B, E}, bị mất do lỗi.

Đối với thuật toán SLRF, thứ tự xử lý các tập mục nhạy cảm là {A, B, E} - {B, C, E} - {A, B, C, DE}. Đối với tập mục {A, B, E}, tiện ích nội tại của mục nạn nhân A của nó bị giảm từ 5 xuống 1 trong giao dịch nạn nhân T4 . Kết quả là cả hai tập mục {A, B, E} và {A, B, C, DE} đều bị che. Để ẩn tập mục {B, C, E}, mục C nạn nhân của nó lần lượt bị xóa khỏi các giao dịch T4 và T6 của nạn nhân. Kết quả là, đối với quy trình dọn dẹp, các tập mục không nhạy cảm, {A, E} và {A, B}, cũng bị ẩn.

Trong thuật toán SDIF, thứ tự xử lý các tập mục nhạy cảm là {B, C, E} - {A, B, E} - {A, B, C, DE}. Trong quá trình xử lý tập mục {B, C, E}, các giao dịch T4 và T6 lần lượt được chọn làm giao dịch nạn nhân. Trong khi đó, mục C của nạn nhân bị xóa khỏi cả hai giao dịch để ẩn tập mục nhạy cảm hiện tại, trong khi tiện ích của tập mục {A, B, C, DE} giảm xuống 0 do nó không còn tồn tại trong giao dịch T4. Theo cách tương tự, tập mục {A, B, E} được che giấu bằng cách loại bỏ mục B nạn nhân của nó khỏi giao dịch nạn nhân T4. Sau khi dọn sạch, chỉ có tập mục không nhạy cảm {A, B} bị ẩn quá mức.

## 6. Đánh giá thực nghiệm

Để đánh giá hiệu quả của các thuật toán SMRF, SLRF và SDIF được đề xuất, nhiều thử nghiệm chuyên sâu đã được tiến hành trên bốn bộ dữ liệu chuẩn, thường được sử dụng để đánh giá hiệu suất trong lĩnh vực khai thác mẫu. Các bộ dữ liệu này được tải xuống từ trang web SPMF1 (Fournier-Viger và cộng sự, 2014), đây là một thư viện khai thác dữ liệu nguồn mở. Bảng 9 tóm tắt các đặc điểm của các bộ dữ liệu được áp dụng. Các bộ dữ liệu này được chọn một cách cụ thể vì chúng có tính chất khác nhau (giao dịch dày đặc, thưa thớt, lớn và dài) và do đó thể hiện tốt các loại dữ liệu chính được thấy trong các ứng dụng trong thế giới thực. Các tính năng trong Bảng 9 bao gồm cho mọi tập dữ liệu: tên tập dữ liệu, số lượng giao dịch, số mục riêng biệt, số mục trung bình trong mỗi giao dịch, độ dài tối đa của giao dịch, loại, tỷ lệ phần trăm mật độ và một nhận xét ngắn tương ứng. Để nhận biết các tập dữ liệu dày đặc và thưa thớt, mật độ được tính bằng cách chia độ dài giao dịch trung bình của tập dữ liệu cho số mục riêng biệt trong đó. Các tập dữ liệu có tỷ lệ phần trăm mật độ lớn hơn 5% được phân loại là dày đặc. Cờ vua là một tập dữ liệu rất dày đặc với mật độ gần 50%. Nấm cũng dày đặc với nhiều vật phẩm khác biệt hơn cờ vua. Tập dữ liệu tai nạn là tập dữ liệu dày đặc nhất vì nó chứa 340.183 giao dịch với trung bình 33,8 mục trên mỗi giao dịch. Ngược lại, tập dữ liệu bán lẻ có tỷ lệ mật độ rất nhỏ vì nó bao gồm số lượng mặt hàng tương đối lớn và giao dịch ngắn.

Hiệu suất tổng thể của các thuật toán đề xuất đã được đánh giá dựa trên các thuật toán cơ bản và hiện đại sau:

1 https://www.philippe-fournier- viger.com/spmf/

Bảng 9 Các đặc điểm của bộ dữ liệu được nghiên cứu.

| Tập dữ liệu | #Trans | #Mặt hàng | Trung bình chiều dài | Tối đa. chiều dài | Kiểu | Tỉ trọng(%) | Bình luận |
|---|---|---|---|---|---|---|---|
| cờ vua | 3196 | 75 | 37 | 37 | dày đặc | 49,3 | Nước cờ hợp pháp của một ván cờ |
| nấm | 8124 | 119 | 23 | 23 | dày đặc | 19.3 | Thông tin về các loại nấm |
| tai nạn | 340.183 | 468 | 33,8 | 51 | Lớn | 7.2 | Dữ liệu tai nạn giao thông ẩn danh trong giai đoạn 19912000 |
| bán lẻ | 88.162 | 16.470 | 10.3 | 76 | thưa thớt | 0,062 | Giao dịch của khách hàng từ một cửa hàng bán lẻ ẩn danh của Bỉ |

Bảng 10 Cài đặt tham số của bộ dữ liệu điểm chuẩn.

| cờ vua | | nấm | | tai nạn | | bán lẻ | |
|---|---|---|---|---|---|---|---|
| #SHI | TRONG(%) | #SHI | TRONG(%) | #SHI | TRONG (%) | #SHI | TRONG(%) |
| đa dạng | 21 | đa dạng | 7 | đa dạng | 1.2 | đa dạng | 0,025 |
| 100 | đa dạng | 50 | đa dạng | 15 | đa dạng | 50 | đa dạng |

HHUIF (Yeh và Hsu, 2010), MSICF (Yeh và Hsu, 2010), MSU-MIU (Lin và cộng sự, 2016), MSU-MAU (Lin và cộng sự, 2016), SMAU (Liu và cộng sự, 2020b ) và Weighted_RoT_DoC (Jangra và Toshniwal, 2022), bao gồm cả hiệu suất của chúng khi thay đổi (1) giá trị minUtil và (2) số lượng tập mục nhạy cảm. Trong quá trình thử nghiệm, các tập mục có tính tiện ích cao được tạo ra lần đầu tiên bằng cách áp dụng thuật toán EFIM (Zida và cộng sự, 2015). Sau đó, các tập mục nhạy cảm được chọn ngẫu nhiên và đồng thời xem xét các tiêu chuẩn văn học (Yun và Kim, 2015). Đại khái, một tập mục nhạy cảm chủ yếu phải thỏa mãn hai điều kiện: (1) đó là tập mục có tính tiện ích cao; và (2) nó bao gồm hai hoặc nhiều mục.

Tất cả các mã được triển khai bằng java và các thử nghiệm được thực hiện trên máy tính được trang bị bộ xử lý intel Core-i7 2,1 GHz thế hệ thứ ba, bộ nhớ 6 GB và chạy HĐH Windows 7. Trong quá trình thử nghiệm, như có thể thấy trong Bảng 10, chúng tôi đặt minUtil (MU) và số lượng tập mục nhạy cảm (SHI) sao cho mỗi tác vụ ẩn có thời gian chạy thích hợp và có thể đo lường được.

### 6.1. Chỉ số đánh giá hiệu quả

Bảy số liệu xác nhận được sử dụng rộng rãi trong tài liệu PPUM (Jangra và Toshniwal, 2022; Lin và cộng sự, 2016; Yeh và Hsu, 2010) được đề xuất cho các thí nghiệm. Chúng như sau:

1) Lỗi ẩn (HF). Số liệu này nhằm mục đích đánh giá tỷ lệ các tập mục nhạy cảm vẫn tồn tại trong tập dữ liệu ngay cả sau khi quá trình dọn dẹp kết thúc. Phần trăm HF được tính như sau:

$$HF = |SH là' | - |SH là| \quad (3)$$

ở đây | SHI | và | SHIs′ | đề cập đến số lượng bộ mục nhạy cảm tương ứng trước và sau quá trình dọn dẹp. Giá trị lý tưởng của HF là 0, điều này cho thấy rằng tất cả các bộ mục nhạy cảm đã bị ẩn sau quá trình dọn dẹp.

2) Chi phí nhân tạo (AC). Số liệu này nhằm mục đích đánh giá tỷ lệ các tập mục có tính tiện ích cao giả được tạo ra do quá trình biến dạng của PPUM. Phần trăm AC được tính như sau:

$$AC = |HUI - HUI' | - |NHÀ' | \quad (4)$$

ở đây | HUI | và | HUI' | đề cập đến số lượng các tập mục có tiện ích cao trước và sau quá trình dọn dẹp tương ứng. Giá trị lý tưởng của AC bằng 0, điều này cho thấy không có mẫu giả nào được tạo ra.

3) Thiếu chi phí (MC). Số liệu này nhằm mục đích đánh giá tỷ lệ các tập mục không nhạy cảm đã bị mất sau khi vệ sinh. quá trình zation. Phần trăm MC được tính như sau: MC = | NHI′ | | NHI | (5)

ở đây | NHI | và | NHI′ | đề cập đến số lượng tập mục không nhạy cảm trước và sau quá trình dọn dẹp tương ứng. Giá trị lý tưởng của MC bằng 0, điều này cho thấy rằng tất cả các tập mục không nhạy cảm vẫn tồn tại trong tập dữ liệu đã được lọc sạch.

4) Tính tương tự của tiện ích tập mục (IUS). Số liệu này phản ánh sự thiếu hụt về tổng tiện ích của các tập mục có tiện ích cao ban đầu do hậu quả của quá trình biến dạng của PPUM. Phần trăm IUS được tính như sau:

$$IUS = \frac{\sum_{Y \in HUIs'} u(Y)}{\sum_{Y \in HUI} u(Y)}. \quad (6)$$

trong đó HUI và HUI ′ lần lượt đề cập đến các tập mục có tiện ích cao được tạo ra trước và sau quá trình dọn dẹp. Nói chung, khi chi phí còn thiếu tăng lên thì IUS sẽ giảm.

5) Tương tự tiện ích tập dữ liệu (DUS). Số liệu này cho thấy sự thiếu hụt về tổng thể tiện ích của tập dữ liệu do quá trình bóp méo PPUM. Phần trăm DUS được tính như sau:

$$BẠN S = \frac{\sum_{Td \in D'} TU(Td)}{\sum_{Td \in DTU} (Td)}. \quad (7)$$

trong đó D và D′ lần lượt đề cập đến tập dữ liệu trước và sau quá trình làm sạch và TU (Td ) đề cập đến tổng tiện ích của giao dịch Td . Rõ ràng, giá trị của DUS càng cao thì chất lượng của tập dữ liệu đã được làm sạch càng tốt.

6) Tỷ lệ sửa đổi giao dịch (TMR). Số liệu này cho biết tỷ lệ phần trăm giao dịch đã bị sai lệch trong quá trình vệ sinh. TMR có thể được tính như sau:

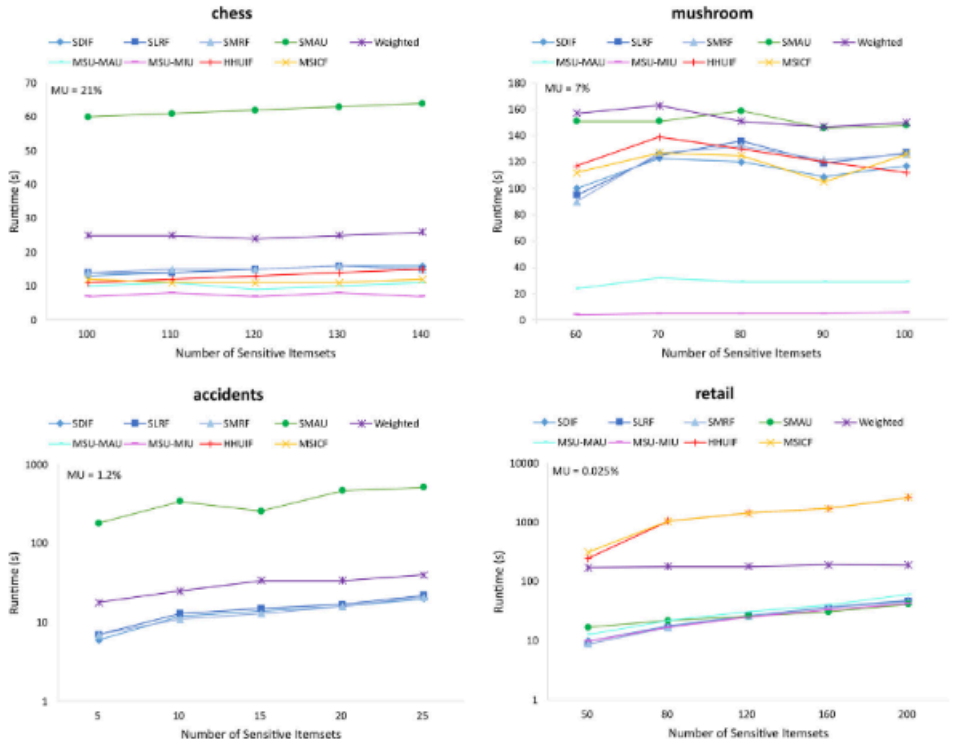
$$T MR = \frac{\text{tổng số giao dịch được sửa đổi}}{\#\text{ giao dịch}}. \quad (8)$$

7) Thời gian chạy (RT). Thời gian xử lý là thước đo quan trọng để đánh giá hiệu quả của bất kỳ thuật toán PPDM nào. Trong PPUM, thời gian ẩn có thể dao động tùy thuộc vào nhiều yếu tố như (1) mật độ tập dữ liệu (dày đặc, thưa thớt), (2) kích thước tập dữ liệu (lớn, nhỏ), (3) số lượng tập mục được yêu cầu ẩn và (4) số lần quét tập dữ liệu cần thiết để hoàn tất quá trình dọn dẹp.

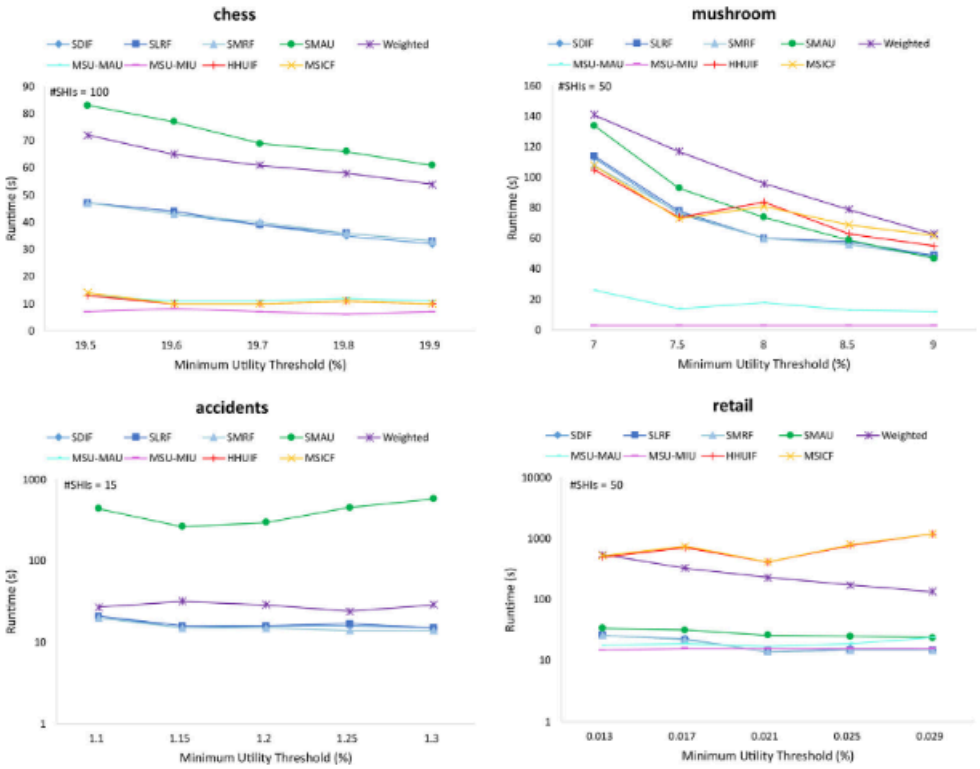
Vì tất cả các thuật toán được so sánh đều tuân theo mô hình làm sạch dựa trên vật phẩm và có thể ẩn thành công tất cả các tập mục có tính hữu ích cao nhạy cảm mà không tạo ra bất kỳ mẫu sai nào, nên chúng tôi thấy không hợp lý khi xem xét các số liệu về Lỗi ẩn và Chi phí nhân tạo trong đánh giá, và do đó kết quả của hai số liệu này không được trình bày trong bài viết này.

### 6.2. Thời gian chạy

Trong phần này, chúng tôi đánh giá thời gian cần thiết để hoàn thành quá trình ẩn trong tất cả các thuật toán được so sánh. Quả sung. 3 và 4 mô tả so sánh thời gian chạy giữa ba thuật toán được đề xuất và các thuật toán khác trong khi thay đổi số lượng




Hình 3. So sánh thời gian chạy bằng cách sử dụng nhiều tập mục nhạy cảm khác nhau.




Hình 4. So sánh thời gian chạy sử dụng các ngưỡng tiện ích tối thiểu khác nhau.

các tập mục nhạy cảm và các giá trị minUtil tương ứng. Như có thể thấy trong Hình 3, thời gian chạy chủ yếu có xu hướng tăng khi số lượng tập mục nhạy cảm tăng lên. Điều này là do càng cần nhiều tập mục cần ẩn thì càng cần nhiều thời gian để hoàn tất quá trình dọn dẹp. Điều này dường như ngược lại trong Hình 4 vì khi ngưỡng minUtil tăng thì số lượng tập mục có tiện ích cao được tạo ra sẽ giảm và do đó cần xem xét ít tập mục không nhạy cảm hơn. Từ các kết quả trong Hình. Như trong Hình 3 và 4, có thể nhận thấy rằng thời gian chạy của các thuật toán được đề xuất có thể dao động tùy thuộc vào loại, kích thước tập dữ liệu và tập mục được chọn để dọn dẹp. Chúng tôi cũng có thể xác nhận rằng hiệu quả thời gian chạy của các thuật toán được đề xuất là tốt nhất so với các thuật toán PPUM gần đây nhất (SMAU và Weighted) trong tất cả các bộ dữ liệu được nghiên cứu. Những lý do như sau. SMAU cần nhiều thời gian hơn để xây dựng và cập nhật cấu trúc dữ liệu đất tiền của nó, trong khi thuật toán Weighted yêu cầu nhiều thời gian hơn cho kỹ thuật sắp xếp kép của nó. Trong tập dữ liệu lớn, các vụ tai nạn, các thuật toán đề xuất hiển thị thời gian chạy nhanh nhất so với các thuật toán PPUM mới nhất. Vì chúng tôi không thể hoàn thành việc chạy các thuật toán cơ bản; HHUIF, MSICF, MSU-MAU và MSU-MIU trong tập dữ liệu về tai nạn, kết quả của chúng không được ghi lại trong tập dữ liệu này. Điều này lần lượt xác nhận rằng các thuật toán PPUM được đề xuất của chúng tôi thực sự có thể được chạy cho các bộ dữ liệu quy mô lớn một cách hiệu quả. Trong tập dữ liệu cờ vua, các thuật toán được đề xuất nhanh hơn gần hai lần so với thuật toán có trọng số và nhanh hơn sáu lần so với SMAU. Trong tập dữ liệu bán lẻ rất thưa thớt, các thuật toán được đề xuất cho thấy hiệu suất gần như tương tự với các thuật toán SMAU, MSU-MIU và MSU-MAU, đồng thời hiệu suất tốt hơn nhiều so với các thuật toán HHUIF, MSICF và Weighted. Điều này là do thuật toán HHUIF và MSICF phải quét tập dữ liệu nhiều lần để hoàn tất quá trình ẩn, trong khi thuật toán Weighted dành nhiều thời gian hơn để sắp xếp kép các giao dịch nhạy cảm. Khoảng cách hiệu suất giữa các thuật toán của chúng tôi và các thuật toán PPUM gần đây nhất là khá rõ ràng trong tập dữ liệu về nấm. Tuy nhiên, các thuật toán cơ bản, MSU-MAU và MSU-MIU, hoạt động tốt hơn các thuật toán được đề xuất trong tập dữ liệu nấm vì chúng chọn các mục và giao dịch nạn nhân chỉ dựa trên khái niệm tiện ích. Tóm lại, các thuật toán được đề xuất đơn giản là nhanh vì những lý do sau. Đầu tiên, họ chỉ yêu cầu quét tập dữ liệu một lần để xác định các giao dịch nhạy cảm và ước tính RISU của các mục nhạy cảm. Thứ hai, bằng cách tận dụng kiến thức được biểu thị bằng các giá trị RISU, họ có thể chỉ định mục nạn nhân lý tưởng cho từng tập mục nhạy cảm và do đó không cần phải đánh giá lại và chọn mục nạn nhân nhiều lần trong mỗi lần lặp của quá trình vệ sinh. quá trình xác định. Thứ ba, các thuật toán được đề xuất áp dụng kỹ thuật Sắp xếp có trọng số để loại bỏ nhu cầu sắp xếp kép các giao dịch hoặc đánh giá lại các giao dịch nhạy cảm trong mỗi lần lặp để tìm ra giao dịch nạn nhân thích hợp.

### 6.3. Thiếu chi phí

Việc mất các tập mục không nhạy cảm là điều phổ biến và gần như không thể tránh khỏi trong PPUM do các mục chồng chéo giữa các tập mục nhạy cảm và không nhạy cảm. Do đó, một trong những thách thức quan trọng trong PPUM là bảo tồn kiến thức hợp pháp của tập dữ liệu gốc trong quá trình chuẩn hóa. Được thúc đẩy bởi điều này, các thuật toán được đề xuất áp dụng các chiến lược khác nhau khi quyết định chọn mục nào của nạn nhân nhằm giảm bớt số lượng mẫu vô hại bị lãng phí. Theo kết quả ở hình. Như được minh họa trong Hình 5 và 6, có thể khẳng định rằng thuật toán SDIF và SLRF được đề xuất đảm bảo hiệu suất tốt nhất trong việc giảm chi phí còn thiếu so với các đối thủ cạnh tranh khác trong tất cả các bộ dữ liệu được nghiên cứu. Các kết quả trong hình. 5 và 6 cũng chỉ ra rằng SDIF được đề xuất thậm chí còn có nhiều khả năng duy trì tập mục không nhạy cảm hơn SLRF được đề xuất trong các tập dữ liệu mật độ cao, trong đó mức độ chồng chéo giữa các tập mục nhạy cảm và không nhạy cảm là rất cao.

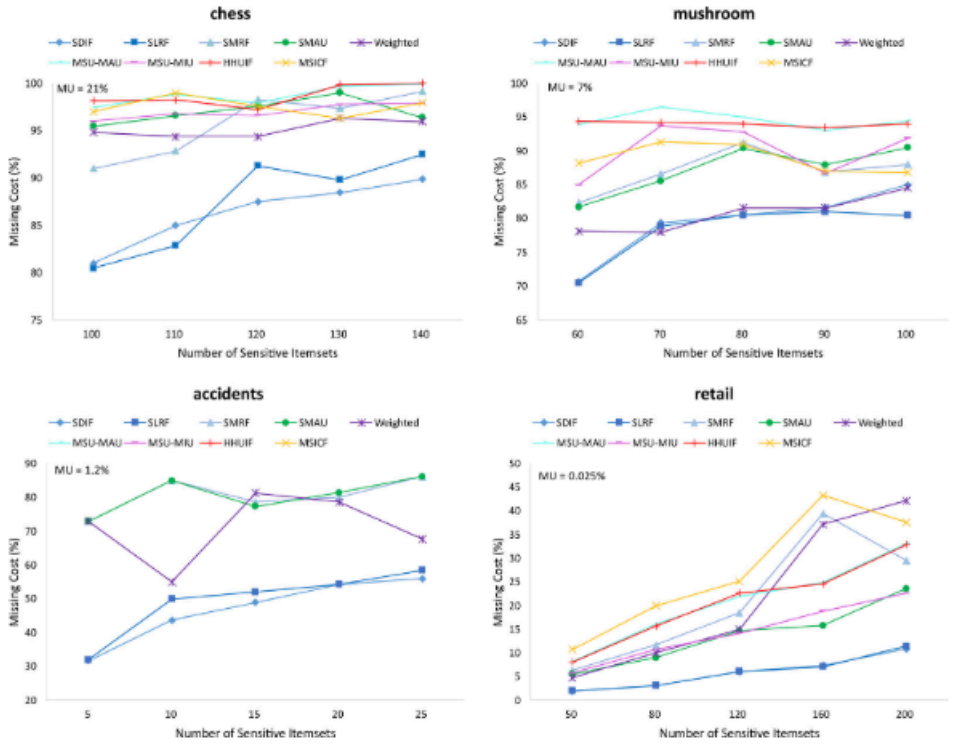
Điều này có thể được quan sát thấy trong tập dữ liệu cờ vua; khi số lượng tập mục nhạy cảm là 120, chi phí còn thiếu của SDIF thấp hơn khoảng 4% so với SLRF và khi minUtil là 19,6%, chi phí còn thiếu của SDIF thấp hơn SLRF khoảng 5%. Lý do chính đằng sau điều này là thuật toán SDIF rất chú ý đến các HUI không nhạy cảm trong khi chọn các mục nạn nhân, do đó nó mất ít kiến thức không nhạy cảm hơn thuật toán SLRF. Thật thú vị khi nhận thấy rằng thuật toán SMRF được đề xuất có hiệu suất tương đối gần với thuật toán SMAU trong hầu hết các bộ dữ liệu. Điều này là do thuật toán SMAU chọn mục có tiện ích tối đa làm mục nạn nhân, trong khi SMRF chọn mục có giá trị RISU tối đa làm mục nạn nhân lý tưởng. Ngoài ra, một kết quả đáng chú ý là chi phí còn thiếu của các thuật toán cơ bản được phát hiện là rất cao trong tập dữ liệu cờ vua. Điều này cho thấy điểm yếu của các thuật toán này trong các tập dữ liệu có mức độ chồng chéo cao giữa các tập mục nhạy cảm và không nhạy cảm. Trong tập dữ liệu về nấm, thuật toán SLRF cho thấy hiệu suất vượt trội so với các đối thủ khác. Điều này là do trong thuật toán SLRF, mục có giá trị RISU tối thiểu được chọn để khử trùng, do đó thiệt hại gây ra trên các tập mục không nhạy cảm cũng được giảm thiểu. Tóm lại, kết quả chi phí còn thiếu chứng thực trực quan rằng việc sử dụng các giá trị RISU, phản ánh ảnh hưởng của từng mục nhạy cảm đối với tất cả các giao dịch nhạy cảm, có thể dẫn đến quy trình vệ sinh thích hợp. Hơn nữa, kỹ thuật Sắp xếp theo Trọng số được đề xuất, xem xét các phạm vi giao dịch nhạy cảm và không nhạy cảm, có thể giảm đáng kể chi phí còn thiếu và làm cho các thuật toán được đề xuất thậm chí còn hiệu quả hơn.

### 6.4. Sự tương đồng về tiện ích của Itemset

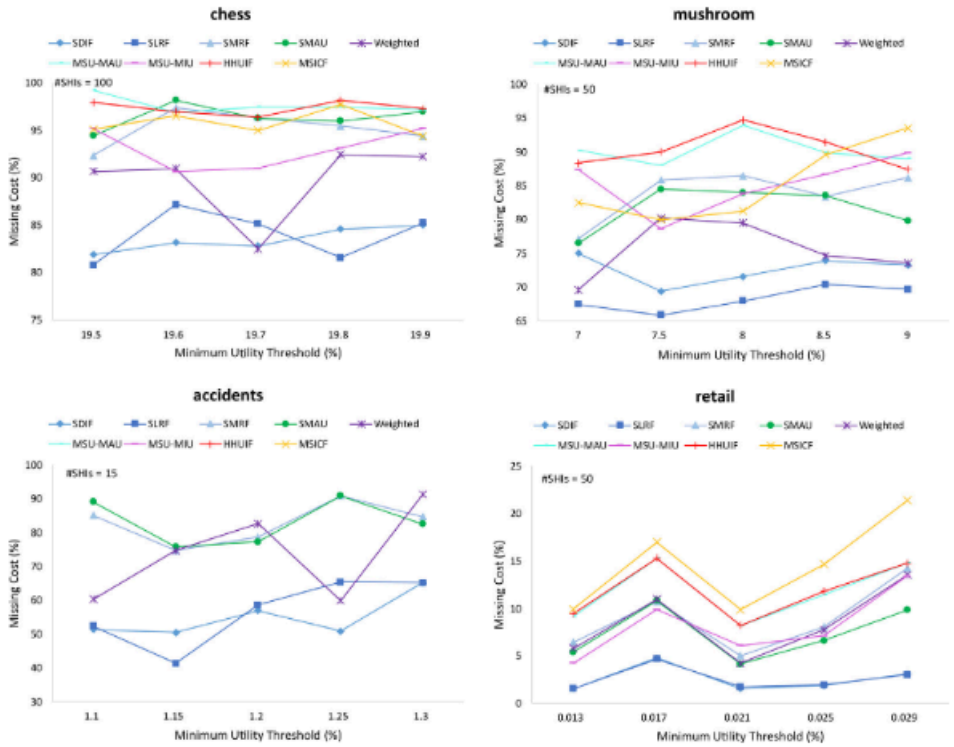
Trong phần này, các giá trị IUS của chín thuật toán được so sánh được phân tích theo số lượng tập mục nhạy cảm và giá trị minUtil khác nhau. Không giống như chi phí còn thiếu, thước đo IUS liên quan đến tổn thất về tổng tiện ích của HUI hơn là tổn thất về số lượng HUI không nhạy cảm. Do đó, có một mối quan hệ nghịch đảo giữa MC và IUS, trong đó việc tăng cái này sẽ làm giảm cái kia. Như thể hiện trong hình. Như được hiển thị trong Hình 7 và 8, thuật toán SDIF và SLRF được đề xuất có hiệu suất vượt trội so với thuật toán PPUM cơ bản và mới nhất. Đây là điều được mong đợi vì cả hai thuật toán SDIF và SLRF đều có chi phí còn thiếu thấp hơn so với các đối thủ cạnh tranh khác. Chúng ta cũng có thể quan sát thấy rằng các thuật toán MSU-MIU và Weighted hoạt động tốt hơn SMRF được đề xuất trong tập dữ liệu về nấm. Lý do là trong quá trình lựa chọn mục nạn nhân của SMRF, mục có giá trị RISU lớn nhất được chọn làm mục nạn nhân lý tưởng. Tuy nhiên, trong tập dữ liệu bán lẻ, thuật toán SMRF cho thấy hiệu suất lạc quan vì mức độ chồng chéo giữa các mục ít hơn trong các tập dữ liệu thưa thớt. Nói tóm lại, kết quả IUS xác nhận rằng các ý tưởng được áp dụng của chúng tôi có thể giảm thiểu một cách hiệu quả các tác động tiêu cực phát sinh từ quá trình vệ sinh.

### 6.5. Sự tương đồng về tiện ích của tập dữ liệu

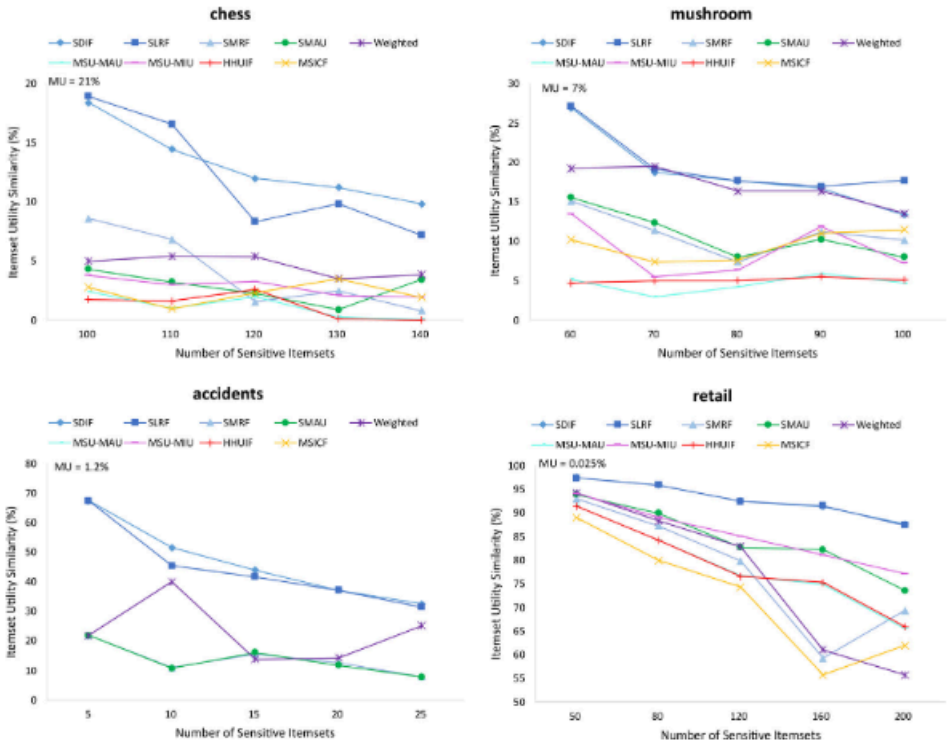
Phần này cân nhắc về ảnh hưởng tiêu cực của các thuật toán làm sạch được đề xuất đối với tổng tiện ích của tập dữ liệu. Quả sung. Hình 9 và 10 minh họa kết quả của thước đo DUS. Có thể thấy, thuật toán SLRF và SDIF được đề xuất mang lại hiệu suất vượt trội so với DUS trong số các thuật toán được so sánh. Tuy nhiên, có một ngoại lệ khi nói đến thuật toán MSU-MIU vì nó mang lại kết quả tốt hơn một chút so với SLRF và SDIF. Lý do tại sao MSU-MIU tiêu thụ ít tiện ích tập dữ liệu hơn là vì nó chọn mục có tiện ích tối thiểu làm mục nạn nhân và do đó, ít tiện ích hơn sẽ bị giảm đi trong tổng tiện ích của tập dữ liệu. Ngoài ra, thuật toán SLRF còn hoạt động tốt hơn thuật toán SDIF vì mục có giá trị RISU ít nhất được chọn trực tiếp để khử trùng. Ngược lại, hiệu suất của các thuật toán HHUIF, SMAU, MSU-MAU và SMRF luôn kém nhất. Những cái này



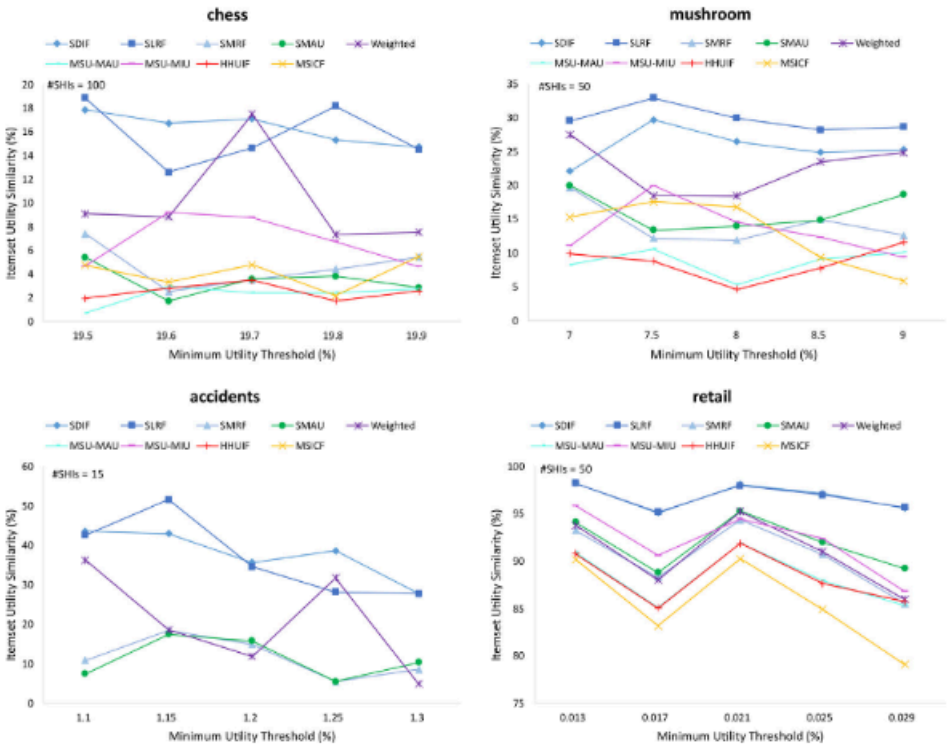

Hình 5. Thiếu chi phí khi sử dụng nhiều tập mục nhạy cảm khác nhau.



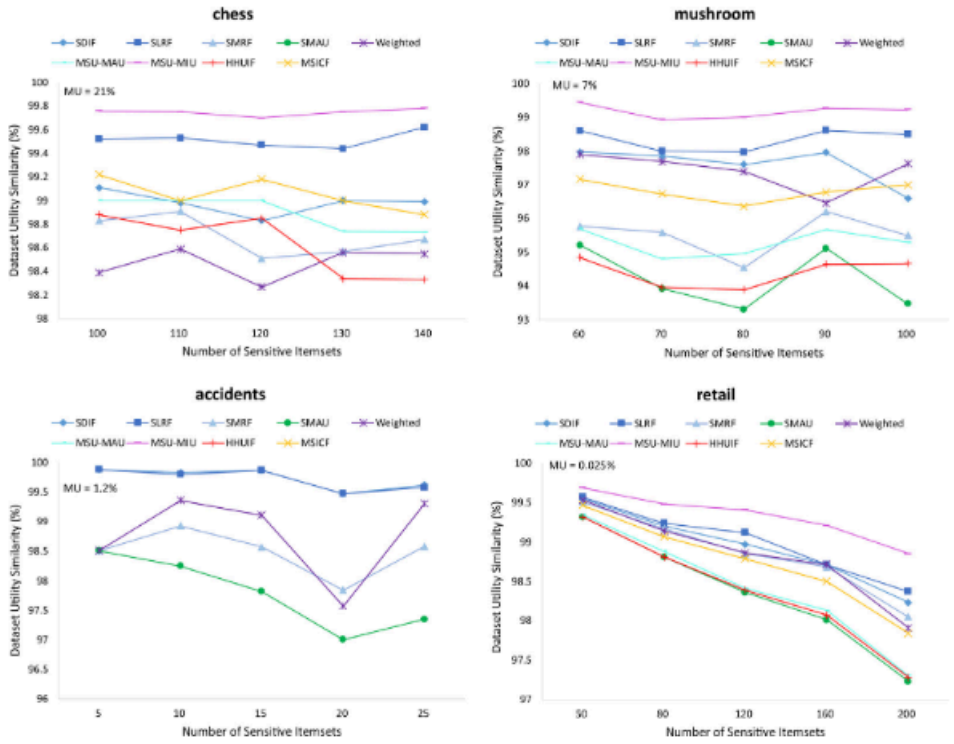

Hình 6. Thiếu chi phí khi sử dụng các ngưỡng tiện ích tối thiểu khác nhau.



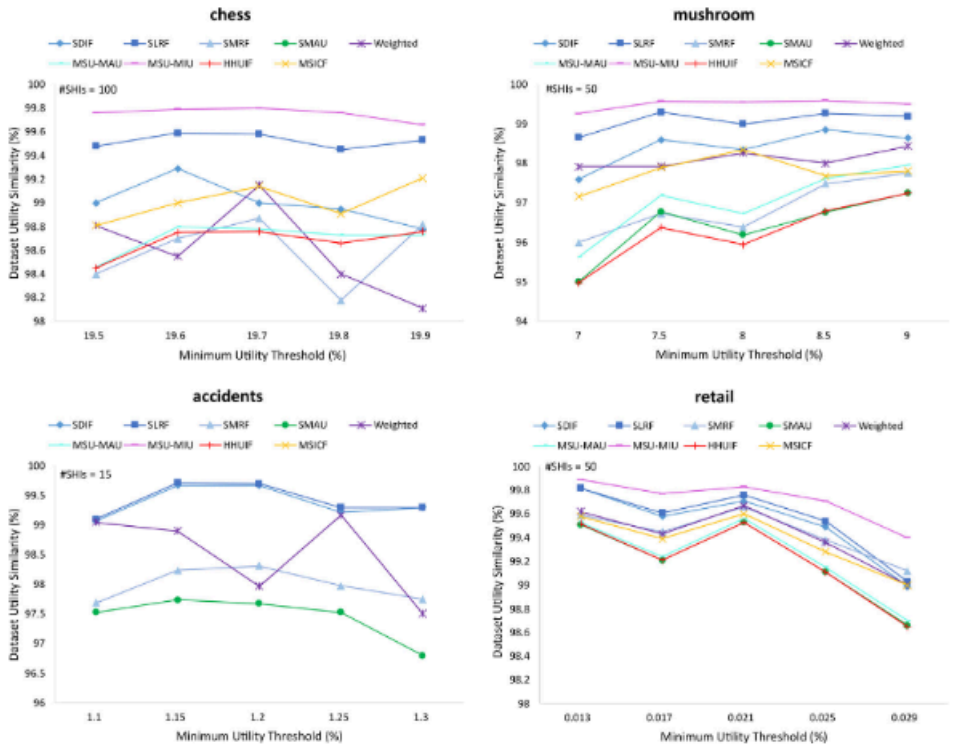

Hình 7. IUS sử dụng nhiều tập mục nhạy cảm khác nhau.



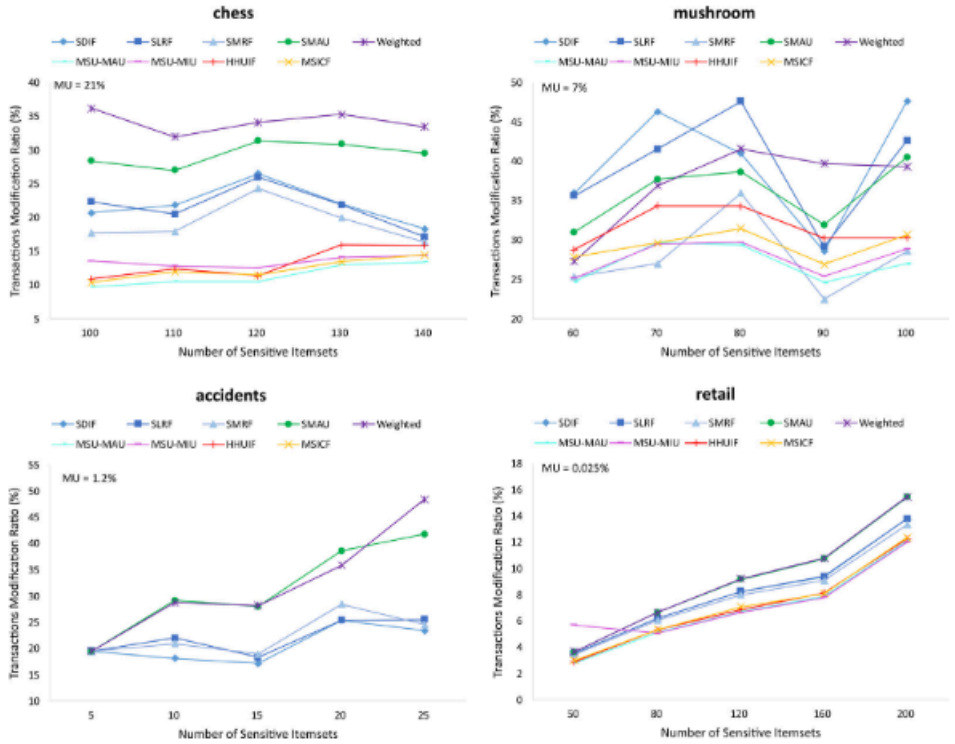

Hình 8. IUS sử dụng các ngưỡng tiện ích tối thiểu khác nhau.



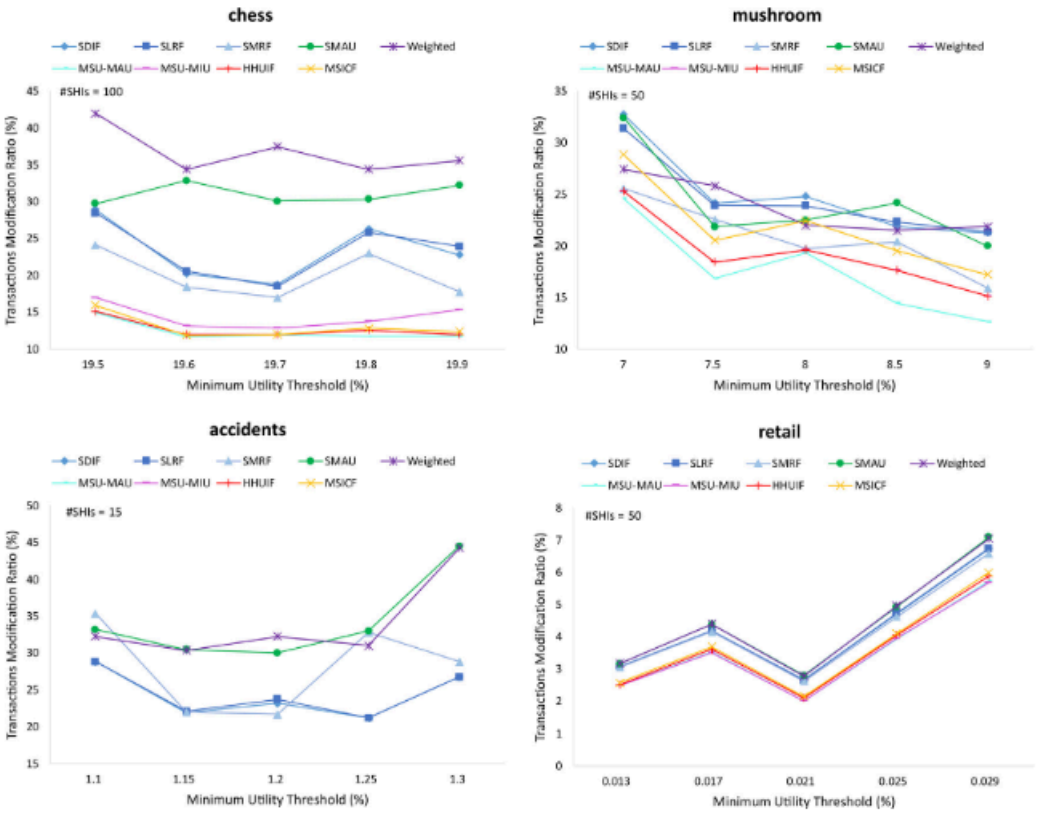

Hình 9. DUS sử dụng nhiều tập mục nhạy cảm khác nhau.




Hình 10. DUS sử dụng các ngưỡng tiện ích tối thiểu khác nhau.




Hình 11. TMR sử dụng nhiều tập mục nhạy cảm khác nhau.




Hình 12. TMR sử dụng các ngưỡng tiện ích tối thiểu khác nhau.

xu hướng hiệu suất là hợp lý, vì mục có tiện ích tối đa hoặc giá trị RISU tối đa luôn được chọn để dọn dẹp trong các thuật toán này. Tuy nhiên, đây không phải là trường hợp của thuật toán MSICF và Weighted vì thuật toán đầu tiên xem xét mức độ chồng chéo của các mục khi chọn mục nạn nhân, trong khi thuật toán thứ hai áp dụng các tiêu chí độc lập về tiện ích để đánh giá các mục nạn nhân ứng cử viên. Nói một cách dễ hiểu, mặc dù thuật toán SDIF không thu được các giá trị tốt nhất của DUS, nhưng nó có xu hướng hoạt động tốt hơn các thuật toán được đề xuất khác, SMRF và SLRF, trong MC và IUS. Điều này là do thuật toán SDIF, không giống như thuật toán SLRF, ưu tiên các mục có vỏ bọc ít nhạy cảm nhất hơn các mục có giá trị RISU ít nhất. Ngoài ra, chúng ta có thể biết rằng thuật toán SMRF có xu hướng tạo ra kết quả tồi tệ nhất khi nói đến các số liệu MC, IUS và DUS. Nguyên nhân là do nó loại bỏ mục có giá trị RISU tối đa dẫn đến ít DUS và IUS hơn và nhiều MC hơn.

### 6.6. Tỷ lệ sửa đổi giao dịch

Cuối cùng nhưng không kém phần quan trọng, thật thú vị khi xem các ý tưởng được áp dụng ảnh hưởng như thế nào đến số lượng giao dịch cần sửa đổi trong quá trình bóp méo dữ liệu. Quả sung. Hình 11 và 12 trình bày tỷ lệ phần trăm các giao dịch được sửa đổi do loại bỏ các tập mục nhạy cảm. Những kết quả này tiết lộ một vài hiểu biết thú vị. Đầu tiên, thuật toán SMRF được đề xuất có xu hướng có tỷ lệ sửa đổi ít hơn so với hai thuật toán được đề xuất còn lại. Lý do là trong các thuật toán SDIF và SLRF, mục có tổng tiện ích nhạy cảm hoặc giá trị RISU ít nhất được chọn để dọn dẹp, do đó cần phải sửa đổi nhiều giao dịch hơn để ẩn tập mục nhạy cảm. Thứ hai, TMR của tất cả các thuật toán được so sánh ở các tập dữ liệu lớn và dày đặc tương đối cao hơn so với tập dữ liệu thưa thớt (bán lẻ). Điều này là do trong các tập dữ liệu thưa thớt, tần suất của các tập mục thường thấp, do đó việc loại bỏ mục nạn nhân khỏi một số giao dịch có thể làm giảm đáng kể tiện ích của các tập mục nhạy cảm. Thứ ba, các thuật toán cơ sở luôn yêu cầu số lần sửa đổi ít hơn so với các thuật toán đề xuất. Điều này chủ yếu là do các thuật toán được đề xuất xử lý tầm quan trọng của các giao dịch một cách khác nhau vì các giao dịch nhạy cảm và không nhạy cảm được tận dụng để chọn ra những sửa đổi tốt nhất. Ngược lại, các thuật toán cơ sở tuân theo các tiêu chí dựa trên tiện ích trong phương pháp lựa chọn giao dịch của chúng. Rõ ràng, các thuật toán được đề xuất thường yêu cầu TMR ít hơn các thuật toán hiện đại (SMAU và Weighted). Kết quả này đạt được nhờ sử dụng hiệu quả các giá trị RISU và kỹ thuật Sắp xếp theo Trọng số. Nhìn chung, các kết quả trước đó cho thấy phương pháp lựa chọn giao dịch cùng với mật độ của tập dữ liệu đóng vai trò chính trong việc xác định tỷ lệ sửa đổi giao dịch.

## 7. Kết luận

Mặc dù việc xác định các mô hình tiện ích cao sẽ mang lại những hiểu biết có giá trị, nhưng những phát hiện trước đây đều đồng ý rằng việc khai thác mô hình theo hướng tiện ích có thể là vũ khí hai lưỡi. Mặc dù một mặt nó có thể được sử dụng để khám phá kiến thức rộng lớn và phong phú, mặt khác nó có thể tiết lộ thông tin bí mật về chủ sở hữu dữ liệu của nhiều bên. Nếu không có biện pháp bảo vệ quyền riêng tư đầy đủ, nhiều nhà cung cấp dữ liệu có thể thận trọng trong việc chia sẻ dữ liệu của họ. Do đó, thuật toán Khai thác tiện ích bảo vệ quyền riêng tư (PPUM) được sử dụng để bảo vệ thông tin riêng tư khỏi các kỹ thuật hướng đến tiện ích. Trong bài báo này, chúng tôi đã phát triển ba thuật toán PPUM, đó là SMRF, SLRF và SDIF, để loại bỏ hiệu quả các tập mục hữu ích nhạy cảm cao dựa trên khái niệm Tiện ích nhạy cảm vật phẩm thực (RISU). Ba thuật toán sử dụng kỹ thuật sắp xếp cho các giao dịch nhạy cảm, được đặt tên là Sắp xếp theo trọng số, để ưu tiên cao hơn cho việc dọn dẹp cho các giao dịch có ít tác dụng phụ hơn. Khái niệm RISU được áp dụng trong quá trình lựa chọn mục nạn nhân để tìm ra mục nạn nhân lý tưởng cho từng tập mục có tính tiện ích cao nhạy cảm trong giai đoạn đầu của quá trình.

quá trình vệ sinh. Để xác thực tính khả thi của các thuật toán được đề xuất, các so sánh mở rộng đã được tiến hành trên bốn bộ dữ liệu chuẩn sử dụng năm thước đo hiệu suất. Các kết quả thu được chỉ ra rằng các thuật toán đề xuất cải tiến các thuật toán tiên tiến nhất trong việc duy trì chất lượng của tập dữ liệu sau quá trình làm sạch. Đối với công việc trong tương lai, chúng tôi sẽ tập trung vào cách hỗ trợ thêm cho việc phát hiện các mục nạn nhân lý tưởng trong các bộ mục nhạy cảm. Ngoài ra, ứng dụng thực tế của PPUM trong các tình huống thực tế cũng rất đáng được quan tâm. Cuối cùng, việc thiết kế một mô hình song song cho nhiệm vụ PPUM trong bộ dữ liệu giao dịch cũng rất thú vị và đầy thách thức.

## Tuyên bố về lợi ích cạnh tranh

Các tác giả tuyên bố rằng họ không có lợi ích tài chính hoặc mối quan hệ cá nhân cạnh tranh nào có thể ảnh hưởng đến công việc được báo cáo trong bài viết này.

## Tuyên bố đóng góp quyền tác giả CRediT

Mohamed Ashraf: Khái niệm hóa, Phương pháp luận, Phần mềm, Trực quan hóa, Điều tra, Viết – bản thảo gốc, Viết – xem xét & chỉnh sửa. Sherine Rady: Xác nhận, Viết – đánh giá và chỉnh sửa. Tamer Abdelkader: Xác nhận, Viết – đánh giá và chỉnh sửa. Tarek F. Gharib: Xác nhận, Viết – đánh giá & chỉnh sửa, Giám sát.

## Tính khả dụng của dữ liệu

Dữ liệu sẽ được cung cấp theo yêu cầu.

## Tài liệu tham khảo


Agrawal, R., Srikant, R., và cộng sự, 1994. Thuật toán nhanh để khai thác các quy tắc kết hợp. Trong: Kỷ yếu của Hội nghị quốc tế lần thứ 20 về cơ sở dữ liệu rất lớn, Tập. 1215. Citeseer, trang 487–499.

Ali, MA, Rady, S., Abdelkader, T., Gharib, TF, 2023. Một phương pháp ẩn hiệu quả để bảo vệ quyền riêng tư khi khai thác tiện ích. Int. J. Trí tuệ. Máy tính.Inf. Khoa học. 23 (1), 69–83. Ashraf, M., Abdelkader, T., Rady, S., Gharib, TF, 2021. TKN: một cách tiếp cận hiệu quả cho

khám phá các tập mục tiện ích cao top-k với lợi nhuận dương hoặc âm. Thông tin Khoa học. (Ny).

Ashraf, M., Rady, S., Abdelkader, T., Gharib, TF, 2022. Một phương pháp bảo vệ quyền riêng tư mạnh mẽ để vệ sinh cơ sở dữ liệu giao dịch khỏi các mẫu tiện ích cao nhạy cảm. Trong: Kỷ yếu của Hội nghị quốc tế lần thứ 8 về Hệ thống thông minh và tin học tiên tiến năm 2022. Springer, trang 381–394.

Bandil, L., Soni, R., Rathi, S., 2018. Một phương pháp mới để bảo vệ quyền riêng tư của các bộ vật phẩm tiện ích bằng cách sử dụng quyền riêng tư khác biệt. Trong: Kỷ yếu hội nghị quốc tế về tiến bộ gần đây về máy tính và truyền thông. Springer, trang 481–487.

Chen, J., Guo, X., Gan, W., Chen, C.-M., Ding, W., Chen, G., 2022. Khai thác tiện ích sẵn có từ cơ sở dữ liệu giao dịch. Anh. ứng dụng. Nghệ thuật. Trí tuệ. 107, 104516.

Dinh, D.-T., Huynh, V.-N., Le, B., Fournier-Viger, P., Huynh, U., Nguyen, Q.-M., 2019. Khảo sát về khai thác tiện ích bảo vệ quyền riêng tư . Trong: Khai thác mẫu tiện ích cao: Lý thuyết, thuật toán và ứng dụng, trang 207–232.

Fournier-Viger, P., Gan, W., Wu, Y., Nouioua, M., Song, W., Trường, T., Dương, H., 2022. Khai thác mô hình: những thách thức và cơ hội hiện tại. Trong: Hệ thống cơ sở dữ liệu cho các ứng dụng nâng cao. Hội thảo quốc tế DASFAA 2022: BDMS, BDQM, GDMA, IWBT, MAQTDS và PMBD, Sự kiện ảo, ngày 11–14 tháng 4 năm 2022, Kỷ yếu. Springer, trang 34–49.

Fournier-Viger, P., Gomariz, A., Gueniche, T., Soltani, A., Wu, C.-W., Tseng, VS, et al., 2014. SPMF: thư viện khai thác mẫu mã nguồn mở java . J. Mach. Học hỏi. Res. 15 (1), 3389–3393.

Fournier-Viger, P., Li, J., Lin, JC-W., Chi, TT, Kiran, RU, 2020. Khai thác hiệu quả về chi phí

các mẫu trong nhật ký sự kiện. Kiến thức. Hệ thống dựa trên. 191, 105241. Gan, W., Chun-Wei, J., Chao, H.-C., Wang, S.-L., Philip, SY, 2018. Bảo vệ quyền riêng tư

khai thác tiện ích: một cuộc khảo sát. Tại: Hội nghị quốc tế IEEE 2018 về dữ liệu lớn (Big Data)

dữ liệu). IEEE, trang 2617–2626. Gan, W., Lin, JC-W., Fournier-Viger, P., Chao, H.-C., Tseng, VS, Philip, SY, 2019. A

khảo sát khai thác mô hình theo định hướng tiện ích. IEEE Trans. Kiến thức. Dữ liệu kỹ thuật số 33 (4), 1306–1327.

Grätzer, G., 2011. Lý thuyết mạng: Nền tảng. Truyền thông Khoa học & Kinh doanh Springer. Holland, JH, 1992. Thích ứng trong các hệ thống tự nhiên và nhân tạo: Giới thiệu

Phân tích với các ứng dụng về Sinh học, Điều khiển và Trí tuệ nhân tạo. Báo chí MIT.

Huynh, U., Le, B., Dinh, D.-T., Fujita, H., 2022. Thuật toán song song đa lõi để ẩn các mẫu tuần tự có tính tiện ích cao. Kiến thức. Hệ thống dựa trên. 237, 107793.

Jangra, S., Toshniwal, D., 2022. Các thuật toán hiệu quả để lựa chọn mục nạn nhân trong khai thác tiện ích bảo vệ quyền riêng tư. Thế hệ tương lai. Máy tính. Hệ thống. 128, 219–234. doi:10.1016/j.future.2021.10.008.

Jisna, J., Salim, A., 2018. Khai thác tiện ích dữ liệu bảo đảm quyền riêng tư bằng cách sử dụng nhiễu loạn. Trong: Hội nghị quốc tế về máy tính phân tán và công nghệ Internet. Springer, trang 112–120.

Kenthapadi, K., Mironov, I., Thakurta, AG, 2019. Khai thác dữ liệu bảo vệ quyền riêng tư trong ngành. Trong: Kỷ yếu của Hội nghị quốc tế ACM lần thứ 12 về Tìm kiếm trên web và Khai thác dữ liệu, trang 840–841.

Krishna, GJ, Ravi, V., 2021. Khai thác tập mục có tiện ích cao bằng cách sử dụng tiến hóa vi sai nhị phân: một ứng dụng để phân khúc khách hàng. Hệ thống chuyên gia ứng dụng. 181, 115122.

Le, B., Dinh, D.-T., Huynh, V.-N., Nguyen, Q.-M., Fournier-Viger, P., 2018. Một thuật toán hiệu quả để ẩn các mẫu tuần tự có tính tiện ích cao. Int. J. Lý luận gần đúng 95, 77–92. doi:10.1016/j.ijar.2018.01.005.

Li, S., Mu, N., Le, J., Liao, X., 2019. Một thuật toán mới để bảo vệ quyền riêng tư khi khai thác tiện ích dựa trên lập trình tuyến tính số nguyên. Anh. ứng dụng. Nghệ thuật. Trí tuệ. 81, 300–312. doi:10.1016/j.engappai.2018.12.006.

Lin, C.-W., Hong, T.-P., Wong, J.-W., Lan, G.-C., Lin, W.-Y., 2014. Cách tiếp cận dựa trên GA để che giấu thông tin nhạy cảm các tập mục có tính tiện ích cao. Khoa học. Thế giới J. 2014.

Lin, JC-W., Djenouri, Y., Srivastava, G., Fournier-Viger, P., 2022. Mô hình tính toán tiến hóa hiệu quả của việc khai thác tập mục tiện ích cao khép kín. ứng dụng. Trí tuệ. 1–13. Lin, JC-W., Hong, T.-P., Fournier-Viger, P., Liu, Q., Wong, J.-W., Zhan, J., 2017. Hiệu quả

che giấu một cách khoa học các tập vật phẩm bí mật có tính tiện ích cao với tác dụng phụ tối thiểu. J.

Exp. Lý thuyết. Artif.Intell. 29 (6), 1225–1245. Lin, JC-W., Wu, T.-Y., Fournier-Viger, P., Lin, G., Zhan, J., Voznak, M., 2016. Fast al-

các thuật toán để ẩn các tập mục hữu ích cao nhạy cảm trong tiện ích bảo vệ quyền riêng tư

khai thác mỏ. Anh. ứng dụng. Nghệ thuật. Trí tuệ. 55, 269–284. doi:10.1016/j.engappai.2016.07.003. Liu, C., Guo, C., 2021. Khai thác các mô hình hoạt động tiện ích cao hàng đầu dành cho tài xế taxi.

Expert Syst Appl 170, 114546. Liu, X., Chen, G., Wen, S., Song, G., 2020. Một thuật toán khử trùng được cải tiến trong

khai thác tiện ích bảo vệ quyền riêng tư. Toán học. Vấn đề. Anh. 2020. Liu, X., Wen, S., Zuo, W., 2020. Các phương pháp vệ sinh hiệu quả để bảo vệ các khu vực nhạy cảm

kiến thức về khai thác tập mục có tính tiện ích cao. ứng dụng. Trí tuệ. 50 (1), 169–191. Liu, X., Xu, F., Lv, X., 2018. Một cách tiếp cận mới để che giấu tiện ích nhạy cảm và thường xuyên

itemset. Trí tuệ. Dữ liệu hậu môn. 22 (6), 1259–1278. Liu, Y., Liao, W.-k., Choudhary, A., 2005. Một thuật toán khai thác tập hợp tiện ích cao nhanh chóng-

ritm. Trong: Kỷ yếu Hội thảo quốc tế lần thứ nhất về Khai thác dữ liệu dựa trên tiện ích, trang 90–99.

Mendes, R., Vilela, JP, 2017. Khai thác dữ liệu bảo vệ quyền riêng tư: phương pháp, số liệu và ứng dụng. Truy cập IEEE 5, 10562–10582.

Qu, J.-F., Fournier-Viger, P., Liu, M., Hang, B., Wang, F., 2020. Khai thác các tập mục tiện ích cao bằng cách sử dụng cấu trúc chuỗi mở rộng và máy tiện ích. Kiến thức. Hệ thống dựa trên. 208, 106457.

Rajalaxmi, R., Natarajan, A., 2012. Các phương pháp khử trùng hiệu quả để che giấu tiện ích nhạy cảm và các tập mục thường xuyên. Trí tuệ. Dữ liệu hậu môn. 16 (6), 933–951.

Segura-Delgado, A., Anguita-Ruiz, A., Alcalá, R., Alcalá-Fdez, J., 2022. Khai thác các quy tắc tuần tự có tính tiện ích trung bình cao để xác định trình tự biểu hiện gen có tính hữu ích cao trong các nghiên cứu theo chiều dọc của con người. Hệ thống chuyên gia ứng dụng. 116411.

Selvaraj, R., Kuthadi, VM, 2013. Thuật toán ẩn mục đầu tiên có tiện ích cao được sửa đổi (HHUIF) với bộ chọn mục (MHIS) để ẩn các tập mục nhạy cảm. J. Đổi mới. Com-put. Inf.Control 9 (12), 4851–4862.

Tran, H.-Y., Hu, J., 2019. Một cuộc khảo sát toàn diện về phân tích dữ liệu lớn bảo vệ quyền riêng tư. J. Phân phối song song. Máy tính. 134, 207–218.

Triệu, VH, Ngọc, CT, Lê Quốc, H., Si, NN, 2018. Thuật toán ẩn luật kết hợp có độ nhạy cao dựa trên mạng giao nhau. Tại: Hội nghị quốc tế lần thứ nhất năm 2018 về Phân tích đa phương tiện và nhận dạng mẫu (MAPR). IEEE, trang 1–6.

Tseng, VS, Shie, B.-E., Wu, C.-W., Philip, SY, 2012. Các thuật toán hiệu quả để khai thác các tập mục có tiện ích cao từ cơ sở dữ liệu giao dịch. IEEE Trans. Kiến thức. Dữ liệu kỹ thuật số 25 (8), 1772–1786.

Tung, N., Nguyen, LT, Nguyen, TD, Võ, B., 2021. Một phương pháp hiệu quả để khai thác các tập mục tiện ích cao đa cấp. ứng dụng. Trí tuệ. 1–22.

Verma, A., Dawar, S., Kumar, R., Navathe, S., Goyal, V., 2021. Khai thác tập hợp vật phẩm đa dạng và tiện ích cao. ứng dụng. Trí tuệ. 1–15.

Yeh, J.-S., Hsu, P.-C., 2010. HHUIF và MSICF: các thuật toán mới để khai thác tiện ích bảo đảm quyền riêng tư. Hệ thống chuyên gia ứng dụng. 37 (7), 4779–4786. doi:10.1016/j.eswa.2009.12.038.

Yun, U., Kim, J., 2015. Thuật toán nhiễu loạn nhanh sử dụng cấu trúc cây để khai thác tiện ích bảo đảm quyền riêng tư. Hệ thống chuyên gia ứng dụng. 42 (3), 1149–1165. doi:10.1016/j.eswa.2014.08.037.

Zhang, C., Han, M., Sun, R., Du, S., Shen, M., 2020. Khảo sát về các công nghệ chính để khai thác các mô hình tiện ích cao. Truy cập IEEE 8, 55798–55814.

Zida, S., Fournier-Viger, P., Lin, JC-W., Wu, C.-W., Tseng, VS, 2015. EFIM: một thuật toán hiệu quả cao để khai thác tập mục có tính tiện ích cao. Trong: Hội nghị quốc tế Mexico về trí tuệ nhân tạo. Springer, trang 530–546.

Mohamed Ashraf nhận bằng Cử nhân và Thạc sĩ về khoa học máy tính và thông tin từ Khoa Khoa học Máy tính và Thông tin (FCIS), Đại học Ain Shams (ASU), Cairo, Ai Cập, lần lượt vào năm 2016 và 2022, nơi ông hiện đang làm việc. đang theo đuổi bằng tiến sĩ về khoa học máy tính và thông tin. Từ năm 2018 đến 2022, ông là Trợ giảng tại Khoa Hệ thống Thông tin, FCIS, ASU. Từ năm 2022, ông là Trợ lý Giảng viên. Mối quan tâm nghiên cứu của ông bao gồm khai thác dữ liệu, học máy, học sâu, công nghệ phần mềm và bảo vệ quyền riêng tư. Các công trình của ông đã được công bố trên các tạp chí quốc tế uy tín và chất lượng như Information Sciences. Email: mohamed.aahis@gmail.com

Sherine Rady nhận bằng Cử nhân kỹ thuật điện (máy tính và hệ thống) và bằng Thạc sĩ về khoa học máy tính và thông tin tại Đại học Ain Shams, Cairo, Ai Cập và bằng Tiến sĩ từ Đại học Mannheim, Đức. Cô hiện là Giáo sư tại Khoa Khoa học Thông tin và Máy tính, Đại học Ain Shams. Mối quan tâm nghiên cứu của cô bao gồm Trí tuệ nhân tạo, Khai thác dữ liệu và Khoa học dữ liệu. Email: srady@cis.asu.edu.eg

Tamer Abdelkader nhận bằng Cử nhân kỹ thuật điện và máy tính và bằng Thạc sĩ về khoa học máy tính và thông tin tại Đại học Ain Shams, Cairo, Ai Cập, năm 2003, đồng thời có bằng Thạc sĩ và Tiến sĩ về kỹ thuật điện và máy tính từ Đại học Ain Shams. Đại học Waterloo, Ontario, ON, Canada, vào năm 2012. Sau khi tốt nghiệp, anh làm việc tại Đại học Waterloo với tư cách là Nhà nghiên cứu sau tiến sĩ và là nhà nghiên cứu Nhà nghiên cứu đến thăm. Ông từng giữ chức vụ Giám đốc Trung tâm Tư vấn Nghiên cứu Công nghệ và Thông tin, Đại học Ain Shams, Cairo, Ai Cập. Ông cũng từng làm Chuyên gia tư vấn về Công nghệ và Thông tin tại một số công ty chính phủ và tư nhân, bao gồm cả Công ty Thông tin và Truyền thông. Dự án Công nghệ cation, Ai Cập và Bộ Điện lực. Ông hiện là Phó Giáo sư và Phó Trưởng khoa về Dịch vụ Cộng đồng và Các vấn đề Môi trường của Khoa Khoa học Thông tin và Máy tính, Đại học Ain Shams. Ông là tác giả của nhiều ấn phẩm trên tạp chí IEEE Transactions cũng như các tạp chí và hội nghị được xếp hạng khác. Mối quan tâm nghiên cứu hiện tại của ông bao gồm mạng và an ninh mạng, bảo vệ quyền riêng tư, mạng chịu được độ trễ, phân bổ tài nguyên trong mạng không dây và các giao thức tiết kiệm năng lượng. Email: tammabde@cis.asu.edu.eg

Tarek F. Gharib hiện là Giáo sư chính thức về Hệ thống Thông tin tại Đại học Ain Shams, Cairo, Ai Cập. Ông nhận bằng Tiến sĩ về Vật lý lý thuyết tại Đại học Ain Shams vào năm 1994. Mối quan tâm nghiên cứu của ông tập trung vào phát triển các kỹ thuật khai thác dữ liệu và học máy mới, đặc biệt cho các ứng dụng khai thác văn bản, mạng xã hội, Tin sinh học và Phân tích dữ liệu. Ông có hơn 90 ấn phẩm. Ông đã nhận được Giải thưởng của Quỹ Khoa học Quốc gia năm 2001. Email: tfgharib@cis.asu.due.eg