Ứng dụng kỹ thuật của trí tuệ nhân tạo 53 (2016) 1–18

Danh sách nội dung có sẵn tại ScienceDirect

Ứng dụng kỹ thuật của trí tuệ nhân tạo

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

Một cách tiếp cận dọn dẹp để ẩn các tập mục nhạy cảm dựa trên tối ưu hóa bầy đàn hạt

Jerry Chun-Wei Lin a,n, Qiankun Liu a, Philippe Fournier-Viger b, Tzung-Pei Hong c , Miroslav Voznak d, Justin Zhan e

a Trường Khoa học và Công nghệ Máy tính, Viện Công nghệ Cáp Nhĩ Tân, Trường Sau đại học Thâm Quyến, Thâm Quyến, Trung Quốc b Trường Khoa học Tự nhiên và Nhân văn, 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, Quốc gia Đại học Cao Hùng, Cao Hùng, Đài Loan d Khoa Kỹ thuật Điện và Khoa học Máy tính, Đại học Kỹ thuật VSB Ostrava, Ostrava-Poruba, Cộng hòa Séc e Khoa Khoa học Máy tính, Đại học Nevada, Las Vegas, Hoa Kỳ

thông tin bài viết

Lịch sử bài viết: Nhận ngày 8 tháng 7 năm 2015 Nhận ở dạng sửa đổi ngày 20 tháng 3 năm 2016

Được chấp nhận ngày 21 tháng 3 năm 2016

Từ khóa: Vệ sinh PPDM Tính toán tiến hóa Tập mục nhạy cảm PSO

trừu tượng

Khai thác dữ liệu bảo vệ quyền riêng tư (PPDM) đã trở thành một lĩnh vực nghiên cứu quan trọng trong những năm gần đây, vì các phương pháp tiếp cận PPDM có thể khám phá thông tin quan trọng trong cơ sở dữ liệu, đồng thời đảm bảo rằng thông tin nhạy cảm không bị tiết lộ. Một số thuật toán đã được đề xuất để ẩn thông tin nhạy cảm trong cơ sở dữ liệu. Chúng áp dụng các phép toán cộng và xóa để làm xáo trộn cơ sở dữ liệu gốc và che giấu thông tin nhạy cảm. Việc tìm kiếm một tập hợp các giao dịch/tập mục thích hợp để che giấu thông tin nhạy cảm trong khi vẫn bảo toàn các thông tin quan trọng khác là một bài toán NP-khó. Trước đây, các phương pháp tiếp cận dựa trên thuật toán di truyền (GA) đã được phát triển để ẩn các tập mục nhạy cảm trong cơ sở dữ liệu gốc thông qua việc xóa giao dịch. Trong bài báo này, thuật toán dựa trên tối ưu hóa bầy đàn (PSO) có tên PSO2DT được phát triển để ẩn các tập mục nhạy cảm đồng thời giảm thiểu tác dụng phụ của quá trình khử trùng. Mỗi phần trong thuật toán PSO2DT được thiết kế đại diện cho một tập hợp các giao dịch sẽ bị xóa. Các hạt được đánh giá bằng cách sử dụng chức năng thích hợp được thiết kế để giảm thiểu tác dụng phụ của quá trình khử trùng. Thuật toán được đề xuất cũng có thể xác định số lượng giao dịch tối đa cần xóa để ẩn các tập mục nhạy cảm một cách hiệu quả, không giống như các phương pháp dựa trên GA tiên tiến nhất. Ngoài ra, một điểm mạnh quan trọng của phương pháp đề xuất là cần đặt ít tham số và vẫn có thể tìm ra giải pháp tốt hơn cho vấn đề vệ sinh so với các phương pháp dựa trên GA. Hơn nữa, khái niệm tiền lớn cũng được áp dụng trong thuật toán được thiết kế để tăng tốc quá trình tiến hóa. Các thử nghiệm thực tế trên cả bộ dữ liệu thực tế và tổng hợp cho thấy thuật toán PSO2DT đề xuất hoạt động tốt hơn thuật toán Greedy và thuật toán dựa trên GA về thời gian chạy, không bị ẩn (FTH), không bị ẩn (NTH).) và độ tương tự cơ sở dữ liệu (DS).

1. Giới thiệu

Với sự phát triển nhanh chóng của công nghệ thông tin và các ứng dụng thương mại điện tử, việc khám phá những thông tin hữu ích và các mối quan hệ thú vị trong lượng dữ liệu khổng lồ ngày càng trở nên dễ dàng. Khai thác dữ liệu, còn được gọi là khám phá tri thức trong cơ sở dữ liệu (KDD), cung cấp một tập hợp các kỹ thuật, thường được sử dụng để phân tích mối quan hệ giữa các sản phẩm được mua để phân tích giỏ hàng thị trường. Tri thức được khám phá bằng kỹ thuật KDD có thể được phân loại tổng quát thành các luật kết hợp (Agrawal và Srikant, 1994b; Chen và cộng sự, 1996; Han và cộng sự, 2004), các mẫu tuần tự (Agrawal và Srikant, 1995; Mooney và Roddick, 2013; Zaki, 2001), các cụm (Murty và Flynn, 1999) và phân loại (Quinlan, 1993). Khai thác quy tắc kết hợp (Agrawal và Srikant, 1994b; Chen và cộng sự, 1996) là một nhiệm vụ KDD cơ bản, bao gồm việc khám phá thông tin và kiến thức thú vị trong các giao dịch của khách hàng.

Vì kỹ thuật khai thác dữ liệu có thể được sử dụng để khám phá thông tin tiềm ẩn trong cơ sở dữ liệu rất lớn nên thông tin riêng tư hoặc bảo mật cũng có thể dễ dàng bị tiết lộ bởi các kỹ thuật đó, chẳng hạn như số thẻ tín dụng, số nhận dạng cá nhân, số điện thoại và dữ liệu bí mật khác. Ngoài ra, một vấn đề quan trọng khác là thông tin được chia sẻ giữa các cộng tác viên kinh doanh cũng có thể được phân tích bằng kỹ thuật khai thác dữ liệu để tiết lộ những kiến thức nhạy cảm sau đó có thể bị rò rỉ cho đối thủ cạnh tranh. Một rủi ro khác là cộng tác viên hiện tại có thể trở thành đối thủ cạnh tranh và đối thủ cạnh tranh này có thể sử dụng kiến thức chiến lược thu được bằng kỹ thuật khai thác dữ liệu để đưa ra quyết định kinh doanh tốt hơn và do đó làm giảm

n Tác giả tương ứng. Địa chỉ email: jerrylin@ieee.org (J.-W. Lin), qiankunliu@ikelab.net (Q. Liu), philfv@hitsz.edu.cn (P. Fournier-Viger), tphong@nuk.edu.tw (T.-P. Hong), miroslav.voznak@vsb.cz (M. Voznak), justin.zhan@unlv. edu (J. Zhan).

http://dx.doi.org/10.1016/j.engappai.2016.03.007 0952-1976/

hiệu quả kinh doanh của nhà cung cấp dữ liệu do tăng cuộc thi. Vì những vấn đề này, việc khai thác dữ liệu đảm bảo quyền riêng tư (PPDM) đã trở thành một vấn đề quan trọng trong những năm gần đây (Ag-garwal và cộng sự, 2006; Dasseni và cộng sự, 2001; Evfimievski và cộng sự, 2002; Lindell và Pinkas, 2000 ; Verykios và cộng sự, 2004). Mục tiêu của PPDM là vệ sinh cơ sở dữ liệu, tức là ẩn và bảo mật thông tin cá nhân, bí mật hoặc nhạy cảm của người tham gia, trong khi vẫn cho phép phân tích dữ liệu. Cách phổ biến nhất để ẩn thông tin nhạy cảm trong cơ sở dữ liệu được thu thập là làm sạch cơ sở dữ liệu thông qua các thao tác xóa hoặc bổ sung. Tuy nhiên, cách tiếp cận này có thể gây ra một số tác dụng phụ như che giấu các mẫu không nhạy cảm hoặc đưa ra các mẫu nhân tạo mới. Đây là một vấn đề tối ưu hóa NP-hard (Ag-garwal và cộng sự, 2006; Verykios và cộng sự, 2004) để chọn một tập hợp hoạt động dọn dẹp thích hợp nhằm che giấu thông tin bí mật, đồng thời giảm thiểu tác dụng phụ.

Agrawal và Srikant lần đầu tiên giới thiệu PPDM (Aggarwal và cộng sự, 2006). Lindell và Pinkas (2000) đã đề cập đến vấn đề PPDM cho việc học cây quyết định bằng thuật toán ID3. Clifton và cộng sự. (2003) đã trình bày một bộ công cụ để giải quyết các vấn đề khác nhau trong PPDM. Để đạt được sự cân bằng tốt giữa quyền riêng tư dữ liệu và tiện ích dữ liệu trong PPDM, Pandya et al. (2014) đã đề xuất một thuật toán nhiễu loạn nhân. Dwork và cộng sự. (2006) đã khái quát hóa công việc trước đó bằng cách xem xét cả cơ sở dữ liệu đa thuộc tính và cơ sở dữ liệu được phân vùng theo chiều dọc, đồng thời thiết kế một số thuật toán để xử lý các số liệu thống kê nhiễu được công bố. Một số thuật toán cũng được thiết kế để ẩn các tập phổ biến nhạy cảm hoặc các quy tắc kết hợp nhạy cảm bằng cách sử dụng quy trình dọn dẹp tùy chỉnh (Evfimievski và cộng sự, 2002; Hong và cộng sự, 2012; Lin và cộng sự, 2013; Wu và cộng sự, 2007).

Các thuật toán PPDM truyền thống gặp khó khăn trong việc đối phó với thách thức tìm kiếm một tập hợp các giao dịch/mục mục thích hợp để dọn dẹp nhằm giảm thiểu tác dụng phụ, đặc biệt khi thông tin nhạy cảm trùng lặp với thông tin quan trọng nhưng không nhạy cảm. Việc ẩn và bảo mật thông tin nhạy cảm có thể đồng thời che giấu thông tin quan trọng. Điện toán tiến hóa là một cách hiệu quả để tìm ra giải pháp gần tối ưu cho các vấn đề NP-khó. Thuật toán di truyền (GA) (Goldberg, 1989; Holland, 1992) là một cách tiếp cận dựa trên dân số tạo điều kiện thuận lợi cho việc tìm kiếm các giải pháp tốt bằng cách áp dụng các nguyên tắc tiến hóa tự nhiên. Nó đã được áp dụng rộng rãi để xử lý các bài toán có cả biến rời rạc và biến liên tục, mục tiêu phi tuyến và hàm ràng buộc không có thông tin gradient. Trước đây, Lin và cộng sự. (2014, 2015a) đã đề xuất một thuật toán dựa trên GA để ẩn các tập mục nhạy cảm bằng cách sử dụng quy trình dọn dẹp được thiết kế. Theo cách tiếp cận này, việc chọn một tập hợp các giao dịch để xóa được thực hiện bằng cách sử dụng khung GA. Người ta đã chứng minh rằng các phương pháp tiếp cận dựa trên GA có thể cung cấp giải pháp tốt hơn cho các vấn đề PPDM với tác dụng phụ thấp hơn so với các thuật toán Greedy truyền thống. Tuy nhiên, các thuật toán đó vẫn yêu cầu thiết lập thủ công số lượng giao dịch cần xóa. Ngoài ra, việc tìm các giá trị (tỷ lệ) thích hợp cho các tham số được sử dụng bởi GA chẳng hạn như tỷ lệ đột biến và tỷ lệ chéo là một nhiệm vụ không hề đơn giản.

Tối ưu hóa bầy hạt (PSO) được phát minh bởi Kennedy và Eberhart (1995). Nó được lấy cảm hứng từ những đàn chim bay đi tìm nguồn thức ăn phong phú. Là GA, PSO là một phương pháp tìm kiếm dựa trên tổng thể, được thiết kế để giải quyết các vấn đề tối ưu hóa. Trong PSO, mỗi hạt đại diện cho một giải pháp và được đánh giá bằng hàm thích ứng được xác định trước. Các hạt tốt nhất cá nhân (pbest) và tốt nhất toàn cầu (gbest) được sử dụng để cập nhật các hạt cũ và tạo ra con cái của quần thể trong quá trình tiến hóa. Vì các phép toán chéo và đột biến trong GA không được sử dụng trong PSO nên việc thực hiện thủ tục PSO để khám phá các giải pháp gần tối ưu sẽ dễ dàng hơn. Ngoài ra, các hạt trong PSO có thể truyền thông tin đến các hạt khác để đẩy nhanh quá trình tiến hóa. Trong bài báo này, thuật toán PSO2DT dựa trên PSO được trình bày để tìm ra các tập hợp giao dịch tốt hơn cần xóa để che giấu thông tin nhạy cảm. Những đóng góp chính của thuật toán được thiết kế được liệt kê dưới đây.

1. Trước đây, một số phương pháp heuristic đã được đề xuất để làm sạch cơ sở dữ liệu nhằm che giấu thông tin nhạy cảm. Hầu hết trong số họ sử dụng khung GA. Đây là bài báo đầu tiên đề cập đến vấn đề ẩn các tập mục nhạy cảm bằng cách sử dụng phương pháp tiếp cận dựa trên PSO.

2. Thuật toán PSO2DT được thiết kế lấy cảm hứng từ PSO rời rạc. Nó gán các hạt và vận tốc của chúng cho tập hợp các mã định danh giao dịch, thể hiện các giao dịch sẽ bị xóa để ẩn các tập mục nhạy cảm. Ưu điểm của thuật toán PSO2DT được thiết kế là nó có ít tham số so với các phương pháp trước đó và vẫn tìm kiếm các giải pháp gần như tối ưu cho vấn đề vệ sinh bằng cách sử dụng phương pháp tiến hóa ngẫu nhiên.

3. Khái niệm tiền lớn cũng được áp dụng trong thuật toán được thiết kế để tránh thực hiện quét cơ sở dữ liệu nhiều lần. Điều này tăng tốc đáng kể việc đánh giá các hạt trong quá trình tiến hóa.

Phần còn lại của bài viết này được tổ chức như sau. Công việc liên quan được xem xét trong Phần 2. Các bước sơ bộ và xác định vấn đề được đề cập trong Phần 3. Thuật toán vệ sinh PSO2DT được trình bày trong Phần 4. Một ví dụ minh họa thuật toán đề xuất được đưa ra trong Phần 5. Kết quả thử nghiệm được báo cáo trong Phần 6. Một kết luận được rút ra và công việc trong tương lai sẽ được thảo luận trong Phần 7.

2. Công việc liên quan

Phần này xem xét các công việc liên quan về GA, PSO và PPDM.

2.1. Thuật toán di truyền

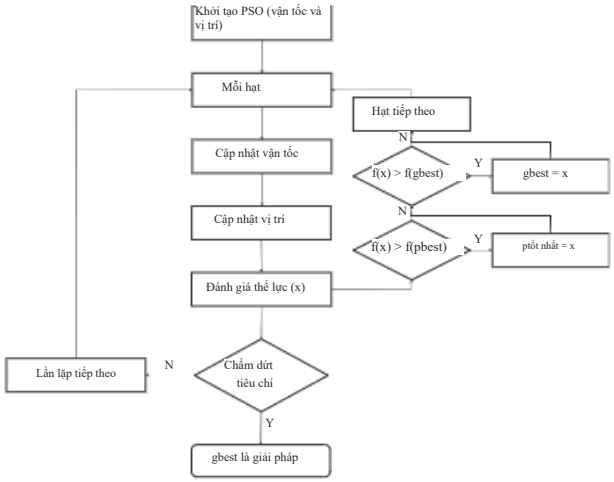
Trong điện toán tiến hóa, các phương pháp tiếp cận dựa trên dân số được sử dụng rộng rãi để tìm ra giải pháp gần tối ưu cho các vấn đề tối ưu hóa. Chúng đặc biệt được sử dụng cho các biến thể của các bài toán NP-khó và các ứng dụng liên quan, trong đó việc tìm ra giải pháp tốt nhất bằng cách đánh giá tất cả các giải pháp là quá tốn kém. GA là cách tiếp cận dựa trên dân số cơ bản nhất. Nó được phát triển vào đầu những năm 1970 bởi Holland (1992). Trong GA, một giải pháp được gọi là nhiễm sắc thể và nó có thể được đánh giá bằng hàm thích ứng được thiết kế. Ba thao tác có tên là chọn lọc, lai ghép và đột biến được GA sử dụng và được mô tả dưới đây.

1. Trao đổi chéo: Hoạt động này hoán đổi một số bit giữa hai nhiễm sắc thể (cá thể) để tạo ra con cái của quần thể. Con cái thừa hưởng các thuộc tính hoặc đặc điểm của hai nhiễm sắc thể bố mẹ của nó.

2. Đột biến: Thao tác này thay đổi ngẫu nhiên một hoặc một số bit của con cái, điều này có thể tạo ra các biến thể về đặc điểm bố mẹ của nó. Hoạt động này được sử dụng để tránh bị mắc kẹt trong các giải pháp tối ưu cục bộ và là điều cho phép quá trình tiến hóa tìm ra các giải pháp gần tối ưu.

3. Lựa chọn: Thao tác này áp dụng hàm thích nghi để chọn ra những con tốt nhất làm nhiễm sắc thể sống sót. Hoạt động này đảm bảo rằng những đặc điểm của con cái tốt nhất có khả năng được truyền lại cho thế hệ tiếp theo.

Các bước chính được thực hiện bởi GA như sau. Bước đầu tiên là xác định loại nhiễm sắc thể đại diện cho các giải pháp khả thi. Nhiễm sắc thể thường được biểu diễn dưới dạng chuỗi bit. Một quần thể ban đầu bao gồm nhiều nhiễm sắc thể, còn được gọi là các cá thể, được xác định và đại diện cho một tập hợp ban đầu các giải pháp khả thi. Các hoạt động lai ghép, đột biến và chọn lọc sau đó được áp dụng cho nhiễm sắc thể để tạo ra thế hệ tiếp theo. Mỗi nhiễm sắc thể được đánh giá bằng chức năng thích hợp được thiết kế để đánh giá mức độ tốt của nhiễm sắc thể. Quá trình này sau đó được lặp lại cho đến khi thỏa mãn tiêu chí kết thúc. Mặc dù GA



Hình 1. Sơ đồ của PSO truyền thống.

thường hoạt động tốt và có nhiều ứng dụng, một nhược điểm quan trọng của GA là việc thiết lập tỷ lệ chéo và đột biến thích hợp là một nhiệm vụ không hề đơn giản.

2.2. Tối ưu hóa bầy hạt

PSO cũng là một cách tiếp cận dựa trên dân số, được Kennedy và Eberhart (1995) giới thiệu. PSO được lấy cảm hứng từ hành vi của các loài chim bay đi tìm nguồn thức ăn tốt hơn. Trong PSO, các hạt được sử dụng để biểu diễn các giải pháp của bài toán, trong đó mỗi hạt có vận tốc biểu thị hướng bay về phía các giải pháp khác. Lưu đồ của PSO truyền thống được hiển thị trong Hình 1. Quy trình PSO trước tiên khởi tạo các hạt một cách ngẫu nhiên. Sau đó, một quá trình tiến hóa lặp đi lặp lại được thực hiện. Trong mỗi lần lặp, mỗi hạt được cập nhật bằng cách sử dụng giá trị tốt nhất cá nhân (pbest) và giá trị tốt nhất toàn cầu (gbest) dựa trên hàm thích ứng được thiết kế. Giá trị pbest là nghiệm tốt nhất của một hạt cho đến nay (theo hàm thích nghi) và giá trị gbest là nghiệm tốt nhất trong số tất cả các giá trị pbest trong tổng thể. Do đó, các hạt và vận tốc tương ứng của chúng được đánh giá và cập nhật bằng cách sử dụng hai giá trị tốt nhất này. Quy trình cập nhật của từng hạt được đưa ra dưới đây:

$$vt_{Tôi}\ 1) = w_1 \times vt_{Tôi} + c_1 \times r_1 \times (\text{tốt nhất} - xt_{Tôi}) + c_2 \times r_2 \times (\text{gbest} - xt_{Tôi}) . \quad (1$$

$$xt_{Tôi}\ 1 \quad xt_{Tôi} + vt_{Tôi}\ 1 . \quad (2$$

Trong các phương trình trên, w1 là hệ số cân bằng tầm quan trọng của tìm kiếm toàn cục và tìm kiếm cục bộ. Ký hiệu vi biểu thị vận tốc của hạt thứ i trong một quần thể. c1 và c2 là các hằng số tương ứng được gọi là trọng số cá nhân và trọng lượng xã hội, xác định tầm quan trọng của giải pháp tốt nhất cá nhân và tốt nhất toàn cầu và thường được đặt thành 2. Cả r1 và r2 đều là các số ngẫu nhiên được tạo ra bởi sự phân bố đồng đều trong phạm vi của [0, 1]. Trong quá trình lặp, vận tốc của mỗi hạt lần đầu tiên được cập nhật bởi biểu thức. (1). Sau đó, mỗi hạt được cập nhật theo hướng giải pháp toàn cục bằng phương trình. (2).

PSO ban đầu được sử dụng để tìm giải pháp cho các vấn đề liên tục và đã được áp dụng trong nhiều ứng dụng. Zuo và cộng sự. (2014) đã đề xuất một khung dựa trên PSO học tập tự thích ứng để lập kế hoạch tài nguyên giữa các đám mây. Kuo và cộng sự. (2011) đã phát triển một phương pháp khám phá các luật kết hợp trong cơ sở dữ liệu giao dịch của khách hàng. Bonam và cộng sự. (2014) đã giải quyết vấn đề khai thác quy tắc kết hợp bảo đảm quyền riêng tư bằng cách phát triển phương pháp tiếp cận dựa trên PSO bằng cách sử dụng biến dạng dữ liệu. Giá trị IR được đề xuất là thước đo mới thú vị để tìm ra các luật kết hợp có ý nghĩa thay thế cho các thước đo truyền thống như độ hỗ trợ và độ tin cậy. Định tuyến và lập kế hoạch trong thế giới thực thường được xem như một vấn đề tối ưu hóa rời rạc. Vì vậy, cách tiếp cận PSO truyền thống không thể được áp dụng trực tiếp để tìm ra giải pháp cho vấn đề này. Để giải quyết vấn đề này, Kennedy và Eberhart (1997) đã phát triển PSO rời rạc để tìm giải pháp gần tối ưu cho các vấn đề tối ưu hóa rời rạc. PSO rời rạc sử dụng phương pháp cập nhật dựa trên vận tốc truyền thống của PSO liên tục, nhưng hàm sigmoid được sử dụng để cập nhật vectơ thứ j của hạt, chứa các giá trị nhị phân (0 hoặc 1). Hàm sigmoid được định nghĩa là

$$\text{ký } vt_{ij}\ \ 1 = \frac{1}{1 + e^{-vt_{ij}}\ 1} \quad (3$$

Ở đâu v1 ij biểu thị vectơ thứ j của vận tốc được tạo ra bởi phương trình. (1). Khi cập nhật một hạt, một số ngẫu nhiên được tạo ra trong khoảng [0, 1] để so sánh. Nếu số ngẫu nhiên nhỏ hơn ((+)) sig vt 1 ij , (+) = xt 1 1 ij ; ngược lại, (+) = xt 1 0 ij .
Sarath và Ravi (2013) sau đó đã thiết kế một phương pháp tối ưu hóa PSO để khai thác các luật kết hợp. Lin và cộng sự. (2015b) đã phát triển thuật toán dựa trên PSO nhị phân để khai thác hiệu quả các tập mục có tiện ích cao. Zhi và cộng sự. (2004) đã phát triển một phương pháp PSO rời rạc để giải quyết vấn đề TSP tổng quát. Thần và cộng sự. (2014) cũng áp dụng PSO rời rạc cho bài toán định tuyến multicast trong mạng truyền thông. Thiên và cộng sự. (2013) đã đề xuất một kế hoạch áp dụng PSO rời rạc để giải quyết vấn đề lập kế hoạch lắp ráp. Thiết kế các phương pháp tiếp cận dựa trên PSO để khai thác các mẫu trong cơ sở dữ liệu là một lĩnh vực nghiên cứu tích cực (Menhas và cộng sự, 2011; Pears và Koh, 2002).

2.3. Khai thác dữ liệu bảo vệ quyền riêng tư

Trong những năm gần đây, khai thác dữ liệu đã trở thành một công nghệ quan trọng đối với các doanh nghiệp vì nó có thể dễ dàng tiết lộ mối quan hệ giữa các sản phẩm được khách hàng mua. Thông tin này sau đó có thể được các nhà quản lý hoặc nhà bán lẻ sử dụng để thực hiện các chiến lược bán hàng hiệu quả và hữu ích. Tuy nhiên, vì các kỹ thuật khai thác dữ liệu được thiết kế để tiết lộ các mối quan hệ và kiến thức tiềm ẩn nên chúng cũng có thể được sử dụng để tiết lộ thông tin bí mật hoặc nhạy cảm. Để tránh vấn đề này, thông tin bí mật hoặc an toàn cần phải được ẩn đi trước khi cơ sở dữ liệu được xuất bản hoặc chia sẻ công khai với các cộng tác viên. Vì lý do này, PPDM gần đây đã nổi lên như một lĩnh vực nghiên cứu trọng điểm (Aggarwal và cộng sự, 2006; Dasseni và cộng sự, 2001; Evfimievski và cộng sự, 2002; Lindell và Pinkas, 2000; Verykios và cộng sự, 2004). Agrawal và Srikant (2000) đã trình bày một quy trình xây dựng lại để ước tính chính xác sự phân bố của các giá trị dữ liệu gốc và xây dựng các bộ phân loại để so sánh độ chính xác của cơ sở dữ liệu đã được làm sạch với cơ sở dữ liệu gốc tương ứng. Verykios và cộng sự. (2004) đã đưa ra một cái nhìn tổng quan và đề xuất cách phân loại theo thứ bậc các kỹ thuật PPDM. Hajian và cộng sự. (2014) đề xuất một cách tiếp cận dựa trên khái quát hóa để vừa bảo vệ quyền riêng tư vừa ngăn chặn sự phân biệt đối xử. Cách tiếp cận này có thể được mở rộng sang các mô hình bảo mật khác nhau và có khả năng mở rộng tốt. Lindell và Pinkas (2000) đã phát triển một phương pháp để ẩn thông tin đặc quyền khỏi quá trình học cây quyết định bằng thuật toán ID3 nổi tiếng, để chia sẻ dữ liệu giữa nhiều bên. Evfimievski và cộng sự. (2002) đã phát triển một số thuật toán để ngẫu nhiên hóa dữ liệu số và phân loại cho PPDM. Họ cũng xác định một biện pháp bảo mật mới để xem xét vấn đề bảo mật từ một góc độ khác với phương pháp mã hóa thông thường. Clifton và cộng sự. (2003) đã cung cấp bộ công cụ và một số kỹ thuật cho các ứng dụng cụ thể của PPDM. Islam và Brankovic (2011) đã đề xuất một khuôn khổ để bảo vệ quyền riêng tư của các cá nhân bằng cách sử dụng tính năng bổ sung tiếng ồn, cả hai đều xem xét dữ liệu số và dữ liệu phân loại. Atallah và cộng sự. (1999) đã trình bày một cách tiếp cận heuristic đối với nhiễu loạn dữ liệu, sửa đổi cơ sở dữ liệu gốc bằng cách giảm mức độ hỗ trợ của các quy tắc nhạy cảm dưới mức do người dùng chỉ định cụ thể.

ngưỡng. Oliveira và Zaane (2002) đã phát triển một khung heuristic với một số thuật toán dọn dẹp để ẩn các tập phổ biến. Các thuật toán được thiết kế dựa trên cách tiếp cận hạn chế mục, do đó tránh được việc thêm nhiễu và hạn chế việc xóa dữ liệu thực. Sweeney (2002) đã thiết kế một thuật toán ẩn danh k tổng quát để bảo vệ và ngăn chặn các thuộc tính nhạy cảm.

Trái ngược với các thuật toán ẩn truyền thống được phát triển cho PPDM, một số thuật toán khử trùng nhằm mục đích giảm thiểu tác dụng phụ của quá trình khử trùng. Vấn đề này có thể được coi là vấn đề tối ưu hóa NP-hard (Aggarwal và cộng sự, 2006; Verykios và cộng sự, 2004). Rất ít thuật toán lấy cảm hứng từ sinh học đã được đề xuất để tìm giải pháp cho vấn đề che giấu thông tin nhạy cảm trong PPDM. Han và Ng (2007) đã phát triển một giao thức an toàn để khám phá một bộ quy tắc đồng thời đảm bảo không tiết lộ dữ liệu riêng tư của họ bằng cách sử dụng GA. Trong công việc này, mức độ tốt của từng quy tắc quyết định được đánh giá bằng cách nhân tỷ lệ dương thực sự với tỷ lệ âm thực sự. Lin và cộng sự. đề xuất các thuật toán sGA2DT, pGA2DT (Lin và cộng sự, 2015a) và cpGA2DT (Lin và cộng sự, 2014) để ẩn các tập mục nhạy cảm bằng cách loại bỏ các giao dịch, sử dụng GA. Mỗi nhiễm sắc thể mã hóa một giải pháp bao gồm một tập hợp các giao dịch sẽ bị xóa. Mức độ tốt của nhiễm sắc thể được đánh giá bằng cách sử dụng chức năng thích ứng được thiết kế, xem xét ba tác dụng phụ của việc vệ sinh. Thuật toán sGA2DT sử dụng GA đơn giản để tìm tập hợp giao dịch thích hợp cần xóa nhằm ẩn các tập mục nhạy cảm. Để tăng tốc quá trình tiến hóa, thuật toán pGA2DT mở rộng thuật toán sGA2DT bằng cách áp dụng khái niệm tiền lớn (Hong và cộng sự, 2001). Sự tối ưu hóa này bao gồm việc duy trì một bộ đệm gồm các tập mục có kích thước lớn trong quá trình tiến hóa, để tránh thực hiện nhiều lần quét cơ sở dữ liệu cho mỗi lần lặp. Để sử dụng các thuật toán dựa trên GA truyền thống, người dùng phải chỉ định số lượng nhiễm sắc thể trong quần thể. Thuật toán cpGA2DT áp dụng cách tiếp cận GA nhỏ gọn (Harik và cộng sự, 1999) để chỉ tạo ra hai cá thể trên mỗi quần thể để cạnh tranh, do đó giảm mức sử dụng bộ nhớ trong quá trình tiến hóa. Tuy nhiên, người dùng vẫn cần chỉ định một số thông số như kích thước nhiễm sắc thể, tỷ lệ đột biến và tỷ lệ chéo. Việc tìm ra các giá trị phù hợp cho các tham số này trong các tình huống thực tế không phải là một nhiệm vụ đơn giản.

3. Sơ bộ và xác định vấn đề

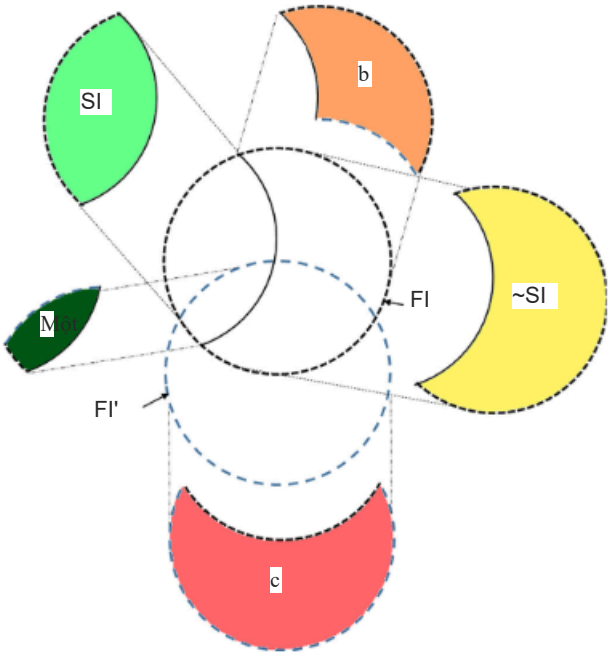
Phần này giới thiệu các bước sơ bộ và sau đó xác định vấn đề ẩn các tập phổ biến nhạy cảm trong khi giảm thiểu các tác dụng phụ.

3.1. Vòng sơ loại

Cho phép TÔI{ T_1, ĩ2 là tập hữu hạn gồm r phần tử phân biệt. Cơ sở dữ liệu D là tập hợp các giao dịch D = { T_1 2 , ở đâu cho mỗi lần chuyển hoạt động TD q , Tq là tập con của I và T q có mã định danh duy nhất q,

Bảng 1 Cơ sở dữ liệu gốc.

THỜI GIAN	Mặt hàng
1	abe, ,
2	bởi vì, ,
3	abce, , ,
4	acd, ,
5	bởi vì, ,
6	bởi vì, ,
7	bcde, , ,
8	abe, ,
9	bụng,
10	cái này,



Hình 2. Mối quan hệ của PPDM.

được gọi là Mã định danh giao dịch (TID) của nó. Ngưỡng hỗ trợ tối thiểu dành riêng cho người dùng δ, được cho là do người dùng hoặc chuyên gia đặt theo cách thủ công (phần trăm). Sau đây, cơ sở dữ liệu hiển thị trong Bảng 1 sẽ được sử dụng làm ví dụ chạy. Nó chứa 10 giao dịch trong đó các mục được thể hiện bằng các chữ cái.

Định nghĩa 1. Tập mục X là tập các mục (X ⊆)TÔI . Sự hỗ trợ số lượng tập mục X trong cơ sở dữ liệu D là số lượng giao dịch chứa X. Đặt số lượng hỗ trợ tối thiểu là tích của ngưỡng hỗ trợ tối thiểu và số lượng giao dịch trong cơ sở dữ liệu. Tập hợp các tập phổ biến được ký hiệu là FI = { f_1, !2 và được định nghĩa là tất cả các tập mục có hỗ trợ số lượng không nhỏ hơn số lượng hỗ trợ tối thiểu. Do đó, một tập mục fi được coi là tập mục phổ biến (∈) f FIs i khi và chỉ khi:

hỗ trợ $_{Tối}$ ≥ |D $_{||}$ × d. (4

Ví dụ: giả sử ngưỡng hỗ trợ tối thiểu δ được đặt thành 40%. Số lượng hỗ trợ tối thiểu được tính như sau (10 × 0,4 (= 4 . Từ Bảng 1, số lượng hỗ trợ của tập mục (bc) là hỗ trợ)(= 5 , vì nó xuất hiện trong năm giao dịch. Vì điều này giá trị lớn hơn số lượng hỗ trợ tối thiểu, ( )bc là một tự do- tập mục quent trong cơ sở dữ liệu này.

Bảng 2 Cơ sở dữ liệu dự kiến của ( )bc Và ()ce .

THỜI GIAN	Mặt hàng
2	bởi vì, ,
3	abce, , ,
5	bởi vì, ,
6	bởi vì, ,
7	bcde, , ,
10	cái này,

Sự định 2. Các bộ của nhạy cảm tập mục là ký hiệu BẰNG SI = { S1 S2 và được định nghĩa là tập hợp con của tập FI (SI DFI). Bộ này có thể được chỉ định bởi người dùng hoặc chuyên gia.

Mục tiêu của PPDM là ẩn càng nhiều thông tin nhạy cảm càng tốt trong cơ sở dữ liệu để dữ liệu bí mật không thể bị phát hiện bằng các kỹ thuật khai thác dữ liệu. Để ẩn các tập mục nhạy cảm, cần phải loại bỏ các giao dịch chứa các tập mục nhạy cảm hoặc loại bỏ các tập mục chứa thông tin bí mật khỏi các giao dịch. Tuy nhiên, việc tìm kiếm một tập hợp các giao dịch hoặc tập mục cần xóa để giảm thiểu tác dụng phụ là một bài toán khó. Ba tác dụng phụ quan trọng là không thể che giấu một số thông tin nhạy cảm (được gọi là không thể ẩn, FTH hoặc thất bại trong việc che giấu), che giấu thông tin quan trọng nhưng không nhạy cảm (được gọi là không thể ẩn, NTH hoặc thiếu chi phí) và giới thiệu thông tin nhân tạo (được gọi là không được tạo ra, NTG hoặc chi phí nhân tạo) (Wu và cộng sự, 2007). Các định nghĩa và giải thích về ba tác dụng phụ này sẽ được đưa ra sau đó.

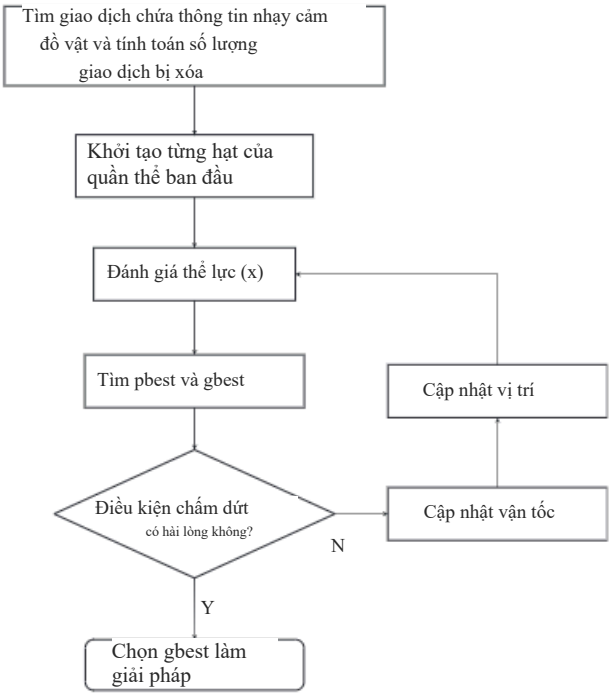
Định nghĩa 3. Hãy để `D là một cơ sở dữ liệu được làm sạch, sao cho 'D đã được thu được bằng cách xóa một số giao dịch/tập mục khỏi cơ sở dữ liệu gốc D.

Mối quan hệ giữa cơ sở dữ liệu gốc (D) và cơ sở dữ liệu đã được làm sạch (′)D được mô tả trong Hình 2. Trong Hình 2, FI biểu thị tập các tập mục phổ biến trong D, SI đại diện cho tập các tập mục nhạy cảm cần ẩn, ∼SIs là tập các tập mục phổ biến không nhạy cảm trong D, và `FIs biểu thị tập các tập mục phổ biến trong cơ sở dữ liệu đã được vệ 'D. sinh

Định nghĩa 4. Tác dụng phụ của FTH được ký hiệu là α và được định nghĩa là số lượng tập mục nhạy cảm xuất hiện trong cơ sở dữ liệu đã được lọc sạch ′D , nghĩa là:

một = |SI ∩ FI. | (5

Định nghĩa 5. Tác dụng phụ của NTH được ký hiệu là β và được định nghĩa là số lượng tập mục không nhạy cảm bị ẩn trong cơ sở dữ liệu đã được làm sạch ′D , nghĩa là:



Hình 3. Lưu đồ của thuật toán PSO2DT được thiết kế.

Bảng 3 Đã khám phá các tập phổ biến.

1 mục	Đếm	Bộ 2 món	Đếm	Bộ 3 món	Đếm
Một	5	bụng	4	bce	5
b	8	bc	5		
c	7	là	7		
e	8	ce	6		

Bảng 4 Tập mục tiền lớn.

Bộ 2 món	Đếm	Bộ 3 món	Đếm
ae	3	abe	3

Bảng 5 Các hạt ban đầu và giá trị thích hợp của chúng.

hạt	LÀN	Sự thích hợp
p1	[2, 5, 0, 7]	1.1
p2	[0, 3, 7, 0]	1.0
p3	[0, 2, 3, 3]	1.0
p4	[0, 5, 6, 0]	0,9
p5	[2, 3, 0, 10]	1.6

b = | ∼ SI − FI | = (FI − SI | − FI. | (6

Định nghĩa 6. Tác dụng phụ của NTG được ký hiệu là γ và được định nghĩa là số lượng tập mục phổ biến trong cơ sở dữ liệu đã được lọc sạch ′D không thường xuyên có trong cơ sở dữ liệu gốc D, đó là:

c = |FI − FI. | (7

3.2. Tuyên bố vấn đề

Vấn đề làm sạch cơ sở dữ liệu D cho PPDM là giảm số lượng hỗ trợ của e tập mục nhạy cảm ∈ s SIs i sao cho số lượng hỗ trợ trở nên nhỏ hơn số lượng hỗ trợ tối thiểu, nghĩa là:

Bảng 6 Vận tốc cập nhật của các hạt.

vận tốc	LÀN
v1	[6]
v2	[5, 6]
v3	[5, 6]
v4	[vô giá trị]
v5	[5, 6]

Bảng 7 Các hạt được cập nhật và giá trị thể lực của chúng.

hạt	LÀN	Sự thích hợp
p1	[6, 3, 0, 7]	0,9
p2	[5, 6, 0, 2]	1.1
p3	[5, 6, 2, 3]	0,1
p4	[2, 0, 5, 3]	0,9
p5	[5, 6, 2, 7]	0,3

Bảng 8 Các thông số của bộ dữ liệu.

#|D|	Tổng số giao dịch
#|Tôi|	Số lượng mặt hàng riêng biệt
DepL	Độ dài giao dịch trung bình
MaxLen	Thời lượng giao dịch tối đa
Kiểu	Loại tập dữ liệu

Bảng 9 Đặc điểm của bộ dữ liệu.

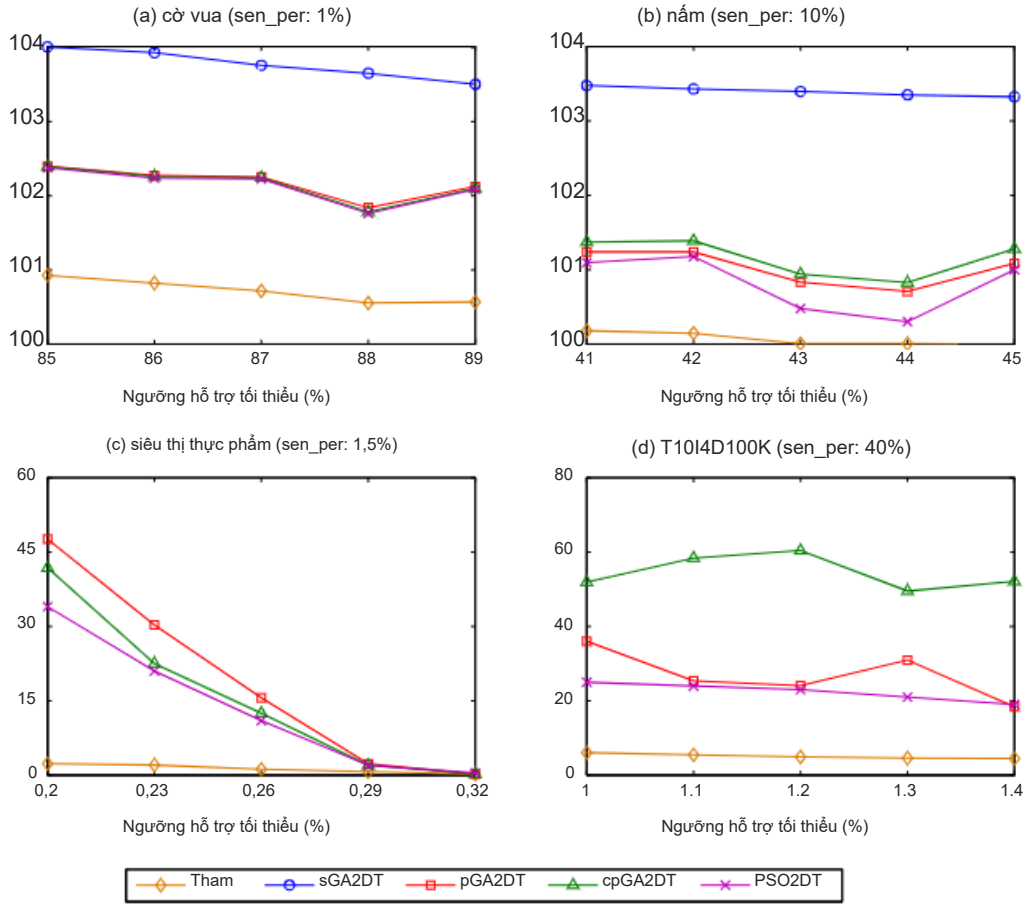
Tập dữ liệu	#|D|	#|Tôi|	DepL	MaxLen	Kiểu
Cờ vua	3196	74	37	37	dày đặc
nấm	8124	119	23	23	dày đặc
Siêu thị thực phẩm	21.556	1559	4	11	thưa thớt
T10I4D100K	100.000	870	10.1	29	thưa thớt

hỗ trợ $_{\text{Tôi}}$ d × |D.| (8

Ba tác dụng phụ nói trên là các biện pháp tiêu chuẩn được sử dụng để đánh giá mức độ tốt của phương pháp vệ sinh. Nếu số FTH quá cao, điều đó có nghĩa là nhiều mẫu nhạy cảm vẫn có thể được phát hiện trong cơ sở dữ liệu đã được lọc sạch. Nếu số lượng NTH quá cao, điều đó cho thấy thông tin đưa ra quyết định quan trọng có thể bị thiếu trong cơ sở dữ liệu đã được làm sạch. Nếu số NTG quá cao có nghĩa là một lượng lớn vô nghĩa thông tin nhân tạo có thể đã được đưa vào bởi quá trình khử trùng. Nếu lượng thông tin nhạy cảm cần ẩn rất lớn thì có khả năng cao là thông tin quan trọng không nhạy cảm cũng có thể bị ẩn bởi quá trình sàng lọc. Ngoài ra, hãy lưu ý rằng khi các giao dịch bị xóa trong cơ sở dữ liệu, số lượng giao dịch sẽ thay đổi và do đó số lượng hỗ trợ tối thiểu cũng thay đổi. Do đó, nhiều tập mục không thường xuyên có thể trở nên thường xuyên do quá trình dọn dẹp. Vì vậy, có mối quan hệ đánh đổi giữa tác dụng phụ của FTH, NTH và NTG. Việc tìm ra giải pháp cho vấn đề PPDM nhằm giảm thiểu ba tác dụng phụ là một vấn đề NP-khó. Trong bài viết này, chúng tôi không chỉ tập trung vào việc che giấu càng nhiều thông tin nhạy cảm càng tốt mà còn giảm thiểu ba tác dụng phụ của việc vệ sinh.

4. Đề xuất thuật toán vệ sinh PSO2DT

Vì PPDM là NP-hard nên các phương pháp heuristic nên được sử dụng để tìm các giải pháp gần như tối ưu thay vì cố gắng tìm các giải pháp thực sự tối ưu. Trong bài báo này, khái niệm PSO rời rạc từ điện toán tiến hóa được áp dụng trong thuật toán được thiết kế để tìm ra một tập hợp các giao dịch cần xóa một cách hiệu quả nhằm giảm thiểu ba tác dụng phụ. Điểm mạnh quan trọng của phương pháp được đề xuất là cần đặt một số tham số so với các phương pháp dựa trên GA cho PPDM, nhưng phương pháp được đề xuất vẫn tìm kiếm các giải pháp bằng cách sử dụng phương pháp tiến hóa ngẫu nhiên. Thuật toán được thiết kế thực hiện các bước sau.



Hình 4. Thời gian chạy ghi các giá trị ngưỡng hỗ trợ tối thiểu khác nhau.

4.1. Khởi tạo

Để ẩn các tập mục nhạy cảm thông qua việc xóa giao dịch, chỉ các giao dịch chứa ít nhất một tập mục nhạy cảm (một tập mục trong tập SI) mới được xem xét xóa.

Định nghĩa 7. Bước đầu tiên là trích xuất cơ sở dữ liệu dự kiến $*D$ từ cơ sở dữ liệu gốc D. Cơ sở dữ liệu dự kiến $*D$ là tập hợp của các giao dịch trong D chứa ít nhất một tập mục nhạy cảm, đó là:

$$*D \leftarrow \{ T_q \mid T_q \in D, \exists S_{Tôi} \in HS, S_{Tôi} \subseteq T_q \}. \quad (9$$

Ví dụ: giả sử rằng hai tập mục nhạy cảm ( )bc và ( )ce cần phải được ẩn giấu. Cơ sở dữ liệu dự kiến tương ứng $*D$ cho ()bc)ce được thể hiện trong Bảng 2. Trong thuật toán PSO2DT được thiết kế, mỗi hạt đại diện cho một

giải pháp và được đánh giá bằng hàm thích hợp được thiết kế. Kích thước hạt được đặt thành số lượng giao dịch thích hợp sẽ bị xóa để ẩn các tập mục nhạy cảm.

Định nghĩa 8. Số lượng giao dịch thích hợp cần xóa được ký hiệu là m và được định nghĩa là sự khác biệt giữa số lượng hỗ trợ lớn nhất trong số tất cả các tập mục nhạy cảm trong SI và số lượng hỗ trợ tối thiểu, đó là:

$$m \text{ Hỗ trợ tối đa } \left\lceil \frac{\dots d \times |D|}{1 \dots} \right\rceil . \quad (10)$$

Định nghĩa 9. Hạt pi là một vectơ m chiều, trong đó mỗi chiều biểu thị một giao dịch cần xóa $\in *$ TD q và

lưu trữ TID của nó. Lưu ý rằng thứ nguyên có thể tùy ý chứa giá trị null, biểu thị không có giao dịch nào.

Ví dụ, kể từ khi ( ) sup bc (¼5) và ( ) hỗ trợ (¼6), khoảng số lượng giao dịch riêng cần xóa được tính như sau:

tôi$\left(= \left\lceil \frac{6 - 0{,}4 \times 10}{1 - 0{,}4} \right\rceil \right)(= 4)$. Như vậy, bốn giao dịch sẽ bị xóa trong cơ sở dữ liệu dự kiến $*D$ để ẩn tập mục nhạy cảm ( )ce . Qua xóa bốn giao dịch, số lượng hỗ trợ của ( )ce có thể tăng- ngày như hỗ trợ cái gì $(= 6 - 4 (= 2$ và kích thước của da- đã được vệ sinh tabase sẽ trở thành ($10 - 4 (= 6$. Có thể nhận thấy rằng ( )ce là ẩn thành công khi bốn giao dịch chứa ( )ce là đã xóa từ cái nguyên bản cơ sở dữ liệu từ ( )(= < × =) phụ 2 0,4 6 2,4 . Như vậy, phương trình thiết kế (10) là hợp lý và có thể chấp nhận được. Do đó, kích thước của mỗi hạt trong quá trình tiến hóa được đặt thành 4. Mỗi hạt được khởi tạo ngẫu nhiên. Trong ví dụ này, một hạt có thể được khởi tạo là:

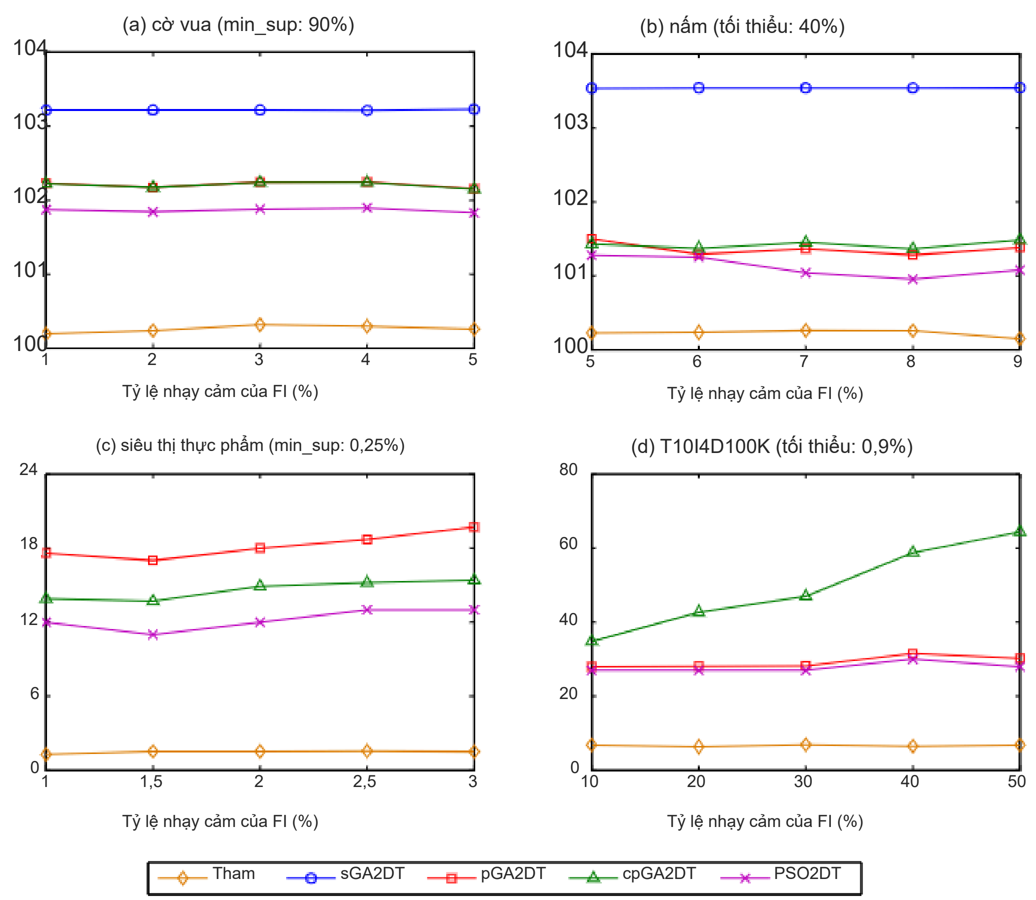
$P_{Tội} = [2, 5, 6, 7]$.

Để đánh giá mức độ tốt của một hạt, một chức năng thích ứng linh hoạt được đề xuất, đo lường ba tác dụng phụ thường được sử dụng của quá trình khử trùng. Trong chức năng này, trọng số được sử dụng để chỉ ra tầm quan trọng tương đối của từng tác dụng phụ đối với người dùng và có thể được đặt theo sở thích của từng người dùng.

Định nghĩa 10. Hàm thích ứng được đề xuất là tổng trọng số của ba tác dụng phụ và được định nghĩa là:

$$\text{thể dục } p_{Tôi} = w_1 \times M + w_2 \times b + w_3 \times c \quad (11)$$

trong đó α là số tập mục FTH; β là số tập mục NTH; và γ là số tập mục NTG; w1, w2 và w3 là



Hình 5. Thời gian chạy ghi các tỷ lệ phần trăm nhạy cảm khác nhau của FI.

trọng số cho thấy tầm quan trọng tương đối của từng tác dụng phụ và người dùng có thể điều chỉnh.

Vì vậy, nếu người dùng muốn ẩn càng nhiều thông tin nhạy cảm càng tốt thì w1 nên được đặt cao hơn. Nếu người dùng muốn đảm bảo rằng thông tin quan trọng không nhạy cảm được lưu giữ để đưa ra quyết định thì w2 nên được đặt cao hơn. Và nếu người dùng muốn giảm việc đưa vào thông tin nhân tạo thì w3 nên được đặt cao hơn. Lưu ý rằng mật độ của cơ sở dữ liệu cũng ảnh hưởng đến tính phù hợp của hạt được đánh giá. Do việc phân phối dữ liệu được cô đọng trong cơ sở dữ liệu dày đặc nên nếu một tập mục nhạy cảm được ẩn thành công thì nhiều tập mục không nhạy cảm nhưng thường xuyên cũng có thể bị ẩn như một tác dụng phụ. Vì vậy, đối với cơ sở dữ liệu dày đặc, w1 nên được đặt cao hơn để có được sự cân bằng tốt giữa ba tác dụng phụ. Ngược lại, việc phân phối dữ liệu trong cơ sở dữ liệu thưa thớt còn thưa thớt. Do đó, nếu một tập mục nhạy cảm bị ẩn thì sẽ có ít tập mục không nhạy cảm nhưng thường xuyên bị ẩn hơn do hiệu ứng phụ. Do đó, w1 có thể được đặt thấp hơn đối với các tập dữ liệu thưa thớt.

Trong thuật toán PSO2DT được thiết kế, một hạt và vận tốc của nó là tập hợp các giao dịch sẽ bị xóa trong ∗D để ẩn các thông tin nhạy cảm.

thông tin và được biểu diễn dưới dạng tập hợp TID. Để cập nhật vận tốc của một hạt, các phép tính hiệu và hợp được áp dụng trên hạt với pbest và gbest, theo phương trình. (12), trong đó phép toán sai phân tính toán sự khác biệt giữa hai hạt và toán tử hợp được áp dụng để thu được hợp của hai tập hợp tạo thành một hạt mới. Bằng cách sử dụng cơ chế cập nhật này, thuật toán PSO2DT được thiết kế xử lý vấn đề vệ sinh như một vấn đề rời rạc. Vì lý do này, một số tham số được sử dụng trong PSO truyền thống như r1, r2, c1 và c2 trong biểu thức. (1)

là không cần thiết trong cách tiếp cận được thiết kế. Để cập nhật một hạt, các TID chứa trong hạt cũ hoặc null được chọn ngẫu nhiên để lấp đầy kích thước của hạt được cập nhật. Kết quả sau đó được tính tổng với vận tốc được cập nhật của hạt, như thể hiện trong biểu thức. (13). Quá trình này làm tăng tính ngẫu nhiên của quá trình tiến hóa và do đó làm tăng khả năng khám phá của nó so với các tham số được sử dụng trong PSO truyền thống. Quá trình cập nhật đầy đủ của các hạt và vận tốc của chúng trong thuật toán PSO2DT được thiết kế được hiển thị bên dưới:

$$vt_{Tôi} \quad 1) = (\text{ tốt nhất } - xt_{Tôi}) \cup (\text{ gbest } - xt_{Tôi}) . \tag{12}$$

$$xt_{Tôi} \quad 1 \quad = \text{ rand } xt_{Tôi} , \ ^{\text{giá trị}}) + vt_{Tôi} \quad 1 . \tag{13}$$

Ví dụ, xem xét kích thước hạt là 4, quá trình cập nhật được áp dụng cho hạt ( ) = [ ] xt 1, 4, 7, 5 , mà

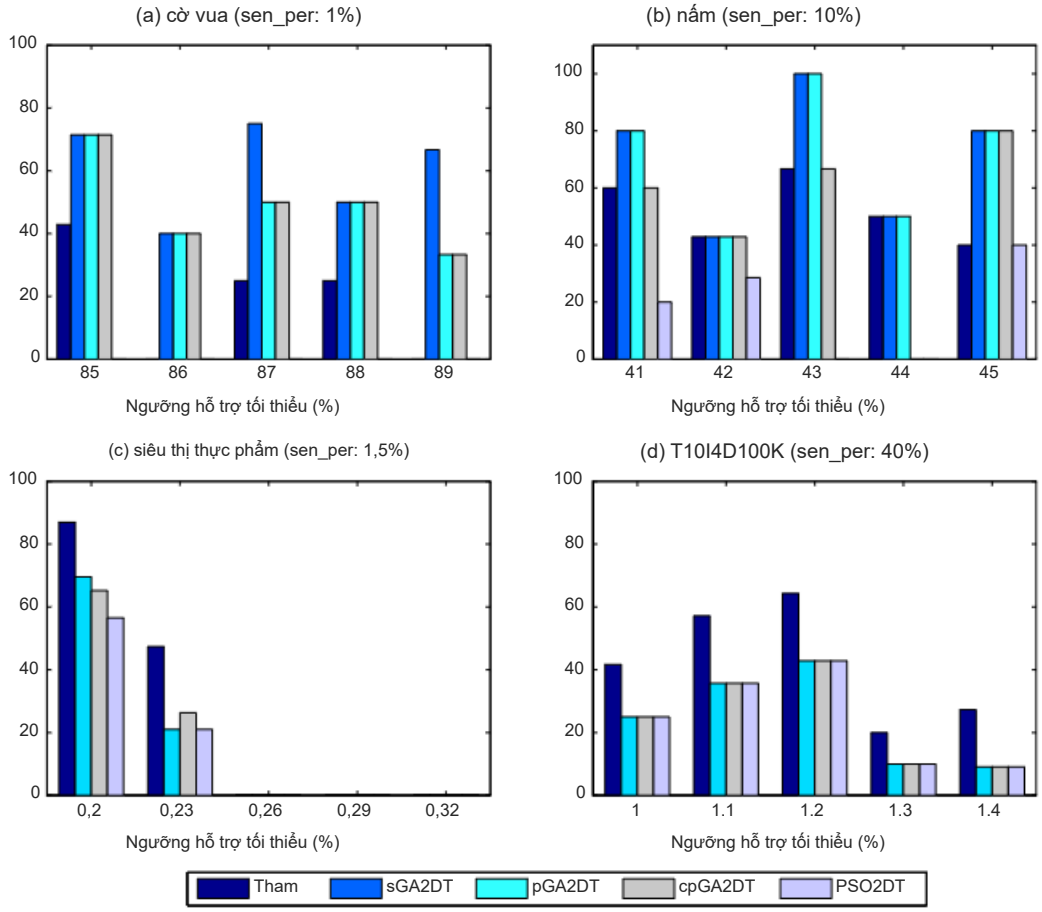
= [ ] pbest 2, 1, 0, 5 và = [ ] gbest 0, 3, 0, 4 . Trong ví dụ này, 0 biểu thị giá trị null. Theo phương trình. (12), vận tốc của hạt cập nhật được tính như (+) = {[ ] − [vt 1 2, 1, 0, 5 1, 4, 7,

5 } ∪ {[0, 3, 0, 4 − [1, 4, 7, 5 } = [2, 0 ∪ [0, 3 = [0, 2, 3 | . Từ có một chiều trống trong hạt được cập nhật, sau đó nó được lấp đầy bằng cách chọn ngẫu nhiên một giá trị từ hạt cũ của nó hoặc giá trị null [¼0]. Giả sử rằng ([ ]) = [ ] rand 1, 4, 7, 5 , 0 1 . sự lên-

do đó hạt có niên đại là [0, 2, 3, 1].

4.2. Thuật toán PSO2DT được đề xuất

Thuật toán PSO2DT được thiết kế được đề xuất để ẩn các tập mục nhạy cảm một cách hiệu quả bằng cách xóa giao dịch. Mã giả được đưa ra trong Thuật toán 1.



Hình 6. FTH ghi các giá trị ngưỡng hỗ trợ tối thiểu khác nhau.

Thuật toán PSO2DT được thiết kế trước tiên tính toán kích thước hạt là sự khác biệt giữa số lượng hỗ trợ của tập mục nhạy cảm thường xuyên nhất và số lượng hỗ trợ tối thiểu, bằng cách áp dụng phương trình. (10) (Dòng 1). Sau đó, hạt M được tạo ra. là quần thể ban đầu cho quá trình tiến hóa lặp đi lặp lại.

Thuật toán 1. Thuật toán PSO2DT đề xuất.

Input: D^*, the set of projected transactions; SIs, a set of sensitive itemsets to be hidden; FIs, the set of frequent itemsets in D; δ, the minimum support threshold; and M, the number of particles to be used for each iteration
Output: D', a sanitized database.

```
1  calculate the size of a particle as m;                    // by equation (10)
2  for i ← 1, M do
3      for i ← 1, m do
4          p_i(t) ← p_i(t) ∪ rand(D*);
5  while termination criterion is not met do
6      for i ← 1, M do
7          if fitness(p_i(t)) ≤ pbest_i(t) then
8              pbest_i(t) ← p_i(t);
9          if pbest_i(t) ≤ gbest then
10             gbest ← pbest_i(t);
11     for i ← 1, M do
12         update the v_i(t + 1) velocities of M particles;  // by equation (12)
13         update the p_i(t + 1) for M particles;            // by equation (13)
14     set t ← t + 1;
15 delete transactions in gbest from D;
16 return the sanitized database D';
```

Mỗi hạt này chứa m TID riêng biệt được chọn ngẫu nhiên từ ∗D (Dòng 2 và 3). Quá trình lặp lại sau đó bắt đầu và

được lặp lại cho đến khi thỏa mãn tiêu chí kết thúc được xác định trước. Trong mỗi lần lặp, giá trị thích hợp của từng hạt được đánh giá bằng hàm thích nghi. Sau đó, pbest i tốt nhất cá nhân của mỗi hạt được tính toán cũng như gbest tốt nhất toàn cầu từ các giá trị M pbest (Dòng 6–10). Sau đó, mỗi hạt được cập nhật theo các phương trình. (12) và (13) (Dòng 11–14). Quá trình lặp lại này sau đó được lặp lại cho đến khi đáp ứng tiêu chí kết thúc (Dòng 5–14). Sau đó, các giao dịch trong gbest được chọn làm giao dịch cần xóa để dọn dẹp (Dòng 15). Cuối cùng, cơ sở dữ liệu đã được làm sạch thu được sẽ được trả về (Dòng 16). Lưu đồ của thuật toán PSO2DT được thiết kế được hiển thị trong Hình 3.

4.3. Một chiến lược được cải thiện

Quá trình phát triển của thuật toán PSO2DT được thiết kế thực hiện nhiều lần quét cơ sở dữ liệu để đánh giá tác dụng phụ khi các giao dịch bị xóa, đặc biệt là để đánh giá tác dụng phụ của NTG. Để tăng tốc độ tính toán tác dụng phụ này và do đó cũng là quá trình tiến hóa, khái niệm tiền lớn (Hong và cộng sự, 2001) đã được áp dụng trong thuật toán được thiết kế. Theo khái niệm tiền lớn, một số tập mục không phổ biến có độ hỗ trợ cao có thể trở nên phổ biến khi các giao dịch bị xóa. Những tập phổ biến nhất này được lưu giữ trong bộ đệm để đánh giá sau này và được gọi là tập phổ biến trước. Sử dụng bộ đệm này, tất cả các tập mục NTG có thể được xác định từ các tập mục có kích thước lớn mà không cần quét cơ sở dữ liệu, vì điều này sẽ được giải thích. ĐẾN

xác định tập mục nào sẽ được thêm vào bộ đệm, ngưỡng hỗ trợ thấp hơn sẽ được xác định. Trong công việc trước đây, khái niệm tiền lớn chủ yếu được sử dụng để khai thác dữ liệu gia tăng nhằm theo dõi các tập phổ biến khi cơ sở dữ liệu được cập nhật thường xuyên bằng các thao tác chèn, xóa hoặc sửa đổi. Khái niệm lớn về việc xóa giao dịch được trình bày sau đó.

Định nghĩa 11. Cho có hai ngưỡng hỗ trợ Su và Sl. Không cần thiết phải quét cơ sở dữ liệu gốc để phát hiện các tập mục NTG nếu số lượng giao dịch bị xóa nhỏ hơn giới hạn an toàn ( )f , được xác định như sau:

$$f S \quad \frac{\text{bạn} - S_{\text{tối}}) \times |D|}{S_{\text{bạn}}} \qquad (14)$$

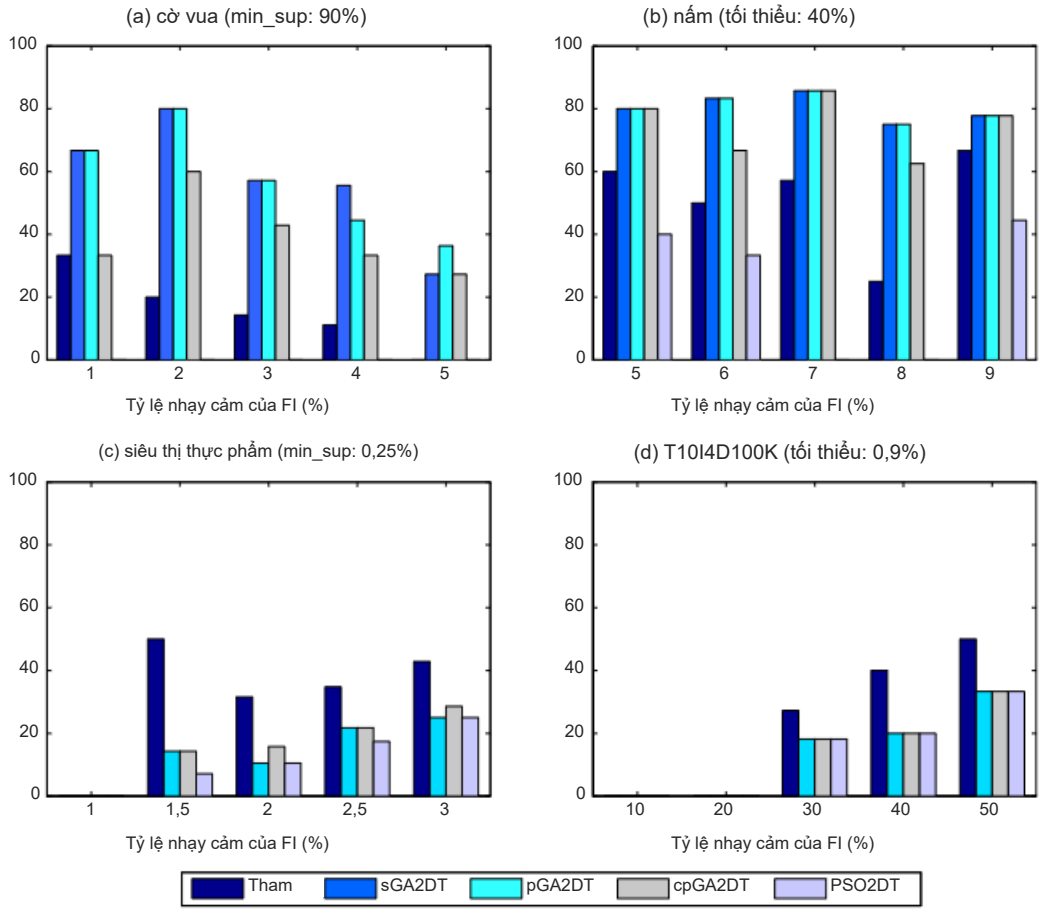
Nói chung, ngưỡng hỗ trợ trên trong khái niệm tiền lớn được xác định là ngưỡng hỗ trợ tối thiểu. Hãy nhớ lại rằng thuật toán PSO2DT được thiết kế sẽ chọn các giao dịch cần xóa khỏi cơ sở dữ liệu để ẩn các tập mục nhạy cảm. Kích thước của mỗi hạt được đặt thành chênh lệch giữa số lượng hỗ trợ của tập mục nhạy cảm thường xuyên nhất và ngưỡng hỗ trợ tối thiểu. Trong PSO2DT, giá trị này được sử dụng làm giới hạn an toàn ( )f để tránh

hình thành nhiều lần quét cơ sở dữ liệu. Do đó, khái niệm tiền lớn có thể được sửa đổi để đạt được ngưỡng hỗ trợ thấp hơn như sau:

$$f\left(= \frac{\text{bệnh đa xơ cứng}}{\text{bạn}} - \frac{S_{\text{tối}}) \times |D|}{S_{\text{bạn}}} \Rightarrow S_{\text{lu}} \quad \frac{\left(D_{,} | - (f \cdot \text{tôi}) \right)}{|D_{,}|}. \qquad (15)$$

Ví dụ trong Bảng 1, ngưỡng hỗ trợ trên Su là 40% và số lượng giao dịch cần xóa là 4. Do đó, ngưỡng hỗ trợ dưới có thể được tính như sau: ( ) = (=) × (−) S 24 % l 40 % 10 4 10 .

Trong quá trình khám phá FI ban đầu, mỗi tập mục có độ hỗ trợ giữa Sl và Su được gọi là tập mục tiền lớn và được



Hình 7. FTH ghi lại các tỷ lệ phần trăm nhạy cảm khác nhau của FI.

do đó được chèn vào bộ đệm. Ví dụ: số lượng hỗ trợ của tập mục ( )ae trong Bảng 1 là 3. Do đó, tập mục này là tập mục tiền lớn.

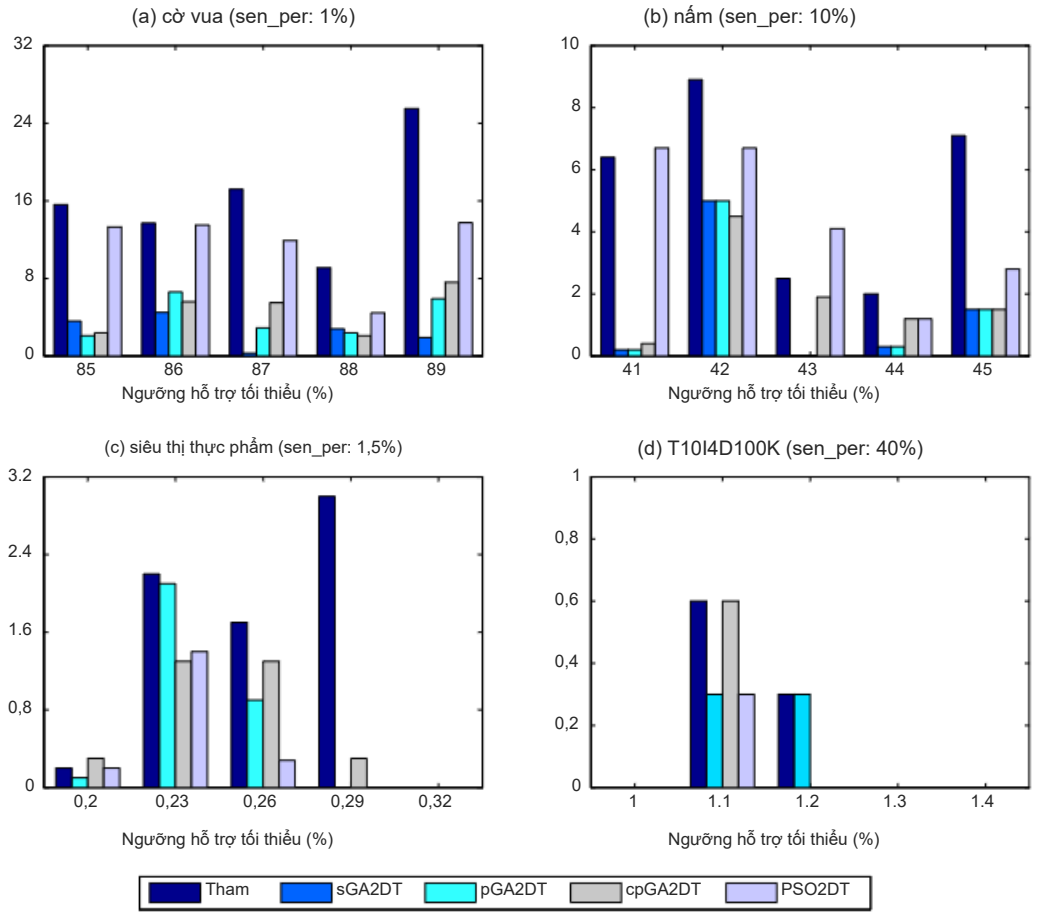
Nhờ khái niệm tiền lớn, thuật toán có thể tránh việc quét cơ sở dữ liệu nhiều lần và điều này do đó đẩy nhanh quá trình phát triển.

5. Ví dụ minh họa

Trong phần này, một ví dụ sử dụng cơ sở dữ liệu trong Bảng 1 được đưa ra để minh họa từng bước cho thuật toán PSO2DT được đề xuất. Trong ví dụ này, ngưỡng hỗ trợ tối thiểu được đặt thành 40%. Các tập mục phổ biến ( )FI được phát hiện được thể hiện trong Bảng 3.

Hãy xem xét các tập mục đó ( )bc và ( )ce là các tập mục nhạy cảm cần được ẩn đi vì mục đích của PPDM. Đầu tiên, số lượng giao dịch cần xóa để ẩn hoàn toàn các tập mục nhạy cảm được tính như sau: − × − ∏ ∏ 6 0,4 10 1 0,4 (¼4). Vì vậy, mức hỗ trợ thấp hơn ngưỡng của khái niệm tiền lớn có thể được tính như sau: $\frac{40\% \times (10 -)4}{10}$ (¼24%). Cơ sở dữ liệu ban đầu sau đó được quét để tìm các tập mục có kích thước lớn, sẽ được sử dụng để tránh thực hiện nhiều lần quét cơ sở dữ liệu trong quá trình tiến hóa, đặc biệt là để tính toán số lượng tập mục NTG. Các tập mục tiền lớn được thể hiện trong Bảng 4.

Sau đó, các giao dịch chứa ít nhất một tập mục nhạy cảm được xác định để tạo cơ sở dữ liệu dự kiến. Cơ sở dữ liệu dự kiến chứa tất cả các giao dịch sẽ được xem xét xóa trong quá trình tiến hóa. Trong ví dụ đang chạy, các giao dịch này có các mã định danh { }2, 3, 5, 6, 7, 10 và được hiển thị trong Bảng 2. Sau đó, các phần tử được khởi tạo trong quần thể ban đầu bằng cách chọn ngẫu nhiên các giá trị từ tập hợp các giao dịch sẽ bị xóa và giá trị null . Giả sử rằng các trọng số của α β và γ trong thể lực được đặt tương ứng thành 0,8, 0,1 và 0,1, có thể điều chỉnh theo sở thích của người dùng. Số lượng hạt trong quần thể được đặt thành 5 và số lần lặp được đặt thành 10. Quá trình tiến hóa sẽ sử dụng các tập mục có kích thước lớn để tránh thực hiện quét nhiều lần cơ sở dữ liệu, đặc biệt là để đánh giá tác dụng phụ NT-G. Hãy lấy hạt p1 trong Bảng 5 làm ví dụ để minh họa quá trình tiến hóa. Hạt p1 đại diện cho giải pháp xóa các giao dịch [2, 5, 7] trong cơ sở dữ liệu gốc. Trong ví dụ này, tỷ lệ hỗ trợ của ( )bc và ( )ce lần lượt là được tính như (7/2 = 28,5% < 40% và (7/3 = 42,8% > 40%. Như vậy, theo giải pháp này, ( )bc bị ẩn hoàn toàn nhưng ( )ce là một FTH. Vì các tập phổ biến đã được phát hiện nên việc duy trì NTH rất dễ dàng. Trong ví dụ này, các tỷ số hỗ trợ của {( ) ( ) ( ) ( ) ( ) ( )} bce bc be ce bce , , , , , , được thể hiện trong Bảng 3 và là được cập nhật tương ứng để xem xét việc xóa các giao dịch. Chỉ ( )bce là NTH vì tỷ lệ hỗ trợ của nó sau khi xóa được tính là: (= <) 2/7 28,5% 40% . Tác dụng phụ của NTH cũng có thể dễ dàng xảy ra được duy trì bằng cách sử dụng các tập mục tiền lớn được trình bày trong Bảng 4. Trong ví dụ này, tỷ lệ hỗ trợ của ( )ae và ( )abe đều tăng và cả hai đều được cập nhật là: (7/3 = 42,8% > 40%. Vì vậy, ( )ae và ( ) abe là NTG. TRONG ví dụ này, giá trị thích hợp của hạt p1 được tính như thể dục p 1) = (0,8 × 1 + 0,1 × 1 + 0,1 × 2) = 1.1 . Các hạt khác được xử lý theo cách tương tự và các hạt được tạo ra ban đầu và các giá trị thích hợp của chúng được thể hiện trong Bảng 5.



Hình 8. NTH cho các giá trị ngưỡng hỗ trợ tối thiểu khác nhau.

Từ kết quả của Bảng 5, có thể thấy rằng p4 có giá trị thích nghi nhỏ nhất. Do đó, p4 được cập nhật bằng cách sử dụng giá trị tốt nhất toàn cầu (gbest) của lần lặp này và giá trị tốt nhất cá nhân của nó (pbest). Theo phương trình. (12), hiệu số tập hợp giữa pbest và từng hạt, cũng như hiệu số giữa gbest và từng hạt, được tính toán tương ứng. Sau đó, sự kết hợp các kết quả của hai tập hợp hiệu được tính cho từng hạt và được sử dụng làm vận tốc của hạt cho lần lặp tiếp theo. Hãy xem p1 trong Bảng 5 làm ví dụ để minh họa các bước này. Trong ví dụ này, pbest của p1 là chính nó và gbest là p4. Sự khác biệt được đặt giữa p1 và pbest được tính như sau:

[2, 5, 0, 7 − [2, 5, 0, 7 = [ ]vô giá trị. Sự khác biệt được đặt giữa p1 và gbest được tính như [0, 5, 6, 0 − [2, 5, 0, 7 = [ ]6 . Liên minh của hai tập hợp khác biệt là [ ] ∪ [ ] = [ ]null 6 . Vận tốc v1 của p1 đối với do đó lần lặp tiếp theo được tính là [6]. Vận tốc còn lại của các hạt khác được tính theo cách tương tự. Vận tốc cập nhật cho tất cả các hạt được thể hiện trong Bảng 6.

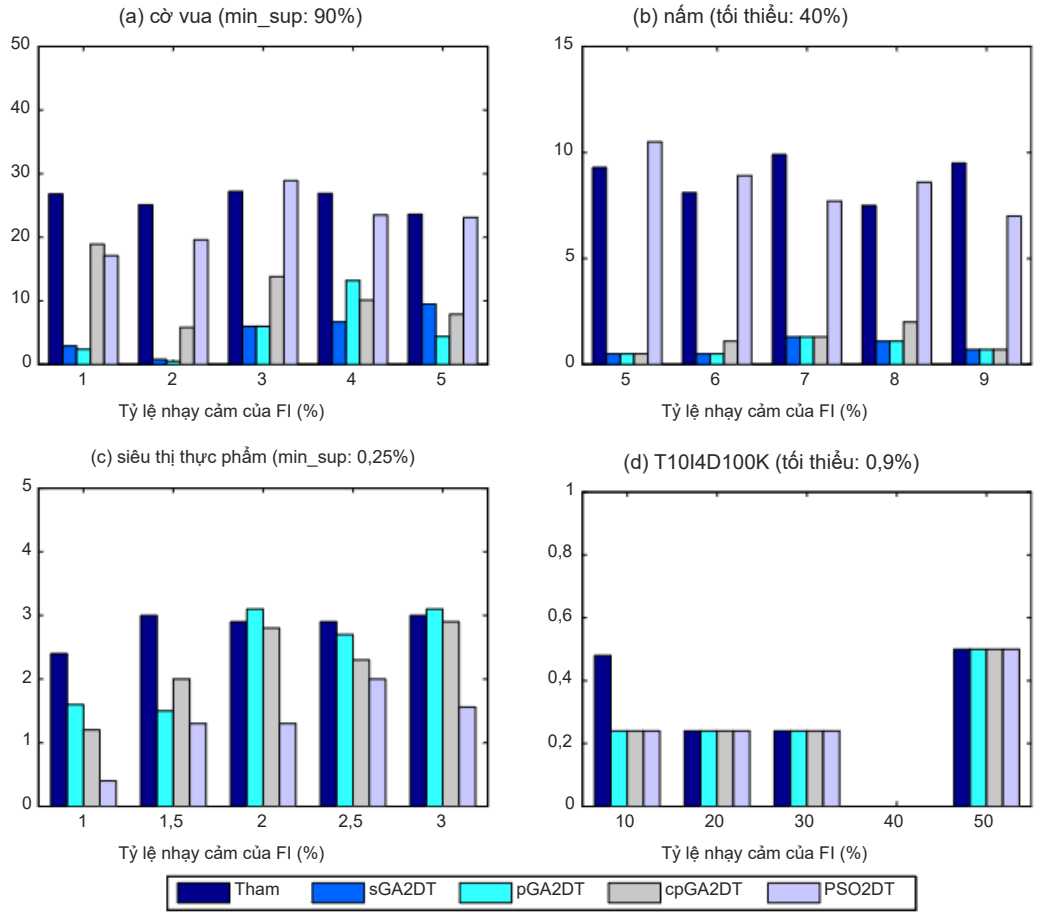
Dựa trên phương trình. (13), TID từ hạt cũ và giá trị null được chọn ngẫu nhiên để duy trì khả năng thăm dò của hạt. Hãy xem v1 trong Bảng 6 làm ví dụ để minh họa quá trình này. Trong ví dụ này, kích thước của v1 là 1 và kích thước của hạt là 4. Ba giao dịch bao gồm giá trị null được chọn ngẫu nhiên từ hạt cũ p1. Kết quả là [6, 3, 0, 7]. Quá trình này được lặp lại với mọi hạt khác theo cách tương tự. Các hạt thu được sau khi áp dụng cơ chế cập nhật này được thể hiện trong Bảng 7.

Trong Bảng 7, có thể thấy rằng p3 có giá trị thích nghi nhỏ nhất. Cái sau này nhỏ hơn gbest hiện tại. Do đó, p3 được đặt làm gbest mới. Pbest của mỗi hạt cũng được cập nhật trong quá trình này sự lặp lại. Quá trình tiến hóa trên được lặp lại cho đến khi điều kiện kết thúc được thỏa mãn. Lưu ý rằng ba tác dụng phụ của hạt được xử lý có thể được tính toán dễ dàng mà không cần thực hiện nhiều lần quét cơ sở dữ liệu (đặc biệt là tác dụng phụ của NT-G khi sử dụng các tập mục có kích thước lớn được phát hiện). Do đó, gbest cuối cùng được đặt làm giải pháp và các giao dịch tương ứng với TID trong gbest cuối cùng sẽ lần lượt bị xóa trong cơ sở dữ liệu gốc để ẩn các tập mục nhạy cảm. Trong ví dụ này, các giao dịch { }2, 3, 5, 6 được chọn để xóa nhằm ẩn các tập mục nhạy cảm

( )bc và ( )ce . Tác dụng phụ FTH, NTH và NTG lần lượt là 0, 1 và 0. Do đó, giá trị thích hợp được tính bằng (× + × + ×) 0,8 0 0,1 1 0,1 0 (¼ 0,1).

6. Kết quả thực nghiệm

Các thí nghiệm đáng kể đã được tiến hành để so sánh tính hiệu quả và hiệu quả của thuật toán PSO2DT được đề xuất với các phương pháp tiếp cận sGA2DT, pGA2DT, cpGA2DT dựa trên GA tiên tiến nhất (Lin và cộng sự, 2015a, 2014) và Thuật toán vệ sinh Greedy không tiến hóa (Lin và cộng sự, 2013). Thuật toán PSO2DT được đề xuất áp dụng khái niệm tiền lớn để tránh thực hiện quét nhiều cơ sở dữ liệu nhằm đánh giá tác dụng phụ trong quá trình phát triển. Các thuật toán trong thử nghiệm được triển khai bằng Java. Các thử nghiệm được thực hiện trên PC trang bị bộ xử lý Intel Core 2 i3-4160 và RAM 4 GB, chạy trên hệ điều hành Microsoft Windows 7 64 bit. Ba bộ dữ liệu trong thế giới thực được gọi là cờ vua (Bộ dữ liệu khai thác tập mục thường xuyên



Hình 9. NTH ghi lại các tỷ lệ phần trăm nhạy cảm khác nhau của FI.

Repository, 2012), Mushroom (Kho lưu trữ tập dữ liệu khai thác tập mục thường xuyên, 2012) và foodmart (Microsoft) đã được sử dụng trong các thử nghiệm. Một tập dữ liệu tổng hợp có tên T10I4D100K cũng đã được tạo bằng trình tạo cơ sở dữ liệu IBM (Agrawal và Srikant, 1994a). Các thông số và đặc điểm của bộ dữ liệu sử dụng trong thí nghiệm lần lượt được trình bày trong Bảng 8 và 9.

Trong các thí nghiệm này, số lần lặp được đặt là 100 và 10 hạt được sử dụng trong mỗi quần thể. Tất cả các thử nghiệm được thực hiện năm lần và các giải pháp có giá trị thích hợp nhỏ nhất được sử dụng để xác định các giao dịch cần xóa để ẩn các tập mục nhạy cảm. Một tỷ lệ phần trăm nhất định của các tập phổ biến được tìm thấy trong mỗi cơ sở dữ liệu được chọn ngẫu nhiên là các tập phổ biến nhạy cảm. Tỷ lệ phần trăm này được gọi là tỷ lệ phần trăm của các tập mục nhạy cảm. Sau đây là mức hỗ trợ tối thiểu

ngưỡng và tỷ lệ phần trăm của các tập mục nhạy cảm lần lượt được ký hiệu là min_sup và sen_per. Trong thực tế, giá trị của min_sup và sen_per có thể được chỉ định bởi người dùng hoặc chuyên gia. Để đánh giá hiệu suất của phương pháp được thiết kế, các giá trị min_sup và sen_per đã được điều chỉnh cho từng tập dữ liệu, để đảm bảo rằng số lượng tập mục phổ biến và tập mục nhạy cảm là phù hợp. Do đó, trong các thử nghiệm được tiến hành, các tham số min_sup và sen_per khác nhau đối với mỗi tập dữ liệu và được điều chỉnh dựa trên đặc điểm của từng tập dữ liệu.

6.1. Thời gian chạy

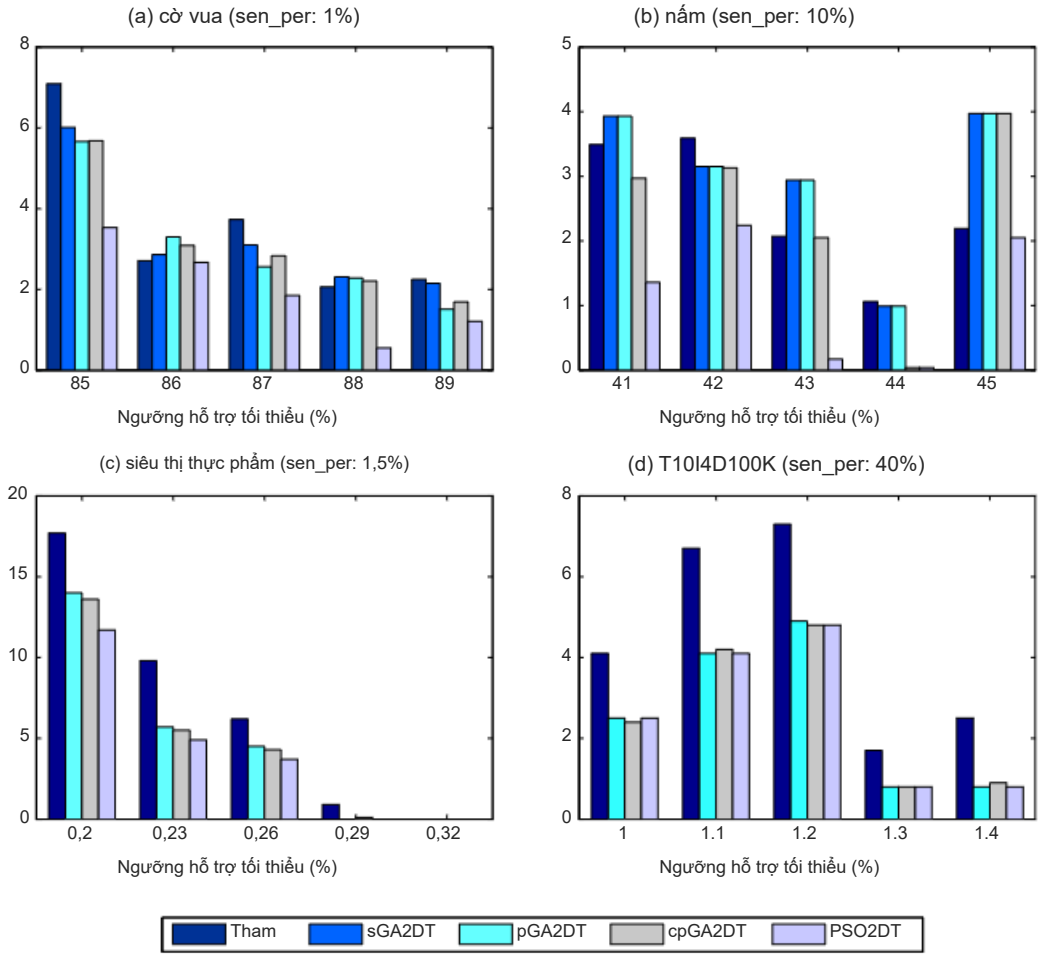
Một thử nghiệm đầu tiên được thực hiện để so sánh thời gian chạy của thuật toán được thiết kế với Greedy, sGA2DT, pGA2DT và thuật toán cpGA2DT, khi ngưỡng hỗ trợ tối thiểu và tỷ lệ phần trăm nhạy cảm thay đổi. Kết quả được hiển thị trong hình. 4 và 5. Vì Greedy không sử dụng cách tiếp cận tiến hóa nên thuật toán này luôn chỉ được thực hiện một lần. Thời gian chạy của Greedy ít hơn các phương pháp tiến hóa trong Hình. 4 và 5. Trên thực tế, không nên so sánh các phương pháp tiếp cận tiến hóa với các phương pháp tiếp cận tiến hóa về mặt thời gian chạy, vì hai loại phương pháp tiếp cận này rất khác nhau. Việc so sánh kết quả của thuật toán không tiến hóa với thuật toán tiến hóa sử dụng phân tích ANOVA hai chiều cũng là vô nghĩa. Vì vậy, sau đây chúng tôi chỉ so sánh kết quả của thuật toán dựa trên GA với thuật toán được thiết kế bằng phân tích ANOVA hai chiều.

Trong Hình 4(c) và (d), thời gian chạy của thuật toán sGA2DT không được hiển thị vì chúng vượt quá 1000 giây cho một lần lặp. Có thể nhận thấy từ những kết quả này rằng thuật toán được đề xuất vượt trội hơn các thuật toán dựa trên GA trên bốn bộ dữ liệu. Một quan sát khác là khi ngưỡng hỗ trợ tối thiểu tăng lên, số lượng FI sẽ giảm và các tập mục ít nhạy cảm hơn cần phải được ẩn đi. Do đó, thời gian chạy của các thuật toán được so sánh sẽ giảm khi ngưỡng hỗ trợ tối thiểu tăng lên. Dựa trên phân tích ANOVA hai chiều, thời gian chạy của thuật toán sGA2DT khác biệt đáng kể so với thời gian chạy của các thuật toán khác (= < F p24,66, 0,001, được hiển thị trong Hình 4(a); = < F p252,395 , 0,001,

thể hiện trong hình 4(b)). Đối với tập dữ liệu foodmart được hiển thị trong Hình 4(c), không có sự khác biệt đáng kể giữa các thời gian chạy

của tất cả các thuật toán so sánh (F p4.040, = 0,061 > 0,05 . Vì

tập dữ liệu T10I4D100K được hiển thị trong Hình 4 (d), thời gian chạy của cpGA2DT khác biệt đáng kể so với hai thuật toán còn lại



Hình 10. Các giá trị thể lực thu được cho các giá trị ngưỡng hỗ trợ tối thiểu khác nhau.

(= <) F p65.295, 0.001 . Trong Hình 4, cũng có thể thấy rằng ngưỡng hỗ trợ tối thiểu có ảnh hưởng nhỏ đến thời gian chạy của các thuật toán được so sánh. Nhìn chung, thuật toán PSO2DT đề xuất luôn vượt trội hơn các thuật toán dựa trên GA khác, tức là nó có thể tìm ra lời giải nhanh hơn các thuật toán khác. Hình 5 cho thấy kết quả thu được từ các thuật toán khi phần trăm độ nhạy của FI thay đổi.

Thời gian chạy của thuật toán sGA2DT không được hiển thị trong Hình 5 (c) và (d) vì chúng lại vượt quá 1000 giây cho một lần lặp. Trong Hình 5, có thể thấy rằng thuật toán PSO2DT được đề xuất vẫn vượt trội hơn các thuật toán dựa trên GA khi tỷ lệ phần trăm nhạy cảm của FI thay đổi, trên bốn bộ dữ liệu. Thuật toán PSO2DT được đề xuất nhanh hơn tới ba lần so với thuật toán pGA2DT và cpGA2DT được so sánh trong Hình 5(a) và (b). Ngoài ra, theo phân tích ANOVA hai chiều, có sự khác biệt đáng kể giữa thời gian chạy của thuật toán được thiết kế và các thuật toán khác (= < F p10146.414, 0,001, được hiển thị trong Hình 5(a)). Sự khác biệt là

giữa thuật toán sGA2DT và các thuật toán so sánh khác cũng có ý nghĩa (= <) F p229189.397, 0,001 . Dành cho siêu thị thực phẩm

và bộ dữ liệu T10I4D100K được hiển thị trong Hình 5 (c) và (d), thời gian chạy của thuật toán PSO2DT được đề xuất ít hơn tới 20% hoặc 30% so với các phương pháp khác. Ngoài ra, thời gian chạy của thuật toán PSO2DT được thiết kế khác biệt đáng kể so với thuật toán cpGA2DT và pGA2DT (= < F p493.437, 0,001, được hiển thị trong Hình 5 (c)). không có

sự khác biệt đáng kể giữa thời gian chạy của thuật toán PSO2DT và pGA2DT nhưng có một sự khác biệt đối với thuật toán cpGA2DT

(F p18.450, = 0,001 < 0,05 . Nhìn chung, phương pháp PSO2DT được đề xuất gorithm vượt trội hơn các thuật toán dựa trên GA tiên tiến nhất.

6.2. Đánh giá tác dụng phụ

Cách tiêu chuẩn để so sánh các thuật toán dọn dẹp trong PPDM là so sánh số lần xuất hiện của ba tác dụng phụ mà mỗi thuật toán tạo ra. Trong phần này, tác dụng phụ của FTH và NTH là thước đo để so sánh hiệu quả của thuật toán được thiết kế với thuật toán khác. Vì tác dụng phụ của NTG khá hiếm đối với tất cả các thuật toán được so sánh nên không cần thiết phải so sánh hiệu suất của các thuật toán ghi lại tác dụng phụ này.

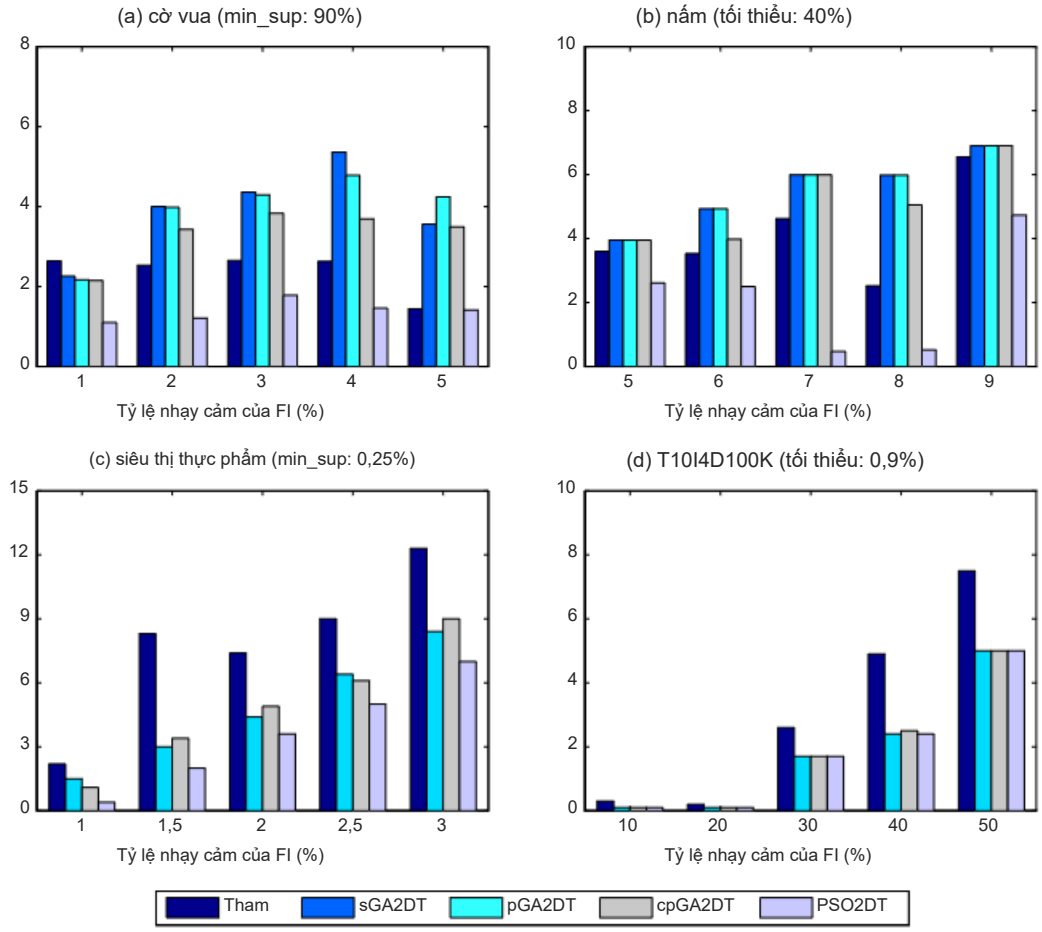
6.2.1. Không thể ẩn được, FTH Tác dụng phụ của FTH đã được đo lường để so sánh mức độ

phần lớn thông tin nhạy cảm đã không được ẩn thành công bởi mỗi thuật toán dọn dẹp. Hãy nhớ lại rằng tác dụng phụ này được định nghĩa là:

$$F - T - H = \left| \frac{*SI}{SI} \right|, \quad (16)$$

trong đó SI là số tập mục nhạy cảm trong cơ sở dữ liệu gốc và ∗SI là số tập mục nhạy cảm vẫn xuất hiện

trong cơ sở dữ liệu đã được vệ sinh. Kết quả cho các giá trị ngưỡng hỗ trợ tối thiểu khác nhau được hiển thị trong Hình 6.



Hình 11. Giá trị phù hợp cho các tỷ lệ phần trăm nhạy cảm khác nhau của FI.

Trong Hình 6, có thể thấy rằng thuật toán PSO2DT được đề xuất đạt được kết quả tốt hơn về mặt FTH so với các thuật toán dựa trên GA và thuật toán Greedy. Ví dụ: thuật toán PSO2DT được đề xuất ẩn hoàn toàn tất cả các tập mục nhạy cảm cho tập dữ liệu cờ vua như trong Hình 6(a) và cũng ẩn thành công tất cả các tập mục nhạy cảm khi ngưỡng hỗ trợ tối thiểu được đặt thành 43% và 44% trên tập dữ liệu nấm, như trong Hình 6(b). Thuật toán được đề xuất có lợi thế rõ ràng so với thuật toán Tham lam đối với tất cả các giá trị ngưỡng tối thiểu được kiểm tra và tất cả các cơ sở dữ liệu, như trong Hình 6(a)–(d). Trên hai bộ dữ liệu cờ vua và nấm dày đặc, kết quả cho Greedy nhìn chung tốt hơn các thuật toán dựa trên GA về mặt FTH, trong khi trên hai bộ dữ liệu foodmart và T10I4D100K thưa thớt, kết quả cho Greedy là kém nhất. Dựa trên phân tích ANOVA hai chiều, kết quả thu được từ thuật toán PSO2DT khác biệt đáng kể so với các thuật toán khác ngoại trừ thuật toán Greedy

(= <) F p15.130, 0,001 , như trong Hình 6(a). Trong Hình 6(b), kết quả của thuật toán PSO2DT được đề xuất khác biệt đáng kể so với các thuật toán khác (= <) F p10.030, 0.05 . Đối với tập dữ liệu thưa thớt

chẳng hạn như foodmart và T10I4D100K, kết quả thu được từ thuật toán PSO2DT được đề xuất tương tự như các thuật toán được so sánh, như trong Hình 6 (d), nhưng PSO2DT vẫn vượt trội hơn so với

Thuật toán pGA2DT và cpGA2DT trong Hình 6(c) khi ngưỡng hỗ trợ tối thiểu được đặt thành 0,2%. Đối với các tập dữ liệu thưa thớt, không có sự khác biệt đáng kể giữa các thuật toán so sánh

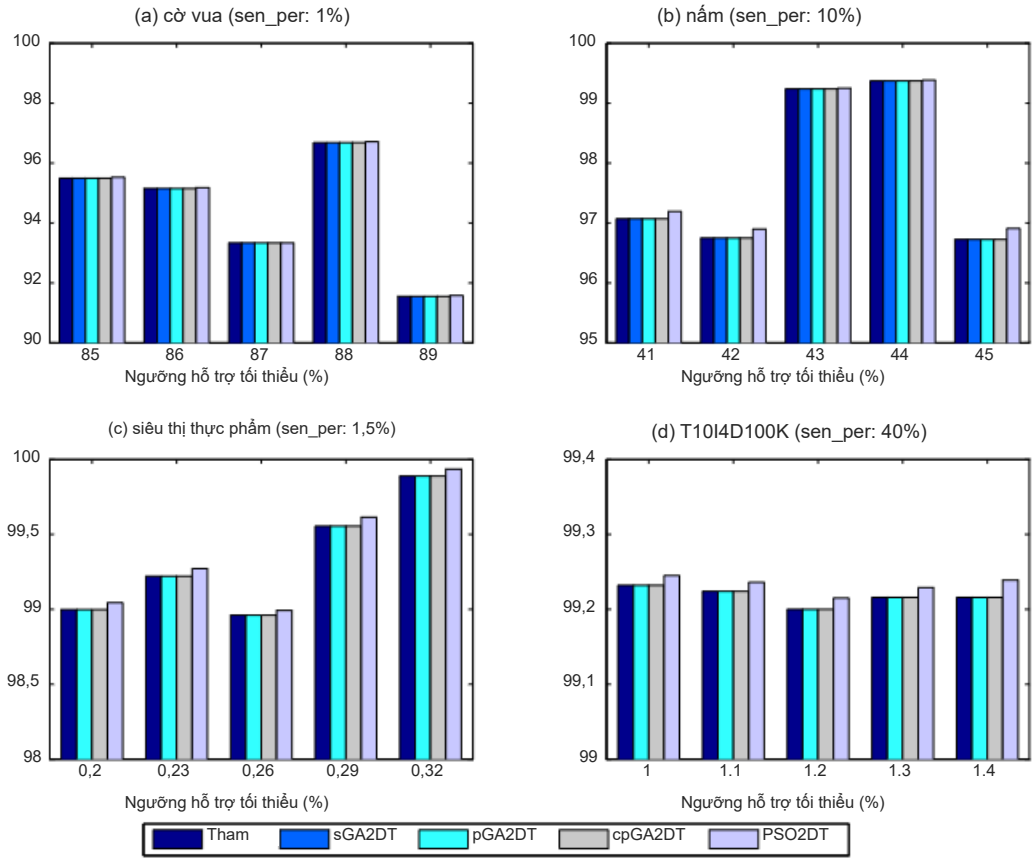
(== > F p2.418, 0.117 0.05, như trong Hình 6(c)). Tuy nhiên, kết quả đạt được bằng thuật toán PSO2DT được đề xuất rất đáng kể.

khác với thuật toán Greedy (F ▫ p34.571, < 0,001, hiển thị trong hình 6(d)). Kết quả thu được dưới dạng FTH, khi thay đổi tỷ lệ phần trăm nhạy cảm của FI được hiển thị trong Hình 7.

Đối với tập dữ liệu cờ vua (kết quả được hiển thị trong Hình 7(a)), thuật toán PSO2DT được đề xuất đã ẩn thành công các tập mục nhạy cảm trong khi các thuật toán dựa trên GA và thuật toán Greedy thì không. Đối với tập dữ liệu nấm, thuật toán PSO2DT được đề xuất cũng đạt được kết quả tốt hơn so với các thuật toán khác, như có thể thấy trong Hình 7(b). Lý do là các tập hợp giao dịch bị xóa để ẩn từng tập mục nhạy cảm có độ chồng chéo cao. Do đó, khi thuật toán PSO2DT đề xuất xóa một giao dịch, một số tập mục nhạy cảm có thể bị ẩn cùng lúc. Theo phân tích ANOVA hai chiều, kết quả của thuật toán PSO2DT được đề xuất khác biệt đáng kể so với các thuật toán khác, ngoại trừ thuật toán Tham lam (= <) F p35.221, 0,001, như được hiển thị trong

Hình 7(a). Đối với tập dữ liệu nấm, kết quả của thuật toán PSO2DT khác biệt đáng kể so với các thuật toán khác (= <) F p18.469, 0,001 , như trong Hình 7(b). Đối với tập dữ liệu thưa thớt

chẳng hạn như foodmart và T10I4D100K, kết quả đạt được bằng thuật toán đề xuất gần như giống hệt với kết quả của thuật toán dựa trên GA. Những kết quả này là hợp lý vì các tập mục nhạy cảm trong các tập dữ liệu thưa thớt thường không xuất hiện trong cùng một giao dịch và do đó, thuật toán PSO2DT được đề xuất không thể ẩn hoàn toàn các tập mục nhạy cảm khi xem xét ba tác dụng phụ. Nhưng thuật toán đề xuất vẫn vượt trội hơn các thuật toán pGA2DT và cpGA2DT dựa trên GA, như trong Hình 7(c). Hơn nữa, kết quả của thuật toán PSO2DT khác biệt đáng kể so với thuật toán Greedy.



Hình 12. DS ghi các giá trị ngưỡng hỗ trợ tối thiểu khác nhau.

thuật toán (F = trang 13.029, < 0,001, được hiển thị trong Hình 7(c)), nhưng không phải từ cái khác so sánh thuật toán (F = p4.261, = 0,029 > 0,05, thể hiện trong hình 7(d)).

6.2.2. Không được giấu kín, NTH Tác dụng phụ của NTH được sử dụng để đánh giá có bao nhiêu

Các tập phổ biến nhạy cảm bị ẩn bởi quá trình dọn dẹp, đó là:

$$N - T - H = \frac{-FI \quad SI - FI^{*}|}{|FI - SI|}, \tag{17}$$

trong đó FI là tập các tập phổ biến được phát hiện trong cơ sở dữ liệu gốc và SI là tập các tập mục nhạy cảm. Do đó, thuật ngữ −FIs SI là số tập mục phổ biến không nhạy cảm trong

cơ sở dữ liệu gốc. ký hiệu ∗FI biểu thị sự thường xuyên

các tập mục vẫn xuất hiện trong cơ sở dữ liệu đã được làm sạch. Do đó, thuật ngữ − − ∗ FIs SIs FI là số lượng tần số không nhạy cảm

các tập mục bị ẩn bởi quá trình dọn dẹp. Ảnh hưởng của việc thay đổi ngưỡng hỗ trợ tối thiểu lên NTH được trình bày trong Hình 8.

Trong Hình 8, có thể thấy rằng thuật toán PSO2DT được đề xuất tạo ra nhiều NTH hơn các thuật toán dựa trên GA, đặc biệt là trong Hình 8 (a) và (b). Lý do là đối với các tập dữ liệu dày đặc như cờ vua và nấm, thuật toán PSO2DT được đề xuất sẽ ẩn hoàn toàn tất cả các tập mục nhạy cảm trong hầu hết các trường hợp, như được hiển thị tương ứng trong Hình 6(a) và (b). Tuy nhiên, có một mối quan hệ đánh đổi giữa FTH và NTH vì nếu các tập mục nhạy cảm hơn bị ẩn đi thì các tập phổ biến hơn sẽ bị bỏ qua. Do đó, một số thông tin không nhạy cảm có thể bị che giấu bởi quá trình dọn dẹp.

xử lý thông qua việc xóa giao dịch. Nhưng quy tắc này không đúng với Greedy. Có thể thấy rằng NTH của Greedy nhìn chung là kém hơn trong số tất cả các thuật toán. FTH của ba thuật toán dựa trên GA trên foodmart và T10I4D100K cũng như FTH của thuật toán PSO2DT được thiết kế trên tất cả các tập dữ liệu đều nhỏ hơn Greedy. Điều này là do Greedy chọn các giao dịch sẽ bị xóa mà không xem xét việc giảm thiểu các tác dụng phụ. Dựa trên phân tích ANOVA hai chiều, số NTH cho thuật toán PSO2DT khác biệt đáng kể so với các thuật toán khác ngoại trừ thuật toán Tham lam (= < F p13.794, 0,05, được hiển thị trong Hình 8 (a)), nhưng

không có sự khác biệt đáng kể giữa tất cả các thuật toán được so sánh (= > F p0.999, 0.05, được hiển thị trong Hình 8(b)). Đối với các tập dữ liệu thưa thớt, nó có thể

có thể thấy rằng thuật toán PSO2DT được đề xuất vượt trội hơn các thuật toán dựa trên GA và tạo ra NTH gần như bằng 0 trong một số trường hợp so với các thuật toán dựa trên GA, như trong Hình 8 (c) và (d). Tuy nhiên, không có sự khác biệt đáng kể giữa các thuật toán được so sánh (== > F p2.648, 0,097 0,05, được hiển thị trong

Hình 8(c); = => F p1, 0,426 0,05, như trong Hình 8(d)). Kết quả thu được dưới dạng NTH, khi thay đổi tỷ lệ phần trăm nhạy cảm của FI được hiển thị trong Hình 9.

Trong Hình 9, có thể thấy rằng thuật toán PSO2DT được đề xuất cũng tạo ra nhiều NTH hơn so với các thuật toán sGA2DT, pGA2DT và cpGA2DT dựa trên GA nhưng ít hơn Greedy trong hầu hết các trường hợp. Điều này có thể được quan sát trong Hình 9 (a) và (b) đối với các tập dữ liệu dày đặc. Lý do giống như Hình 8. Đối với các tập dữ liệu thưa thớt như foodmart (kết quả được hiển thị trong Hình 9(c)), thuật toán PSO2DT được đề xuất vẫn vượt trội hơn các thuật toán dựa trên GA và kết quả tương tự đối với các thuật toán khác (như trong Hình 9(d)). Theo phân tích ANOVA hai chiều, kết quả của thuật toán PSO2DT được đề xuất khác biệt đáng kể so với các thuật toán khác.

các thuật toán ngoại trừ thuật toán Greedy (F = p33.525, < 0,05, thể hiện trong hình 9(a); F = p119.184, < 0,001, được hiển thị trong Hình 9(b); = < F p11,88, 0,05, được hiển thị trong Hình 9 (c)). Tuy nhiên, không có sự khác biệt đáng kể so với các thuật toán được so sánh trong Hình 9(d) (= = >) F p1.000, 0,426 0,05, vì kết quả là giống nhau đối với giảm bớt thuật toán. Tổng thể, cái đề xuất PSO2DT thuật toán đạt được kết quả tốt hơn trên các tập dữ liệu thưa thớt.

6.2.3. Giá trị phù hợp Để so sánh thêm thuật toán được thiết kế với thuật toán dựa trên GA

gorithms và Greedy, giá trị thích hợp của các nghiệm cuối cùng cũng được so sánh. Trọng số của ba tác dụng phụ trong chức năng thể dục có thể được điều chỉnh theo sở thích của người dùng. Để ẩn càng nhiều thông tin nhạy cảm càng tốt, w1 được đặt cao hơn nhiều so với hai trọng số còn lại trong thử nghiệm của chúng tôi. Do đó, w1, w, 2 và w3 lần lượt được đặt thành 0,98, 0,01 và 0,01 cho các bộ dữ liệu cờ vua và nấm dày đặc. w1, w, 2 và w3

các trọng số lần lượt được đặt thành 0,8, 0,1 và 0,1 cho các bộ dữ liệu foodmart và T104D100K thưa thớt. Các giá trị thích hợp thu được cho các giá trị ngưỡng hỗ trợ tối thiểu khác nhau được hiển thị trong Hình 10.

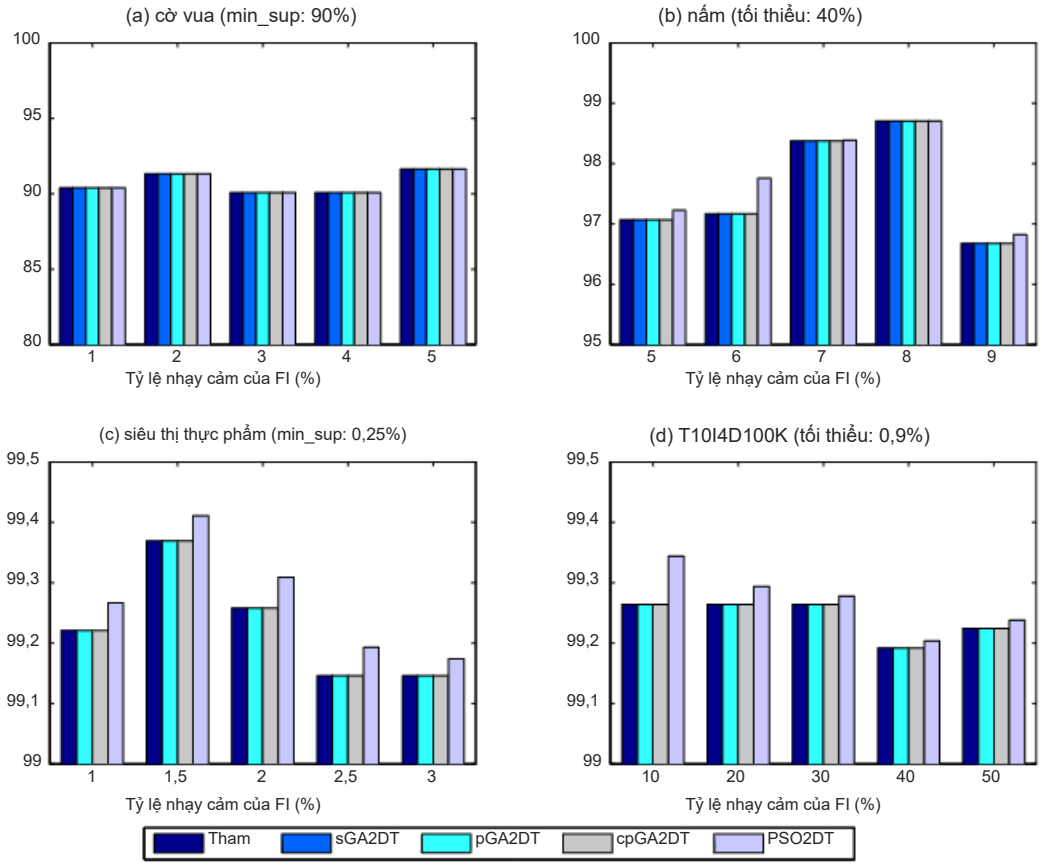
Trong Hình 10, rõ ràng là các giá trị thích hợp thu được bằng thuật toán được thiết kế nhìn chung tốt hơn so với các giá trị thu được bằng bốn thuật toán khác. Đối với tập dữ liệu cờ vua (Hình 10(a)), lý do là tác dụng phụ của FTH, vì thuật toán PSO2DT được thiết kế luôn có thể ẩn thành công tất cả các tập mục nhạy cảm cho tập dữ liệu đó. Mặc dù Greedy có thể ẩn các tập mục nhạy cảm hơn ba thuật toán dựa trên GA cho hai tập dữ liệu dày đặc (Hình 6 và 7), giá trị thích hợp của nó cao hơn nhiều so với các thuật toán dựa trên GA trong hầu hết các trường hợp đối với cờ vua và nấm. bộ dữ liệu. Điều này là do Greedy tạo ra một lượng lớn tập mục NTH. Vì có một quan hệ đánh đổi giữa FTH và NTH, tác dụng phụ của NTH đã xuất hiện. Sự cố này xảy ra đối với các tập dữ liệu rất dày đặc. Mặc dù số lượng tập mục FTH được tạo ra bởi thuật toán PSO2DT được thiết kế tương tự như các phương pháp khác, PSO2DT vẫn có thể thu được ít NTH và NTG hơn và giá trị thích hợp của các giải pháp cuối cùng của nó nhỏ hơn đối với hai tập dữ liệu thưa thớt. Điều này có thể được quan sát trong Hình 10(a) và (b). Theo phân tích ANOVA hai chiều, kết quả của thuật toán PSO2DT khác biệt đáng kể so với các thuật toán khác (= < F p6.767, 0,05, được hiển thị trong Hình 10(a);

= < F p9,283, 0,05, như trong Hình 10(b)). Đối với các tập dữ liệu thưa thớt (kết quả được hiển thị trong Hình 10 (c) và (d)), kết quả của thuật toán PSO2DT được đề xuất không khác biệt đáng kể so với các thuật toán khác ngoại trừ thuật toán Greedy (= < F p6.197, 0,05, thể hiện trong Hình 10(c); = < F p35.7, 0,001, được hiển thị trong Hình 10(d)). Các giá trị thích hợp thu được cho các tỷ lệ phần trăm nhạy cảm khác nhau của FI được hiển thị trong Hình 11.

Trong Hình 11, có thể quan sát được các kết quả tương tự như Hình 10. Giá trị thích hợp của thuật toán được thiết kế thấp hơn khoảng hai hoặc ba lần so với các thuật toán khác, đặc biệt đối với các bộ dữ liệu dày đặc. Dựa trên phân tích ANOVA hai chiều, các giá trị phù hợp thu được từ thuật toán PSO2DT được đề xuất khác biệt đáng kể so với các thuật toán khác ngoại trừ thuật toán Tham lam

(= < F p17.267, 0,001, được hiển thị trong Hình 11(a)). Ngoài ra, kết quả của thuật toán PSO2DT được đề xuất khác biệt đáng kể so với tất cả các thuật toán được so sánh (= < F p11.644, 0,05, được hiển thị trong Hình 11(b)).

Đối với các bộ dữ liệu foodmart và T10I4D100K thưa thớt, thuật toán PSO2DT được đề xuất vẫn hoạt động tốt hơn các thuật toán dựa trên GA, như được hiển thị trong Hình 11(c) và có các giá trị phù hợp gần như giống như các phương pháp tiếp cận khác, như được hiển thị trong Hình 11(d) . Các giá trị thích hợp của thuật toán PSO2DT khác biệt đáng kể so với thuật toán Greedy



Hình 13. DS ghi lại các tỷ lệ phần trăm nhạy cảm khác nhau của FI.

thuật toán chứ không phải thuật toán khác (F ▫ p24.013, < 0,001, như trong Hình 11 (c); = < F p5.438, 0.05, như trong Hình 11(d)). Từ các cuộc thảo luận ở trên, có thể kết luận rằng thuật toán PSO2DT được đề xuất có thể thu được kết quả tốt hơn vì nó xem xét ba tác dụng phụ, không giống như các thuật toán dựa trên GA khác.

6.3. Sự tương đồng về cơ sở dữ liệu

Bên cạnh ba tác dụng phụ, cần đánh giá độ tương tự cơ sở dữ liệu (DS) giữa cơ sở dữ liệu gốc và cơ sở dữ liệu đã được làm sạch để kiểm tra sự khác biệt về kích thước giữa hai cơ sở dữ liệu này. Nếu sự khác biệt của thuật toán nhỏ hơn so với các thuật toán khác, điều đó có nghĩa là thuật toán đã chọn tập hợp giao dịch tốt hơn để xóa và do đó tránh xóa các giao dịch không liên quan có thể dẫn đến ẩn nhiều tập mục không nhạy cảm nhưng thường xuyên hơn. DS được định nghĩa là:

$$\text{DS ĐĐ}, \qquad (18)$$

Trong đó D là số lượng giao dịch trong cơ sở dữ liệu gốc D và ∗D là số lượng giao dịch trong cơ sở dữ liệu đã được lọc sạch ∗D .

Các kết quả ghi lại các giá trị ngưỡng hỗ trợ tối thiểu khác nhau và tỷ lệ phần trăm nhạy cảm của FI tương ứng được hiển thị trong Hình. 12 và 13.

Từ các kết quả được hiển thị trong Hình. 12 và 13, có thể thấy rằng tất cả các thuật toán được so sánh đều đạt được tính toàn vẹn cơ sở dữ liệu tốt vì giá trị DS không bao giờ nhỏ hơn 90%. Vì số lượng giao dịch bị xóa tối đa nhất định cho thuật toán Greedy là số lượng giao dịch mà thuật toán dựa trên GA xóa, DS của Greedy giống với thuật toán dựa trên GA. Thuật toán PSO2DT được đề xuất vượt trội hơn các thuật toán dựa trên GA về mặt DS. Điều này chỉ ra rằng thuật toán PSO2DT được đề xuất có thể ẩn hoàn toàn các tập mục nhạy cảm trong khi vẫn duy trì tính toàn vẹn cơ sở dữ liệu tốt. Dựa trên các thử nghiệm đã tiến hành, thuật toán PSO2DT được thiết kế có thể tìm thấy các giao dịch tốt hơn để xóa, giúp giảm thiểu ba tác dụng phụ và cũng duy trì tính tương tự cơ sở dữ liệu cao thông qua việc xóa giao dịch. Dựa trên phân tích ANOVA hai chiều, kết quả của thuật toán PSO2DT khác biệt đáng kể so với các thuật toán khác (= < F p15.980, 0,001, được hiển thị trong Hình 12(a); = < F p7.431, 0,05, được hiển thị trong Hình 12(b); F ▫ p107.233, < 0,001, được hiển thị trong Hình 12(c)). Tuy nhiên, không có sự khác biệt đáng kể giữa các thuật toán được so sánh vì tất cả chúng đều bảo toàn tính toàn vẹn cơ sở dữ liệu tốt khi ẩn các tập mục nhạy cảm (== > F p3.093, 0.223 0.05, được hiển thị trong Hình 13(a); => F p0,057, 0,05, như trong Hình 13(b)). Ngoài ra, DS cho thuật toán PSO2DT không khác biệt đáng kể so với các thuật toán được so sánh cho các tập dữ liệu thưa thớt (= < F p116.381, 0,001, được hiển thị trong Hình 13(c); = < F p5.396, 0.05, như trong Hình 13(d)). Tóm lại, thuật toán PSO2DT được đề xuất có thể chấp nhận được để ẩn các tập mục nhạy cảm thông qua việc xóa giao dịch.

7. Kết luận và công việc tiếp theo

Trước đây, các thuật toán dựa trên GA đã được đề xuất để ẩn các tập mục nhạy cảm thông qua việc xóa giao dịch. Đây là bài báo đầu tiên thiết kế phương pháp tiếp cận dựa trên PSO để ẩn các tập mục nhạy cảm thông qua việc xóa giao dịch, đồng thời giảm thiểu tác dụng phụ. Trong thuật toán PSO2DT được thiết kế, số lượng giao dịch tối đa cần xóa sẽ tự động được xác định và đặt làm kích thước hạt. Thuật toán được đề xuất linh hoạt hơn so với các phương pháp tiếp cận dựa trên GA trước đây và thuật toán Tham lam không tiến hóa, vì người dùng cần ít tham số hơn. Hơn nữa, thuật toán PSO2DT được thiết kế áp dụng khái niệm tiền lớn để tránh thực hiện nhiều lần quét cơ sở dữ liệu và do đó tăng tốc quá trình phát triển. Các thí nghiệm cơ bản có được thực hiện để so sánh hiệu suất của thuật toán được thiết kế với các phương pháp tiếp cận dựa trên GA tiên tiến nhất và thuật toán Greedy về ba tác dụng phụ và độ tương tự của cơ sở dữ liệu (tính toàn vẹn). Kết quả thử nghiệm cho thấy thuật toán được thiết kế nhanh hơn nhiều và tạo ra tác dụng phụ NTH nhỏ hơn so với các phương pháp dựa trên GA. Do có mối quan hệ cân bằng giữa tác dụng phụ của FTH và NTH, nên thuật toán đề xuất tạo ra nhiều NTH hơn so với các phương pháp tiếp cận dựa trên GA đối với các tập dữ liệu dày đặc nhưng vẫn mang lại kết quả tốt cho các tập dữ liệu thưa thớt. Ngoài ra, so với cách tiếp cận Greedy, thuật toán PSO2DT được thiết kế mang lại kết quả xuất sắc trong mọi trường hợp. Hơn nữa, thuật toán PSO2DT được đề xuất đạt được tính toàn vẹn cơ sở dữ liệu cao hơn so với các phương pháp dựa trên GA trên tất cả các bộ dữ liệu.

Trong bài báo này, một cách tiếp cận dựa trên PSO đã được thiết kế. Tuy nhiên, vẫn có thể thiết kế các thuật toán dọn dẹp dựa trên các phương pháp tiến hóa khác như tối ưu hóa đàn kiến (ACO) và chiến lược tiến hóa thích ứng ma trận hiệp phương sai (CMA-ES) để tìm ra giải pháp tốt hơn cho việc ẩn các tập mục nhạy cảm. Tuy nhiên, việc xác định các công thức để ẩn các tập mục nhạy cảm thông qua việc xóa giao dịch bằng khung ACO là một nhiệm vụ không hề đơn giản (đặc biệt là quy trình cập nhật pheromone). Một cơ hội nghiên cứu khác cho công việc trong tương lai là coi vấn đề bảo vệ quyền riêng tư được thảo luận trong bài viết này là một vấn đề tối ưu hóa đa mục tiêu trong đó các tiêu chí như FTH, NTH và DS được coi là một đối tượng. Các thuật toán tiến hóa như GA, PSO, ACO và CMA-ES phù hợp cho vấn đề này. Ngoài ra, cách tiếp cận Pareto sử dụng trong các bài toán tối ưu đa mục tiêu có thể được xem xét để tìm lời giải cân bằng thỏa mãn tiêu chí quan tâm. Hàm vectơ đa thành phần cũng có thể được coi là một giải pháp khác cho PPDM vì nó có thể dễ dàng thực hiện và tích hợp hơn trong phương pháp PSO. Các vấn đề trên sẽ được xem xét trong công việc tương lai của chúng tôi.

Lời cảm ơn

Nghiên cứu này được hỗ trợ một phần bởi Quỹ khoa học tự nhiên quốc gia Trung Quốc (NSFC) theo Grant no. 61503092, bởi Quỹ đổi mới nghiên cứu khoa học tự nhiên tại Viện công nghệ Cáp Nhĩ Tân dưới sự tài trợ của HIT.NSRIF.2014100, và bởi Chương trình các ngành công nghiệp mới nổi chiến lược Thâm Quyến dưới sự tài trợ của ZDSY20120613125016389.

Tài liệu tham khảo

Agrawal, R., Srikant, R., 1994a. Trình tạo dữ liệu tổng hợp Quest. IBM Almaden Re-
Trung tâm tìm kiếm 〈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. Trong: Hội nghị quốc tế về cơ sở dữ liệu rất lớn, San Francisco, CA, Hoa Kỳ, trang 487–499.

Agrawal, R., Srikant, R., 1995. Khai thác các mẫu tuần tự. Trong: Quốc tế
Hội nghị về Kỹ thuật dữ liệu, trang 3–14. Agrawal, R., Srikant, R., 2000. Khai thác dữ liệu bảo vệ quyền riêng tư. ACM SIGMOD Rec.

29, 439–450. Aggarwal, CC, Pei, J., Zhang, B., 2006. Về bảo vệ quyền riêng tư chống lại việc khai thác dữ liệu đối nghịch. Trong: Hội nghị quốc tế ACM SIGKDD về Khám phá tri thức và khai thác dữ liệu, trang 510–516.

Atallah, M., Bertino, E., Elmagarmid, A., Ibrahim, M., Verykios, V., 1999. Tiết lộ
hạn chế của các quy tắc nhạy cảm. Trong: Hội thảo về trao đổi kiến thức và kỹ thuật dữ liệu, trang 45–52.

Bonam, J., Reddy, AR, Kalyani, G., 2014. Bảo đảm quyền riêng tư trong khai thác quy tắc kết hợp bằng cách bóp méo dữ liệu bằng PSO. Trong: CNTT và Cơ sở hạ tầng quan trọng: Kỷ yếu của Hội nghị thường niên lần thứ 48 của Hiệp hội Máy tính Ấn Độ—tập. II, trang 551–558. Chen, MS, Han, J., Yu, PS, 1996. Khai thác dữ liệu: tổng quan từ cơ sở dữ liệu
quan sát. IEEE Trans. Kiến thức. Dữ liệu kỹ thuật số 8 (6), 866–883. Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X., Zhu, MY, 2003. Công cụ bảo mật
bảo tồn khai thác dữ liệu phân tán. ACM SIGKDD Explorer. 4, 1–7. Dasseni, E., Verykios, VS, Elmagarmid, AK, Bertino, E., 2001. Ẩn các luật kết hợp bằng cách sử dụng độ tin cậy và độ hỗ trợ. Trong: Hội thảo quốc tế về che giấu thông tin, trang 369–383.

Dwork, C., McSherry, F., Nissim, K., Smith, A., 2006. Hiệu chỉnh tiếng ồn theo độ nhạy trong phân tích dữ liệu riêng tư. Trong: Bài giảng Khoa học Máy tính, tập. 3876, trang 265–284.

Evfimievski, A., Srikant, R., Agrawal, R., Gehrke, J., 2002. Bảo vệ quyền riêng tư khi khai thác các quy tắc kết hợp. Trong: Hội nghị quốc tế ACM SIGKDD về khám phá tri thức và khai thác dữ liệu, trang 217–228.

Kho lưu trữ tập dữ liệu khai thác tập mục thường xuyên 〈http://fimi.ua.ac.be/data/〉, 2012.

Goldberg, DE, 1989. Thuật toán di truyền trong tìm kiếm, tối ưu hóa và máy

Học hỏi. Nhà xuất bản Addison-Wesley Longman Co., Inc, Boston, MA, USA. Hajian, S., Domingo-Ferrer, J., Farrs, O., 2014. Quyền riêng tư dựa trên khái quát hóa trước

phục vụ và ngăn chặn phân biệt đối xử trong xuất bản và khai thác dữ liệu. Dữ liệu tối thiểu. Kiến thức. Discov. 28 (5–6), 1158–1188.

Han, J., Pei, J., Yin, Y., Mao, R., 2004. Khai thác các mẫu phổ biến mà không tạo ra ứng cử viên: cách tiếp cận cây mẫu thường xuyên. Dữ liệu tối thiểu. Kiến thức. Discov. 8 (1), 53–87.

Han, S., Ng, WK, 2007. Thuật toán di truyền bảo vệ quyền riêng tư để khám phá quy tắc. TRONG:

Ghi chú bài giảng về Khoa học máy tính, trang 407–417. Harik, GR, Lobo, FG, Goldberg, DE, 1999. Thuật toán di truyền nhỏ gọn. IEEE

Chuyển. Tiến hóa. Máy tính. 3 (4), 287–297. Holland, JH, 1992. Thích ứng trong các hệ thống tự nhiên và nhân tạo. Báo chí MIT. Hong, TP, Wang, CY, Tao, YH, 2001. Thuật toán khai thác dữ liệu gia tăng mới

bằng cách sử dụng các tập mục lớn. Trí tuệ. Dữ liệu hậu môn. 5, 111–129. Hong, TP, Lin, CW, Yang, KT, Wang, SL, 2012. Sử dụng TF-IDF để che giấu sự nhạy cảm

itemset. ứng dụng. Trí tuệ. 38 (4), 502–510. Islam, Z., Brankovic, L., 2011. Khai thác dữ liệu bảo vệ quyền riêng tư: một sự bổ sung tiếng ồn

framework sử dụng một kỹ thuật phân cụm mới. Hệ thống dựa trên kiến thức 24 (8), 1214–1223.

Kennedy, J., Eberhart, R., 1995. Tối ưu hóa bầy hạt. Trong: IEEE quốc tế

Hội nghị về Mạng thần kinh, trang 1942–1948. Kennedy, J., Eberhart, R., 1997. Một phiên bản nhị phân rời rạc của thuật toán bầy hạt

ritm. Trong: Hội nghị quốc tế của IEEE về Hệ thống, Con người và Điều khiển học, trang 4104–4108.

Kuo, RJ, Chao, CM, Chiu, YT, 2011. Ứng dụng tối ưu hóa bầy hạt để

khai phá luật kết hợp. ứng dụng. Máy tính mềm. 11(1), 326–336. Lindell, Y., Pinkas, B., 2000. Khai thác dữ liệu bảo đảm quyền riêng tư. Trong: Báo cáo thường niên

Hội nghị Mật mã học quốc gia về những tiến bộ trong Mật mã học, trang 36–54. Lin, CW, Hong, TP, Chang, CC, Wang, SL, 2013. Cách tiếp cận dựa trên lòng tham đối với

ẩn các tập mục nhạy cảm bằng cách chèn giao dịch. J. Inf. Ẩn Multimed. Quá trình tín hiệu. 4, 201–227.

Lin, CW, Hong, TP, Yang, KT, Wang, SL, 2015a. Các thuật toán dựa trên GA cho

tối ưu hóa việc ẩn các tập mục nhạy cảm thông qua việc xóa giao dịch. ứng dụng. Trí tuệ. 42 (2), 210–230.

Lin, CW, Yang, L., Fournier-Viger, P., Wu, MT, Hong, TP, Wang, SL, 2015b. MỘT

cách tiếp cận dựa trên bầy đàn để khai thác các tập mục có tính tiện ích cao. Đa ngành. Sóc. Mạng. Nghị quyết, 572–581.

Lin, CW, Zhang, B., Yang, KT, Hong, TP, 2014. Ẩn hiệu quả các tập mục nhạy cảm bằng cách xóa giao dịch dựa trên thuật toán di truyền. Khoa học. Thế giới J., 13 (ID bài viết 398269).

Menhas, MI, Fei, M., Wang, L., Fu, X., 2011. Một thuật toán PSO nhị phân lai mới. TRONG:

Ghi chú bài giảng về Khoa học máy tính, tập. 6728, trang 93–100. Microsoft. Cơ sở dữ liệu mẫu Foodmart của Dịch vụ phân tích của Microsoft 〈http://msdn.

microsoft.com/en-us/library/aa217032(SQL.80).aspx〉. Murty, MN, Flynn, PJ, 1999. Phân cụm dữ liệu: đánh giá. Máy tính ACM. Sống sót. 31 (3),

264–323. Mooney, CH, Roddick, JF, 2013. Khai thác mẫu tuần tự—các phương pháp tiếp cận và

các thuật toán. Máy tính ACM. Sống sót. 45 (2), 1–39. Oliveira, SRM, Zaane, Osmar R., 2002. Bảo mật quyền riêng tư trong việc khai thác tập mục thường xuyên.

Trong: Hội nghị quốc tế của IEEE về quyền riêng tư, bảo mật và khai thác dữ liệu, trang 43–54.

Pandya, BK, Dixit, K., Singh, UK, Bunkar, K., 2014. Hiệu quả của nhiễu loạn dữ liệu nhân lên để bảo vệ quyền riêng tư trong khai thác dữ liệu. Int. J. Khuyến cáo. Res. Máy tính. Khoa học. 5 (6), 112–115.

Pears, R., Koh, YS, 2012. Khai thác quy tắc kết hợp có trọng số bằng cách sử dụng tối ưu hóa bầy đàn hạt. Trong: Bài giảng Khoa học Máy tính, tập. 7104, trang 327–338.

Quinlan, JR, 1993. C4.5: Các chương trình dành cho học máy. Quán rượu Morgan Kaufmann-

lishers Inc, San Francisco, CA, Hoa Kỳ. Sweeney, L., 2002. k-anonymity: một mô hình bảo vệ quyền riêng tư. Int. J. Không chắc chắn.

Hệ thống dựa trên kiến thức mờ, 557–570. Sarath, KNVD, Ravi, V., 2013. Khai thác quy tắc kết hợp bằng cách sử dụng bầy hạt nhị phân

tối ưu hóa. Anh. ứng dụng. Nghệ thuật. Trí tuệ. 26, 1832–1840. Shen, M., Zhan, ZH, Chen, WN, Gong, YJ, Zhang, J., Li, Y., 2014. Hai vận tốc rời rạc

Tối ưu hóa bầy đàn hạt và ứng dụng của nó vào vấn đề định tuyến multicast trong mạng truyền thông. IEEE Trans. Điện tử Ấn Độ. 61 (12), 7141–7151. Tian, Y., Liu, D., Yuan, D., Wang, K., 2013. Một PSO rời rạc cho lắp ráp hai giai đoạn

vấn đề lập kế hoạch. Int. J. Adv. Technol. 66 (1–4), 481–499. Verykios, VS, Bertino, E., Fovino, IN, Provenza, LP, Saygin, Y., Theodoridis, Y.,

2004. Công nghệ khai thác dữ liệu tiên tiến nhất đảm bảo quyền riêng tư. ACM SIGMOD Rec. 33, 50–57.

Wu, YH, Chiang, CM, Chen, ALP, 2007. Ẩn các quy tắc kết hợp nhạy cảm với

tác dụng phụ hạn chế. IEEE Trans. Kiến thức. Dữ liệu kỹ thuật số 19, 29–42. Zaki, M., 2001. SPADE: một thuật toán hiệu quả để khai thác chuỗi phổ biến. Mach.

Học hỏi. 42 (1–2), 31–60. Zuo, X., Zhang, G., Tan, W., 2014. Thời hạn học tập dựa trên PSO tự thích ứng

lập lịch tác vụ căng thẳng cho đám mây IaaS lai. IEEE Trans. Tự động. Khoa học. Anh. 11 (2), 564–573.

Zhi, XH, Xing, XL, Qang, QX, Zhang, LH, 2004. Một phương pháp PSO rời rạc cho bài toán TSP tổng quát. Trong: Hội nghị quốc tế của IEEE về Học máy và Điều khiển học, trang 2378–2383.