TẠP CHÍ KHOA HỌC ĐẠI HỌC ĐÀ LẠT Tập 8, Số 2, 2018 124–138

PHÁT HIỆN TẬP PHỔ BIẾN GÂY NHẦM LẪN

Huỳnh Thành Lộc[a*]

[a]*Khoa Công nghệ Thông tin, Trường Đại học Ngoại ngữ - Tin học TP. Hồ Chí Minh, TP. Hồ Chí Minh, Việt Nam*
[*]*Tác giả liên hệ: Email: lochuynh@huflit.edu.vn*

Lịch sử bài báo
Nhận ngày 15 tháng 03 năm 2018
Chỉnh sửa ngày 04 tháng 06 năm 2018 | Chấp nhận đăng ngày 13 tháng 06 năm 2018

Tóm tắt

Khai thác tập phổ biến là một trong những hướng nghiên cứu quan trọng trong lĩnh vực khai thác luật kết hợp. Việc khai thác tập phổ biến ở các mức độ tổng quát khác nhau của dữ liệu sẽ đem lại nhiều tri thức có giá trị. Tuy nhiên, trong các tập phổ biến tổng quát đó có thể tồn tại những tập phổ biến phản ảnh tri thức trái ngược so với những tri thức mà các tập phổ biến con của nó phản ánh. Những tập phổ biến như vậy được gọi là tập phổ biến gây nhầm lẫn. Việc xác định được các tập phổ biến gây nhầm lẫn giúp cho các nhà phân tích có thêm cơ sở để đưa ra những lời khuyến nghị chính xác hơn. Bài viết này sẽ giới thiệu khái niệm tập phổ biến gây nhầm lẫn, nghiên cứu việc áp dụng các kỹ thuật khai thác tập phổ biến hiện có vào bài toán khai thác tập phổ biến gây nhầm lẫn và định nghĩa độ đo dùng để đánh giá độ lý thú của một tập phổ biến gây nhầm lẫn.

Từ khóa: Cây phân loại; Khai thác dữ liệu; Sự tương quan; Tập phổ biến.

Mã số định danh bài báo: http://tckh.dlu.edu.vn/index.php/tckhdhdl/article/view/440

Loại bài báo: Bài báo nghiên cứu gốc có bình duyệt

DISCOVERING CONFUSING FREQUENT ITEMSETS

Huynh Thanh Loc[a*]

[a]*The Faculty of Information Technology, Hochiminh City University of Foreign Languages - Information Technology, Hochiminh City, Vietnam*
[*]*Corresponding author: Email: lochuynh@huflit.edu.vn*

Article history
Received: March 15th, 2018
Received in revised form: June 04th, 2018 | Accepted: June 13th, 2018

Abstract

Frequent itemset mining is one of the most important research area in the field of association rule mining. Exploiting frequent itemsets at different abstraction levels of data will yield valuable knowledge. However, some Confusing Frequent Itemsets (CFIs) could be included in the mined set. These CFIs represent contrasting knowledge with their low-level descendants. Experts need to analyze CFIs from traditional frequent itemsets to make more accurate recommendations. In this paper we presented a definition of a CFI, CFI's interestingness measure and how to apply existing frequent itemset mining techniques to discover CFIs from data by exploiting a taxonomy.

Keywords: Correlation; Data mining; Frequent itemset; Taxonomy.

Article identifier: http://tckh.dlu.edu.vn/index.php/tckhdhdl/article/view/440

Article type: (peer-reviewed) Full-length research article

1. GIỚI THIỆU

Khai thác tập phổ biến được khởi xướng bởi Agrawal và Srikant (1994). Hướng nghiên cứu này đóng vai trò quan trọng trong lĩnh vực khai thác luật kết hợp và ngày càng được nhiều nhà nghiên cứu quan tâm. Các tập phổ biến là cơ sở để các chuyên gia đưa ra những thông tin dự đoán từ dữ liệu dựa trên mối quan hệ giữa các hạng mục xuất hiện trong tập phổ biến đó. Để đánh giá mối quan hệ này, nghiên cứu của Brin, Motwani, và Silverstein (1997) dùng độ đo sự tương quan χ^2. Tuy nhiên, độ đo χ^2 là độ đo có giá trị bị ảnh hưởng bởi kích thước của tập dữ liệu. Chính vì vậy, các nghiên cứu của Tan, Kumar, và Srivastava (2002); và Wu, Chen, và Han (2007, 2010) đã chỉ ra những độ đo tương quan bất biến với những giao dịch rỗng. Đây là những độ đo mà giá trị của chúng không bị ảnh hưởng bởi kích thước của tập dữ liệu, giúp việc đánh giá sự tương quan giữa các hạng mục trong một tập phổ biến chính xác hơn. Tùy theo độ đo được sử dụng mà phạm vị giá trị của độ tương quan sẽ khác nhau, nhưng nhìn chung sự tương quan giữa các hạng mục trong một tập phổ biến thường được chia làm ba nhóm: i) Tương quan dương; ii) Tương quan âm; và iii) Không tương quan.

Việc khai thác tập phổ biến ở mức thấp có thể bỏ qua nhiều tri thức có giá trị mà khi phân tích dữ liệu ở các mức độ tổng quát hơn mới có thể thấy được nên các nhà nghiên cứu tiếp tục đề xuất việc khai thác tập phổ biến ở các mức độ tổng quát khác nhau dựa trên cây phân loại của của dữ liệu nhằm khám phá ra thêm nhiều tri thức đáng giá ở mức cao hơn (Srikant & Agrawal, 1995). Gần đây, các nghiên cứu của Barsky, Kim, Weninger, và Han (2011); và Cagliero, Cerquitelli, Garza, và Grimaudo (2014) sử dụng một độ đo tương quan bất biến với những giao dịch rỗng gọi là Kulczynsky để chỉ ra các tập phổ biến tổng quát có kiểu tương quan trái ngược so với các tập phổ biến ở mức thấp hơn.

Khi một tập phổ biến tổng quát có kiểu tương quan trái ngược so với các tập phổ biến ở mức thấp hơn nghĩa là chúng đang phản ánh các xu hướng trái ngược nhau của dữ liệu. Nếu sử dụng tập phổ biến tổng quát này để đưa ra những nhận định mà không xem xét đến các tập phổ biến cụ thể ở mức thấp hơn thì có thể những nhận định đó sẽ thiếu chính xác. Những tập phổ biến như thế gọi là tập phổ biến gây nhầm lẫn. Việc xác định các tập phổ biến gây nhầm lẫn khi khai thác dữ liệu ở nhiều mức độ tổng quát khác nhau là cần thiết, nó giúp cho các nhà phân tích có thêm cơ sở để đưa ra những nhận định chính xác hơn.

Bài báo này sẽ giới thiệu khái niệm tập phổ biến gây nhầm lẫn, đồng thời nghiên cứu việc áp dụng các kỹ thuật khai thác tập phổ biến hiện có vào bài toán khai thác tập phổ biến gây nhầm lẫn, định nghĩa độ đo dùng để đánh giá độ lý thú của một tập phổ biến gây nhầm lẫn và cài đặt thực nghiệm trên bộ dữ liệu UCI (Dheeru & Karra, 2017). Phần còn lại của bài báo này được chia thành bốn mục: Mục 2 trình bày các khái niệm cơ bản trong khai thác tập phổ biến, giới thiệu bài toán và phương pháp khai thác tập phổ biến gây nhầm lẫn từ dữ liệu; Mục 3 cài đặt thực nghiệm; Mục 4 kết quả và thảo luận; và Mục 5 là kết luận và hướng phát triển.

2. NỘI DUNG NGHIÊN CỨU

2.1. Các kiến thức cơ bản

Gọi $B=\{i_1,i_2,...,i_m\}$ là một cơ sở hạng mục mà mỗi hạng mục có thể là một sản phẩm, một dịch vụ hay giá trị của một thuộc tính nào đó. Một tập $X=\{i_1,...,i_k\}\subseteq B$ được gọi là tập hạng mục. Tập chứa k hạng mục được gọi là tập k hạng mục. Một giao dịch trên B ký hiệu là $T_i=(tid, I)$ với tid là mã định danh của giao dịch và I là tập hạng mục có trong giao dịch đó. Một giao dịch được gọi là chứa tập hạng mục X nếu $X\subseteq I$. Một cơ sở dữ liệu giao tác D là một tập các giao dịch trên B: $D=\{T_1,T_2,...,T_n\}$.

Tập bao phủ của tập hạng mục X trong D là tập hợp tất cả các giao dịch chứa I trong D: $cover(X,D):=\{tid|(tid,I)\in T,X\subseteq I\}$. Độ phổ biến của một tập hạng mục X trong D là tổng số giao dịch trong tập bao phủ của X: $support(X,D):=|cover(X,D)|$.

Bảng 1. Cơ sở dữ liệu giao tác *D*

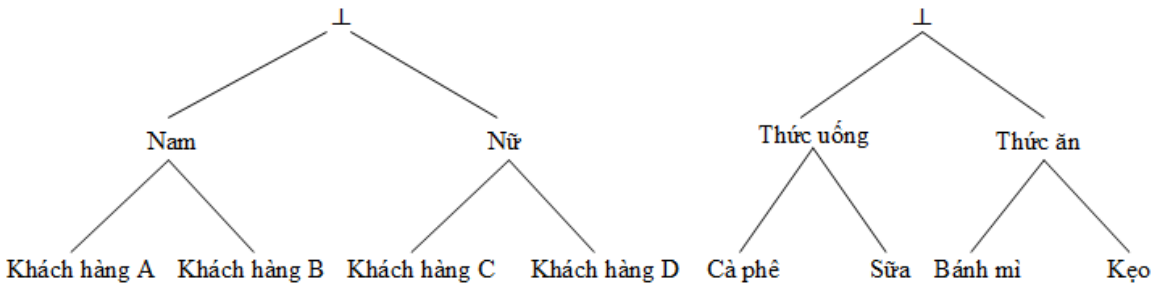
TID	Khách hàng	Sản phẩm
T_1	Khách hàng A	Cà phê
T_2	Khách hàng A	Cà phê
T_3	Khách hàng A	Sữa
T_4	Khách hàng B	Bánh mì
T_5	Khách hàng B	Bánh mì
T_6	Khách hàng C	Kẹo
T_7	Khách hàng A	Cà phê

Trong cơ sở liệu D ở Bảng 1, *B = {Khách hàng A, Khách hàng B, Cà phê, Sữa, Bánh mì, Kẹo}* là một cơ sở hạng mục. Trong đó, *X = {Khách hàng A, Cà phê}* là một tập hai hạng mục và *T1, T2, T7* là các giao dịch chứa tập hạng mục X. Như vậy, *cover(X,D) = {T1, T2, T7}* và *support(X,D) = 3*.

Trên một tập dữ liệu D cụ thể thì *support(X,D)* có thể viết gọn là *sup(X)*. Số lượng giao dịch trong D được ký hiệu là: $|D|$. Một tập hạng mục được gọi là phổ biến (gọi tắt là tập phổ biến) khi độ phổ biến của tập hạng mục đó không nhỏ hơn một ngưỡng phổ biến tối thiểu *minsup* ($0\leq minsup\leq|D|$) do người dùng đặt ra: Tập hạng mục X phổ biến khi $support(X,D)\geq minsup$. Trong cơ sở liệu D ở Bảng 1, nếu chọn *minsup = 2*, thì có 2 tập phổ biến là *{Khách hàng A, Cà phê} (sup = 3)* và *{Khách hàng B, Bánh mì} (sup = 2)*.

Cây phân loại là tập các cây phân cấp được xây dựng dựa trên các biến của dữ liệu. Một thuộc tính có thể có một hay nhiều cây phân cấp mà tại đó những giá trị thuộc cùng một miền giá trị sẽ được tổng hợp thành khái niệm mức cao hơn. Hình 1 minh họa

cây phân loại của hai biến *Khách hàng* và *Sản phẩm* dựa trên cơ sở dữ liệu được cho ở Bảng 1.



Hình 1. Cây phân loại trên thuộc tính *Khách hàng* và *Sản phẩm*

Mối quan hệ "Cha – Con" của hai tập hạng mục: Cho hai tập hạng mục I_1 và I_2, I_1 được gọi là con của I_2 nếu với mọi hạng mục $i_j \epsilon I_1$ đều tồn tại hạng mục $i_k \epsilon I_2$ sao cho $i_j = i_k$ hoặc i_j là con của i_k dựa trên cây phân loại. Khi đó, I_2 được gọi là cha của I_1. Ví dụ: tập hạng mục *{Khách hàng A, Bánh mì}* là con của tập hạng mục *{Nam, Thức ăn}*.

2.2. Bài toán khai thác tập phổ biến gây nhầm lẫn

Độ tương quan dùng để đánh giá mối liên hệ giữa các hạng mục trong một tập hạng mục. Các công trình nghiên cứu của Tan và ctg. (2002); và Wu và ctg. (2007, 2010) đã chỉ ra những độ đo thích hợp cho việc đánh giá sự tương quan giữa các hạng mục trong một cơ sở dữ liệu giao tác lớn. Mặc dù có nhiều độ đo nhưng nhìn chung sự tương quan giữa các hạng mục trong một tập hạng mục thường được chia làm ba nhóm:

- *Tương quan âm (-)*: Nếu độ tương quan có giá trị nhỏ hơn hoặc bằng một ngưỡng tối đa tương quan âm cho trước *(max_neg_cor)*;
- *Tương quan dương (+)*: Nếu độ tương quan có giá trị lớn hơn hoặc bằng một ngưỡng tối thiểu tương quan dương cho trước (*min_pos_cor*);
- *Không tương quan*: Nếu không thuộc về hai kiểu trên.

Hai tập phổ biến I_1 và I_2 được gọi là có kiểu tương quan trái ngược nhau khi I_1 có kiểu tương quan dương và I_2 có kiểu tương quan âm hoặc ngược lại I_1 có kiểu tương quan âm và I_2 có kiểu tương quan dương.

Định nghĩa tập phổ biến gây nhầm lẫn (*Confusing Frequent Itemset - CFI*): Cho một tập phổ biến X và tập *Desc(X)* chứa các tập phổ biến con của X. Tập phổ biến X được gọi là tập phổ biến gây nhầm lẫn nếu tồn tại ít nhất một tập phổ biến thuộc *Desc(X)* có kiểu tương quan trái ngược với X. Mỗi CFI là một mẫu được ký hiệu $X \rhd \varepsilon$, với X là tập phổ biến gây nhầm lẫn và ε là tập các tập phổ biến con của X mà có kiểu tương quan trái ngược với X.

Ví dụ: Với tập dữ liệu *D* ở Bảng 1 và cây phân loại Γ như Hình 1, nếu dùng độ đo Kulczynsky (Kulc) (Wu & ctg., 2010) để đánh giá sự tương quan giữa các hạng mục trong một tập hạng mục. Ta có công thức tính độ đo Kulc như sau:

$$Kulc(I)=\frac{1}{k}\sum_{j=1}^{k}\frac{sup(I,D)}{sup(i_j,D)} \quad (1)$$

Trong đó: *sup(I,D)* là độ hỗ trợ của tập hạng mục *I* trong *D.*; i_j [$1 \le j \le k$] là hạng mục thứ *j* trong *I.*

Cho ngưỡng *min_sup=1*, *max_neg_cor=0.65* và *min_pos_cor=0.8*. Có thể tìm được tập phổ biến *X* ở mức 2: *{(Khách hàng, Nam), (Sản phẩm, Thức uống)}.* Ta tính độ tương quan của tập *X*:

$$sup(\{(Khách\ hàng,\ Nam),\ (Sản\ phẩm,\ Thức\ uống)\},D)=\frac{4}{7}$$

$$sup(\{(Khách\ hàng,\ Nam)\},D)=\frac{6}{7};\ sup(\{(Sản\ phẩm,\ Thức\ uống)\},D)=\frac{4}{7}$$

$$Kulc(X)=\frac{1}{2}\left(\frac{4/7}{6/7}+\frac{4/7}{4/7}\right)=\frac{5}{6}=0.83>min\_pos\_cor$$

Do đó, tập phổ biến *X* có kiểu *tương quan dương*.

Tương tự, ta có các tập phổ biến con của *X* ở mức 1: I_1{*(Khách hàng, Khách hàng A), (Sản phẩm, Cà phê)*} với $Kulc(I_1)=\frac{7}{8}=0.88>min\_pos\_cor$ (nên I_1 có kiểu *tương quan dương*) và I_2: *{(Khách hàng, Khách hàng A), (Sản phẩm, Sữa)}* với $Kulc(I_2)=\frac{5}{8}=0.63<max\_neg\_cor$ (nên I_2 có kiểu *tương quan âm*).

Vì *X* và I_2 có kiểu tương quan trái ngược nhau nên *X* là một tập phổ biến gây nhầm lẫn có dạng:

{(Khách hàng, Nam), (Sản phẩm, Thức uống)} ⊳ *{(Khách hàng, Khách hàng A), (Sản phẩm, Sữa)}*

Bảng 2 liệt kê các tập phổ biến ở mức 1 và mức 2 và các CFI tìm được từ tập dữ liệu *D* với ngưỡng *min_sup=1*, *max_neg_cor=0.65* và *min_pos_cor=0.8*.

Bảng 2. Các CFI từ tập dữ liệu *D*

Tập phổ biến ở mức 2 [Kiểu tương quan (giá trị Kulc)]	Tập phổ biến con ở mức 1 [Kiểu tương quan (giá trị Kulc)]	NOD (%)
{(Khách hàng, Nam)} [+ (1)]	{(Khách hàng, Khách hàng A)} [+ (1)] {(Khách hàng, Khách hàng B)} [+ (1)]	-
{(Khách hàng, Nữ)} [+ (1)]	{(Khách hàng, Khách hàng C)} [+ (1)]	-
{(Sản phẩm, Thức uống)} [+ (1)]	{(Sản phẩm, Sữa)} [+ (1)] {(Sản phẩm, Cà phê)} [+ (1)]	-
{(Sản phẩm, Thức ăn)} [+ (1)]	{(Sản phẩm, Bánh mì)} [+ (1)] {(Sản phẩm, Kẹo)} [+ (1)]	-
{(Khách hàng, Nữ), (Sản phẩm, Thức ăn)} [không tương quan (2/3=0.66)	{(Khách hàng, Khách hàng C), (Sản phẩm, Kẹo)} [+ (1)]	-
{(Khách hàng, Nam), (Sản phẩm, Thức ăn)} [- (1/2=0.5)]	{(Khách hàng, Khách hàng B), (Sản phẩm, Bánh mì)} [+ (1)]	0.00
{(Khách hàng, Nam), (Sản phẩm, Thức uống)} [+ (5/6=0.83)]	{(Khách hàng, Khách hàng A), (Sản phẩm, Sữa)} [- (5/8=0.63)] {(Khách hàng, Khách hàng A), (Sản phẩm, Cà phê)} [+ (7/8=0.88)]	0.75

Về mặt ý nghĩa, có thể rút ra nhận xét: Những khách hàng giới tính nam có mối tương quan dương với các sản phẩm thuộc nhóm Thức uống. Tuy nhiên, khi nhìn ở một mức cụ thể hơn thì những khách hàng này có mối tương quan dương với các sản phẩm Cà phê nhưng lại có mối tương quan âm đối với các sản phẩm Sữa. Từ nhận xét này, các chuyên gia có thể đưa ra những lời khuyến nghị chính xác và phù hợp với thực tế hơn. Độ thú vị của một CFI $X \rhd \varepsilon$ được đánh giá dựa trên mức độ trùng lắp giữa các tập phổ biến trong ε so với tập phổ biến X. Mức độ trùng lắp càng lớn nghĩa là càng có nhiều tập phổ biến con của X có sự tương quan trái ngược với X, CFI đó càng thể hiện rõ sự nhầm lẫn của mình.

Định nghĩa độ đo NOD (*Not Overlapping Degree*): Với một CFI có dạng $X \rhd \varepsilon$. Gọi $sup(X,D)$ là độ hỗ trợ của X trong tập dữ liệu D và $cov(\varepsilon,D)$ là độ phủ của ε trong tập dữ liệu D. Độ đo NOD được định nghĩa bởi công thức (2).

$$NOD(X \rhd \varepsilon) = \frac{sup(X,D) - cov(\varepsilon,D)}{sup(X,D)} \quad (2)$$

Vì $sup(X,D) - cov(\varepsilon,D) \geq 0$ nên giá trị của độ đo NOD thuộc luôn thuộc đoạn [0, 1]. NOD càng nhỏ thì mức độ trùng lắp càng lớn, CFI càng có giá trị. Để hạn chế số CFI được sinh ra và tìm được các CFI thật sự đáng giá, chúng ta thường chỉ xem xét những CFI có giá trị NOD tương đối nhỏ (nhỏ hơn một ngưỡng *max_NOD* nào đó) vì những CFI này thể hiện rõ sự nhầm lẫn đối với tri thức mà nó phản ánh.

2.3. Thuật toán khai thác tập phổ biến gây nhầm lẫn

- *Đầu vào*: *D*, *Γ*, *minsup*, *max_neg_cor*, *min_pos_cor*, *max_NOD*;
- *Đầu ra*: CFIs;
- *Chi tiết thuật toán*:

```
1   F = Tập các tập phổ biến ở các mức khác nhau của dữ liệu
2   CFIs= Ø;
3   /*Sinh tập ứng viên CFI*/
4   for l=2 to maxlevel do
5           for all X in F[l] do
6                   Thêm ứng viên CFI có dạng X⊳ε vào tập C[l];
7           end for
8           for all it in F[l-1] do
9                   genit = Cha của tập phổ biến it.
10                  cor_type_genit = Kiểu tương quan của genit;
11                  cor_type_it = Kiểu tương quan của it;
12                  if cor_type_genit ≠ cor_type_it then
13                          Thêm it vào tập ε của genit;
14                  end if
15          end for
16          for all c in C[l] do
17                  c.NOD = Giá trị NOD của ứng viên CFI c.
18                  if c.NOD≤max_NOD them
19                          Thêm c vào tập kết quả CFIs[l];
20                  end if
21          end for
22  end for
23  return CFIs
```

2.3.1. Khai thác tập phổ biến theo từng mức độ tổng quát trên cây phân loại

Từ tập dữ liệu ban đầu, tập dữ liệu tương ứng với từng mức sẽ được tạo ra dựa trên cây phân loại. Hiện đã có rất nhiều thuật toán khai thác tập phổ biến đã được đề xuất (Fournier và ctg., 2017). Nghiên cứu này sử dụng một số thuật toán nổi tiếng và được đánh giá tốt như: FP-Growth (Han, Pei, & Yin, 2000); LCMv2 (Uno, Kiyomi, & Arimura, 2004); Eclat, dEclat (Zaki & Gouda, 2003) để khai thác tập phổ biến trên các tập dữ liệu này. Kết quả của bước này là một tập *F* chứa tất cả các tập phổ biến ở theo từng mức tương ứng với cây phân loại (Dòng 1).

2.3.2. Xác định các CFI

Sau khi đã khai thác tập phổ biến theo từng mức độ tổng quát dựa trên cây phân loại. Tại mỗi mức *l≥2*, các ứng viên CFI sẽ được tạo ra dưới dạng *X⊳ε*, nếu ứng viên nào có giá trị NOD *≤max_NOD* thì sẽ đưa vào tập kết quả (Dòng 4-22).

Cụ thể, tại mức *l*, tất cả các tập phổ biến *X* ở mức *l* sẽ được đưa vào tập ứng viên *C[l]* (Dòng 5 - 7). Các tập phổ biến *it* ở mức *(l-1)* sẽ được xem xét và xác định kiểu tương quan. Nếu *it* có kiểu tương quan trái ngược với *genit* (*genit*$\in$*C[l]*) thì *it* sẽ được thêm vào tập ε của *genit* (Dòng 8 - 15). Cuối cùng, nếu các ứng viên trong tập *C[l]* có giá trị NOD $\leq$ *max_NOD* thì sẽ đưa vào tập kết quả (Dòng 16 - 21).

2.3.3. Ví dụ minh họa

Với tập dữ liệu *D* ở Bảng 1 và cây phân loại Γ như Hình 1, tập dữ liệu ở mức 2 sẽ được tạo ra bằng cách thay thế các hạng mục ở mức 1 bằng hạng mục ở mức 2 tương ứng trên cây phân loại.

Bảng 3. Tập dữ liệu ở mức 2 của tập dữ liệu *D*

TID	Khách hàng	Sản phẩm
T_1	Nam	Thức uống
T_2	Nam	Thức uống
T_3	Nam	Thức uống
T_4	Nam	Thức ăn
T_5	Nam	Thức ăn
T_6	Nữ	Thức ăn
T_7	Nam	Thức uống

Lần lượt áp dụng thuật toán khai thác tập phổ biến trên tập dữ liệu ở mức 1 và mức 2 của tập dữ liệu *D* sẽ được tập hợp các tập phổ biến ở mức 1 (*F[1]*) và mức 2 (*F[2]*) như trong Bảng 2. Sau đó tất cả các tập phổ biến thuộc *F[2]* sẽ được đưa vào tập ứng viên CFI mức 2 *C[2]*. Với mỗi tập phổ biến *it* trong *F[1]* sẽ xác định kiểu tương quan *it* và tập phổ biến cha tương ứng của *it* (*genit*) trong *F[2]*. Nếu chúng có kiểu tương quan trái ngược nhau thì thêm *it* vào tập ε của *genit*. Ví dụ cụ thể về việc xác định kiểu tương quan của một tập hạng mục đã được trình bày cụ thể trong Mục 2.2.

Cuối cùng, tính độ đo NOD của các ứng viên CFI trong tập *C[2]*. Ứng viên nào có giá trị NOD $\leq$*max_NOD* sẽ được đưa vào tập kết quả. Kết quả các CFI khai thác được từ tập dữ liệu *D* được cho trong Bảng 2.

3. CÀI ĐẶT THỰC NGHIỆM

Nghiên cứu này cài đặt thực nghiệm thuật toán khai thác các CFI và áp dụng trên các tập dữ liệu đã được đánh giá từ UCI (Dheeru & Karra, 2017). Để khai thác được các CFI từ các tập dữ liệu UCI cần phải xây dựng cây phân loại của tập dữ liệu. Cây phân loại được xây dựng như sau:

- Đối với các biến định danh: Cây phân loại do nhà phân tích cung cấp;
- Đối với các biến liên tục: Cây phân loại trên các biến liên tục được xây dựng bằng cách rời rạc hóa dữ liệu theo kiểu bằng nhau về tần số (Tan, Steinbach, & Kumar, 2005).

Kết quả khai thác CFI từ các tập dữ liệu UCI được trình bày trong Bảng 4. Trong Bảng 4, cây phân loại được xây dựng gồm 3 mức, các biến liên tục được rời rạc hóa thành 10 nhóm ở mức 1, 5 nhóm ở mức 2 và tất cả được gom thành 1 nhóm ở mức 3. Những biến không có ý nghĩa phân loại sẽ được tổng hợp trực tiếp tới nút gốc.

Bảng 4. Kết quả khai thác CFI từ các tập dữ liệu UCI

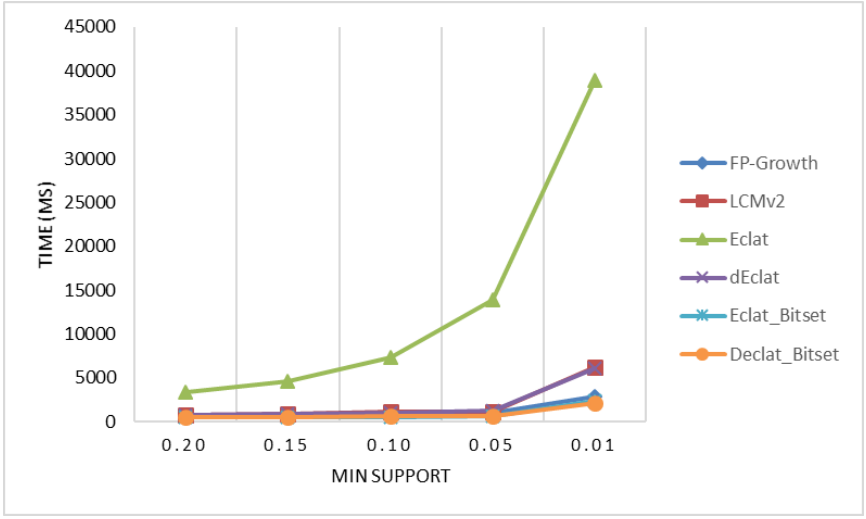
Tập dữ liệu	Số giao dịch	Số thuộc tính	max_NOD (%)	Số CFI
Adult	32,561	15	1	15
			5	22
Breast	699	11	1	8
			5	32
Glass	214	11	1	14
			5	24
Pima	768	768	1	3
			5	3
Shuttle	43,500	10	1	81
			5	108

Ghi chú: *min_sup = 1*; *max_neg_cor=0.6*; và *min_pos_cor=0.7*.

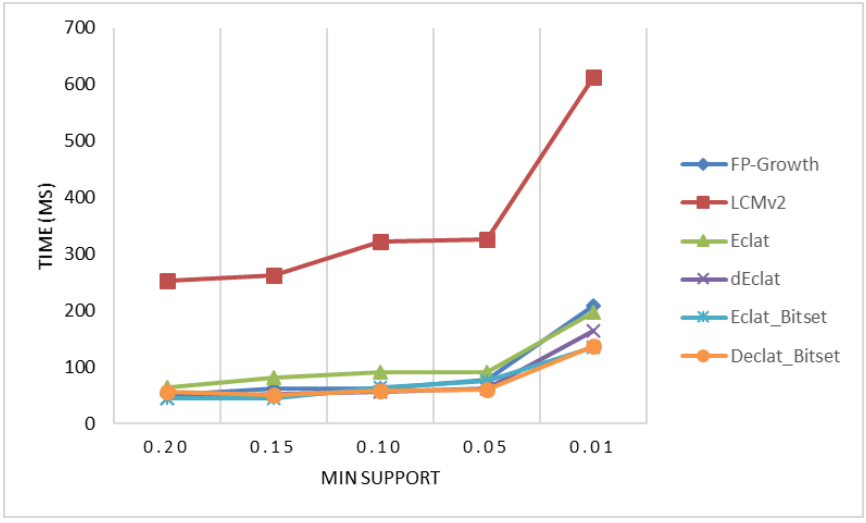
4. KẾT QUẢ VÀ THẢO LUẬN

4.1. So sánh các thuật toán khai tập phổ biến

Hình 2 thể hiện thời gian để xác định tập phổ biến ở từng mức trên hai tập dữ liệu là *Adult* (32,561 giao dịch) và *Breast* (699 giao dịch) bằng nhiều thuật toán khai thác tập phổ biến khác nhau như: FP-Growth (Han & ctg., 2000); LCMv2 (Uno & ctg., 2004); và Eclat, dEclat kết hợp với bitset (Zaki & Gouda, 2003). Nhìn chung, thời gian thực thi của các thuật toán sẽ tăng dần khi giảm ngưỡng *min_sup*. Nhưng nếu so sánh hiệu quả của các thuật toán trên với cùng một ngưỡng *min_sup* thì hai thuật toán *Eclat_Bitset* và *dEclat_Bitset* cho kết quả tốt nhất do không phải duyệt lại cơ sở dữ liệu nhiều lần và sử dụng phép AND trên bitset để tính nhanh độ hỗ trợ của các tập hạng mục.



(a)



(b)

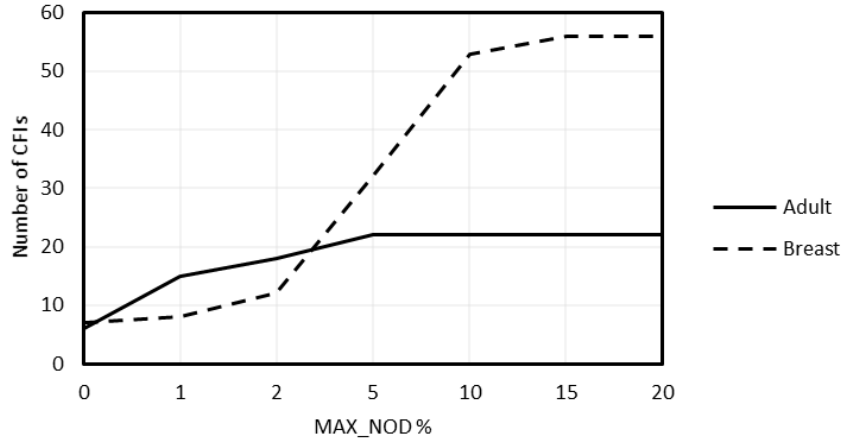
Hình 2. Thời gian thực thi của các thuật toán khai thác tập phổ biến khác nhau

Ghi chú: a) *Adult*; b) *Breast*.

4.2. Phân tích ảnh hưởng của các tham số trong quá trình khai thác CFI

4.2.1. Ảnh hưởng của ngưỡng max_NOD

Ngưỡng *max_NOD* giúp lựa chọn những CFI thật sự có giá trị. Ngưỡng *max_NOD* càng nhỏ thì số lượng CFI sinh ra càng giảm đi, điều này thể hiện rõ trong Hình 3 khi chạy thực nghiệm với các ngưỡng *max_NOD* có giá trị nằm trong khoảng [0, 10%] trên bộ dữ liệu *Adult* và *Breast*.

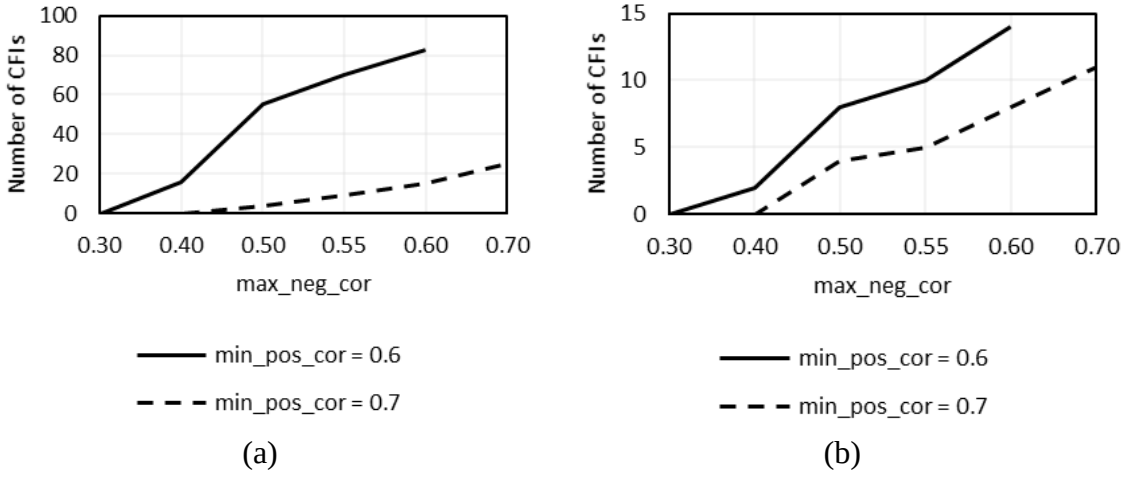


Hình 3. Ảnh hưởng của ngưỡng *max_NOD* đối với số lượng CFI

Ghi chú: *min_sup*=1%; *max_neg_cor*=0.6; và *min_pos_cor*=0.7.

4.2.2. Ảnh hưởng của các ngưỡng tương quan

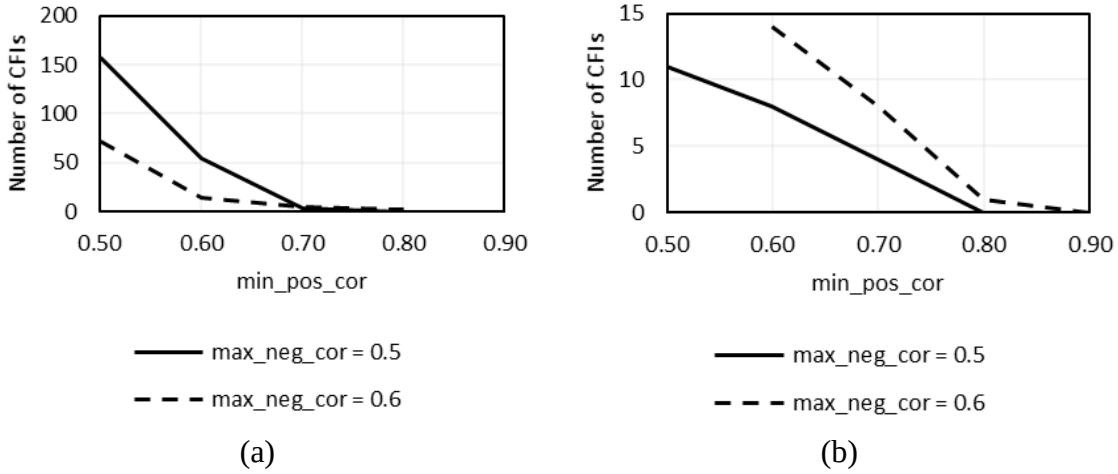
Thay đổi các ngưỡng tương quan cũng làm ảnh hưởng tới số lượng CFI khai thác được. Không nên đặt giá trị của các ngưỡng *max_neg_cor* và *min_pos_cor* quá gần nhau vì khi đó hai tập phổ biến có giá trị tương quan chênh lệch không quá lớn nhưng lại thuộc hai kiểu tương quan khác nhau dẫn đến có nhiều CFI được tìm ra nhưng sẽ có những CFI không đáng giá. Hình 4 và Hình 5 thể hiện sự ảnh hưởng của ngưỡng *max_neg_cor* và *min_pos_cor* đối với số lượng CFI khai thác được.



(a) (b)

Hình 4. Ảnh hưởng của các ngưỡng *max_neg_cor* đối với số lượng CFI

Ghi chú: (a) *Adult*; b) *Breast*; *min_sup*=1%; và *max_NOD*=1%.



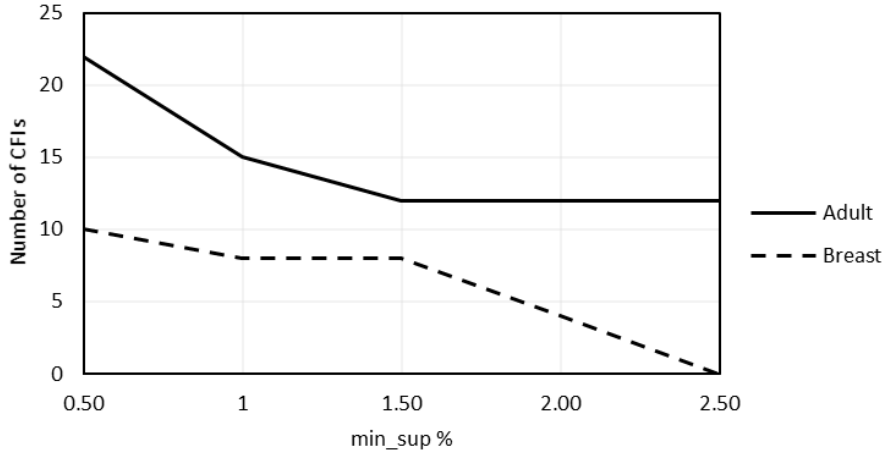
(a) (b)

Hình 5. Ảnh hưởng của các ngưỡng *min_pos_cor* đối với số lượng CFI

Ghi chú: (a) *Adult*; b) *Breast*; *min_sup*=1%; và *max_NOD*=1%.

4.2.3. Ảnh hưởng của ngưỡng min_sup

Do CFI được xác định từ các tập phổ biến khai thác từ dữ liệu nên ngưỡng *min_sup* cũng ảnh hưởng đến số lượng CFI tìm được. Số lượng CFI sẽ càng tăng khi giá trị của ngưỡng *min_sup* càng nhỏ bởi vì việc giảm ngưỡng *min_sup* sẽ kéo theo tăng số lượng các tập phổ biến khai thác được.



Hình 6. Ảnh hưởng của các ngưỡng *min_sup* đối với số lượng CFI

Ghi chú: *max_NOD* = 1%; *max_neg_cor* = 0.6; và *min_pos_cor* = 0.7.

5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Hiện nay, đã có rất nhiều công trình nghiên cứu để giải quyết vấn đề khai thác tập phổ biến từ dữ liệu. Tuy nhiên, đa phần các nghiên cứu tập trung vào việc cải tiến

hiệu quả của các thuật toán khai thác tập phổ biến hay tìm tập phổ biến trên các dạng dữ liệu khác nhau mà có rất ít nghiên cứu về phát hiện những tập phổ biến gây nhầm lẫn từ các tập phổ biến đã khai thác được. Việc xác định được các tập phổ biến gây nhầm lẫn góp phần tạo cơ sở cho những chuyên gia đưa ra các nhận định, phân tích đúng đắn hơn. Nghiên cứu này đã giới thiệu khái niệm và độ đo để đánh giá độ lý thú của một tập phổ biến gây nhầm lẫn và cài đặt thực nghiệm thuật toán khai thác tập phổ biến gây nhầm lẫn trên các tập dữ liệu chuẩn của UCI. Nghiên cứu này có thể được tiếp tục mở rộng theo các hướng sau: (i) Thêm ràng buộc vào quá trình khai thác CFI để tăng hiệu quả của thuật toán; (ii) Sử dụng nhiều ngưỡng tối thiểu khác nhau khi khai thác tập phổ biến ở các mức khác nhau của dữ liệu; và (iii) Song song hóa quá trình khai thác CFI.

TÀI LIỆU THAM KHẢO

Agrawal, R., & Srikant, R. (1994). *Fast algorithms for mining association rules*. Paper presented at The 20th International Conference on Very Large Data Bases, Chile.

Barsky, M., Kim, S., Weninger, T., & Han, J. (2011). *Mining flipping correlations from large datasets with taxonomies*. Paper presented at The 38th International Conference on Very Large Data Bases, Turkey.

Brin, S., Motwani, R., & Silverstein, C. (1997). *Beyond market baskets generalizing association rules to correlations*. Paper presented at The ACM SIGMOD International Conference on Management of Data, USA.

Cagliero, L., Cerquitelli, T., Garza, P., & Grimaudo, L. (2014). Misleading generalized itemset discovery. *Expert Systems with Applications, 41*(4), 1400-1410.

Dheeru, D., & Karra, T. E. (2017). *Machine learning repository*. Retrieved from http:// archive.ics.uci.edu/ml.

Fournier, V. P., Lin, J. C., Vo, B., Truong, C. T., Zhang, J., & Le, H. B. (2017). A survey of itemset mining. *WIREs: Data Mining and Knowledge Discovery, 7*(4), 1-18.

Han, J., Pei, J., & Yin, Y. (2000). *Mining frequent patterns without candidate generation*. Paper presented at The ACM SIGMOD International Conference on Management of Data, Canada.

Srikant, R., & Agrawal, R. (1995). Mining generalized association rules. *Future Generation Computer Systems, 13*(2-3), 161-180.

Tan, P. N., Kumar, V., & Srivastava, J. (2002). *Selecting the right interestingness measure for association patterns*. Paper presented at The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Tan, P. N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining* (2nd ed.). Boston, USA: Pearson Addison Wesley.

Uno, T., Kiyomi, M., & Arimura, H. (2004). *LCM ver. 2: Efficient mining algorithms*

for frequent/closed/maximal itemsets. Paper presented at The IEEE ICDM Workshop Frequent Itemset Mining Implementations, USA.

Wu, T., Chen, Y., & Han, J. (2007). *Association mining in large databases: A re-examination of its measures*. Paper presented at The European Conference on Principles of Data Mining and Knowledge Discovery, Germany.

Wu, T., Chen, Y., & Han, J. (2010). Re-examination of interestingness measures in pattern mining: A unified framework. *Data Mining and Knowledge Discovery*, *21*(3), 371-397.

Zaki, M. J., & Gouda, K. (2003). *Fast vertical mining using diffsets*. Paper presented at The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, USA.