Tạp chí Tin học và Điều khiển học, T.30, S.2 (2014), 163–176

BIỂU DIỄN PHỤ THUỘC HÀM XẤP XỈ THEO PHÂN HOẠCH, MA TRẬN PHÂN BIỆT ĐƯỢC VÀ LUẬT KẾT HỢP

TRẦN DUY ANH

Trường Cao Đẳng Sư Phạm Thừa Thiên Huế; duyanh208@gmail.com

Tóm tắt. Các phụ thuộc hàm xấp xỉ và luật kết hợp là những tri thức thực sự có ý nghĩa trong khai phá dữ liệu. Trong bài báo này, đầu tiên, chúng tôi nhắc lại một số khái niệm cơ bản của lý thuyết tập thô, các độ đo lỗi g_1, g_2, g_3 của phụ thuộc hàm. Sau đó, chúng tôi đề xuất độ đo lỗi g_4 dựa trên phân hoạch và kỳ vọng trong lý thuyết xác suất. Phần tiếp theo chúng tôi xây dựng ma trận phân biệt theo một cách khác và biểu diễn các độ đo lỗi g_1, g_2, độ phụ thuộc γ và ý nghĩa thuộc tính σ theo ma trận phân biệt được. Cuối cùng, chúng tôi đưa ra mối liên hệ giữa phụ thuộc hàm xấp xỉ và luật kết hợp thông qua độ đo lỗi g_4 và độ tin cậy *Confidence*.

Từ khóa. Phụ thuộc hàm xấp xỉ, luật kết hợp

Abstract. Approximate Functional Dependencies (AFD) and Association Rules are really meaningful knowledge in data mining. In this article, we first recall some basic concepts of rough set theory, error measures g_1, g_2 and g_3 for functional dependencies. Then, based on the method of partitions and expectation in probability theory, we propose an error measure g_4 to construct the discernibility matrix in a different way, defined error measures g_1, g_2, dependency degree γ and significance of Attributes σ from the discernibility matrix. Finally, a relationship between AFD and Association Rules via error measure g_4 and confidence is presented.

Key words. Approximate Functional Dependencies, association rules.

1. MỞ ĐẦU

Phụ thuộc hàm xấp xỉ (Approximate Functional Dependencies) là tri thức biểu diễn sự phụ thuộc một phần giữa các thuộc tính. Nó là một mở rộng của phụ thuộc hàm, phụ thuộc hàm xấp xỉ cho phép có một số lượng lỗi nhất định của các bộ dữ liệu đối với phụ thuộc hàm. Để nghiên cứu về loại phụ thuộc này Kivinen, Mannila [5] đã đưa ra các độ đo lỗi g_1, g_2, g_3 đối với phụ thuộc hàm. Sau đó đã có nhiều tác giả nghiên cứu các thuật toán để phát hiện các phụ thuộc hàm xấp xỉ như Huhtala, Karkkainen, Porkka, Toivonnen [4], Stéphane Lopes, Jean-Marc Petit, Lotfi Lakhal [6], Daniel Sánchez, José María Serano, Ignacio Blanco, Maria José Martin-Bautista, María Amparo Vila [7], ... Phụ thuộc hàm xấp xỉ đã có nhiều ứng dụng trong phân tích dữ liệu và đánh giá thông tin như rút gọn các thuộc tính dư thừa[11], tìm kiếm xấp xỉ [3],... Ngoài ra, những tri thức tiềm ẩn trong cơ sở dữ liệu chẳng hạn như: "Khách hàng khi mua sữa và bánh mì thường mua thêm bơ", "Những du khách đến du lịch ở Huế khi mua tôm chua và kẹo mè xững thường mua thêm bánh lọc"... Những tri thức này chính là luật kết hợp (Association Rules). Luật kết hợp được đưa ra bởi nhà nghiên cứu Agrawal và SriKant vào năm 1994 [1] và đã có nhiều thuật toán để phát hiện luật kết hợp như thuật toán Apriori [1], Eclat [8], FP-Growth [12].

Trong bài báo này, đầu tiên, chúng tôi tìm hiểu các độ đo lỗi g_1, g_2, g_3 của Kivinen, Mannila [5]. Sau đó, chúng tôi đề xuất một độ đo lỗi g_4 đối với phụ thuộc hàm và tìm mối liên hệ giữa phụ thuộc hàm xấp xỉ và luật kết hợp thông qua g_4. Tiếp theo, chúng tôi xây dựng ma trận phân biệt được theo một cách khác, từ đó biểu diễn các độ đo lỗi g_1, g_2, độ phụ thuộc γ và ý nghĩa thuộc tính σ thông qua ma trận phân biệt này.

2. MỘT SỐ KHÁI NIỆM CƠ BẢN CỦA LÝ THUYẾT TẬP THÔ

Định nghĩa 2.1. [1, 9] (*Quan hệ không phân biệt được*) Cho $r(R)$. Khi đó, với bất kỳ $X \subseteq R$, tồn tại một quan hệ không phân biệt được $\phi(X)$ trên r được định nghĩa như sau:

$$\forall t,\ u \in r,\ (t,u) \in \phi(X) \Leftrightarrow t[X] = u[X].$$

Định nghĩa 2.2. [1, 9] (*Lớp tương đương và phân hoạch*) Quan hệ $\phi(X)$ sẽ phân hoạch r thành các lớp tương đương. Lớp tương đương của bộ $t \in r$ ứng với tập $X \subseteq R$, ký hiệu $[t]_X$, được định nghĩa như sau:

$$[t]_X = \{u \in r | t[A] = u[A]\ \forall A \in X\},\ [t]_X \neq \emptyset.$$

Khi đó, $\pi_X = \{[t]_X | t \in r\}$ là một phân hoạch của r ứng với X. Lực lượng của π, ký hiệu $|\pi|$, là số lớp tương đương của π.

Cho $U \in \pi_X$. Khi đó, ta quan niệm rằng, U thỏa phụ thuộc hàm $X \to Y$, ký hiệu là $U| = X \to Y$ nếu với mọi t, $u \in U$ sao cho $t[X] = u[X]$, thì $t[Y] = u[Y]$.

Bổ đề 2.1. [4] *$X \to Y$ đúng khi và chỉ khi $|\pi_X| = |\pi_{XY}|$.*

Định nghĩa 2.3. [4] (*Phân hoạch thu gọn*) Phân hoạch thu gọn của π, ký hiệu là $\hat{\pi}$ nếu $\hat{\pi} = \{U \in \pi || U| > 1\}$. Để giảm độ phức tạp tính toán khi làm việc với các phân hoạch, ta dùng các phân hoạch thu gọn thay cho các phân hoạch.

Định nghĩa 2.4. [1] (*Không gian dương*) Không gian dương của tập thuộc tính X ứng với tập thuộc tính Y được định nghĩa như sau:

$$POS(X,Y) = \cup\{U \in \pi_X | \exists V \in \pi_Y : U \subseteq V\}$$

Định nghĩa 2.5. [1] (*Độ phụ thuộc*) Tập thuộc tính Y phụ thuộc vào tập thuộc tính X với mức độ $\gamma(X,Y) \in [0,1]$, ký hiệu là $X \longrightarrow^{\gamma(X,Y)} Y$, trong đó $\gamma(X,Y)$ được xác định như sau:

$$\gamma(X,Y) = \frac{|\,POS\,(X,Y)\,|}{|r|}$$

Định nghĩa 2.6. [9](*Bảng quyết định*) Bảng quyết định $S = (r,R)$ là bảng dữ liệu với các cột tương ứng với tập các thuộc tính R và các hàng là tập các đối tượng (bộ) r. Tập thuộc tính R được phân thành tập thuộc tính điều kiện C và tập thuộc tính quyết định D, $R = C \cup D$, $C \cap D = \emptyset$.

Định nghĩa 2.7. [9] (*Ý nghĩa của thuộc tính*) Ý nghĩa của các thuộc tính đo độ quan trọng của các thuộc tính trong bảng dữ liệu, nghĩa là ta xem xét độ phụ thuộc $\gamma(C,D)$ thay đổi

như thế nào khi ta loại bỏ một thuộc tính A_i khỏi tập thuộc tính điều kiện C. Từ đó, ý nghĩa của thuộc tính A_i được định nghĩa như sau:

$$\sigma_{C\cup D}\left(\mathrm{A_i}\right) = \frac{\gamma\left(C,D\right) - \gamma(C - \{A_i\}\,,D)}{\gamma\left(C,D\right)} = 1 - \frac{\gamma(C - \{A_i\}\ ,D)}{\gamma\left(C,D\right)}$$

Định nghĩa 2.8. [9] (*Ma trận phân biệt được*) Cho $r = \{t_1, t_2, \ldots, t_n\}$. Ma trận phân biệt được của $S = (r, R)$, ký hiệu $M(S) = (m_{ij})_{|r|\times|r|}$ là ma trận đối xứng mà mỗi phần tử của nó là một tập hợp các thuộc tính, được xác định như sau:

$$m_{ij} = \begin{cases} \{A_i \in C | t_i(A_i) \neq t_j(A_i)\} & t_i(D) \neq t_j(D) \\ \emptyset & t_i(D) = t_j(D) \end{cases} \quad \text{với } i, j = \overline{1, n}.$$

3. CÁC ĐỘ ĐO LỖI CỦA PHỤ THUỘC HÀM

Để xác định một phụ thuộc hàm xấp xỉ, Kivinen và Mannila [5] đã đưa ra một số độ đo để tính toán lỗi của một phụ thuộc hàm như sau:

Định nghĩa 3.1. [5] (*Độ đo lỗi g_1*) Cho quan hệ $r(R)$. Khi đó, độ đo lỗi g_1 của một phụ thuộc hàm $X \to Y$ trên r được xác định như sau:

$$g_1(X \to Y, r) = \frac{|\{(t_i, t_j) | t_i, t_j \in r,\ t_i[X] = t_j[X],\ t_i[Y] \neq t_j[Y]\}|}{|r|^2}$$

Định nghĩa 3.2. [5] (*Độ đo lỗi g_2*) Cho quan hệ $r(R)$. Khi đó, độ đo lỗi g_2 của một phụ thuộc hàm $X \to Y$ trên r được xác định như sau:

$$g_2(X \to Y, r) = \frac{|\{\, t_i \mid t_i \in r, \exists t_j \in r : t_i\left[X\right] = t_j\left[X\right] , t_i\left[Y\right] \neq t_j\left[Y\right] \ \} \,|}{|r|}.$$

Định nghĩa 3.3. [5](Độ đo lỗi g_3) Cho quan hệ $r(R)$. Khi đó, độ đo lỗi g_3 của một phụ thuộc hàm $X \to Y$ trên r được xác định như sau:

$$g_3(X \to Y, r) = 1 - \frac{\max\left\{|s| \middle| s \subseteq r, s \middle| = X \to Y\right\}}{|r|}$$

4. BIỂU DIỄN PHỤ THUỘC HÀM XẤP XỈ THEO PHÂN HOẠCH

Độ phụ thuộc γ rất thuận tiện trong việc xem xét hệ tiên đề Armstrong và một số phép toán đại số quan hệ đối với phụ thuộc hàm xấp xỉ trong [1]. Tuy nhiên các thuật toán [4, 10] dùng độ đo lỗi g_3 để phát hiện phụ thuộc hàm xấp xỉ. Trong các thuật toán này độ đo lỗi g_3 được tính theo các phân hoạch dựa vào Bổ đề 2.1 như sau:

Định nghĩa 4.1. [4] (*Độ đo lỗi g_3 theo phân hoạch*) Cho quan hệ $r(R)$. Khi đó, độ đo lỗi của một phụ thuộc hàm $X \to Y$ được xác định như sau:

$$g_3(X \to Y, r) = \frac{|r| - \sum\limits_{U \in \pi_X} \max\left\{\, |V| \,|\, V \in \pi_{XY}\ ,\ V \subseteq U \,\right\}}{|r|}$$

Tính chất 4.1. [10](Mối liên hệ giữa g_3 và γ) Cho độ phụ thuộc

$$\gamma(X,Y) = \frac{|POS\,(X,Y)\,|}{|r|}.$$

và độ đo lỗi $g_3(X \to Y, r)$ của phụ thuộc hàm $X \to Y$. Khi đó, ta có:

$$g_3(X \to Y, r) = 1 - \gamma(X,Y) - \frac{\sum\limits_{U \in \pi_X} \max\{\,|V|\,|\,V \in \pi_{XY}\,, V \subset U\}}{|r|}$$

Định nghĩa 4.2. [10](*Độ đo lỗi g_3 theo phân hoạch thu gọn*) Độ đo lỗi $g_3(X \to Y, r)$ từ các phân hoạch thu gọn được xác định như sau:

$$g_3(X \to Y, r) = \frac{\sum\limits_{U \in \hat{\pi}_X} (|U| - \max\{|V|\,|V \in \hat{\pi}_{XY}, V \subset U\}) + \sum\limits_{U \in \hat{\pi}_X} \{(|U|\,|\,\nexists V \in \hat{\pi}_{XY}, V \subset U) - 1\}}{|r|}$$

Bây giờ chúng tôi đưa ra một số tính chất và nhận xét để xây dựng độ đo lỗi g_4 của phụ thuộc hàm $X \to Y$.

Tính chất 4.2. Cho một quan hệ $r(R)$ và $X, Y \subseteq R$. Khi đó, $X \to Y$ là một phụ thuộc hàm khi và chỉ khi $\sum\limits_{U \in \pi_X} |U|^2 = \sum\limits_{V \in \pi_{XY}} |V|^2$

Chứng minh. Giả sử phân hoạch π_X gồm các lớp tương đương $U_i, i = 1, ..., |\pi_X|$ và phân hoạch π_{XY} gồm các lớp tương đương $V_j,\ j = 1, ..., |\pi_{XY}|$. Gọi $E(\pi_X)$ là kỳ vọng của tổng số bộ ứng với các lớp tương đương $U_i,\ i = 1, ..., |\pi_X|$. Gọi $E(\pi_{XY})$ là kỳ vọng của tổng số bộ ứng với các lớp tương đương $V_j,\ j = 1, ..., |\pi_{XY}|$.

Khi đó

$$\begin{aligned}
E(\pi_X) &= \sum_{i=1}^{|\pi_X|} |U_i|.P\,(U_i)\,, \text{ với } P(U_i)\text{: khả năng phân bố các bộ của } r \text{ vào } U_i \\
&= \sum_{i=1}^{|\pi_X|} |U_i|.\frac{|U_i|}{|r|} = \frac{1}{|r|} \sum_{U \in \pi_X} |U|^2, \\
E(\pi_{XY}) &= \sum_{j=1}^{|\pi_{XY}|} |V_j|.P\,(V_j), \text{ với } P(V_j)\text{: khả năng phân bố các bộ của } r \text{ vào } V_j \\
&= \sum_{j=1}^{|\pi_{XY}|} |V_j|.\frac{|V_j|}{|r|} = \frac{1}{|r|} \sum_{V \in \pi_{XY}} |V|^2.
\end{aligned}$$

Ta có $E(\pi_X) = E(\pi_{XY})$ khi và chỉ khi có sự phân bố các bộ vào các $U \in \pi_X$ giống sự phân bố các bộ vào các $V \in \pi_{XY}$. Do vậy $X \to Y$ là một phụ thuộc hàm khi và chỉ khi $\sum\limits_{U \in \pi_X} |U|^2 = \sum\limits_{V \in \pi_{XY}} |V|^2$ ■

Nhận xét 4.1. Ta có thể đặt $\delta(X,Y) = \frac{\sum\limits_{V \in \pi_{XY}} |V|^2}{\sum\limits_{U \in \pi_X} |U|^2}$. Khi đó, $0 < \delta(X,Y) \leq 1$ và $\delta(X,Y)$ tăng thì khả năng xảy ra lỗi của phụ thuộc hàm càng ít.

Ví dụ 4.1. Cho quan hệ $r(R)$ sau:

Bảng 1. Một quan hệ trên tập thuộc tính $R = \{A_1, ..., A_4\}$

A_1	A_2	A_3	A_4
0	1	0	0
1	0	1	1
0	1	1	2
2	1	0	0
0	1	0	1
1	0	2	2

Khi đó, ta có: $\delta(A_1, A_2) = 1$, $\delta(A_1, A_3) = 4/7$, $\delta(A_1, A_4) = 3/7$.

Từ Tính chất 4.2 và Nhận xét 4.1, ta có Định nghĩa 4.3 như sau:

Định nghĩa 4.3. (*Độ đo lỗi g_4 theo phân hoạch*) Cho quan hệ $r(R)$. Khi đó, độ đo lỗi $g_4(X \to Y, r)$ từ các phân hoạch được tính như sau:

$$g_4(X \to Y, r) = 1 - \delta(X, Y) = 1 - \frac{\sum\limits_{V \in \pi_{XY}} |V|^2}{\sum\limits_{U \in \pi_X} |U|^2}.$$

Với Bảng 1, ta có $g_4(A_1 \to A_2, r) = 0$, $g_4(A_1 \to A_3, r) = 3/7$, $g_4(A_1 \to A_4, r) = 4/7$.

Nhận xét 4.2. Từ Tính chất 4.2, Nhận xét 4.1 và Định nghĩa 4.3, ta thấy rằng $g_4(X \to Y, r)$ có quan hệ mật thiết với sự phân bố của các bộ dữ liệu vào các $V \in \pi_{XY}$ ứng với các $U \in \pi_X$. Tuy nhiên $g_2(X \to Y, r)$ và $g_3(X \to Y, r)$ không biểu diễn được cho sự phân bố này.

Ví dụ 4.2. Cho quan hệ $r(R)$ sau:

Bảng 2. Một quan hệ trên tập thuộc tính {Hoten, Trieuchung, Benh}

Hoten	Trieuchung	Benh
P_1	2	4
P_2	1	1
P_3	1	2
P_4	2	2
P_5	1	4
P_6	2	3
P_7	1	1
P_8	2	2
P_9	2	2
P_{10}	1	3
P_{11}	2	1

Với quan hệ ở Bảng 2, ta có $g_2(Trieuchung \to Benh, r) = 0$, $g_3(Trieuchung \to Benh, r) = \frac{6}{11}$. Nếu chúng ta thay đổi sự phân bố của các bộ dữ liệu, chẳng hạn ở người có Hoten là P_6, ta thay thế giá trị 3 thành giá trị 1 ứng với thuộc tính Benh thì $g_2(Trieuchung \to Benh, r)$ và $g_3(Trieuchung \to Benh, r)$ vẫn không thay đổi.

Nhận xét 4.3. Bây giờ, ta thu hẹp tập dữ liệu của quan hệ r ứng với phép chọn, khi đó $g_4(X \to Y, \sigma_{X=x_i}(r))$ với $x_i \in \mathrm{dom}(X)$ đo được mức độ tập trung của các bộ dữ liệu trong $\sigma_{X=x_i}(r)$ vào các lớp tương đương $V \in \pi_{XY}$.

Thật vậy, ta có $g_4(X \to Y, \sigma_{X=x_i}(r)) = 1 - \frac{\sum\limits_{V \in \pi_{XY}} |V|^2}{|\sigma_{X=x_i(r)}|^2}$. Do đó $g_4(X \to Y, \sigma_{X=x_i}(r))$ càng nhỏ khi chỉ khi $\frac{\sum\limits_{V \in \pi_{XY}} |V|^2}{|\sigma_{X=x_i}(r)|^2}$ càng lớn. Hay $g_4(X \to Y, \sigma_{X=x_i}(r))$ càng nhỏ khi chỉ khi mức độ tập trung các bộ $\sigma_{X=x_i}(r)$ vào một hay một số lớp tương đương $V \in \pi_{XY}$ là càng lớn.

Với quan hệ ở Bảng 2, ta có: $g_4(Trieuchung \to Benh, \sigma_{Trieuchung=1}(r)) = \frac{18}{25}$, $g_4(Trieuchung \to Benh, \sigma_{Trieuchung=2}(r)) = \frac{2}{3}$. Nếu ở người có Hoten là P_6, ta thay đổi giá trị 3 thành giá trị 1 ứng với thuộc tính Benh thì $g_4(Trieuchung \to Benh, \sigma_{Trieuchung=2}(r)) = \frac{11}{8}$.

Như vậy nếu $g_4(Trieuchung \to Benh, \sigma_{Trieuchung=x_i}(r))$ càng nhỏ ứng với triệu chứng x_i thì mức độ phân bố tập trung bệnh nhân trong $\sigma_{Trieuchung=x_i}(r)$ vào một hoặc một số bệnh nào đó càng lớn và ngược lại. Điều này góp một phần trong việc dự đoán được bệnh của bệnh nhân thông qua các triệu chứng.

Nhận xét 4.4. Ta thấy rằng $g_4(X \to Y, r) \geq g_1(X \to Y, r)$, nghĩa là $g_4(X \to Y, r)$ nghiêm ngặt hơn $g_1(X \to Y, r)$. Tuy nhiên $g_1(X \to Y, r)$ là không tốt cho việc đo lỗi của $X \to Y$ và mức độ tập trung của các bộ dữ liệu trong r vào các lớp tương đương $V \in \pi_{XY}$. Thật vậy, ta có

$$\begin{aligned} g_4(X \to Y, r) =& 1 - \frac{\sum\limits_{V \in \pi_{XY}} |V|^2}{\sum\limits_{U \in \pi_X} |U|^2} = \frac{|r|^2}{\sum\limits_{U \in \pi_X} |U|^2} \cdot \frac{\sum\limits_{U \in \pi_X} \left(|U|^2 - \sum\limits_{\substack{V \in \pi_{XY} \\ V \subseteq U}} |V|^2 \right)}{|r|^2} \\ =& \frac{|r|^2}{\sum\limits_{U \in \pi_X} |U|^2} \cdot \frac{\left| \left\{ (t_i, t_j) \middle| t_i, t_j \in r, t_i[X] = t_j[X], t_i[Y] \neq t_j[Y] \right\} \right|}{|r|^2} \\ =& \frac{|r|^2}{\sum\limits_{U \in \pi_X} |U|^2} \cdot g_1(X \to Y, r). \end{aligned}$$

Suy ra $g_4(X \to Y, r) \geq g_1(X \to Y, r)$.

Bây giờ ta xét một quan hệ ở Ví dụ 4.3 dưới đây.

Ví dụ 4.3. Cho quan hệ $r(R)$ sau:

Bảng 3. Một quan hệ trên tập thuộc tính $\{A, B, C\}$

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

$g_4(A \to B, r) = \frac{44}{59} \simeq 0.756$, $g_3(A \to B, r) = \frac{11}{15} \simeq 0.733$, $g_1(A \to B, r) = \frac{44}{225} \simeq 0.196$.

Ta thấy rằng độ đo lỗi $g_1(A \to B, r)$ là quá nhỏ so với $g_4(A \to B, r)$, trong khi đó lỗi của phụ thuộc hàm $A \to B$ là tương đối lớn và mức độ phân bố tập trung các bộ dữ liệu vào các lớp tương đương $V \in \pi_{AB}$ ứng với $U \in \pi_A$ là nhỏ.

Tính chất 4.3. (*Độ đo g_4 trên phân hoạch thu gọn*) Cho quan hệ $r(R)$. Khi đó, độ đo lỗi $g_4(X \to Y)$ từ các phân hoạch thu gọn được xác định như sau:

$$g_4(X \to Y, r) = \frac{\sum\limits_{U \in \hat{\pi}_X} |U|^2 - \sum\limits_{V \in \hat{\pi}_{XY}} |V|^2 - \sum\limits_{U \in \hat{\pi}_X} |U| + \sum\limits_{U \in \hat{\pi}_{XY}} |V|}{\sum\limits_{U \in \hat{\pi}_X} |U|^2 + |r| - \sum\limits_{U \in \hat{\pi}_X} |U|}.$$

Chứng minh. Ta có $|r| - \sum\limits_{U \in \hat{\pi}_X} |U|$: là số lớp tương đương một phần tử bị loại khỏi π_X để có $\hat{\pi}_X$. Suy ra

$$\sum_{U \in \pi_X} |U|^2 = \sum_{U \in \hat{\pi}_X} |U|^2 + |r| - \sum_{U \in \hat{\pi}_X} |U|$$

Tương tự ta có $\sum\limits_{V \in \pi_{XY}} |V|^2 = \sum\limits_{V \in \hat{\pi}_{XY}} |V|^2 + |r| - \sum\limits_{V \in \hat{\pi}_{XY}} |V|$. Do đó:

$$\begin{aligned} g_4(X \to Y, r) &= 1 - \frac{\sum\limits_{V \in \hat{\pi}_{XY}} |V|^2 + |r| - \sum\limits_{V \in \hat{\pi}_{XY}} |V|}{\sum\limits_{U \in \hat{\pi}_X} |U|^2 + |r| - \sum\limits_{U \in \hat{\pi}_X} |U|} \\ &= \frac{\sum\limits_{U \in \hat{\pi}_X} |U|^2 - \sum\limits_{V \in \hat{\pi}_{XY}} |V|^2 - \sum\limits_{U \in \hat{\pi}_X} |U| + \sum\limits_{U \in \hat{\pi}_{XY}} |V|}{\sum\limits_{U \in \hat{\pi}_X} |U|^2 + |r| - \sum\limits_{U \in \hat{\pi}_X} |U|} \end{aligned}$$ ■

Ví dụ 4.4. Với Bảng 1, ta có:

$$\hat{\pi}_{A_1} = \{\{t_1, t_2, t_3\},\ \{t_4, t_5\}\},\ \hat{\pi}_{A_1A_3} = \{\{t_1, t_2\}\}$$

Độ đo lỗi $g_4(A_1 \to A_3, r)$ được tính từ các phân hoạch thu gọn và $g_4(A_1 \to A_3, r) = \dfrac{3}{7}$.

5. BIỂU DIỄN PHỤ THUỘC HÀM XẤP XỈ THEO MA TRẬN PHÂN BIỆT ĐƯỢC

Trong phần này chúng tôi xây dựng ma trận phân biệt được $M(S)$ theo 1 cách khác và đề xuất cách biểu diễn độ đo lỗi g_1, g_2, độ phụ thuộc γ và ý nghĩa thuộc tính σ thông qua $M(S)$.

Định nghĩa 5.1. (Ma trận phân biệt được) Cho $r = t_1, t_2, \ldots, t_n$. Ma trận phân biệt được của $S = (r, R)$, ký hiệu $M(S) = (m_{ij})_{|r| \times |r|}$ là ma trận đối xứng mà mỗi phần tử của nó là một tập hợp các thuộc tính, được xác định như sau:

$$mij = \begin{cases} \{A_k \in C | t_i(A_k) \neq t_j(A_k)\} & t_i(D) \neq t_j(D) \\ \emptyset & t_i(D) = t_j(D) \\ \{\beta \,|\, \nexists A_k \in C : t_i(A_k) \neq t_j(A_k)\} & t_i(D) \neq t_j(D), \beta \notin C \end{cases} \quad \text{với } i, j = \overline{1, |r|}.$$

Định nghĩa 5.2. Cho ma trận phân biệt được $M(S) = (m_{ij})_{|r|\times|r|}$. Khi đó số lần X xuất hiện trong $M(S)$ ký hiệu là $SL(X)$ và được định nghĩa như sau:

$$SL\,(X) = \left|\left\{ m_{ij} \,\middle|\, (X \cap m_{ij}) \neq \emptyset, X \subseteq C, \forall i,j = \overline{1,|r|} \right\}\right|$$

Nếu $X = \emptyset$ thì $SL(\emptyset) = \left|\left\{ m_{ij} \,\middle|\, m_{ij} = \emptyset\ , \forall i,j = \overline{1,|r|} \right\}\right|$.

Tính chất 5.1. Cho ma trân phân biệt được $M(S) = (m_{ij})_{|r|\times|r|}$. Khi đó, $X \to D$ với $X \subseteq C$ là một phụ thuộc hàm khi và chỉ khi $SL(X) = |r|^2 - SL(\emptyset)$.

Chứng minh. Ta có với mỗi $m_{ij} = \emptyset$ thì $t_i[D] = t_j[D]$. Do đó $SL(\emptyset)$ là số cặp bộ (t_i, t_j), $i,j = \overline{1,|r|}$ không vi phạm phụ thuộc hàm $X \to D$.

Mặt khác, với mỗi m_{ij} sao cho $(X \cap m_{ij}) \neq \emptyset, X \subseteq C$ thì $t_i[X \cap m_{ij}] \neq t_j[X \cap m_{ij}]$ và $t_i[D] \neq t_j[D]$ suy ra $t_i[X] \neq t_j[X]$ và $t_i[D] \neq t_j[D]$. Do đó $SL(X)$ là số cặp bộ (t_i, t_j), $i,j = \overline{1,|r|}$ không vi phạm phụ thuộc hàm $X \to D$.

Vậy $SL(X) + SL(\emptyset) = |r|^2$ với $X \subseteq C \Leftrightarrow \nexists t_i, t_j \in r : t_i[D] \neq t_j[D]$ và $t_i[X] = t_j[X] \Leftrightarrow X \to D$ là một phụ thuộc hàm. ■

Tính chất 5.2. Độ đo lỗi g_1 của một phụ thuộc hàm $X \to D$, với $X \subseteq C$ trên r được xác định như sau:

$$g_1(X \to D, r) = 1 - \frac{SL(\emptyset) + SL(X)}{|r|^2}$$

Chứng minh. Theo Tính chất 5.1, ta có $SL(X) + SL(\emptyset)$ là số cặp bộ (t_i, t_j), $i,j = \overline{1,|r|}$ không vi phạm phụ thuộc hàm $X \to D$. Do đó: $g_1(X \to D, r) = \frac{|r|^2 - SL(\emptyset) - SL(X)}{|r|^2} = 1 - \frac{SL(\emptyset)+SL(X)}{|r|^2}$
■

Hệ quả 5.1. *Độ đo lỗi g_1 của một phụ thuộc hàm $X \to D$, với $X \subseteq C$ trên r được xác định như sau:*

$$g_1(X \to D, r) = \frac{\left|\left\{ m_{ij} \,|\, ((X \cap m_{ij}) = \emptyset) \wedge (m_{ij} \neq \emptyset)\ ,\ i,j = \overline{1,|r|} \right\}\right|}{|r|^2}$$

Chứng minh. - Ta có $((X \cap m_{ij}) = \emptyset) \wedge (m_{ij} \neq \emptyset) \Leftrightarrow \nexists A \in X : t_i[A] \neq t_j[A]$ và $t_i[D] \neq t_j[D] \Leftrightarrow t_i[X] = t_j[X]$ và $t_i[D] \neq t_j[D] \Leftrightarrow$ cặp (t_i, t_j) vi phạm phụ thuộc $X \to D$.

- Trường hợp $m_{ij} = \beta$ suy ra $((X \cap m_{ij}) = \emptyset) \wedge (m_{ij} \neq \emptyset)$ (vì $\beta \notin X$) $\Leftrightarrow$ cặp (t_i, t_j) vi phạm phụ thuộc $X \to D$

Vậy $\left|\left\{ m_{ij} \,|\, ((X \cap m_{ij}) = \emptyset) \wedge (m_{ij} \neq \emptyset)\ ,\ i,j = \overline{1,|r|} \right\}\right|$ = số cặp (t_i, t_j) vi phạm phụ thuộc $X \to D$. Do đó: $g_1(X \to D, r) = \frac{\left|\left\{ m_{ij} \,|((X \cap m_{ij}) = \emptyset) \wedge (m_{ij} \neq \emptyset)\ , i,j = \overline{1,|r|} \right\}\right|}{|r|^2}$. ■

Tính chất 5.3. Độ đo lỗi g_2 của một phụ thuộc hàm $X \to D$, với $X \subseteq C$ trên r được xác định như sau:

$$g_2(X \to D, r) = 1 - \frac{\sum_{i=1}^{|r|} hang_i(X)}{|r|}$$

Trong đó: $hang_i(X) = \begin{cases} 1 & (X \cap m_{ij}) \neq \emptyset \quad \forall m_{ij} \neq \emptyset, j = \overline{1,|r|} \\ 0 & \exists j : (m_{ij} \neq \emptyset) \wedge ((X \cap m_{ij}) = \emptyset) \end{cases}$.

Chứng minh. Ta có:

$$g_2(X \to D, r) = \frac{|\{\, t_i \,|t_i \in r, \exists t_j \in r : t_i\,[X] = t_j\,[X]\ \ , t_i\,[D] \neq t_j\,[D]\ \}\,|}{|r|}.$$

Gọi q là số bộ vi phạm phụ thuộc hàm $X \to D$ trên r. Khi đó:

$$q = |\{\, t_i \,|t_i \in r, \exists t_j \in r : t_i\,[X] = t_j\,[X]\ \ , t_i\,[D] \neq t_j\,[D]\ \}\,|$$

Ta có với mỗi bộ $t_i \in r$:

- *Xét trường hợp, nếu* $(X \cap m_{ij}) \neq \emptyset \quad \forall m_{ij} \neq \emptyset,\ \ j = \overline{1, |r|}$

Ta có $(X \cap m_{ij}) \neq \emptyset \quad \forall m_{ij} \neq \emptyset, j = \overline{1, |r|} \Leftrightarrow \nexists t_j \in r : t_i[X] = t_j[X]$ và $t_i[D] \neq t_j[D]$ (vì nếu $\exists t_j : t_i[X] = t_j[X]$ và $t_i[D] \neq t_j[D]$ thì $(X \cap m_{ij}) = \emptyset$. Điều này mâu thuẫn với giả thiết) $\Leftrightarrow t_i$ là bộ thỏa phụ thuộc $X \to D \Leftrightarrow hang_i(X) = 1$.

- Xét trường hợp, nếu $\exists j : (m_{ij} \neq \emptyset) \wedge ((X \cap m_{ij}) = \emptyset)$:

Ta có $\exists j : (m_{ij} \neq \emptyset) \wedge ((X \cap m_{ij}) = \emptyset) \Leftrightarrow t_i[X] = t_j[X]$ và $t_i[D] \neq t_j[D] \Leftrightarrow t_i$ là bộ vi phạm phụ thuộc $X \to D \Leftrightarrow hang_i(X) = 0$.

- *Xét trường hợp* $m_{ij} = \beta$

Ta có $m_{ij} = \beta$ suy ra $(m_{ij} \neq \emptyset) \wedge ((X \cap m_{ij}) = \emptyset) \Leftrightarrow t_i$ là bộ vi phạm phụ thuộc $X \to D \Leftrightarrow hang_i(X) = 0$. Do đó $\sum\limits_{i=1}^{|r|} hang_i(X) =$ số bộ thỏa phụ thuộc hàm $X \to D$ nên $q = |r| - \sum\limits_{i=1}^{|r|} hang_i(X)$. Suy ra $g_2(X \to D, r) = 1 - \frac{\sum\limits_{i=1}^{|r|} hang_i(X)}{|r|}$. ■

Nhận xét 5.1.

- Độ phụ thuộc $\gamma(X, D)$ được tính thông qua công thức $\gamma(X, D) = 1 - g_2(X \to D, r)$ [7], với $X \subseteq C$ và $g_2(X \to D, r) = 1 - \frac{\sum\limits_{i=1}^{|r|} hang_i(X)}{|r|}$ (Tính chất 5.3) như sau: $\gamma(X, D) = \frac{\sum\limits_{i=1}^{|r|} hang_i(X)}{|r|}$.

- Ý nghĩa thuộc tính được tính dựa trên ma trận phân biệt được thông qua công thức:

$\sigma_{C \cup D}(A_i) = 1 - \frac{\gamma(C - \{A_i\}, D)}{\gamma(C, D)}$ [9], với $\gamma(X, D) = \frac{\sum\limits_{i=1}^{|r|} hang_i(X)}{|r|}$ và $X \subseteq C$.

6. LUẬT KẾT HỢP

Định nghĩa 6.1. [2] (CSDL giao dịch) Cho $I(items) = \{i_1, i_2, ..., i_m\}$ là tập mục, một CSDL giao dịch, được ký hiệu là TD gồm các giao dịch $T \in TD$, với mỗi giao dịch (Transaction) T được định nghĩa là tập con của tập mục $I (T \subseteq I)$ và có một định danh duy nhất $\langle TID, i_1, i_2, ..., i_k \rangle$.

Định nghĩa 6.2. [2] (Luật kết hợp) Cho cơ sở dữ liệu TD gồm các giao dịch T ứng với tập mục $I = \{i_1, i_2, \ldots, i_m\}$. Khi đó, một luật kết hợp giữa I_X và I_Y có dạng là $I_X \Rightarrow I_Y$, với I_X, $I_Y \subset I$ và $I_X \cap I_Y = \emptyset$.

Định nghĩa 6.3. [2] (Độ hỗ trợ của một tập mục) Cho tập mục $I_X \subseteq I$, độ hỗ trợ của tập mục I_X được ký hiệu là $Support(I_X, TD)$ và được định nghĩa như sau: $Support(I_X, TD) = \frac{|\{T \in TD | I_X \subseteq T\}|}{|TD|}$ hay $Support(I_X, TD)$ là tỷ lệ phần trăm giữa các giao dịch chứa I_X trên tổng các giao dịch có trong cơ sở dữ liệu TD.

Định nghĩa 6.4. [2] (Độ hỗ trợ của luật kết hợp) Độ hỗ trợ của luật kết hợp $I_X \Rightarrow I_Y$, (ký hiệu là $Support(I_X \Rightarrow I_Y, TD)$) bằng tỷ lệ phần trăm giữa các giao dịch chứa $I_X \cup I_Y$ trên tổng số các giao dịch trong cơ sở dữ liệu TD:

$$Support(I_X \Rightarrow I_Y, TD) = Support(I_X \cup I_Y, TD) = \frac{|\{T \in TD | I_X \cup I_Y \subseteq T\}|}{|TD|}.$$

Định nghĩa 6.5. [2] (Độ tin cậy của luật kết hợp) Độ tin cậy của luật kết hợp $I_X \Rightarrow I_Y$, (ký hiệu là $Confidence(I_X \Rightarrow I_Y, TD)$) bằng tỷ lệ phần trăm giữa các giao dịch chứa $I_X \cup I_Y$ trên số giao dịch có chứa I_X:

$$Confidence(I_X \Rightarrow I_Y, TD) = \frac{Support(I_X \cup I_Y, TD)}{Support(I_X, TD)}$$

7. MỐI QUAN HỆ GIỮA PHỤ THUỘC HÀM XẤP XỈ VÀ LUẬT KẾT HỢP

7.1. Một số ký hiệu [7]

- $r(R)$: một quan hệ trên lược đồ R, với $R = \{A_1, A_2, ..., A_m\}$
- Miền trị của $X \subset R$, ký hiệu là $\mathrm{dom}(X) = \{x_1, x_2, ..., x_k\}$
- Miền trị của $Y \subset R$, ký hiệu là $\mathrm{dom}(Y) = \{y_1, y_2, ..., y_l\}$
- $n = |r|$ và $n_{x_i} = |\{t \in r | t[X] = x_i\}|$, $n_{y_j} = |\{t \in r | t[Y] = y_j\}|$,
  $n_{x_i y_j} = \left|\{t \in r | t[X] = x_i \text{ và } t[Y] = y_j\}\right|$
- Cho $R = \{A_1, A_2, ..., A_m\}$. Khi đó, với $A_k \in R$ thì i_{A_k} là một mục (item) trong luật kết hợp.
- Cho $X \subseteq R$. Khi đó, tập mục của X ký hiệu là $I_X = \{i_{A_k} | A_k \in X\}$

7.2. Định nghĩa mới của phụ thuộc hàm xấp xỉ

Định nghĩa 7.1. [7] Cho quan hệ $r(R)$ với $R = \{A_1, A_2, ..., A_m\}$. Một cơ sở dữ liệu giao dịch TD được định nghĩa như sau: mỗi cặp $(t, s) \in r \times r$ sao cho $t, s \in r$ là một giao dịch $ts \in TD$ và được xác định như sau: $i_{A_k} \in ts \Leftrightarrow t[A_k] = s[A_k]$. Khi đó: $ts.i_{A_k} = \begin{cases} 1 & \text{nếu } i_{A_k} \in ts \\ 0 & \text{nếu } i_{A_k} \notin ts \end{cases}$

Ví dụ 7.1. Cho quan hệ sau:

Bảng 4. Một quan hệ r

Masv	Quequan	Truong	Ketqua
1	Hue	QH	Dau
2	Hue	NH	Dau
3	Hue	QH	Rot

Ta có thể biểu diễn Bảng 4 thành cơ sở dữ liệu giao dịch TD như trong Bảng 5.

Bảng 5. Một cơ sở dữ liệu giao dịch *TD* của r

ts	i_{Masv}	$i_{Quequan}$	i_{Truong}	i_{Ketqua}
(1,1)	1	1	1	1
(1,2)	0	1	0	1
(1,3)	0	1	1	0
(2,1)	0	1	0	1
(2,2)	1	1	1	1
(2,3)	0	1	0	0
(3,1)	0	1	1	0
(3,2)	0	1	0	0
(3,3)	1	1	1	1

Định nghĩa 7.2. [7] Cho $X, Y \subset R$ sao cho $X \cap Y = \emptyset$. Khi đó một phụ thuộc hàm xấp xỉ $X \to Y$ trên quan hệ r là một luật kết hợp $I_X \Rightarrow I_Y$ trên cơ sở giao dịch *TD*. Và ta có độ hỗ trợ và độ tin cậy như sau: $Support\,(X \to Y, r) = Support(I_X \Rightarrow I_Y, TD)$

$$Confidence\,(X \to Y, r) = Confidence(I_X \Rightarrow I_Y, TD)$$

Theo cách biểu diễn này thì độ hỗ trợ của tập thuộc tính X là $Support(X, r) = Support(I_X, TD)$.

Tính chất 7.1. [7] $X,\ Y \subset R,\ X \to Y$ là một phụ thuộc hàm khi và chỉ khi $Confidence(I_X \Rightarrow I_Y, TD) = 1$.

Định nghĩa 7.3. [7] Cho $R = \{A_1, A_2, ..., A_m\}$, $X, Y \subset R$. Khi đó, độ hỗ trợ của X và $X \to Y$ tương ứng là $Support(X, r) = \frac{1}{n^2} \sum_{i=1}^{k} n_{x_i}^2$ và $Support(X \to Y, r) = \frac{1}{n^2} \sum_{i=1}^{k} \sum_{j=1}^{l} n_{x_i y_j}^2$.

Ví dụ 7.2. Trong Bảng 4, ta có: $Support\,(Truong, r) = 5/9$; $Support\,(\{Truong,\ Ketqua\}, r) = 3/9$.

Khi đó, ta có một số luật kết hợp trong *TD* tương ứng với các phụ thuộc hàm xấp xỉ trong r như sau:

Bảng 6. Một số luật kết hợp tương ứng với phụ thuộc hàm xấp xỉ

Luật kết hợp	Độ hỗ trợ	Độ tin cậy	Phụ thuộc hàm xấp xỉ
$\{i_{Quequan}\} \Rightarrow \{i_{Truong}\}$	5/9	5/9	$\{Quequan\} \to \{Truong\}$
$\{i_{Quequan},\ i_{Truong}\} \Rightarrow \{i_{Ketqua}\}$	1/3	3/5	$\{Quequan, Truong\} \to \{Ketqua\}$

7.3. Độ hỗ trợ, độ tin cậy của phụ thuộc hàm xấp xỉ

Gọi $ARS_{[X \to Y]} = \left\{AR_{ij} | \exists t \in r : t\,[X] = x_i \text{ và } t\,[Y] = y_j\right\} \quad \forall i = \overline{1,k}; \forall j = \overline{1,l}$, trong đó AR_{ij} là một luật kết hợp có dạng $(X = x_i) \Rightarrow (Y = y_j)$, với S_{ij}, C_{ij} tương ứng là độ hỗ trợ, độ tin cậy của AR_{ij}.

Tính chất 7.2. [7] Độ hỗ trợ của phụ thuộc hàm xấp xỉ $X \to Y$ được tính theo công thức như sau:

$$Support(X \to Y, r) = \frac{1}{n^2} \sum_{AR_{ij} \in ARS_{[X \to Y]}} n_{x_i y_j}^2 = \sum_{AR_{ij} \in ARS_{[X \to Y]}} S_{ij}^2$$

Tính chất 7.3. [7] Độ tin cậy của một phụ thuộc hàm xấp xỉ được tính theo công thức sau:

$$\frac{1}{Confidence(X \to Y, r)} = \sum_{AR_{ij} \in ARS_{[X \to Y]}} \frac{S_{ij}^2}{\sum\limits_{AR_{pq} \in ARS_{[X \to Y]}} S_{pq}^2} \cdot \frac{1}{C_{ij}}.$$

Ví dụ 7.3. Đối với Bảng 4, ta có:

$$\frac{1}{Confidence(\{Quequan, Truong\} \to \{Ketqua\}, r)} = \frac{5}{3}$$
$$\Rightarrow Confidence(\{Quequan, Truong\} \to \{Ketqua\}, r) = \frac{3}{5}$$

7.4. Biểu diễn độ phụ thuộc, độ đo lỗi thông qua luật kết hợp

Tính chất 7.4. [7] Cho phụ thuộc hàm xấp xỉ $X \overset{\gamma(X,Y)}{\to} Y$, trong đó độ phụ thuộc $\gamma(X,Y) = \frac{|POS(X,Y)|}{|r|}$. Khi đó $\gamma(X,Y) = \sum_{AR_{ij} \in ARS_{[X \to Y]} | C_{ij}=1} S_{ij}$.

Tính chất 7.5. [7] Độ đo lỗi g_1 của phụ thuộc hàm $X \to Y$ ở Định nghĩa 3.1 được biểu diễn như sau:

$$g_1(X \to Y, r) = Support\ (I_X, TD) - Support\ (I_X \Rightarrow I_Y, TD)$$

Tính chất 7.6. [7] Độ đo lỗi g_2 của phụ thuộc hàm $X \to Y$ ở Định nghĩa 3.2 được biểu diễn như sau:

$$g_2(X \to Y, r) = 1 - \sum\nolimits_{AR_{ij} \in ARS_{[X \to Y]} | C_{ij}=1} S_{ij}$$

Tính chất 7.7. [7] Cho phụ thuộc hàm xấp xỉ $X \to Y$ với

$$g_3(X \to Y, r) = \frac{|r| - \sum\limits_{U \in \pi_X} \max \left\{ |V| \,\middle|\, V \in \pi_{XY}, V \subseteq U \right\}}{|r|}$$

Khi đó $g_3(X \to Y, r) = 1 - \sum\limits_{i=1}^{K} \max\limits_{j=\overline{1,l}} \{S_{ij} | AR_{ij} \in ARS_{[X \to Y]}\}$.

Tính chất 7.8. Cho $g_4(X \to Y, r)$ là độ đo lỗi của phụ thuộc hàm $X \to Y$, *Confidence*($I_X \Rightarrow I_Y$, *TD*) là độ tin cậy của luật kết hợp $I_X \Rightarrow I_Y$. Khi đó $g_4(X \to Y, r) = 1 -$ *Confidence*($I_X \Rightarrow I_Y$, *TD*).

Chứng minh. Ta có

$$\begin{aligned} g_4(X \to Y, r) =& 1 - \frac{\sum\limits_{V \in \pi_{XY}} |V|^2}{\sum\limits_{U \in \pi_X} |U|^2} = 1 - \frac{\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{l} |\{t \in r | (t[X] = x_i) \wedge (t[Y] = y_i)\}|^2}{\sum\limits_{i=1}^{k} |\{t \in r | t[X] = x_i\}|^2} \\ =& 1 - \frac{\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{l} n_{x_i y_j}^2}{\sum\limits_{i=1}^{k} n_{x_i}^2} = 1 - \frac{Support(I_X \Rightarrow I_Y, TD)}{Support(I_X, TD)} \\ =& 1 - Confidence(I_X \Rightarrow I_Y, TD). \end{aligned}$$

Vậy $g_4(X \to Y, r) = 1 - Confidence(I_X \Rightarrow I_Y, TD)$ ■

Ví dụ 7.4. Đối với Bảng 1 ta có $g_4(A_1 \to A_3, r) = 3/7$ và $Confidence(I_{A_1} \Rightarrow I_{A_3}, TD) = 4/7$. Do đó, $g_4(A_1 \to A_3, r) = 1 - Confidence(I_{A_1} \Rightarrow I_{A_3}, TD)$.

Tính chất 7.9. Mối liên hệ giữa $g_4(X \to Y, r)$ và $ARS_{[X \to Y]}$

$$g_4(X \to Y, r) = 1 - \frac{\sum_{AR_{pq} \in ARS_{[X \to Y]}} S_{pq}^2}{\sum_{AR_{ij} \in ARS_{[X \to Y]}} \left(\frac{S_{ij}^2}{C_{ij}}\right)}$$

Chứng minh. Ta có $g_4(X \to Y, r) = 1 - Confidence(I_X \Rightarrow I_Y, TD)$. Mà

$$\frac{1}{Confidence(X \to Y, r)} = \sum_{AR_{ij} \in ARS_{[X \to Y]}} \frac{S_{ij}^2}{\sum\limits_{AR_{pq} \in ARS_{[X \to Y]}} S_{pq}^2} \cdot \frac{1}{C_{ij}} \text{ (Tính chất 7.3)}$$

suy ra $g_4(X \to Y) = \dfrac{\frac{1}{\sum_{AR_{pq} \in ARS_{[X \to Y]}} S_{pq}^2} \sum_{AR_{ij} \in ARS_{[X \to Y]}} \left(\frac{S_{ij}^2}{C_{ij}}\right) - 1}{\frac{1}{\sum_{AR_{pq} \in ARS_{[X \to Y]}} S_{pq}^2} \sum_{AR_{ij} \in ARS_{[X \to Y]}} \left(\frac{S_{ij}^2}{C_{ij}}\right)}$.

Do đó, $g_4(X \to Y, r) = 1 - \dfrac{\sum_{AR_{pq} \in ARS_{[X \to Y]}} S_{pq}^2}{\sum_{AR_{ij} \in ARS_{[X \to Y]}} \left(\frac{S_{ij}^2}{C_{ij}}\right)}$. ■

Ví dụ 7.5. Đối với Bảng 1 ta có $g_4(A_1 \to A_3, r) = 3/7$ (theo Định nghĩa 4.3) và

$$1 - \frac{\sum_{AR_{pq} \in ARS_{[X \to Y]}} S_{pq}^2}{\sum_{AR_{ij} \in ARS_{[X \to Y]}} \left(\frac{S_{ij}^2}{C_{ij}}\right)} = 1 - 8/14 = 3/7.$$

8. KẾT LUẬN

Trong bài báo này chúng tôi đã nghiên cứu phụ thuộc hàm xấp xỉ dựa vào phân hoạch, ma trận phân biệt được và luật kết hợp và đã đạt được một số kết quả sau:

1. Đề xuất độ đo lỗi g_4 đối với phụ thuộc hàm và phân tích những thuận lợi của nó so với các độ đo lỗi g_1, g_2, g_3. Độ đo g_4 không những đo được lỗi của phụ thuộc hàm mà còn đo được mức độ tập trung hay không tập trung của các bộ dữ liệu. Sau đó xây dựng g_4 trên phân hoạch thu gọn và biểu diễn mối quan hệ giữa nó với độ tin cậy của luật kết hợp. Từ đó chúng ta có thể sử dụng các thuật toán phát hiện luật kết hợp để phát hiện các phụ thuộc hàm xấp xỉ và ngược lại.

2. Đưa ra một định nghĩa mới về ma trận phân biệt được. Từ đó xem xét phụ thuộc hàm, biểu diễn độ đo lỗi g_1, độ đo lỗi g_2, độ phụ thuộc và ý nghĩa thuộc tính thông qua ma trận phân biệt được. Điều này làm cơ sở để nghiên cứu tiếp tục thuật toán rút gọn các thuộc tính dư thừa thông qua các độ đo lỗi này.

TÀI LIỆU THAM KHẢO

[1] L. B. Cristofor, A Rough Set Based Generalization of Functional Denpendencies, Department of Math and Computer Science, UMass/Boston, 2000.

[2] Rakesh Agrawal and Ramakrishnan Srikant, Fast algorithms for mining association rules in large databases, *In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB, Santiago, Chile (1994) 487-499.

[3] Ullas Nambiar, Subbarao Kamhampati, Mining Approximate Functional Dependencies and Concept Similarities to Answer Imprecise Queries, Department of Computer Science, Arizona State University, USA, 2004.

[4] Y. Huhtala, J Karkkainen, P. Porkka, H. Toivonnen, Tane: An Efficient Algorithm for Discovery Functional and Approxiamate Dependencies, *The Computer Journal*, **42** (3) (1999) 100-111.

[5] J. Kivinen, H. Mannila, Approximate Inference of Functional Dependencies from Relations, *Theoretical Computer Science*, **149** (1) (1995) 129-149.

[6] Stéphane Lopes, Jean-Marc Petit, Lotfi Lakhal, Functional and approximate dependency mining: database and FCA points of view, *J. Exp. Theor. Artif. Intell.*, **14**(2-3) (2002) 93-114.

[7] Daniel Sánchez, José María Serano, Ignacio Blanco, Maria José Martin-Bautista, María Amparo Vila, Using association rules to mine for strong approximate dependencies, *Data mining knowledge discovery*, *Springer Science* (2008) 313-348.

[8] Mohammed J. Zaki, Scalable algorithms for association mining. IEEE Transactions on *Knowledge and Data Engineering*, **12**(3) (2000) 372-390.

[9] Jan Komorowski, Lech Polkowski, Andrzej Skowron, Rough Set: A Tutorial, Institute of Mathematics, Warsaw University, 2000.

[10] Trần Duy Anh, Phát hiện các phụ thuộc hàm xấp xỉ theo cách tiếp cận tập thô, *Tạp chí tin học và điều khiển học*, **23**(3) (2007) 284-295

[11] Keyun Hu, Yuchang Lu, Chunyi Shi, Feature ranking in rough set, Department of Computer science, Tsinghua University Beijing 100084, P.R.China, 2003.

[12] Jiawei Han, Jian Pei, Yiwen Yin, and Runying Mao, Mining frequent patterns without candidate generation, *Data Mining and Knowledge Discovery* **8** (2004) 53-87.

Ngày nhận bài 21 - 3 - 2013
Nhận lại sau sửa ngày 8 - 9 - 2013