MỘT ĐỘ ĐO MỚI ĐO ĐỘ PHỤ THUỘC THUỘC TÍNH

Nguyễn Minh Huy [1], Đỗ Sĩ Trường [2], Nguyễn Huy Đức [3], Nguyễn Thanh Tùng [2]

[1]Trường Đại học Thủ đô Hà nội

[2] Trường Đại học Lạc Hồng

[3] Trường Cao đẳng Sư phạm Trung ương

nguyenminhhuy86@gmail.com,truongds@gmail.com,ducnghuy@yahoo.com, nttung@lhu.edu.vn

TÓM TẮT-*Trong bài báo này, chúng tôi trình bày phương pháp xây dựng một độ đo mới, gọi là độ phụ thuộc Gamma, đo độ phụ thuộc giữa các tập thuộc tính phạm trù (categorical attributes) trong một hệ thông tin. Độ đo này được xây dựng dựa trên khái niệm entropy bù (complementary entropy) do Jiye Liang và cộng sự đề xuất. Với hai tập thuộc tính X và Y, độ đo này sẽ gán cho chúng một số thực thuộc khoảng đóng [0,1] phản ánh mức độ phụ thuộc của Y vào X. Giá trị độ đo bằng 1 khi và chỉ khi tồn tại phụ thuộc hàm $X \to Y$. Và như thế, giá trị của nó càng gần bằng 1 thì sự phụ thuộc của Y vào X trong hệ thông tin càng gần phụ thuộc hàm $X \to Y$. Các tính chất của độ đo phụ thuộc đề xuất và mối liên hệ của nó với phụ thuộc hàm cũng được nghiên cứu. Các tính chất này cho thấy có thể xem nó là sự mở rộng của khái niệm phụ thuộc hàm, và độ phụ thuộc Gamma có thể được sử dụng như là một độ đo phụ thuộc hàm xấp xỉ.*

Từ khóa- *Entropy bù, Độ phụ thuộc thuộc tính Gamma, Phụ thuộc hàm, Khai phá dữ liệu.*

I. MỞ ĐẦU

Trong một cơ sở dữ liệu, tập thuộc tính Y phụ thuộc hàm vào tập thuộc tính X nếu giá trị của các thuộc tính trong Y được xác định duy nhất bởi giá trị của các thuộc tính trong X. Trong những năm gần đây, vấn đề khai phá sự phụ thuộc giữa các thuộc tính (các biến) trong cơ sở dữ liệu đã trở thành đề tài thu hút sự quan tâm của nhiều nhà nghiên cứu. Mục tiêu của khai phá phụ thuộc thuộc tính là nhằm phát hiện ra các mối quan hệ giữa các thuộc tính trong một cơ sở dữ liệu. Các phụ thuộc thuộc tính phát hiện được sẽ được sử dụng vào việc thực hiện các nhiệm vụ khác trong khai phá dữ liệu như lựa chọn thuộc tính (đặc trưng) trong nhận dạng, phân lớp dữ liệu, khai phá luật kết hợp, rời rạc hóa dữ liệu, … [10, 17, 23].

Để phát hiện hiệu quả các phụ thuộc thuộc tính thì việc xây dựng các độ đo (các hàm) cho phép đánh giá đúng mức độ phụ thuộc là điều rất quan trọng. Trong những năm qua, nhiều độ đo đã được đề xuất hoặc phát triển nhằm đo đạc mức độ phụ thuộc giữa các thuộc tính. Hệ số tương quan Pearson [9] là độ đo kinh điển, được xây dựng nhằm đánh giá mức độ tương quan tuyến tính giữa các biến số ngẫu nhiên. Dễ thấy, có một số hạn chế khi sử dụng hệ số này. Thứ nhất, hệ số tương quan chỉ phản ánh được sự phụ thuộc tuyến tính, trong khi trên thực tế, các mối quan hệ giữa các biến thường không phải là tuyến tính. Thứ hai, hệ số tương quan không cho phép đo đạc mức độ quan hệ giữa một tập biến này với một tập biến khác. Như đã biết, khi giải quyết vấn đề lựa chọn thuộc tính, ta thường phải tính toán mối quan hệ giữa một thuộc tính ứng viên và một tập thuộc tính đã được lựa chọn. Hơn nữa, hệ số tương quan Pearson có thể trở nên không hiệu quả khi phải tính toán độ phụ thuộc giữa các thuộc tính phạm trù (như quốc tịch, màu sắc,…). Để giải quyết những vấn đề nêu trên, các nhà nghiên cứu đã đề xuất nhiều độ đo mới. Chẳng hạn, độ đo dựa vào thông tin tương hỗ [2], độ đo độ nhất quán trong lựa chọn thuộc tính [6], Chi 2 trong lựa chọn thuộc tính và rời rạc hóa [17], Relief và ReliefF để ước lượng các thuộc tính [22], độ đo độ phụ thuộc riêng phần trong lý thuyết tập thô [20, 18, 19, 11].

Trong lý thuyết tập thô, dựa trên quan hệ bất khả phân biệt, Pawlak đã đề xuất một mô hình toán học, gọi là độ phụ thuộc riêng phần γ để tính mức độ phụ thuộc của một tập thuộc tính này vào một tập thuộc tính khác [18]. Các tính chất đại số của mô hình này cũng đã được nhiều nhà nghiên cứu bàn luận [20, 18, 11, 7, 8, 6],. Khi dữ liệu chứa các giá trị phạm trù, độ phụ thuộc riêng phần γ thường được sử dụng vào việc tính toán các tập thuộc tính rút gọn, giải quyết bài toán lựa chọn thuộc tính [11, 19, 23]. Tuy nhiên, trong [8] Düntsch và Gediga đã chỉ ra rằng mô hình của Pawlak là không hoàn chỉnh (inadequate) cho việc tính toán độ phụ thuộc. Vấn đề gặp phải ở đây là, trong một số trường hợp, một thuộc tính có sự phụ thuộc vào một thuộc tính khác ở mức độ nào đó nhưng mô hình Pawlak lại cho độ phụ thuộc γ bằng 0. Chi tiết về vấn đề này có thể tham khảo các tài liệu [8, 24].

Trong những năm qua, một số mô hình tính toán độ phụ thuộc kiểu Pawlak cũng đã được đề xuất. Bhatt và Gopal [3] đã đề xuất mô hình độ phụ thuộc dựa vào xấp xỉ tập thô mờ. Mô hình này là sự mở rộng mô hình Pawlak và có thể áp dụng cho cả dữ liệu giá trị thực, tuy nhiên về bản chất nó cũng giống như mô hình của Pawlak, do đó cũng gặp phải vấn đề vừa nêu trên. Trong [4] Chen và cộng sự cũng đã đề nghị một mô hình dựa trên các tập thô mờ, trong đó độ phụ thuộc được tính toán theo một quan hệ T-tương tự mờ. Tuy nhiên, mô hình này trở thành mô hình giống như mô hình Pawlak khi quan hệ T-tương tự mờ là quan hệ tương tự rõ. Và như thế, mô hình của Chen và cộng sự cũng gặp phải vấn đề như mô hình của Pawlak. Trong [13] Hu và cộng sự đã trình bày mô hình tập thô dựa trên khoảng cách và hàm phụ thuộc giống như của Pawlak. Trong [21] Sakai và Okuma đã đề xuất một mô hình tính toán độ phụ thuộc trong bảng quyết định không nhất quán (có chứa cả giá trị tập hợp và giá trị khoảng). Thuật toán này đòi hỏi hai giá trị ngưỡng mà nếu chúng không được nạp vào một cách đúng đắn sẽ cho ra độ phụ thuộc sai lệch. Việc xác định các ngưỡng thế nào cho đúng không được bàn trong [21]. Ziarko [25,26] cũng đã đề xuất một mô hình phụ thuộc thuộc tính, gọi là hàm k-phụ thuộc, dựa vào xác suất. Mô hình này đòi hỏi một tập đích để xấp xỉ tập thô và độ phụ thuộc

được tính dựa vào tập đích đã chọn. Thế nhưng, việc xác định tập đích ra sao không được bàn tới trong [25,26]. Gần đây, Yamaguchi [24] đã đề xuất một mô hình mới tính toán độ phụ thuộc bằng cách xét đến độ hiệu quả dữ liệu. Dựa vào ma trận khả phân biệt đối với quyết định, mô hình này xem xét số lần các thuộc tính điều kiện được sử dụng để xác định giá trị của thuộc tính quyết định.

Mặc dù một số mô hình phụ thuộc đã được đề xuất như vừa trình bày trên đây, vấn đề nêu ra trong [8] hầu như vẫn chưa được giải quyết một cách triệt để.

Trong bài báo này, chúng tôi trình bày phương pháp xây dựng một độ đo mới, gọi là độ phụ thuộc Gamma, đo độ phụ thuộc giữa các tập thuộc tính phạm trù (categorical attributes) trong một hệ thông tin. Độ đo này được xây dựng dựa trên khái niệm entropy bù (complementary entropy) do Jiye Liang và cộng sự đề xuất [14, 15]. Với hai tập thuộc tính X và Y, độ đo này sẽ gán cho chúng một số thực thuộc khoảng đóng [0,1] phản ánh mức độ phụ thuộc của Y vào X. Giá trị độ đo bằng 1 khi và chỉ khi tồn tại phụ thuộc hàm $X \to Y$ trong quan hệ. Và như thế, giá trị của nó càng gần bằng 1 thì sự phụ thuộc của Y vào X trong quan hệ càng gần phụ thuộc hàm $X \to Y$. Các tính chất của độ đo phụ thuộc đề xuất và mối liên hệ của nó với phụ thuộc hàm cũng được nghiên cứu. Các tính chất này cho thấy có thể xem phụ thuộc Gamma là sự mở rộng của khái niệm phụ thuộc hàm, và độ phụ thuộc Gamma có thể được sử dụng như là một độ đo phụ thuộc hàm xấp xỉ.

Nội dung phần còn lại của bài báo này là như sau. Mục II trình bày vắn tắt một số kiến thức liên quan; mục III đưa ra định nghĩa về độ phụ thuộc Gamma và nghiên cứu các tính chất của nó; mục IV trình bày mối liên hệ giữa phụ thuộc Gamma và phụ thuộc hàm; mục V là phần kết luận trong đó nêu cả hướng nghiên cứu tiếp theo. Cuối bài báo là danh sách các tài liệu tham khảo.

II. MỘT SỐ KIẾN THỨC LIÊN QUAN

Nếu không nói gì khác, tất cả các tập hợp xét đến trong phần còn lại của bài báo là hữu hạn.

A. Phân hoạch của một tập hợp hữu hạn

Cho U là một tập hợp khác rỗng các đối tượng. Một phân hoạch của U là một họ khác rỗng các tập con $\pi = \{B_1, B_2, \dots, B_s\}$ thỏa mãn $\sum_{i=1}^{s} B_i = U$ và $B_i \cap B_j = \emptyset$ với mọi $i \neq j$. Mỗi tập con B_i được gọi là một khối hay một lớp của π. Dưới đây sẽ ký hiệu họ tất cả các phân hoạch của U là PART(U).

Trên họ các phân hoạch của một tập hợp có thể định nghĩa một quan hệ thứ tự bộ phận như sau: cho $\pi, \sigma \in$ PART(U), ta nói π mịn hơn σ và viết $\pi \leq \sigma$ nếu mỗi khối B của π đều tồn tại một khối C của σ sao cho $B \subseteq C$; nói cách khác, nếu mỗi khối C thuộc σ là hợp của một số khối thuộc π. Người ta đã chứng minh được rằng, quan hệ riêng phần này sinh ra một dàn trên PART(U), nghĩa là với hai phân hoạch bất kỳ $\pi, \sigma \in$ PART(U) luôn tồn tại một phân hoạch mịn nhất θ_1 sao cho $\pi \leq \theta_1, \sigma \leq \theta_1$ và một phân hoạch thô nhất θ_2 thỏa mãn $\theta_2 \leq \pi, \theta_2 \leq \sigma$.

B. Khái niệm entropy bù

Lý thuyết tập thô do Z. Pawlak đề xuất vào những năm đầu thập niên 80 thế kỷ XX là một công cụ cho việc xử lý dữ liệu không chắc chắn, không đầy đủ. Trong lý thuyết tập thô, một bảng dữ liệu gồm m cột ứng với m thuộc tính phạm trù, n hàng ứng với n đối tượng (bộ dữ liệu) được gọi là một hệ thống thông tin. Nếu gọi U là tập tất cả các đối tượng, A là tập tất cả các thuộc tính thì một hệ thông tin thường được ký hiệu là bộ đôi $S = (U, A)$.

Để đo đạc sự không chắc chắn và tính mờ trong lý thuyết tập thô, trong [14,15] Jiye Liang và cộng sự đã đưa ra khái niệm entropy bù (Complementary entropy) của các phân hoạch như sau.

Cho $\pi, \sigma \in$ PART(U) và giả sử $\pi = \{X_1, X_2, \dots, X_m\}$, $\sigma = \{Y_1, Y, \dots, Y_n\}$.

Định nghĩa 1 (*Entropy bù*) [14,15]. Entropy bù của phân hoạch π là đại lượng $E(\pi)$ định nghĩa bởi

$$E(\pi) = \sum_{i=1}^{m} \frac{|X_i|}{|U|} \frac{|\overline{X}_i|}{|U|},$$

trong đó $|.|$ chỉ số phần tử của một tập hợp và $\overline{X}$ là phần bù của X in U.

Dễ thấy, $E(\pi)$ có thể được viết lại như sau:

$$E(\pi) = \sum_{i=1}^{m} \frac{|X_i|}{|U|}\left(1 - \frac{|X_i|}{|U|}\right) = 1 - \frac{1}{|U|^2}\sum_{i=1}^{m} |X_i|^2.$$

Định nghĩa 2 (*Entropy bù có điều kiện*) [14,15]. Entropy bù có điều kiện của π khi đã biết σ được định nghĩa bởi:

$$E(\sigma|\pi) = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{|X_i \cap Y_j|}{|U|} \frac{|\overline{Y}_j - \overline{X}_i|}{|U|}.$$

Vì $|\overline{Y}_j - \overline{X}_i| = |\overline{Y}_j \cap X_i| = |X_i| - |X_i \cap Y_j|$, $E(\pi|\sigma)$ có thể được viết lại như sau:

$$E(\sigma|\pi) = \frac{1}{|U|^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\left|X_i \cap Y_j\right|\left|\overline{Y}_j \cap X_i\right| = \frac{1}{|U|^2}\left(\sum_{i=1}^{m}|X_i|^2 - \sum_{i=1}^{m}\sum_{j=1}^{n}\left|X_i \cap Y_j\right|^2\right).$$

Định nghĩa 3 (*Entropy bù đồng thời*) [14]. Entropy bù đồng thời của π và σ được định nghĩa bởi:

$$E(\pi,\sigma) = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left|X_i \cap Y_j\right|}{|U|}\frac{\left|\overline{X_i \cap Y_j}\right|}{|U|} = 1 - \frac{1}{|U|^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\left|X_i \cap Y_j\right|^2.$$

Từ định nghĩa, suy ra $E(\pi,\sigma) = E(\sigma,\pi)$. Và nếu đặt

$$\pi \wedge \sigma = \left\{X_i \cap Y_j \,\middle|\, i = 1,\dots,m; j = 1,\dots,n, X_i \cap Y_j \neq \emptyset\right\}$$

thì $\pi \wedge \sigma$ là một phân hoạch của U. Rõ ràng $\pi \wedge \sigma \leq \pi, \sigma$ và ta có:

$$E(\pi,\sigma) = E(\pi \wedge \sigma).$$

Định nghĩa 4 (*Entropy bù tương hỗ*) [14]. Entropy bù tương hỗ của π và σ được định nghĩa bởi:

$$I(\pi;\sigma) = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left|X_i \cap Y_j\right|}{|U|}\frac{\left|\overline{X}_i \cap \overline{Y}_j\right|}{|U|}.$$

Dễ thấy $I(\pi;\sigma)$ có tính đối xứng và

$$I(\pi;\sigma) = E(\sigma) - E(\sigma|\pi) = E(\pi) - E(\pi|\sigma).$$

Cũng như Shannon entropy [27], entropy bù E có các tính chất sau đây.

Mệnh đề 1 (*Giá trị nhỏ nhất, lớn nhất*) [1,14]. Với mọi $\pi \in \mathrm{PART}(U)$, ta đều có $0 \leq E(U) \leq 1 - 1/|U|$. Giá trị nhỏ nhất 0 đạt được khi và chỉ khi $\pi = \omega_U = \{U\}$, còn giá trị lớn nhất $1 - 1/|U|$ đạt được khi và chỉ khi $\pi = \alpha_U = \{\{x\} \,|\, x \in U\}$.

Mệnh đề 2 (*Tính đơn điệu*) [1,14]. Cho $\pi, \sigma \in \mathrm{PART}(U)$.

a) Nếu $\pi < \sigma$ thì $E(\pi) > E(\sigma)$.

b) Nếu $\pi \leq \sigma$ và $E(\pi) = E(\sigma)$ thì $\pi = \sigma$.

Chú ý rằng, nói chung nếu chỉ có $E(\pi) = E(\sigma)$ thì chưa suy ra được $\pi = \sigma$.

Mệnh đề 3 [1]. Cho $\pi, \sigma \in \mathrm{PART}(U)$. Ta có

$$E(\pi,\sigma) = E(\pi) + E(\sigma|\pi) = E(\sigma) + E(\pi|\sigma)\,.$$

Mệnh đề 4 [1]. Cho $\pi, \sigma \in \mathrm{PART}(U)$. Ta có

$$I(\pi;\sigma) = I(\sigma;\pi) = E(\pi) + E(\sigma) - E(\pi,\sigma)\,.$$

Mệnh đề 5 (*Giá trị nhỏ nhất, lớn nhất của entropy bù có điều kiện*)**.** Với mọi $\pi, \sigma \in \mathrm{PART}(U)$ ta đều có

$$0 \leq\ E(\sigma|\pi) \leq 1 - \frac{1}{|U|}\,;$$

$E(\sigma|\pi) = 0$ khi và chỉ khi $\pi \leq \sigma$; $E(\sigma|\pi) = 1 - 1/|U|$ khi và chỉ khi $\sigma = \alpha_U$ và $\pi = \omega_U$.

Chứng minh. Hiển nhiên ta có $E(\sigma|\pi) \geq 0$. Theo Mệnh đề 3,

$$E(\sigma|\pi) = E(\pi,\sigma) - E(\pi).$$

Thế thì

$$E(\sigma|\pi) = 0 \Leftrightarrow\ E(\pi,\sigma) = E(\pi) \Leftrightarrow\ E(\pi \wedge \sigma) = E(\pi).$$

Vì $\pi \wedge \sigma \leq \pi$, theo Mệnh đề 2, ta có

$$E(\pi \wedge \sigma) = E(\pi) \Leftrightarrow\ \pi \wedge \sigma = \pi\ \Leftrightarrow\ \pi \leq \sigma\,.$$

Vậy, $E(\sigma|\pi) = 0$ khi và chỉ khi $\pi \leq \sigma$.

Mặt khác, theo Mệnh đề 1,

$$E(\pi,\sigma) = E(\pi \wedge \sigma) \leq 1 = \frac{1}{|U|} \text{ and } E(\sigma) \geq 0\,.$$

Suy ra

$$E(\sigma|\pi) = E(\pi,\sigma) - E(\pi) \le 1 - \frac{1}{|U|}.$$

Dấu "=" xảy ra khi và chỉ khi

$$\begin{cases} E(\pi) = 0 \\ E(\pi \wedge \sigma) = 1 - \dfrac{1}{|U|} \end{cases} \Leftrightarrow \begin{cases} \pi = \omega_U \\ \sigma = \alpha_U \end{cases} \quad \square$$

Mệnh đề 6 (*Giá trị nhỏ nhất, lớn nhất của entropy bù đồng thời*). Cho $\pi, \sigma \in \mathrm{PART}(U)$. Khi đó

$$\max(E(\pi), E(\sigma)) \le E(\pi,\sigma) \le E(\pi) + E(\sigma).$$

Chứng minh. Vế trái $\max(E(\pi), E(\sigma)) \le E(\pi,\sigma)$ suy ra từ các Mệnh đề 1, 3 và 5. Vế phải $E(\pi,\sigma) \le E(\pi) + E(\sigma)$ suy ra từ Mệnh đề 4 và Định nghĩa 4.□

III. ĐỘ ĐO ĐỘ PHỤ THUỘC GAMMA

A. Định nghĩa độ phụ thuộc Gamma

Cho hệ thống thông tin $S = (U, A)$, trong đó U là tập tất cả các đối tượng, A là tập tất cả các thuộc tính. Các tập con thuộc tính trong A có mối liên kết tự nhiên với các phân hoạch của U: mỗi tập con thuộc tính tạo ra một phân hoạch trên U, trong đó hai đối tượng sẽ thuộc vào cùng một khối nếu chúng có cùng giá trị về tập thuộc tính đó.

Dưới đây, để cho tiện, ta sẽ viết hợp của các tập con thuộc tính, chẳng hạn của X và Y là XY. Phân hoạch trên U sinh ra bởi tập thuộc tính X là π_X.

Chú ý rằng đối với một cơ sở dữ liệu quan hệ, π_X là phân hoạch của tập các hàng trong một bảng có thể thu được bằng cách sử dụng tùy chọn **group by *X*** trong SQL.

Cho hai tập con thuộc tính $X, Y \subseteq A$. Giả sử các phân hoạch trên U sinh bởi X và Y lần lượt là $\pi_X = \{X_1, X_2, \dots, X_m\}$ và $\pi_Y = \{Y_1, Y_2, \dots, Y_n\}$. Khi đó, phân hoạch trên U sinh bởi XY sẽ là

$$\pi_{XY} = \pi_X \wedge \pi_Y = \{X_i \cap Y_j \mid i = 1, \dots, m; j = 1, \dots, n, X_i \cap Y_j \ne \emptyset\}.$$

Định nghĩa 5. Cho hai tập con thuộc tính $X, Y \subseteq A$. Giả sử các phân hoạch trên U sinh bởi X và Y lần lượt là $\pi_X = \{X_1, X_2, \dots, X_m\}$ và $\pi_Y = \{Y_1, Y_2, \dots, Y_n\}$. Ta gọi độ phụ thuộc của Y vào X là đại lượng $\Gamma(X,Y)$ xác định như sau:

$$\Gamma(X,Y) = 1 - \frac{|U|}{|U|-1} E(\pi_Y|\pi_X) = 1 - \frac{1}{|U|(|U|-1)} \left(\sum_{i=1}^{m} |X_i|^2 - \sum_{i=1}^{m}\sum_{j=1}^{n} |X_i \cap Y_j|^2 \right).$$

Ví dụ: Xét bảng quyết định cho trong Bảng 1.

Bảng 1. Bảng quyết định của Düntsch [8].

x	c_1	c_2	d
x_1	0	0	0
x_2	0	2	0
x_3	0	2	0
x_4	1	1	0
x_5	1	0	1
x_6	1	2	1
x_7	1	2	1
x_8	0	1	1

Ở đây, ta có: $|U| = 8$,

$$\pi_{c1} = \{X_1 = \{x_1, x_2, x_3, x_8\},\ X_2 = \{x_4, x_5, x_6, x_7\}\},\ \pi_d = \{Y_1 = \{x_1, x_2, x_3, x_4\}, Y_2 = \{x_5, x_6, x_7, x_8\}\}.$$

$$\Gamma(\{c_1\},\{d\}) = 1 - \frac{1}{|U|(|U|-1)} \left(\sum_{i=1}^{m} |X_i|^2 - \sum_{i=1}^{m}\sum_{j=1}^{n} |X_i \cap Y_j|^2 \right)$$

$$= 1 - \frac{1}{8 \times 7} \left((4^2 + 4^2) - \left((3^2 + 1^2) + (1^2 + 3^2) \right) \right) = \frac{11}{14}.$$

Chú ý rằng, nếu tính theo mô hình Pawlak, ta có $\gamma(\{c_1\},\{d\}) = 0$ (xem [8]).

B. Các tính chất

Mệnh đề 7 (*Giá trị nhỏ nhất, lớn nhất của độ phụ thuộc Gamma*). $0 \le \Gamma(X,Y) \le 1$.

Chứng minh. Theo Mệnh đề 6: $E(\pi_Y|\pi_X) = 0$ khi và chỉ khi $\pi_X \le \pi_Y$; $E(\pi_Y|\pi_X) = 0$ khi và chỉ khi $\pi_X = \omega_U$ và $\pi_Y = \alpha_U$. Suy ra, $\Gamma(X,Y) = 1$ khi và chỉ khi $\pi_X \le \pi_Y$; $\Gamma(X,Y) = 0$ khi và chỉ khi $\pi_X = \omega_U$ và $\pi_Y = \alpha_U$.□

Mệnh đề 8 (*Quy tắc phản xạ*). Nếu $Y \subseteq X \subseteq A$ thì $\Gamma(X,Y) = 1$.

Chứng minh. Nếu $Y \subseteq X$ thì $\pi_X \le \pi_Y$. Vậy theo Mệnh đề 7, $\Gamma(X,Y) = 1$. □

Mệnh đề 9. Cho ba tập con thuộc tính $X, Y, Z \subseteq A$. Ta có $\Gamma(XZ, YZ) = \Gamma(XZ, Y)$.

Chứng minh.

$$\begin{aligned} E(\pi_{YZ}|\pi_{XZ}) &= E(\pi_{XYZ}) - E(\pi_{XZ}) \text{ (Mệnh đề 3)} \\ &= E(\pi_Y|\pi_{XZ}) + E(\pi_{XZ}) - E(\pi_{XZ}) \text{ (Mệnh đề 3)} \\ &= E(\pi_Y|\pi_{XZ}). \end{aligned}$$

Suy ra,

$$\Gamma(XZ, YZ) = 1 - \frac{|U|}{|U|-1} E(\pi_{YZ}|\pi_{XZ}) = 1 - \frac{|U|}{|U|-1} E(\pi_Y|\pi_{XZ}) = \Gamma(XZ, Y)\,.\square$$

Mệnh đề 10 (*Quy tắc hợp phải*). Cho ba tập con thuộc tính $X, Y, Z \subseteq A$. Giả sử $\pi_X = \{X_1, X_2, \dots, X_m\}$, $\pi_Y = \{Y_1, Y_2, \dots, Y_n\}$ và $\pi_Z = \{Z_1, Z_2, \dots, Z_l\}$. Khi đó,

$$\Gamma(X,Y) + \Gamma(X,Z) \le \Gamma(X, YZ) + 1.$$

Chứng minh. Theo Định nghĩa 2, ta có

$$\begin{aligned} E(\pi_Y|\pi_X) + E(\pi_Z|\pi_X) &= \frac{1}{|U|^2}\sum_{i=1}^{m}\sum_{j=1}^{n}|X_i \cap Y_j| \cdot |\overline{Y}_j \cap X_i| + \frac{1}{|U|^2}\sum_{i=1}^{m}\sum_{k=1}^{l}|X_i \cap Z_k| \cdot |\overline{Z}_k \cap X_i| \\ &= \frac{1}{|U|^2}\sum_{i=1}^{m}\sum_{j=1}^{n}|\overline{Y}_j \cap X_i| \cdot \sum_{k=1}^{l}|X_i \cap Y_j \cap Z_k| + \frac{1}{|U|^2}\sum_{i=1}^{m}\sum_{k=1}^{l}|\overline{Z}_k \cap X_i| \sum_{j=1}^{n}|X_i \cap Y_j \cap Z_k| \\ &= \frac{1}{|U|^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{l}|X_i \cap Y_j \cap Z_k|\left(|\overline{Y}_j \cap X_i| + |\overline{Z}_k \cap X_i|\right) \\ &\ge \frac{1}{|U|^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{l}|X_i \cap (Y_j \cap Z_k)||(\overline{Y}_j \cup \overline{Z}_k) \cap X_i| \\ &= \frac{1}{|U|^2}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{l}|X_i \cap (Y_j \cap Z_k)||(\overline{Y_j \cap Z_k}) \cap X_i| \\ &= E(\pi_{YZ}|\pi_X)\,. \end{aligned}$$

Do đó

$$\frac{|U|}{|U|-1}E(\pi_Y|\pi_X) + \frac{|U|}{|U|-1}E(\pi_Z|\pi_X) \ge \frac{|U|}{|U|-1}E(\pi_{YZ}|\pi_X)$$

$$\left(1 - \frac{|U|}{|U|-1}E(\pi_Y|\pi_X)\right) + \left(1 - \frac{|U|}{|U|-1}E(\pi_Z|\pi_X)\right) \le 1 + \left(1 - \frac{|U|}{|U|-1}E(\pi_{YZ}|\pi_X)\right)$$

$\Gamma(X,Y) + \Gamma(X,Z) \le \Gamma(X, YZ) + 1\,.\square$

Mệnh đề 11 (*Quy tắc xích*). $\Gamma(X, YZ) = \Gamma(X,Y) + \Gamma(XY, Z) - 1$.

Chứng minh. Áp dụng liên tiếp Mệnh đề 3:

$$E(\pi_{YZ}|\pi_X) = E(\pi_{XYZ}) - E(\pi_X)$$

$$= E(\pi_Z|\pi_{XY}) + E(\pi_{XY}) - E(\pi_X)$$

$$= E(\pi_Z|\pi_{XY}) + E(\pi_Y|\pi_X).$$

Suy ra,

$$\Gamma(X,YZ) = 1 - \frac{|U|}{|U|-1}E(\pi_{YZ}|\pi_X) = 1 - \frac{|U|}{|U|-1}\big(E(\pi_Z|\pi_{XY}) + E(\pi_Y|\pi_X)\big)$$

$$= \left(1 - \frac{|U|}{|U|-1}\,E(\pi_Z|\pi_{XY})\right) + \left(1 - \frac{|U|}{|U|-1}E(\pi_Y|\pi_X)\right) - 1$$

$$= \Gamma(X,Y) + \Gamma(XY,Z) - 1\,.\square$$

Mệnh đề 12. $\Gamma(XY,Z) \geq \Gamma(X,Z)$.

Chứng minh.

$$\Gamma(XY,Z) = \Gamma(X,YZ) - \Gamma(X,Y) + 1 \text{ (Mệnh đề 11)}$$

$$\geq \Gamma(X,Y) + \Gamma(X,Z) - \Gamma(X,Y) \text{ (Quy tắc hợp phải, Mệnh đề 10)}$$

$$= \Gamma(X,Z)\ .\square$$

Mệnh đề 13 (*Quy tắc hợp trái*)**.** $\text{Max}\big(\Gamma(X,Z),\Gamma(Y,Z)\big) \leq \Gamma(XY,Z)$.

Chứng minh. Theo Mệnh đề 12:

$$\Gamma(X,Z) \leq \Gamma(XY,Z),$$

$$\Gamma(Y,Z) \leq \Gamma(XY,Z).$$

Vậy, $\text{Max}\big(\Gamma(X,Z),\Gamma(Y,Z)\big) \leq \Gamma(XY,Z)$. □

Mệnh đề 14 (*Quy tắc gia tăng*)**.** $\Gamma(XZ,YZ) \geq \Gamma(X,Y)$.

Chứng minh. Ta có:

$$\Gamma(XZ,YZ) = \Gamma(XZ,Y) \text{ (Mệnh đề 9)}$$

$$\geq \Gamma(X,Y) \text{ (Mệnh đề 12) .}\square$$

Mệnh đề 15 (*Quy tắc bắc cầu*)**.** $\Gamma(X,Y) + \Gamma(Y,Z) \leq \Gamma(X,Z) + 1$.

Chứng minh. Ta có:

$$\Gamma(X,Y) + \Gamma(Y,Z) \leq \Gamma(X,Y) + \Gamma(XY,XZ) \text{ (Quy tắc gia tăng, Mệnh đề 14)}$$

$$= \left(1 - \frac{|U|}{|U|-1}E(\pi_Y|\pi_X)\right) + \left(1 - \frac{|U|}{|U|-1}E(\pi_{XZ}|\pi_{XY})\right)$$

$$= \left(1 - \frac{|U|}{|U|-1}\big(E(\pi_{XY}) - E(\pi_X)\big)\right) + \left(1 - \frac{|U|}{|U|-1}\big(E(\pi_{XYZ}) - E(\pi_{XY})\big)\right) \text{(Mệnh đề 3)}$$

$$= 1 + \left(1 - \frac{|U|}{|U|-1}\big(E(\pi_{XYZ}) - E(\pi_X)\big)\right)$$

$$\leq 1 + \left(1 - \frac{|U|}{|U|-1}\big(E(\pi_{XZ}) - E(\pi_X)\big)\right) \text{(Mệnh đề 6)}$$

$$= 1 + \left(1 - \frac{|U|}{|U|-1}E(\pi_Z|\pi_X)\right) \text{(Mệnh đề 3)}$$

$$= \Gamma(X,Z) + 1\ .\square$$

Mệnh đề 16 (*Quy tắc hợp toàn phần*)**.** $\Gamma(X,Y) + \Gamma(W,Z) \leq \Gamma(XW,YZ) + 1$.

Chứng minh.

$$\Gamma(X,Y) + \Gamma(W,Z) \le \Gamma(XW,YW) + \Gamma(WY,YZ) \text{ (Quy tắc gia tăng, Mệnh đề 14)}$$

$$\le \Gamma(XW,YZ) + 1 \text{ (Quy tắc bắc cầu, Mệnh đề 15). } \square$$

Mệnh đề 17 (*Quy tắc tách*). Nếu $Z \subseteq Y$ thì $\Gamma(X,YZ) \le \Gamma(X,Z)$.

Chứng minh.

$$\Gamma(X,YZ) + \Gamma(YZ,Z) \le \Gamma(X,Z) + 1 \text{ (Quy tắc bắc cầu, Mệnh đề 15) .}$$

Vì $\Gamma(YZ,Z) = 1$ (Quy tắc phản xạ, Mệnh đề 8), ta có $\Gamma(X,YZ) \le \Gamma(X,Z)$. □

Mệnh đề 18 (*Quy tắc giả bắc cầu*). $\Gamma(X,Y) + \Gamma(WY,Z) \le \Gamma(XW,Z) + 1$.

Chứng minh.

$$\Gamma(X,Y) + \Gamma(WY,Z) \le \Gamma(XW,YW) + \Gamma(WY,Z) \text{ (Quy tắc gia tăng, Mệnh đề 14)}$$

$$\le \Gamma(WY,Z) + 1 \text{ (Quy tắc bắc cầu, Mệnh đề 15) . } \square$$

IV. MỐI LIÊN HỆ GIỮA PHỤ THUỘC GAMMA VÀ PHỤ THUỘC HÀM

Một quan hệ r xác định trên tập thuộc tính A có thể được xem như một hệ thông tin $S = (U, A)$. Tuy nhiên, khái niệm hệ thống thông tin là tổng quát hơn, do các đối tượng ở đây được xem là những phần tử của U thay vì là những bộ giá trị gồm $|A|$ thành phần [20].

Các phụ thuộc hàm đã được nghiên cứu kỹ trong nhiều tài liệu. Cho quan hệ r xác định trên tập thuộc tính A. Với hai tập thuộc tính $X, Y \subseteq A$, ta nói Y phụ thuộc hàm vào X, viết $X \to Y$, nếu mỗi bộ giá trị X cho ta một bộ giá trị duy nhất của Y. Có thể thấy sự phụ thuộc Gamma nghiên cứu trên đây là một mở rộng của phụ thuộc hàm.

A . Mối liên hệ

Mệnh đề 19. Cho hai tập thuộc tính $X, Y \subseteq A$. $X \to Y$ thỏa mãn khi và chỉ khi $\Gamma(X,Y) = 1$.

Chứng minh. Giả sử các phân hoạch trên U sinh ra bởi X và Y lần lượt là $\pi_X = \{X_1, X_2, \dots, X_m\}$, $\pi_Y = \{Y, Y_2, \dots, Y_n\}$.

$\Rightarrow$: Nếu $X \to Y$ thì với mỗi bộ giá trị $x_i \in dom(X)$ chỉ có tương ứng một bộ giá trị duy nhất $y_j \in dom(Y)$. Suy ra, $\pi_X \le \pi_Y$. Tức là với mỗi khối $X_i \in \pi_X$ chỉ tồn tại duy nhất một khối $Y_j \in \pi_Y$ thỏa mãn $X_i \subseteq Y_j$. Do đó,

$$\sum_{j=1}^{n} |X_i \cap Y_j|^2 = |X_i|^2$$

Khi đó,

$$E(\pi_Y | \pi_X) = \frac{1}{|U|^2} \sum_{i=1}^{m} \left(|X_i|^2 - \sum_{j=1}^{n} |X_i \cap Y_j|^2 \right) = \frac{1}{|U|^2} \sum_{i=1}^{m} (|X_i|^2 - |X_i|^2) = 0 .$$

Suy ra

$$\Gamma(X,Y) = 1 - \frac{|U|}{|U|-1} E(Y|X) = 1$$

$\Leftarrow$: Nếu $\Gamma(X,Y) = 1$ thì $E(\pi_Y|X) = 0$. Suy ra,

$$|X_i|^2 - \sum_{j=1}^{m} |X_i \cap Y_j|^2 = 0$$

với mọi $i = 1,2, \dots, n$. Điều này chỉ có thể xảy ra nếu mỗi khối $X_i \in \pi_X$ chỉ tồn tại duy nhất một khối $Y_j \in \pi_Y$ thỏa mãn $X_i \subseteq Y_j$. Tức là có phụ thuộc hàm $X \to Y$.□

B . Các tiên đề Armstrong

Các tiên đề Armstrong là rất quan trọng đối với lý thuyết phụ thuộc hàm vì chúng cung cấp cơ sở cho hệ thống suy diễn phụ thuộc. Thông thường các tiên đề Armstrong bao gồm 3 quy Quy tắc chính sau đây [5].

1. *Quy tắc phản xạ:* Nếu $Y \subseteq X$ thì $X \to Y$
2. *Quy tắc tăng trưởng*: Nếu $X \to Y$ thì $XZ \to YZ$
3. *Quy tắc bắc cầu*: Nếu $X \to Y$ và $Y \to Z$ thì $X \to Z$.

Mệnh đề 20. Các tiên đề Armstrong suy ra trực tiếp từ các bất đẳng thức phụ thuộc Gamma.

Chứng minh. 1. Tính phản xạ: Theo Mệnh đề 4, nếu $Y \subseteq X$ thì $\Gamma(X,Y) = 1$. Lại theo Mệnh đề 19, từ $\Gamma(X,Y) = 1$ suy ra $X \to Y$.

2. Quy tắc tăng trưởng: Nếu $X \to Y$ thì theo Mệnh đề 15 ta có $\Gamma(X,Y) = 1$. Do $\Gamma(XZ,YZ) \geq \Gamma(X,Y)$ (theo Mệnh đề 10), suy ra $\Gamma(XZ,YZ) = 1$ (vì $\Gamma(XZ,YZ) \leq 1$). Lại áp dụng Mệnh đề 19, suy ra $XZ \to YZ$.

3. Quy tắc bắc cầu: Nếu $X \to Y$ và $Y \to Z$ thì $\Gamma(X,Y) = \Gamma(Y,Z) = 1$ (theo Mệnh đề 19). Vì $\Gamma(X,Y) + \Gamma(Y,Z) \leq \Gamma(X,Z) + 1$ (theo Mệnh đề 11), suy ra $\Gamma(X,Z) = 1$. Do đó $X \to Z$ (theo Mệnh đề 19). □

V. KẾT LUẬN

Phụ thuộc giữa các tập thuộc tính trong một cơ sở dữ liệu là một dạng tri thức hữu ích tiềm ẩn. Để phát hiện hiệu quả các phụ thuộc thì việc xây dựng các độ đo (các hàm) cho phép đánh giá đúng mức độ phụ thuộc là điều rất quan trọng. Trong báo cáo này, chúng tôi trình bày phương pháp xây dựng một độ đo mới, gọi là độ phụ thuộc Gamma, đo độ phụ thuộc giữa các tập thuộc tính phạm trù trong một hệ thông tin. Độ đo này được xây dựng dựa trên khái niệm entropy bù (complementary entropy) do Jiye Liang và cộng sự đề xuất. Với hai tập thuộc tính X và Y, độ đo này sẽ gán cho chúng một số thực thuộc khoảng đóng [0,1] phản ánh mức độ phụ thuộc của Y vào X. Giá trị độ đo bằng 1 khi và chỉ khi tồn tại phụ thuộc hàm $X \to Y$ trong quan hệ. Và như thế, giá trị của nó càng gần bằng 1 thì sự phụ thuộc của Y vào X trong quan hệ càng gần phụ thuộc hàm $X \to Y$. Các tính chất của độ đo phụ thuộc đề xuất và mối liên hệ của nó với phụ thuộc hàm cũng được nghiên cứu. Các tính chất này cho thấy có thể xem nó là sự mở rộng của khái niệm phụ thuộc hàm, và độ phụ thuộc Gamma có thể được sử dụng như là một độ đo phụ thuộc hàm xấp xỉ.

Dựa trên các kết quả đã nghiên cứu được về độ đo độ phụ thuộc Gamma, trong thời gian tới, chúng tôi sẽ nghiên cứu thuật toán khai phá các phụ thuộc Gamma với ngưỡng phụ thuộc cho trước; tiến hành thử nghiệm sử dụng độ đo Gamma thay cho information gain trong thuật toán xây dựng cây quyết định C4.5.

VI. TÀI LIỆU THAM KHẢO

[1] Nguyễn Thanh Tùng, Về một metric trên họ các phân hoạch của một tập hợp hữu hạn. *Tạp chí Tin học và Điều khiển học*, Vol. 26, Nr. 1, pp. 75-87, 2010.

[2] Battiti, R., Using mutual information for Selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5, pp. 537–550, 1994.

[3] R. B. Bhatt, M. Gopal, On the extension of functional dependency degree from crisp to fuzzy partitions, *Pattern Recognition Letters* 27,pp. 487–491, 2006.

[4] D. Chen, W. Yang, F. Li, Measures of general fuzzy rough sets on a probabilistic space, *Information Sciences* 178, pp. 3177–3187, 2008.

[5] E. F. Codd, A relational model of data for large shared data banks, *Communications of the ACM* 13,pp. 377–387, 1970.

[6] Dash, M., & Liu, H., Consistency-based search in feature selection. *Artificial Intelligence*, 151(1–2), pp. 155–176, 2003.

[7] I. Düntsch, G. Gediga, Algebraic aspects of attribute dependencies, *Fundamenta Informaticae* 29, pp. 119–133, 1997.

[8] I. Düntsch, G. Gediga, Statistical evaluation of rough set dependency analysis, *International Journal of Human–Computer Studies* 46, pp. 589–604, 1997.

[9] Hall, M. A., Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings 17th international conference on machine learning*, pp. 359–366, 2000.

[10] Han J., and Kamber M., *Data Mining: Concepts and Techniques*, Morgan Kanufmann, 2012.

[11] Hu, X. H., & Cercone, N. Learning in relational databases: A rough set approach. *Computational Intelligence*, 12(2), pp. 323–338, 1995.

[12] Hu, Q. H., Xie, Z. X., & Yu, D. R., Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation. *Pattern Recognition*, 40(12), pp. 3509–3521, 2007.

[13] Q. Hu, D. Yu, J. Liu, C. Wu, Neighborhood rough set based heterogeneous feature subset selection, *Information Sciences,* 2008, doi:10.1016/j.ins.2008.05.024.

[14] J. Liang, K. S. Chin, Chuangyin Dang, Richard C. M. Yam, A new method for measuring uncertainty and fuzziness in rough set theory. *International Journal of General Systems*, Vol. 31 (4), pp. 331-342, 2002.

[15] Jiye Liang, Uncertainty and Feature Selection in Rough Set Theory. In J. T. Yao et al. (Eds): *RSKT 2011*, *LNCS 6954*, pp. 8–15, 2011.

[16] Kivinen, J., Mannila, H., "Approximate inference of functional dependencies from relations", *Theoretical Computer Science* 149(1),pp. 129-149, 1997.

[17] Liu, H., & Setiono, R., Feature selection via discretization of numeric attributes.*IEEE Transactions on Knowledge and Data Engineering*, 9(4), pp. 642–645.

[18] M. Novotny, Z. Pawlak, Partial dependency of attributes, *Bulletin of the Polish Academy of Sciences Mathematics* 36,pp. 453–458, 1988.

[19] Z. Pawlak, Rough Sets: *Theoretical Aspects of Reasoning About Data*, Kluwer Academic Publishers, 1991.

[20] Z. Pawlak, C. Rauszer, Dependency of attributes in information systems, *Bulletin of the Polish Academy of Sciences Mathematics* 33,pp. 551–559, 1985.

[21] H. Sakai, A. Okuma, An algorithm for checking dependencies of attributes in a table with non-deterministic information: a rough sets based approach, in: R. Mizoguchi, J. Slaney (Eds.), *Proceedings of Sixth Pacific Rim International Conference on Artificial Intelligence,* PRICAI2000, LNAI1886, pp. 219-229, 2000.

[22] Sikonja, M. R., & Kononenko, I., Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, 53, pp. 23–69, 2003.

[23] R. W. Swiniarski, A. Skowron, Rough set methods in feature selection and recognition, Pattern Recognition Letters 24, pp. 833–849, 2003.

[24] D. Yamaguchi, Attribute dependency functions considering data efficiency. *International Journal of Approximate Reasoning*, 51, pp. 89-98, 2009.

[25] W. Ziarko, Dependencies in structures of decision tables, in Krzyszkievicz et al. (Eds.), *Proceeding of the International Conference on Rough Sets and Emerging Intelligent Systems paradigms*, RSEISP'07, Warsaw, Poland, LNAI4585, pp. 113-121, 2007.

[26] W. Ziarko, Probabilistic approach to Rough sets, *International Journal of Approximate Reasoning*, 49, 2008, pp. 272-284, 2008.

[27] Dalkilic, M. M., Robertson, E. L. "Information dependencies", In: *Proceedings of ACM PODS,* 245-253, 2000.

A NEW MEASURE FOR MEASURING ATTRIBUTE DEPENDENCIES

Nguyen Minh Huy, Đo Si Truong, Nguyen Huy Duc, Nguyen Thanh Tung

ABSTRACT *- In this paper, we propose a new dependency measure, called Gamma, to measure dependency degree between two given sets of categorical attributes in an information system. The proposed measure is based on the concept of complementary entropy introduced by Jiye Liang et al. For two sets of attributes X and Y, this measure maps them to a real number in the closed interval [0;1] describing the dependency degree of Y on X. The mapped number equal to 1 if and only if there exists functional dependency* $X \rightarrow Y$*. Hence, the smaller the number to which X and Y are mapped, the "closer"* $X \rightarrow Y$ *is to being a functional dependency in the information system. The properties of the proposed measure and its relationship with functional dependency have also been investigated. These properties show that we can consider Gamma dependency as an extension of the concept of functional dependency, and it can be used as an approximation measure for functional dependencies.*