Tạp chí Tin học và Điều khiển học, T.26, S.1 (2010), 73–85

VỀ MỘT METRIC TRÊN HỌ CÁC PHÂN HOẠCH CỦA MỘT TẬP HỢP HỮU HẠN

NGUYỄN THANH TÙNG

Viện Công nghệ thông tin, Viện Khoa học và Công nghệ Việt Nam

Abstract. In a database, partitions of the set of objects are naturally associated with their attributes; each attribute induces a partition of the set of objects, where two objects belong to the same block if they have identical values for that attribute. So, any metric defined on the set of partitions of the set of objects generates a metric on the set of attributes. Once a metric is defined, we can evaluate how far these attributes are, cluster the attributes, find centrally located attributes and so on. All these possibilities can be exploited for improving existing data mining algorithms and for formulating new ones.

The purpose of this paper is to define a metric on the set of partitions of a finite set starting from the concept of information entropy proposed by Jiye Liang et. al.

Tóm tắt. Trong một cơ sở dữ liệu, các phân hoạch của tập các đối tượng có mối liên kết tự nhiên với các thuộc tính; mỗi thuộc tính tạo ra một phân hoạch của tập các đối tượng, trong đó hai đối tượng sẽ thuộc vào cùng một khối nếu chúng có chung giá trị về thuộc tính đó. Như vậy, khi một metric nào đó được định nghĩa trên tập các phân hoạch của tập các đối tượng thì cũng có nghĩa là một metric đã được xác lập trên tập các thuộc tính. Một khi đã có metric, ta có thể đánh giá độ gần nhau, phân cụm các thuộc tính, xác định thuộc tính trung tâm, thuộc tính quan trọng,... Tất cả khả năng này có thể được khai thác, sử dụng vào việc nâng cao độ hiệu quả của các thuật toán khai phá dữ liệu đã có hay tạo ra những thuật toán mới.

Bài báo đề xuất phương pháp xây dựng một metric trên tập các phân hoạch của một tập hợp hữu hạn các đối tượng, xuất phát từ độ đo entropy thông tin do Jiye Liang và cộng sự đề xuất.

1. MỞ ĐẦU

Kỹ thuật sử dụng metric đóng vai trò quan trọng trong khai phá dữ liệu. Trong những năm gần đây, kỹ thuật này được nhiều người quan tâm nghiên cứu và áp dụng vào việc giải quyết những vấn đề lớn của khai phá dữ liệu như phân lớp, phân cụm dữ liệu, lựa chọn đặc trưng, rời rạc hóa dữ liệu,...

Khi tất cả các thuộc tính của các đối tượng trong cơ sở dữ liệu cần khai phá đều là những thuộc tính giá trị thực, các đối tượng hay các thuộc tính có thể được biểu diễn bằng các điểm trong một không gian $\mathbb{R}^n$ và để đánh giá độ giống nhau giữa các đối tượng, thuộc tính, một số metric có thể được sử dụng, thường là khoảng cách Euclide. Tuy nhiên, trong thực

*Nghiên cứu này được hoàn thành dưới sự hỗ trợ từ quỹ NAFOSTED

hành, một cơ sở dữ liệu thường có nhiều loại thuộc tính khác nhau: định lượng, định tính (như màu sắc, hình dạng, kết quả xét nghiệm,...). Đối với những trường hợp dữ liệu như thế, hiện nay, ngoài khoảng cách Hamming, có rất ít metric có thể lựa chọn. Việc nghiên cứu tìm ra metric thích hợp để giải quyết một nhiệm vụ cụ thể luôn một vấn đề cần được quan tâm.

Người đầu tiên ứng dụng kỹ thuật metric giữa các phân hoạch của một tập hợp vào khai phá dữ liệu là R. López de Mántaras [5]. Ông đã xây dựng một metric giữa các phân hoạch của một tập hợp hữu hạn các đối tượng, từ đó đề xuất một tiêu chuẩn lựa chọn thuộc tính phân chia các đối tượng tại mỗi nút trong quá trình xây dựng cây quyết định giải bài toán phân lớp. Kết quả phân lớp thu được sử dụng tiêu chuẩn này, trong nhiều trường hợp, tốt hơn so với kết quả dựa vào tiêu chuẩn entropy gain (hay entropy gain ratio).

Việc ứng dụng kỹ thuật metric giữa các phân hoạch trong khai phá dữ liệu nảy sinh từ một ý tưởng rất đơn giản, đó là mỗi thuộc tính đều sinh ra một phân hoạch trên tập các đối tượng, trong đó hai đối tượng sẽ thuộc vào cùng một khối nếu chúng có cùng giá trị về thuộc tính đó. Vì vậy, khi một metric nào đó được định nghiã trên tập các phân hoạch của tập các đối tượng thì nó cũng sẽ là một metric trên tập các thuộc tính. Khi một metric đã được định nghiã, ta có thể sử dụng nó đánh giá được sự khác nhau giữa các thuộc tính, phân cụm các thuộc tính, phát hiện các thuộc tính quan trọng,... Nhờ đó, có thể cải thiện độ hiệu quả của các thuật toán đã có hoặc xây dựng các thuật toán mới giải quyết các bài toán khai phá dữ liệu.

Những đóng góp quan trọng trong nghiên cứu xây dựng metric trên họ các phân hoạch của một tập hữu hạn phải kể đến các công trình của J. P. Barthélemy [2], B. Monjardet [4], Barthélemy và Leclerc [3], trong đó metric trên tập các phân hoạch của một tập hợp được xây dựng dựa trên các lớp tương đương giữa các đối tượng hoặc các độ đo thông tin của các phân hoạch.

Tư tưởng chung trong việc lựa chọn thuộc tính phân chia các đối tượng tại mỗi nút trong quá trình xây dựng cây quyết định là sử dụng thuộc tính xấp xỉ tốt nhất thuộc tính quyết định (nhãn lớp) trên tập các đối tượng cần phân chia. Điều này có thể thực hiện bằng cách sử dụng một metric định nghĩa trên các phân hoạch. Gần đây, xuất phát từ khái niệm entropy có điều kiện tổng quát do D. A. Simovici đề xuất, D. A. Simovici và S. Jaroszewicz [6, 8], D. A. Simovici, Singla et. al. [7], R. A. Butterworth [10] đã xây dựng được một họ các metric phụ thuộc vào một tham số. Tùy thuộc vào đặc trưng của cơ sở dữ liệu cần khai phá, một metric thích hợp được sử dụng làm tiêu chuẩn lựa chọn thuộc tính tốt nhất phân chia các đối tượng tại mỗi nút trong quá trình xây dựng cây quyết định, cũng như để phân cụm hay rời rạc hóa dữ liệu.

Bài báo đề xuất phương pháp xây dựng một metric trên tập các phân hoạch của một tập hợp hữu hạn các đối tượng, xuất phát từ độ đo entropy thông tin của các phân hoạch do Jiye Liang và cộng sự [1] đề xuất.

Phần còn lại của bài báo gồm: Mục 2 trình bày không gian các phân hoạch của một tập

hữu hạn cùng với các phép toán đại số; Mục 3 nêu các khái niệm và tính chất của entropy thông tin; Mục 4 đề xuất phương pháp xây dựng metric; Mục 5 đưa ra một ví dụ tính toán; Cuối cùng là kết luận và hướng nghiên cứu tiếp theo.

2. PHÂN HOẠCH CỦA MỘT TẬP HỢP VÀ MỘT SỐ PHÉP TOÁN

Sau đây, mọi tập hợp được xét đều là những tập hợp hữu hạn.

Cho tập hợp không rỗng các phần tử U, một phân hoạch của U là một họ các tập con $\pi = \{X_1, ..., X_m\}$ thỏa mãn $X_i \cap X_j = \phi,\ \forall i \neq j$. Ta gọi các tập con $X_1, ..., X_m$ là các khối của π.

Ký hiệu tập tất cả các phân hoạch có thể của U là $\mathrm{PART}(U)$. Trên $\mathrm{PART}(U)$ có thể định nghĩa một quan hệ thứ tự bộ phận $(\mathrm{PART}(U), \leq)$, như sau.

Định nghĩa 2.1. Cho hai phân hoạch $\pi,\ \sigma \in \mathrm{PART}(U)$. Ta nói π mịn hơn σ (hay σ thô hơn π), ký hiệu $\pi \leq \sigma$, nếu mỗi khối X của π đều tồn tại một khối Y của σ sao cho $X \subseteq Y$. Nói cách khác, $\pi \leq \sigma$ nếu mỗi khối của σ đều là hợp của một số khối nào đó trong π. Trường hợp $\pi \leq \sigma$, nhưng $\pi \neq \sigma$, ta nói π mịn thực sự hơn σ và viết $\pi < \sigma$.

Định nghĩa 2.2. Phần tử nhỏ nhất trong $(\mathrm{PART}(U), \leq)$ là phân hoạch α_U với mỗi khối chỉ bao gồm một phần tử của U. Phần tử lớn nhất trong $(\mathrm{PART}(U), \leq)$ là phân hoạch ω_U bao gồm một khối duy nhất là toàn bộ tập U.

Định nghĩa 2.3. [6, 8] Cho hai phân hoạch $\pi,\ \sigma \in \mathrm{PART}(U)$. Ta nói σ phủ π, ký hiệu $\pi \prec \sigma$, nếu $\pi \leq \sigma$ và không tồn tại $\tau \in \mathrm{PART}(U)$ sao cho $\pi < \tau < \sigma$.

Dễ thấy, $\pi \prec \sigma$ khi và chỉ khi σ thu được từ π bằng cách hợp nhất hai khối của π.

Định nghĩa 2.4. [6, 8] Cho hai phân hoạch $\pi,\ \sigma \in \mathrm{PART}(U)$. Ta gọi *infimum* của π và σ là phân hoạch

$$\inf\{\pi,\ \sigma\} = \{X \cap Y | X \in \pi,\ Y \in \sigma, X \cap Y = \phi\}.$$

Để đơn giản, $\inf\{\pi,\ \sigma\}$ được ký hiệu là $\pi \wedge \sigma$.

Định nghĩa 2.5. [6, 8] Cho hai phân hoạch $\pi,\ \sigma \in \mathrm{PART}(U)$. $G(\pi, \sigma)$ là đồ thị hai phía, trong đó các đỉnh của một phía là các khối của π, các đỉnh ở phía bên kia là các khối của σ, và nếu $X \cap Y = \phi$ với $X \in \pi,\ Y \in \sigma$ thì sẽ có một cạnh (X, Y). Khi đó, ta gọi *supremum* của π và σ là phân hoạch trong đó mỗi khối là hợp của các đỉnh thuộc một thành phần liên thông của đồ thị hai phía $G(\pi, \sigma)$. $sup\{\pi,\ \sigma\}$ được ký hiệu bằng $\pi \vee \sigma$.

Với khái niệm *infimum* và *supremum* giữa hai phân hoạch định nghĩa như trên thì tập $\mathrm{PART}(U)$ tạo thành một giàn, nghĩa là với mọi cặp $\pi,\ \sigma \in \mathrm{PART}(U)$ đều tồn tại $\inf\{\pi, \sigma\}$ lẫn $\sup\{\pi, \sigma\}$.

3. ENTROPY THÔNG TIN CỦA PHÂN HOẠCH VÀ CÁC TÍNH CHẤT

Trong [1], Jiye Liang và cộng sự đã đưa ra các định nghĩa sau đây về entropy thông tin của các phân hoạch.

Cho hai phân hoạch $\pi,\ \sigma \in \mathrm{PART}(U)$. Giả sử $\pi = \{X_1, X_2, ..., X_m\}$ và $\sigma = \{Y_1, Y_2, ..., Y_n\}$.

Định nghĩa 3.1. [1] Entropy thông tin của π là đại lượng $E(\pi)$, xác định như sau

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

trong đó $|X|$ chỉ lực lượng của tập hợp $X, \overline{X}$ là phần bù của X trong U.

Dễ thấy, $E(U)$ có thể viết dưới dạng

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

Định nghĩa 3.2. [1] Entropy thông tin có điều kiện của σ khi đã biết π được định nghĩa bởi

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

Để ý $|X_i \cap \overline{Y}_j| = |X_i| - |X_i \cap Y_j|$, có thể viết $E(\sigma|\pi)$ dưới dạng

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

Định nghĩa 3.3. [1] Entropy thông tin đồng thời của π và σ được định nghĩ bởi

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

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

Định nghĩa 3.4. [1] Entropy thông tin tương hỗ giữa π và σ được định nghĩ bởi

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

Dễ thấy, $I(\pi;\sigma)$ có thể viết dưới dạng

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

Giống như Shannon entropy, entropy thông tin $E(\pi)$ của phân hoạch π có các tính chất sau.

Mệnh đề 3.1. (Giới nội) [1] *Cho phân hoạch* $\pi \in \mathrm{PART}(U)$. *Nếu* $\pi = \{X_1, X_2, ..., X_m\}$ *thì*

$$0 \leq E(\pi) \leq 1 - \frac{1}{m},$$

$E(\pi) = 0$ *khi và chỉ khi* $m = 1$, $E(\pi) = 1 - \frac{1}{m}$ *khi* $|X_1| = |X_2| = ... = |X_m| = \frac{|U|}{m}$.

Mệnh đề 3.2. (Đơn điệu) [1] *Cho hai phân hoạch* $\pi, \sigma \in \mathrm{PART}(U)$.

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

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

Chú ý: Nếu chỉ có $E(\pi) = E(\sigma)$ *thì chưa thể suy ra* $\pi < \sigma$.

Từ Mệnh đề 3.1 và Mệnh đề 3.2 suy ra mệnh đề sau.

Mệnh đề 3.3. (Maximum-Minimum) [1] *Cho tập hợp hữu hạn* U. *Hàm entropy thông tin* $E : \mathrm{PART}(U) \to R$ *có giá trị cực đại là* $1 - (1/|U|)$ *khi* $\pi = \alpha_U$, *có giá trị cực tiểu là 0 khi* $\pi = \omega_U$.

Mệnh đề 3.4. *Cho hai phân hoạch* $\pi, \sigma \in \mathrm{PART}(U)$. *Giả sử* $\pi = \{X_1, X_2, ..., X_m\}$ *và* $\sigma = \{Y_1, Y_2, ..., Y_n\}$. *Ta có*

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

Chứng minh

$$\begin{aligned} E(\pi, \sigma) = E(\pi \wedge \sigma) &= 1 - \frac{1}{|U|^2} \sum_{i=1}^{m} \sum_{j=1}^{n} |X_i \cap Y_j|^2 \\ &= 1 - \frac{1}{|U|^2} \sum_{i=1}^{m} |X_i|^2 + \frac{1}{|U|^2} \sum_{i=1}^{m} |X_i|^2 - \frac{1}{|U|^2} \sum_{i=1}^{m} \sum_{j=1}^{n} |X_i \cap Y_j|^2 \\ &= E(\pi) + E(\sigma|\pi). \end{aligned}$$

Do tính đối xứng của $E(\pi, \sigma)$, ta cũng có $E(\pi, \sigma) = E(\sigma) + E(\pi|\sigma)$. ■

Mệnh đề 3.5. *Cho hai phân hoạch* $\pi, \sigma \in \mathrm{PART}(U)$. *Giả sử* $\pi = \{X_1, X_2, ..., X_m\}$ *và* $\sigma = \{Y_1, Y_2, ..., Y_n\}$. *Ta có*

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

và như thế $I(\pi, \sigma)$ *cũng là hàm đối xứng của* π *và* σ.

Chứng minh

Suy ra từ Định nghĩa 3.4. và Mệnh đề 3.4.

Mệnh đề 3.6. *Cho hai phân hoạch* $\pi, \sigma \in \mathrm{PART}(U)$. *Giả sử* $\pi = \{X_1, X_2, ..., X_m\}$ *và* $\sigma = \{Y_1, Y_2, ..., Y_n\}$. *Ta có*

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

$E(\sigma|\pi) = 0$ khi và chỉ khi $\pi \leq \sigma$, $E(\sigma|\pi) = 1 - \dfrac{1}{mn}$ khi và chỉ khi $m = 1$ và

$$|Y_1| = |Y_2| = ... = |Y_n| = \frac{|U|}{n}.$$

Chứng minh

Hiển nhiên $E(\sigma|\pi) \geq 0$. Theo Mệnh đề 3.4,

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

Suy ra

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

Vì $\pi \wedge \sigma \subseteq \pi$ nên theo Mệnh đề 3.2

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

Mặt khác, theo Mệnh đề 3.4 và Mệnh đề 3.1, ta có

$$E(\sigma|\pi) = E(\pi, \sigma) - E(\pi),\ E(\pi, \sigma) = E(\pi \wedge \sigma) \leq 1 - \frac{1}{mn},\ E(\pi) \geq 0.$$

Suy ra

$$E(\sigma|\pi) \leq 1 - \frac{1}{mn},$$

dấu "=" xảy ra khi và chỉ khi

$$(E(\pi) = 0) \wedge \left(E(\pi \wedge \sigma) = 1 - \frac{1}{mn}\right) \Leftrightarrow (m = 1) \wedge \left(|Y_1| = |Y_2| = ... = |Y_n| = \frac{|U|}{n}\right).$$

■

Mệnh đề 3.7. *Cho ba phân hoạch $\pi, \sigma, \theta \in \mathrm{PART}(U)$. Nếu $\pi < \sigma$ thì $E(\theta|\sigma) \geq E(\theta|\pi)$.*

Chứng minh

Do $\pi < \sigma$, mỗi $Y_j \in \sigma$ sẽ là hợp của một số khối thuộc π. Để chứng minh mệnh đề, chỉ cần chỉ ra rằng, nếu $\pi, \pi' \in \mathrm{PART}(U)$ và π' phủ π, nghĩa là các khối của π' trùng với các khối của π, ngoại trừ chỉ một khối của π' là hợp của hai khối trong π

$$\pi = \{X_1, ..., X_{g-1}, X_g, X_{g+1}, ..., X_{h-1}, X_h, X_{h+1}, ..., X_m\},$$

$$\pi' = \{X_1, ..., X_{g-1}, X_{g+1}, ..., X_{h-1}, X_{h+1}, ..., X_m,\ X_g \cup X_h\},$$

thì $E(\theta|\pi') \geq E(\theta|\pi)$. Thật vậy, giả sử $\theta = \{Z_1, Z_2, ..., Z_l\}$, ta có

$$\begin{aligned}
E(\theta|\pi') - E(\theta|\pi) &= \frac{1}{|U|^2}\sum_{k=1}^{l}|(X_g \cup X_h)\cap Z_k||(X_g \cup X_h)\cap \overline{Z}_k| \\
&\quad - \frac{1}{|U|^2}\sum_{k=1}^{l}|X_g \cap Z_k||X_g \cup \overline{Z}_k| - \frac{1}{|U|^2}\sum_{k=1}^{l}|X_h \cap Z_k||X_h \cup \overline{Z}_k| \\
&= \frac{1}{|U|^2}\sum_{k=1}^{l}(|X_g \cap Z_k| + |X_h \cap Z_k|)(|X_g \cap \overline{Z}_k| + |X_h \cap \overline{Z}_k|) \\
&\quad - \frac{1}{|U|^2}\sum_{k=1}^{l}|X_g \cap Z_k||X_g \cup \overline{Z}_k| - \frac{1}{|U|^2}\sum_{r=1}^{l}|X_h \cap Z_k||X_h \cup \overline{Z}_k| \\
&= \frac{1}{|U|^2}\sum_{k=1}^{l}|X_g \cap Z_k||X_g \cup \overline{Z}_k| + \frac{1}{|U|^2}\sum_{k=1}^{l}|X_g \cap Z_k||X_h \cup \overline{Z}_k| \geq 0
\end{aligned}$$

$E(\theta|\pi') = E(\theta|\pi)$ khi $|X_g \cap Z_k||X_g \cup \overline{Z}_k| = 0$ và $|X_g \cap Z_k||X_h \cup \overline{Z}_k| = 0$ với mọi $k \in \{1, 2, ..., l\}$, tức là khi tồn tại một khối $Z_p \in \theta$ sao cho X_h và X_g đều thuộc Z_p. ■

Mệnh đề 3.8. *Cho ba phân hoạch $\pi, \sigma, \theta \in \mathrm{PART}(U)$. Nếu $\pi < \sigma$ thì $E(\pi|\theta) \geq E(\sigma|\theta)$.*

Chứng minh

Vì $\pi < \sigma$, nên $\pi \wedge \theta \leq \sigma \wedge \theta$. Do đó, theo Mệnh đề 3.2 có

$$E(\pi \wedge \theta) \geq E(\sigma \wedge \theta).$$

Suy ra,

$$E(\pi|\theta) = E(\pi \wedge \theta) - E(\theta) \geq E(\sigma \wedge \theta) - E(\theta) = E(\sigma|\theta).$$

■

4. XÂY DỰNG METRIC TRÊN TẬP CÁC PHÂN HOẠCH

Một metric trên tập hợp U là một ánh xạ $d : U \times U \to [0, \infty)$ thỏa mãn các điều kiện sau

(P1) $d(x, y) = 0$ khi và chỉ khi $x = y$.

(P2) $d(x, y) = d(y, x)$.

(P3) $d(x, y) + d(y, z) \geq d(x, z)$ với mọi $x, y, z \in U$.

Điều kiện (P3) được gọi là *tiên đề bất đẳng thức tam giác*. Bộ đôi (U, d) được gọi là một không gian metric.

Cho hai phân hoạch $\pi, \sigma \in \mathrm{PART}(U)$, với $\pi = \{X_1, X_2, ..., X_m\}$ và $\sigma = \{Y_1, Y_2, ..., Y_n\}$. Shannon entropy có điều kiện của π khi biết σ là đại lượng $I(\pi|\sigma)$ định nghĩa bởi

$$I(\pi|\sigma) = -\sum_{j=1}^{n}\frac{|Y_j|}{|U|}\sum_{i=1}^{m}\frac{|X_i \cap Y_j|}{|Y_j|}\log_2\frac{|X_i \cap Y_j|}{|Y_j|}.$$

Shannon entropy đồng thời của π và σ là đại lượng $I(\pi,\sigma)$ xác định bởi

$$I(\pi,\sigma) = -\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|X_i\cap Y_j|}{|U|}\log_2\frac{|X_i\cap Y_j|}{|U|}.$$

Trong [5], López de Mántaras đã chứng minh rằng ánh xạ

$$d_E : \mathrm{PART}(U)\times\mathrm{PART}(U)\to[0,\infty)$$

xác định bởi

$$d_N(\pi,\sigma) = \frac{I(\pi|\sigma)+I(\sigma|\pi)}{I(\pi,\sigma)}$$

với mọi $\pi,\sigma\in\mathrm{PART}(U)$ là một metric trên tập $\mathrm{PART}(U)$.

Nghiên cứu các tính chất của Shannon entropy, D. A. Simovici và S. Jaroszewicz [6, 8] đã đề xuất khái niệm entropy có điều kiện mở rộng của π khi biết σ. Đó là đại lượng $H_\beta(\pi|\sigma)$ phụ thuộc tham số $\beta>0$, xác định bởi

$$H_\beta(\pi|\sigma) = \begin{cases} \dfrac{1}{(2^{1-\beta}-1)|U|^\beta}\displaystyle\sum_{i=1}^{m}\sum_{j=1}^{n}\Big(\sum_{j=1}^{n}|Y_j|^\beta - |X_i\cap Y_j|^\beta\Big) & \text{khi } 0<\beta<1\\ -\displaystyle\sum_{j=1}^{n}\frac{|Y_j|}{|U|}\sum_{i=1}^{m}\frac{|X_i\cap Y_j|}{|Y_j|}\log_2\frac{|X_i\cap Y_j|}{|Y_j|} & \text{khi } \beta=1\\ \dfrac{1}{(1-2^{1-\beta})|U|^\beta}\displaystyle\sum_{i=1}^{m}\sum_{j=1}^{n}\Big(\sum_{j=1}^{n}|Y_j|^\beta - |X_i\cap Y_j|^\beta\Big) & \text{khi } \beta>1 \end{cases}$$

Các tác giả cũng đã chứng minh được rằng ánh xạ

$$d_E : \mathrm{PART}(U)\times\mathrm{PART}(U)\to[0,\infty)$$

xác định bởi

$$d_\beta(\pi,\sigma) = H_\beta(\pi|\sigma)+H_\beta(\sigma|\pi)$$

với mọi $\pi,\sigma\in\mathrm{PART}(U)$ là một metric trên tập $\mathrm{PART}(U)$.

Dưới đây là một đề xuất phương pháp mới xây dựng metric trên tập các phân hoạch của một tập hữu hạn, sử dụng độ đo entropy thông tin có điều kiện do Jiye Liang và cộng sự đề xuất.

Mệnh đề 4.1. *Với mọi $\pi,\sigma,\theta\in\mathrm{PART}(U)$, ta đều có*

$$E(\pi|\theta)+E(\sigma|\pi\wedge\theta) = E(\pi\wedge\sigma|\theta).$$

Chứng minh

Thật vậy, giả sử $\pi = \{X_1, X_2, ..., X_m\}$, $\sigma = \{Y_1, Y_2, ..., Y_n\}$ và $\theta = \{Z_1, Z_2, ..., Z_l\}$, ta có

$$E(\pi|\theta) = \frac{1}{|U|^2}\sum_{p=1}^{l}|Z_p|^2 - \frac{1}{|U|^2}\sum_{p=1}^{l}\sum_{i=1}^{n}|X_i \cap Z_p|^2;$$

$$\begin{aligned}
E(\sigma|\pi \wedge \theta) &= \frac{1}{|U|^2}\sum_{i=1}^{m}\sum_{p=1}^{l}|X_i \cap Z_p|^2 - \frac{1}{|U|^2}\sum_{i=1}^{m}\sum_{p=1}^{l}\sum_{j=1}^{n}|(X_i \cap Z_p) \cap Y_j|^2 \\
&= \frac{1}{|U|^2}\sum_{i=1}^{m}\sum_{p=1}^{l}|X_i \cap Z_p|^2 - \frac{1}{|U|^2}\sum_{i=1}^{m}\sum_{p=1}^{l}\sum_{j=1}^{n}|(X_i \cap Z_p) \cap (\bigcup_{q=1}^{l}(Y_j \cap Z_q))|^2 \\
&= \frac{1}{|U|^2}\sum_{i=1}^{m}\sum_{p=1}^{l}|X_i \cap Z_p|^2 - \frac{1}{|U|^2}\sum_{i=1}^{m}\sum_{p=1}^{l}\sum_{j=1}^{n}|\bigcup_{q=1}^{l}((X_i \cap Z_p) \cap (Y_j \cap Z_q))|^2 \\
&= \frac{1}{|U|^2}\sum_{i=1}^{m}\sum_{p=1}^{l}|X_i \cap Z_p|^2 - \frac{1}{|U|^2}\sum_{i=1}^{m}\sum_{p=1}^{l}\sum_{j=1}^{n}|(X_i \cap Z_p) \cap (Y_j \cap Z_p)|^2
\end{aligned}$$

(vì với $p \neq q$ thì $(X_i \cap Z_p) \cap (Y_j \cap Z_q) = \phi$).

Tính toán tương tự, thu được

$$E(\pi \wedge \sigma|\theta) = \frac{1}{|U|^2}\sum_{p=1}^{l}|Z_p|^2 - \frac{1}{|U|^2}\sum_{p=1}^{l}\sum_{i=1}^{m}\sum_{j=1}^{n}|(X_i \cap Z_p) \cap (Y_j \cap Z_p)|^2.$$

Vậy,

$$\begin{aligned}
E(\sigma|\pi \wedge \theta) + E(\pi|\theta) &= \frac{1}{|U|^2}\sum_{i=1}^{m}\sum_{p=1}^{l}|X_i \cap Z_p|^2 - \frac{1}{|U|^2}\sum_{i=1}^{m}\sum_{p=1}^{l}\sum_{j=1}^{n}|(X_i \cap Z_p) \cap (Y_j \cap Z_p)|^2 \\
&+ \frac{1}{|U|^2}\sum_{p=1}^{l}|Z_p|^2 - \frac{1}{|U|^2}\sum_{p=1}^{l}\sum_{i=1}^{m}|X_i \cap Z_p|^2 \\
&= \frac{1}{|U|^2}\sum_{p=1}^{l}|Z_p|^2 - \frac{1}{|U|^2}\sum_{i=1}^{m}\sum_{p=1}^{l}\sum_{j=1}^{n}|(X_i \cap Z_p) \cap (Y_j \cap Z_p)|^2 \\
&= E(\pi \wedge \sigma|\theta).
\end{aligned}$$

■

Mệnh đề 4.2. *Với mọi bộ ba phân hoạch $\pi, \sigma, \theta \in \mathrm{PART}(U)$, ta đều có*

$$E(\sigma|\pi) + E(\pi|\theta) \geq E(\sigma|\theta).$$

Chứng minh

Để ý, $\pi \wedge \theta \leq \pi,\ \pi \wedge \sigma \leq \sigma$, áp dụng lần lượt các Mệnh đề 3.7, 4.1 và 3.8, ta có

$$E(\sigma|\pi) + E(\pi|\theta) \geq E(\sigma|\pi \wedge \theta) + E(\pi|\theta) = E(\pi \wedge \sigma|\theta) \geq E(\sigma|\theta).$$

■

Định lý 4.1. *Ánh xạ* $d_E : \mathrm{PART}(U) \times \mathrm{PART}(U) \to [0, \infty)$ *xác định bởi*

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

với mọi $\pi, \sigma \in \mathrm{PART}(U)$ *là một metric trên tập* $\mathrm{PART}(U)$.

Chứng minh

Theo Mệnh đề 3.6

(P1) $d_E(\pi, \sigma) \geq 0$ với mọi $\pi, \sigma \in \mathrm{PART}(U)$

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

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

Từ định nghĩa của d_E suy ra

(P2) $d_E(\pi, \sigma) = d_E(\sigma, \pi)$ với mọi $\pi, \sigma \in \mathrm{PART}(U)$.

Lại theo Mệnh đề 4.2, với mọi $\pi, \sigma, \theta \in \mathrm{PART}(U)$, ta có

$$E(\sigma|\pi) + E(\pi|\theta) \geq E(\sigma|\theta),$$

$$E(\theta|\pi) + E(\pi|\sigma) \geq E(\theta|\sigma).$$

Cộng hai bất đẳng thức trên, vế với vế, thu được

(P3) $d_E(\sigma, \pi) + d_E(\pi, \theta) \geq d_E(\sigma, \theta)$. ■

5. VÍ DỤ

Xét bảng dữ liệu sau đây về các điều kiện thời tiết và quyết định chơi golf, lấy từ kho dữ liệu UCI [11] với thuộc tính nhiệt độ đã được rời rạc hóa thành ba mức 1,2,3 và thuộc tính độ ẩm được rời rạc hóa thành hai mức 1,2.

U	Quang cảnh $a1$	Nhiệt độ $a2$	Độ ẩm $a3$	Gió $a4$	Chơi golf d
u_1	nắng	3	2	không	không
u_2	nắng	3	2	có	không
u_3	u ám	3	2	không	có
u_4	mưa	2	2	không	có
u_5	mưa	1	2	không	có
u_6	mưa	1	1	có	không
u_7	u ám	1	1	có	có
u_8	nắng	2	2	không	không
u_9	nắng	1	1	không	có
u_{10}	mưa	2	2	không	có
u_{11}	nắng	2	1	có	có
u_{12}	u ám	2	2	có	có
u_{13}	u ám	3	1	không	có
u_{14}	mưa	2	2	có	không

Ta có

$$U = \{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}, u_{11}, u_{12}, u_{13}, u_{14}\},\ |U| = 14.$$

Phân hoạch của U theo $a1, a2, a3, a4$ và d lần lượt là
$\pi_{a1} = \{X_1, X_2, X_3\} = \{\{u_1, u_2, u_8, u_9, u_{11}\},\ \{u_3, u_7, u_{12}, u_{13}\},\ \{u_4, u_5, u_6, u_{10}, u_{14}\}\}$,
$\pi_{a2} = \{Y_1, Y_2, Y_3\} = \{\{u_1, u_2, u_3, u_{13}\},\ \{u_4, u_8, u_{10}, u_{11}, u_{12}, u_{14}\}, \{u_5, u_6, u_7, u_9\}\}$,
$\pi_{a3} = \{Z_1, Z_2, Z_3\} = \{\{u_1, u_2, u_3, u_4, u_5, u_8, u_{10}, u_{12}, u_{14}\},\ \{u_6, u_7, u_9, u_{11}, u_{13}\}\}$,
$\pi_{a4} = \{W_1, W_2\} = \{\{u_1, u_3, u_4, u_5, u_8, u_9, u_{10}, u_{13}\},\ \{u_2, u_6, u_7, u_{11}, u_{12}, u_{14}\}\}$,
$\pi_d = \{D_1, D_2\} = \{\{u_1, u_2, u_6, u_8, u_{14}\},\ \{u_3, u_4, u_5, u_7, u_9, u_{10}, u_{11}, u_{12}, u_{13}\}\}$,
$X_1 \cap D_1 = \{u_1, u_2, u_8\},\ |X_1 \cap D_1| = 3$,
$X_1 \cap D_2 = \{u_9, u_{11}\},\ |X_1 \cap D_2| = 2$,
$X_2 \cap D_1 = \phi,\ |X_2 \cap D_1| = 0$,
$X_2 \cap D_2 = \{u_3, u_7, u_{12}, u_{13}\},\ |X_2 \cap D_2| = 4$,
$X_3 \cap D_1 = \{u_6, u_{14}\},\ |X_3 \cap D_1| = 2$,
$X_3 \cap D_2 = \{u_4, u_5, u_{10}\},\ |X_3 \cap D_2| = 3$.
Các entropy thông tin có điều kiện theo Jiye Liang

$$\begin{aligned} E(\pi_d|\pi_{a1}) &= \frac{1}{|U|^2}\Big(\sum_{i=1}^{3}|X_i|^2 - \sum_{i=1}^{3}\sum_{j=1}^{2}|X_i \cap D_j|^2\Big) \\ &= \frac{1}{14^2}\{(5^2+4^2+5^2) - (3^2+2^2+4^2+2^2+3^2)\} \\ &= \frac{24}{196} \end{aligned}$$

$$E(\pi_{a1}|\pi_d) = \frac{1}{|U|^2}\Big(\sum_{j=1}^{2}|D_j|^2 - \sum_{i=1}^{3}\sum_{j=1}^{2}|X_i \cap D_j|^2\Big)$$
$$= \frac{1}{14^2}\{(5^2+9^2) - (3^2+2^2+4^2+2^2+3^2)\}$$
$$= \frac{64}{196}$$

Vậy, $d_E(\pi_{a1}, \pi_d) = E(\pi_{a1}|\pi_d) + E(\pi_d|\pi_{a1}) = \frac{24}{196} + \frac{64}{196} = \frac{88}{196}$.

Tính toán tương tự, thu được

$$E(\pi_{a2}|\pi_d) = \frac{68}{196},\ E(\pi_d|\pi_{a2}) = \frac{30}{196},$$

$$d_E(\pi_{a2}, \pi_d) = E(\pi_{a2}|\pi_d) + E(\pi_d|\pi_{a2}) = \frac{68}{196} + \frac{30}{196} = \frac{98}{196}.$$

$$E(\pi_{a3}|\pi_d) = \frac{48}{196},\ E(\pi_d|\pi_{a3}) = \frac{48}{196},$$

$$d_E(\pi_{a3}, \pi_d) = E(\pi_{a3}|\pi_d) + E(\pi_d|\pi_{a3}) = \frac{48}{196} + \frac{48}{196} = \frac{96}{196}.$$

$$E(\pi_{a4}|\pi_d) = \frac{48}{196},\ E(\pi_d|\pi_{a4}) = \frac{42}{196},$$

$$d_E(\pi_{a4}, \pi_d) = E(\pi_{a4}|\pi_d) + E(\pi_d|\pi_{a4}) = \frac{48}{196} + \frac{42}{196} = \frac{90}{196}.$$

Ta có $d_E(\pi_{a1}, \sigma_d) < d_E(\pi_{a4}, \sigma_d) < d_E(\pi_{a3}, \sigma_d) < d_E(\pi_{a2}, \sigma_d)$. Điều này phù hợp với thực tế có thể quan sát thấy về sự kết hợp dữ liệu trên các đối tượng của mỗi thuộc tính $a1, a2, a3, a4$ với thuộc tính quyết định d. Nếu cần lựa chọn thuộc tính cho việc dự đoán quyết định d thì thứ tự ưu tiên sẽ là $a1, a4, a3, a2$.

6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo đã đề xuất phương pháp xây dựng một metric trên họ các phân hoạch của một tập hợp hữu hạn các đối tượng, và như vậy cũng là một metric trên họ các thuộc tính cùng xác định trên tập hợp đó. Dựa trên các công thức tính cho trong Mục 3, có thể thấy, metric chúng tôi xây dựng đòi hỏi khối lượng tính toán nhỏ hơn so với metric đề xuất bởi López de Mántaras và metric đề xuất bởi D. A. Simovici và S. Jaroszewicz.

Một khi đã có metric, ta có thể đánh giá độ gần nhau, phân cụm các thuộc tính, xác định thuộc tính quan trọng, lựa chọn thuộc tính,... Do đó, công việc tiếp theo cần phải thực hiện là sử dụng metric đã xây dựng vào việc giải quyết các bài toán như lựa chọn thuộc tính xây dựng cây quyết định, rời rác hóa dữ liệu,...; tính toán thực nghiệm trên các tập dữ liệu khác nhau và so sánh kết quả với kết quả thu được bằng các metric khác hiện có.

TÀI LIỆU THAM KHẢO

[1] Jiye 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* **31** (4) (2002) 331–342.

[2] J. P. Barthélemy, Notes on the metric properties of ordered sets, *Mathematics and Human Sciences* Vol. **61** (1976) 39–60 (tiếng Pháp).

[3] J. P. Barthélemy & B. Leclerc, The median procedure for partitions, *Partitioning data sets*, American Mathematical Society, Providence, 1995 (pp.3–34).

[4] B. Monjardet, A metrics on parially ordered sets - A survey, *Discrete Mathematics* Vol. **35** (1981) 173–184.

[5] R. López de Mántaras, A distance-based attribute selection measure for decision tree induction, *Machine Learning* Vol. **6** (1991) 81–92.

[6] D. A. Simovici & S. Jaroszewicz, A new metric splitting criterion for decision trees, *International Journal of Parallel Emergent and Distributed Systems* Vol. **21** (4) (2006) 239–256.

[7] D. A. Simovici, Singla et. al., "Metric incremental clustering of nominal data", Univer-sity of Massachusetts - Boston, USA, 2005.

[8] D. A. Simovici & S. Jaroszewicz, Generalized conditional entropy and decision trees, *Proceeding of EGC*, Lyon, France, 2003 (369–380).

[9] D. A. Simovici, Metric methods in data mining, 2008: https://www.infosci-online.com/downloads/.../IGR3997_EIbV9PDbbL.pdf.

[10] Richard A. Butterworth, "Contributions to metric methods in data mining", PhD Disser-tation, University of Massachusetts Boston, USA, 2006.

[11] Murphy P., Aha W. UCI repository of machine learning databases: http://www.ics.ici.edu/∼mlearn

Nhận bài ngày 13 - 4 - 2010