TẬP THÔ VÀ ĐÁNH GIÁ HỆ THÔNG TIN NỀN

Hà Quang Thụy
Đại học Khoa học Tự nhiên - ĐHQG Hà Nội

ho hệ thông tin [6] $S =< \Omega, A, V, r >$, trong đó Ω là tập đối tượng, A tập tên thuộc V là tập các giá trị còn r là ánh xạ từ $\Omega \times A$ tới V.

ới mỗi $a \in A$, xây dựng quan hệ tương đương $a*$ trên Ω : $xa^*y \iff r(x,a) = r(y,a)$.

ương đương theo quan hệ $S^* = \bigcap_{a \in A} a^*$ được gọi là *tập cơ bản* trong S. Trong bài

ày, thường dùng chữ cái e để chỉ tập cơ bản.

ập hợp tất cả các tập cơ bản trong S được kí hiệu là E_S (được viết tắt là E).

nhiều trường hợp, dùng kí hiệu (Ω, E) để chỉ hệ thông tin S.

ập con của Ω là hợp của một số tập cơ bản trong S được gọi là *tập mô tả được.*

ợp tất cả các tập mô tả được của S được kí hiệu là D_S. Trong [6] đưa ra ngôn $_S$ để hình thức hóa câu hỏi liên quan đến hệ thông tin. Câu trả lời của mỗi câu ong L_S hoặc là tập rỗng hoặc là một tập mô tả được trong D_S. Trong [2], [6] đưa ật toán tìm câu trả lời trong hệ thông tin.

uy vậy, trong nhiều trường hợp không chỉ giới hạn xem xét các tập mô tả được n cần quan tâm đến các tập hợp con khác trong Ω : nảy sinh vấn đề xấp xỉ một ất kỳ qua tập mô tả được. Có một số cách xây dựng xấp xỉ trong hệ thông tin. iết này được định hướng theo khái niệm "tập thô". Các khái niệm được trình bày đây được dùng để hình thức hóa quá trình xấp xỉ các tập không mô tả được thông ác tập mô tả được.

I. TẬP GIỚI HẠN

ho X là tập con của Ω. Các *tập giới hạn* liên quan tới X được định nghĩa như ây:

nghĩa: *Tập giới hạn trên của X là tập mô tả được bé nhất chứa X. Tập giới hạn (của X) là tập mô tả được lớn nhất bị chứa trong X.*

í hiệu tập giới hạn trên của X là X^*, tập giới hạn dưới là X_*. Hiển nhiên, khi X tả được thì X cũng chính là các tập giới hạn của nó. Ta có,

đề : $X^* = \bigcup_{e \in E, e \cap X \neq \emptyset} e; \quad X_* = \bigcup_{e \in E, e \subset X} e;$

Chứng minh: Chúng ta chứng minh trường hợp X^* (trường hợp X_* là tương tự).

$= \bigcup_{e \in E, e \cap X \neq \emptyset} e$. Hiển nhiên $X^* \subset B^*$. Cần kiểm chứng chỉ xẩy ra dấu đẳng thức.

ự ngược lại có nghĩa là $\exists x \in B \setminus X^*$. Theo cách xây dựng tập B, ta phải có tập cơ * nào đó để cho $x \in e^*$ mà $e^* \cap X \neq \emptyset$. Suy ra $e^* \cap X^* \neq \emptyset$.

o X^* mô tả được nên ta có $e^* \subset X^*$, hay cũng vậy $x \in X^*$ mâu thuẫn với giả thiết chứng. Vậy ta có $B = X^*$ (điều phải chứng minh). $\Diamond$

ụ : Xét hệ thông tin chứa bệnh án của một số bệnh nhân trẻ em cho trong bảng

dưới đây:

Đối tượng	Dịch não tủy lần 1	BK dịch dạ dày	Thân nhiệt	BK dịch não tủy
A	vàng	âm	rất cao	homo ân
B	trong	dương	nhẹ	homo du
C	đục	âm	cao	homo ân
D	vàng	âm	nhẹ	homo ân
E	vàng	âm	cao	homo ân
F	đục	âm	cao	homo ân
G	vàng	dương	hơi cao	homo ân
H	vàng	âm	rất cao	homo ân
I	đục	âm	cao	homo ân
J	vàng	âm	nhẹ	homo ân

Trong hệ thông tin này, các tập cơ bản là:

$$e_1 = \{A, H\},\ e_2 = \{B\},\ e_3 = \{C, F, I\},\ e_4 = \{D, J\},\ e_5 = \{E\},\ e_6 = \{G\}.$$

Giả sử, các bệnh nhân A, C, E, H, I, J đã thực sự bị lao màng não còn các bệnh khác chưa thể kết luận. Nếu ký hiệu X để chỉ tập bệnh nhân lao màng não, thì cá giới hạn liên quan đến X là:

$$X_* = e_1 \cup e_5 \qquad X^* = e_1 \cup e_3 \cup e_4 \cup e_5$$

II. TẬP THÔ

Khái niệm *tập thô* được A. Marek và Z. Pawlak đưa ra trong [7,8]:

Cho (Ω, E) là hệ thông tin với tập đối tượng Ω và tập các tập cơ bản E. Cho tập con của Ω. *Tập thô* (rough set) đối với tập X được định nghĩa như dưới đây:

Theo cách biểu diễn tập hợp:

$$R_*(X) = \bigcup_{e \subset X} e \qquad \text{tập thô } \textit{chắc chắn} \text{ (thuộc mạnh)}$$

$$R^*(X) = \bigcup_{e \cap X \neq \emptyset} e \qquad \text{tập thô } \textit{có thể} \text{ (thuộc yếu)}$$

hoặc theo cách biểu diễn chỉ bằng duy nhất hàm thành viên f_X (tương ứng với X Ω vào R^+ thì tập thô được xác định như sau: $\forall x \in \Omega,\ x \in e\ (e \in E)$, ta có

$$f_X(x) = \begin{cases} 0 & \text{nếu } e \cap X = \emptyset \\ 1 & \text{nếu } e \subset X \quad \text{hay } x \in R_*(X) \\ 1/2 & \text{trường hợp còn lại hay } x \in R^*(X) - R_*(X) \end{cases}$$

Theo cách biểu diễn tập thô theo ngôn ngữ hàm thành viên, khi đối sánh với định tập mờ của Zadeh, theo Pawlak [7], định nghĩa tập thô tuy có một số điều tương

cũng có những khác biệt quan trọng, nhất là đối với một số phép toán thao tác các tập.

Nếu quan niệm hàm thành viên đối với tập mờ chỉ cần thỏa mãn các điều kiện ng hoặc s-dạng) thì các khác biệt đó không phải là khác biệt cơ bản.

Mặt khác, tồn tại sự tương ứng giữa *tập thô chắc chắn* với tập là hợp các tập con định nghĩa *xác suất dưới* thông qua các xác suất cơ sở của lý thuyết Dempster-r (hay cũng vậy, *tập thô có thể* $\longleftrightarrow$ *xác suất trên*) [4].

III. TẬP THÔ TRONG KHÔNG GIAN ĐO ĐƯỢC

Biểu diễn tập thô

Trong trường hợp hệ thông tin trên tập đối tượng là không gian đo được, chúng ra ra một công thức "mịn hơn" cho định nghĩa tập thô (theo ngôn ngữ hàm thành :

Cho (Ω, μ) là không gian đo được với độ đo μ, lớp các tập μ-đo được được kí hiệu (σ–đại số cấc tập đo được). Ta giả thiết với mọi tập cơ bản e là ν- đo được và > 0; giả thiết X là ν-đo được. Định nghĩa tập thô (theo thuật ngữ hàm thành viên) thay bởi công thức như sau (giả sử $x \in e$ tập cơ *bản nào đó*):

$$f_X(x) = \frac{\mu(X \cap e)}{\mu(e)} \tag{2}$$

Ω là hữu hạn thì ta có

$$f_X(x) = \frac{n(X \cap e)}{n(e)} \tag{2*}$$

ở đó $n(\cdot)$ là số lượng phần tử có trong tập đối số.

Về hình thức, định nghĩa này cho biểu diễn gọn hơn so định nghĩa ban đầu. Về nội , trong định nghĩa này hàm thành viên xấp xỉ mịn hơn.

Như vậy với mọi tập con X là ν-đo được trên Ω, có một hàm xấp xỉ tương ứng thành viên) cũng được gọi là *tập thô.*

Chúng ta cần xác định (làm rõ) tập thô, song thực chất chưa biết đầy đủ thông tin ích xác định nó. Như vậy, việc đưa ra các lớp tương đương trong Ω, cho phép thay việc chưa có đầy đủ thông tin về giá trị hàm thành viên đối với mỗi đối tượng cụ bằng một thông tin chung cho lớp nào đó chứa đối tượng đó, dù rằng việc phân lối tượng tương ứng với tập còn theo một cách "thô" nào đó.

Ụ Trong y học, người ta quan tâm đến một nhóm các căn bệnh (một số bệnh có liên với nhau), mỗi căn bệnh tương ứng với một tập người (bệnh nhân) là một tập con ập các đối tượng Ω {tập tất cả mọi người }. Có thể coi mỗi tập người bị một căn nào đó là một tập chưa thể xác định. Bởi vậy, với từng căn bệnh, có thể phân lớp ượng theo những cách mịn hay thô nào đó mà thường quan tâm đến các thông tin a tuổi, nghề nghiệp, tình trạng kinh tế, địa dư sinh sống, thói quen sinh hoạt, các chứng liên quan đến căn bệnh v.v. Đấy chính là các tiêu chuẩn đầu tiên để phân Như vậy, các đối tượng trong một lớp tương đương e được đánh giá là đồng nhất hả năng nhiễm một căn bệnh nào đó (cùng điều kiện để nhiễm bệnh). Các thông ề triệu chứng lâm sàng cho ta một cách phân hoạch để tạo ra các tập cơ bản (xây ệ thông tin).

Điểm qua các tính chất của ánh xạ

$$F : \Sigma \longrightarrow \quad \{\text{các hàm thô trên } \Omega\}.$$

tính chất này nhận được một cách tự nhiên từ định nghĩa của hàm thô và tính chất độ đo.

a. $f_X + f_Y = f_{X \cup Y} + f_{X \cap Y}$

b. Tính đơn điệu dưới theo phép hợp: $\max\{f_X, f_Y\} \leq f_{X \cup Y} \leq f_X + f_Y$

c. Tính đơn điệu trên theo phép giao: $\min\{f_X, f_Y\} \geq f_{X \cap Y}$

Như vậy một số tính chất của hàm thành viên trong tập mờ không còn đúng tập thô, chẳng hạn hai tính chất đặc trưng sau (m biểu thị hàm mờ):

$$m(X \cup Y) = \max\{m(X), m(Y)\}$$
$$m(X \cap Y) = \min\{m(X), m(Y)\}$$

b. Tính giá trị hàm thô

Đáng tiếc, trong cả hai định nghĩa (1) và (2), một số điều chưa được làm rõ. X đang được xem xét nói chung là chưa xác định được (dù có thể quan sát được số phần tử của nó). Điều đó cũng có nghĩa cách xác định giá trị của hàm thô trong hai định nghĩa mang tính chất áp đặt. Lẽ tự nhiên, chúng ta có thể giải trình, t định nghĩa (1) cũng như (2), các đánh giá trong so sánh tập hợp hoặc độ đo là có ước lượng được và thay vì những giá trị đúng, chúng ta cho các giá trị ước lượng. ước lượng này có thể nhận được bằng những thống kế toán học hoặc những kết của một quá trình thu thập ý kiến chuyên gia hoặc cả hai biện pháp. Qua cách tiến hành như vậy, đối với một tập X dù không thể (hoặc chưa thể) xác định hoàn được thì để biểu diễn nó, ta xác định các tập thô tương ứng.

IV. HỆ THÔNG TIN NỀN

Không gian các đối tượng (Ω) của hệ thông tin có thể coi là vũ trụ không thay và tập thô tồn tại khách quan. Như vậy, có thể quan niệm phân hoạch các tập cơ mới chỉ là một cách làm thô trong quá trình xác định các tập chưa xác định. Với tập chưa xác định, có thể có nhiều cách "làm thô" nó. Đặt ra vấn đề, có thể hay kl so sánh các cách làm thô nói trên? Để trả lời câu hỏi này, chúng ta xây dựng định n *cách làm thô tốt nhất* để hướng tới cách xây dựng hệ thông tin có ý nghĩa trong áp dụng. Một trong những vấn đề là để so sánh cần dựa trên các đối chứng. Nhu cùng với tập mờ, chúng ta có một mẫu đối chứng (có thể quan niệm tương tự như niệm mẫu quan trắc trong thống kê). Cho $\{< x_i, V(x_i) >, V(x_i) \in \{0,1\}\}$, trong đó tập mờ (thô) đang được quan tâm.

Cho hệ thông tin (Ω, E) mà Ω có hữu hạn phần tử, có các đại lượng:

$$n(e) = ||x_i \in e||$$
$$\rho(e) = ||x_i \in e, V(x_i) = 1||/n(e)$$

Cho hai hệ thông tin (Ω, E_1) và (Ω, E_2). Nói hệ thông tin (Ω, E_2). là *mịn hơn* so (Ω, E_1). Nếu thỏa mãn điều kiện mọi tập cơ bản trong hệ thông tin (Ω, E_1) đều là tậ tả được trong (Ω, E_2). Nếu (Ω, E_2). là *mịn hơn* so với (Ω, E_1). thì gọi $(\Omega, E_1$ là *thô hơ* với (Ω, E_2). .

Cho hai hệ thông tin (Ω, E_1), (Ω, E_2) với (Ω, E_2). là mịn hơn (Ω, E_1) và một lớp tập chưa tường minh trên Ω.

Gọi hệ thông tin (Ω, E_2). *làm thô tốt hơn* hệ thông tin (Ω, E_1) nếu như:

$$\Sigma n(e)|f_{E_2}(e) - \rho(e)| \leq \Sigma n(e')|f_{E_1}(e') - \rho(e')|$$

trong đó f_{E_2}, f_{E_1} để chỉ cùng một hàm thô trên các hệ thông tin (Ω, E_2), (Ω, E_1) t ứng.

Chú ý: Trong định nghĩa có thể bỏ qua điều kiện về tính mịn của một hệ đối với h lại.

iải thích định nghĩa: Trong định nghĩa, các hệ số $n(e)$, $n(e')$ để đánh giá trọng số
n thô đối với các tập cơ sở. Giá thành xấp xỉ (làm thô) không chỉ phụ thuộc vào
của xấp xỉ mà còn phụ thuộc vào giá trị của mỗi tập cơ bản (không phải các tập
ı cùng có một trọng số khi xem xét). Về hình thức, có thể đặt vấn đề là *càng làm*
ệ thông tin nền thì *càng nhận* được thông tin *tốt hơn.*
ặt khác, bản thân trong định nghĩa tập thô lại bao hàm ý nghĩa là càng làm "thô"
bao nhiêu thì càng tốt bấy nhiêu chỉ cần thỏa mãn điều kiện đảm bảo tính xấp
thiết. Như vậy việc làm quá mịn hệ thông tin không phải là hướng đi duy nhất
vì như thế có thể dẫn đến điều giá trị thu nhận được không tương xứng với công
ỏ ra khi làm mịn. Và như thế, lại có thể đặt ra vấn đề là: làm thô hệ thông tin
nức độ nào và theo những tiêu chuẩn nào để đảm bảo được giá trị của việc làm
ó theo hai khía cạnh: giá thành và độ tin cậy. Bài toán đặt ra là hoàn toàn hợp lý
thực tế thông qua ví dụ sau đây:
'heo các nghiên cứu trong [1], khi theo dõi bệnh lao màng não của trẻ nhỏ, chúng
ận được các thông tin (với khoảng vài chục lớp như trong bảng thống kê [1]) về
nhân được mô hình qua một hệ thông tin S. Các thông tin trong hệ phổ dụng S
tập hợp từ nhiều nguồn, theo các điều kiện phân cấp (các tuyến khám bệnh, xét
m và điều trị). Tuy vậy, ngay cả trong cấp độ cao nhất (có đủ các thuộc tính)
ş tin về bệnh lao màng não cũng vẫn chưa đầy đủ. Vấn đề đặt ra, có mối liên hệ
ɔ giữa các triệu chứng (các thông tin về bệnh nhân) với bệnh. Thông qua mối liên
ɣ, bài viết [1] cho một phương pháp nhận biết thông tin tình trạng bệnh thông qua
riệu chứng. Phương pháp sử dụng trong bài viết đó là tính toán thống kê các giá
ơng quan.
ở đây chúng ta quan tâm đến một cách đề cập khác. Tập trẻ em bị bệnh lao màng
à một *tập thô* trên hệ thông tin S. Với các bác sĩ ở các tuyến dưới (chỉ quan sát
một số thông tin nào đó trong tập con các tiêu chuẩn theo dõi), có nghĩa hệ thông
ược xem xét là thô hơn so với hệ thông tin của các bệnh viên trung ương. Một vấn
n được xem xét: tập thô biểu thị bệnh nói trên (nhận được khi xem xét S) khi hệ
ị rút gọn như vậy còn sử dụng được không? Việc khẳng định khả năng sử dụng
ệ thu gọn sẽ được đánh giá theo các tiêu chuẩn nào? Liên quan đến mô hình hóa
ống đã cho ở [1], chúng ta còn phải xây dựng các hệ thông tin với thông tin chưa
ủ (chưa xác định). Giải quyết cụ thể các vấn đề đó ra sao?

V. HỆ THÔNG TIN THU GỌN VÀ THUỘC TÍNH XÁC ĐỊNH TẬP THÔ

nghĩa: Cho hệ thông tin S với tập các tập cơ bản E. Hệ thông tin S' thô hơn
ợc gọi là *hệ thu gọn của* S đối với lớp tập thô F nếu thỏa điều kiện :Với mọi
: $f|_S = f|_{S'}$.
' nghĩa của định nghĩa này nêu lên: Tính phân bố đều của các hàm thô thuộc F
ới các thuộc tính bị lược bỏ.
ı: Xét hệ thông tin S là hệ thông tin liên quan đến tất cả các thuộc tính được cho
thống kê tần suất bệnh nhân viêm màng não có trong [1]. Có thể chỉ ra một S' là
ông tin nhận được từ S khi bỏ đi các thuộc tính chẳng hạn thuộc tính *ho.*
)ịnh nghĩa trên định hướng thu gọn hệ thông tin (tinh giản nó thông qua việc giảm
ợng các thuộc tính) để làm đơn giản khi giải các bài toán.
Khái niệm hệ thông tin thu gọn nhằm đạt được mục đích thu gọn hệ thông tin thông
iệc bỏ bớt các thuộc tính "vô ích". Tuy nhiên, trong nhiều trường hợp, không chỉ
ở mức quan tâm đến một lớp tập thô, mà quan trọng hơn là nhận được các chỉ
phân biệt tập thô này với tập thô khác hay cũng vậy nhận biết một tập thô trong
ã cho nào đó.
nghĩa: *Cho S là một hệ thông tin còn a là một thuộc tính ($a \in A$); cho L là một*

lớp các tập thô trong S với $f \in L$. Nói rằng a là xác định f nếu như:

$$\forall Z = \sigma(a, v_a) \quad f|_e \gg g|_e \cdot \text{ hoặc } \cdot \; g|_e \gg f|_e \quad (\forall e \in E_S;\; e \subset Z;\; \forall g \in L,\; g \neq f)$$

Kí hiệu $\gg$ để chỉ quan hệ "lớn hơn rất nhiều".

Chú ý rằng, ý nghĩa của định nghĩa này ở chỗ thông tin về tập thô f sẽ được hẳn so với các tập thô còn lại.

Ví dụ: Trong hệ thông tin về bệnh viêm màng não của trẻ em thì thuộc tính *dịch tủy lần 1* xác định bệnh viêm màng não.

Cho S là một hệ thông tin còn S' là một hệ thông tin thô hơn S; cho L là mộ các tập thô trong S với $f \in L$. Nói rằng S' *là xác định* f nếu như:

$$\frac{f}{g}|_{S'} \gg \frac{f}{g}|_S \quad \forall g \in L,\; g \neq f,\; g \neq 0$$

(trong công thức trên bỏ qua mọi trường hợp phân số vô nghĩa).

Ví dụ, trong hệ thống các bệnh nhân trẻ em có triệu chứng của bệnh lao màng Tập các tập chưa xác định trên S là $L =$ { bệnh lao màng não, bệnh sốt cao, bệnh }. Chúng ta chú ý hệ thông tin thô S' bao gồm các thuộc tính: *sốt cao, dịch não tủ* 1. Lúc đó, S'xác định bệnh lao màng não.

TÀI LIỆU THAM KHẢO

1. Phạm Kim Thanh, Đặng Ứng Vận. Sử dụng máy vi tính trong việc xây dựng chuẩn chẩn đoán lao màng não trẻ em. *Tạp chí khoa học*, Đại học Quốc gia Hà số 1-1994.
2. Hà Quang Thụy. Một thuật toán tìm ngữ nghĩa một term trong hệ thông tin. *chí khoa học Đại học Tổng hợp Hà Nội*, số 3 - 1987.
3. Dubois Didier, Prade Henri. Possibility Therory: An Approach to Comput Processing of Uncertainty. CNRS, *Languages and Computer Systems (LSI), versity of Toulouse* III (1986). Translated in English by University of Camb (1986).
4. D.D. Stephen, S. S. Jeffery, Valtorta Marco. Statistical Consistency With Dem Rule on Diagnostic Trees Having Uncertain Performance Parameters. *Interna Journal of Approximate Reasoning*. No. 6, Vol. 1, 1992.
5. Kacprzyk Janusz and Ziólkowski Andrzej. Database Querries with Fuzzy Ling Quantifiers. *IEEE Transactions on systems, Man and Cybernetics*, Vol. SM No 3 May/June, 1986.

VNU. JOURNAL OF SCIENCE, Nat. Sci., t. XII, n°3, 1996

ROUGH SETS ESTIMATION OF INFORMATION SYSTEMS

Ha Quang Thuy
College of Natural Sciences - VNU

Information systems and the concept of rough sets have been carried out by Pa For information systems, Pawlak gave the question language for problems of se objects which answer is a description set. He also concerned with rough sets.

In this article, we introduce a definition of rough sets in a measurable spa objects. Since the determination of rough sets depends on information systems d them then the comparison problem of information systems is dealt with. For a c rough set, it is possible to diminish the size of information systems by saving the ca for recognizing the rough set.