Tạp chí Tin học và Điều khiển học, T.25, S.1 (2009), 69–78

MỘT THUẬT TOÁN TÌM TẬP RÚT GỌN TRONG BẢNG QUYẾT ĐỊNH KHÔNG ĐẦY ĐỦ

HOÀNG THỊ LAN GIAO

Khoa CNTT, Trường Đại học Khoa học- Đại học Huế

Abstract. The aim of this work is to generalize the concept of knowledge reduction to class of incomplete decision tables. By innovating the method of J. Liang and Z. Xu [3], which based on rough entropy, we establish a heuristic algorithm for finding a reduct of incomplete decision table. The time complexity of this algorithm is $O(kn^2 \log n)$, where k is the number of conditional attributes and n is the number of objects in data table.

Tóm tắt. Bài báo mở rộng khái niệm rút gọn tri thức lên lớp các bảng quyết định không đầy đủ. Bằng cách cải tiến phương pháp của Jiye Liang và Zongben Xu [3], dựa vào entropy thô đã thiết lập một thuật tóan Heuristic để tìm rút gọn của bảng quyết định không đầy đủ. Độ phức tạp của thuật toán này là $O(kn^2 \log n)$, với k là số thuộc tính điều kiện và n là số đối tượng trong bảng.

1. MỞ ĐẦU

Đối với các mô hình dữ liệu lớn, rất dễ xãy ra tình trạng dữ liệu bị thiếu bởi nhiều lý do khác nhau. Việc rút gọn tri thức sẽ gặp khó khăn trong các hệ thống thông tin không đầy đủ nói chung và đặc biệt trong bảng quyết định, các luật quyết định đưa ra sẽ không đầy đủ. Thông thường, dữ liệu thiếu thông tin rơi vào một trong ba trường hợp [9]: không có thông tin về dữ liệu (không biết một người nào đó có số điện thoại hay không), không tồn tại (không có chức vụ), tồn tại nhưng không biết (ngày sinh). Trong khuôn khổ bài báo, ta chỉ nghiên cứu bảng quyết định không đầy đủ với giá trị thuộc tính bị mất do tồn tại nhưng không biết. Một thuật toán tìm tập rút gọn dựa vào khái niêm entropy thô với bảng chứa dữ liệu bị thiếu loại này cũng đã được đề xuất.

2. CÁC KHÁI NIỆM

2.1. Hệ thống thông tin không đầy đủ

Cho $\mathcal{A} = (U, A)$ là một hệ thống thông tin với U là tập hữu hạn, khác rỗng các đối tượng và A là tập hữu hạn, khác rỗng các thuộc tính. Với mỗi $u \in U$ và $a \in A$ ta ký hiệu $u(a)$ là giá trị thuộc tính a của đối tượng u. Nếu $X \subseteq A$ là một tập các thuộc tính ta ký hiệu $u(X)$ là bộ gồm các giá trị $u(a)$ với $a \in X$. Vì vậy, nếu u và v là hai đối tượng thuộc U, ta sẽ nói $u(X) = v(X)$ nếu $u(a) = v(a)$ với mọi thuộc tính $a \in X$. Tập tất cả các giá trị của thuộc tính a là V_a.

Bảng quyết định là một hệ thống thông tin với tập thuộc tính A gồm hai tập con rời nhau C và D, trong đó tập C được gọi là tập thuộc tính điều kiện và D là tập thuộc tính quyết

định.

Hệ thống thông tin $\mathcal{A}$ được gọi là không đầy đủ nếu tồn tại thuộc tính $a \in A$ và đối tương $u \in U$ mà giá trị $u(a)$ bị mất hay nói cách khác V_a chứa giá trị null. Giá trị này trong bảng chúng ta sẽ ký hiệu bởi ký tự "*".

Tương tự như vậy, ta có khái niệm bảng quyết định không đầy đủ.

2.2. Quan hệ không phân biệt được trên hệ thống thông tin không đầy đủ

Cho hệ thống thông tin $\mathbb{A} = (U, A)$. Với mỗi tập con các thuộc tính $B \subseteq A$, ta định nghĩa quan hệ hai ngôi $\mathrm{IND}(B)$ trên U xác định bởi:

$$\mathrm{IND}(B) = \{(u, v) \in U \times U \mid u(B) = v(B)\}.$$

$\mathrm{IND}(B)$ được gọi là quan hệ $B-$không phân biệt được. Dễ kiểm chứng được rằng đây là một quan hệ tương đương trên U.

Nếu $(u, v) \in \mathrm{IND}(B)$ thì hai đối tượng u và v không phân biệt được bởi các thuộc tính trong B. Lớp tương đương chứa phần tử u được ký hiệu $[u]_B$. Khi đó quan hệ $\mathrm{IND}(B)$ được xác định hoàn toàn bởi các lớp tương đương $[u]_B$, $u \in U$. Tập hợp thương của quan hệ $\mathrm{IND}(B)$ được ký hiệu U/B, tức là $U/B = \{[u]_B \mid u \in U\}$.

Trong trường hợp hệ thống không đầy đủ, ta định nghĩa quan hệ hai ngôi trên U, ký hiệu $\mathrm{SIM}(B)$, với mỗi $B \subseteq A$.

Định nghĩa 2.1. [3]

$$\mathrm{SIM}(B) = \{(u, v) \in U \times U \mid \forall a \in B, u(a) = v(a) \text{ hoặc } u(a) = * \text{ hoặc } v(a) = *\}.$$

Rõ ràng,

$$\mathrm{SIM}(B) = \bigcap_{b \in B} \mathrm{SIM}(\{b\}).$$

Ký hiệu $S_B(u) = \{v \in U \mid (u, v) \in \mathrm{SIM}(B)\}$ và gọi là lớp tolerance của quan hệ $\mathrm{SIM}(B)$, $S_B(u)$ là tập tối đa các đối tượng v không phân biệt được với u bằng tập thuộc tính B. Khi đó trên U ta phân lớp các đối tượng dựa vào quan hệ $\mathrm{SIM}(B)$, mỗi lớp là một tập $S_B(u), u \in U$. Họ các lớp này được ký hiệu $U/\mathrm{SIM}(B)$, đây là một phủ của U ($\bigcup_{u \in U} S_B(u) = U$), và mỗi phần tử trong họ này đều khác rỗng do quan hệ $\mathrm{SIM}(B)$ có tính phản xạ ($S_B(u) \supseteq \{u\}$). Nói chung, đây không phải là một phân hoạch của U và $\mathrm{SIM}(B)$ cũng không phải là một quan hệ tương đương. Tuy nhiên, trong trường hợp hệ thống đầy đủ, quan hệ $\mathrm{SIM}(B)$ trùng với quan hệ $\mathrm{IND}(B)$, và khi đó $S_B(u) = [u]_B, \forall u \in U$ hay $U/\mathrm{SIM}(B) = U/B$.

Ví dụ 2.1. Xét hệ thống thông tin không đầy đủ với ba thuộc tính: $A =$\{Thân nhiệt, Đau đầu, Đau cơ\} cho trong Bảng 1.

Bảng 1

U	Thân nhiệt	Đau đầu	Đau cơ
1	cao	*	không
2	rất cao	có	có
3	*	không	không
4	cao	có	có
5	cao	*	có
6	bình thường	có	không
7	bình thường	không	có
8	*	có	*

Ta có:

$S_A(1) = \{1, 3, 8\}; S_A(2) = \{2, 8\};$

$S_A(3) = \{1, 3\}; S_A(4) = \{4, 5, 8\};$

$S_A(5) = \{4, 5, 8\}; S_A(6) = \{6, 8\};$

$S_A(7) = \{7\}; S_A(8) = \{1, 2, 4, 5, 6, 8\}.$

$U/\text{SIM}(A) = \{\{1, 8\}; \{2, 8\}; \{1, 3\}; \{4, 5, 8\}; \{4, 5, 8\}; \{6, 8\}; \{7, 8\}; \{1, 2, 4, 5, 6, 8\}\}.$

Với $B = \{$ Thân nhiệt, Đau cơ$\}$.

$S_B(1) = \{1, 3, 8\}; S_B(2) = \{2, 8\};$

$S_B(3) = \{1, 3, 6, 8\}; S_B(4) = \{4, 5, 8\};$

$S_B(5) = \{4, 5, 8\}; S_B(6) = \{3, 6, 8\};$

$S_B(7) = \{7, 8\}; S_B(8) = \{1, 2, 3, 4, 5, 6, 7, 8\}.$

Định nghĩa 2.2. Thuộc tính điều kiện $c \in C$ được gọi là *không cốt yếu* trong bảng quyết định $\mathbb{T}$ nếu $U/\text{SIM}(C \setminus \{c\}) = U/\text{SIM}(C \setminus \{c\} \cup D)$. Ngược lại, c được gọi là cốt yếu.

Bảng quyết định không đầy đủ $\mathbb{T}$ được gọi là độc lập nếu mọi thuộc tính $c \in C$ đều cốt yếu. Tập tất cả các thuộc tính cốt yếu trong $\mathbb{T}$ được gọi là lõi và được ký hiệu $\text{Core}(C)$. Từ đây, ta quy ước viết Core thay cho $\text{Core}(C)$.

Định nghĩa 2.3. Tập các thuộc tính $R \subseteq C$ được gọi là một rút gọn của tập thuộc tính điều kiện C nếu $\mathbb{T}' = (U, R \cup D)$ là độc lập và $U/\text{SIM}(R) = U/\text{SIM}(R \cup D)$.

Rõ ràng là có thể có nhiều tập rút gọn của C. Ta ký hiệu $\text{RED}(C)$ là tập tất cả các rút gọn của C trong $\mathbb{T}$. Một thuộc tính là cốt yếu khi và chỉ khi nó thuộc vào mọi tập rút gọn của C.

Hai định nghĩa trên là sự mở rộng của các định nghĩa về thuộc tính không cốt yếu và rút gọn trong bảng quyết định (đầy đủ). Rõ ràng khi T đầy đủ, ta có

$$
\begin{aligned}
& U/SIM(C \setminus \{c\}) = U/SIM((C \setminus \{c\}) \cup D) \\
\Leftrightarrow \quad & U/(C \setminus \{c\}) = U/((C \setminus \{c\}) \cup D) \\
\Leftrightarrow \quad & \text{Card}(\prod(C \setminus \{c\})) = \text{Card}(\prod(C \setminus \{c\}) \cup D) \\
\Leftrightarrow \quad & c \text{ là thuộc tính không cốt yếu} \quad (\textit{theo } [2]).
\end{aligned}
$$

3. ENTROPY THÔ

Khái niệm entropy thô đã được giới thiệu trong [1, 6], ở đây, ta nhắc lại khái niệm entropy thô của tri thức trong hệ thống không đầy đủ do Jiye Liang và Zongben Xu đề xuất.

Định nghĩa 3.1. Cho hệ thống thông tin không đầy đủ $\mathbb{A} = (U, A)$ và$B \subseteq A$. Entropy thô của tri thức B là giá trị

$$E(B) = -\sum_{i=1}^{n} \frac{|S_B(x_i)|}{n} \log \frac{1}{|S_B(x_i)|},$$

với $\text{Card}(U) = n$, $|S_B(x)|$ là ký hiệu của $\text{Card}(S_B(x))$ và $\log x$ là ký hiệu của $\log_2 x$. Khi đó $\frac{|S_B(x_i)|}{n}$ là xác suất để một đối tượng trong U thuộc lớp $S_B(x_i)$ và $\frac{1}{|S_B(x_i)|}$ là xác suất của một đối tượng trong lớp $S_B(x_i)$ và bằng x_i.

Từ định nghĩa trên, ta có các tính chất của entropy thô trong hệ thống thông tin không đầy đủ $\mathbb{A} = (U, A)$ [3].

Tính chất 3.1. *Cho $P, Q \subseteq A$. Nếu tồn tại một song ánh $h : U/\text{SIM}(P) \to U/\text{SIM}(Q)$ sao cho*

$$|h(S_P(x_i))| = |S_P(x_i))|, i = 1, 2, \cdots, |U|,$$

thì

$$E(P) = E(Q).$$

Điều này có nghĩa là entropy thô của tri thức là bất biến đối với tập các lớp $U/\text{SIM}(P)$ đẳng cấu.

Tính chất 3.2. (Đơn điệu giảm) *Cho $B, C \subseteq A$. Nếu $B \subseteq C$ (rõ ràng khi đó $U/\text{SIM}(C) \subseteq U/\text{SIM}(B)$) thì $E(C) \leqslant E(B)$.*

Tính chất này chỉ ra rằng số lớp tolerance trong $U/\text{SIM}(B)$ càng lớn thì entropy thô của tri thức càng giảm.

Tính chất 3.4. (Tương đương) *Cho $C \subseteq A$. Khi đó $U/\text{SIM}(C) = U/\text{SIM}(A)$ nếu và chỉ nếu $E(C) = E(A)$.*

Như vậy, nếu tập $C \subseteq A$ thì sự phân lớp của quan hệ $\text{SIM}(C)$ tương đương với sự phân lớp của quan hệ $\text{SIM}(A)$ khi và chỉ khi entropy của hai tập thuộc tính này bằng nhau.

Tính chất 3.5. (Cực đại) *Cho $C \subseteq A$. Giá trị cực đại của entropy thô của tri thức C bằng $|U|\log|U|$, đạt được khi $S_C(x) = U \forall x \in U$. Lúc đó $U/\text{SIM}(C) = \{U\}$.*

Entropy đạt cực đại khi mọi lớp của $U/\text{SIM}(C)$ đều bằng U. Điều đó cũng có nghĩa là thông tin nhận được từ tập thuộc tính tương ứng có độ tin cậy bé nhất.

Tính chất 3.5. (Cực tiểu) *Cho $C \subseteq A$. Giá trị cực tiểu của entropy thô của tri thức C bằng 0, đạt được khi $S_C(x) = \{x\} \forall x \in U$.*

Entropy đạt cực tiểu khi mỗi lớp tolerance chỉ chứa đúng một phần tử và như vậy thông tin nhận được dựa vào tập thuộc tính tương ứng chắc chắn nhất.

Mệnh đề 3.1. *Cho $\mathbb{T} = (U, C \cup D)$ là bảng quyết định không đầy đủ. Khi đó, tập $R \subseteq C$ là một rút gọn của tập thuộc tính điều kiện C trong bảng nếu và chỉ nếu R là tập tối tiểu thoả*

mãn $E(R) = E(R \cup D)$.
Chứng minh. Đặt $\mathbb{T}' = (U, R \cup D)$.

R là tập rút gọn $\Leftrightarrow$ T'độc lập và$U/SIM(R) = U/SIM(R \cup D)$
$\Leftrightarrow$ $\forall c \in R, U/SIM(R \setminus \{c\}) \neq U/SIM(R \setminus \{c\} \cup D)$ và $U/SIM(R) = U/SIM(R \cup D)$
$\Leftrightarrow$ R là tập tối tiểu và $E(R) = E(R \cup D)$ (theo tính chất tương đương)

■

4. Ý NGHĨA CỦA THUỘC TÍNH

Định nghĩa 4.1. Cho $\mathbb{T} = (U, C \cup D)$ là bảng quyết định không đầy đủ. Ý nghĩa của thuộc tính c trong C, ký hiệu $\text{sig}_{C\setminus\{c\}}(c)$, được xác định

$$sig_{C\setminus\{c\}}(c) = E(C \setminus \{c\}) - E(C \setminus \{c\} \cup D).$$

Mệnh đề 4.1.
a) $0 \leqslant sig_{C\setminus\{c\}}(c) \leqslant |U|\log|U|$.
b) $c \in C$ *là thuộc tính cốt yếu trong* C *nếu và chỉ nếu* $\text{sig}_{C\setminus\{c\}}(c) > 0$. *Khi đó,*

$$\text{Core}(C) = \{c \in C \mid \text{sig}_{C\setminus\{c\}}(c) > 0\}.$$

Chứng minh.
a) $0 \leqslant E(C \setminus \{c\}) - E(C \setminus \{c\} \cup D)$ vì$C \setminus \{c\} \subseteq C \setminus \{c\} \cup D$. Mặt khác, theo tính chất cực tiểu $E(C \setminus \{c\} \cup D) \geqslant 0$ và tính chất cực đại thì $E(C \setminus \{c\}) \leqslant |U|log|U|$. Ta có $|U|log|U| \geqslant E(C \setminus \{c\}) \geqslant E(C \setminus \{c\}) - E(C \setminus \{c\} \cup D) \geqslant 0$ hay $0 \leqslant sig_{C\setminus\{c\}}(c) \leqslant |U|\log|U|$.
b) $c \in C$ là thuộc tính cốt yếu trong C nếu và chỉ nếu $U/\text{SIM}(C\setminus\{c\}) \neq U/\text{SIM}(C\setminus\{c\}\cup D)$. Tức là, $E(C \setminus \{c\}) - E(C \setminus \{c\} \cup D) > 0$ hay $\text{sig}_{C\setminus\{c\}}(c) > 0$. Rõ ràng, khi đó $\text{Core}(C) = \{c \in C \mid \text{sig}_{C\setminus\{c\}}(c) > 0\}$. ■

Định nghĩa 4.2. Cho $\mathbb{T} = (U, C \cup D)$ là bảng quyết định không đầy đủ, $R \subseteq C$ và $c \in C \setminus R$. Ý nghĩa của thuộc tính c đối với R, ký hiệu $\text{sig}_R(c)$, được xác định

$$sig_R(c) = E(R \cup D) - E(R \cup \{c\} \cup D).$$

5. THUẬT TOÁN TÌM TẬP RÚT GỌN

Dựa vào các tính chất của entropy thô và ý nghĩa của một thuộc tính, bài báo đề xuất một thuật toán Heuristic tìm tập rút gọn trong bảng quyết định không đầy đủ. Thuật toán này xuất phát từ tập lõi (bởi vì mọi tập rút gọn đều chứa tập lõi) và tìm cách bổ sung các thuộc tính cho đến khi nhận được tập rút gọn thực sự. Thuộc tính được ưu tiên chọn bổ

sung tại mỗi bước là thuộc tính có ý nghĩa lớn nhất. Cụ thể, thuật toán có thể được trình bày chi tiết như sau.

Vào: Bảng quyết định không đầy đủ $\mathbb{T} = (U, C \cup D)$.

Ra: Một rút gọn R của $\mathbb{T}$.

Phương pháp

B1. Tính $\text{Core}(C) := \{c \in C \mid \text{sig}_{C\setminus\{c\}}(c) > 0\}$

B2. $R := \text{Core}(C)$

B3. Tính $E(R)$ và $E(R \cup D)$

B4. While $E(R) \neq E(R \cup D)$ do

1. For $c \in C \setminus R$ do
   Tính $\text{sig}_R(c)$
2. Chọn c sao cho $\text{sig}_R(c) = max\{\text{sig}_R(c') \mid c' \in C \setminus R\}$
   $R := R \cup \{c\}$
3. Tính $E(R), E(R \cup D)$

B5. $R := R \setminus \text{Core}(C)$

B6. For $c \in R$ do

1. If $E((R \setminus \{c\}) \cup \text{Core}(C)) = E((R \setminus \{c\}) \cup \text{Core}(C) \cup D)$ then
2. $R := R \setminus \{c\}$

B7. $R := R \cup \text{Core}(C)$

Độ phức tạp tính toán của thuật toán này được xác định bởi vòng lặp while ở $B4$. (vòng lặp này thực hiện tối đa $\text{Card}(C)$ lần, do sau mỗi bước, số thuộc tính được chọn tăng lên 1). Tại mỗi bước của vòng lặp, ta cần tính $\text{sig}_R(c)$, với mỗi $c \in C \setminus R$ và tính $E(R), E(R \cup D)$. Việc tính $\text{sig}_R(c)$ cũng đưa về tính entropy thô. Vì vậy, nếu gọi n là số đối tượng trong bảng quyết định và k là số thuộc tính điều kiện tương ứng, thì tại mỗi bước của vòng lặp ta cần tính tối đa k giá trị entropy thô. Mặt khác, dựa vào công thức của entropy thô, ta đánh giá được độ phức tập tính toán của phép toán này là $O(n^2 \log n)$, vì mỗi lần tìm một lớp $U/\text{SIM}(R)$, ta phải sắp xếp các đối tượng trong U với độ phức tạp của phép sắp xếp là $O(n \log n)$. Như vậy có thể khẳng định độ phức tạp tính toán của thuật toán là $O(kn^2 \log n)$.

Ví dụ 5.1. Xét bảng quyết định trong [11] được cho bởi Bảng 2.

Bảng 2

U	c_1	c_2	c_3	c_4	d
u_1	low	2	compact	4	high
u_2	low	4	sub	6	low
u_3	medium	4	compact	4	high
u_4	high	2	compact	6	low
u_5	high	4	compact	4	low
u_6	low	4	compact	4	high
u_7	high	4	sub	6	low
u_8	low	2	sub	6	low

Bảng 2 lưu thông tin về 8 chiếc xe hơi nhận được thông qua các thuộc tính điều kiện $C = \{c_1(Weight), c_2(Door), c_3(Size), c_4(Cylinder)\}$ và thuộc tính quyết định $D = \{d(Mileage)\}$.

Trong bảng này, ta có tập thuộc tính lõi là $\{c_1\}$ và có hai tập rút gọn $R_1 = \{c_1, c_3\}$, $R_2 = \{c_1, c_4\}$.

Giả sử vì một lý do nào đó ta không có được thông tin đầy đủ như trên, một số giá trị của các thuộc tính bị mất (Bảng 3)

Bảng 3

U	c_1	c_2	c_3	c_4	d
u_1	low	2	compact	4	high
u_2	*	4	sub	6	low
u_3	medium	4	compact	4	high
u_4	high	2	compact	*	low
u_5	high	4	*	4	low
u_6	low	4	compact	4	high
u_7	high	*	sub	6	low
u_8	low	2	sub	6	low

Thực hiện từng bước thuật toán trên ta thu được kết quả sau:

B1. Tìm lõi:

Xét c_1

$$\begin{aligned}
S_{C\setminus\{c_1\}}(u_1) &= \{u_1, u_4\}; & S_{(C\cup D)\setminus\{c_1\}}(u_1) &= \{u_1\},\\
S_{C\setminus\{c_1\}}(u_2) &= \{u_2, u_7\}; & S_{(C\cup D)\setminus\{c_1\}}(u_2) &= \{u_2, u_7\},\\
S_{C\setminus\{c_1\}}(u_3) &= \{u_3, u_5, u_6\}; & S_{(C\cup D)\setminus\{c_1\}}(u_3) &= \{u_3, u_6\},\\
S_{C\setminus\{c_1\}}(u_4) &= \{u_1, u_4\}; & S_{(C\cup D)\setminus\{c_1\}}(u_4) &= \{u_4\},\\
S_{C\setminus\{c_1\}}(u_5) &= \{u_3, u_5, u_6\}; & S_{(C\cup D)\setminus\{c_1\}}(u_5) &= \{u_5, u_6\},\\
S_{C\setminus\{c_1\}}(u_6) &= \{u_3, u_5, u_6\}; & S_{(C\cup D)\setminus\{c_1\}}(u_6) &= \{u_3, u_6\},\\
S_{C\setminus\{c_1\}}(u_7) &= \{u_2, u_7, u_8\}; & S_{(C\cup D)\setminus\{c_1\}}(u_7) &= \{u_2, u_7, u_8\},\\
S_{C\setminus\{c_1\}}(u_8) &= \{u_7, u_8\}; & S_{(C\cup D)\setminus\{c_1\}}(u_8) &= \{u_7, u_8\}.
\end{aligned}$$

$$\begin{aligned}
Sig_{C\setminus\{c_1\}}(c_1) &= E(C\setminus\{c_1\}) - E((C\setminus\{c_1\})\cup D)\\
&= -\sum_{i=1}^{n}\frac{|S_{C\setminus\{c_1\}}(u_i)|}{n}\log\frac{1}{|S_{C\setminus\{c_1\}}(u_i)|} + \sum_{i=1}^{n}\frac{|S_{(C\setminus\{c_1\})\cup D}(u_i)|}{n}\log\frac{1}{|S_{(C\setminus\{c_1\})\cup D}(u_i)|}\\
&= \frac{9}{8}\log 3 - \frac{1}{4} > 0.
\end{aligned}$$

Do đó c_1 là thuộc tính cốt yếu.

Xét c_2

$$\begin{aligned}
S_{C\setminus\{c_2\}}(u_1) &= \{u_1, u_6\}; & S_{(C\cup D)\setminus\{c_2\}}(u_1) &= \{u_1, u_6\},\\
S_{C\setminus\{c_2\}}(u_2) &= \{u_2, u_7, u_8\}; & S_{(C\cup D)\setminus\{c_2\}}(u_2) &= \{u_2, u_7, u_8\},\\
S_{C\setminus\{c_2\}}(u_3) &= \{u_3\}; & S_{(C\cup D)\setminus\{c_2\}}(u_3) &= \{u_3\},\\
S_{C\setminus\{c_2\}}(u_4) &= \{u_4, u_5\}; & S_{(C\cup D)\setminus\{c_2\}}(u_4) &= \{u_4, u_5\},
\end{aligned}$$

$$\begin{aligned}
S_{C\setminus\{c_2\}}(u_5) &= \{u_4,u_5\}; & S_{(C\cup D)\setminus\{c_2\}}(u_5) &= \{u_4,u_5\},\\
S_{C\setminus\{c_2\}}(u_6) &= \{u_1,u_6\}; & S_{(C\cup D)\setminus\{c_2\}}(u_6) &= \{u_1,u_6\},\\
S_{C\setminus\{c_2\}}(u_7) &= \{u_2,u_7\}; & S_{(C\cup D)\setminus\{c_2\}}(u_7) &= \{u_2,u_7\},\\
S_{C\setminus\{c_2\}}(u_8) &= \{u_2,u_8\}; & S_{(C\cup D)\setminus\{c_2\}}(u_8) &= \{u_2,u_8\}.
\end{aligned}$$

$$Sig_{C\setminus\{c_2\}}(c_2) = E(C\setminus\{c_2\}) - E((C\setminus\{c_2\})\cup D) = 0.$$

Do đó c_2 không phải là thuộc tính cốt yếu, tương tự, c_3 vàc_4 cũng không phải là thuộc tính cốt yếu. Hay nói cách khác Core $= \{c_1\}$.

B2. Đặt $R := \{c_1\}$.

B3. Tính $E(R)$ và $E(R\cup D)$.

$$\begin{aligned}
S_R(u_1) &= \{u_1,u_2,u_6,u_8\}; & S_{R\cup D}(u_1) &= \{u_1,u_6\},\\
S_R(u_2) &= \{u_1,u_2,u_3,u_4,u_5,u_6,u_7,u_8\}; & S_{R\cup D}(u_2) &= \{u_2,u_4,u_5,u_7,u_8\},\\
S_R(u_3) &= \{u_2,u_3\}; & S_{R\cup D}(u_3) &= \{u_3\},\\
S_R(u_4) &= \{u_2,u_4,u_5,u_7\}; & S_{R\cup D}(u_4) &= \{u_2,u_4,u_5,u_7\},\\
S_R(u_5) &= \{u_2,u_4,u_5,u_7\}; & S_{R\cup D}(u_5) &= \{u_2,u_4,u_5,u_7\},\\
S_R(u_6) &= \{u_1,u_6,u_8\}; & S_{R\cup D}(u_6) &= \{u_1,u_6\},\\
S_R(u_7) &= \{u_2,u_4,u_5,u_7\}; & S_{R\cup D}(u_7) &= \{u_2,u_4u_5,u_7\},\\
S_R(u_8) &= \{u_1,u_2,u_6,u_8\}; & S_{R\cup D}(u_8) &= \{u_2,u_8\}.
\end{aligned}$$

B4. Ta có, $E(R)\neq E(R\cup D)$. Vì vậy chúng ta sẽ kiểm tra xem trong ba thuộc tính còn lại thuộc tính nào có ý nghĩa hơn đối với R để chọn bổ sung vào R.

Trước hết ta xét $Sig_R(c_2) = E(R\cup D) - E(R\cup\{c_2\}\cup D)$.

$$\begin{aligned}
S_{R\cup\{c_2\}\cup D}(u_1) &= \{u_1\},\\
S_{R\cup\{c_2\}\cup D}(u_2) &= \{u_2,u_5,u_7\},\\
S_{R\cup\{c_2\}\cup D}(u_3) &= \{u_3\},\\
S_{R\cup\{c_2\}\cup D}(u_4) &= \{u_4,u_7\},\\
S_{R\cup\{c_2\}\cup D}(u_5) &= \{u_2,u_5,u_7\},\\
S_{R\cup\{c_2\}\cup D}(u_6) &= \{u_6\},\\
S_{R\cup\{c_2\}\cup D}(u_7) &= \{u_2,u_4,u_5,u_7\},\\
S_{R\cup\{c_2\}\cup D}(u_8) &= \{u_8\}.
\end{aligned}$$

$$Sig_R(c_2) \quad = \frac{1}{8}((3\times 4\times \ln 4 + 3\times 2\times \ln 2 + 5\times \ln 5) - (2\times 3\times \ln 3 + 4\times \ln 4 + 2\ln 2)).$$

Tương tự ta có

$$Sig_R(c_3) = \frac{1}{8}((3\times 4\times \ln 4 + 3\times 2\times \ln 2 + 5\times \ln 5) - (3\times \ln 3 + 2\times 4\times \ln 4 + 4\times 2\times \ln 2)),$$

$$Sig_R(c_4) = \frac{1}{8}((3 \times 4 \times \ln 4 + 3 \times 2 \times \ln 2 + 5 \times \ln 5) - (3 \times \ln 3 + 2 \times 4 \times \ln 4 + 4 \times 2 \times \ln 2)).$$

Như vậy $sig_R(c_2)$ lớn nhất và ta chọn c_2 bổ sung vào R.

Bây giờ ta xét $R = \{c_1, c_2\}$, ta có $E(R) \neq E(R \cup D)$. Do đó lại tính tiếp $Sig_R(c_3)$ và $Sig_R(c_4)$.

$$\begin{aligned}
S_{R\cup\{c_3\}\cup D}(u_1) &= \{u_1\}; & S_{R\cup\{c_3\}\cup D}(u_2) &= \{u_2, u_5, u_7\},\\
S_{R\cup\{c_3\}\cup D}(u_3) &= \{u_3\}; & S_{R\cup\{c_3\}\cup D}(u_4) &= \{u_4\},\\
S_{R\cup\{c_3\}\cup D}(u_5) &= \{u_2, u_5, u_7\}; & S_{R\cup\{c_3\}\cup D}(u_6) &= \{u_6\},\\
S_{R\cup\{c_3\}\cup D}(u_7) &= \{u_2, u_5, u_7\}; & S_{R\cup\{c_3\}\cup D}(u_8) &= \{u_8\}.
\end{aligned}$$

$$\begin{aligned}
S_{R\cup\{c_4\}\cup D}(u_1) &= \{u_1\}; & S_{R\cup\{c_4\}\cup D}(u_2) &= \{u_2, u_7\},\\
S_{R\cup\{c_4\}\cup D}(u_3) &= \{u_3\}; & S_{R\cup\{c_4\}\cup D}(u_4) &= \{u_4, u_7\},\\
S_{R\cup\{c_4\}\cup D}(u_5) &= \{u_5\}; & S_{R\cup\{c_4\}\cup D}(u_6) &= \{u_6\},\\
S_{R\cup\{c_4\}\cup D}(u_7) &= \{u_2, u_4, u_7\}; & S_{R\cup\{c_4\}\cup D}(u_8) &= \{u_8\}.
\end{aligned}$$

$$Sig_R(c_3) = \frac{1}{8}((2 \times 3 \times \ln 3 + 4 \times \ln 4 + 2 \times \ln 2) - (3 \times 3 \times \ln 3))$$

và

$$Sig_R(c_4) = \frac{1}{8}((2 \times 3 \times \ln 3 + 4 \times \ln 4 + 2 \times \ln 2) - (2 \times 2 \times \ln 2 + 3 \times \ln 3)).$$

Nên ta chọn C_4 vào R. Lúc này, với $R = \{c_1, c_2, c_4\}$ ta có $E(R) = E(R \cup D)$.

B5 và B6. Theo trên ta có $\{c_1, c_2\}$ không phải là tập rút gọn. Ta chỉ còn xét $\{c_1, c_4\}$.

$$\begin{aligned}
S_{\{c_1,c_4\}}(u_1) &= \{u_1, u_6\}; & S_{\{c_1,c_4\}\cup D}(u_1) &= \{u_1, u_6\},\\
S_{\{c_1,c_4\}}(u_2) &= \{u_2, u_4, u_7, u_8\}; & S_{\{c_1,c_4\}\cup D}(u_2) &= \{u_2, u_4, u_7, u_8\},\\
S_{\{c_1,c_4\}}(u_3) &= \{u_3\}; & S_{\{c_1,c_4\}\cup D}(u_3) &= \{u_3\},\\
S_{\{c_1,c_4\}}(u_4) &= \{u_2, u_4, u_5, u_7\}; & S_{\{c_1,c_4\}\cup D}(u_4) &= \{u_2, u_4, u_5, u_7\},\\
S_{\{c_1,c_4\}}(u_5) &= \{u_4, u_5\}; & S_{\{c_1,c_4\}\cup D}(u_5) &= \{u_4, u_5\},\\
S_{\{c_1,c_4\}}(u_6) &= \{u_1, u_6\}; & S_{\{c_1,c_4\}\cup D}(u_6) &= \{u_1, u_6\},\\
S_{\{c_1,c_4\}}(u_7) &= \{u_2, u_4, u_7\}; & S_{\{c_1,c_4\}\cup D}(u_7) &= \{u_2, u_4, u_7\},\\
S_{\{c_1,c_4\}}(u_8) &= \{u_2, u_8\}; & S_{\{c_1,c_4\}\cup D}(u_8) &= \{u_2, u_8\}.
\end{aligned}$$

Rõ ràng, $E(\{c_1, c_4\}) = E(\{c_1, c_4\} \cup D)$. Vì c_1 là thuộc tính cốt yếu và $\{c_1\}$ không phải là tập rút gọn. Vậy tập rút gọn đối với bảng quyết định không đầy đủ này là $R = \{c_1, c_4\}$.

TÀI LIỆU THAM KHẢO

[1] T. Beaubouef, P. E. Petry, G. Arora, Information-theoretic measures of uncertainty for rought sets and rough relational databases, *Information Sciences* **109** (1998) 535–563.

[2] Hoàng Thị Lan Giao, Một số thuật toán tìm tập rút gọn của bảng quyết định sử dụng các phép toán của đại số quan hệ, *Tạp chí Khoa học, Khoa học Tự nhiên và Công nghệ*, Đại học Quốc gia HàNội, **21** (2PT) (2005) 41–48.

[3] Jiye Liang, Zongben Xu, The algorithm on knowledge reduction in incomplete information systems, *International Journal of Uncertainty, Fizziness and Knowledge-Based Systems* **10** (1) (2002) 93–103.

[4] Jerzy W. Grzymala Busse, *Data with missing Attribute Values: Generalization of Indiscernibility Relation and Rule Induction*, Vol. 1, Published in Transactions on Rough Sets, Springer - Verlag, 2004 (78–95).

[5] Z.Pawlak, *Rough Sets - Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, Dordrecht, 1991.

[6] D.Slezak, Approximate reducts in decision table, *Pro. of IPMU'96*, Granada, July 1-5, 1996 (1159–1164).

[7] J. Stefanowski and A. Tsoukias, Incomplete information tables and rough classification, *Computational Intelligence* **17** (2001) 545–566.

[8] A. Skowron, C. Rauszer, The discernibility matrices and functions in information systems, Intelligent Decision Support, *Handbook of Applications and Advances of the Rough Sets Theory*, Kluwer, Dordrecht, 1992 (331–362).

[9] Hồ Thuần, Hoàng Thị Lan Giao, Mở rộng một số toán tử quan hệ lên cơ sở dữ liệu thiếu thông tin, *Tạp chí Tin học và Điều khiển học* **19** (4) (2003) 359–365.

[10] Hồ Thuần, Hoàng Thị Lan Giao, Một thuật toán tìm tập rút gọn sử dụng ma trận phân biệt được, *Chuyên san Các công trình nghiên cứu triển khai Viễn thông và CNTT* (15) (tháng 12/2005) 83–87.

[11] Xiaohua Hu, Jianchao han, T. Y. Lin, "A new rough sets model based on database systems, *Fundamenta Informaticae* **XX** (2004) 1–18.

Nhận bài ngày 5 - 10 - 2008