Tạp chí Tin học và Điều khiển học, T.27, S.2 (2011), 166–175

MỘT PHƯƠNG PHÁP RÚT GỌN THUỘC TÍNH TRONG BẢNG QUYẾT ĐỊNH DỰA TRÊN ENTROPY CẢI TIẾN

NGUYỄN LONG GIANG, VŨ ĐỨC THI,

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

Tóm tắt. Trong lý thuyết tập thô, nhiều thuật toán rút gọn thuộc tính dựa trên Entropy thông tin được đề xuất. Mặc dù cải tiến đáng kể độ phức tạp thời gian tính toán, tuy nhiên các thuật toán này đều chưa tìm được tập rút gọn tối thiểu trong các bảng quyết định không nhất quán. Trong bài báo này, chúng tôi đề xuất một độ đo Entropy cải tiến và chứng minh tập rút gọn dựa trên độ đo này tương đương với tập rút gọn dựa trên miền dương của Pawlak. Trên cơ sở đó, chúng tôi xây dựng một thuật toán rút gọn thuộc tính dựa trên Entropy cải tiến với độ phức tạp $O\left(|C|^2 |U|\right)$ với $|C|$ là số thuộc tính điều kiện và $|U|$ là số đối tượng.

Abstract. In rough set theory, many attribute reduction algorithms based on information entropy have been proposed. Although these algorithms reduce the time complexity, they can not obtain the minimal reduction in inconsistent decision tables. In this paper, we propose an improved entropy and we prove that the attribute reduction based on this entropy is equivalent to Pawlak's reduction in inconsistent decision tables. As a result, a complete heuristic algorithm is designed to find the attribute reduction based on improved entropy and its time complexity is $O\left(|C|^2 |U|\right)$, where $|C|$ is the number of condition attributes, and $|U|$ is the number of objects.

1. MỞ ĐẦU

Rút gọn thuộc tính là bài toán quan trọng nhất trong lý thuyết tập thô. Trong những năm gần đây, nhiều khái niệm khác nhau về tập rút gọn được đề xuất với mục tiêu tìm kiếm các thuật toán rút gọn thuộc tính hiệu quả phù hợp với các nhiệm vụ khai phá dữ liệu.

Xuất phát từ định nghĩa tập rút gọn của Pawlak [8], Skowron [1], Hu [10] đề xuất tập rút gọn dựa trên ma trận phân biệt và Wang [3, 4] đề xuất tập rút gọn dựa trên Entropy của Shannon. Trong [2, 4, 9] đã chứng minh tập rút gọn dựa trên Entropy của Shannon và tập rút gọn dựa trên ma trận phân biệt không tương đương với tập rút gọn của Pawlak trong trường hợp bảng quyết định không nhất quán. Trong [6], đã nghiên cứu mối liên hệ giữa ba khái niệm tập rút gọn nêu trên.

Nhằm giảm thiểu độ phức tạp trong công thức tính toán Entropy của Shannon, J.Liang và các cộng sự [5] đề xuất độ đo Entropy mới, P.Luo [7] đưa ra khái niệm tập rút gọn dựa trên Entropy mới và [13] đã chứng minh tập rút gọn này tương đương với tập rút gọn dựa trên ma

trận phân biệt. Với độ đo Entropy mới, trong [7] cũng chỉ ra tập rút gọn dựa trên Entropy mới không tương đương với tập rút gọn của Pawlak trong bảng quyết định không nhất quán. Do đó, các thuật toán tìm tập rút gọn dựa trên Entropy mới, mặc dù giảm thiểu độ phức tạp tính toán, tuy nhiên không tìm được tập rút gọn tối thiểu theo định nghĩa của Pawlak.

Nhằm khắc phục nhược điểm trên, bài báo xây dựng công thức Entropy cải tiến xuất phát từ công thức Entropy mới, đề xuất khái niệm tập rút gọn dựa trên Entropy cải tiến và chứng minh tập rút gọn này tương đương với tập rút gọn của Pawlak trong bảng quyết định không nhất quán. Trên cơ sở đó, xây dựng thuật toán rút gọn thuộc tính hiệu quả xuất phát từ tập thuộc tính nhân.

2. CÁC KHÁI NIỆM CƠ BẢN

Phần này sẽ trình bày một số khái niệm cơ bản trong lý thuyết tập thô [9] và một số khái niệm khác nhau về tập rút gọn.

Bảng quyết định là một bộ tứ $S = (U, C \cup D, V, f)$ trong đó $U = \{u_1, u_2, ..., u_n\}$ là tập khác rỗng, hữu hạn các đối tượng; $C = \{c_1, c_2, ..., c_m\}$ là tập các thuộc tính điều kiện; D là tập các thuộc tính quyết định với $C \cap D = \emptyset$. $V = \prod_{a \in C \cup D} V_a$ với V_a là tập giá trị của thuộc tính $a \in A$; $f : U \times (C \cup D) \to V$ là hàm thông tin, với $\forall a \in C \cup D, u \in U$ hàm f cho giá trị $f(u, a) \in V_a$.

Mỗi tập con $P \subseteq C \cup D$ xác định một quan hệ không phân biệt được, gọi là quan hệ tương đương:

$$IND(P) = \{(x, y) \in U \times U \,|\forall a \in P, f(x, a) = f(y, a)\},$$

IND(P) xác định một phân hoạch của U, ký hiệu là $U/P = \{P_1, P_2, ..., P_m\}$. Một phần tử $[x]_P = \{y \in U \,|(x, y) \in IND(P)\}$ trong U/P gọi là một lớp tương đương.

Với $B \subseteq C$ và $X \subseteq U$, $B-$xấp xỉ trên của X là tập $\overline{B}X = \{u \in U \,|[u]_B \cap X \neq \emptyset\}$, $B-$xấp xỉ dưới của X là tập $\underline{B}X = \{u \in U \,|[u]_B \subseteq X\}$, $B-$miền biên của X là tập $BN_B(X) = \overline{B}X \backslash \underline{B}X$ và $B-$miền dương của D là tập $POS_B(D) = \bigcup_{X \in U/D} (\underline{B}X)$. Bảng quyết định DT được gọi là nhất quán khi và chỉ khi $POS_C(D) = U$. Ngược lại DT là không nhất quán.

Trong lý thuyết tập thô [8], Pawlak đưa ra khái niệm ban đầu về tập rút gọn của bảng quyết định dựa trên miền dương, còn gọi là tập rút gọn của Pawlak.

Định nghĩa 2.1. [8] Cho bảng quyết định $DT = (U, C \cup d, V, f)$. Nếu $B \subseteq C$ thỏa mãn:

1) $POS_B(\{d\}) = POS_C(\{d\})$,

2) $\forall B' \subset B \left(POS_{B'}(\{d\}) \neq POS_C(\{d\})\right)$.

thì B được gọi là một tập rút gọn của C. Ký hiệu **PRED(C)** là tập tất cả các rút gọn của bảng quyết định DT dựa trên miền dương.

Từ định nghĩa tập rút gọn của Pawlak, có một số công trình nghiên cứu đề xuất các khái niệm khác nhau về tập rút gọn.

Định nghĩa 2.2. [1] Cho bảng quyết định $DT = (U, C, D, V, f)$, $M = (m_{ij})_{n\times n}$ là ma trận phân biệt của Skowron. Nếu $B \subseteq C$ thỏa mãn:

1) $\forall \emptyset \neq m_{ij} \in M, B \cap m_{ij} \neq \emptyset$

2) $\forall b \in B, B - \{b\}$ không thỏa mãn 1)

thì B gọi là một rút gọn của C dựa trên ma trận phân biệt của Skowron. Ký hiệu **SRED(C)** là tập tất cả các rút gọn của bảng quyết định DT dựa trên ma trận phân biệt.

Theo tiếp cận lý thuyết thông tin, G.Y. Wang [3, 4] đưa ra khái niệm tập rút gọn dựa trên Entropy.

Định nghĩa 2.3. [3, 4] Cho bảng quyết định $DS = (U, C \cup D, V, f)$, tập thuộc tính $R \subseteq C$ gọi là một rút gọn của DS dựa trên Entropy của Shannon nếu thỏa mãn hai điều kiện sau:

1) $H(D/C) = H(D/R)$,

2) Với $\forall r \in R, H(D/R) \neq H(D/R - \{r\})$.

Ký hiệu **ERED(C)** là tập tất cả các rút gọn của bảng quyết định DT dựa trên Entropy của Shannon.

Nhằm giảm thiểu độ phức tạp trong công thức tính toán Entropy của Shannon, J.Liang và các cộng sự [5] đề xuất độ đo Entropy mới.

Định nghĩa 2.4. [5] Cho bảng quyết định $DT = (U, C \cup D, V, f)$ với $P \subseteq C$, $U/P = \{X_1, X_2, ..., X_m\}$, $U/D = \{D_1, D_2, ..., D_n\}$. Entropy mới có điều kiện của D đối với P là $H'(D/P)$ được định nghĩa:

$$H'(D/P) = \sum_{i=1}^{m} \left(\frac{|X_i|}{|U|}\right)^2 \sum_{j=1}^{n} \frac{|X_i \cap D_j|}{|X_i|} \left(1 - \frac{|X_i \cap D_j|}{|X_i|}\right).$$

Giống như Entropy Shannon có điều kiện, trong [7] đã chứng minh định lý sau đây về tính phản đơn điệu của Entropy mới có điều kiện.

Định lý 2.1. *[7] Cho bảng quyết định $S = (U, C \cup D, V, f)$, $\forall A_1 \subseteq A_2 \subseteq C$, giả sử $U/D = \{D_1, D_2, ..., D_s\}$, $U/A_1 = \{Y_1, Y_2, ..., Y_n\}$, $U/A_2 = \{X_1, X_2, ..., X_m\}$. Ta có $H'(D/A_1) \geq H'(D/A_2)$. Điều kiện cần và đủ của dấu đẳng thức là $\forall X_i, \forall X_j \in U/A_2, X_i \neq X_j$, nếu $(X_i \cup X_j) \subseteq Y_t \in U/A_1$ thì:*

$|D_k \cap X_i|\,|D_k^c \cap X_j| = 0$ *và* $|D_k \cap X_j|\,|D_k^c \cap X_i| = 0$ *với mọi* $k = 1, 2, ..., s$.

hoặc

$|X_i \cap D_k|\,|X_j - X_j \cap D_k| = 0$ *và* $|X_j \cap D_k|\,|X_i - X_i \cap D_k| = 0$ *với mọi* $k = 1, 2, ..., s$.

Từ độ đo Entropy mới, P.Luo [7] đưa ra khái niệm tập rút gọn dựa trên Entropy mới.

Định nghĩa 2.5. [7] Cho bảng quyết định $DT = (U, C \cup D, V, f)$, một tập thuộc tính $B \subseteq C$ gọi là một rút gọn của DT dựa trên Entropy thông tin mới nếu thỏa mãn hai điều kiện sau:

1) $H^{'}(D/B) = H^{'}(D/C)$,

2) Với $\forall b \in B, H^{'}(D/B - \{b\}) > H^{'}(D/C)$.

Ký hiệu **NERED(C)** là tập tất cả các rút gọn của bảng quyết định DT dựa trên Entropy thông tin mới.

3. MỐI LIÊN HỆ GIỮA CÁC KHÁI NIỆM TẬP RÚT GỌN

Những đóng góp quan trọng trong nghiên cứu mỗi liên hệ giữa các khái niệm tập rút gọn phải kể đến các công trình [6, 7, 13]. Trên bảng quyết định nhất quán, các công trình trên đã chứng minh các khái niệm tập rút gọn là như nhau, nghĩa là $PRED(C) = ERED(C) = SRED(C) = NERED(C)$. Tuy nhiên, với bảng quyết định không nhất quán: trong [6] đã chứng minh nếu $R_S \in SRED(C)$ thì tồn tại $R_E \in ERED(C)$ sao cho $R_E \subseteq R_S$ và tồn tại $R_P \in PRED(C)$ sao cho $R_P \subseteq R_E$, nghĩa là $R_P \subseteq R_E \subseteq R_S$; trong [13] đã chứng minh $NERED(C) = SRED(C)$. Với tập rút gọn dựa trên Entropy của J.Liang, [7] đã chỉ ra nếu $R_{NE} \in NRED(C)$ thì tồn tại $R_E \in ERED(C)$ sao cho $R_E \subseteq R_{NE}$, kết quả này phù hợp với kết quả trong [13] và cho thấy tập rút gọn dựa trên Entropy của J.Liang không phải là tập rút gọn tối thiểu theo định nghĩa của Pawlak trong bảng quyết định không nhất quán. Do đó, các thuật toán tìm tập rút gọn dựa trên Entropy của J.Liang, mặc dù giảm thiểu độ phức tạp tính toán, tuy nhiên không tìm được tập rút gọn tối thiểu theo định nghĩa của Pawlak.

Phần tiếp theo sẽ trình bày phương pháp xây dựng công thức Entropy cải tiến dựa trên Entropy mới và chứng minh tập rút gọn dựa trên Entropy cải tiến tương đương với tập rút gọn của Pawlak.

4. TẬP RÚT GỌN DỰA TRÊN ENTROPY CẢI TIẾN

Cho bảng quyết định $S = (U, C \cup D, V, f)$ với $A_1 \subseteq A_2 \subseteq C$, $U/D = \{D_1, D_2, ..., D_s\}$, $U/A_1 = \{Y_1, Y_2, ..., Y_m\}$, $U/A_2 = \{X_1, X_2, ..., X_n\}$. Ý tưởng thuật toán rút gọn thuộc tính của Pawlak là bảo toàn các đối tượng thuộc miền dương $POS_C(D)$ (các đối tượng nhất quán) khi thực hiện bổ sung hoặc loại bỏ thuộc tính $b \in C$. Dựa vào ý tưởng trên và theo Định lý 2.1, để bảo toàn Entropy mới $H^{'}(D/A_1) = H^{'}(D/A_2)$ với $\forall X_i, \forall X_j \in U/A_2, X_i \neq X_j$, nếu $(X_i \cup X_j) \subseteq Y_t \in U/A_1$ thì $|X_i \cap D_k||X_j - X_j \cap D_k| = 0$ và $|X_j \cap D_k||X_i - X_i \cap D_k| = 0$ với mọi $k = 1, 2, ..., s$, nghĩa là $X_i, X_j \subseteq D_k$. Như vậy các D_k phải là hợp của các lớp nhất quán hoặc các lớp không nhất quán X_i, chứ D_k không phải thuộc phân hoạch U/D như hiện nay.

Từ ý tưởng trên, với $DS = (U, C, D, V, f)$ không nhất quán, gọi $D_0 = U - POS_C(D)$ là tập các đối tượng không nhất quán, hiển nhiên $\underline{C}(D_0) = D_0$. Thay vì sử dụng phân hoạch $U/D = \{D_1, D_2, ..., D_m\}$, ta sử dụng phân hoạch mới, ký hiệu là U/R_d như sau: $U/R_D = \{\underline{C}D_0, \underline{C}D_1, \underline{C}D_2, ..., \underline{C}D_m\}$. Phân hoạch mới này tách các đối tượng không nhất quán vào lớp D_0, còn lại là các lớp nhất quán $\underline{C}(D_i)$ với $1 < i \leq m$. Dựa trên phân hoạch mới này, tương tự Định nghĩa 2.4, ta định nghĩa Entropy có điều kiện cải tiến, ký hiệu là $H^{'}(R_D/P)$ như sau.

$$H'(R_D/P) = \sum_{i=1}^{m} \left(\frac{|X_i|}{|U|}\right)^2 \sum_{j=0}^{n} \frac{|X_i \cap \underline{C} D_j|}{|X_i|} \left(1 - \frac{|X_i \cap \underline{C} D_j|}{|X_i|}\right).$$

Khi đó, tập rút gọn dựa trên Entropy cải tiến được định nghĩa như sau.

Định nghĩa 4.1. Cho bảng quyết định $DS = (U, C \cup D, V, f)$ và tập thuộc tính $B \subseteq C$, nếu thuộc tính $b \in B$ thỏa mãn $H'(R_D/B - \{b\}) = H'(R_D/B)$ thì b gọi là dư thừa (có thể rút gọn) trong B đối với D dựa trên Entropy cải tiến.

Định nghĩa 4.2. Cho bảng quyết định $DS = (U, C \cup D, V, f)$, một tập thuộc tính $B \subseteq C$ gọi là một rút gọn của DS dựa trên Entropy cải tiến nếu thỏa mãn hai điều kiện sau:

1) $H'(R_D/B) = H'(R_D/C)$,

2) Với $\forall b \in B, H'(R_D/B - \{b\}) > H'(R_D/C)$.

Ký hiệu **INERED(C)** là tập tất cả các rút gọn của bảng quyết định DS dựa trên Entropy cải tiến.

Tiếp theo, ta sẽ chứng minh tập rút gọn dựa trên Entropy cải tiến và tập rút gọn của Pawlak là như nhau.

Định lý 4.1. *Cho bảng quyết định $S = (U, C, D, V, f)$, $B \subseteq C$. Nếu $H'(R_D/B) = H'(R_D/C)$ thì $POS_B(D) = POS_C(D)$.*

Chứng minh. Giả sử $POS_B(D) \neq POS_C(D)$, khi đó chắc chắn $\exists Y \in U/B$ sao cho $Y \notin POS_B(D)$ và $Y \in POS_C(D)$. Từ đó suy ra chắc chắn $\exists x_{k_0}, x_{k_1} \in Y, x_{k_0} \neq x_{k_1}$ sao cho $f(x_{k_0}, D) \neq f(x_{k_1}, D)$. Đặt $X_{i_0} = [x_{k_0}]_B, X_{j_0} = [x_{k_1}]_B$, do $Y \in POS_C(D)$ và $f(x_{k_0}, D) \neq f(x_{k_1}, D)$ nên x_{k_0}, x_{k_1} phải thuộc hai lớp tương đương khác nhau trong phân hoạch U/C, nghĩa là $X_{i_0} \neq X_{j_0}$ và $X_{i_0}, X_{j_0} \subseteq POS_C(D)$ (1). Hơn nữa, theo tính chất của lớp tương đương ta có $[x_{k_0}]_C \subseteq [x_{k_0}]_B$ và $[x_{k_1}]_C \subseteq [x_{k_1}]_B$ mà $Y \in U/B$ nên $[x_{k_0}]_C \subseteq Y, [x_{k_1}]_C \subseteq Y$ từ đó suy ra $X_{i_0} \cup X_{j_0} \subseteq Y \in U/B$ (2).

Mặt khác, từ (1) $X_{i_0}, X_{j_0} \subseteq POS_C(D)$ nên tồn tại $D_k \in U/R_D$ với $D_k \neq D_0$ sao cho: (a) $X_{i_0} \cap D_k = X_{i_0}$, $X_{j_0} \cap D_k = \emptyset$ hoặc (b) $X_{i_0} \cap D_k = \emptyset$, $X_{j_0} \cap D_k = X_{j_0}$. Cả hai trường hợp (a), (b) ta đều có: $|X_{i_0} \cap D_k|\,|X_{j_0} - X_{j_0} \cap D_k| \neq 0$ và $|X_{j_0} \cap D_k|\,|X_{i_0} - X_{i_0} \cap D_k| \neq 0$ (3).

Từ (1), (2), (3) theo Định lý 2.1 ta có $H'(R_D/B) > H'(R_D/C)$, điều này mâu thuẫn với giả thiết. Vậy giả sử là sai và định lý được chứng minh.

■

Định lý 4.2. *Cho bảng quyết định $S = (U, C, D, V, f)$, $B \subseteq C$. Nếu $POS_B(D) = POS_C(D)$ thì $H'(R_D/B) = H'(R_D/C)$.*

Chứng minh. Từ $B \subseteq C$, theo Định lý 2.1 ta có $H'(D/B) \geq H'(D/C)$.

Giả sử $H'(D/B) \neq H'(D/C)$, khi đó $C - B \neq \emptyset$. Mặt khác, từ Định lý 2.1 suy ra tồn tại ít nhất $X_{i_0}, X_{j_0} \in U/C, X_{i_0} \neq X_{j_0}$ và $(X_{i_0} \cup X_{j_0}) \subseteq Y_0 \in U/B$ sao cho biểu

thức logic $|X_i \cap D_k|\,|X_j - X_j \cap D_k| = 0$ và $|X_j \cap D_k|\,|X_i - X_i \cap D_k| = 0$ là sai với mọi $k = 0, 1, 2, ..., s$, từ đó suy ra $X_{i_0}, X_{j_0} \in D_k$ với $1 \leq k \leq s$ và $X_{i_0}, X_{j_0} \notin D_0$ (1) (nghĩa là X_{i_0}, X_{j_0} là các lớp nhất quán), vì nếu $X_{i_0}, X_{j_0} \in D_0$ thì $|X_i \cap D_k|\,|X_j - X_j \cap D_k| = 0$ và $|X_j \cap D_k|\,|X_i - X_i \cap D_k| = 0$ với mọi $k = 0, 1, 2, ..., s$.

Khi $|X_i \cap D_k|\,|X_j - X_j \cap D_k| = 0$ sai với mọi $k = 1, 2, ..., s$, nghĩa là tồn tại ít nhất $D_p \in U/R_D$ để $|X_{i_o} \cap D_p|\,|X_{j_o} - X_{j_o} \cap D_p| \neq 0$. Điều này suy ra chắc chắn tồn tại $x_{k_0} \in X_{i_0}, x_{k_1} \in X_{j_0}$ để $f(x_{k_0}, D) \neq f(x_{k_1}, D)$ (2) (bởi vì với $\forall x \in X_{i_0}, \forall y \in X_{j_0}$ nếu $f(x, D) = f(y, D)$ thì $(X_{i_0} \cup X_{j_0}) \subseteq D_p \in U/D$ suy ra $|X_{i_o} \cap D_p|\,|X_{j_o} - X_{j_o} \cap D_p| = 0$) và $f(x_{k_0}, C) \neq f(x_{k_1}, C)$ (3).

Mặt khác, từ giả thiết $(X_{i_0} \cup X_{j_0}) \subseteq Y_0 \in U/B$, $X_{i_0} \neq X_{j_0}$ nên với $\forall b \in B$ ta có $f(x_{k_0}, b) = f(x_{k_1}, b)$ (4). Từ (3),(4) suy ra $X_{i_0} \cup X_{j_0} \not\subset POS_B(D)$ (5).

Cũng từ (1) ta có $X_{i_0}, X_{j_0} \in D_k$ suy ra $X_{i_0} \cup X_{j_0} \subseteq POS_C(D)$, kết hợp với (5) ta thu được $POS_B(D) \neq POS_C(D)$, điều này mâu thuẫn với giả thiết, vậy giả sử là sai và ta có điều phải chứng minh.

■

Từ Định lý 4.1 và Định lý 4.2 ta thu được hệ quả sau.

Hệ quả 4.1. *Cho bảng quyết định $S = (U, C, D, V, f)$, với $B \subseteq C$, $b \in B$ ta có:*

(1) *Nếu $H'(R_D/B - \{b\}) \neq H'(R_D/B)$ thì $POS_{B-\{b\}}(D) \neq POS_B(D)$.*

(2) *Nếu $POS_{B-\{b\}}(D) \neq POS_B(D)$ thì $H'(R_D/B - \{b\}) \neq H'(R_D/B)$.*

Dễ dàng chứng minh (1) dựa vào cách chứng minh Định lý 4.2. Tương tự chứng minh (2) dựa vào cách chứng minh Định lý 4.1.

Từ các kết quả trên ta có định lý về mối quan hệ giữa **PRED(C)** và **INERED(C)**.

Định lý 4.3. *Cho bảng quyết định $S = (U, C, D, V, f)$, ta luôn có **INERED(C) = PRED(C)**.*

5. THUẬT TOÁN RÚT GỌN THUỘC TÍNH DỰA TRÊN ENTROPY CẢI TIẾN

Để xây dựng thuật toán rút gọn thuộc tính xuất phát từ tập nhân theo hướng tiếp cận heuristic, chúng tôi đưa ra một số định nghĩa về độ quan trọng của thuộc tính dựa trên Entropy cải tiến.

Định nghĩa 5.1. Cho bảng quyết định $DT = (U, C \cup D, V, f)$ và tập thuộc tính $B \subseteq C$. Độ quan trọng của thuộc tính $b \in B$ đối với D theo tiếp cận Entropy cải tiến, ký hiệu là $ISGF(b, B, D)$, được xác định như sau:

(1) $ISGF(b, B, D) = H'(R_D/B - \{b\}) - H'(R_D/B)$

Tương tự, độ quan trọng của $b \in C - B$ đối với D theo tiếp cận Entropy cải tiến như sau:

(2) $ISGF(b, B, D) = H'(R_D/B) - H'(R_D/B \cup \{b\})$

Dễ thấy $ISGF(b,B,D) \geq 0$, $ISGF(b,B,D) = 0$ khi b là dư thừa trong B. Nếu thêm b vào B làm cho số đối tượng nhất quán càng ít, nghĩa là $H'(R_D/B \cup \{b\})$ càng nhỏ hay $SGF(b,B,D)$ càng lớn thì b càng quan trọng và ngược lại.

Thuật toán 5.1. Tìm tập nhân dựa trên Entropy cải tiến

Đầu vào: $DT = (U, C \cup D, V, f)$.

Đầu ra: Tập nhân Core.

1. $CORE := \emptyset$;
2. Tính phân hoạch U/R_D;
3. Tính $H'(R_D/C)$;
4. For $a \in C$ do

   4.1. Tính $H'(R_D/C - \{a\})$;

   4.2. If $H'(R_D/C) \neq H'(R_D/C - \{a\})$ then $CORE := CORE \cup \{a\}$;

Sử dụng thuật toán cải tiến trong [12] để tính U/C, độ phức tạp thời gian là $O(|C||U|)$. Do đó, độ phức tạp để tính $H'(R_D/C)$ là $O(|C||U|)$. Vì vậy, độ phức tạp thời gian của thuật toán tìm tập nhân là $O(|C|^2|U|)$.

Thuật toán 5.2. Tìm tập rút gọn dựa trên Entropy cải tiến

Đầu vào: Bảng quyết định $DT = (U, C \cup D, V, f)$.

Đầu ra: Tập rút gọn R.

1. Tìm Core theo Thuật toán 5.1;
2. R:= Core;
3. Tính $H'(R_D/R)$;
4. While $H'(R_D/R) \neq H'(R_D/C)$ do

   4.1. Tìm $ISGF(c_k,R,D) = \max\{ISGF(c_j,R,D)/c_j \in C\backslash R\}$;

   4.2. $R := R \cup \{c_k\}$;

   4.3. Tính $H'(R_D/R)$;

5. Return R;

Như vậy, xuất phát từ R = Core, thuật toán thực hiện tối đa k vòng lặp để bổ sung $R = R \cup \{c_k\}$. Trong mỗi vòng lặp này, ta phải tính lại tất cả các $ISGF(c_j,R,D)$ và chọn ra thuộc tính có độ quan trọng lớn nhất. Độ phức tạp để tính Entropy có điều kiện cải tiến phụ thuộc vào độ phức tạp tính phân hoạch U/R. Sử dụng thuật toán cải tiến tính U/R trong [12], độ phức tạp để tính $H'(R_D/R)$ là $O(|C||U|)$ và độ phức tạp của thuật toán là $O\left(|C|^2|U|\right)$.

Bảng 6.1. Bảng quyết định không nhất quán

U	c_1	c_2	c_3	c_4	c_5	d
u_1	1	1	0	1	0	1
u_2	0	0	1	0	0	1
u_3	1	0	0	0	1	1
u_4	0	0	1	0	0	0
u_5	1	1	1	1	0	1
u_6	1	0	0	0	0	0
u_7	0	0	1	0	1	0
u_8	1	0	0	0	0	1
u_9	1	0	0	0	1	0
u_10	0	0	0	1	0	0

6. MINH HỌA THUẬT TOÁN

Ví dụ 6.1. Cho bảng quyết định không nhất quán $DT = (U, C \cup D, V, f)$ với $U = \{u_1, u_2, ..., u_{10}\}$, $C = \{c_1, c_2, c_3, c_4, c_5\}$ và $D = d$ như Bảng 6.1.

Ta có:

$U/\{d\} = \{\{u_1, u_2, u_3, u_5, u_8\}, \{u_4, u_6, u_7, u_9, u_{10}\}\}$,

$U/C = \{\{u_1\}, \{u_5\}, \{u_7\}, \{u_{10}\}, \{u_2, u_4\}, \{u_3, u_9\}, \{u_6, u_8\}\}$, do đó:

$U/R_D = \{\{u_2, u_3, u_4, u_6, u_8, u_9\}, \{u_1, u_5\}, \{u_7, u_{10}\}\}$ và $H'(R_D/C) = 0$.

Áp dụng Thuật toán 5.1 tìm tập nhân, ta có:

$H'(R_D/C - \{c_1\}) = H'(R_D/\{c_2, c_3, c_4, c_5\}) = 0$ vậy thuộc tính $\{c_1\}$ dư thừa.

$H'(R_D/C - \{c_2\}) = H'(R_D/\{c_1, c_3, c_4, c_5\}) = 0$ vậy thuộc tính $\{c_2\}$ dư thừa.

$H'(R_D/C - \{c_3\}) = H'(R_D/\{c_1, c_2, c_4, c_5\}) = 0$ vậy thuộc tính $\{c_3\}$ dư thừa.

$H'(R_D/C - \{c_4\}) = H'(R_D/\{c_1, c_2, c_3, c_5\}) = 0$ vậy thuộc tính $\{c_4\}$ dư thừa.

$H'(R_D/C - \{c_5\}) = H'(R_D/\{c_1, c_2, c_3, c_4\}) <> 0$, vậy $CORE = \{c_5\}$.

Áp dụng Thuật toán 5.2 tìm tập rút gọn, ta có:

Ban đầu: $R = CORE = \{c_5\}$, $H'(R_D/\{c_5\}) = 8/25 = 0.32$.

Tính $ISGF(c_1, \{c_5\}, D) = H'(R_D/\{c_5\}) - H'(R_D/\{c_1, c_5\}) = 8/25$ - $3/25 = 5/25 = 0.2$.

Tính $ISGF(c_2, \{c_5\}, D) = H'(R_D/\{c_5\}) - H'(R_D/\{c_2, c_5\}) = 8/25$ - $3/25 = 5/25 = 0.2$.

Tính $ISGF(c_3, \{c_5\}, D) = H'(R_D/\{c_5\}) - H'(R_D/\{c_3, c_5\}) = 8/25$ - $14/100 = 0.18$.

Tính $ISGF(c_4, \{c_5\}, D) = H'(R_D/\{c_5\}) - H'(R_D/\{c_4, c_5\}) = 8/25$ - $2/25 = 0.24$.

Bảng 6.2. Độ phức tạp thời gian của các thuật toán

Thuật toán	Tập rút gọn	Độ phức tạp thời gian
A1	$\{c_1, c_4, c_5\}$	$O\left(\|C\|^2 \|U\| \log \|U\|\right)$
A2	$\{c_1, c_2, c_3, c_5\}$	$O\left(\|C\| \|U\|^2\right) + O\left(\|U\|^3\right)$
A3	$\{c_1, c_4, c_5\}$	$O\left(\|C\|^2 \|U\|\right)$

Vậy $R = R \cup \{c_4\} = \{c_4, c_5\}$ với $H'(R_D/\{c_4, c_5\}) = 2/25 = 0.08$.

Tương tự như trên, tính $ISGF(c_1, \{c_4, c_5\}, D) = H'(R_D/\{c_4, c_5\}) - H'(R_D/\{c_1, c_4, c_5\}) = 0.08 - 0 = 0.08$,

$ISGF(c_2, \{c_4, c_5\}, D)$ =2/25 - 1/25 = 1/25 = 0.004,

$ISGF(c_3, \{c_4, c_5\}, D) = 2/25$ - 2/100 = 0.06.

Như vậy $R = R \cup \{c_1\} = \{c_1, c_4, c_5\}$ và $H'(R_D/\{c_1, c_4, c_5\}) = H'(R_D/C)$. Vậy thuật toán dừng. Do đó tập rút gọn của Pawlak trên bảng quyết định DT là $R = \{c_1, c_4, c_5\}$.

Nhận xét 6.1. Gọi Thuật toán trong [15] tìm tập rút gọn dựa trên miền dương là A1, Thuật toán CEBARKCC trong [14] tìm tập rút gọn theo Entropy của Shannon là A2 và Thuật toán được đề xuất là A3. Với bảng quyết định ở ví dụ trên, kết quả tìm tập rút gọn được mô tả trong Bảng 6.2.

Như vậy, thuật toán đề xuất tìm được tập rút gọn của Pawlak và hiệu quả hơn so với các thuật toán theo Entropy của Shannon.

7. KẾT LUẬN

Bài báo đã đề xuất phương pháp xây dựng Entropy cải tiến dựa trên Entropy mới của J.Liang [5] và chứng minh tập rút gọn dựa trên Entropy cải tiến tương đương với tập rút gọn của Pawlak trong bảng quyết định không nhất quán. Đây là kết quả quan trọng làm cơ sở để xây dựng thuật toán tìm tập rút gọn của Pawlak trong bảng quyết định không nhất quán dựa trên Entropy cải tiến. Thuật toán được xây dựng có độ phức tạp $O\left(|C|^2 |U|\right)$.

TÀI LIỆU THAM KHẢO

[1] Andrzej Skowron, Rauszer C., 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).

[2] D.Y. Ye, Z.J. Chen, *Inconsistency classification and discernibility matrix based approaches for computing an attribute*, Core, G. Wang et al. (Eds.): RSFDGrC 2003, LNAI 2639, Springer-Verlag, Berlin Heidelberg, 2003 (269-273).

[3] G.Y. Wang, Algebra view and information view of rough sets theory, Dasarathy BV,editor. Data mining and knowledge discovery: Theory, tools, and technology III, *Proceedings of SPIE*, 2001 (200–207).

[4] DG.Y.Wang, Rough reduction in algebra view and information view, *International Journal of Intelligent System* **18** (2003) 679–688.

[5] J.Y. Liang, K.S. Chin, C.Y. Dang, C.M. YAM. Richard, New method for measuring uncertainty and fuzziness in rough set theory, *International Journal of General Systems* **31** (2002) 331–342.

[6] Nguyễn Long Giang, Nguyễn Thanh Tùng, Nghiên cứu mối liên hệ giữa ba khái niệm tập rút gọn trong lý thuyết tập thô, *Kỷ yếu Một số vấn đề chọn lọc của CNTT và Truyền thông*, Đồng Nai, 2009 (282–293).

[7] P. Luo, Q. He, Z.Z. Shi, Theoretical study on a new information entropy and its use in attribute reduction, *ICCI* (2005) 73–79.

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

[9] T.R. Li, K.R. Qing, N. Yang, Y. Xu, *Study on reduct and core computation in incompatible information systems*, S. Tsumoto et al. (Eds.): RSCTC 2004, LNAI 3066, Springer-Verlag, Berlin Heidelberg, 2004 (471–476).

[10] X.H. Hu, N. Cercone, Learning in relational databases: a rough set approach, *International Journal of Computational Intelligence* (1995) 323–338.

[11] Xiaohua, Jianchao Han, T.Y.Lin, A new rough sets model based on database systems, *Fundamenta Informatica* **20** (2004) 1–18.

[12] Yue-jin Lv, Jin-hai Li, A Quick Algorithmfor Reduction of Attribute in Information Systems, *The First International Symposium on Data, Privacy, and E-Commerce (ISDPE 2007)*, 2007 (98–100).

[13] Zhangyan Xu, Wenbin Qian, Liyu Huang, Bingru Yang, Comparative Comparative Research of Attribute Reduction based on the New Information Entropy and on Skowron's Discernibility Matrix, *International Symposium on Computation Intelligence and Design*, 2008 (129–132).

[14] G.Y.Wang, H.Yu, D.C.Yang, Decision table reduction based on conditional information entropy, *Journal of Computers* **25** (7) (2002) 759–766.

[15] S. Liu, Q. Sheng, B. Wu, Z. Shi, F. Hu, Research on efficient algorithms for rough set methods, *Chinese Journal of Computers* **25** (5) (2003) 524–529.

Ngày nhận bài 8 - 6 - 2011