RÚT GỌN THUỘC TÍNH CỦA BẢNG QUYẾT ĐỊNH SỬ DỤNG MIỀN DƯƠNG MỜ

Cao Chính Nghĩa[1], **Vũ Đức Thi**[2], **Tân Hạnh**[3], **Nguyễn Long Giang**[4]
[1]Học viện Cảnh sát nhân dân, Bộ Công an
[2]Đại học Sư phạm Kỹ thuật Hưng Yên
[3]Học viện Công nghệ Bưu chính Viễn thông
[4]Viện Công nghệ thông tin

Tóm tắt: Các phương pháp rút gọn thuộc tính theo tiếp cận tập thô truyền thống khi áp dụng trên các bảng quyết định có miền giá trị số thực cần phải rời rạc hóa dữ liệu. Việc rời rạc hóa dữ liệu dẫn đến mất mát thông tin làm ảnh hưởng đến độ chính xác phân lớp dữ liệu. Các phương pháp rút gọn thuộc tính trực tiếp trên bảng quyết định có miền giá trị số thực theo tiếp cận tập thô mờ khắc phục được hạn chế trên. Trong bài báo này, chúng tôi đề xuất hai phương pháp rút gọn thuộc tính sử dụng miền dương mờ dựa trên phân hoạch mờ và quan hệ tương tự mờ**.** Phân tích đánh giá từng phương pháp, kết luận phương pháp sử dụng quan hệ tương tự mờ có khả năng áp dụng thực tế.

Từ khóa: tập thô mờ, bảng quyết định mờ, phân hoạch mờ, quan hệ tương tự mờ, miền dương mờ, rút gọn thuộc tính, tập rút gọn.

I. MỞ ĐẦU

Rút gọn thuộc tính là bài toán quan trọng trong bước tiền xử lý số liệu với mục tiêu là giảm số chiều dữ liệu (số thuộc tính) nhằm tăng tính hiệu quả của các thuật toán khai phá dữ liệu. Các phương pháp rút gọn thuộc tính theo tiếp cận lý thuyết tập thô truyền thống của Pawlak được thực hiện trên các bảng quyết định có miền giá trị rời rạc [6]. Đối với các bảng quyết định có miền giá trị số thực, cần phải rời rạc hóa dữ liệu trước khi áp dụng các phương pháp rút gọn thuộc tính theo tiếp cận tập thô truyền thống dẫn đến mất mát thông tin, làm hạn chế độ chính xác phân lớp của dữ liệu. Để khắc phục vấn đề này, D. Dubois và các cộng sự đề xuất mô hình tập thô mờ (fuzzy rough set) kết hợp giữa lý thuyết tập thô và lý thuyết tập mờ [1]. Lý thuyết tập mờ đóng vai trò bảo toàn ngữ nghĩa của dữ liệu, còn lý thuyết tập thô bảo toàn tính không phân biệt được của dữ liệu.

Các công trình nghiên cứu về rút gọn thuộc tính theo tiếp cận tập thô mờ đang được phát triển [2, 5, 7, 9, 10] và cần nhiều hơn nữa sự đóng góp kết quả của cộng đồng nghiên cứu.

Trong bài báo này, chúng tôi đề xuất hai phương pháp rút gọn thuộc tính điển hình sử dụng miền dương mờ theo tiếp cận tập thô mờ dựa trên phân hoạch mờ và quan hệ tương tự mờ. Phương pháp dựa trên phân hoạch mờ áp dụng cho lớp bài toán rút gọn thuộc tính của bảng quyết định mờ. Phương pháp sử dụng quan hệ tương tự mờ áp dụng cho lớp bài toán rút gọn thuộc tính của bảng quyết định có miền giá trị số thực. Dựa trên phân tích đánh giá từng phương pháp đi đến kết luận là các phương pháp rút gọn thuộc tính của bảng quyết định dựa trên quan hệ tương tự mờ là có khả năng áp dụng thực tế và được cộng đồng quan tâm nghiên cứu sôi động lâu nay. Dưới đây trình bày một số khái niệm cơ bản về tập thô mờ; các phương pháp rút gọn thuộc tính của bảng quyết định sử dụng miền dương mờ và kết quả thực nghiệm; kết luận và hướng phát triển tiếp theo.

Tác giả liên hệ: Cao Chính Nghĩa
Email: ccnghia@gmail.com
Đến tòa soạn: 23/7/2016, chỉnh sửa: 30/8/2016, chấp nhận đăng: 03/9/2016.

II. MỘT SỐ KHÁI NIỆM CƠ BẢN VỀ TẬP THÔ MỜ

Hệ thông tin là một cặp $IS = (U, A)$ trong đó U là tập hữu hạn, khác rỗng các đối tượng; A là tập khác rỗng, hữu hạn các thuộc tính. Mỗi thuộc tính $a \in A$ xác định một ánh xạ: $a : U \to V_a$ với V_a là tập giá trị của thuộc tính $a \in A$. Bảng quyết định là dạng đặc biệt của hệ thông tin trong đó tập các thuộc tính A bao gồm hai tập con tách biệt nhau: tập các thuộc tính điều kiện C và tập các thuộc tính quyết định D, ký hiệu là $DS = (U, C \cup D)$ với $C \cap D = \varnothing$. Bảng quyết định mờ là bảng quyết định mà các giá trị của thuộc tính là các tập mờ.

A. Quan hệ tương tự mờ

Cho hệ thông tin là một cặp *IS = (U, A)*. Một quan hệ *R* xác định trên *U* được gọi là quan hệ tương tự mờ (fuzzy similarity relation) nếu thỏa mãn các điều kiện sau đây [5].

1) Tính phản xạ: $R(x,x) = 1, \forall x \in U$

2) Tính đối xứng: $R(x,y) = R(y,x), \forall x, y \in U$

3) Tính bắc cầu max - min:

$R(x,z) \geq \min\{R(x,y), R(y,z)\}$ với mọi $x, y, z \in U$.

B. Phân hoạch mờ

Cho bảng quyết định $DS = (U, C \cup D)$, mỗi tập thuộc tính $P \subseteq C$ xác định một quan hệ tương tự (tương đương) mờ. Tương tự trong lý thuyết tập thô truyền thống, dựa trên quan hệ tương tự mờ, mỗi tập thuộc tính $P \subseteq C$ xác định một phân hoạch mờ như sau:

với $U / P = \otimes\{a \in P : U / IND(\{a\})\}$ (1)

$$A \otimes B = \{X \cap Y : \forall X \in A, \forall Y \in B, X \cap Y \neq \varnothing\}.$$

Mỗi phần tử thuộc *U/P* là một lớp tương đương mờ (fuzzy equivalence class) với $\mu_{[x]_P}(y) = \mu_P(x,y)$.

Ví dụ, nếu $P = \{a, b\}$, $U / IND(\{a\}) = \{N_a, Z_a\}$ và $U/IND(\{b\}) = \{N_b, Z_b\}$, khi đó

$$U / P = \{N_a \cap N_b, N_a \cap Z_b, Z_a \cap N_b, Z_a \cap Z_b\}.$$

Hàm thành viên của các đối tượng trong lớp tương đương mờ được xác định dựa trên lý thuyết tập mờ [4].

$$\mu_{F_1 \cap ... \cap F_n}(x) = min\left(\mu_{F_1}(x), \mu_{F_2}(x), ..., \mu_{F_n}(x)\right) \quad (2)$$

C. Tập thô mờ định nghĩa theo phân hoạch mờ

Dựa vào các lớp tương đương mờ, khái niệm tập xấp xỉ dưới và xấp xỉ trên được mở rộng thành tập xấp xỉ dưới mờ (fuzzy lower approximation) và xấp xỉ trên mờ (fuzzy upper approximation). Với tập thuộc tính $P \subseteq C$, hàm thành viên của các đối tượng thuộc tập xấp xỉ dưới mờ và tập xấp xỉ trên mờ được xác định [1,7].

$$\mu_{\underline{PX}}(x) = \sup_{F \in U/P} min\left(\mu_F(x), \inf_{y \in U} max\{1 - \mu_F(y), \mu_X(y)\}\right) \quad (3)$$

$$\mu_{\overline{PX}}(x) = \sup_{F \in U/P} min\left(\mu_F(x), \sup_{y \in U} min\{\mu_F(y), \mu_X(y)\}\right) \quad (4)$$

với ký hiệu inf *X*, sup *X* tương ứng là cận dưới đúng và cận trên đúng của tập hợp *X*. *F* là các lớp tương đương mờ của phân hoạch mờ U / P. Bộ $\left\langle \underline{PX}, \overline{PX} \right\rangle$ được gọi là một tập thô mờ.

III. RÚT GỌN THUỘC TÍNH CỦA BẢNG QUYẾT ĐỊNH SỬ DỤNG MIỀN DƯƠNG MỜ

A. Rút gọn thuộc tính của bảng quyết định sử dụng miền dương mờ dựa trên phân hoạch mờ

Phương pháp này được áp dụng cho lớp bài toán có bảng quyết định mờ, giá trị hàm thuộc của các đối tượng nằm trong khoảng [0,1]. Trong lý thuyết tập thô truyền thống, khái niệm miền dương được định nghĩa là giao của tất cả các tập xấp xỉ dưới. Với $P, Q \subseteq A$, hàm thành viên của đối tượng thuộc miền dương mờ trong mô hình tập thô mờ được xác định [7].

$$\mu_{POS_P(Q)}(x) = \sup_{X \in U/Q} \mu_{\underline{PX}}(x) \quad (5)$$

Lực lượng của miền dương mờ được tính theo công thức [7].

$$\left|\mu_{POS_P(Q)}(x)\right| = \sum_{x\in U} \mu_{POS_P(Q)}(x) \quad (6)$$

Phương pháp heuristic tìm một tập rút gọn nhỏ nhất của bảng quyết định mờ bao gồm các bước: Định nghĩa tập rút gọn, định nghĩa độ quan trọng của thuộc tính và xây dựng thuật toán heuristic tìm một tập rút gọn nhỏ nhất dựa trên độ quan trọng của thuộc tính.

Định nghĩa 1. Cho bảng quyết định $DS = (U, C \cup D)$ và tập thuộc tính $P \subseteq C$. Nếu

$$\begin{aligned} &1)\ \left|\mu_{\mathrm{POS}_P(D)}(x)\right| = \left|\mu_{\mathrm{POS}_C(D)}(x)\right| \\ &2)\ \forall p \in P, \left|\mu_{\mathrm{POS}_{P-\{p\}}(D)}(x)\right| \neq \left|\mu_{\mathrm{POS}_C(D)}(x)\right| \end{aligned} \quad (7)$$

thì P là một tập rút gọn của C dựa trên miền dương mờ.

Định nghĩa 2. Cho bảng quyết định $DS = (U, C \cup D)$, $B \subset C$ và $b \in C - B$. Độ quan trọng của thuộc tính b đối với tập thuộc tính B được định nghĩa:

$$SIG_B(b) = \left|\mu_{\mathrm{POS}_{B\cup\{b\}}(D)}(x)\right| - \left|\mu_{\mathrm{POS}_B(D)}(x)\right| \quad (8)$$

Thuật toán tìm một tập rút gọn nhỏ nhất của bảng quyết định sử dụng miền dương mờ ở công thức (6) dựa trên phân hoạch mờ được mô tả như sau:

Thuật toán F_RSAR 1 *(Fuzzy Rough Set based Attribute Reduction)*.

Đầu vào: Bảng quyết định mờ $DS = (U, C \cup D)$

Đầu ra: Một tập rút gọn nhỏ nhất P.

1. $P \leftarrow \varnothing$; $|\mu_{\mathrm{POS}_\varnothing(D)}(x)| = 0$;

2. Tính $\left|\mu_{\mathrm{POS}_C(D)}(x)\right|$;

3. While $\left|\mu_{\mathrm{POS}_P(D)}(x)\right| \neq \left|\mu_{\mathrm{POS}_C(D)}(x)\right|$ Do

4. Begin

5. For $c \in C - P$ tính

$SIG_P(c) = \left|\mu_{\mathrm{POS}_{P\cup\{c\}}(D)}(x)\right| - \left|\mu_{\mathrm{POS}_P(D)}(x)\right|$;

6. Chọn $c_m \in C - P$ sao cho $SIG_P(c_m) = \underset{c\in C-P}{Max}\{SIG_P(c)\}$;

7. $P \leftarrow P \cup \{c_m\}$;

8. End;

//Loại bỏ các thuộc tính dư thừa trong P nếu có

9. For each $a \in P$

10. Begin

11. Tính $\left|\mu_{POS_{P-\{a\}}(D)}(x)\right|$;

12. If $\left|\mu_{\mathrm{POS}_{P-\{a\}}(D)}(x)\right| = \left|\mu_{\mathrm{POS}_C(D)}(x)\right|$ then $P = P - \{a\}$;

13. End;

14. Return P;

Ví dụ 1. Cho bảng quyết định $DS = (C \cup D)$ với $C = \{a, b, c\}$, $D = \{d\}$ trong công trình [7] được mô tả ở Bảng 1. Giá trị các thuộc tính a, b, c được biểu diễn bởi hai tập mờ N và Z tương ứng là (Na, Za), (Nb, Zb), (Nc, Zc) [7].

Bảng I. Bảng quyết định mô tả Ví dụ 1

Đối tượng	a		b		c		D
	Na	Za	Nb	Zb	Nc	Zc	
	c_1	c_2	c_3	c_4	c_5	c_6	
1	0,8	0,2	0,6	0,4	1	0	No
2	0,8	0,2	0	0,6	0,2	0,8	Yes
3	0,6	0,4	0,8	0,2	0,6	0,4	No
4	0	0,4	0,6	0,4	0	1	Yes
5	0	0,6	0,6	0,4	0	1	Yes
6	0	0,6	0	1	0	1	No

Áp dụng các bước của Thuật toán F_RSAR 1 tìm một tập rút gọn của bảng quyết định, đầu tiên tính $U/D = \{\{1,3,6\},\{2,4,5\}\}$, $\left|\mu_{\mathrm{POS}_C(D)}(x)\right| = 3.4$, tiếp theo tính các tập xấp xỉ dưới đối với các thuộc tính a, b và c. Xét thuộc tính a, với lớp tương đương $X = \{1,3,6\}$, $\mu_{\underline{a}\{1,3,6\}}(x)$ được tính:

$$\mu_{\underline{a}\{1,3,6\}}(x) = \sup_{F\in U/a} min\left(\mu_F(x), \inf_{y\in U} max\left\{1-\mu_F(y), \mu_{\{1,3,6\}}(y)\right\}\right)$$

Xét lớp tương đương mờ N_a trên thuộc tính a:

$$min\left(\mu_{N_a}(x), \inf_{y\in U} max\left\{1-\mu_{N_a}(y), \mu_{\{1,3,6\}}(y)\right\}\right)$$

Đối tượng 1 được tính:

$$min\left(0.8, \inf\left\{1, 0.2, 1, 0.4, 1, 1\right\}\right) = 0.2$$

Vì vậy $\mu_{\underline{Na}\{1,3,6\}}(1) = 0.2$. Tương tự với $X = \{2,4,5\}$, tính được $\mu_{\underline{Na}\{2,4,5\}}(1) = 0.2$. Theo công thức (5) tính được tương tự với $\mu_{POS_{Na}(D)}(1) = 0.2$, tương tự $\mu_{POS_{Za}(D)}(1) = 0.2$, vậy $\mu_{POS_a(D)}(1) = 0.2$. Tương tự ta có: $\mu_{POS_a(D)}(2) = 0.2$, $\mu_{POS_a(D)}(3) = 0.4$, $\mu_{POS_a(D)}(4) = 0.4$, $\mu_{POS_a(D)}(5) = 0.4$, $\mu_{POS_a(D)}(6) = 0.4$. Theo công thức (6), hàm thuộc $\left|\mu_{POS_a(D)}(x)\right| = 2$. Tính tương tự $\left|\mu_{POS_b(D)}(x)\right| = 2.4$, $\left|\mu_{POS_c(D)}(x)\right| = 1.6$, theo F_RSAR 1 ta có *P={b}*.

Áp dụng tiếp các bước của F_RSAR 1 ta có *P={a,b}*.

Thuật toán F_RSAR 1 tìm một tập rút gọn nhỏ nhất bảo toàn miền dương mờ, có độ phức tạp tính toán là $O(|U|\times c^{|A|})$, với $|U|$ số lượng đối tượng, $|A|$ là số lượng thuộc tính điều kiện, c là số lượng các tập mờ biểu diễn cho mỗi thuộc tính điều kiện [2]. Trong khi đó, tập rút gọn tìm được bởi thuật toán FuzzyQuickReduct của R. Jensen và các cộng sự [7] không bảo toàn miền dương mờ. Tuy nhiên, F_RSAR 1 luôn có độ phức tạp là hàm mũ của số thuộc tính điều kiện khi tính $\left|\mu_{\mathrm{POS}_C(D)}(x)\right|$, bằng FuzzyQuickReduct trong trường hợp xấu nhất. Do vậy, thuật toán F_RSAR 1 chỉ mang tính học thuật, không khả thi khi áp dụng thực tế.

B. Rút gọn thuộc tính của bảng quyết định sử dụng miền dương mờ dựa trên quan hệ tương tự mờ

Phương pháp sử dụng quan hệ tương tự mờ giải quyết bài toán rút gọn trực tiếp thuộc tính trên bảng quyết định có miền giá trị số thực. Theo hướng tiếp cận này, giá trị hàm thuộc của các đối tượng trên các tập mờ được xem là các giá trị cụ thể của ma trận quan hệ mờ; ma trận quan hệ mờ được định nghĩa mềm dẻo bằng một quan hệ tương tự mờ nào đó trên các tập thuộc tính điều kiện. Phương pháp tiếp cận này không cần phải tính tất cả các phân hoạch mờ, do vậy tránh được độ phức tạp tính toán là hàm mũ của thuộc tính điều kiện theo cách tiếp cận dựa trên phân hoạch mờ. Do vậy, cách tiếp cận này được nghiên cứu sâu rộng, có tính ứng dụng thực tế cao.

1. Ma trận quan hệ mờ

Cho $U = \{x_1, ..., x_n\}$ là tập hữu hạn, khác rỗng *n* đối tượng, R là một quan hệ tương tự mờ trên U . Ma trận quan hệ của R trên U , ký hiệu là $M(R)$, được xác định $M(R) = r_{ij} = R(x_i, x_j)$ là giá trị của quan hệ giữa đối tượng x_i và x_j , $r_{ij} \in [0,1]$ với mọi *i,j=1..n*.

Quan hệ tương tự mờ *R* xác định một phân hoạch mờ (fuzzy partition) trên *U*, ký hiệu $U/R = \{[x_i]_R\}_{i=1}^n$, với $[x_i]_R$ là một tập mờ, được gọi là lớp tương đương mờ. $[x_i]_R$ được biểu diễn dựa trên lý thuyết tập mờ là: $[x_i]_R = \left\{\frac{r_{i1}}{x_1} + \frac{r_{i2}}{x_2} + ... + \frac{r_{in}}{x_n}\right\}$; lực lượng của tập mờ $[x_i]_R$ được tính: $\left|[x_i]_R\right| = \sum_{j=1}^{n} r_{ij}$.

2. Tập thô mờ định nghĩa theo quan hệ tương tự mờ

Giả sử *F* là một tập mờ trên *U* và *R* là một quan hệ tương tự mờ, khi đó tập mờ xấp xỉ dưới $\underline{R}(F)$ và tập mờ xấp xỉ trên $\overline{R}(F)$ của *F* là các tập mờ và hàm thuộc của các đối tượng $x \in U$ được xác định như sau [1, 10]:

$$\mu_{\underline{R}(F)}(x) = \inf_{y\in U} \max\left(1-\mu_{[x]_R}(y), \mu_F(y)\right) \quad (9)$$

$$\mu_{\overline{R}(F)}(x) = \sup_{y\in U} min\left(\mu_{[x]_R}(y), \mu_F(y)\right) \quad (10)$$

Với $\mu_{[x]_R}(y) = \mu_R(x,y) = R(x,y)$ [1,10], cặp $\left(\underline{R}(F), \overline{R}(F)\right)$ được gọi là tập thô mờ.

Cho bảng quyết định có miền giá trị thuộc tính số thực $DS = (U, C \cup D)$ với $U = \{u_1, ..., u_n\}$, $C = \{c_1, ..., c_m\}$. Giả sử một quan hệ tương tự mờ R xác định trên miền giá trị của thuộc tính $c_k \in C$, ký hiệu $R(\{c_k\})$ với k = 1... m. Khi đó $R(C)$ là quan hệ R xác định trên tập thuộc tính điều kiện C. Khái niệm miền dương $POS_c(D)$ trong lý thuyết tập thô truyền thống được mở rộng thành khái niệm miền dương mờ của tập thuộc tính D đối với tập thuộc tính C dựa trên quan hệ R, ký hiệu là $POS_{R(C)}(D)$. $POS_{R(C)}(D)$ là một tập mờ mà hàm thuộc của các đối tượng $x \in U$ được định nghĩa [10].

$$\mu_{POS_{R(C)}(D)}(x) = \sup_{X \in U/D} \mu_{\underline{R(C)}(X)}(x) \quad (11)$$

Lực lượng của miền dương mờ dựa trên quan hệ R được xác định [10].

$$\left|\mu_{POS_{R(C)}(D)}(x)\right| = \sum\nolimits_{x \in U} \mu_{POS_{R(C)}(D)}(x) \quad (12)$$

Cho bảng quyết định có miền giá trị thuộc tính số $DS = (U, C \cup D)$ với $P, Q \subseteq C$ và $R(P)$, $R(Q)$, là quan hệ R trên tập thuộc tính P, Q tương ứng. Khi đó ta có $R(P \cup Q) = R(P) \cap R(Q)$ [5], nghĩa là với mọi $x, y \in U$, $R(P \cup Q)(x,y) = \min\{R(P)(x,y), R(Q)(x,y)\}$. Giả sử $M(R(P)) = \left[r_{ij}^{R(P)}\right]_{n \times n}$ và $M(R(Q)) = \left[r_{ij}^{R(Q)}\right]_{n \times n}$ là các ma trận quan hệ của R trên tập thuộc tính P và Q tương ứng, khi đó ma trận quan hệ của R trên tập thuộc tính $P \cup Q$ là:

$$M(R(P \cup Q)) = \left[r_{ij}^{R(P \cup Q)}\right]_{n \times n} \text{ với } r_{ij}^{R(P \cup Q)} = \min\left\{r_{ij}^{R(P)}, r_{ij}^{R(Q)}\right\} \quad (13)$$

Tiếp theo, chúng tôi đề xuất phương pháp heuristic tìm một tập rút gọn nhỏ nhất của bảng quyết định có miền giá trị số thực dựa trên quan hệ tương tự mờ, bao gồm các bước: định nghĩa tập rút gọn dựa trên miền dương mờ, định nghĩa độ quan trọng của thuộc tính và xây dựng thuật toán heuristic tìm tập rút gọn nhỏ nhất dựa trên tiêu chuẩn độ quan trọng của thuộc tính.

Định nghĩa 3. Cho bảng quyết định có miền giá trị thuộc tính số $DS = (U, C \cup D)$, quan hệ tương tự mờ R và tập thuộc tính $P \subseteq C$. Nếu

$$\begin{aligned} &1)\ \left|\mu_{POS_{R(P)}(D)}(x)\right| = \left|\mu_{POS_{R(C)}(D)}(x)\right| \\ &2)\ \forall p \in P, \left|\mu_{POS_{R(P-\{p\})}(D)}(x)\right| \neq \left|\mu_{POS_{R(C)}(D)}(x)\right| \end{aligned} \quad (14)$$

thì P là một tập rút gọn nhỏ nhất của C dựa trên miền dương mờ của thuộc tính.

Định nghĩa 4. Cho bảng quyết định có miền giá trị thuộc tính số $DS = (U, C \cup D)$ và quan hệ tương tự mờ R xác định trên miền giá trị thuộc tính. Với $B \subset C$, độ quan trọng của thuộc tính $b \in C - B$ đối với tập thuộc tính B dựa trên quan hệ R được định nghĩa:

$$SIG_{R(B)}(b) = \left|\mu_{\mathrm{POS}_{R(B \cup \{b\})}(D)}(x)\right| - \left|\mu_{\mathrm{POS}_{R(B)}(D)}(x)\right| \quad (15)$$

Độ quan trọng của thuộc tính ở công thức (15) được sử dụng làm tiêu chuẩn lựa chọn thuộc tính cho thuật toán heuristic tìm tập rút gọn nhỏ nhất dựa trên miền dương mờ như sau.

Thuật toán F_RSAR 2 *(Fuzzy Rough Set based Attribute Reduction)*

Đầu vào: Bảng quyết định giá trị thuộc tính số $DS = (U, C \cup D)$, quan hệ tương tự mờ R.

Đầu ra: Một tập rút gọn nhỏ nhất P.

1. $P \leftarrow \varnothing$; $|\mu_{\mathrm{POS}_{R(\varnothing)}(D)}(x)| = 0$;
2. Tính $\left|\mu_{\mathrm{POS}_{R(C)}(D)}(x)\right|$;
3. While $\left|\mu_{\mathrm{POS}_{R(P)}(D)}(x)\right| \neq \left|\mu_{\mathrm{POS}_{R(C)}(D)}(x)\right|$ Do
4. Begin
5. For $c \in C - P$ tính

$SIG_P(c) = \left|\mu_{\mathrm{POS}_{R(P\cup\{c\})}(D)}(x)\right| - \left|\mu_{\mathrm{POS}_{R(P)}(D)}(x)\right|$;

6. Chọn $c_m \in C - P$ sao cho

$$SIG_P(c_m) = \underset{c\in C-P}{Max}\{SIG_P(c)\};$$

7. $P \leftarrow P \cup \{c_m\}$;
8. End;

//Loại bỏ các thuộc tính dư thừa trong P nếu có

9. For each $a \in P$
10. Begin
11. Tính $\left|\mu_{POS_{R(P-\{a\})}(D)}(x)\right|$;
12. If $\left|\mu_{\mathrm{POS}_{R(P-\{a\})}(D)}(x)\right| = \left|\mu_{\mathrm{POS}_{R(C)}(D)}(x)\right|$
13. then $P = P - \{a\}$;
14. End;
15. Return P;

Ví dụ 2. Xét bảng quyết định $DS = (U, C \cup D)$, $C = \{c_1, c_2, c_3, c_4, c_5, c_6\}$ cho ở Ví dụ 1 (Bảng 1). Định nghĩa quan hệ tương tự mờ $R(\{c_k\}); k = 1..6$ trên $c_k \in C$ như sau:

$$R(\{c_k\})(x_i, x_j) = \begin{cases} 1 - 4 * \dfrac{|x_i - x_j|}{|\max(c_k) - \min(c_k)|}, & \dfrac{|x_i - x_j|}{|\max(c_k) - \min(c_k)|} \le 0.25 \\ 0, \quad otherwise \end{cases} \quad (16)$$

*M(R({C}))*được tính thông qua các ma trận quan hệ tương tự mờ trên các thuộc tính điều kiện $M(R(\{c_k\}))$; *k = 1 ... 6*. Từ đó tính được $\left|\mu_{\mathrm{POS}_{R(C)}(D)}(x)\right|$.

$$M(R(C)) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\left|\mu_{POS_{R(C)}(D)}(x)\right| = \sum\nolimits_{x\in U} \mu_{POS_{R(C)}(D)}(x) = 6$$

Áp dụng các bước của Thuật toán F_RSAR 2 tìm được tập rút gọn nhỏ nhất $P = \{c_4, c_1\}$, tương ứng với *P = {b, a}* của các thuộc tính trước khi mờ hóa.

Thuật toán F_RSAR 2 tìm được một tập rút gọn nhỏ nhất dựa trên độ quan trọng của thuộc tính sử dụng miền dương mờ. F_RSAR 2 tính toán miền dương mờ của thuộc tính thông qua ma trận quan hệ mờ, có độ phức tạp tính toán $O(|C|^3|U|^2)$ với $|U|$ số lượng đối tượng, $|C|$ là số lượng thuộc tính điều kiện, tránh được độ phức tạp tính toán là hàm mũ của số thuộc tính điều kiện như F_RSAR 1. Dễ thấy rằng, tập rút gọn thu được của Thuật toán F_RSAR 2 cũng bảo toàn miền dương.

C. Thực nghiệm

Để đánh giá khả năng ứng dụng trong thực tế của F_RSAR 2, chúng tôi tiến hành cài đặt thuật toán F_RSAR 2 và thuật toán GAIN_RATIO_AS_FRS (gọi tắt là thuật toán GRAF) tìm tập rút gọn sử dụng lượng thông tin tăng thêm (gain ratio) theo tiếp cận tập thô mờ trong công trình [3] để so sánh với thuật toán đề xuất F_RSAR 2 về thời gian thực hiện và số lượng thuộc tính của tập rút gọn thu được. Chúng tôi chọn GRAF[3] vì đây là đây là công bố mới, có kết quả tốt nhất về phương pháp tìm một tập rút gọn tốt nhất đã được công bố cho đến thời điểm hiện nay theo tiếp cận tập thô mờ. Chúng tôi không cài đặt F_RSAR 1 để so sánh với F_RSAR 2 vì sự so sánh này không có ý nghĩa khi đã kết luận độ phức tạp tính toán của F_RSAR 1 là hàm mũ của số thuộc tính điều kiện, không khả thi khi ứng dụng thực tế. Để tiến hành thử nghiệm, chúng tôi cài đặt cả hai thuật toán bằng ngôn ngữ C# trên máy Pentium 2 Duo 2.20 GHz CPU, 2 GB RAM, hệ điều hành Windows 7, chạy thử nghiệm với 5 bộ số liệu lấy từ kho dữ liệu UCI[8].

Với mỗi bộ số liệu, giả sử $|U|$ là số đối tượng, $|C|$ là số thuộc tính điều kiện, $|R|$ là số thuộc tính của tập rút gọn, *t* là thời gian thực hiện thuật toán (đơn vị là giây), các thuộc tính điều kiện được đánh số là 1, 2,... , $|C|$. Kết quả thực hiện được mô tả ở Bảng II.

Bảng II. Kết quả thực hiện Thuật toán F_RSAR 2 và GRAF[3]

TT	Bộ số liệu	|U|	|C|	Thuật toán F_RSAR 2		Thuật toán GRAF[3]	
				|R|	t	|R|	t
1	Ionosphere	351	34	12	0,96	12	1,01
2	Wpbc	198	33	15	0,61	17	0,65
3	Wine	178	13	6	0,23	6	0,25
4	Glass	214	9	7	0,40	8	0,45
5	Magic04	19020	10	9	5,96	9	6,25

Kết quả thử nghiệm cho thấy:

- Trên các bộ số liệu Ionosphere.data, Wine.data, Magic04.data, tập rút gọn thu được bởi Thuật toán F_RSAR 2 và Thuật toán GRAF[3] là như nhau. Tuy nhiên, với bộ số liệu Wpbc.data, Glass.data, tập rút gọn thu được bởi Thuật toán F_RSAR 2 tối thiểu hơn tập rút gọn thu được bởi Thuật toán GRAF[3].
- Thời gian thực hiện của F_RSAR 2 nhỏ hơn GRAF[3], đặc biệt là trên các bộ số liệu lớn thì sự chênh lệnh này càng nhiều do thuật toán GRAF[3] phải tính logarit trong các công thức tính entropy shanon.

IV. KẾT LUẬN

Mô hình tập thô mờ do D. Dubois và các cộng sự [1] đề xuất là công cụ hiệu quả để giải quyết bài toán rút gọn thuộc tính trực tiếp trên các bảng quyết định có miền giá trị thuộc tính số thực. Trong bài báo này, chúng tôi đề xuất hai phương pháp heuristic tìm một tập rút gọn nhỏ nhất của bảng quyết định có miền giá trị thuộc tính số thực sử dụng miền dương mờ dựa trên phân hoạch mờ và quan hệ tương tự mờ. Miền dương mờ trong F_RSAR 1 được xác định dựa trên phân hoạch mờ ở công thức (6). Miền dương mờ trong F_RSAR 2 được xác định dựa trên ma trận quan hệ tương tự mờ ở công thức (12). Thực nghiệm trên các bộ số liệu UCI[8] cho thấy, F_RSAR 2 có khả năng ứng dụng thực tế. Định hướng nghiên cứu tiếp theo là tìm kiếm các độ đo hiệu quả để giải quyết bài toán rút gọn thuộc tính theo tiếp cận tập thô mờ.

TÀI LIỆU THAM KHẢO

[1] D. Dubois and H. Prade. Rough fuzzy sets and fuzzy rough sets. International Journal of General Systems, 17, pp. 191-209, 1990.

[2] C.C. Eric Tsang, Degang Chen, S. Daniel Yeung, Xi-Zhao Wang, and W.T. John Lee. Attributes Reduction Using Fuzzy Rough Sets, IEEE Transactions On Fuzzy Systems, Vol. 16, No. 5, October 2008.

[3] Jianhua Dai and Qing Xu. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. Applied Soft Computing 13, pp. 211–221, 2013.

[4] Lotfi Aliasker Zadeh. Fuzzy sets. Information and Control, pp. 338-353, 1965.

[5] Qinghua Hu, Daren Yu, Zongxia Xie. Information-preserving hybrid data reduction based on fuzzy-rough techniques. Pattern Recognition Letters 27, 2006, pp. 414-423.

[6] Z. Pawlak. Rough sets. International Journal of Computer and Information Sciences, 11(5): 341-356, 1982.

[7] Richard Jensen and Qiang Shen. Fuzzy–rough attribute reduction with application to web categorization. Fuzzy Sets and Systems, Volume 141, Issue 3, pp. 469-485, 2004.

[8] The UCI machine learning repository, http://archive.ics.uci.edu/ml/datasets.html.

[9] Xiao Zhang, Changlin Mei, Degang Chen, Jinhai Li, Feature selection in mixed data: A method using a novel fuzzy rough set-based information entropy, PatternRecognition 56, pp.1-15, 2016.

[10] Yi Cheng. Forward approximation and backward approximation in fuzzy rough sets. Neurocomputing, Volume 148, pp. 340-353, 2014.

FUZZY POSITIVE REGION BASED ATTRIBUTE REDUCTION IN DECISION TABLES

Abstract: Traditional rough set based attribute reduction methods has performed on the decision tables with real value attribute domain needs to be discretized data. The discretized data can be lost information which will affect the quality of data classification. To overcome this drawback, attribute reduction performs directly on the decision table with real value attribute according to fuzzy rough set approach has proved effective. In this paper, we propose two attribute reduction methods using fuzzy positive region based on fuzzy partition and fuzzy similarity relation. Analyzing and evaluating for each method which concludes the method using fuzzy similarity relation has practical application.

Keywords: Fuzzy rough set, fuzzy decision table, fuzzy partition, fuzzy similarity relation, fuzzy positive region, attribute reduction, reduct.

Cao Chính Nghĩa, nhận bằng Thạc sĩ năm 2006 tại Đại học Công nghệ, Đại học Quốc gia Hà Nội. Hiện công tác tại Học viện Cảnh sát nhân dân, Bộ Công an. Lĩnh vực nghiên cứu: cơ sở dữ liệu, khai phá dữ liệu và học máy.

Vũ Đức Thi, nhận bằng Tiến sĩ năm 1987 tại Học viện Khoa học Hungary, học hàm Phó Giáo sư năm 1991, Giáo sư năm 2009. Hiện đang công tác tại Đại học Sư phạm Kỹ thuật Hưng Yên. Lĩnh vực nghiên cứu: cơ sở dữ liệu, khai phá dữ liệu và học máy.

Nguyễn Long Giang, nhận bằng Tiến sĩ năm 2012 tại Viện Công nghệ thông tin, Viện Hàn lâm khoa học Việt Nam. Hiện đang công tác tại Viện Công nghệ thông tin, Viện Hàn lâm khoa học Việt Nam. Lĩnh vực nghiên cứu: cơ sở dữ liệu, khai phá dữ liệu và học máy.

Tân Hạnh, nhận bằng Tiến sĩ năm 2009 tại Viện Công nghệ Grenoble, Pháp. Hiện đang công tác tại Học viện Công nghệ Bưu chính Viễn Thông, Thành phố Hồ Chí Minh. Lĩnh vực nghiên cứu: cơ sở dữ liệu, tìm kiếm thông tin, phục hồi thông tin, hệ thống phân tán.