Về phương pháp rút gọn thuộc tính trong bảng quyết định với miền trị thuộc tính nhận giá trị số theo tiếp cận tập thô mờ

Fuzzy Rough Set based Attribute Reduction in Numeric Domain Decision Tables

Nguyễn Văn Thiện, Nguyễn Long Giang, Nguyễn Như Sơn

Abstract: *Attributes reduction based on rough set is interesting research area. However, the attributes reduction algorithms based on rough set is done on the discrete domain decision tables (that is applied discretization methods). In recent years, some researchs on fuzzy rough set based directly attribute reduction in numeric domain decision tables have been studied. This paper proposes fuzzy rough set based directly attribute reduction method in numeric domain decision tables. The experiment results showed that the fuzzy rough set method has better classification accuracy than rough set theory.*

Keywords: *rough set, fuzzy rough set, decision table, fuzzy similarity relation, attribute reduction, reduct.*

I. GIỚI THIỆU

Rút gọn thuộc tính là bài toán quan trọng trong bước tiền xử lý số liệu với mục tiêu là loại bỏ các thuộc tính dư thừa nhằm nâng cao tính hiệu quả của các thuật toán khai phá dữ liệu. Lý thuyết tập thô [13] là công cụ hiệu quả giải quyết bài toán rút gọn thuộc tính trong bảng quyết định và được cộng đồng nghiên cứu về tập thô thực hiện lâu nay. Để thực hiện các phương pháp rút gọn thuộc tính theo tiếp cận tập thô, các thuộc tính có miền giá trị số, liên tục cần được rời rạc hóa. Tuy nhiên, các phương pháp rời rạc hóa dữ liệu không bảo toàn được sự khác nhau ban đầu giữa các giá trị thuộc tính. Ví dụ, với thuộc tính "*Nhiệt độ cơ thể*", giả sử hai đối tượng *x* và *y* có nhiệt độ cơ thể tương ứng là 39.5 độ và 39.6 độ được rời rạc hóa thành một giá trị "*Nhiệt độ cao*". Trên bảng quyết định mới, hai đối tượng *x* và *y* có giá trị bằng nhau trên thuộc tính "*Nhiệt độ cơ thể*" và không bảo toàn được sự khác nhau là 0.1 độ trên bảng quyết định ban đầu. Do đó, các phương pháp rời rạc hóa dữ liệu không bảo toàn "ngữ nghĩa" của dữ liệu gốc và có thể làm giảm độ chính xác phân lớp trên dữ liệu gốc. Để giải quyết bài toán rút gọn thuộc tính trực tiếp trên các bảng quyết định có miền trị thuộc tính nhận giá trị số, liên tục nhằm khắc phục nhược điểm trên, trong mấy năm gần đây các công trình nghiên cứu đề xuất hướng tiếp cận mới sử dụng lý thuyết tập thô mờ

Lý thuyết tập thô mờ (Fuzzy Rough Set) do D. Dubois và các cộng sự [4, 5] đề xuất mở ra một hướng nghiên cứu mới về rút gọn thuộc tính trên các bảng quyết định mờ và các bảng quyết định có miền trị thuộc tính nhận giá trị số, liên tục. Lý thuyết tập thô mờ là sự kết hợp của lý thuyết tập thô [13] và lý thuyết tập mờ [11] nhằm xấp xỉ các tập mờ dựa trên một quan hệ tương tự (similarity relation) được xác định trên miền giá trị thuộc tính. Trong lý thuyết tập thô, hai đối tượng là tương đương trên tập thuộc tính *R*, hay độ tương tự là 1, nếu giá trị thuộc tính của chúng bằng nhau trên tất cả các thuộc tính trong *R*. Ngược lại, chúng là không tương đương, hay độ tương tự là 0. Trong lý thuyết tập thô mờ, quan hệ tương tự thay thế quan hệ tương đương nhằm xác định độ tương tự của hai đối tượng. Độ tương tự là một giá trị nằm trong khoảng [0, 1] cho thấy tính gần nhau, hay tính tương

tự, của hai đối tượng. Các nghiên cứu liên quan đến rút gọn thuộc tính theo tiếp cận tập thô mờ tập trung vào hai hướng chính: hướng thứ nhất là sử dụng tập thô mờ để giải quyết bài toán rút gọn thuộc tính trên các bảng quyết định mờ (bảng quyết định với giá trị thuộc tính là các tập mờ) trước khi thực hiện các thuật toán trích lọc hệ luật mờ với các các công bố điển hình là [6, 7, 8, 15, 18, 19, 20, 24]; hướng thứ hai giải quyết bài toán rút gọn thuộc tính trực tiếp trên các bảng quyết định có miền trị thuộc tính nhận giá trị số, đây là hướng nghiên cứu của bài báo này.

Theo hướng tiếp cận rút gọn thuộc tính trực tiếp trên các bảng quyết định có miền trị thuộc tính nhận giá trị số, trước hết một quan hệ tương tự được định nghĩa trên miền giá trị thuộc tính. Tiếp theo, các ma trận quan hệ được xây dựng. Ma trận quan hệ cho phép xác định giá trị hàm thuộc của các đối tượng đối với mỗi lớp tương tự mờ. Từ đó, hàm thuộc của các tập xấp xỉ dưới mờ, xấp xỉ trên mờ, miền dương mờ được tính dựa vào các toán tử xấp xỉ trong lý thuyết tập thô mờ [4, 5]. Trên cơ sở đó, các phương pháp rút gọn thuộc tính được xây dựng dựa trên nền tảng mở rộng các phương pháp rút gọn thuộc tính theo tiếp cận tập thô truyền thống. Đóng góp quan trọng về hướng nghiên cứu này phải kể đến các công trình [1, 2, 3, 21, 22, 23, 25]. Trong các công trình này, các tác giả xây dựng ma trận phân biệt mờ dựa trên ma trận quan hệ và đề xuất thuật toán tìm tất cả các tập rút gọn bằng cách mở rộng phương pháp rút gọn thuộc tính dựa trên ma trận phân biệt trong lý thuyết tập thô truyền thống. Tuy nhiên, các tác giả chưa công bố thuật toán heuristic tìm một tập rút gọn tốt nhất dựa trên tiêu chuẩn chất lượng phân lớp hay độ quan trọng của thuộc tính. Trong công trình [8, 17], các tác giả xây dựng các phương pháp rút gọn thuộc tính dựa trên entropy Shannon. Các tác giả trong [8, 17] cũng minh chứng bằng thực nghiệm rằng, phương pháp rút gọn thuộc tính theo tiếp cận tập thô mờ có độ chính xác phân lớp tốt hơn phương pháp rút gọn theo tiếp cận lý thuyết tập thô truyền thống (sau khi rời rạc hóa dữ liệu) trên một số bộ dữ liệu thử nghiệm.

Trong bài báo này, chúng tôi đề xuất thuật toán heuristic rút gọn thuộc tính trong bảng quyết định có miền trị thuộc tính nhận giá trị số sử dụng độ phụ thuộc mờ của thuộc tính trong tập thô mờ. Thuật toán đề xuất tìm một tập rút gọn tốt nhất theo tiêu chuẩn chất lượng phân lớp (độ quan trọng của thuộc tính), do đó hiệu quả hơn các công bố trong [1, 2, 3, 21, 22, 23, 25]. Do sử dụng độ phụ thuộc mờ của các thuộc tính nên thuật toán đề xuất có khối lượng tính toán nhỏ hơn thuật toán trong [8, 17] sử dụng công thức entropy Shannon mờ. Kết quả thử nghiệm trên một số bộ số liệu cho thấy, phương pháp đề xuất có độ chính xác phân lớp tốt hơn so với phương pháp sử dụng độ phụ thuộc của thuộc tính theo tiếp cận lý thuyết tập thô truyền thống. Hơn nữa, phương pháp đề xuất có độ chính xác phân lớp tốt hơn các phương pháp dựa trên entropy Shannon mờ trong [8, 17]. Cấu trúc bài báo như sau. Phần 2 trình bày một số khái niệm cơ bản trong lý thuyết tập thô mờ. Phần 3 trình bày phương pháp rút gọn thuộc tính sử dụng độ phụ thuộc mờ của thuộc tính theo tiếp cận tập thô mờ. Phần 4 trình bày kết quả thử nghiệm. Cuối cùng là kết luận và hướng phát triển tiếp theo

II. CÁC KHÁI NIỆM CƠ BẢN

Phần này trình bày một số khái niệm cơ bản trong lý thuyết tập thô truyền thống của Pawlak [13] và lý thuyết tập thô mờ do D. Dubois và các cộng sự [4, 5] đề xuất.

Mô hình tập thô truyền thống do Pawlak [13] đề xuất dựa trên quan hệ tương đương để xấp xỉ tập hợp. Xét bảng quyết định $DS = (U, C \cup D)$, Mỗi tập thuộc tính $P \subseteq C$ xác định một quan hệ tương đương $IND(P) = \{(u,v) \in U \times U \mid \forall a \in P, a(u) = a(v)\}$. Nếu $(u,v) \in IND(P)$ thì *u* và *v* không phân biệt được nhau bởi các thuộc tính trong *P*. Ký hiệu lớp tương đương chứa đối tượng *u* là $[u]_P$, khi đó $[u]_P = \{v \in U \mid (u,v) \in IND(P)\}$. Với $X \subseteq U$, các tập $\underline{C}X = \{u \in U \mid [u]_C \subseteq X\}$ và $\overline{C}X = \{u \in U \mid [u]_C \cap X \neq \varnothing\}$ tương

ứng gọi là *C-xấp xỉ dưới*, *C-xấp xỉ trên* của *X*. Ta gọi tập $POS_C(D) = \bigcup_{X \in U/D} (\underline{C}X)$ là *C-miền dương* của *D*. Dễ thấy $POS_C(D)$ là tập các đối tượng trong *U* được phân lớp đúng vào các lớp của U/D sử dụng tập thuộc tính *C*. Độ phụ thuộc của tập thuộc tính *C* vào tập thuộc tính *D* trong lý thuyết tập thô truyền thống, ký hiệu là $\gamma_C(D)$, được định nghĩa như sau:

$$k = \gamma_C(D) = \frac{|POS_C(D)|}{|U|}$$

với $|S|$ là lực lượng của tập *S*. Nếu *k =1* thì *D* phụ thuộc hoàn toàn vào *C*. Nếu $0 < k < 1$, *D* phụ thuộc bộ phận vào *C*. Tiếp theo, chúng tôi trình bày các khái niệm trên trong lý thuyết tập thô mờ trong [4, 5].

Trong mô hình tập thô mờ, một quan hệ tương tự (similarity relation) được sử dụng thay thế quan hệ tương đương để xấp xỉ các tập mờ. Cho *U* là tập các đối tượng, một quan hệ *R* được định nghĩa trên *U* được gọi là quan hệ tương tự nếu *R* thỏa mãn các tính chất: tính phản xạ (reflexive) $R(x,x)=1$, tính đối xứng (symetric) $R(x,y)=R(y,x)$ và tính bắc cầu sup-min (sup-min transitive) $R(x,z) \geq \min\{R(x,y), R(y,z)\})$ với mọi $x, y, z \in U$. Quan hệ tương tự *R* xác định một phân hoạch mờ trên *U*, ký hiệu $U/R = \{[x]_R | x \in U\}$, trong đó $[x]_R$ là một lớp tương đương mờ tương ứng với đối tượng *x*, hàm thuộc được xác định bởi công thức $\mu_{[x]_R}(y) = \mu_R(x,y) = R(x,y)$ với mọi $y \in U$.

Giả sử *F* là một tập mờ và *R* là một quan hệ tương tự xác định trên *U*, 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:

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

$$\mu_{\overline{R}(F)}(x) = \sup_{y \in U} min\left(\mu_R(x,y), \mu_F(y)\right) \quad (2)$$

với $\mu_{[x]_R}(y) = \mu_R(x,y)$, khi đó tập mờ xấp xỉ dưới $\underline{R}(F)$ và tập mờ xấp xỉ trên $\overline{R}(F)$ được viết lại như sau:

$$\mu_{\underline{R}(F)}(x) = \inf_{y \in U} \max\left(1 - \mu_{[x]_R}(y), \mu_F(y)\right) \quad (3)$$

$$\mu_{\overline{R}(F)}(x) = \sup_{y \in U} min\left(\mu_{[x]_R}(y), \mu_F(y)\right) \quad (4)$$

Cặp $\left(\underline{R}(F), \overline{R}(F)\right)$ được gọi là tập thô mờ. Dễ thấy rằng một tập hợp (tập rõ) bất kỳ $X \subseteq U$ có thể xem là một tập mờ, hàm thuộc $\mu_X(y) = 1$ với $y \in X$ và $\mu_X(y) = 0$ với $y \notin X$. Mô hình tập thô mờ có thể xem là việc sử dụng quan hệ tương tự để xấp xỉ tập mờ (hoặc tập rõ) bằng tập mờ xấp xỉ dưới và tập mờ xấp xỉ trên.

Cho bảng quyết định có miền trị thuộc tính nhận giá trị số $DS = (U, C \cup D)$ với $U = \{u_1, ..., u_n\}$, $C = \{c_1, ..., c_m\}$. Giả sử một quan hệ tương tự R bất kỳ xác định trên miền giá trị của các thuộc tính điều kiện $c \in C$. Ta ký hiệu c_k là quan hệ *R* xác định trên thuộc tính điều kiện $c_k \in C, k = 1..m$. Khi đó, 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 $POS_C(D)$. $POS_C(D)$ là một tập mờ mà hàm thuộc của các đối tượng $x \in U$ được định nghĩa như sau [18].

$$\mu_{POS_C(D)}(x) = \sup_{X \in U/D} \mu_{\underline{C}(X)}(x) \quad (5)$$

Dựa trên khái niệm miền dương mờ, độ phụ thuộc của tập thuộc tính điều kiện *C* vào tập thuộc tính quyết định *D* dựa trên quan hệ *R* được định nghĩa theo tiếp cận tập thô mờ như sau [18].

$$\gamma_C(D) = \frac{\left|\mu_{POS_C(D)}(x)\right|}{|U|} = \frac{\sum_{x \in U} \mu_{POS_C(D)}(x)}{|U|} \quad (6)$$

Theo hướng tiếp cận rút gọn thuộc tính trực tiếp trên các bảng quyết định thuộc tính số, ma trận quan

hệ xác định trên thuộc tính c_k là ma trận vuông cấp *n* được định nghĩa như sau [21]

$$M\left(c_k\right)=\left[m_{ij}^{c_k}\right]_{n\times n}$$

với $m_{ij}^{c_k}=c_k\left(u_i,u_j\right),\ i=1..n,\ j=1..n$. Ở đây, $c_k\left(u_i,u_j\right)$ là giá trị của quan hệ giữa hai đối tượng u_i và u_j. Khi đó, ma trận quan hệ mờ xác định trên tập thuộc tính điều kiện C được định nghĩa [21]

$$M\left(C\right)=\left[m_{ij}^{C}\right]_{n\times n}$$

với $m_{ij}^{C}=\min\limits_{k=1..m}\left\{m_{ij}^{c_k}\right\}=\min\left\{m_{ij}^{c_1},m_{ij}^{c_2},...,m_{ij}^{c_m}\right\}$

Dễ thấy rằng, với phần tử m_{ij}^{C} của ma trận $M\left(C\right)$ ta có $m_{ij}^{C}=C\left(u_i,u_j\right),\ i=1..n,\ j=1..n$, với $C\left(u_i,u_j\right)=\mu_C\left(u_i,u_j\right)=\mu_{[u_i]_C}\left(u_j\right)$. Từ ma trận $M\left(C\right)$ cho phép xác định được các lớp tương đương mờ $\left[u_i\right]_C=\frac{m_{i1}^{C}}{u_1}+...+\frac{m_{i1}^{C}}{u_n}$ thuộc phân hoạch mờ $U/C=\left\{\left[u_i\right]_C\middle|u_i\in U\right\}$. Khi đó, hàm thuộc của các tập mờ xấp xỉ dưới, tập mờ xấp xỉ trên, miền dương mờ có thể được tính dựa vào các công thức (3), (4), (5) tương ứng và từ đó tính được độ phụ thuộc của tập thuộc tính điều kiện *C* vào tập thuộc tính quyết định *D* theo công thức (6).

III. RÚT GỌN THUỘC TÍNH TRONG BẢNG QUYẾT ĐỊNH VỚI MIỀN TRỊ THUỘC TÍNH NHẬN GIÁ TRỊ SỐ

Như đã trình bày ở phần I, trên lớp bài toán rút gọn thuộc tính trực tiếp trên bảng quyết định với miền trị thuộc tính nhận giá trị số theo hướng tiếp cận heuristic sử dụng tập thô mờ, trong công trình [17], các tác giả đã định nghĩa tập rút gọn dựa trên entropy Shannon mờ và xây dựng thuật toán heuristic tìm một tập rút gọn tốt nhất dựa trên entropy Shannon mờ. Trong công trình [8], các tác giả định nghĩa tập rút gọn dưa trên độ đo lượng thông tin tăng thêm mờ (fuzzy information gain). Lượng thông tin tăng thêm mờ được xây dựng dựa trên entropy Shannon mờ. Các tác giả trong [8] cũng xây dựng thuật toán heuristic tìm một tập rút gọn tốt nhất dựa trên lượng thông tin tăng thêm mờ. Trong cả hai công trình [8, 17], các tác giả đều chứng minh bằng thực nghiệm rằng rút gọn thuộc tính theo tiếp cận thô mờ có độ chính xác phân lớp tốt hơn rút gọn thuộc tính theo tiếp cận lý thuyết tập thô truyền thống. Tuy nhiên, nhược điểm của hai phương pháp này là đều sử dụng công thức entropy Shannon để xây dựng các định nghĩa tập rút gọn, do đó thời gian thực hiện kém hiệu quả do phải tính toán các biểu thức logarit. Trong bài báo này, chúng tôi sử dụng độ phụ thuộc mờ của thuộc tính thay cho độ đo entropy Shannon để định nghĩa tập rút gọn và xây dựng thuật toán heuristic tìm một tập rút gọn tốt nhất. Vì độ phụ thuộc mờ của thuộc tính không phải tính toán các biểu thức logarit nên hiệu quả hơn các phương pháp dựa trên entropy Shannon trong [8, 17]. Chúng tôi cũng chứng minh bằng thực nghiệm rằng phương pháp đề xuất có độ chính xác phân lớp cao hơn các phương pháp dựa trên entropy Shannon trong [8, 17].

Tương tự phương pháp rút gọn thuộc tính trong lý thuyết tập thô truyền thống, phương pháp đề xuất bao gồm các bước: định nghĩa tập rút gọn dựa trên độ phụ thuộc mờ của thuộc tính, định nghĩa độ quan trọng của thuộc tính đặc trưng cho chất lượng phân lớp của thuộc tính và xây dựng thuật toán heuristic tìm tập rút gọn tốt nhất dựa trên tiêu chuẩn độ quan trọng của thuộc tính.

Định nghĩa 1. Cho bảng quyết định $DS=\left(U,C\cup D\right)$ có miền trị thuộc tính nhận giá trị số , một quan hệ tương tự *R* được xác định trên miền giá trị của các thuộc tính. Với $P\subseteq C$, nếu

1) $\gamma_P(D)=\gamma_C(D)$

2) $\forall p\in P,\gamma_{P-\{p\}}(D)\neq\gamma_C(D)$

thì *P* là một tập rút gọn của *C* dựa trên độ phụ thuộc mờ của thuộc tính.

Từ Định nghĩa 1, dễ thấy rằng tập rút gọn dựa trên độ phụ thuộc mờ của thuộc tính tương đương với tập rút gọn dựa trên miền dương mờ, tập rút gọn dựa trên

miền dương mờ là mở rộng tập rút gọn dựa trên miền dương của Pawlak theo tiếp cận lý thuyết tập thô mờ.

Định nghĩa 2. Cho bảng quyết định $DS=(U,C\cup D)$ có miền giá trị thuộc tính số và quan hệ tương tự *R* xác định trên miền giá trị thuộc tính. Với $B\subset C$, độ quan trọng mờ của thuộc tính $b\in C-B$ đối với tập thuộc tính *B* được định nghĩa:

$$SIG_B(b)=\gamma_{B\cup\{b\}}(D)-\gamma_B(D) \quad (7)$$

Độ quan trọng của thuộc tính đặc trưng cho chất lượng phân lớp của thuộc tính điều kiện vào thuộc tính quyết định và được sử dụng làm tiêu chuẩn lựa chọn thuộc tính cho thuật toán heuristic tìm tập rút gọn sau đây.

Thuật toán F_RSAR (Fuzzy Rough Set based Attribute Reduction). Thuật toán heuristic tìm một tập rút gọn sử dụng độ phụ thuộc mờ của thuộc tính.

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

Đầu ra: Một tập rút gọn tốt nhất P.

1. $P\leftarrow\varnothing$;
2. $\gamma_{\varnothing}(D)=0$;
3. Tính ma trận quan hệ $M(C)$;
4. Tính $\gamma_C(D)$;
5. While $\gamma_P(D)\neq\gamma_C(D)$ do
6. Begin
7. For $c\in C-P$ tính
   $SIG_P(c)=\gamma_{P\cup\{c\}}(D)-\gamma_P(D)$;
8. Chọn $c_m\in C-P$ sao cho
   $SIG_P(c_m)=\underset{c\in C-P}{Max}\{SIG_P(c)\}$;
9. $P\leftarrow P\cup\{c_m\}$;
10. Tính $\gamma_P(D)$;
11. End;
12. Return *P*;

Ví dụ 1. Xét bảng quyết định có miền giá trị thuộc tính số $DS=(U,C\cup\{d\})$ cho ở Bảng 1 với $U=\{u_1,u_2,u_3,u_4,u_5,u_6\}$, $C=\{c_1,c_2,c_3,c_4,c_5,c_6\}$.

Bảng 1. Bảng quyết định mô tả Ví dụ 1

	c_1	c_2	c_3	c_4	c_5	c_6	d
u_1	0.8	0.2	0.6	0.4	1	0	No
u_2	0.8	0.2	0	0.6	0.2	0.8	Yes
u_3	0.6	0.4	0.8	0.2	0.6	0.4	No
u_4	0	0.4	0.6	0.4	0	1	Yes
u_5	0	0.6	0.6	0.4	0	1	Yes
u_6	0	0.6	0	1	0	1	No

Giả sử trên miền giá trị của thuộc tính $c_k\in C$, quan hệ tương tự c_k được định nghĩa như sau [8]

$$c_k(u_i,u_j)=\begin{cases}1-4*\dfrac{|c_k(u_i)-c_k(u_j)|}{|\max(c_k)-\min(c_k)|}, & \dfrac{|c_k(u_i)-c_k(u_j)|}{|\max(c_k)-\min(c_k)|}\le 0.25\\ 0, otherwise\end{cases} \quad (8)$$

Với $\max(c_k),\min(c_k)$ tương ứng là giá trị lớn nhất và giá trị nhỏ nhất của miền giá trị thuộc tính c_k

Áp dụng các bước của Thuật toán F_RSAR tìm một tập rút gọn của bảng quyết định. Trước hết, tính các ma trận quan hệ trên các thuộc tính điều kiện $M(c_1)$, $M(c_2)$, $M(c_3)$, $M(c_4)$, $M(c_5)$, $M(c_6)$. Từ đó, tính ma trận $M(C)$:

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

Ta có $U/\{d\}=\{\{u_1,u_3,u_6\},\{u_2,u_4,u_5\}\}$. Xét $X=\{u_1,u_3,u_6\}$, xấp xỉ mờ dưới $\underline{C}(X)$ là tập mờ với hàm thuộc của $x\in U$ tính bởi

$$\mu_{\underline{C}(\{u_1,u_3,u_6\})}(x)=\inf_{y\in U}\max\left(1-\mu_{[x]_C}(y),\mu_{\{u_1,u_3,u_6\}}(y)\right)$$

Từ ma trận $M(C)$ ta có

$$[u_1]_C = \frac{1}{u_1} + \frac{0}{u_2} + \frac{0}{u_3} + \frac{0}{u_4} + \frac{0}{u_5} + \frac{0}{u_6}$$

Do đó $\mu_{\underline{C}(\{u_1,u_3,u_6\})}(u_1) = \inf\{1,1,1,1,1,1\} = 1$, tương tự ta có $\mu_{\underline{C}(\{u_1,u_3,u_6\})}(u_2) = 0$, $\mu_{\underline{C}(\{u_1,u_3,u_6\})}(u_3) = 1$, $\mu_{\underline{C}(\{u_1,u_3,u_6\})}(u_4) = 0$, $\mu_{\underline{C}(\{u_1,u_3,u_6\})}(u_5) = 0$, $\mu_{\underline{C}(\{u_1,u_3,u_6\})}(u_6) = 1$, $\mu_{\underline{C}(\{u_2,u_4,u_5\})}(u_1) = 0$, $\mu_{\underline{C}(\{u_2,u_4,u_5\})}(u_2) = 1$, $\mu_{\underline{C}(\{u_2,u_4,u_5\})}(u_3) = 0$, $\mu_{\underline{C}(\{u_2,u_4,u_5\})}(u_4) = 1$ $\mu_{\underline{C}(\{u_2,u_4,u_5\})}(u_5) = 1$, $\mu_{\underline{C}(\{u_2,u_4,u_5\})}(u_6) = 0$.

Từ đó, hàm thuộc của các đối tượng đối với miền dương mờ $POS_C(\{d\})$ là: $\mu_{POS_C(\{d\})}(u_1) = \sup\left\{\mu_{\underline{C}(\{u_1,u_3,u_6\})}(u_1), \mu_{\underline{C}(\{u_2,u_4,u_5\})}(u_1)\right\} = 1$, $\mu_{POS_C(\{d\})}(u_2) = 1$, $\mu_{POS_C(\{d\})}(u_3) = 1$, $\mu_{POS_C(\{d\})}(u_4) = 1$, $\mu_{POS_C(\{d\})}(u_5) = 1$, $\mu_{POS_C(\{d\})}(u_6) = 1$. Từ đó: $\gamma_C(\{d\}) = 1$

Áp dụng các bước của Thuật toán F_RSAR ta có $\gamma_{c_1}(\{d\}) = 0.167$, $\gamma_{c_2}(\{d\}) = 0$, $\gamma_{c_3}(\{d\}) = 0.167$, $\gamma_{c_4}(\{d\}) = 0.5$, $\gamma_{c_5}(\{d\}) = 0.467$, $\gamma_{c_6}(\{d\}) = 0.467$. Chọn thuộc tính c_4 có độ quan trọng lớn nhất và $P = \{c_4\}$. Thực hiện vòng lặp While. Xét các thuộc tính c_1 ta có:

$SIG_{c_4}(c_1) = \gamma_{c_4,c_1}(\{d\}) - \gamma_{c_4}(\{d\}) = 1 - 0.5 = 0.5$. Tương tự $SIG_{c_4}(c_2) = 0.5$, $SIG_{c_4}(c_3) = 0$, $SIG_{c_4}(c_5) = 0.5$, $SIG_{c_4}(c_6) = 0.5$. Không mất tính tổng quát, chọn thuộc tính c_1 có độ quan trọng lớn nhất và $P = \{c_1, c_4\}$. Khi đó ta có $\gamma_{c_1,c_4}(\{d\}) = 1 = \gamma_C(\{d\})$, do đó thuật toán dừng và $P = \{c_1, c_4\}$ là một tập rút gọn tốt nhất của bảng quyết định *DS*.

Thuật toán F_RSAR tìm được một tập rút gọn của bảng quyết định dựa trên độ quan trọng của thuộc tính (đặc trưng cho chất lượng phân lớp của thuộc tính) nên hiệu quả hơn các thuật toán tìm tất cả các tập rút gọn theo hướng tiếp cận ma trận phân biệt trong các công trình [1, 2, 3, 21, 22, 23, 25]. Thuật toán F_RSAR sử dụng độ phụ thuộc của thuộc tính để tìm tập rút gọn, do đó có khối lượng tính toán nhỏ hơn các thuật toán theo hướng tiếp cận entropy Shannon trong [8, 17]. Dễ thấy rằng, tập rút gọn thu được của Thuật toán F_RSAR bảo toàn miền dương mờ. Phần tiếp theo, chúng tôi tiến hành thử nghiệm phương pháp đề xuất trên một số bộ dữ liệu thử nghiệm để làm rõ hai vấn đề sau: 1) Tính hiệu quả của hướng tiếp cận tập thô mờ so với hướng tiếp cận tập thô truyền thống về độ chính xác phân lớp sau khi rút gọn thuộc tính; 2) Tính hiệu quả của thuật toán đề xuất với thuật toán trong công trình [8] về độ chính xác phân lớp.

VI. KẾT QUẢ THỰC NGHIỆM

Chúng tôi chọn 8 bộ dữ liệu mẫu từ lấy từ kho dữ liệu UCI [26] có miền trị thuộc tính nhận giá trị số cho ở Bảng 2 để tiến hành thử nghiệm. Môi trường thử nghiệm là máy tính PC với cấu hình Pentium dual core 2.13 GHz CPU, 1GB bộ nhớ RAM, sử dụng hệ điều hành Windows 7.

Bảng 2. Bộ số liệu thử nghiệm

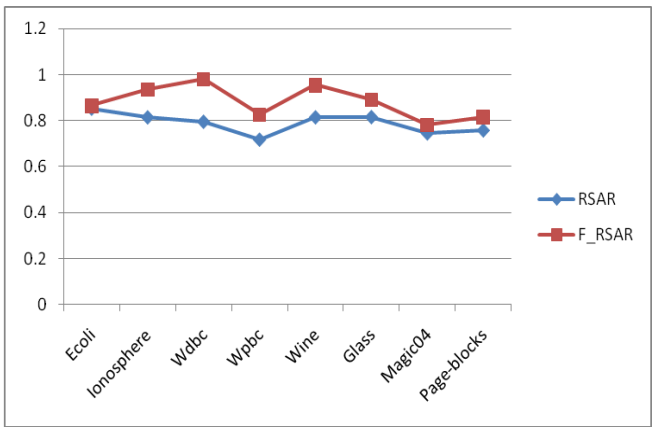
TT	Bộ dữ liệu	Số thuộc tính điều kiện	Số đối tượng
1	Ecoli	7	336
2	Ionosphere	34	351
3	Wdbc	30	569
4	Wpbc	33	198
5	Wine	13	178
6	Glass	9	214
7	Magic04	10	19020
8	Page-blocks	10	5473

Trước hết, chúng tôi tiến hành thử nghiệm nhằm đánh giá độ chính xác phân lớp trên các bộ số liệu

mẫu sau khi thực hiện thuật toán F_RSAR và thuật toán rút gọn thuộc tính sử dụng độ phụ thuộc của thuộc tính trong lý thuyết tập thô truyền thống (gọi tắt là thuật toán RSAR). Để tiến hành thử nghiệm, chúng tôi thực hiện các công việc sau:

- Cài đặt thuật toán rời rạc hóa dữ liệu bằng phương pháp equal-width [12], thuật toán RSAR, thuật toán F_RSAR sử dụng quan hệ tương tự trong [8], thuật toán phân lớp SVM [9], C4.5 [10] bằng công cụ Java.
- Thực hiện thuật toán rời rạc hóa equal-width và thuật toán RSAR để tìm tập rút gọn theo tiếp cận tập thô.
- Thực hiện thuật toán F_RSAR để tìm tập rút gọn trực tiếp từ bảng quyết định ban đầu theo tiếp cận tập thô mờ sử dụng quan hệ tương tự ở công thức (8) trong công trình [8]
- Trên bảng quyết định thu được của hai cách tiếp cận, chọn 2/3 đối tượng đầu tiên để làm tập huấn luyện (training), 1/3 đối tượng còn lại làm tập kiểm tra (test). Thực hiện thuật toán SVM, C4.5 trên tập huấn luyện và đánh giá độ chính xác phân lớp trên tập kiểm tra. Từ đó, đánh giá độ chính xác phân lớp của hai cách tiếp cận.

Bảng 3 là kết quả thử nghiệm trên 8 bộ số liệu được chọn với $|U|$ là số đối tượng, $|C|$ là số thuộc tính điều kiện, $|R|$ là số thuộc tính của tập rút gọn.



Hình 1. Độ chính xác phân lớp F_RSAR và RSAR

Bảng 3. Kết quả thử nghiệm rút gọn thuộc tính theo tiếp cận tập thô và tập thô mờ

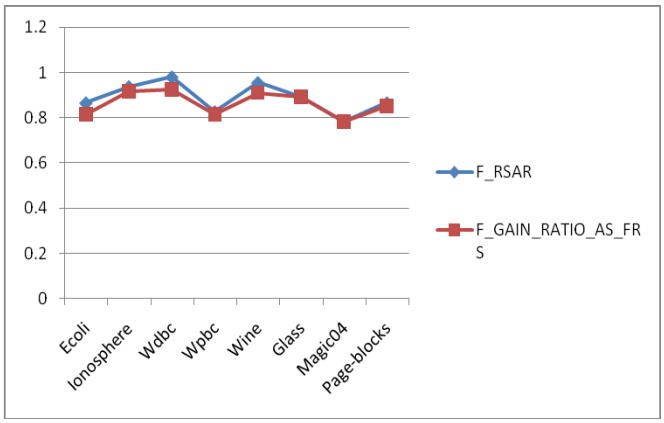
TT	Bộ số liệu	$\|U\|$	$\|C\|$	Rút gọn thuộc tính theo tiếp cận tập thô (RSAR)			Rút gọn thuộc tính theo tiếp cận tập thô mờ (F_RSAR)		
				$\|R\|$	Độ chính xác phân lớp SVM	Độ chính xác phân lớp C4.5	$\|R\|$	Độ chính xác phân lớp SVM	Độ chính xác phân lớp C4.5
1	Ecoli	336			0.851	0.819	7	0.865	0.855
2	Ionosphere	351	4	0	0.814	0.802	15	0.937	0.915
3	Wdbc	569	0		0.795	0.784	19	0.980	0.975
4	Wpbc	198	3		0.718	0.704	19	0.825	0.818
5	Wine	178	3		0.814	0.802	10	0.955	0.920
6	Glass	214			0.815	0.795	7	0.891	0.882
7	Magic04	19020	0		0.745	0.715	6	0.782	0.765
8	Page-blocks	5473	0		0.758	0.725	7	0.865	0.855

Từ Bảng 3 và Hình 1 ta thấy, tập rút gọn của F_RSAR nhiều thuộc tính hơn RSAR. Độ chính xác phân lớp sau khi rút gọn thuộc tính theo tiếp cận tập thô mờ (F_RSAR) cao hơn độ chính xác phân lớp theo tiếp cận tập thô truyền thống (RSAR).

Tiếp theo, chúng tôi tiến hành thử nghiệm để đánh giá thuật toán đề xuất (F_RSAR) với thuật toán tìm tập rút gọn theo tiếp cận tập thô mờ sử dụng lượng thông tin tăng thêm (information gain) dựa trên entropy Shannon, gọi là thuật toán GAIN_RATIO_AS_FRS [8]. Để tiến hành thử nghiệm, chúng tôi cài đặt thuật toán GAIN_RATIO_AS_FRS trong [8] và thuật toán F_RSAR. Cả hai thuật toán đều dùng quan hệ tương tự ở công thức (8) trong công trình [8]. Chạy 02 thuật toán trên 8 bộ dữ liệu thử nghiệm. Kết quả thử nghiệm cho ở Bảng 4.

Bảng 4. Kết quả thử nghiệm thuật toán GAIN_RATIO_AS_FRS và thuật toán F_RSAR

TT	Bộ số liệu	$\|U\|$	$\|C\|$	Thuật toán GAIN_RATIO_AS_FRS			Thuật toán F_RSAR		
				$\|R\|$	Độ chính xác phân lớp SVM	Độ chính xác phân lớp C4.5	$\|R\|$	Độ chính xác phân lớp SVM	Độ chính xác phân lớp C4.5
1	Ecoli	336	7	6	0.814	0.802	7	0.865	0.855
2	Ionosphere	351	34	13	0.916	0.904	15	0.937	0.915
3	Wdbc	569	30	17	0.925	0.917	19	0.980	0.975
4	Wpbc	198	33	17	0.815	0.804	19	0.825	0.818
5	Wine	178	13	9	0.910	0.902	10	0.955	0.920
6	Glass	214	9	7	0.891	0.882	7	0.891	0.882
7	Magic04	19020	10	6	0.782	0.765	6	0.782	0.765
8	Page-blocks	5473	10	6	0.852	0.848	7	0.865	0.855



Hình 2. Độ chính xác phân lớp của GAIN_RATIO_AS_FRS và F_RSAR

Từ Bảng 4 và Hình 2 ta thấy, trên cùng một quan hệ tương tự được sử dụng, độ chính xác phân lớp sau khi thực hiện thuật toán đề xuất F_RSAR cao hơn độ chính xác phân lớp sau khi thực hiện thuật toán GAIN_RATIO_AS_FRS trong [8]. Bảng 4 cũng cho thấy, tập rút gọn của thuật toán đề xuất F_RSAR bảo toàn miền dương mờ và nhiều thuộc tính hơn so với thuật toán GAIN_RATIO_AS_FRS trong [8].

V. KẾT LUẬN

Mô hình tập thô mờ do D. Dubois và các cộng sự [4, 5] 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 trị thuộc tính nhận giá trị số. Trong bài báo này, chúng tôi đề xuất thuật toán heuristic tìm một tập rút gọn của bảng quyết định có miền trị thuộc tính nhận giá trị số sử dụng độ phụ thuộc mờ của thuộc tính theo tiếp cận tập thô mờ.

Độ phụ thuộc mờ của thuộc tính được xác định dựa trên ma trận quan hệ sinh bởi một quan hệ tương tự xác định trên miền giá trị thuộc tính. Thực nghiệm trên các bộ số liệu UCI cho thấy, độ chính xác phân lớp của tập dữ liệu sau khi thực hiện phương pháp đề xuất cao hơn độ chính xác phân lớp sau khi thực hiện rút gọn thuộc tính theo tiếp cận tập thô truyền thống.

Hơn nữa, phương pháp đề xuất có độ chính xác phân lớp cao hơn phương pháp tiếp cận dựa trên entropy Shannon trong công trình [8]. Mặt khác, phương pháp đề xuất không phải tính các công thức logarit của entropy Shannon nên thời gian thực hiện hiệu quả hơn phương pháp trong [8].

Về định hướng nghiên cứu tiếp theo, thứ nhất là tìm kiếm các độ đo hiệu quả để giải quyết bài toán rút gọn thuộc tính theo tiếp cận tập thô mờ nhằm nâng cao độ chính xác phân lớp, thứ hai là tìm kiếm các quan hệ tương tự khác nhằm nâng cao độ chính xác phân lớp sau khi rút gọn thuộc tính.

TÀI LIỆU THAM KHẢO

[1] CHEN, D.G., TSANG E.C.C. and ZHAO, S.Y, *An approach of attributes reduction based on fuzzy TL rough sets*, IEEE International Conference on Systems, Man and Cybernetics, pp. 486-491, 2007.

[2] CHEN D.G, ZHAO S.Y., *Local reduction of decision system with fuzzy rough sets*, Fuzzy Sets and Systems 161, pp. 1871-1883, 2010.

[3] CHEN D.G, LEI Z, SUYUN Z, QINGHUA H, and PENGFEI Z, *A Novel Algorithm for Finding Reducts With Fuzzy Rough Sets*, IEEE Transaction on Fuzzy Systems, Vol. 20, No. 2, pp. 385 - 389 , 2012.

[4] DUBOIS D, PRADE H, *Putting rough sets and fuzzy sets together*, Intelligent Decision Support, Kluwer Academic Publishers,Dordrecht, 1992.

[5] DUBOIS D, PRADE H, *Rough fuzzy sets and fuzzy rough sets*, International Journal of General Systems, 17, pp. 191-209, 1990.

[6] F. F. XU, D. Q. MIAO and L. WEI, *An Approach for Fuzzy-Rough Sets Attributes Reduction via Mutual Information*, Fourth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2007, Volume 3, pp. 107 – 112, 2007.

[7] F.F. XU, D.Q. MIAO and L. WEI, *Fuzzy-rough attribute reduction via mutual information with an application to cancer classification*, Computers and Mathematics with Applications 57, pp. 1010 -1017, 2009.

[8] J. DAI, QING X, *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.

[9] J. NEUMANN, C. SCHNORR, G. STEILD, *Combined SVM-based feature selection and classification*, Mach. Learn. 61 (2005), pp. 129-150.

[10] J. QUINLAN, *C4.5: Programs For Machine Learning*, Morgan kaufmann, 1993.

[11] L. A. ZADEH, *Fuzzy sets*, Information and Control, 8:338-353, 1965.

[12] M.R. CHMIELEWSKI, J.W. GRZYMALABUSSE, *Global discretization of continuous attributes as preprocessing for machine learning*, Int. J. Approx. reasoning 15 (4), 1996, pp. 319–331.

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

[14] PAWLAK Z., *Rough sets: Theoretical Aspects of Reasoning About Data*, Kluwer Aca-demic Publishers, 1991.

[15] Q. SHEN, R. JENSEN, *Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring*, Pattern Recognition, Volume 37, Issue 7, pp. 1351–1363, 2004.

[16] QIANG HE, CONGXIN WU, DEGANG CHEN, SUYUN ZHAO, *Fuzzy rough set based attribute reduction for information systems with fuzzy decisions*, Knowledge-Based Systems 24 (2011), pp. 689–696, 2011.

[17] QINGHUA HU, DAREN YU, ZONGXIA XIE, *Information-preserving hybrid data reduction based on fuzzy-rough techniques*, Pattern Recognition Letters 27, 2006, pp. 414-423.

[18] R. JENSEN, Q. SHEN, *Fuzzy-Rough Sets for Descriptive Dimensionality Reduction*, Proceedings of the 2002 IEEE International Conference on Fuzzy Systems ,FUZZ-IEEE'02, pp. 29-34, 2002.

[19] R. JENSEN, Q. SHEN, *Fuzzy–rough attribute reduction with application to web categorization*, Fuzzy Sets and Systems, Volume 141, Issue 3, pp. 469-485,2004.

[20] RAJEN, B. BHATT, M. GOPAL., *On fuzzy-rough sets approach to feature selection*, Pattern Recognition Letters 26, pp. 965–975, 2005.

[21] TSANG G.C.Y., CHEN DEGANG., TSANG E.C.C, LEE J.W.T and DANIEL S. YEUNGA, *On attributes reduction with fuzzy rough sets*, Proceedings of 2005 IEEE International Conference on Systems, Man and Cybernetics ,Volume 3, pp. 2775 - 2780, 2005.

[22] TSANG E.C.C, DE GANG CHEN, *The Fuzzy Rough Set Approaches of Fuzzy Reasoning*, Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, pp. 1642-1646, 2006.

[23] TSANG E.C.C, DEGANG CHEN. YEUNG D.S., XI ZHAO WANG and JOHN W. T. LEE, *Attributes Reduction Using Fuzzy Rough Sets*, IEEE Transactions on Fuzzy Systems, Volume16, Issue 5 , pp. 1130 - 1141, 2008.

[24] YI CHENG, *Forward approximation and backward approximation in fuzzy rough sets*, Neurocomputing, Volume 148, pp. 340-353, 2015.

[25] ZHAO MING, YAN ZHENGBO, ZHOU LIUKUN, WANG HUIJIE and XU XIAOGANG, *The Extraction Method of the Energy Consumption Characteristics Based on Fuzzy Rough Set*, Proceedings of Conference on Computational Intelligence and Bioinformatics (AASRI), pp. 142 – 149, 2012.

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

Ngày nhận bài: 29/02/2016

SƠ LƯỢC VỀ TÁC GIẢ

NGUYỄN VĂN THIỆN

Sinh năm 1970 tại Phú Thọ.

Tốt nghiệp ĐH Bách khoa Hà Nội năm 1996. Tốt nghiệp Thạc sỹ tại trường ĐH Sư phạm Hà Nội năm 2000.

Hiện đang công tác tại : Trường ĐH Công nghiệp Hà Nội.

Hướng nghiên cứu: Hệ thống thông tin, Cơ sở dữ liệu, Khai phá dữ liệu.

Điện thoại: 0902416668.

Email: nvthien1970@gmail.com

NGUYỄN LONG GIANG

Sinh năm 1975 tại Hà Nội.

Tốt nghiệp ĐH Bách khoa Hà Nội năm 1997, thạc sĩ CNTT tại ĐH Công nghệ, ĐH Quốc gia Hà Nội năm 2003. Nhận bằng tiến sỹ tại Viện CNTT, Viện Hàn lâm KH&CN Việt Nam năm 2012.

Hiện công tác tại: Viện CNTT, Viện Hàn lâm KH&CN Việt Nam.

Hướng nghiên cứu: Cơ sở dữ liệu, khai phá dữ liệu và học máy.

Điện thoại: 0904739189.

Email: nlgiang@ioit.ac.vn

NGUYỄN NHƯ SƠN

Sinh năm 1974 tại Nghệ An.

Tốt nghiệp ĐH Bách khoa Hà Nội năm 1995, thạc sĩ CNTT tại trường ĐH Bách khoa Hà Nội năm 2001. Nhận bằng tiến sỹ tại ĐH Queensland - Australia năm 2007, chuyên ngành Khoa học máy tính.

Hiện công tác tại: Viện CNTT, Viện Hàn lâm KH&CN Việt Nam.

Hướng nghiên cứu: Hệ thống thông tin, Cơ sở dữ liệu, Khai phá dữ liệu, Tính toán đám mây.

Điện thoại: 0987039966.

Email: nnson@ioit.ac.vn