Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.000101

VỀ PHƯƠNG PHÁP RÚT GỌN THUỘC TÍNH TRỰC TIẾP TRÊN BẢNG QUYẾT ĐỊNH SỬ DỤNG KHOẢNG CÁCH MỜ

Nguyễn Long Giang[1], Nguyễn Văn Thiện[2], Cao Chính Nghĩa[3]

[1] Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam
[2] Trường Đại học Công nghiệp Hà Nội
[3]Học viện Cảnh sát nhân dân, Bộ Công an

nlgiang@ioit.ac.vn, nguyenthien@haui.edu.vn, ccnghia@gmail.com

TÓM TẮT *— Các phương pháp rút gọn thuộc tính theo tiếp cận lý thuyết tập thô truyền thống đều thực hiện trên các bảng quyết định có miền giá trị rời rạc, là bảng quyết định thu được sau khi thực hiện các phương pháp rời rạc hóa dữ liệu. Để giải quyết bài toán rút gọn thuộc tính trực tiếp trên các bảng quyết định có miền giá trị thực, liên tục, trong mấy năm gần đây các nhà nghiên cứu đã đề xuất một số phương pháp theo tiếp cận lý thuyết tập thô mờ. Trong bài báo này, chúng tôi đề xuất phương pháp rút gọn thuộc tính trực tiếp trên bảng quyết định có miền giá trị thực, liên tục sử dụng độ đo khoảng cách mờ. Kết quả thực nghiệm cho thấy, độ chính xác phân lớp của phương pháp đề xuất hiệu quả hơn một số phương pháp sử dụng miền dương mờ và entropy mờ.*

Từ khóa *— Tập thô mờ, quan hệ tương đương mờ, khoảng cách mờ, bảng quyết định, rút gọn thuộc tính, tập rút gọn.*

I. MỞ ĐẦU

Rút gọn thuộc tính là bài toán quan trọng của bước tiền xử lý số liệu trong quá trình khai phá dữ liệu, phát hiện tri thức. Mục tiêu của rút gọn thuộc tính 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 do Pawlak đề xuất [12, 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. Các phương pháp rút gọn thuộc tính theo tiếp cận lý thuyết tập thô đều thực hiện trên các bảng quyết định có miền giá trị rời rạc. Trong thực tế, miền giá trị thuộc tính của các bảng quyết định thường chứa giá trị thực, liên tục. Ví dụ, thuộc tính trọng lượng cơ thể và huyết áp trong bảng dữ liệu bệnh nhân thường là các giá trị thực, liên tục. Để thực hiện các phương pháp rút gọn thuộc tính theo tiếp cận tập thô, miền giá trị thuộc tính liên tục cần được rời rạc hóa. Tuy nhiên, các phương pháp rời rạc hóa không bảo toàn sự khác nhau ban đầu giữa các đối tượng trong dữ liệu gốc và do đó có khả năng làm giảm độ chính xác phân lớp sau khi rút gọn thuộc tính. Để giải quyết bài toán rút gọn thuộc tính trực tiếp trên các bảng quyết định có miền giá trị thực, liên tục, trong mấy năm gần đây các nhà 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 [1] đề xuất là sự kết hợp của lý thuyết tập thô và lý thuyết tập mờ nhằm xấp xỉ các tập mờ dựa trên một quan hệ tương đương mờ (fuzzy equivalent relation) được xác định trên miền giá trị thuộc tính. Lý thuyết tập thô truyền thống dựa trên quan hệ tương đương để xấp xỉ tập hợp, trong đó độ tương tự của hai đối tượng là 1 nếu chúng tương đương, ngược lại là 0 nếu chúng không tương đương. Lý thuyết tập thô mờ sử dụng quan hệ tương đương mờ thay thế quan hệ tương đương, độ tương tự của hai đối tượng là một giá trị nằm trong khoảng [0, 1] cho thấy tính gần nhau, hay khả năng phân biệt giữa hai đối tượng. Do đó, quan hệ tương đương mờ bảo toàn sự khác nhau, hay độ tương tự, giữa các đối tượng và các phương pháp rút gọn thuộc tính theo tiếp cận tập thô mờ có tiềm năng trong việc bảo toàn độ chính xác phân lớp sau khi thực hiện các phương pháp rút gọn thuộc tính.

Chủ đề nghiên cứu về rút gọn thuộc tính theo tiếp cận tập thô mờ đã thu hút sự quan tâm của các nhà nghiên cứu trong mấy năm gần đây [2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. Với bài toán rút gọn thuộc tính trực tiếp trên bảng quyết định theo tiếp cận tập thô mờ, các nghiên cứu liên quan tập trung vào hai hướng tiếp cận chính: hướng tiếp cận miền dương mờ và hướng tiếp cận entropy mờ. Theo hướng tiếp cận miền dương mờ, Hu và các cộng sự [5] đề xuất thuật toán FAR-VPFRS tìm tập rút gọn miền dương mờ sử dụng hàm thuộc mờ. Thực nghiệm trên một số bộ số liệu mẫu cho thấy, độ chính xác phân lớp của thuật toán FAR-VPFRS cao hơn độ chính xác phân lớp của thuật toán sử dụng hàm thuộc theo tiếp cận lý thuyết tập thô truyền thống. Qian và các cộng sự [11] đề xuất thuật toán FA_FPR, là cải tiến của thuật toán FAR-VPFRS [5] về thời gian thực hiện. Theo hướng tiếp cận entropy mờ, Hu và các cộng sự [4] đề xuất entropy mờ dựa trên entropy Shannon và xây dựng thuật toán FSCE tìm tập rút gọn sử dụng entropy mờ. Dai và các cộng sự [3] xây dựng độ đo lượng thông tin tăng thêm mờ (fuzzy gain ratio) dựa trên entropy mờ và xây dựng thuật toán GAIN_RATION_AS_FRS tìm tập rút gọn sử dụng lượng thông tin tăng thêm mờ. Thực nghiệm trên một số bộ số liệu mẫu cho thấy, độ chính xác phân lớp của các thuật toán FSCE, GAIN_RATION_AS_FRS cao hơn độ chính xác phân lớp của các thuật toán sử dụng entropy, lượng thông tin tăng thêm (gain ratio) theo tiếp cận tập thô truyền thống. Qian và các cộng sự [11] đề xuất thuật toán FA_FSCE, là cải tiến của thuật toán FSCE [4] về thời gian thực hiện. Trong cả hai hướng tiếp cận, các tác giả trong [11] chưa đánh giá độ chính xác của mô hình phân lớp sau khi thực hiện các thuật toán cải tiến FA_FPR, FA_FSCE. Với bài toán rút gọn thuộc tính trực tiếp trên bảng quyết định miền giá trị thực theo tiếp cận tập thô mờ, mục tiêu của bài báo là đề xuất thuật toán mới nhằm nâng cao độ chính xác của mô hình phân lớp so với các thuật toán đã công bố.

Trong bài báo này, chúng tôi đề xuất thuật toán rút gọn thuộc tính trên bảng quyết định miền giá trị thực sử dụng khoảng cách mờ. Khoảng cách mờ giữa hai tập thuộc tính được xây dựng dựa trên khoảng cách mờ giữa hai tập mờ. Kết quả thực nghiệm trên một số bộ số liệu mẫu cho thấy, thuật toán đề xuất cải thiện độ chính xác của mô hình phân lớp so với các thuật toán FA_FSCE và FA_FSCE [11]. Cấu trúc bài báo như sau. Phần II trình bày một số khái niệm cơ bản trong lý thuyết tập thô mờ. Phần III trình bày phương pháp xây dựng khoảng cách mờ giữa hai tập thuộc tính. Phần IV trình bày phương pháp rút gọn thuộc tính sử dụng độ đo khoảng cách mờ. Phần V trình bày kết quả thử nghiệm. Cuối cùng là kết luận và hướng phát triển tiếp theo.

II. MỘT SỐ KHÁI NIỆM CƠ BẢN

Trong phần này, chúng tôi trình bày một số vấn đề về lý thuyết tập thô, tập thô mờ và một số khái niệm liên quan đến không gian phân hoạch mờ.

Bảng quyết định là một cặp $DS = (U, C \cup D)$ trong đó U là tập hữu hạn, khác rỗng các đối tượng; C là tập thuộc tính điều kiện, D là tập thuộc tính quyết định với $C \cap D = \varnothing$. DS được gọi là bảng quyết định miền giá trị thực nếu với mọi $c \in C$, miền giá trị của c là số thực.

Lý thuyết tập thô truyền thống của Pawlak [12] sử dụng quan hệ tương đương để xấp xỉ tập hợp. Mỗi tập con thuộc tính $P \subseteq C$ xác định một quan hệ tương đương trên miền giá trị thuộc tính, ký hiệu là $IND(P)$.

$$IND(P) = \{(u,v) \in U \times U \mid \forall a \in P, a(u) = a(v)\}$$

Ký hiệu $a(v)$ là giá trị thuộc tính a tại đối tượng v. Quan hệ $IND(P)$ xác định một phân hoạch trên U, ký hiệu là $U / IND(P)$ và lớp tương đương của đối tượng u ký hiệu là $[u]_P$. Tập xấp xỉ dưới và xấp xỉ trên của $X \subseteq U$ đối với $P \subseteq C$ được định nghĩa $\underline{P}X = \{u \in U \mid [u]_P \subseteq X\}$ và $\overline{P}X = \{u \in U \mid [u]_P \cap X \neq \varnothing\}$.

Lý thuyết tập thô mờ do D. Dubois và các cộng sự [1] đề xuất sử dụng quan hệ tương đương mờ để xấp xỉ các tập mờ. Xét bảng quyết định miền giá trị thực $DS = (U, C \cup D)$, một quan hệ R xác định trên miền giá trị thuộc tính được gọi là quan hệ tương đương mờ nếu thỏa mãn các điều kiện:

1) Tính phản xạ (reflexive): $R(x,x) = 1$;

2) Tính đối xứng (symetric): $R(x,y) = R(y,x)$;

3) Tính bắc cầu max-min (max-min transitive): $R(x,z) \geq \min\{R(x,y), R(y,z)\}$) với mọi $x, y, z \in U$.

Cho hai quan hệ tương đương mờ R_P và R_Q xác định trên tập thuộc tính P và Q, khi đó với mọi $x, y \in U$ ta có [11]:

1) $R_P = R_Q \Leftrightarrow R_P(x,y) = R_Q(x,y)$

2) $R = R_P \cup R_Q \Leftrightarrow R(x,y) = \max\{R_P(x,y), R_Q(x,y)\}$

3) $R = R_P \cap R_Q \Leftrightarrow R(x,y) = \min\{R_P(x,y), R_Q(x,y)\}$

4) $R_P \subseteq R_Q \Leftrightarrow R_P(x,y) \leq R_Q(x,y)$

Quan hệ R_P được biểu diễn bởi ma trận tương đương mờ $M(R_P) = [p_{ij}]_{n \times n}$ như sau:

$$M(R_P) = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix}$$

với $p_{ij} = R_P(x_i, x_j)$ là giá trị của quan hệ giữa hai đối tượng x_i và x_j trên tập thuộc tính P, $p_{ij} \in [0,1]$.

Cho bảng quyết định miền giá trị thực $DS=\left(U,C\cup D\right)$ và $P,Q\subseteq C$. Theo [11] ta có $R_P=\cap_{a\in P}R_a$ và $R_{P\cup Q}=R_P\cap R_Q$, nghĩa là với mọi $x,y\in U$, $R_{P\cup Q}\left(x,y\right)=\min\left\{R_P\left(x,y\right),R_Q\left(x,y\right)\right\}$. Giả sử $M\left(R_P\right)=\left[p_{\text{ij}}\right]_{n\times n}$ và $M(R_Q)=\left[q_{\text{ij}}\right]_{n\times n}$ là ma trận quan hệ của R_P, R_Q, khi đó ma trận quan hệ trên tập thuộc tính $S=P\cup Q$ là:

$$M(R_S)=M\left(R_{P\cup Q}\right)=\left[s_{\text{ij}}\right]_{n\times n} \text{ với } s_{\text{ij}}=\min\left\{p_{\text{ij}},q_{\text{ij}}\right\}$$

Với $P\subseteq C$, $U=\left\{x_1,x_2,...,x_n\right\}$, quan hệ tương đương mờ R_P xác định một phân hoạch mờ $\pi\left(P\right)=U/R_P$ trên U

$$\pi\left(R_P\right)=U/R_P=\left\{\left[x_i\right]_{R_P}\right\}_{i=1}^{n}=\left\{\left[x_1\right]_{R_P},...,\left[x_n\right]_{R_P}\right\}$$

với $\left[x_i\right]_{R_P}=p_{i1}/x_1+p_{i2}/x_2+...+p_{in}/x_n$ là một tập mờ đóng vai trò là một lớp tương đương mờ của đối tượng x_i. Hàm thuộc của các đối tượng xác định bởi $\mu_{\left[x_i\right]_{R_P}}\left(x_j\right)=\mu_{R_P}\left(x_i,x_j\right)=R_P\left(x_i,x_j\right)=p_{ij}$ với mọi $x_j\in U$. Khi đó, lực lượng của lớp đương đương mờ $\left[x_i\right]_{R_P}$ được tính bởi [11]:

$$\left|\left[x_i\right]_{R_P}\right|=\sum_{j=1}^{n}p_{ij}$$

Gọi $\mathcal{P}$ là tập tất cả các phân hoạch mờ trên U xác định bởi các quan hệ tương tự mờ trên các tập thuộc tính, khi đó $\mathcal{P}$ được gọi là một không gian phân hoạch mờ trên U. Như vậy, không gian phân hoạch mờ được xác định bởi quan hệ tương đương mờ được chọn trên miền giá trị thuộc tính. Xét phân hoạch mờ $\pi\left(R_P\right)=\left\{\left[x_1\right]_{R_P},...,\left[x_n\right]_{R_P}\right\}$ với $\left[x_i\right]_{R_P}=p_{i1}/x_1+...+p_{in}/x_n$. Trường hợp đặc biệt, nếu $p_{ij}=0$ với $i,j\le n$ thì $\left|\left[x_i\right]_{R_P}\right|=0$ và khi đó phân hoạch mờ $\pi\left(R_P\right)$ được gọi là mịn nhất, ký hiệu là $\pi\left(\omega\right)$. Khi đó $\pi\left(\omega\right)=\left\{\left[x_1\right]_{\omega},...,\left[x_n\right]_{\omega}\right\}$ với $\left[x_i\right]_{\omega}=\sum_{j=1}^{n}\omega_{ij}/x_j,\forall i,j\le n,\omega_{ij}=0$. Nếu $p_{ij}=1$ với $i,j\le n$ thì $\left|\left[x_i\right]_{R_P}\right|=\left|U\right|$ với $i\le n$ và khi đó phân hoạch mờ $\pi\left(R_P\right)$ được gọi là thô nhất, ký hiệu là $\pi\left(\delta\right)$. Khi đó $\pi\left(\delta\right)=\left(\left[x_1\right]_{\delta},...,\left[x_n\right]_{\delta}\right)$ với $\left[x_i\right]_{\delta}=\sum_{j=1}^{n}\delta_{ij}/x_j,\forall i,j\le n,\delta_{ij}=1.$

Cho $\mathcal{P}$ là một không gian phân hoạch mờ trên U, với $\pi\left(R_P\right),\pi\left(R_Q\right)\in\mathcal{P}$ ta định nghĩa một quan hệ thứ tự bộ phận $\preceq$: $\pi\left(R_P\right)\preceq\pi\left(R_Q\right)\Leftrightarrow\left[x_i\right]_{R_P}\subseteq\left[x_i\right]_{R_Q},i\le n\Leftrightarrow p_{ij}\le q_{ij},i,j\le n$, viết tắt là $R_P\preceq R_Q$. Dấu đẳng thức $\pi\left(R_P\right)=\pi\left(R_Q\right)\Leftrightarrow\left[x_i\right]_{R_P}=\left[x_i\right]_{R_Q},i\le n\Leftrightarrow p_{ij}=q_{ij},i,j\le n$, viết tắt là $R_P=R_Q$. $\pi\left(R_P\right)\prec\pi\left(R_Q\right)\Leftrightarrow\pi\left(R_P\right)\preceq\pi\left(R_Q\right)$ và $\pi\left(R_P\right)\ne\pi\left(R_Q\right)$, viết tắt là $R_P\prec R_Q$.

Ví dụ 1. Cho $U=\left\{x_1,x_2\right\}$, $\pi\left(R_P\right)=\left(\left[x_1\right]_{R_P},\left[x_2\right]_{R_P}\right)$, $\pi\left(R_Q\right)=\left(\left[x_1\right]_{R_Q},\left[x_2\right]_{R_Q}\right)$,
$\pi\left(R_S\right)=\left(\left[x_1\right]_{R_S},\left[x_2\right]_{R_S}\right)$ với $\left[x_1\right]_{R_P}=0.1/x_1+0.2/x_2$, $\left[x_2\right]_{R_P}=0.2/x_1+0.3/x_2$,
$\left[x_1\right]_{R_Q}=0.2/x_1+0.3/x_2$, $\left[x_2\right]_{R_Q}=0.3/x_1+0.4/x_2$, $\left[x_1\right]_{R_S}=0.3/x_1+0.4/x_2$,
$\left[x_2\right]_{R_S}=0.4/x_1+0.6/x_2$. Khi đó ta có:

$\left|\left[x_1\right]_{R_P}\right| = 0.1+0.2 = 0.3,\ \left|\left[x_2\right]_{R_P}\right| = 0.2+0.3 = 0.5,\ \left|\left[x_1\right]_{R_Q}\right| = 0.2+0.3 = 0.5,$

$\left|\left[x_2\right]_{R_Q}\right| = 0.3+0.4 = 0.7,\ \left|\left[x_1\right]_{R_S}\right| = 0.3+0.4 = 0.7,\ \left|\left[x_2\right]_{R_S}\right| = 0.4+0.6 = 1,$

$\left|\left[x_1\right]_{R_P} \cap \left[x_1\right]_{R_Q}\right| = 0.3,\ \left|\left[x_2\right]_{R_P} \cap \left[x_2\right]_{R_Q}\right| = 0.5,\ \left|\left[x_1\right]_{R_Q} \cap \left[x_1\right]_{R_S}\right| = 0.5,\ \left|\left[x_2\right]_{R_Q} \cap \left[x_2\right]_{R_S}\right| = 0.7,$

$\left|\left[x_1\right]_{R_P} \cap \left[x_1\right]_{R_S}\right| = 0.3,\ \left|\left[x_2\right]_{R_P} \cap \left[x_2\right]_{R_S}\right| = 0.5$

III. KHOẢNG CÁCH MỜ GIỮA HAI PHÂN HOẠCH MỜ VÀ CÁC TÍNH CHẤT

3.1. Khoảng cách mờ giữa hai tập mờ

Trước hết, trong mục này chúng tôi xây dựng một độ đo khoảng cách giữa hai tập mờ, gọi là khoảng cách mờ.

Bổ đề 1. *Cho ba số thực a, b, m với* $a \ge b$. *Khi đó ta có* $a-b \ge \min(a,m) - \min(b,m)$

Chứng minh. Dễ thấy rằng $a-b \ge \min(a,m) - \min(b,m)$ thỏa mãn với ba trường hợp: $m \ge a, b \le m < a, m < b$. Vậy Bổ đề 1 được chứng minh.

Bổ đề 2. *Cho ba tập mờ* A, B, C *trên cùng tập đối tượng U. Khi đó ta có:*

1) Nếu $A \subseteq B$ *thì* $\left|B\right| - \left|B \cap C\right| \ge \left|A\right| - \left|A \cap C\right|$.

2) Nếu $A \subseteq B$ *thì* $\left|C\right| - \left|C \cap A\right| \ge \left|C\right| - \left|C \cap B\right|$.

3) $\left|A\right| - \left|A \cap B\right| + \left|C\right| - \left|C \cap A\right| \ge \left|C\right| - \left|C \cap B\right|$

Chứng minh.

1) Vì $A \subseteq B$, với mọi $x_i \in U$ ta có $\mu_B(x_i) \ge \mu_A(x_i)$. Áp dụng Bổ đề 1 ta có:

$\mu_B(x_i) - \mu_A(x_i) \ge \min\left(\mu_B(x_i), \mu_C(x_i)\right) - \min\left(\mu_A(x_i), \mu_C(x_i)\right)$

$\Leftrightarrow \sum_{i=1}^{|U|} \mu_B(x_i) - \sum_{i=1}^{|U|} \mu_A(x_i) \ge \sum_{i=1}^{|U|} \min\left(\mu_B(x_i), \mu_C(x_i)\right) - \sum_{i=1}^{|U|} \min\left(\mu_A(x_i), \mu_C(x_i)\right)$

$\Leftrightarrow \left|B\right| - \left|A\right| \ge \left|B \cap C\right| - \left|A \cap C\right| \Leftrightarrow \left|B\right| - \left|B \cap C\right| \ge \left|A\right| - \left|A \cap C\right|$

2) Vì $A \subseteq B$, với mọi $x_i \in U$ ta có $\mu_B(x_i) \ge \mu_A(x_i)$

$\Leftrightarrow \min\left(\mu_B(x_i), \mu_C(x_i)\right) \ge \min\left(\mu_A(x_i), \mu_C(x_i)\right)$.

$\Leftrightarrow \mu_C(x_i) - \min\left(\mu_A(x_i), \mu_C(x_i)\right) \ge \mu_C(x_i) - \min\left(\mu_B(x_i), \mu_C(x_i)\right)$

$\Leftrightarrow \sum_{i=1}^{|U|} \mu_C(x_i) - \sum_{i=1}^{|U|} \min\left(\mu_A(x_i), \mu_C(x_i)\right) \ge \sum_{i=1}^{|U|} \mu_C(x_i) - \sum_{i=1}^{|U|} \min\left(\mu_B(x_i), \mu_C(x_i)\right)$

$\Leftrightarrow \left|C\right| - \left|C \cap A\right| \ge \left|C\right| - \left|C \cap B\right|$.

3) Từ $A \cap C \subseteq A$, áp dụng tính chất 1) ta có $\left|A\right| - \left|A \cap B\right| \ge \left|A \cap C\right| - \left|A \cap C \cap B\right|$ (*)

Mặt khác, từ $A \cap B \subseteq B$, áp dụng tính chất 2) ta có $\left|C\right| - \left|C \cap A \cap B\right| \ge \left|C\right| - \left|C \cap B\right|$ (**)

Từ (*) và (**) ta có:

$\left|A\right| - \left|A \cap B\right| + \left|C\right| - \left|C \cap A\right| \ge \left|A \cap C\right| - \left|A \cap C \cap B\right| + \left|C\right| - \left|C \cap A\right| =$

$= \left|C\right| - \left|A \cap B \cap C\right| \ge \left|C\right| - \left|C \cap B\right|$.

Mệnh đề 1. *Cho hai tập mờ* A,B *trên cùng tập đối tượng U. Khi đó* $d\left(A,B\right)=\left|A\right|+\left|B\right|-2\left|A\cap B\right|$ *là một độ đo khoảng cách giữa* A *và* B.

Chứng minh. Rõ ràng $\left|A\right|\geq\left|A\cap B\right|$ và $\left|B\right|\geq\left|A\cap B\right|$ nên $d\left(A,B\right)\geq 0$. Hơn nữa, $d\left(A,B\right)=d\left(B,A\right)$. Tiếp theo, ta cần chứng minh bất đẳng thức tam giác. Không mất tính chất tổng quát ta chứng minh $d\left(A,B\right)+d\left(A,C\right)\geq d\left(B,C\right)$. Theo Bổ đề 2 (phần 3) ta có:

$$\left|A\right|-\left|A\cap B\right|+\left|C\right|-\left|C\cap A\right|\geq\left|C\right|-\left|C\cap B\right| \quad (***)$$

$$\left|A\right|-\left|A\cap C\right|+\left|B\right|-\left|B\cap A\right|\geq\left|B\right|-\left|B\cap C\right| \quad (****)$$

Cộng (***) với (****), vế với vế ta được:

$$\left(\left|A\right|+\left|B\right|-2\left|A\cap B\right|\right)+\left(\left|A\right|+\left|C\right|-2\left|A\cap C\right|\right)\geq\left|B\right|+\left|C\right|-2\left|B\cap C\right|,$$ hay

$$d\left(A,B\right)+d\left(A,C\right)\geq d\left(B,C\right)$$

Từ đó, $d\left(A,B\right)$ là một khoảng cách giữa hai tập mờ A và B, gọi là khoảng cách mờ. Dựa trên khoảng cách mờ này, mục tiếp theo chúng tôi xây dựng khoảng cách giữa hai phân hoạch mờ.

3.2. Khoảng cách mờ giữa hai phân hoạch mờ và các tính chất

Định lý 1. *Xét bảng quyết định* $DS=\left(U,C\cup D\right)$ *với* $U=\left\{x_1,x_2,...,x_n\right\}$ *và* $\pi\left(R_P\right)$, $\pi\left(R_Q\right)$ *là hai phân hoạch mờ sinh bởi hai quan hệ tương đương mờ* R_P, R_Q *trên* $P,Q\subseteq C$. *Khi đó:*

$$D\left(\pi\left(R_P\right),\pi\left(R_Q\right)\right)=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|\left[x_i\right]_{R_P}\right|+\left|\left[x_i\right]_{R_Q}\right|-2\left|\left[x_i\right]_{R_P}\cap\left[x_i\right]_{R_Q}\right|}{n}\right) \quad (1)$$

là một khoảng cách mờ giữa $\pi\left(R_P\right)$ *và* $\pi\left(R_Q\right)$.

Chứng minh. Rõ ràng $D\left(\pi\left(R_P\right),\pi\left(R_Q\right)\right)\geq 0$ và $D\left(\pi\left(R_P\right),\pi\left(R_Q\right)\right)=D\left(\pi\left(R_Q\right),\pi\left(R_P\right)\right)$. Ta cần chứng minh bất đẳng thức tam giác. Không mất tính chất tổng quát, với mọi $\pi\left(R_P\right),\pi\left(R_Q\right),\pi\left(R_S\right)\in\mathcal{P}$ ta chứng minh $D\left(\pi\left(R_P\right),\pi\left(R_Q\right)\right)+D\left(\pi\left(R_P\right),\pi\left(R_S\right)\right)\geq D\left(\pi\left(R_Q\right),\pi\left(R_S\right)\right)$. Từ Mệnh đề 1, với mọi $x_i\in U$ ta có: $d\left(\left[x_i\right]_{R_P},\left[x_i\right]_{R_Q}\right)+d\left(\left[x_i\right]_{R_P},\left[x_i\right]_{R_S}\right)\geq d\left(\left[x_i\right]_{R_Q},\left[x_i\right]_{R_S}\right)$. Từ đó:

$$\begin{aligned}
&D\left(\pi\left(R_P\right),\pi\left(R_Q\right)\right)+D\left(\pi\left(R_P\right),\pi\left(R_S\right)\right)\\
&=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|\left[x_i\right]_{R_P}\right|+\left|\left[x_i\right]_{R_Q}\right|-2\left|\left[x_i\right]_{R_P}\cap\left[x_i\right]_{R_Q}\right|}{n}\right)+\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|\left[x_i\right]_{R_P}\right|+\left|\left[x_i\right]_{R_S}\right|-2\left|\left[x_i\right]_{R_P}\cap\left[x_i\right]_{R_S}\right|}{n}\right)\\
&=\frac{1}{n}\sum_{i=1}^{n}\frac{d\left(\left[x_i\right]_{R_P},\left[x_i\right]_{R_Q}\right)}{n}+\frac{1}{n}\sum_{i=1}^{n}\frac{d\left(\left[x_i\right]_{R_P},\left[x_i\right]_{R_S}\right)}{n}\geq\frac{1}{n}\sum_{i=1}^{n}\frac{d\left(\left[x_i\right]_{R_Q},\left[x_i\right]_{R_S}\right)}{n}=D\left(\pi\left(R_Q\right),\pi\left(R_S\right)\right)
\end{aligned}$$

Dễ thấy rằng, $D\left(\pi\left(R_P\right),\pi\left(R_Q\right)\right)$ đạt giá trị nhỏ nhất là 0 khi và chỉ khi $\pi\left(R_P\right)=\pi\left(R_Q\right)$ và $D\left(\pi\left(R_P\right),\pi\left(R_Q\right)\right)$ đạt giá trị lớn nhất là 1 khi và chỉ khi $\pi\left(R_P\right)=\pi\left(\omega\right)$ và $\pi\left(R_Q\right)=\pi\left(\delta\right)$ (hoặc $\pi\left(R_P\right)=\pi\left(\delta\right)$ và $\pi\left(R_Q\right)=\pi\left(\omega\right)$ Do đó, $0\le D\left(\pi\left(R_P\right),\pi\left(R_Q\right)\right)\le 1$.

Mệnh đề 2. *Cho* $\pi\left(R_P\right)\in\mathcal{P}$ *là một phân hoạch mờ trên* $\mathcal{P}$*, khi đó ta có:* $D\left(\pi\left(R_P\right),\pi\left(\delta\right)\right)+D\left(\pi\left(R_P\right),\pi\left(\omega\right)\right)=1$

Chứng minh. Giả sử $\pi\left(R_P\right)=\left\{\left[x_1\right]_{R_P},\left[x_2\right]_{R_P},...,\left[x_n\right]_{R_P}\right\}$. Khi đó $D\left(\pi\left(R_P\right),\pi\left(\omega\right)\right)=\frac{1}{n^2}\sum_{i=1}^{n}\left|\left[x_i\right]_{R_P}\right|$, $D\left(\pi\left(R_P\right),K\left(\delta\right)\right)=\frac{1}{n^2}\sum_{i=1}^{n}\left(n-\left|\left[x_i\right]_{R_P}\right|\right)$. Từ đó ta có $D\left(\pi\left(R_P\right),\pi\left(\delta\right)\right)+D\left(\pi\left(R_P\right),\pi\left(\omega\right)\right)=1$.

Ví dụ 2. Tiếp tục Ví dụ 1, theo Định lý 1 ta có $D\left(\pi\left(R_P\right),\pi\left(R_Q\right)\right)=0.1$, $D\left(\pi\left(R_Q\right),\pi\left(R_S\right)\right)=0.125$, $D\left(\pi\left(R_P\right),\pi\left(R_S\right)\right)=0.225$. Do đó:

$$D\left(\pi\left(R_P\right),\pi\left(R_Q\right)\right)+D\left(\pi\left(R_Q\right),\pi\left(R_S\right)\right)=D\left(\pi\left(R_P\right),\pi\left(R_S\right)\right)$$
$$D\left(\pi\left(R_P\right),\pi\left(R_Q\right)\right)+D\left(\pi\left(R_P\right),\pi\left(R_S\right)\right)>D\left(\pi\left(R_Q\right),\pi\left(R_S\right)\right)$$
$$D\left(\pi\left(R_Q\right),\pi\left(R_S\right)\right)+D\left(\pi\left(R_P\right),\pi\left(R_S\right)\right)>D\left(\pi\left(R_P\right),\pi\left(R_Q\right)\right)$$

IV. RÚT GỌN THUỘC TÍNH TRONG BẢNG QUYẾT ĐỊNH MIỀN GIÁ TRỊ THỰC DỰA TRÊN KHOẢNG CÁCH MỜ

Trong phần này, chúng tôi trình bày phương pháp rút gọn thuộc tính trực tiếp trên bảng quyết định miền giá trị thực sử dụng khoảng cách mờ định nghĩa giữa hai phân hoạch mờ được trình bày ở phần 3.

Cho bảng quyết định miền giá trị thực $DS=\left(U,C\cup D\right)$ với $U=\left\{x_1,x_2,...,x_n\right\}$. Trên tập thuộc tính điều kiện chúng tôi sử dụng một quan hệ tương đương mờ xác định trên miền giá trị thuộc tính. Với $p\in C$, quan hệ tương đương mờ R_p thường được sử dụng với ma trận quan hệ $M\left(R_P\right)=\left[p_{\text{ij}}\right]_{n\times n}$ được xác định như sau [3]:

$$p_{ij}=\begin{cases}1-4*\dfrac{\left|p\left(x_i\right)-p\left(x_j\right)\right|}{\left|p_{\max}-p_{\min}\right|}, & \dfrac{\left|p\left(x_i\right)-p\left(x_j\right)\right|}{\left|p_{\max}-p_{\min}\right|}\le 0.25\\ 0, \quad otherwise\end{cases} \tag{2}$$

với $p\left(x_i\right)$ là giá trị của thuộc tính *p* tại đối tượng x_i, $p_{\max}$, $p_{\min}$ tương ứng là giá trị lớn nhất, nhỏ nhất của thuộc tính *p*.

Trên tập thuộc tính quyết định chúng tôi sử dụng quan hệ tương đương $IND\left(D\right)$ với ma trận tương đương $M\left(IND\left(D\right)\right)=\left[d_{\text{ij}}\right]_{n\times n}$, $d_{ij}=1$ nếu $x_j\in\left[x_i\right]_D$ và $d_{ij}=0$ nếu $x_j\notin\left[x_i\right]_D$. Nói cách khác, lớp tương đương $\left[x_i\right]_D$ có thể xem là lớp đương đương mờ, ký hiệu là $\left[x_i\right]_D$, với hàm thuộc $\mu_{\left[x_i\right]_D}\left(x_j\right)=1$ nếu $x_j\in\left[x_i\right]_D$ và $\mu_{\left[x_i\right]_D}\left(x_j\right)=0$ nếu $x_j\notin\left[x_i\right]_D$. Khi đó, ký hiệu phân hoạch mờ $\pi\left(D\right)=\left\{\left[x_i\right]_D\right\}_{i=1}^{n}=\left\{\left[x_1\right]_D,...,\left[x_n\right]_D\right\}$.

Dựa trên các quan hệ được xác định, chúng tôi xây dựng khoảng cách mờ giữa tập thuộc tính điều kiện và tập thuộc tính quyết định. Như đã trình bày ở phần 3, mỗi tập thuộc tính $P\subseteq C$ xác định một phân hoạch mờ $\pi\left(R_P\right)$. Do đó, để đơn giản chúng tôi sử dụng khái niệm khoảng cách mờ giữa hai tập thuộc tính thay cho khái niệm khoảng cách mờ giữa hai phân hoạch mờ bởi Định nghĩa 1 sau đây.

Định nghĩa 1. Cho bảng quyết định miền giá trị thực $DS=(U,C\cup D)$ với $\pi(R_P)$, $\pi(R_Q)$ là hai phân hoạch mờ sinh bởi hai quan hệ tương đương mờ R_P, R_Q trên $P,Q\subseteq C$. Khi đó, khoảng cách mờ giữa hai tập thuộc tính P và Q, ký hiệu là $F(P,Q)$, được định nghĩa là khoảng cách mờ giữa hai phân hoạch mờ $\pi(R_P)$ và $\pi(R_Q)$, nghĩa là $F(P,Q)=D(\pi(R_P),\pi(R_Q))$.

Mệnh đề 3. *Cho bảng quyết định miền giá trị thực* $DS=(U,C\cup D)$ *với* $U=\{x_1,x_2,...,x_n\}$ *và* R *là quan hệ tương đương mờ xác định trên miền giá trị tập thuộc tính điều kiện, khi đó khoảng cách mờ giữa hai tập thuộc tính* C *và* $C\cup D$ *được xác định như sau:*

$$F(\mathrm{C},\mathrm{C}\cup\mathrm{D})=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|[x_i]_{R_C}\right|-\left|[x_i]_{R_C}\cap[x_i]_D\right|}{n}\right) \tag{3}$$

Chứng minh. Từ Định nghĩa 1 và Định lý 1 ta có:

$$F(\mathrm{C},\mathrm{C}\cup\mathrm{D})=D(\pi(R_C),\pi(R_{C\cup D}))=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|[x_i]_{R_C}\right|+\left|[x_i]_{R_{C\cup D}}\right|-2\left|[x_i]_{R_C}\cap[x_i]_{R_{C\cup D}}\right|}{n}\right)$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|[x_i]_{R_C}\right|+\left|[x_i]_{R_C}\cap[x_i]_{R_D}\right|-2\left|[x_i]_{R_C}\cap[x_i]_{R_D}\right|}{n}\right)=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|[x_i]_{R_C}\right|-\left|[x_i]_{R_C}\cap[x_i]_{R_D}\right|}{n}\right)$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|[x_i]_{R_C}\right|-\left|[x_i]_{R_C}\cap[x_i]_D\right|}{n}\right)$$

Dễ thấy rằng $0\le F(C,C\cup D)\le 1-\frac{1}{n}$. $F(C,C\cup D)=0$ khi $\pi(R_C)\preceq\pi(D)$ và $F(C,C\cup D)=1-\frac{1}{n}$ khi $\pi(R_C)=\pi(\delta)$ và $[x_i]_D=\{x_i\}$ với $1\le i\le n$.

Mệnh đề 4. *Cho bảng quyết định miền giá trị thực* $DS=(U,C\cup D)$ *với* $U=\{x_1,x_2,...,x_n\}$, $B\subseteq C$ *và* R *là quan hệ tương đương mờ xác định trên miền giá trị tập thuộc tính điều kiện. Khi đó* $F(B,B\cup D)\ge F(C,C\cup D)$.

Chứng minh: Từ $B\subseteq C$, theo [11] ta có $\pi(R_C)\preceq\pi(R_B)$, nghĩa là $[x_i]_{R_C}\subseteq[x_i]_{R_B}$ với $1\le i\le n$, suy ra $\left|[x_i]_{R_C}\right|\le\left|[x_i]_{R_B}\right|$ với $1\le i\le n$. Xét đối tượng $x_i\in U$ ta có:

$$\left|[x_i]_{R_C}\right|-\left|[x_i]_{R_C}\cap[x_i]_D\right|=\sum_{j=1}^{n}\mu_{[x_i]_{R_C}}(x_j)-\sum_{j=1}^{n}\min\left\{\mu_{[x_i]_{R_C}}(x_j),\mu_{[x_i]_D}(x_j)\right\}$$

$$\left|[x_i]_{R_B}\right|-\left|[x_i]_{R_B}\cap[x_i]_D\right|=\sum_{j=1}^{n}\mu_{[x_i]_{R_B}}(x_j)-\sum_{j=1}^{n}\min\left\{\mu_{[x_i]_{R_B}}(x_j),\mu_{[x_i]_D}(x_j)\right\}$$

(1) Với $x_j\in[x_i]_D$ ta có $\mu_{[x_i]_D}(x_j)=1$, do đó $\left|[x_i]_{R_C}\right|-\left|[x_i]_{R_C}\cap[x_i]_D\right|=0=\left|[x_i]_{R_B}\right|-\left|[x_i]_{R_B}\cap[x_i]_D\right|$

(2) Với $x_j\notin[x_i]_D$ ta có $\mu_{[x_i]_D}(x_j)=0$, do đó $\left|[x_i]_{R_C}\right|-\left|[x_i]_{R_C}\cap[x_i]_D\right|=\left|[x_i]_{R_C}\right|\le\left|[x_i]_{R_B}\right|=\left|[x_i]_{R_B}\right|-\left|[x_i]_{R_B}\cap[x_i]_D\right|$.

Từ (1), (2) ta có:

$$\left|[x_i]_{R_B}\right|-\left|[x_i]_{R_B}\cap[x_i]_D\right|\ge\left|[x_i]_{R_C}\right|-\left|[x_i]_{R_C}\cap[x_i]_D\right|$$

$$\Leftrightarrow \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|[x_i]_{R_B}\right|-\left|[x_i]_{R_B}\cap[x_i]_D\right|}{n}\right)\geq\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\left|[x_i]_{R_C}\right|-\left|[x_i]_{R_C}\cap[x_i]_D\right|}{n}\right)$$

$$\Leftrightarrow F\left(B,B\cup D\right)\geq F\left(C,C\cup D\right).$$

Dễ thấy rằng dấu đẳng thức $F\left(B,B\cup D\right)=F\left(C,C\cup D\right)$ xảy ra khi và chỉ khi $\left|[x_i]_{R_B}\right|=\left|[x_i]_{R_C}\right|$ với mọi $x_i\in U$.

Tiếp theo, chúng tôi trình bày phương pháp rút gọn thuộc tính sử dụng khoảng cách mờ trong Mệnh đề 3, bao gồm các bước: định nghĩa tập rút gọn, định nghĩa độ quan trọng của thuộc tính dựa trên khoảng cách mờ và xây dựng thuật toán heuristic tìm một tập rút gọn dựa trên độ quan trọng của thuộc tính.

Định nghĩa 2. Cho bảng quyết định miền giá trị thực $DS=\left(U,C\cup D\right)$ với $B\subseteq C$ và R là quan hệ tương đương mờ xác định trên miền giá trị tập thuộc tính điều kiện. Nếu

1) $F\left(B,B\cup D\right)=F\left(C,C\cup D\right)$

2) $\forall b\in B,\ F(\{B-\{b\}\},\{B-\{b\}\}\cup D))\neq F(C,C\cup D)$

thì B là một tập rút gọn của C dựa trên khoảng cách mờ.

Định nghĩa 3. Cho bảng quyết định miền giá trị thực $DS=\left(U,C\cup D\right)$ với $B\subset C$ và $b\in C-B$. Độ quan trọng của thuộc tính b đối với B được định nghĩa bởi

$$SIG_B\left(b\right)=F\left(B,B\cup D\right)-F\left(B\cup\{b\},B\cup\{b\}\cup D\right)$$

Từ Mệnh đề 4 ta có $SIG_B\left(b\right)\geq 0$. Độ quan trọng $SIG_B\left(b\right)$ đặc trưng cho chất lượng phân lớp của thuộc tính *b* vào thuộc tính quyết định *D* và được sử dụng làm tiêu chuẩn lựa chọn thuộc tính cho thuật toán heuristic tìm tập rút gọn sau đây.

Thuật toán NF_DBAR (New Fuzzy Distance based Attribute Reduction): Thuật toán heuristic tìm một tập rút gọn sử dụng khoảng cách mờ.

Đầu vào: Bảng quyết định miền giá trị thực $DS=\left(U,C\cup D\right)$, quan hệ tương đương mờ R

Đầu ra: Một tập rút gọn B

1. $B\leftarrow\varnothing$; $M(R_B)=[1]_{n\times n}$;
2. Tính ma trận tương đương mờ $M(R_C)$, ma trận tương đương $M(IND(D))$, khoảng cách mờ $F\left(C,C\cup D\right)$;

// Thêm dần vào B các thuộc tính có độ quan trọng lớn nhất

3. While $F\left(B,B\cup D\right)\neq F\left(C,C\cup D\right)$ do
4. Begin
5. For each $a\in C-B$ tính $SIG_B\left(a\right)=F\left(B,B\cup D\right)-F\left(B\cup\{a\},B\cup\{a\}\cup D\right)$
6. Chọn $a_m\in C-B$ sao cho $SIG_B\left(a_m\right)=\underset{a\in C-B}{Max}\{SIG_B\left(a\right)\}$;
7. $B=B\cup\{a_m\}$;
8. End;

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

9. For each $a\in B$
10. Begin
11. Tính $F\left(K(B-\{a\}),K(B-\{a\}\cup D)\right)$;

12. If $F\left(K(B-\{a\}),K(B-\{a\}\cup D\right)=F\left(K(C,C\cup D)\right)$ then $B=B-\{a\}$;

13. End;

Return B;

Ví dụ 3. Xét bảng quyết định miền giá trị thực $DS=\left(U,C\cup\{d\}\right)$ cho ở Bảng 1 với $U=\{u_1,u_2,u_3,u_4\}$, $C=\{c_1,c_2,c_3,c_4\}, D=\{d\}$, quan hệ tương đương mờ R cho ở công thức (7).

Bảng 1. Bảng quyết định miền giá trị thực

	c_1	c_2	c_3	c_4	***d***
u_1	2.5045	5.4072	1.4741	5.9308	0
u_2	1.9559	4.0554	7.6407	9.4846	1
u_3	4.3517	9.5647	3.4221	4.7597	1
u_4	2.7831	9.2830	4.8055	9.8475	1

Áp dụng các bước của thuật toán NF_DBAR tìm một tập rút gọn ta có:

Khởi tạo $B\leftarrow\varnothing$; $M(R_B)=[1]_{n\times n}$; $F\left(\varnothing,\varnothing\cup\{d\}\right)=0.375$; tính các ma trận tương đương mờ $M(R_{c_1}),M(R_{c_2}),M(R_{c_3}),M(R_{c_4}),M(R_C)$, ma trận tương đương $M(IND(\{d\}))$:

$$M(R_{c_1})=\begin{bmatrix}1 & 0.0841 & 0 & 0.5349\\ 0.0841 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0.5349 & 0 & 0 & 1\end{bmatrix},\ M(R_{c_2})=\begin{bmatrix}1 & 0.0185 & 0 & 0\\ 0.0185 & 1 & 0 & 0\\ 0 & 0 & 1 & 0.7955\\ 0 & 0 & 0.7955 & 1\end{bmatrix}$$

$$M(R_{c_3})=\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0.1026\\ 0 & 0 & 0.1026 & 1\end{bmatrix},\ M(R_{c_4})=\begin{bmatrix}1 & 0 & 0.0793 & 0\\ 0 & 1 & 0 & 0.7147\\ 0.0793 & 0 & 1 & 0\\ 0 & 0.7147 & 0 & 1\end{bmatrix},$$

$$M(R_C)=\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix},\ M(IND(\{d\}))=\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 1 & 1\\ 0 & 1 & 1 & 1\\ 0 & 1 & 1 & 1\end{bmatrix}$$

Từ đó ta có: $F\left(C,C\cup\{d\}\right)=0$, $F\left(\{c_1\},\{c_1\}\cup\{d\}\right)=0.0774$, $F\left(\{c_2\},\{c_2\}\cup\{d\}\right)=0.0023$, $F\left(\{c_3\},\{c_3\}\cup\{d\}\right)=0$, $F\left(\{c_4\},\{c_4\}\cup\{d\}\right)=0.0099$; $SIG_B(c_1)=0.2976$, $SIG_B(c_2)=0.3727$, $SIG_B(c_3)=0.375$, $SIG_B(c_4)=0.3651$. Thuộc tính $\{c_3\}$ được chọn; kiểm tra $F\left(C,C\cup\{d\}\right)=F\left(\{c_3\},\{c_3\}\cup\{d\}\right)=0$, do đó thuật toán dừng và $B=\{c_3\}$ là tập rút gọn tìm được của thuật toán.

V. THỬ NGHIỆM

Chúng tôi chọn 8 bộ dữ liệu mẫu từ lấy từ kho dữ liệu UCI [14] có miền giá trị thực cho ở Bảng 2 để tiến hành thử nghiệm. Môi trường thử nghiệm là máy tính PC với cấu hình Pentium dual core 2.13GHz CPU, 2GB bộ nhớ RAM, sử dụng hệ điều hành Windows 7.

Bảng 2. Bộ dữ liệu thử nghiệm

STT	Bộ dữ liệu	Số thuộc tính điều kiện	Số đối tượng
1	Ecoli	7	336
2	Ionosphere	34	351
3	Wdbc (Breast Cancer Wisconsin)	30	569
4	Wpbc (Breast Cancer Wisconsin)	32	198
5	Wine	13	178
6	Glass	9	214
7	Sonar (Connectionist Bench)	60	208
8	Heart	13	270

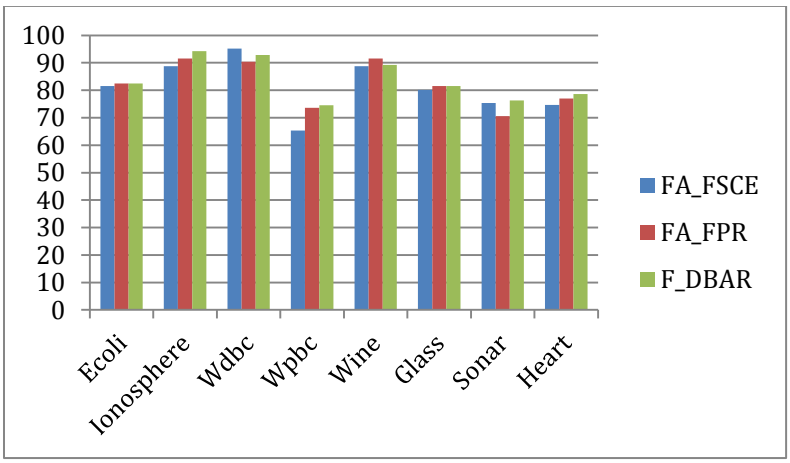
Chúng tôi chọn thuật toán FA_FPR (tìm tập rút gọn dựa trên miền dương mờ) và thuật toán FA_FSCE (tìm tập rút gọn dựa trên entropy mờ) trong công trình [11] để so sánh với thuật toán đề xuất NF_DBAR về độ chính xác phân lớp sau khi rút gọn thuộc tính. Thuật toán FA_FPR là cải tiến của thuật toán FAR-VPFRS trong [5] về thời gian thực hiện, còn thuật toán FA_FSCE là cải tiến của thuật toán FSCE trong [4] về thời gian thực hiện. Theo hướng tiếp cận tập thô mờ, độ chính xác phân lớp sau khi thực hiện các thuật toán FAR-VPFRS [5], FSCE [4] đều cao hơn so với hướng tiếp cận tập thô truyền thống sau khi rời rạc hóa dữ liệu. Tuy nhiên, trong công trình [11] tác giả chưa đánh giá độ chính xác phân lớp đối với các thuật toán cải tiến FA_FPR và FA_FSCE. Để tiến hành thử nghiệm, chúng tôi thực hiện các công việc sau:

1) Cài đặt các thuật toán FA_FPR, FA_FSCE và NF_DBAR bằng ngôn ngữ Java, các thuật toán đều sử dụng quan hệ tương đương mờ trong công thức (2).
2) Thực hiện 03 thuật toán trên 8 bộ dữ liệu mẫu với môi trường thử nghiệm được chọn.
3) Sử dụng thuật toán C4.5 trong WEKA [15] để đánh giá độ chính xác phân lớp của 03 thuật toán bằng cách chọn 2/3 đối tượng đầu tiên để làm tập huấn luyện (training set), 1/3 đối tượng còn lại làm tập kiểm tra (testing set).

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 với mỗi thuật toán.

Bảng 3. Kết quả thử nghiệm 03 thuật toán FA_FSCE, FA_FPR, NF_DBAR

STT	Bộ số liệu	$\|U\|$	$\|C\|$	Thuật toán FA_FSCE		Thuật toán FA_FPR		Thuật toán NF_DBAR	
				$\|R\|$	Độ chính xác phân lớp C4.5 (%)	$\|R\|$	Độ chính xác phân lớp C4.5 (%)	$\|R\|$	Độ chính xác phân lớp C4.5 (%)
1	Ecoli	336	7	6	81.50	7	82.45	7	82.45
2	Ionosphere	351	34	11	88.72	13	91.52	15	94.25
3	Wdbc	569	30	16	95.2	17	90.46	19	92.84
4	Wpbc	198	32	16	65.32	17	73.60	18	74.60
5	Wine	178	13	5	88.72	9	91.57	10	89.25
6	Glass	214	9	6	80.15	7	81.56	7	81.56
7	Sonar	208	60	8	75.40	12	70.60	13	76.25
8	Heart	270	13	8	74.62	9	76.95	10	78.65
Độ chính xác phân lớp trung bình C4.5					**81.2**		**82.33**		**83.73**



Hình 1. Độ chính xác phân lớp C4.5 của FA_FSCE, FA_FPR và NF_DBAR

Kết quả thử nghiệm ở Bảng 3 và Hình 1 cho thấy, trên 8 bộ dữ liệu thử nghiệm, độ chính xác phân lớp trung bình của NF_DBAR (sử dụng khoảng cách mờ) là lớn nhất, tiếp theo đến FA_FPR (sử dụng miền dương mờ) và thấp nhất là FA_FSCE (sử dụngh entropy mờ). Trên từng bộ dữ liệu cụ thể, độ chính xác phân lớp của 03 thuật toán là khác nhau, tuy nhiên về cơ bản thuật toán NF_DBAR có độ chính xác phân lớp tốt nhất trong 03 thuật toán.

VI. KẾT LUẬN

Một trong những mục tiêu của rút gọn thuộc tính trong bảng quyết định là nâng cao độ chính xác của mô hình phân lớp. Trên lớp bài toán rút gọn thuộc tính trong bảng quyết định miền giá trị thực, các nghiên cứu liên quan cho thấy các phương pháp rút gọn thuộc tính theo tiếp cận tập thô mờ có độ chính xác phân lớp cao hơn phương pháp rút gọn thuộc tính theo tiếp cận tập thô truyền thống. Trong bài báo này, chúng tôi xây dựng phương pháp rút gọn thuộc tính trực tiếp trên bảng quyết định miền giá trị thực sử dụng khoảng cách mờ theo tiếp cận tập thô mờ. Nghiên cứu của chúng tôi bao gồm các nội dung: xây dựng một khoảng cách mờ giữa hai phân hoạch mờ, định nghĩa tập rút gọn và độ quan trọng của thuộc tính dựa trên khoảng cách mờ và xây dựng thuật toán heuristic tìm một tập rút gọn. Kết quả thử nghiệm trên một số bộ dữ liệu mẫu cho thấy, độ chính xác phân lớp của phương pháp khoảng cách mờ tốt hơn độ chính xác phân lớp của các phương pháp sử dụng miền dương mờ và entropy mờ. Định hướng nghiên cứu tiếp theo là nghiên cứu mối liên hệ giữa các tập rút gọn của các phương pháp để phân nhóm và đánh giá tổng thể về các phương pháp theo tiếp cận tập thô mờ.

LỜI CẢM ƠN

Kết quả nghiên cứu này được tài trợ bởi Đề tài nghiên cứu mã số VAST01.08/16-17, cấp Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

TÀI LIỆU THAM KHẢO

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

[2] E.C.C. Tsang, D.G. Chen, D.S. Yeung, X.Z. Wang, J.W.T. Lee, Attributes reduction using fuzzy rough sets, IEEETrans. Fuzzy Syst. 16 (2008) 1130–1141.

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

[4] Q. Hu, D.R. Yu, Z.X. Xie, Information-preserving hybrid data reduction based on fuzzy-rough techniques, Pattern Recognit. Lett. 27(5) (2006) 414–423.

[5] Q. Hu, Z.X. Xie, D.R. Yu, Hybrid attribute reduction based on a novel fuzzy-rough model and information granulation, Pattern Recognit. 40 (2007) 3509–3521.

[6] R. Jensen, Q. Shen, Semantics-preserving dimensionality reduction: rough and fuzzy-rough-based approaches, IEEE Trans. Knowl. Data Eng. 16(12) (2004) 1457–1471.

[7] R. Jensen, Q. Shen, Fuzzy-rough attribute reduction with application to web categorization, Fuzzy Sets Syst. 141 (2004) 469-485.

[8] R. Jensen, Q. Shen, Fuzzy-rough sets assisted attribute reduction, IEEE Trans. Fuzzy Syst. 15(1) (2007) 73–89.

[9] R. Jensen, Q. Shen, New approaches to fuzzy-rough feature selection, IEEE Trans. Fuzzy Syst. 17(4) (2009) 824–838.

[10] R.B. Bhatt, M. Gopal, On fuzzy-rough sets approach to feature selection, Pattern Recognit. Lett. 26 (2005) 965–975.

[11] Y.H. Qian, Q. Wang, H.H. Cheng, J.Y. Liang, C.Y. Dang, Fuzzy-rough feature selection accelerator, Fuzzy Sets and Systems 258 (2015) 61–78.

[12] Z. Pawlak, Rough sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publisher, London, 1991.

[13] Z. Pawlak, J.W. Grzymala-Busse, R. Slowiski, W. Ziako, Rough sets, Commun. ACM 38(11) (1995) 89-95.

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

[15] https://sourceforge.net/projects/weka/.

FUZZY DISTANCE BASED ATTRIBUTE REDUCTION IN DECISION TABLES

Nguyen Long Giang, Nguyen Van Thien, Cao Chinh Nghia

ABSTRACT *— Traditional rough set based attribute reduction methods has performed on the decision tables with discretized value attribute domain. In recent years, many researchers has proposed some attribute reduction methods on the decision table with real attribute value domain based on fuzzy rough set. In this paper, we propose an attribute reduction method which performs directly on the decision table with real value domain using fuzzy distance. The experiment from UCI data sets showed that the accuracy classification of the proposed method is more efficient than the ones based on fuzzy positive region and fuzzy entropy.*

Keywords— *Fuzzy rough set, fuzzy equivalence relation, fuzzy distance, fuzzy decision table, attribute reduction, reduct.*