Phương pháp rút gọn thuộc tính trong bảng quyết định không đầy đủ sử dụng khoảng cách phân hoạch

Partition Distance Based Attribute Reduction in Incomplete Decision Tables

Vũ Văn Định, Vũ Đức Thi, Ngô Quốc Tạo, Nguyễn Long Giang

Abstract: *Tolerance based attribute reduction in incomplete decision tables is a hot topic which has attracted the attention of researchers in recent years. In this paper, we develop a distance based attribute reduction method in incomplete decision tables. The distance between the conditional attribute and the decision attribute has determined based on a partition distance. By theoretically and experimentally, we compare the proposed method with others methods on the time complexity and the obtained reduct.*

Keyword*: Tolerance rough set, incomplete decision table, attribute reduction, reduct, partition distance.*

I. GIỚI THIỆU

Rút gọn thuộc tính trên các hệ thông tin đầy đủ là chủ đề nghiên cứu quan trọng nhất trong lý thuyết tập thô truyền thống của Pawlak [8]. Trong thực tế, các hệ thông tin thường thiếu giá trị trên miền giá trị của thuộc tính, goi là các hệ thông tin không đầy đủ. Nhằm giải quyết bài toán rút gọn thuộc tính và khai phá luật trên các hệ thông tin đầy đủ, Kryszkiewicz [3] đã mở rộng quan hệ tương đương trong lý thuyết tập thô truyền thống thành quan hệ dung sai và xây dựng mô hình tập thô dung sai. Trong mấy năm gần đây, nhiều phương pháp rút gọn thuộc tính trong bảng quyết định không đầy đủ theo tiếp cận mô hình tâp thô dung sai đã được công bố. Mỗi phương pháp đều đưa ra khái niệm về tập rút gọn dựa trên một độ đo được chọn và xây dựng thuật toán heuristic tìm một tập rút gọn tốt nhất dựa trên tiêu chuẩn chất lượng phân lớp của thuộc tính, còn gọi là độ quan trọng của thuộc tính. Một số tập rút gọn của các phương pháp có thể kể đến là: tập rút gọn dựa trên hàm quyết định suy rộng [3], tập rút gọn miền dương [10], tập rút gọn dựa trên lượng thông tin [1], tập rút gọn dựa trên metric [5], tập rút gọn phân bố (distribution reduct), tập rút gọn ấn định (assignment reduct) [9,11], tập rút gọn dựa trên ma trận phân biệt [7], tập rút gọn dựa trên ma trận dung sai [2]. Trong công trình [7], các tác giả đã phân nhóm các phương pháp rút gọn thuộc tính dựa vào tập rút gọn và nghiên cứu mối liên hệ giữa các tập rút gọn của các phương pháp nhằm so sánh, đánh giá tính hiệu quả của các phương pháp.

Trong bài báo này chúng tôi xây dựng một phương pháp rút gọn thuộc tính trong bảng quyết định không đầy đủ sử dụng khoảng cách phân hoạch. Trước hết, chúng tôi định nghĩa một khoảng cách phân hoạch xác định bởi một tập đối tượng U và một tập thuộc tính P dựa vào khoảng cách Jaccard giữa hai tập hợp hữu hạn. Dựa trên khoảng cách phân hoạch, chúng tôi xây dựng một độ đo khoảng cách giữa một tập thuộc tính điều kiện và thuộc tính quyết định, trên cơ sở đó xây dựng phương pháp rút gọn thuộc tính sử dụng khoảng cách. Tương tự như các phương pháp heuristic khác, phương pháp của chúng tôi cũng bao gồm các bước: định nghĩa tập rút gọn dựa trên khoảng cách, định nghĩa độ quan trọng của thuộc tính dựa trên khoảng cách và xây dựng một thuật toán heuristic tìm một tập rút gọn tốt nhất dựa trên tiêu chí đánh giá là độ quan trọng của thuộc tính. Bằng lý thuyết và thực nghiệm, chúng tôi so sánh, đánh giá phương pháp sử dụng khoảng cách đề xuất với các phương pháp khác đã công bố trên hai tiêu chuẩn: độ phức tạp thời gian và tập rút gọn thu được. Cấu trúc của bài báo như sau:

Phần II trình bày một số khái niệm cơ bản về mô hình tập thô dung sai và một số kết quả về rút gọn thuộc tính trong bảng quyết định không đầy đủ. Phần III trình bày phương pháp xây dựng khoảng cách. Phần IV trình bày phương pháp rút gọn thuộc tính sử dụng khoảng cách. Phần V trình bày kết quả thử nghiệm thuật toán. 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 về mô hình tập thô dung sai [3] và một số kết quả nghiên cứu về các phương pháp rút gọn thuộc tính trong bảng quyết định không đầy đủ theo tiếp cận mô hình tập thô dung sai.

Hệ thông tin là một cặp $IS=(U,A)$ trong đó *U* là tập khác rỗng, hữu hạn 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 $a\in A$. Nếu V_a chứa *giá trị thiếu* (missing value) thì *IS* được gọi là *hệ thông tin không đầy đủ*, ngược lại là *hệ thông tin đầy đủ*, giá trị thiếu được biểu diễn là '*'. Bảng quyết định không đầy đủ là hệ thông tin không đầy đủ $IDS=(U,A\cup\{d\})$ với $d,\ d\notin A$ và $*\notin V_d$, là thuộc tính quyết định, tập thuộc tính *A* gọi là tập thuộc tính điều kiện.

Với mỗi tập con thuộc tính $P\subseteq A$, ta định nghĩa một quan hệ nhị phân trên *U* như sau:

$$SIM(P)=\left\{(u,v)\in U\times U \left| \begin{array}{l}\forall a\in P,\\ f(u,a)=f(v,a)\vee f(u,a)\\ ='*'\vee f(v,a)='*'\end{array}\right.\right\} \quad (1)$$

$SIM(P)$ là *quan hệ dung sai* (tolerance relation) trên *U* vì chúng có *tính phản xạ, đối xứng* nhưng không có *tính bắc cầu*. Dễ thấy $SIM(P)=\bigcap_{a\in P}SIM(\{a\})$. Ký hiệu $U/SIM(P)=\{S_P(u)|u\in U\}$ với $S_P(u)=\{v\in U|(u,v)\in SIM(P)\}$. $S_P(u)$ là tập các đối tượng không phân biệt được với *u* đối với quan hệ dung sai trên tập thuộc tính *P*, còn được gọi là một *lớp dung sai* hay một *hạt thông tin*. Rõ ràng các lớp dung sai trong $U/SIM(P)$ không phải là một phân hoạch của *U* mà hình thành một phủ của *U* vì chúng có thể giao nhau, nghĩa là $S_P(u)\neq\varnothing$ với mọi $u\in U$ và $\bigcup_{u\in U}S_P(u)=U$.

Với $B\subseteq A$, $X\subseteq U$, *B-xấp xỉ dưới* của X là tập $\underline{B}X=\{u\in U|S_B(u)\subseteq X\}=\{u\in X|S_B(u)\subseteq X\}$, *B-xấp xỉ trên của* X là tập $\overline{B}X=\{u\in U|S_B(u)\cap X\neq\varnothing\}=\bigcup\{S_B(u)|u\in U\}$, *B-miền biên* của *X* là tập $BN_P(X)=\overline{P}X-\underline{P}X$. Với các tập xấp xỉ như vậy, ta gọi *B-miền dương* đối với *{d}* là tập:

$$POS_B(\{d\})=\bigcup_{X\in U/\{d\}}(\underline{B}X) \quad (2)$$

Cho bảng quyết định không đầy đủ $IDS=(U,A\cup\{d\})$. Với $B\subseteq A$ và $u\in U$, $\partial_B(u)=\{f_d(v)|v\in S_B(u)\}$ được gọi là hàm quyết định suy rộng của *IDS*. Nếu $|\partial_C(u)|=1$ với mọi $u\in U$ thì *IDS* là *nhất quán*, trái lại *IDS* là *không nhất quán*. Theo định nghĩa miền dương, *IDS* nhất quán khi và chỉ khi $POS_A(\{d\})=U$, trái lại *IDS* là không nhất quán.

Kể từ khi Kryszkiewicz [3] đề xuất mô hình tập thô dung sai, nhiều phương pháp heuristic rút gọn thuộc tính trong bảng quyết định được công bố. Mỗi phương pháp đều đưa ra khái niệm về tập rút gọn dựa trên một độ đo được chọn và xây dựng thuật toán heuristic tìm một tập rút gọn tốt nhất dựa trên tiêu chuẩn chất lượng phân lớp của thuộc tính, còn gọi là độ quan trọng của thuộc tính. Các phương pháp rút gọn thuộc tính điển hình và các tập rút gọn được trình bày trong Bảng 1.

Trong công trình [7], các tác giả đã phân nhóm các tập rút gọn trong bảng quyết định không nhất quán

thành 04 nhóm theo nguyên tắc các tập rút gọn giống nhau được phân vào một nhóm:

Nhóm 1: Bao gồm tập rút gọn R_P.

Nhóm 2: Bao gồm các tập rút gọn R_∂, R_δ, R_M.

Nhóm 3: Bao gồm các tập rút gọn R_I, R_{TM}, R_H.

Nhóm 4: Bao gồm tập rút gọn R_μ.

Mối liên hệ giữa các tập rút gọn trong các nhóm như sau:

(1) Nếu R_3 là một tập rút gọn thuộc nhóm 3 thì tồn tại một tập rút gọn R_2 thuộc nhóm 2 và một tập rút gọn R_1 thuộc nhóm 1 sao cho $R_1 \subseteq R_2 \subseteq R_3$.

(2) Nếu R_4 là một tập rút gọn thuộc nhóm 4 thì tồn tại một tập rút gọn R_2 thuộc nhóm 2 và một tập rút gọn R_1 thuộc nhóm 1 sao cho $R_1 \subseteq R_2 \subseteq R_4$.

***Bảng 1**. Các phương pháp rút gọn thuộc tính và tập rút gọn*

STT	Phương pháp rút gọn thuộc tính	Ký hiệu
1	Phương pháp miền dương [10]	R_P
2	Phương pháp sử dụng hàm quyết định suy rộng [3]	R_∂
3	Phương pháp sử dụng hàm ấn định (assignment) [11]	R_δ
4	Phương pháp sử dụng ma trận phân biệt [7]	R_M
5	Phương pháp sử dụng độ đo lượng thông tin [1]	R_I
6	Phương pháp sử dụng ma trận dung sai [2]	R_{TM}
7	Phương pháp sử dụng metric [5]	R_H
8	Phương pháp sử dụng hàm phân bố (distribution) [9]	R_μ

Trên cơ sở đó, các phương pháp rút gọn thuộc tính cũng được phân thành 04 nhóm tương ứng. Trong công trình [6], các tác giả đã nghiên cứu sự thay đổi các độ đo đánh giá tập luật quyết định trên các tập rút gọn. Trên bảng quyết định không nhất quán, tập rút gọn thuộc *nhóm 2* là tốt nhất vì có số thuộc tính tối thiểu nhất.

Phần tiếp theo, chúng tôi xây dựng phương pháp rút gọn thuộc tính trong bảng quyết định không đầy đủ sử dụng một độ đo khoảng cách xác định giữa tập thuộc tính điều kiện và thuộc tính quyết định.

III. XÂY DỰNG ĐỘ ĐO KHOẢNG CÁCH TRONG BẢNG QUYẾT ĐỊNH KHÔNG ĐẦY ĐỦ

III.1. Khoảng cách phân hoạch và độ đo thông tin

Cho *U* là tập hữu hạn các đối tượng và $X, Y \subseteq U$.

Biểu thức: $$D(X,Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|}$$

được gọi là khoảng cách Jaccard (Jaccard distance) giữa hai tập hợp *X* và *Y* [4]. Dựa vào khoảng cách Jaccard, chúng tôi xây dựng khoảng cách phân hoạch.

Cho hệ thông tin $IS = (U, A)$, giả sử $K(P) = U/P = \{P_1, ..., P_k\}$ là phân hoạch sinh bởi tập thuộc tính $P \subseteq A$ và $K(\delta) = \{\delta_1, ..., \delta_k\}$ với $\delta_i = U, i = 1..k$. Khi đó, khoảng cách phân hoạch giữa $K(\delta)$ và $K(P)$, ta gọi là khoảng cách phân hoạch xác định bởi tập đối tượng *U* và tập thuộc tính *P*, được tính bằng tổng khoảng cách Jaccard trung bình giữa các phần tử tương ứng thuộc $K(\delta)$ và $K(P)$ như sau:

$$d(K(\delta), K(P)) = \frac{1}{k}\sum_{i=1}^{k}\left(1 - \frac{|U \cap P_i|}{|U \cup P_i|}\right) \quad (3)$$

Mệnh đề 1. *Cho hệ thông tin* $IS = (U, A)$ *với* $P \subseteq A$ *và* $U = \{u_1, ..., u_n\}$. *Giả sử* $K(P) = \{P_1, ..., P_k\}$, $K(\delta) = \{\delta_1, ..., \delta_k\}$ *với* $\delta_i = U, i = 1..k$. *Khi đó ta có:*

1) $d\left(K(\delta),K(P)\right)=1-\frac{1}{k}$

2) $d\left(K(\delta),K(P)\right)$ *đạt giá trị lớn nhất là* $1-\frac{1}{n}$ *khi* $K(P)=\omega=\{\{u_1\},...,\{u_n\}\}$. $d\left(K(\delta),K(P)\right)$ *đạt giá trị nhỏ nhất là 0 khi* $K(P)=\delta=\{U\}$.

Chứng minh.

1) Từ công thức (3) ta có:

$$d(K(\delta),K(P))=\frac{1}{k}\sum_{i=1}^{k}\left(1-\frac{|P_i|}{|U|}\right)=$$

$$=\frac{1}{k}\left(k-\frac{|P_1|+...+|P_k|}{|U|}\right)=\frac{k-1}{k}=1-\frac{1}{k}$$

2) Dễ thấy rằng $d\left(K(\delta),K(P)\right)$ đạt giá trị lớn nhất khi $\frac{1}{k}$ đạt giá trị nhỏ nhất, nghĩa là $k=n$ hay $K(P)=\omega=\{\{u_1\},...,\{u_n\}\}$. $d\left(K(\delta),K(P)\right)$ đạt giá trị nhỏ nhất khi $k=1$, nghĩa là $K(P)=\delta=\{U\}$.

Từ khoảng cách phân hoạch xác định bởi tập đối tượng *U* và tập thuộc tính *P* nêu trên, mục tiếp theo chúng tôi xây dựng khoảng cách giữa tập thuộc tính điều kiện *P* và thuộc tính quyết định $\{d\}$ trong bảng quyết định không đầy đủ.

III.2. Xây dựng khoảng cách trong bảng quyết định không đầy đủ

Cho bảng quyết định không đầy đủ $IDS=\left(U,A\cup\{d\}\right)$ với $U=\{u_1,...,u_n\}$ và tập thuộc tính $P\subseteq A$. Với mỗi lớp dung sai $S_P(u_i), u_i\in U$, ta ký hiệu $K_P^i\left(\{d\}\right)=S_P(u_i)/\{d\}=\left\{S_1^i,S_2^i,...,S_{k_P^i}^i\right\}$ là phân hoạch của lớp dung sai $S_P(u_i)$ trên thuộc tính quyết định $\{d\}$, và $K_P^i(\delta)=\left\{\delta_1^i,\delta_2^i,...,\delta_{k_P^i}^i\right\}$ với $\delta_j^i=S_P(u_i),\ j=1..k_P^i$. Khi đó, khoảng cách phân hoạch xác định bởi lớp dung sai $S_P(u_i)$ và thuộc tính quyết định $\{d\}$ là :

$$d\left(K_P^i(\delta),K_P^i\left(\{d\}\right)\right)=1-\frac{1}{k_P^i} \qquad (4)$$

Cho bảng quyết định không đầy đủ $IDS=\left(U,A\cup\{d\}\right)$ và $U=\{u_1,...,u_n\}$, với $P\subseteq A$ ta có $U/SIM(P)=\left\{S_P(u_i)\middle|u_i\in U,i=1..n\right\}$ là một phủ của *U*. Khi đó, ta xây dựng khoảng cách giữa tập thuộc tính điều kiện *P* và thuộc tính quyết định $\{d\}$, ký hiệu là $D\left(P,\{d\}\right)$, là trung bình cộng của các khoảng cách phân hoạch thành phần xác định bởi các lớp dung sai $S_P(u_i)$ và $\{d\}$, khoảng cách đó được định nghĩa bởi công thức (5) sau đây :

$$D(P,\{d\})=\frac{1}{n}\sum_{i=1}^{n}d\left(K_P^i(\delta),K_P^i\left(\{d\}\right)\right)=$$
$$=\frac{1}{n}\sum_{i=1}^{n}\left(1-\frac{1}{k_P^i}\right)=1-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{k_P^i}\right) \qquad (5)$$

Với *n* là số đối tượng của bảng quyết định và k_P^i là số lớp tương của phân hoạch $S_P(u_i)/\{d\}$ với $u_i\in U$

Mệnh đề 2. *Cho bảng quyết định không đầy đủ* $IDS=\left(U,A\cup\{d\}\right)$ *và* $P,Q\subseteq A$. *Nếu* $P\subseteq Q$ *thì* $D\left(P,\{d\}\right)\geq D\left(Q,\{d\}\right)$. *Dấu đẳng thức* $D\left(P,\{d\}\right)=D\left(Q,\{d\}\right)$ *khi và chỉ khi* $\partial_P(u)=\partial_Q(u)$ *với mọi* $u\in U$.

Chứng minh.

Xét bảng quyết định không đầy đủ $IDS=\left(U,A\cup\{d\}\right)$ với $U=\{u_1,...,u_n\}$. Nếu $P\subseteq Q$ thì $S_Q(u_i)\subseteq S_P(u_i)$ với mọi $u_i\in U$. Giả sử với $u_i\in U$ ta có $S_P(u_i)/\{d\}=\left\{S_1^i,S_2^i,...,S_{k_P^i}^i\right\}$,

$S_Q(u_i)/\{d\} = \{S_1^i, S_2^i, ..., S_{k_Q^i}^i\}$, khi đó rõ ràng $k_Q^i \le k_P^i$. Vì vậy, $1 - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k_P^i} \ge 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k_Q^i}$, nghĩa là $D(P,\{d\}) \ge D(Q,\{d\})$.

Dấu đẳng thức $D(P,\{d\}) = D(Q,\{d\})$ khi và chỉ khi $k_P^i = k_Q^i$ với mọi $u_i \in U$, theo định nghĩa hàm quyết định suy rộng ta có $|\partial_P(u_i)| = |\partial_Q(u_i)|$ với mọi $u_i \in U$. Từ $S_Q(u_i) \subseteq S_P(u_i)$ ta suy ra $\partial_P(u_i) = \partial_Q(u_i)$ với mọi $u_i \in U$.

Mệnh đề 2 chứng minh tính phản đơn điệu của khoảng cách đối với lực lượng của tập thuộc tính điều kiện. Nghĩa là tập thuộc tính điều kiện *P* càng nhỏ thì phủ sinh bởi *P* càng thô và khoảng cách từ *P* tới thuộc tính quyết định *{d}* càng lớn và ngược lại. Mệnh đề này rất quan trọng và cho ta cơ sở để xây dựng phương pháp rút gọn thuộc tính sử dụng khoảng cách.

Mệnh đề 3. *Cho bảng quyết định không đầy đủ* $IDS = (U, A \cup \{d\})$ *và* $P \subseteq A$. *Khi đó ta có:*

1) $D(P,\{d\})$ *đạt giá trị lớn nhất là* $1 - \frac{1}{n}$ *khi* $|\partial_P(u_i)| = n$ *với mọi* $u_i \in U$.

2) $D(P,\{d\})$ *đạt giá trị nhỏ nhất là* 0 *khi* $|\partial_P(u_i)| = 1$ *với mọi* $u_i \in U$ *(Bảng quyết định IDS nhất quán trên tập thuộc tính P)*

Chứng minh.

1) Từ công thức (5) ta thấy $D(P,\{d\})$ đạt giá trị lớn nhất khi k_P^i đạt giá trị lớn nhất là *n* với mọi $u_i \in U$, xảy ra khi $S_P(u_i) = U$ và phân hoạch $S_P(u_i)/\{d\} = \{\{u_i\} | u_i \in U\}$ (phân hoạch rời rạc), nghĩa là $|\partial_P(u_i)| = n$. Khi đó, giá trị lớn nhất là $1 - \frac{1}{n}\left(\sum_{i=1}^{n}\frac{1}{n}\right) = 1 - \frac{1}{n}$.

2) Tương tự, $D(P,\{d\})$ đạt giá trị nhỏ nhất khi k_P^i đạt giá trị nhỏ nhất là 1 với mọi $u_i \in U$, xảy ra khi phân hoạch $S_P(u_i)/\{d\} = \{S_P(u_i)\}$ (phân hoạch khối), nghĩa là $|\partial_P(u_i)| = 1$ với mọi $u_i \in U$, khi đó *IDS* là bảng quyết định nhất quán trên tập thuộc tính điều kiện *P*.

Mệnh đề 4. *Cho bảng quyết định không đầy đủ* $IDS = (U, A \cup \{d\})$. *Khi đó ta có:*

$$d(A,\{d\}) + \alpha(IDS) = 1 \qquad (6)$$

với $\alpha(IDS)$ *là độ chắc chắn của bảng quyết định IDS trong công trình* [6].

Mệnh đề 4 dễ dàng được suy ra từ công thức tính khoảng cách (5) và công thức tính độ chắc chắn của bảng quyết định $\alpha(IDS)$ trong công trình [6]. Mệnh đề 4 cho thấy khoảng cách từ tập thuộc tính điều kiện *A* đến thuộc tính quyết định *{d}* là đại lượng đối ngẫu với độ chắc chắn của bảng quyết định. Nếu khoảng cách này càng lớn (thuộc tính điều kiện càng xa thuộc tính quyết định) thì độ chắc chắn của bảng quyết định càng nhỏ và ngược lại.

IV. RÚT GỌN THUỘC TÍNH TRONG BẢNG QUYẾT ĐỊNH KHÔNG ĐẦY ĐỦ SỬ DỤNG KHOẢNG CÁCH

Trong phần này, chúng tôi trình bày một phương pháp heuristic rút gọn thuộc tính trong bảng quyết định không đầy đủ sử dụng khoảng cách. Giống như các phương pháp heuristic khác, phương pháp của chúng tôi cũng bao gồm các bước: định nghĩa tập rút gọn dựa trên khoảng cách, định nghĩa độ quan trọng của thuộc tính dựa trên khoảng cách và xây dựng một thuật toán heuristic tìm một tập rút gọn tốt nhất dựa trên tiêu chí đánh giá là độ quan trọng của thuộc tính.

Định nghĩa 1. Cho bảng quyết định không đầy đủ $IDS = (U, A \cup \{d\})$ và tập thuộc tính $R \subseteq C$. Nếu

1) $D(R,\{d\}) = D(A,\{d\})$

2) $\forall r \in R,\ D(R-\{r\},\{d\}) \neq D(A,\{d\})$

thì *R* là một tập rút gọn của *C* dựa trên khoảng cách.

Từ Mệnh đề 2 ta thấy tập rút gọn dựa trên khoảng cách và tập rút gọn dựa trên hàm quyết định suy rộng là như nhau. Từ kết quả phân nhóm các phương pháp rút gọn thuộc tính trong [7] ta có, phương pháp rút gọn khoảng cách được xây dựng thuộc Nhóm 2. Do đó, tập rút gọn của phương pháp đề xuất tương đương tập rút gọn của các phương pháp thuộc Nhóm 2 và hiệu quả hơn về chất lương phân lớp (tối thiểu hơn) các phương pháp thuộc Nhóm 3 và Nhóm 4. Điều đó có nghĩa rằng tập rút gọn của phương pháp đề xuất thuộc nhóm phương pháp tốt nhất về chất lượng phân lớp.

Định nghĩa 2. Cho bảng quyết định không đầy đủ $IDS = (U, A \cup \{d\})$, $B \subset A$ và $b \in A - B$. Độ quan trọng của thuộc tính b đối với tập thuộc tính *B* được định nghĩa bởi:

$$SIG_B(b) = D(B,\{d\}) - D(B \cup \{b\},\{d\}) \quad (7)$$

Theo Mệnh đề 2, $D(B,\{d\}) \geq D(B \cup \{b\},\{d\})$ nên $SIG_B(b) \geq 0$. $SIG_B(b)$ được tính bởi lượng thay đổi khoảng cách giữa tập thuộc tính điều kiện *B* và thuộc tính quyết định *{d}* khi thêm thuộc tính *b* vào *B* và $SIG_B(b)$ càng lớn thì lượng thay đổi khoảng cách càng lớn, hay thuộc tính *b* càng quan trọng và ngược lại. Độ quan trọng của thuộc tính này là tiêu chuẩn lựa chọn thuộc tính trong thuật toán heuristic tìm tập rút gọn của bảng quyết định.

Ý tưởng của thuật toán heuristic tìm một tập rút gọn tốt nhất sử dụng khoảng cách là xuất phát từ tập rỗng $R = \varnothing$, lần lượt bổ sung thêm vào *R* các thuộc tính có độ quan trọng lớn nhất cho đến khi tìm được tập rút gọn.

Thuật toán 1. Thuật toán heuristic tìm một tập rút gọn tốt nhất sử dụng khoảng cách.

Đầu vào: Bảng quyết định không đầy đủ $IDS = (U, A \cup \{d\})$

Đầu ra: Một tập rút gọn tốt nhất R.

1. $R = \varnothing$;
2. Tính khoảng cách $D(R,\{d\})$ và $D(A,\{d\})$;

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

3. While $D(R,\{d\}) \neq D(A,\{d\})$ do
4. Begin
5. For $a \in A - R$ tính $SIG_R(a) = D(R,\{d\}) - D(R \cup \{a\},\{d\})$;
6. Chọn $a_m \in A - R$ sao cho $SIG_R(a_m) = \underset{a \in A-R}{Max}\{SIG_R(a)\}$;
7. $R = R \cup \{a_m\}$;
8. Tính khoảng cách $D(R,\{d\})$;
9. End;

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

10. For each $a \in R$ do
11. Begin
12. Tính khoảng cách $D(R-\{a\},\{d\})$;
13. If $D(R-\{a\},\{d\}) = D(R,\{d\})$ then $R = R - \{a\}$;
14. End;
15. Return R;

Xét vòng lặp While từ dòng lệnh 3 đến 9, để tính $SIG_R(a)$ ta cần phải tính phải tính $D(R \cup \{a\},\{d\})$ vì $D(R,\{d\})$ đã được tính ở bước trước, nghĩa là cần phải tính $S_{R\cup\{a\}}(u_i)$ và phân hoạch $S_{R\cup\{a\}}(u_i)/\{d\}$. Trong công trình [5], độ phức tạp để tính $S_{R\cup\{a\}}(u_i)$ với mọi $u_i \in U$ khi $S_R(u_i)$ đã được tính là $O(|U|^2)$, độ phức tạp để tính phân hoạch $S_{R\cup\{a\}}(u_i)/\{d\}$ với mọi $u_i \in U$ là

$O\left(|U|^2\right)$. Do đó, độ phức tạp thời gian để tính tất cả các $SIG_R(a)$ ở dòng lệnh số 5 là:

$$\left(|A|+(|A|-1)+...+1\right)*|U|^2=\left(|A|*(|A|-)/2\right)*|U|^2=O\left(|A|^2|U|^2\right)$$

với $|A|$ là số thuộc tính điều kiện và $|U|$ là số đối tượng. Độ phức tạp thời gian để chọn thuộc tính có độ quan trọng lớn nhất ở dòng lệnh số 6 là: $|A|+(|A|-1)+...+1=|A|*(|A|-1)/2=O\left(|A|^2\right)$. Do đó, độ phức tạp thời gian của vòng lặp While là $O\left(|A|^2|U|^2\right)$. Tương tự, độ phức tạp của vòng lặp For từ dòng lệnh số 10 đến 14 là $O\left(|A|^2|U|^2\right)$. Vì vậy, độ phức tạp thời gian của Thuật toán 1 là $O\left(|A|^2|U|^2\right)$.

Độ phức tạp này tương đương với độ phức tạp của các phương pháp sử dụng độ đo trong Nhóm 2 và Nhóm 3 và hiệu quả hơn các phương pháp theo tiếp cận tính toán ma trận trong Nhóm 2 và Nhóm 3.

Ví dụ 1. Xét bảng quyết định không đầy đủ mô tả dữ liệu về các xe hơi cho ở Bảng 2 [1] $IDS=\left(U,A\cup\{d\}\right)$ với $U=\{u_1,u_2,u_3,u_4,u_5,u_6\}$ và *A = {Car, Price, Mileage, Size, Max-speed}*

***Bảng 2**. Bảng mô tả về các xe hơi*

Car	Price	Mileage	Size	Max-speed	d
u_1	High	High	Full	Low	Good
u_2	Low	*	Full	Low	Good
u_3	*	*	Compact	High	Poor
u_4	High	*	Full	High	Good
u_5	*	*	Full	High	Excellent
u_6	Low	High	Full	*	Good

Ta khởi tạo $R=\varnothing$ khi đó từ công thức $S_P(u)=\{v\in U|(u,v)\in SIM(P)\}$ ta có $S_R(u_1)=S_R(u_2)=S_R(u_3)=S_R(u_4)=S_R(u_5)=S_R(u_6)=U$ Từ đó: $S_R(u_1)/\{d\}=S_R(u_2)/\{d\}=S_R(u_3)/\{d\}=S_R(u_4)/\{d\}=S_R(u_5)/\{d\}=S_R(u_6)/\{d\}$ $=U/\{d\}=\{(u_1,u_2,u_4,u_6),(u_3),(u_5)\}$.

Tính $D(R,\{d\})$, từ công thức :

$$D(P,\{d\})=\frac{1}{n}\sum_{i=1}^{n}d\left(K_P^i(\delta),K_P^i(\{d\})\right)=$$

$$=\frac{1}{n}\sum_{i=1}^{n}\left(1-\frac{1}{k_P^i}\right)=1-\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{k_P^i}\right)$$

ta có $D(R,\{d\})$ = 1/6 { (1-1/3)+ (1-1/3)+(1-1/3)+ (1-1/3)+ (1-1/3)+(1-1/3)}=2/3

Tiếp tục tính $D(A,\{d\})$, ta có $S_A(u_1)=u_1$, $S_A(u_2)=(u_2,u_6)$, $S_A(u_3)=u_3$, $S_A(u_4)=(u_4,u_5)$, $S_A(u_5)=(u_4,u_5,u_6)$, $S_A(u_6)=(u_2,u_5,u_6)$.

Khi đó $S_A(u_1)/\{d\}=\{u_1\}, S_A(u_2)/\{d\}=\{(u_2,u_6)\}$, $S_A(u_3)/\{d\}=\{u_3\}, S_A(u_4)/\{d\}=\{(u_4),(u_5)\}$ $S_A(u_5)/\{d\}=\{(u_4,u_6),(u_5)\}$, $S_A(u_6)/\{d\}=\{(u_2,u_6),(u_5)\}$

Từ công thức (5) ta có:

$$D(A,\{d\})=1/6\{(1-1)+(1-1)+(1-1)+(1-1/2)+(1-1/2)+(1-1/2)\}=1/4$$

Vì vậy: $D(R,\{d\})\neq D(A,\{d\})$

Tiếp tục thực hiện vòng lặp While. Tính tương tự ta có:

$$D(R\cup\{a_1\},\{d\})=$$
$$=1/6\{(1-1/3)+(1-1/3)+(1-1/3)+(1-1/3)+(1-1/3)+(1-1/3)\}$$
$$=2/3$$

Từ đó

$$SIG_R(a_1)=D(R,\{d\})-D(R\cup\{a_1\},\{d\})=2/3-2/3=0$$

Tương tự ,

$$SIG_R(a_2)=D(R,\{d\})-D(R\cup\{a_2\},\{d\})=2/3-2/3=0$$

$$SIG_R(a_3)=D(R,\{d\})-D(R\cup\{a_3\},\{d\})=2/3-5/12=1/4$$

$SIG_R(a_4) = D(R,\{d\}) - D(R \cup \{a_4\},\{d\}) = 2/3 - 4/9 = 2/9$

Vậy $SIG_R(a_3)$ lớn nhất do đó $R = R \cup \{a_3\}$. Từ đó ta có $D(R,\{d\}) = 5/12$

Tiếp tục tính:

$SIG_R(a_1) = D(R,\{d\}) - D(R \cup \{a_1\},\{d\}) = 5/12 - 5/12 = 0$

$SIG_R(a_2) = D(R,\{d\}) - D(R \cup \{a_2\},\{d\}) = 5/12 - 5/12 = 0$

$SIG_R(a_4) = D(R,\{d\}) - D(R \cup \{a_4\},\{d\}) = 5/12 - 1/4 = 1/6$

Vậy $SIG_R(a_4)$ lớn nhất do đó ta có $R = R \cup \{a_4\}$, Vậy $D(R,\{d\}) = 1/4$

$D(R,\{d\}) = D(A,\{d\})$ dừng vòng lặp,

vậy $R = \{a_3, a_4\}$

Loại bỏ thuộc tính dư thừa trong *R*

Ta có $D(R - \{a_4\},\{d\}) = 5/12$.

do đó $D(R - \{a_4\},\{d\}) \neq D(R,\{d\})$

không loại bỏ a_4

Ta có $D(R - \{a_3\},\{d\}) = 4/9$.

do đó $D(R - \{a_4\},\{d\}) \neq D(R,\{d\})$

không loại bỏ a_3

Vậy tập rút gọn là $R = \{a_3, a_4\}$

V. THỰC NGHIỆM THUẬT TOÁN

Chúng tôi chọn thuật toán IQBAR tìm tập rút gọn của bảng quyết định không đầy đủ sử dụng độ đo lượng thông tin (Information Quantity) trong [1] để so sánh với thuật toán đề xuất (Thuật toán 1) về thời gian thực hiện và kết quả thực hiện. Sở dĩ chọn thuật toán IQBAR vì theo lý thuyết đã trình bày, tập rút gọn của Thuật toán 1 (Nhóm 2) tối thiểu hơn tập rút gọn của thuật toán IQBAR (Nhóm 3). Để tiến hành thử nghiệm, chúng tôi thực hiện các công việc sau:

1) Cài đặt thuật toán IQBAR và Thuật toán 1 bằng ngôn ngữ C#. Cả hai thuật toán đều sử dụng thuật toán trong [6] để tính các lớp dung sai $S_B(u_i)$ với $u_i \in U$.

2) Trên máy tính PC với cấu hình Pentium dual core 2.13 GHz CPU, 1GB bộ nhớ RAM, sử dụng hệ điều hành Windows XP Proessional, chạy thử nghiệm hai thuật toán với 6 bộ số liệu lấy từ kho dữ liệu UCI [12]. 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 *s*), các thuộc tính điều kiện được đánh số là 1, 2,…, $|C|$. Kết quả thực hiện của hai thuật toán được mô tả ở Bảng 3 và Bảng 4:

***Bảng 3**. Kết quả thực hiện thuật toán IQBAR và Thuật toán 1*

TT	Bộ số liệu	$\|U\|$	$\|C\|$	Thuật toán IQBAR		Thuật toán 1	
				$\|R\|$	*T*	$\|R\|$	*t*
1	Hepatitis.data	155	19	4	1.3	4	1.29
2	Lung-cancer.data	32	56	4	0.17	4	0.17
3	Automobile.data	205	25	5	1.7	5	1.68
4	Anneal.data	798	38	9	179	8	178
5	Congressional Voting Records	435	16	15	16.5	13	16.73
6	Credit Approval	690	15	7	16.2	7	15.68

***Bảng 4**. Tập rút gọn của thuật toán IQBAR và Thuật toán 1*

TT	Tập dữ liệu	Tập rút gọn của Thuật toán IQBAR	Tập rút gọn của Thuật toán 1
1	Hepatitis.data	{1, 2, 4, 17}	{1, 2, 4, 17}
2	Lung-cancer.data	{3, 4, 9, 43}	{3, 4, 9, 43}
3	Automobile.data	{1, 13, 14, 20, 21}	{1, 13, 14, 20, 21}
4	Anneal.data	{1, 3, 4, 5, 8, 9, 33, 34, 35}	{1, 3, 4, 5, 8, 9, 34, 35}
5	Congressional Voting Records	{1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16}	{1, 2, 3, 4, 5, 8, 10, 11, 12, 13, 14, 15, 16}
6	Credit Approval	{1, 2, 3, 4, 5, 6, 8}	{1, 2, 3, 4, 5, 6, 8}

Kết quả thử nghiệm cho thấy:

- Trên các bộ số liệu Hepatitis.data, Lung-cancer.data, Automobile.data, Credit Approval, tập rút gọn thu được bởi Thuật toán 1 và Thuật toán IQBAR là như nhau. Tuy nhiên, với bộ số liệu Anneal.data, Congressional Voting Records, tập rút gọn thu được bởi Thuật toán 1 tối thiểu hơn tập rút gọn thu được bởi Thuật toán IQBAR. Điều này cũng phù hợp với kết quả nghiên cứu về lý thuyết.

- Thời gian thực hiện Thuật toán 1 và Thuật toán IQBAR về cơ bản là tương đương nhau.

VI. KẾT LUẬN

Các nghiên cứu về rút gọn thuộc tính trong bảng quyết định không đầy đủ theo tiếp cận mô hình tập thô dung sai khá sôi động trong mấy năm gần đây. Trong bài báo này, chúng tôi đề xuất phương pháp heuristic rút gọn thuộc tính trong bảng quyết định không đầy đủ sử dụng độ đo khoảng cách phân hoạch, bao gồm các bước: xây dựng độ đo khoảng cách giữa tập thuộc tính điều kiện và thuộc tính quyết định; định nghĩa tập rút gọn dựa trên khoảng cách; định nghĩa độ quan trọng của thuộc tính dựa trên khoảng cách; xây dựng thuật toán heuristic tìm một tập rút gọn tốt nhất sử dụng khoảng cách. Chúng tôi chứng minh tập rút gọn dựa trên khoảng cách thuộc Nhóm 2. Do đó về chất lượng phân lớp, phương pháp sử dụng khoảng cách tương đương với các phương pháp thuộc Nhóm 2 và hiệu quả hơn các phương pháp thuộc Nhóm 3, Nhóm 4. Về độ phức tạp thời gian, phương pháp sử dụng khoảng cách tương đương với các phương pháp khác sử dụng độ đo và hiệu quả hơn các phương pháp theo tiếp cận ma trận trong Nhóm 2 và Nhóm 3. Kết quả thu được của bài báo bổ sung thêm các phương phương pháp rút gọn thuộc tính trong bảng quyết định không đầy đủ. Hướng phát triển tiếp theo của nhóm tác giả là nghiên cứu các phương pháp rút gọn trên bảng quyết định không đầy đủ với dữ liệu thay đổi.

TÀI LIỆU THAM KHẢO

[1] HUANG B., LI H. X. AND ZHOU X. Z., "*Attribute Reduction Based on Information Quantity under Incomplete Information Systems*", Systems Application Theory & Practice, Vol. 34, 2005, pp. 55-60.

[2] HUASHENG ZOU, CHANGSHENG ZHANG, "*Efficient Algorithm for Knowledge Reduction in Incomplete Information System*", Journal of Computational Information Systems 8: 6, 2012, pp. 2531-2538.

[3] KRYSZKIEWICZ M., "*Rough set approach to incomplete information systems*", Information Science, Vol. 112, 1998, pp. 39-49.

[4] LONG GIANG NGUYEN, "Metric Based Attribute Reduction in Decision Tables", Federated Conference on Computer Science and Information System (FEDCSIS), Wroclaw, Poland, IEEE, 2012, pp. 311-316.

[5] LONG GIANG NGUYEN, HUNG SON NGUYEN, "*Metric Based Attribute Reduction in Incomplete Decision Tables*", Proceedings of 14th International Conference, Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, RSFDGrC 2013, Halifax, NS, Canada, LNCS, SpingerLink, Vol. 8170, 2013, pp. 99-110.

[6] NGUYỄN LONG GIANG, VŨ VĂN ĐINH, "*Nghiên cứu sự thay đổi giá trị các độ đo đánh giá hiệu năng tập luật quyết định trên các tập rút gọn của bảng quyết định không đầy đủ*", Fundamental and Applied IT Research, Vol. 52, 2013, pp.394 – 402.

[7] NGUYEN LONG GIANG, VU VAN DINH, "*Relationships Among the Concepts of Reduct in Incomplete Decision Tables*", Frontiers in Artificial Intelligence and Applications, Volume 252: Advanced Methods and Technologies for Agent and Multi-Agent Systems, IOS Press, 2013, pp. 417-426.

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

[9] RENPU LI, DAO HUANG, "*Reducts in incomplete decision tables*", Proceedings of the First international conference on Advanced Data Mining and Applications, ADMA'05, 2005, pp. 165-174.

[10] ZUQIANG MENG, ZHONGZHI SHI, "*A fast approach to attribute reduction in incomplete decision*

systems with tolerance relation-based rough sets", Information Sciences, Vol. 179, 2009, pp. 2774-2793.

[11] ZHOU, X.Z., HUANG, B, "*Rough Set-based Attribute Reduction under Incomplete Information Systems*", Journal of Nanjing University of Science and Technology, 27(2003), pp. 630-635.

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

Ngày nhận bài: 04/12/2014

SƠ LƯỢC VỀ TÁC GIẢ

VŨ VĂN ĐỊNH

Sinh ngày 22/08/1977 tại Hải Phòng

Tốt nghiệp Trương ĐH Khoa học Tự nhiên, ĐH Quốc gia Hà Nội năm 2003, chuyên ngành Toán tin ứng dụng. Bảo vệ luận án Thạc sĩ tại ĐH Công nghệ Thông tin năm 2007, chuyên ngành Khoa học máy tính.

Hướng nghiên cứu: Khai phá dữ liệu, cơ sở dữ liệu và mô hình hóa hệ thống thông tin.

Email: dinhvv@epu.edu.vn

VŨ ĐỨC THI

Sinh ngày 07/04/1949 tại Hải Dương.

Tốt nghiệp ĐH Tổng hợp Hà Nội năm 1971. Bảo vệ luận án tiến sỹ tại Viện Hàn lâm Khoa học Hungary, năm 1987, chuyên ngành Cơ sở dữ liệu, CNTT. Nhận học hàm Phó giáo sư năm 1991, Giáo sư năm 2009.

Hướng nghiên cứu: Cơ sở dữ liệu và hệ thống thông tin, khai phá dữ liệu và học máy.

Email: vdthi@vnu.edu.vn

NGUYỄN LONG GIANG

Sinh ngày 05/06/1975 tại Hà Tây.

Tốt nghiệp Trường ĐH Bách khoa Hà Nội năm 1997, thạc sĩ tại Trường ĐH Công nghệ, ĐH Quốc gia Hà Nội năm 2003. Bảo vệ luận án tiến sỹ tại Viện CNTT, Viện Hàn lâm KH&CN Việt Nam năm 2012, chuyên ngành: Đảm bảo toán học cho máy tính và các hệ thống tính toán.

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

Email: nlgiang@ioit.ac.vn

NGÔ QUỐC TẠO

Tốt nghiệp: Khoa Toán lý, Trường ĐH Bách khoa Hà Nội năm 1982, chuyên ngành Toán Máy tính. Nhận bằng Tiến sỹ Toán lý năm 1997, Chuyên ngành đảm báo toán học cho các hệ thống tính toán. Được phong Phó Giáo sư Tin học năm 2002.

Lĩnh vực nghiên cứu: Nhận dạng, xử lý ảnh, nhập liệu tự động, trí tuệ nhân tạo, khai phá dữ liệu.

Email: nqtao@ioit.ncst.ac.vn