KHAI PHÁ DỮ LIỆU SỬ DỤNG LÝ THUYẾT TẬP THÔ

Ninh Văn Thọ*
Trường Đại học Kỹ thuật Công nghiệp – ĐH Thái Nguyên

TÓM TẮT

Lý thuyết tập thô đã được sử dụng hiệu quả trong các bước của quá trình khai phá dữ liệu và khám phá tri thức. Trong đó bài toán rút gọn thuộc tính theo tiếp cận lý thuyết tập thô là bài toán quan trọng trong khai thác dữ liệu nói chung và trong rút gọn các thuộc tính nói riêng. Trong thực tế dữ liệu thường đa dạng, phong phú nhưng nhiều khi có thể dư thừa hoặc không đầy đủ, điều này ảnh hưởng đến việc khám phá tri thức từ dữ liệu. Trong bài báo này, tác giả sử dụng ma trận phân biệt mở rộng trong mô hình tập thô dung sai để xây dựng thuật toán cho bài toán rút gọn thuộc tính trong hệ thông tin đa trị và minh họa kết quả thuật toán qua thực nghiệm.

Từ khóa: *Tập thô dung sai, tập thô, hệ quyết định đa trị, rút gọn thuộc tính, tập rút gọn*

MỞ ĐẦU

Rút gọn thuộc tính trong hệ quyết định đa trị là tìm ra tập thuộc tính nhỏ nhất có thể được để biểu diễn dữ liệu nhưng vẫn giữ được mối quan hệ ngữ nghĩa giữa các tập thuộc tính. Rút gọn thuộc tính vừa làm giảm khối lượng tính toán do quá trình xử lý dữ liệu chỉ thao tác trên một dung lượng dữ liệu nhỏ hơn, làm cho kết quả thu được từ quá trình xử lý trở nên cô đọng và dễ hiểu hơn. Trên hệ thông tin đa trị, Yan Yong Guan và cộng sự [2] đã mở rộng quan hệ tương tương trong lý thuyết tập thô truyền thống thành quan hệ dung sai và xây dựng mô hình tập thô dung sai bằng cách mở rộng các định nghĩa xấp xỉ trên, xấp xỉ dưới, miền dương...dựa trên quan hệ dung sai. Theo hướng tiếp cận mô hình tập thô dung sai, một số công trình nghiên cứu đáng chú ý về rút gọn thuộc tính trên hệ quyết định đa trị và hệ quyết định đa trị xếp thứ tự có thể kể đến [1, 6, 9]. Trong công trình [11], sử dụng phương pháp ma trận các tác giả đã nghiên cứu sự thay đổi của các tập xấp xỉ khi bổ sung và loại bỏ tập thuộc tính. Tuy nhiên, các kết quả nghiên cứu về rút gọn thuộc tính trong trường hợp hệ quyết định đa trị vẫn còn hạn chế và đòi hỏi nhiều nỗ lực nghiên cứu hơn nữa.

Dựa trên ý tưởng ma trận phân biệt và hàm phân biệt trong lý thuyết tập thô truyền thống do Skowron đề xuất [8], trong bài báo này tác giả xây dựng ma trận phân biệt mở rộng và hàm phân biệt mở rộng. Sử dụng hàm phân biệt mở rộng, tác giả xây dựng phương pháp rút gọn thuộc tính trong trường hợp hệ quyết định đa trị không thay đổi.

Cấu trúc bài báo như sau. Phần 2 trình bày một số khái niệm về hệ quyết định đa trị và các khái niệm tập rút gọn. Phần 3 trình bày phương pháp rút gọn thuộc tính sử dụng hàm phân biệt mở rộng. Phần 4 là kết luận và định hướng nghiên cứu tiếp theo.

CÁC KHÁI NIỆM CƠ BẢN

Trong phần này, bài báo sẽ đưa ra một số khái niệm cơ bản về hệ thông tin đa trị được tham khảo trong [2].

Định nghĩa 1 [2]. Hệ thông tin đa trị là một bộ bốn $IS = (U, A, V, f)$ trong đó U là tập hữu hạn, khác rỗng được gọi là tập vũ trụ hoặc tập các đối tượng; A là tập là hữu hạn khác rỗng các thuộc tính; f là hàm thông tin, $f : U \times A \to 2^V$ là ánh xạ tập giá trị.

Ví dụ 1. Bảng 1 [6] minh họa một hệ thông tin đa trị (bỏ qua cột thuộc tính dec) với mười đối tượng

$U = \{u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}\}$, bốn thuộc tính giá trị tập

$A = \{Audition,\ Spoken\ Language,\ Reading,\ Writing\}$ và tập giá trị $V = \{E, F, G\}$.

* *Tel: 0914 770072, Email: Thodtu.nd@gmail.com*

Bảng 1. *Hệ thông tin đa trị*

U	**Audition (A)**	**Spoken Language (S)**	**Reading (R)**	**Writing (W)**	**Dec**
u_1	$\{E\}$	$\{E\}$	$\{F,G\}$	$\{F,G\}$	No
u_2	$\{E,F,G\}$	$\{E,F,G\}$	$\{F,G\}$	$\{E,F,G\}$	No
u_3	$\{E,G\}$	$\{E,F\}$	$\{F,G\}$	$\{F,G\}$	No
u_4	$\{E,F\}$	$\{E,G\}$	$\{F,G\}$	$\{F\}$	No
u_5	$\{F,G\}$	$\{F,G\}$	$\{F,G\}$	$\{F\}$	No
u_6	$\{F\}$	$\{F\}$	$\{E,F\}$	$\{E,F\}$	Yes
u_7	$\{E,F,G\}$	$\{E,F,G\}$	$\{E,G\}$	$\{E,F,G\}$	Yes
u_8	$\{E,F\}$	$\{F,G\}$	$\{E,F,G\}$	$\{E,G\}$	Yes
u_9	$\{F,G\}$	$\{G\}$	$\{F,G\}$	$\{F,G\}$	Yes
u_{10}	$\{E,F\}$	$\{E,G\}$	$\{F,G\}$	$\{E,F\}$	Yes

Định nghĩa 2. (*Quan hệ dung sai trong hệ thông tin đa trị*)

Cho hệ thông tin đa trị $IS=(U,A,V,f)$. Với mỗi tập con thuộc tính $B\subseteq A$, quan hệ $T_B=\{(u,v)\in U\times U \mid \forall b\in B, u(b)\cap v(b)\neq\varnothing\}$ là một quan hệ dung sai và được gọi là quan hệ dung

sai tương ứng với B.

Đặt $[u]_{T_B}=\{v\in U \mid (u,v)\in T_B\}$ thì $[u]_{T_B}$ được gọi là một lớp dung sai tương ứng với quan hệ T_B.

Ký hiệu $U/T_B=\{[u]_{T_B} \mid u\in U\}$ biểu diễn tập tất cả các lớp dung sai tương ứng với quan hệ T_B, khi đó U/T_B hình thành một *phủ* của U vì các lớp dung sai trong U/T_B *có thể giao nhau và* $\bigcup_{u\in U}[u]_{T_B}$ =U.

Rõ ràng là nếu $C\subseteq B$ thì $[u]_{T_B}\subseteq[u]_{T_C}$ với mọi $u\in U$.

Định nghĩa 3 . Bảng quyết định đa trị (còn được gọi là hệ quyết định đa trị)

Hệ quyết định đa trị là hệ thông tin đa trị $DS=(U,AT\cup\{d\})$ trong đó AT là các thuộc tính điều kiện và d là thuộc tính quyết định, với giả thiết $d(u)$ chứa *một giá trị* với mọi $u\in U$. Với $u\in U$, $\partial_{AT}(u)=\{d(v)\mid v\in T_{AT}(u)\}$ được gọi là *hàm quyết định suy rộng* của đối tượng u trên tập thuộc tính AT.

Nếu $|\partial_{AT}(u)|=1$ với mọi $u\in U$ thì DS là *nhất quán*, trái lại DS là *không nhất quán*.

Tương tự hệ quyết định không đầy đủ [3], tập rút gọn của hệ quyết định đa trị được định nghĩa như sau:

Định nghĩa 4. Cho hệ quyết định đa trị $DS=(U,AT\cup\{d\})$. Nếu $R\subseteq AT$ thỏa mãn:

(1) $\partial_R(u)=\partial_{AT}(u)$ với mọi $u\in U$

(2) với mọi $\forall R'\subset R$, tồn tại $u\in U$ sao cho $\partial_{R'}(u)\neq\partial_{AT}(u)$

thì R được gọi là một *tập rút gọn của DS dựa trên hàm quyết định suy rộng.*

RÚT GỌN THUỘC TÍNH TRONG HỆ QUYẾT ĐỊNH ĐA TRỊ SỬ DỤNG HÀM PHÂN BIỆT MỞ RỘNG

Rút gọn thuộc tính trong hệ quyết định là quá trình lựa chọn tập con nhỏ nhất của tập thuộc tính điều kiện mà bảo toàn thông tin phân lớp của bảng quyết định. Trong lý thuyết tập thô truyền thống, Skowron [8] đã đưa ra khái niệm ma trận phân biệt và hàm phân biệt để tìm tập rút gọn. Dựa trên cách tiếp cận này, tác giả đưa ra khái niệm ma trận phân biệt mở rộng và hàm phân biệt mở rộng để tìm tập rút gọn của hệ quyết định đa trị.

Định nghĩa 5. Cho hệ quyết định đa trị $DS=(U,AT\cup\{d\})$ với $A\subseteq AT$ và $|U|=n$.

Ma trận phân biệt mở rộng của DS trên tập thuộc chính A, ký hiệu $M_A=(m_{ij})_{n\times n}$, là ma trận vuông cấp n, mỗi phần tử có giá trị 0 hoặc 1, được định nghĩa như sau:

(1) $m_{ij}=1$ nếu $d(u_j)\notin\partial_A(u_i)$

(2) $m_{ij}=0$ nếu $d(u_j)\in\partial_A(u_i)$.

Chú ý: Nếu $A=\varnothing$ thì quy ước $m_{ij}=0$ và M_A không phải là ma trận đối xứng vì

$d(u_j) \notin \partial_A(u_i)$ vẫn có thể $d(u_i) \in \partial_A(u_j)$ với $i = 1,...,n;\ j = 1,...,n$.

Ví dụ 2. Với hệ quyết định đa trị cho ở ví dụ 1, ma trận phân biệt mở rộng của DS trên tập thuộc tính AT như sau:

$$M_{AT} = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Ví dụ 3. Tiếp tục Ví dụ 2, giả sử $A \subseteq AT$ với $A = \{a_1, a_2, a_3\}$, khi đó ta tính được

$$M_A = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Rõ ràng, $M_A \preceq M_{AT}$

Định nghĩa 6. Cho hệ quyết định tập giá trị $DS = (U, AT \cup \{d\})$, $A \subseteq AT$ và $M_A = (m_{ij})_{n \times n}$ là ma trận phân biệt mở rộng của DS trên tập thuộc tính A. Khi đó, hàm phân biệt mở rộng của DS trên A, ký hiệu là $DIS(A)$, được định nghĩa như sau:

$$DIS(A) = \sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij} \text{ với}$$

$1 \le i \le n, 1 \le j \le n$.

Ví dụ 4. Tiếp tục Ví dụ 2, với ma trận phân biệt M_{AT}, hàm phân biệt là:

$$DIS(AT) = 2+2+5+1+1+1 = 12$$

Từ Định nghĩa 4 ta có mệnh đề sau:

Mệnh đề 1. Cho hệ quyết định đa trị $DS = (U, AT \cup \{d\})$ với $P, Q \subseteq AT$.

Nếu $P \subseteq Q$ thì $DIS(Q) \ge DIS(P)$.

Phần tiếp theo, bài báo trình bày phương pháp rút gọn thuộc tính trong hệ quyết định đa trị sử dụng hàm phân biệt mở rộng. Giống như các phương pháp rút gọn thuộc tính trong lý thuyết tập thô truyền thống, phương pháp của bài báo bao gồm các bước: định nghĩa tập rút gọn, định nghĩa độ quan trọng của thuộc tính và xây dựng thuật toán heuristic tìm một tập rút gọn tốt nhất dựa trên độ quan trọng của thuộc tính. Định nghĩa 4 cho thấy, hàm phân biệt mở rộng $DIS(A)$ đặc trưng cho khả năng phân lớp của tập thuộc tính $A \subseteq AT$ vào các lớp quyết định sinh bởi thuộc tính d, do đó bài báo sử dụng hàm phân biệt mở rộng làm tiêu chuẩn lựa chọn thuộc tính trong thuật toán heuristic tìm tập rút gọn, gọi là độ quan trọng của thuộc tính.

Định nghĩa 7. Cho hệ quyết định đa trị $DS = (U, AT \cup \{d\})$. Nếu $R \subseteq AT$ thỏa mãn:

(1) $DIS(R) = DIS(AT)$

(2) $\forall R' \subset R,\ DIS(R') \ne DIS(AT)$

thì *R* được gọi là một *tập rút gọn của DS dựa trên hàm phân biệt mở rộng.*

Định nghĩa 8. Cho hệ quyết định đa trị $DS = (U, AT \cup \{d\})$, $A \subset AT$ và $a \in AT - A$. Độ quan trọng của thuộc tính a đối với tập thuộc tính A được định nghĩa bởi

$$SIG_A^{out}(a) = DIS(A \cup \{a\}) - DIS(A)$$

Định nghĩa 9. Cho hệ quyết định đa trị $DS = (U, AT \cup \{d\})$, $A \subset AT$ và $a \in A$. Độ quan trọng của thuộc tính a trong tập thuộc tính A được định nghĩa bởi

$$SIG_A^{\text{in}}(a) = DIS(A) - DIS(A - \{a\})$$

Từ Mệnh đề 1 ta có $SIG_A^{out}(a) \ge 0$ và $SIG_A^{in}(a) \ge 0$ Do đó, $SIG_A^{out}(a)$ và $SIG_A^{in}(a)$ được tính bởi lượng thay đổi hàm phân biệt mở rộng khi thêm thuộc tính *a* vào *A* hoặc loại bỏ *a* khỏi *A* và $SIG_A^{out}(a)$,

$SIG_A^{in}(a)$ càng lớn thì lượng thay đổi này càng lớn, hay thuộc tính *a* càng quan trọng và ngược lại.

Tiếp theo, bài báo đề xuất thuật toán heuristic tìm một tập rút gọn tốt nhất theo tiêu chuẩn đánh giá độ quan trọng của thuộc tính.

Ý tưởng của thuật toán là xuất phát từ tập thuộc tính rỗng $R := \{\varnothing\}$, lần lượt bổ sung vào tập *R* các thuộc tính có độ quan trọng lớn nhất cho đến khi tìm được tập rút gọn.

Thuật toán heuristic tìm một tập rút gọn sử dụng hàm phân biệt mở rộng

Đầu vào: Hệ quyết định đa trị $DS = (U, AT \cup \{d\})$.

Đầu ra: Một tập rút gọn R.

1. $R = \varnothing$;

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

2. While $DIS(R) \neq DIS(AT)$ do
3. Begin
4. For each $a \in AT - R$ tính

$SIG_R^{out}(a) = DIS(R \cup \{a\}) - DIS(R)$;

5. Chọn $a_m \in AT - R$ sao cho

$SIG_R^{out}(a_m) = \underset{a \in AT-R}{Max}\{SIG_R^{out}(a)\}$;

6. $R = R \cup \{a_m\}$;
7. End;

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

8. For each $a \in R$
9. If $DIS(R - \{a\}) = DIS(R)$ then

$R = R - \{a\}$;

10. Return R;

Đánh giá độ phức tạp của thuật toán:

Giả sử *k* là số thuộc tính điều kiện và n là số đối tượng.

Ta có độ phức tạp để tính M_{AT} là $O(kn^2)$, do đó độ phức tạp tính $DIS(AT)$ là $O(kn^2)$. Xét vòng lặp While từ dòng lệnh 2 đến dòng lệnh 7, độ phức tạp để tính tất cả các $SIG_R(a)$ là

$(k + (k-1) + ... + 1) * kn^2 = (k * (k-1)/2) * kn^2 = O(k^3 n^2)$

Độ phức tạp thời gian để chọn thuộc tính có độ quan trọng lớn nhất là

$k + (k-1) + ... + 1 = k * (k-1)/2 = O(k^2)$.

Do đó, độ phức tạp của vòng lặp While là $O(k^3 n^2)$. Tương tự, độ phức tạp của vòng lặp For là $O(k^2 n^2)$. Vì vậy, độ phức tạp của Thuật toán là là $O(k^3 n^2)$.

Ví dụ 5. Xét hệ quyết định đa trị $DS = (U, AT \cup \{d\})$ cho ở Ví dụ 1. Áp dụng Thuật toán tìm tập rút gọn R ta có:

Đặt $R = \varnothing$ và tính:

$SIG_{\varnothing}^{out}(a_1) = DIS(\{a_1\}) - DIS(\varnothing) = DIS(\{a_1\}) = 0$

$SIG_{\varnothing}^{out}(a_2) = DIS(\{a_2\}) - DIS(\varnothing) = DIS(\{a_2\}) = 0$

$SIG_{\varnothing}^{out}(a_3) = DIS(\{a_3\}) - DIS(\varnothing) = DIS(\{a_3\}) = 10$

$SIG_{\varnothing}^{out}(a_4) = DIS(\{a_4\}) - DIS(\varnothing) = DIS(\{a_4\}) = 4$

Chọn thuộc tính a_3 có độ quan trọng lớn nhất và $R = \{a_3\}$. Từ Ví dụ 4 ta có $DIS(AT) = 12$, do đó $DIS(R) \neq DIS(AT)$.

Chuyển vòng lặp thứ 2 và tính:

$SIG_{\{a_3\}}^{out}(a_1) = DIS(\{a_1, a_3\}) - DIS(\{a_3\}) = 10 - 10 = 0$

$SIG_{\{a_3\}}^{out}(a_2) = DIS(\{a_2, a_3\}) - DIS(\{a_3\}) = 10 - 10 = 0$

$SIG_{\{a_3\}}^{out}(a_4) = DIS(\{a_3, a_4\}) - DIS(\{a_3\}) = 12 - 10 = 2$

Chọn thuộc tính a_4 có độ quan trọng lớn nhất, và $R = \{a_3, a_4\}$.

Ta thấy $DIS(\{a_3, a_4\}) = DIS(AT) = 12$, chuyển đến vòng lặp *For* thực hiện kiểm tra tập *R* thu được.

Theo tính toán ở trên, $DIS(\{a_4\}) \neq DIS(AT)$ và $DIS(\{a_3\}) \neq DIS(AT)$. Do đó thuật toán

kết thúc và $R=\{a_3,a_4\}$ là một rút gọn "tốt nhất" của *AT*.

Kết quả thực nghiệm thuật toán tìm tập rút gọn

Môi trường thực 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 XP Professional. bộ số liệu đa trị được chuyển đổi từ bộ số liệu trong kho dữ liệu [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 thứ tự từ 1 đến $|C|$.

STT	Bộ số liệu	$\|U\|$	$\|C\|$	Thuật toán		Tập rút gọn của thuật toán
				$\|R\|$	T	
1	Hepatitis.data	155	19	3	0.735	{2, 15, 16}
2	Automobile.data	205	25	6	4.567	{1, 2, 7, 14, 20, 21}

KẾT LUẬN

Dựa trên ý tưởng ma trận phân biệt và hàm phân biệt [8] trong lý thuyết tập thô truyền thống, trong bài báo này tác giả xây dựng công cụ ma trận phân biệt mở rộng và hàm phân biệt mở rộng để xây dựng thuật toán tìm tập rút gọn của hệ quyết định đa trị.

Định hướng nghiên cứu tiếp theo là xây dựng các thuật toán gia tăng tìm tập rút gọn của hệ quyết định tập giá trị trong trường hợp bổ sung hoặc loại bỏ tập thuộc tính.

TÀI LIỆU THAM KHẢO

1. Chen Z. C, Shi P., Liu P. G., Pei Z., Criteria Reduction of Set-Valued Ordered Decision System Based on Approximation Quanlity, *International Journal of Innovative Computing, Information and Control*, Vol 9, N 6, 2013, pp. 2393-24-4.

2. Guan Y. Y., Wang H. K., Set-valued information systems, *Information Sciences* 176, 2006, pp. 2507–2525.

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

4. Pawlak Z., Rough sets, *International Journal of Information and Computer Sciences*, 11(5), 1982, pp. 341-356.

5. Pawlak Z., Rough sets: Theoretical Aspects of Reasoning About Data, Kluwer Aca-demic Publishers, 1991.

6. Qian Y. H., Dang C. Y., Liang J. Y., Tang D. W., Set-valued ordered information systems, *Information Sciences* 179, 2009, pp. 2809-2832.

7. Shifei D., Hao D., Research and Development of Attribute Reduction Algorithm Based on Rough Set, *IEEE, CCDC2010*, 2010, pp. 648-653.

8. Skowron A., Rauszer C., The Discernibility Matrices and Functions in Information systems, Interlligent Decision Support, *Handbook of Applications and Advances of the Rough Sets Theory*, Kluwer, Dordrecht, 1992, pp. 331-362.

9. Y. H. Qian Y. H. , Liang J. Y., On Dominance Relations in Disjunctive Set-Valued Ordered Information Systems, *International Journal of Information Technology & Decision Making* Vol. 9, No. 1, 2010, pp. 9–33.

10. Yao Y.Y., Zhao Y., Wang J., On reduct construction algorithms, *Proceedings of International Conference on Rough Sets and Knowledge Technology*, 2006, pp. 297-304.

11. Zhang J. B., Li T. R., Ruan D., Liu D., Rough sets based matrix approaches with dynamic attribute variation in set-valued information systems, *International Journal of Approximate Reasoning* 53, 2012, pp. 620–635. The UCI machine learning repository, http://archive.ics.uci.edu/ml/datasets.html

SUMMARY

DATA MINING USING ROUGH SETS THEORY

Ninh Van Tho*
College of Technology - TNU

Rough set theory has been used effectively in the steps of the process of data mining and knowledge discovery. In this simplified attributes reduction by rough set theory was important problem in data mining in general and in shortened particular attributes. In fact the data are often diverse, rich but sometimes can be excessive or inadequate, which affects knowledge discovery from data. In this paper, the author use a generalized discernibility matrix rough set model tolerances to build algorithms attribute reduction in set-valued information systems and illustrate the results of the algorithm through experimental program.

Keywords: *Tolerance-based rough set, rough set, decision system, set-valued decision system, attribute reduction, reduct*

Ngày nhận bài:25/12/2014; Ngày phản biện:20/01/2015; Ngày duyệt đăng: 31/5/2015
Phản biện khoa học: *TS. Vũ Vinh Quang – Trường Đại học Công nghệ Thông tin & Truyền thông - ĐHTN*

* *Tel: 0914 770072, Email: Thodtu.nd@gmail.com*