KHAI PHÁ DỮ LIỆU TRÊN HỆ THÔNG TIN ĐA TRỊ

Phùng Thị Thu Hiền*
Trường Đại học Kinh tế Kỹ thuật Công nghiệp

TÓM TẮT

Dựa trên ý tưởng thu nhỏ kích thước tập dữ liệu ban đầu, trong bài báo này tác giả đề xuất phương pháp lựa chọn tập đối tượng đại diện, gọi tắt là mẫu đại diện, từ tập đối tượng ban đầu cho bài toán tìm tập thuộc tính tối ưu của hệ thông tin đa trị. Tác giả chứng minh tập thuộc tính tối ưu trên tập đối tượng ban đầu và tập thuộc tính tối ưu trên mẫu đại diện là tương đương, từ đó khẳng định tính đúng đắn của phương pháp. Vì kích thước mẫu đại diện nhỏ hơn kích thước tập đối tượng ban đầu nên thời gian thực hiện các thuật toán tìm tập thuộc tính tối ưu trên mẫu đại diện giảm thiểu đáng kể. Kích thước mẫu đại diện được chọn lớn hay nhỏ phụ thuộc vào đặc thù mỗi hệ thông tin đa trị trong thực tế. Đồng thời bài báo trình bày phương pháp khai phá luật xếp thứ tự bằng cách chuyển đổi hệ thông tin đơn trị xếp thứ tự thành hệ thông tin đơn trị nhị phân và áp dụng các kỹ thuật sinh luật trong lý thuyết tập thô trên hệ thông tin đơn trị nhị phân thu được.

Từ khóa: *Hệ thông tin đa trị, tập thô, tập thuộc tính tối ưu, quan hệ dung sai*

MỞ ĐẦU

Lý thuyết tập thô truyền thống do Pawlak [1], [2] đề xuất được xây dựng dựa trên quan hệ tương đương nhằm giải quyết bài toán tìm tập thuộc tính tối ưu và sinh luật quyết định trên các hệ thông tin đơn trị. Trong các bài toán thực tế, giá trị một đối tượng tại một thuộc tính trên hệ thông tin có thể là một tập hợp nhiều giá trị.

Trên cả hệ thông tin đơn trị và hệ thông tin đa trị, tìm tập thuộc tính tối ưu là bài toán quan trọng nhất, đã và đang thu hút sự quan tâm của cộng đồng nghiên cứu về tập thô. Với bài toán tìm tập thuộc tính tối ưu, vấn đề đang được các nhà nghiên cứu quan tâm hàng đầu là xây dựng các phương pháp pháp nhằm tối ưu thời gian thực hiện các thuật toán, nhờ đó có thể áp dụng trên các hệ thông tin kích thước lớn. Trên hệ thông tin đơn trị, cho đến nay nhiều phương pháp tìm tập thuộc tính tối ưu đã được công bố [3], tuy nhiên các phương pháp này đều thực hiện trên tập đối tượng ban đầu. Trên hệ thông tin đa trị, các công trình nghiên cứu [4], [5], [6] đã đề xuất giải pháp nén dữ liệu với mục đích thu nhỏ kích thước tập dữ liệu ban đầu nhằm giảm thiểu thời gian thực hiện các thuật toán.

* *Tel: 0914 770070, Email: Thuhiencn1@gmail.com*

Bài báo này tác giả đề xuất phương pháp lựa chọn tập đối tượng đại diện, gọi tắt là mẫu đại diện, từ tập đối tượng ban đầu cho bài toán tìm tập thuộc tính tối ưu của hệ thông tin đa trị, và trình bày phương pháp khai phá luật xếp thứ tự.

Cấu trúc bài báo như sau. Phần 2 trình bày một số khái niệm cơ bản và một số kết quả trên hệ thông tin đa trị và phương pháp khai phá luật xếp thứ tự trên hệ thông đơn trị. Phần 3 đề xuất phương pháp chọn mẫu đại diện trên hệ thông tin đa trị. Phần 4 là kết luận và định hướng nghiên cứu tiếp theo

CÁC KHÁI NIỆM CƠ BẢN

Hệ thông tin đa trị

Hệ thông tin đa trị [7], [8] là một bộ bốn $IS=(U,AT,V,f)$ trong đó U là tập hữu hạn, khác rỗng được gọi là tập vũ trụ hoặc tập các đối tượng; AT 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 tương ứng mỗi cặp (u,a) tới một tập giá trị thuộc V.

Bài báo quy ước viết tắt $IS=(U,AT,V,f)$ là $IS=(U,AT)$.

Ký hiệu giá trị của thuộc tính $a\in AT$ tại đối tượng $u\in U$ là $a(u)$, khi đó mỗi tập con thuộc tính $A\subseteq AT$ xác định một quan hệ tương đương:

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

Định nghĩa 2.1.[7]**.** Quan hệ dung sai trong hệ thông tin đa trị

Cho hệ thông tin đa trị $IS = (U, AT)$. Với mỗi tập con thuộc tính $B \subseteq AT$, quan hệ $S_B = \{(u,v) \in U \times U \mid \forall b \in B, u(b) \cap v(b) \neq \varnothing\}$ là một quan hệ dung sai và được gọi là quan hệ dung sai tương ứng với *B*. Rõ ràng là $\forall B \subseteq AT:\ S_B = \bigcap_{b \in B} S_b$.

Đặt $[u]_{S_B} = \{v \in U \mid (u,v) \in S_B\}$ thì $[u]_{S_B}$ được gọi là một lớp dung sai tương ứng với quan hệ S_B. Ký hiệu $U / S_B = \{[u]_{S_B} \mid u \in U\}$ biểu diễn tập tất cả các lớp dung sai tương ứng với quan hệ S_B, khi đó U / S_B hình thành một phủ của *U* vì các lớp dung sai trong U / S_B có thể giao nhau và $\bigcup_{u \in U} [u]_{S_B} = U$. Rõ ràng là nếu $C \subseteq B$ thì $[u]_{S_B} \subset [u]_{S_C}$ với mọi $u \in U$.

Tương tự trong hệ thông tin không đầy đủ [9], với hệ thông tin đa trị $IS = (U, AT)$, tập thuộc tính $R \subseteq AT$ được gọi là tập thuộc tính tối ưu của *IS* nếu $S_R = S_{AT}$ và $\forall B \subset R,\ S_B \neq S_{AT}$, điều này tương đương với $S_R(u) = S_{AT}(u)$ với mọi $u \in U$ và $\forall B \subset R$ tồn tại $u \in U$ sao cho $S_B(u) \neq S_{AT}(u)$.

Hệ quyết định đa trị là hệ thống gồm các thành phần $DS = (U, AT \cup \{d\})$ 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 $u \in U$, $\partial_{AT}(u) = \{d(v) | v \in S_{AT}(u)\}$ 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.

Từ $S_A = \bigcap_{a \in A} S_a$, theo định nghĩa hàm quyết định suy rộng ta suy ra $\partial_{AT}(u) = \bigcap_{a \in AT} \partial_{AT}(u)$ với mọi $u \in U$.

Nếu $B \subseteq A$ thì từ $S_A(u) \subseteq S_B(u)$ ta dễ dàng suy ra $\partial_A(u) \subseteq \partial_B(u)$ với mọi $u \in U$.

Tương tự hệ quyết định không đầy đủ [9], với hệ quyết định đa trị $DS = (U, AT \cup \{d\})$, tập thuộc tính $R \subseteq AT$ được gọi là tập thuộc tính tối ưu của *DS* nếu $\partial_R(u) = \partial_{AT}(u)$ với mọi $u \in U$ và $\forall B \subset R$ tồn tại $u \in U$ sao cho $\partial_B(u) \neq \partial_{AT}(u)$.

Hệ thông tin đơn trị xếp thứ tự

Hệ thông tin đơn trị (IIS) là hệ thống gồm các thành phần $T = (U, A \cup D, F, G)$ với:

$U = (x_1, x_2, ..., x_n)$ là tập hữu hạn khác rỗng các đối tượng; $A \cup D$ là tập hữu hạn khác rỗng các thuộc tính; $A = (a_1, a_2, ..., a_p)$ là tập các thuộc tính điều kiện; $D = (d_1, d_2, ..., d_p)$ là tập các thuộc tính quyết định, và $A \cap D \neq \phi$; $F = \{f_k | U \to V_k, k \leq p\}$, $f_k(x)$ là giá trị của a_k trên $x \in U$, V_k là miền giá trị của a_k, $a_k \in A$;

$G = \{g_{k'} | U \to V_{k'},\ k' \leq p\}, g_{k'}(x)$ là giá trị của $d_{k'}$ trên $x \in U$, $V_{k'}$ là miền giá trị của $d_{k'}$, $d_{k'} \in D$;

Nếu miền giá trị của một thuộc tính được xếp theo ưu tiên tăng dần hoặc giảm dần thì thuộc tính đó gọi là một tiêu thức.

Định nghĩa 2.2. [10] Một hệ thông tin đơn trị được gọi là xếp thứ tự *(OIIS)* nếu tất cả các thuộc tính điều kiện là các tiêu thức.

Giả sử rằng một quan hệ xếp thứ tự $\preccurlyeq_a$ được định nghĩa trên miền giá trị của một tiêu thức $a \in A$; $x \preccurlyeq_a y$ có nghĩ là *x* ít nhất tốt bằng *y* đối với tiêu thức *a*, hay *x* trội hơn *y*. Không mất tính tổng quát, ta xét thuộc tính điều kiện và quyết định có miền giá trị số và theo ưu tiên tăng dần, nghĩa là $V_a \in R$ (*R* là tập số thực). Với $a \in A,\ x, y \in U$, ta định nghĩa

$x \preccurlyeq y \Leftrightarrow f(x,a) \geq f(y,a).$

Với một tập con thuộc tính $B \subseteq A$, ta định nghĩa $x \preccurlyeq_B y \Leftrightarrow \forall a \in B,\ x \preccurlyeq_a y$, có nghĩa là *x* trội hơn *y* đối với tất cả các thuộc tính trong *B*, ta ký hiệu $xR_B^{\geq}y$. Do vậy, hệ thông tin đơn trị xếp thứ tự theo ưu tiên tăng dần được biểu diễn $T^{\preccurlyeq} = (U, A \cup D, F, G)$.

Cho $T^{\phi} = (U, A \cup D, F, G)$ là hệ thông tin đơn trị xếp thứ tự, với $B \subseteq A$, ký hiệu:

$$R_B^{\succeq} = \left\{ (x_i, x_j) \in U \times U \mid f_l(x_i) \geq f_l(x_j), \forall a_l \in B \right\} \quad (1)$$

$$R_D^{\succeq} = \left\{ (x_i, x_j) \in U \times U \mid g_m(x_i) \geq g_m(x_j), \forall d_m \in D \right\} \quad (2)$$

(2) R_B^{ϕ} và R_D^{ϕ} được gọi là quan hệ trội của hệ thông tin T^{ϕ}. Nếu ta biểu diễn

$$[x_i]_B^{\mathrm{f}} = \left\{ x_j \in U \mid (x_j, x_i) \in R_B^{\succeq} \right\} = \left\{ x_j \in U \mid f_l(x_j) \geq f_l(x_i), \forall a_l \in B \right\}$$

$$[x_i]_D^{\mathrm{f}} = \left\{ x_j \in U \mid (x_j, x_i) \in R_D^{\succeq} \right\} = \left\{ x_j \in U \mid g_{ml}(x_j) \geq g_m(x_i), \forall d_m \in D \right\}$$

Thì ta thu được các tính chất sau đây của quan hệ trội:

Tính chất 2.1 [10] *Cho* $R_A^{\succeq}$ *là quan hệ trội*

(1) $R_A^{\succeq}$ không phải là quan hệ tương đương, vì chúng có tính phản xạ, bắc cầu nhưng không đối xứng.

(2) Nếu $B \subseteq A$ thì $R_B^{\succeq} \supseteq R_A^{\mathrm{f}}$.

(3) Nếu $B \subseteq A$ thì $[\mathrm{x}]_B^{\mathrm{f}} \supseteq [\mathrm{x}]_A^{\mathrm{f}}$.

(4) Nếu $\mathrm{x}_j \in [\mathrm{x}_i]_A^{\mathrm{f}}$ thì $[\mathrm{x}_j]_A^{\mathrm{f}} \subseteq [\mathrm{x}_i]_A^{\mathrm{f}}$ và

$$[\mathrm{x}_i]_A^{\mathrm{f}} = \cup \left\{ [\mathrm{x}_j]_A^{\mathrm{f}} : \mathrm{x}_j \in [\mathrm{x}_i]_A^{\mathrm{f}} \right\}.$$

(5) $[\mathrm{x}_j]_A^{\mathrm{f}} = [\mathrm{x}_i]_A^{\mathrm{f}}$ nếu và chỉ nếu $f(x_i, a) = f(x_j, a) \quad (\forall a \in A)$.

(6) $T = \cup \left\{ [x]_A^{\mathrm{f}} \mid x \in U \right\}$; tạo thành một bao phủ của *U*.

Với $X \subseteq U$ và $A \subseteq T^{\phi}$, xấp xỉ trên và xấp xỉ dưới của *X* đối với quan hệ trội R_A^{ϕ} được định nghĩa như sau:

$$\underline{R_A^{\mathrm{f}}}(X) = \left\{ x \in U \mid [x]_A^{\mathrm{f}} \subseteq X \right\};$$

$$\overline{R_A^{\mathrm{f}}}(X) = \left\{ x \in U \mid [x]_A^{\mathrm{f}} \cap X \neq \varphi \right\};$$

Các tập xấp xỉ trên quan hệ trội cũng có một số đặc tính tương tự như các tập xấp xỉ trên quan hệ tương đương trong lý thuyết tập thô truyền thống.

Khai phá luật xếp thứ tự

Mục tiêu của bài toán khai phá dữ liệu trên hệ thông tin đơn trị xếp thứ tự là tìm kiếm các luật xếp thứ tự về mặt ngữ nghĩa trên miền giá trị các thuộc tính.

Trong một *OIS*, một biểu thức nguyên tố trên thuộc tính *a* được định nghĩa (a, f) hoặc $(a, \underline{\mathrm{p}})$. Với tập thuộc tính $B \subseteq A$, một biểu thức trên *B* trong *OIS* được định nghĩa $\Lambda_{a \in B} e(a)$, với $e(a)$ là một biểu thức nguyên tố trên *a*. Tập các biểu thức trên *B* trong *OIS* ký hiệu là $E(B)$. Các biểu thức kết nối với nhau bởi các toán tử logic như $\neg$ và $\vee$, tuy nhiên, để đơn giản, ta chỉ dùng $\wedge$.

Xét các cặp đối tượng trong *OIS,* tập vũ trụ

$$(U \times U)^+ = U \times U - \{(x, x) \mid x \in U\} = \{(x, y) \mid x, y \in U, x \neq y\}$$

Ký hiệu tập $m(\phi)$ bao gồm tất cả các cặp đối tượng thỏa mãn biểu thức ϕ, ta có:

$m(a, \phi) = \{(x, y) \in (U \times U)^+ \mid f_a(x) \ \phi \ f_a(y)\}$

$m(a, \underline{\pi}) = \{(x, y) \in (U \times U)^+ \mid f_a(x) \ \underline{\pi} \ f_a(y)\}$,

$m(\underset{a \in A}{\Lambda} e(a)) = \underset{a \in A}{\cap} m(e(a))$.

Một cặp đối tượng *x, y* thỏa mãn biểu thức ϕ, viết là $(x, y) \models \phi$, nếu thứ tự xác định bởi biểu thức ϕ là (x, y). Với tập biểu thức E(A), họ $\{m(\phi) \neq \varnothing \mid \phi \in E(A)\}$ tạo thành một phân hoạch của $(U \times U)^+$, ký hiệu là P(A). Mỗi cặp đối tượng thỏa mãn một và chỉ một biểu thức trong E(A).

Định nghĩa 2.3. Cho $T = (U, A \cup D, F, G)$ là hệ thông tin đơn trị xếp thứ tự. Xét hai tập thuộc tính $B, C \subseteq A \cup D$.

Với hai biểu thức $\phi \in E(B)$ và $\psi \in E(C)$, một luật xếp thứ tự đọc là "Nếu ϕ thì ψ", ký hiệu $\phi \Rightarrow \psi$. Biểu thức ϕ gọi là tiền tố (vế trái) của luật, biểu thức ψ gọi là hậu tố (vế phải) của luật.

Một luật xếp thứ tự diễn tả thứ tự các đối tượng trên tập thuộc tính B xác định thứ tự các đối tượng trên tập thuộc tính C.

Ví dụ, một luật xếp thứ tự:

$(a,\ \mathrm{f}) \wedge (b,\ \underline{\mathrm{p}}) \Rightarrow (c,\ \mathrm{f})$.

được diễn giải

$x \phi_{\{a\}}\ y \wedge x \underline{\pi}_{\{b\}} y \Rightarrow x \phi_{\{c\}}\ y$.

Nghĩa là, với hai đối tương *x* và *y* tùy ý, nếu *x* xếp trên *y* đối với thuộc tính *a*, và *x* xếp dưới *y* đối với thuộc tính *b* thì *x* xếp trên *y* đối với thuộc tính *c*.

Định nghĩa 2.4. Độ chính xác và độ bao phủ của một luật xếp thứ tự, $\phi \Rightarrow \psi$, được định nghĩa như sau [3], [11]:

Độ chính xác ($\phi \Rightarrow \psi$) = $\dfrac{|m(\varphi \wedge \psi)|}{|m(\varphi)|}$ (3)

Độ bao phủ ($\phi \Rightarrow \psi$) = $\dfrac{|m(\varphi \wedge \psi)|}{|m(\psi)|}$ (4)

Với $|\ |$ biểu diễn lực lượng của tập hợp.

Độ chính xác ($\phi \Rightarrow \psi$) là độ đo về sự đúng đắn của luật, và độ bao phủ ($\phi \Rightarrow \psi$) là độ đo về tính ứng dụng của luật. Một luật có độ bao phủ cao ngụ ý rằng luật thỏa mãn tiêu thức xếp thứ tự của nhiều cặp đối tượng. Độ chính xác và độ bao phủ không độc lập với nhau, chúng đều liên quan đến số lượng $|m(\phi \wedge \psi)|$. Một luật có độ bao phủ cao hơn có thể có độ chính xác thấp hơn và một luật có độ chính xác cao hơn có thể có độ bao phủ thấp hơn.

Để khai phá luật xếp thứ tự từ bảng thông tin đơn trị xếp thứ tự, ta sử dụng cách tiếp cận lý thuyết tập thô. Từ bảng thông tin đơn trị xếp thứ tự, ta xây dựng bảng thông tin nhị phân. Trong bảng thông tin nhị phân, ta xét tất cả các cặp đối tượng thuộc tích đề các $U \times U$. Hàm chuyển được định nghĩa như sau:

$$I_a((x,y)) = \begin{cases} 1, & x\ \mathrm{f}_{\{a\}}\ y \\ 0, & x \underline{\mathrm{p}}_{\{a\}} y \end{cases} \quad (5)$$

Các biểu diễn luật trên bảng thông tin xếp thứ tự được chuyển đổi thành các biểu diễn luật trên bảng thông tin nhị phân. Ví dụ: $x \phi_{\{a\}}\ y$ được chuyển thành $I_a((x,y)) = 1$. Trong quá trình chuyển đổi, ta không xét các cặp đối tượng (x, x).

Trong bảng thông tin nhị phân, ta định nghĩa một quan hệ tương đương E_B đối với tập con thuộc tính $B \subseteq A$:

$(x,y) E_B (x',y') \Leftrightarrow (\forall a \in B) I_a(x,y) = I_a(x',y')$.

Thuộc tính phân lớp xếp thứ tự $o \in D$ phân hoạch các cặp đối tượng thành hai lớp rời nhau Cl_o và Cl_1. Xấp xỉ trên và xấp xỉ dưới của Cl_i $(i = 1,2)$ trên tập thuộc tính *B* được xác định như sau:

$$\underline{apr}(Cl_i) = \cup \left\{ \left[(x,y)\right]_B \middle| \left[(x,y)\right]_B \subseteq Cl_i \right\},$$

$$\overline{apr}(Cl_i) = \cup \left\{ \left[(x,y)\right]_B \middle| \left[(x,y)\right]_B \cap Cl_i \neq o \right\},$$

với $\left[(x,y)\right]_B$ là lớp tương đương chứa (x,y) theo quan hệ tương đương E_B.

Với mỗi lớp tương đương $\left[(x,y)\right]_B \in \underline{apr}(Cl_i)$, ta có thể rút ra một luật xếp thứ tự chắc chắn như sau: $Des(\left[(x,y)\right]_B) \Rightarrow Des(Cl_i)$.

Với $Des(\left[(x,y)\right]_B)$ và $Des(Cl_i)$ biểu diễn mô tả của các lớp tương đương tương ứng.

Với mỗi thuộc tính xếp thứ tự $a \in B$, ta có thể lấy một biểu thức nguyên tố trong $Des(\left[(x,y)\right]_B) : (a, \mathrm{f})$ nếu $I_a((x,y)) = 1$, và $(a, \underline{\mathrm{p}})$ nếu $I_a((x,y)) = 0$. Sự kết hợp của các biểu thức nguyên tố như vậy $Des(\left[(x,y)\right]_B)$. *Des(Cl$_i$)* biểu diễn một trong hai biểu thức nguyên tố đối với thứ tự phân lớp: (o, f) nếu $i = 1$ và $(a, \underline{\mathrm{p}})$ nếu $i = 0$.

CHỌN MẪU ĐẠI DIỆN TRÊN HỆ THÔNG TIN ĐA TRỊ

Chọn mẫu đại diện thực chất là bước tiền xử lý dữ liệu trước khi thực hiện các thuật toán tìm tập thuộc tính tối ưu. Thay vì tìm tập thuộc tính tối ưu trên toàn bộ tập đối tượng ban đầu, chúng tôi tìm tập thuộc tính tối ưu trên tập đối tượng đại diện (chúng tôi gọi là mẫu đại diện) và chứng minh bằng lý thuyết tập thuộc tính tối ưu thu được từ mẫu đại diện tương đương với tập thuộc tính tối ưu thu được từ tập đối tượng ban đầu. Vì kích cỡ mẫu đại diện nhỏ

hơn nhiều so với kích cỡ tập dữ liệu ban đầu nên thời gian thực hiện thuật toán tìm tập thuộc tính tối ưu trên mẫu đại diện giảm thiểu đáng kể. Mẫu đại diện bao gồm các đối tượng đại diện, mỗi đối tượng đại diện được lựa chọn như sau:

Xét hệ thông tin đa trị $IS=(U,AT)$, trước hết chúng tôi phân hoạch tập đối tượng *U* ban đầu trên tập thuộc tính AT thành các lớp tương đương.

Hai đối tượng $u,v\in U$ thuộc cùng một lớp tương đương nếu $S_{\{a\}}(u)=S_{\{a\}}(v)$ với mọi $a\in AT$.

Với mỗi lớp tương đương, chúng tôi chọn ra một đối tượng đại diện cho lớp tương đương đó, không mất tính chất tổng quát, chúng tôi chọn đối tượng đầu tiên làm đại diện. Tập các đối tượng đại diện là mẫu đại diện được chọn.

Thuật toán chọn mẫu đại diện của hệ thông tin đa trị được mô tả như sau:

Thuật toán 1. Chọn mẫu đại diện của hệ thông tin đa trị.

Đầu vào: Hệ thông tin đa trị ban đầu $IS=(U,AT)$ với $U=\{u_1,...,u_n\}$, $AT=\{a_1,...,a_m\}$.

Đầu ra: Hệ thông tin đa trị mẫu $IS_P=(U_P,AT)$ với $U_P\subseteq U$ là một mẫu đại diện.

Bước 1: Đặt $U_P=\varnothing$;

Bước 2: Với mỗi $a_i\in AT, i=1..m$, tính phân hoạch $U/\{a_i\}=\left\{[u]_{\{a_i\}}\mid u\in U\right\}$

với $[u]_{\{a_i\}}=\left\{v\in U\mid S_{\{a_i\}}(u)=S_{\{a_i\}}(v)\right\}$.

Bước 3: Tính phân hoạch $U/AT=\left\{[u]_{AT}\mid u\in U\right\}$ với

$$[u]_{AT}=[u]_{\{a_1\}}\cap...\cap[u]_{\{a_m\}}=\bigcap_{i=1}^{m}[u]_{\{a_i\}}.$$

Giả sử $U/AT=\{X_1,...,X_k\}$ và $X_i=\{u_{i_1},...,u_{i_l}\}$ với $i=1..k$.

Bước 4: Với mọi $X_i\in U/AT$, $i=1..k$, đặt $U_P:=U_P\cup\{u_{i_1}\}$;

Bước 5: Return $IS_P=(U_P,AT)$;

Ví dụ 1. Cho hệ thông tin đa trị như (bảng 1)

Bảng 1. *Hệ thông tin đa trị*

U	a_1	a_2	a_3	a_4
u_1	{1}	{ 1}	{1}	{0}
u_2	{0}	{0, 1}	{1}	{0}
u_3	{0, 1}	{0, 1}	{0}	{1}
u_4	{1}	{0, 1}	{1}	{1}
u_5	{0, 1}	{0, 1}	{1}	{1}
u_6	{0}	{1}	{1}	{0, 1}
u_7	{0, 1}	{1}	{0}	{0, 1}
u_8	{0}	{1}	{1}	{0}
u_9	{0, 1}	{0, 1}	{0}	{1}

Ta có:

$S_{\{a_1\}}(u_1)=S_{\{a_1\}}(u_4)=\{u_1,u_3,u_4,u_5,u_7,u_9\}$,

$S_{\{a_1\}}(u_3)=S_{\{a_1\}}(u_5)=S_{\{a_1\}}(u_7)=S_{\{a_1\}}(u_9)=U$,

$S_{\{a_1\}}(u_2)=S_{\{a_1\}}(u_6)=S_{\{a_1\}}(u_8)$

$=\{u_2,u_3,u_5,u_6,u_7,u_8,u_9\}$

Do đó:

$U/\{a_1\}=\{\{u_1,u_4\},\{u_2,u_6,u_8\},\{u_3,u_5,u_7,u_9\}\}$

Tính toán tương tự, ta có $U/\{a_2\}=U$,

$U/\{a_3\}=\{\{u_1,u_2,u_4,u_5,u_6,u_8\},\{u_3,u_7,u_9\}\}$,

$U/\{a_4\}=\{\{u_1,u_2,u_8\},\{u_3,u_4,u_5,u_9\},\{u_6,u_7\}\}$

Từ đó ta có

$$U/AT=\left\{\begin{matrix}\{\{u_1\},\{u_2,u_8\},\{u_3,u_9\},\{u_4\},\\ \{u_5\},\{u_6\},\{u_7\}\end{matrix}\right\}$$

Tập đối tượng đại diện được chọn là $U_P=\{u_1,u_2,u_3,u_4,u_5,u_6,u_7\}$ và hệ thông tin đa trị đại diện $IS_P=(U_P,AT)$ được chọn ở Bảng 2.

Đánh giá độ phức tạp thuật toán:

Giả sử *k* là số thuộc tính điều kiện, *n* là số đối tượng. Xét *Bước 2*, với mỗi $a_i\in A, i=1..m$, độ phức tạp $S_{\{a_i\}}(u), u\in U$ là $O(n^2)$, độ phức tạp để tính phân hoạch $U/\{a_i\}$ là $O(n\log n)$. Do đó, độ phức tạp của Bước 2 là $O(kn^2)$. Độ phức tạp của Bước 3 khi bước 2 đã được tính là $O(n)$. Độ phức tạp của bước

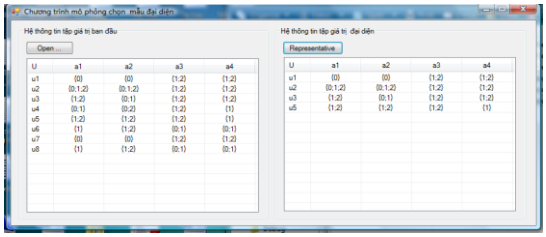
4 là $O(n\log n)$. Do đó, độ phức tạp của Thuật toán là $O(kn^2)$.

Bảng 2. *Hệ thông tin đa trị mẫu từ Bảng 1*

U	a_1	a_2	a_3	a_4
u_1	{1}	{ 1}	{1}	{0}
u_2	{0}	{0, 1}	{1}	{0}
u_3	{0, 1}	{0, 1}	{0}	{1}
u_4	{1}	{0, 1}	{1}	{1}
u_5	{0, 1}	{0, 1}	{1}	{1}
u_6	{0}	{1}	{1}	{0, 1}
u_7	{0, 1}	{1}	{0}	{0, 1}

Thực nghiệm minh họa thuật toán

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. Việc thực nghiệm Thuật toán 1 được thực hiện trên bộ số liệu tập giá 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, $|A|$ là số thuộc tính điều kiện. Các thuộc tính điều kiện được đánh số thứ tự từ 1 đến $|A|$.



Cho hệ thông tin đa trị ban đầu $IS=(U,AT)$ và hệ thông tin đa trị mẫu $IS_P=(U_P,AT)$, trước hết bài báo chứng minh bổ đề sau:

Bổ đề 1. *Nếu* $u_p\in U$ *là một đối tượng đại diện được chọn trên* $IS=(U,AT)$ *sao cho* $S_B(u_p)\neq S_{AT}(u_p)$ *với* $B\subset AT$ *thì ta cũng có* $S_B(u_p)\neq S_{AT}(u_p)$ *trên* $IS_P=(U_P,AT)$ *với* $u_p\in U_p$.

Chứng minh. Trên $IS=(U,AT)$, giả sử $S_{AT}(u_p)=[u_p]_{AT}\cup X$, khi đó với mọi $u\in[u_p]_{AT}$ ta đều có $S_{AT}(u)=S_{AT}(u_p)$.

Từ $S_B(u_p)\neq S_{AT}(u_p)$ suy ra $S_B(u_p)=S_{AT}(u_p)\cup Y$. Xét đối tượng bất kỳ $y\in Y$, vì $y\notin S_{AT}(u_p)$ nên $y\notin S_{AT}(u)$ với mọi $u\in[u_p]_{AT}$, do đó $S_{AT}(y)$ không chứa u với mọi $u\in[u_p]_{AT}$, nghĩa là trên $IS_P=(U_P,AT)$, $S_{AT}(y_p)$ không chứa u_p với y_p là đối tượng đại diện của lớp tương đương chứa y trên $IS=(U,AT)$ **(i)**.

Mặt khác, từ giả thiết $S_{AT}(u_p)=[u_p]_{AT}\cup X$, với $x\in X$ thì $x\in S_{AT}(u)$ với mọi $u\in[u_p]_{AT}$, hay $S_{AT}(x)$ chứa u với mọi $u\in[u_p]_{AT}$. Với đối tượng y được xét ở trên rõ ràng $y\notin[u_p]_{AT}$, giả sử $y\in[x]_{AT}$ với $x\in X$ khi đó $S_{AT}(y)=S_{AT}(x)$ và $S_{AT}(y)$ chứa u với mọi $u\in[u_p]_{AT}$, nghĩa là trên $IS_P=(U_P,AT)$, $S_{AT}(y_p)$ chứa u_p với y_p là đối tượng đại diện của lớp tương đương chứa y, điều này mâu thuẫn với **(i)**. Do đó $y\notin[x]_{AT}$ với mọi $x\in X$.

Với giả thiết $S_{AT}(u_p)=[u_p]_{AT}\cup X$ thì trên $IS_P=(U_P,AT)$, $S_{AT}(u_p)=\{u_p\}\cup X_p$ với X_p là tập các đối tượng đại diện của các đối tượng thuộc X. Với giả thiết $S_B(u_p)=S_{AT}(u_p)\cup Y$ và kết quả chứng minh $y\in Y$, $y\notin[x]_{AT}$ với mọi $x\in X$ thì trên $IS_P=(U_P,AT)$, $S_B(u_p)=\{u_p\}\cup X_p\cup Y_p$ với $y_p\in Y_p$ và y_p là đối tượng đại diện của

$y \in Y$. Do đó ta kết luận trên $IS_P = (U_P, AT)$, $S_B(u_p) \neq S_{AT}(u_p)$, (đpcm)

Từ kết quả của Bổ đề 1, tác giả chứng minh rằng tập thuộc tính tối ưu của hệ thông tin đa trị ban đầu và tập thuộc tính tối ưu của hệ thông tin đa trị mẫu là như nhau.

Giả sử $R \subseteq AT$ là tập thuộc tính tối ưu của hệ thông tin đa trị ban đầu $IS = (U, AT)$, khi đó $S_R(u) = S_{AT}(u)$ với mọi $u \in U$ và $\forall B \subset R$ tồn tại $u \in U$ sao cho $S_B(u) \neq S_{AT}(u)$.

a) Từ $S_R(u) = S_{AT}(u)$ với mọi $u \in U$ trên $IS = (U, AT)$ dễ dàng suy ra $S_R(u_p) = S_{AT}(u_p)$ với mọi $u_p \in U_P$ trên $IS_P = (U_P, AT)$.

b) Không mất tính tổng quát, giả sử $B \subset R$ và tồn tại $u \in U$ sao cho $S_B(u) \neq S_{AT}(u)$ trên $IS = (U, AT)$

Nếu u là đối tượng đại diện được chọn thì $u = u_p$ và $S_B(u) \neq S_{AT}(u)$ trên $IS = (U, AT)$, theo Bổ đề 1 thì $S_B(u_p) \neq S_{AT}(u_p)$ trên $IS_P = (U_P, AT)$ **(i)**.

Nếu u không phải đối tượng đại diện thì trên $IS = (U, AT)$, giả sử u_p là đối tượng đại diện của lớp tương đương $[u_p]_{AT}$ chứa u và u_p, khi đó $[u_p]_{AT} = [u]_{AT}$. Do $B \subset R \subseteq AT$ nên từ $[u_p]_{AT} = [u]_{AT}$ ta cũng suy ra $[u_p]_B = [u]_B$. Từ $[u_p]_{AT} = [u]_{AT}$ ta có $[u_p]_{\{a_i\}} = [u]_{\{a_i\}}$ với mọi $a_i \in AT$, theo cách xây dựng phân hoạch ta có $S_{\{a_i\}}(u_p) = S_{\{a_i\}}(u)$ với mọi $a_i \in AT$, do đó

$$S_{AT}(u_p) = \bigcap_{i=1}^{m} S_{\{a_i\}}(u_p) = \bigcap_{i=1}^{m} S_{\{a_i\}}(u) = S_{AT}(u).$$

Từ $[u_p]_B = [u]_B$, bằng cách tương tự ta suy ra $S_B(u_p) = S_B(u)$. Theo giả thiết, $S_B(u) \neq S_{AT}(u)$ nên ta thu được $S_B(u_p) \neq S_{AT}(u_p)$ trên $IS = (U, AT)$, theo Bổ đề 1 thì ta cũng có $S_B(u_p) \neq S_{AT}(u_p)$ trên $IS_P = (U_P, AT)$ **(ii)**

Như vậy, cả hai trường hợp **(i)** và **(ii)** ta đều có $S_B(u_p) \neq S_{AT}(u_p)$ trên $IS_P = (U_P, AT)$, từ đó kết luận tồn tại $B \subset R$ sao cho $S_B(u_p) \neq S_{AT}(u_p)$. Từ a) và b) theo định nghĩa ta có $R \subseteq AT$ là một tập thuộc tính tối ưu của hệ thông tin đa trị mẫu $IS_P = (U_P, AT)$.

KẾT LUẬN

Bài báo đã đề xuất thuật toán chọn mẫu đại diện trong hệ thông tin đa trị sử dụng lý thuyết tập thô. Đồng thời bài báo trình bày khai phá các luật xếp thứ tự bằng phương pháp chuyển đổi hệ thông tin đơn trị xếp thứ tự thành hệ thông tin nhị phân, từ đó áp dụng các kỹ thuật khai phá luật sử dụng lý thuyết tập thô truyền thống. Định hướng nghiên cứu tiếp theo là đề xuất các phương pháp tìm tập thuộc tính tối ưu hiệu quả trên hệ quyết định đa trị.

TÀI LIỆU THAM KHẢO

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

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

3. S. Tsumoto, Modelling medical diagnostic rules based on rough sets, *Rough Sets and Current Trends in Computing, Lecture Notes in Artificial Intelligence*, 1424, Springer-Verlag, Berlin, pp. 475-482, 1998.

4. Lang G. M., Lia Q. G., Data compression of dynamic set-valued information systems, *CoRR abs/1209.6509*, 2012.

5. Wang C. Z., Chen D. G., Wuc C., Hu Q. H., Data compression with homomorphism in covering information systems, *International Journal of Approximate Reasoning* 52, 2011, pp. 519–525.

6. Wang C. Z., Wua C. X., Chenb D. G., Duc W. J., Some properties of relation information

systems under homomorphisms, *Applied Mathematics Letters* 21, 2008, pp. 940–945.
7. Guan Y. Y., Wang H. K, Set-valued information systems, *Information Sciences* 176, 2006, pp. 2507–2525.
8. Qian Y. H., Dang C. Y., Liang J. Y., Tang D. W., Set-valued ordered information systems, *Information Sciences* 179, 2009, pp. 2809-2832.
9. Kryszkiewicz M., Rough set approach to incomplete information systems, *Information Science*, Vol. 112, 1998, pp. 39-49.
10. W.X. Zhang, W.Z. Wu, J.Y. Liang, D.Y.Li, *Theory Method of Rough sets*, Science Press, Beijing, 2001.
11. Y.Y. Yao, N. Zhong, An analysis of quantitative measures associated with rules, *Proceedings of PAKDD'99*, 479-488, 1999
12. The UCI machine learning repository, http://archive.ics.uci.edu/ml/datasets.html

SUMMARY

DATA MINING ON SET- VALUED INFORMATION SYSTEMS

Phung Thi Thu Hien*
University of Economic and Technical Industries

Based on the idea of minimizing the original data set, in this paper, we propose a method of selecting representative object set from initial object set to the solve optimal set of attributes problem in set-valued information systems. We demonstrate that the optimal set of attributes on the original objects and the optimal set of attributes on the representative one are equivalent, therefore we confirm the correctness of the method. Because the representative sample size is smaller than the original object's size, the execution time of algorithms for finding the optimal attribute set on the representative sample is significantly reduced. Representative sample size is large or small depending on the specificity of each real-time information system. At the same time, the article presents the method of exploring ordinal law by converting ordinal monopole information system into binary monopole information system and applying the law biotechnology technique in the systematic set theory based on the binary monotherapy obtained.

Keywords: *Set-valued information system, rough set, the optimal set of attributes, tolerance relation.*

Ngày nhận bài: 30/7/2018; Ngày phản biện: 5/8/2018; Ngày duyệt đăng: 16/9/2018

* *Tel: 0914 770070, Email: Thuhiencn1@gmail.com*