ISSN 2354-0575

CÁC THUẬT TOÁN TÌM CÁC RÚT GỌN CHO HỆ TIN ĐƠN TRỊ VÀ ĐA TRỊ SỬ DỤNG KHÁI NIỆM VÙNG DƯƠNG

Nguyễn Hữu Đông[1], **Nguyễn Bá Tường**[1], **Nguyễn Đức Thọ**[2]
1 Trường Đại học Sư phạm Kỹ thuật Hưng Yên
2 Học viện Kỹ thuật Quân sự

Ngày nhận: 05/4/2016
Ngày sửa chữa: 03/6/2016
Ngày xét duyệt: 20/6/2016

Tóm tắt:

Trong bài này chúng tôi trình bày một số khái niệm và tính chất liên quan đến vùng dương trong lý thuyết tập thô của Pawlak. Trên cơ sở các tính chất của vùng dương, chúng tôi nêu ra một số các ràng buộc giữa các thuộc tính và đặc biệt giữa các thuộc tính điều kiện trong hệ quyết định để làm tiền đề cho các thuật toán tìm rút gọn cho hệ tin giá trị đơn và hệ tin giá trị tập.

Đồng thời trong bài viết này chúng tôi cũng đã minh chứng hệ tin đa trị (a set-value information system) cũng có thể xét như một hệ tin đơn trị.

Từ khóa: *Tập thô, vùng dương, hệ quyết định, hệ thống thông tin, khai thác dữ liệu.*

Mở đầu

Trong [1] Guangming Lang và cộng sự đã dùng phương pháp nén như là một cách rút gọn dữ liệu trong hệ tin giá trị tập. Trong bài này chúng tôi trình bày một số khái niệm và tính chất liên quan đến vùng dương trong lý thuyết tập thô của Pawlak. Trên cơ sở các tính chất của vùng dương, chúng tôi nêu ra một số các ràng buộc giữa các thuộc tính và đặc biệt giữa các thuộc tính điều kiện trong hệ quyết định để làm tiền đề cho các thuật toán tìm rút gọn cho hệ tin giá trị đơn và hệ tin giá trị tập.

1. Một số khái niệm cơ bản

Định nghĩa 1. Hệ thống thông tin

Hệ thống thông tin (information system) là S = (U, A); trong đó U là tập hữu hạn khác rỗng các đối tượng; A là tập hữu hạn khác rỗng các thuộc tính. Mỗi thuộc tính a $\in$ A, V_a là tập giá trị của a và u $\in$ U a(u) là giá trị của u tại thuộc tính a.

Chú ý: Nếu $\forall$a $\in$ A, $\forall o \in U$ a(o) chỉ có một giá trị thì $S = (U, A)$ là *hệ tin đơn trị*, ngược lại $S = (U, A)$ gọi là *hệ tin đa trị* hay *hệ tin giá trị tập (set-value information system)*.

Ví dụ Bảng 1 là hệ tin đơn trị, Bảng 2 là hệ tin đa trị.

Trong bài viết này khi ta nói cho hệ tin $S = (U, A)$ thì S có thể là đơn trị hoặc đa trị.

Cho hệ tin $S = (U, A)$, B $\subseteq$ A.

Định nghĩa 2. Quan hệ bất khả phân biệt

Quan hệ $IND(B) \subseteq U \times U$ được gọi là quan hệ bất khả phân biệt trên U nếu với mọi cặp đối tượng $o, o' \in U$ thì $o\ IND(B)\ o'$ khi và chỉ khi a(o) = a(o') với mọi a$\in$B.

Dễ dàng thấy rằng quan hệ $IND(B)$ là quan hệ tương đương trên U. Phân hoạch $U/IND(B) = U/B$ là phân hoạch tương đương.

Chú ý: Chúng ta sẽ ký hiệu U/B là phân hoạch của U/IND(B) và U/B = {$[o]_B$: o$\in$U} là các nhóm tương đương. Với $[o]_B$ là nhóm các đối tượng quan hệ với nhau.

Định nghĩa 3. Hệ quyết định

Hệ quyết định là hệ tin S mà trong tập thuộc tính A có thuộc tính quyết định D.

Vậy hệ quyết định $T = (U, A)$; trong đó $A = C \cup D$; $C \cap D \neq \phi$. Tập C được gọi là tập thuộc tính điều kiện, D là thuộc tính quyết định.

Ví dụ:

Bảng 1. *Hệ quyết định đơn trị*

U	C_1	C_2	C_3	C_4	C_5	D
u_1	1	2	1	2	1	0
u_2	1	2	1	2	2	0
u_3	1	2	1	2	1	0
u_4	2	2	1	2	2	0
u_5	2	3	4	3	1	0
u_6	3	3	4	3	2	2
u_7	3	3	4	3	1	0

Bảng 2. *Hệ quyết định đa trị*

U	C_1	C_2	C_3	C_4	C_5	D
u_1	{1}	{1,2}	{2}	{1,2}	{1}	0

u_2	{1}	{1,2}	{2}	{1,2}	{2}	0
u_3	{1}	{1,2}	{2}	{1,2}	{1}	0
u_4	{1,2}	{1,2}	{1,2}	{1,2}	{2}	0
u_5	{1,2,3}	{1,3}	{1,4}	{1,3}	{1}	1
u_6	{1,2,3}	{1,3}	{1,4}	{1,3}	{2}	1
u_7	{1,2,3}	{1,3}	{1,4}	{1,3}	{1}	1

Chú ý: Trong hệ quyết định đa trị $\forall o \in U\ o[D]$ chỉ có một giá trị.

Định nghĩa 4. Hệ quyết định nhất quán

Hệ quyết định $T = (U, C \cup D)$ là nhất quán nếu mọi cặp $x, y \in U$ mà $x[C] = y[C]$ thì $x[D] = y[D]$. Nói cách khác T là nhất quán nếu các đối tượng giống nhau trên C thì giống nhau trên D. Ví dụ các Bảng 1, 2 là các hệ quyết định nhất quán.

Định nghĩa 5. Vùng dương của hai tập thuộc tính B, B'

Cho hệ tin $S = (U, A)$, $B, B' \subseteq A$.

Vùng dương (positive region) của B và B', với $B, B' \subseteq A$, ký hiệu $POS(B, B')$ là hợp của các nhóm của U/B được bao hàm trong các nhóm của U/B'. Hay $POS(B, B') = \bigcup\{E_i \in U/B\text{: } \exists P_j \in U/B'$ *sao cho* $E_i \subseteq P_j\}$.

Chú ý: Vùng dương của B và B' cho ta một khung nhìn về độ bao hàm của các tập sơ cấp trong hai không gian $Apr = (U, U/B)$ và $Apr' = (U, U/B')$.

Định nghĩa 6. Phụ thuộc hàm với độ phụ thuộc k(B,B')

Cho hệ tin $S = (U, A)$, $B, B' \subseteq A$.

Tập B' được gọi là phụ thuộc hàm độ $k(B,B')$ vào B, ký hiệu $B \xrightarrow{k(B,B')} B'$ nếu

$$k(B,B') = \frac{Card(POS(B,B'))}{Card(U)}$$

Định nghĩa 7. Rút gọn tập thuộc tính

Cho hệ tin $S = (U, A)$. Tập $R \subseteq A$ được gọi là *tập rút gọn của A* nếu R là tập tối thiểu thỏa mãn $U/R = U/A$.

R tối thiểu theo nghĩa với mọi $b \in R$ thì $U/(R \setminus \{b\}) \neq U/A$.

Định nghĩa 8. Rút gọn tập thuộc tính điều kiện

Cho hệ quyết định $T = (U, C \cup D)$. Tập $R \subseteq C$ được gọi là *tập rút gọn của C* nếu R là tập tối thiểu thỏa mãn $U/R = U/C$.

R tối thiểu theo nghĩa với mọi $b \in R$ thì $U/(R \setminus \{b\}) \neq U/C$.

2. Một số tính chất cơ bản của vùng dương và rút gọn

Tính chất 1. Sự bao nhau của các nhóm trên các tập thuộc tính bao nhau

Cho hệ tin $S = (U, A)$. Nếu $B \subseteq B' \subseteq A$ thì mọi $o \in U$ ta luôn có $[o]_{B'} \subseteq [o]_B$.

Chứng minh: Lấy $o' \in [o]_{B'}$ khi đó vì o' và o giống nhau (bất khả phân biệt) trên B' và $B \subseteq B'$ nên o và o' giống nhau trên B, hay $o' \in [o]_{B'}$ nên $[o]_{B'} \subseteq [o]_B$.

Tính chất 2. Cho hệ tin $S = (U, A)$. Với mọi $o \in U$ thì $o \in POS(B, B')$ khi và chỉ khi $[o]_B \subseteq [o]_{B'}$.

Chứng minh: tính chất 2 được suy trực tiếp từ định nghĩa vùng dương.

Tính chất 3. Biểu diễn vùng dương qua xấp xỉ dưới

Nếu đặt $E = U/B = \{E_1, E_2, ..., E_k\}$; $AprE = (U, E)$ và $P = U/B' = \{P_1, P_2, ..., P_l\}$; $AprP = (U, P)$ thì $POS(B,B') = \bigcup_{P_j \in P} (P_j)_E$ và $POS(B,B') = \bigcup_{E_i \in E} (E_i)_P$.

Chứng minh: Tính chất 3 được suy trực tiếp từ định nghĩa vùng dương và xấp xỉ dưới.

Tính chất 4. Vùng dương của rút gọn R và D bằng vùng dương của C và D

Cho hệ quyết định $T = (U, C \cup D)$.

Nếu R là rút gọn của C thì $POS(R, D) = POS(C, D)$.

Chứng minh: Giả sử $P = U/D = \{P_1, P_2, ..., P_l\}$ là *phân hoạch quyết định*.

Vì R là rút gọn của C nên $E = U/R = U/C = \{E_1, E_2, ..., E_k\}$. Khi đó theo tính chất 1 ta có $POS(C,D) = \bigcup_{p_j \in P} (P_j)_E = POS(R,D)$.

Tính chất 5. Độ phụ thuộc của tập rút gọn

Cho hệ quyết định $T = (U, C \cup D)$. Nếu R là rút gọn của C thì $k(R, D) = k(C, D)$.

Chứng minh: tính chất 5 suy trực tiếp từ tính chất 4.

Tính chất 6. Số các nhóm đối tượng liên quan đến các tập thuộc tính

Cho hệ tin $S = (U, A)$.

Nếu B và B' là hai tập thuộc tính thỏa mãn $B \subseteq B'$ thì $card(U/B) \leq card(U/B')$.

Chứng minh: Vì mỗi nhóm của U/B' là một nhóm con của U/B nên số nhóm của U/B không thể vượt quá số nhóm của U/B'.

Tính chất 7. Sự đồng biến của hàm độ đo phụ thuộc

Cho hệ quyết định $T = (U, C \cup D)$. Hàm

$k(B,D): 2^C \to [0, 1]$ với 2^C là họ các tập con của C và $k(B,D) = \frac{Card(POS(B,D)}{Card(U)}$ là hàm đồng biến.

Chứng minh: Để chứng minh tính chất 7 ta chỉ cần chứng minh với mọi cặp tập thuộc tính điều kiện *B, B'* mà $B \subseteq B'$ thì $POS(B, D) = POS(B', D)$.

Lấy $o \in POS(B, D)$ khi đó $[o]_B \subseteq [o]_D$. Mặt khác vì $B \subseteq B'$ nên theo tính chất 1 ta có $[o]_{B'} \subseteq [o]_B$. Vậy $[o]_{B'} \subseteq [o]_D$ hay $o \in POS(B', D)$.

Tính chất 8. Cho hệ quyết định $T = (U, C \cup D)$. Nếu đặt $w(c) = k(\{c\}, D)$ là trọng số của thuộc tính $c \in C$ và $w(B) = k(B, D)$ là trọng số của tập thuộc tính $B (B \subseteq C)$ thì $w(c) \le w(B)$ với mọi $c \in B$.

Chứng minh tính chất 8 suy ra từ tính chất 7.

3. Một số thuật toán tìm rút gọn

Cho hệ quyết định $T = (U, C \cup D)$.

Từ tính chất 7 hàm $k(B,D) = \frac{Card(POS(B,D))}{Card(U)}$ là hàm đồng biến.

Nên $k(C,D) = \frac{Card(POS(C,D))}{Card(U)}$ đạt giá trị cực đại.

Nếu *R* là rút gọn của *C* thì từ tính chất 4 ta có $k(R,D) = k(C,D)$.

Đặt $k = k(C,D)$. Đặt $w(c) = k(\{c\}, D)$ với $c \in C$ là trọng số của c.

Thuật toán 1. Tính xấp xỉ dưới X_E của X trong không gian Apr = (U, E)

Input Tập đối tượng *U*; phân hoạch $E = \{E_1, E_2, ..., E_k\}$; $X \subseteq U$.

Output xấp xỉ dưới X_E của *X* trong *Apr* = (*U, E*)

Algorithm

1. $X_E = \phi$

2. for i = 1 to k thực hiện if $E_i \subseteq X$ thì $X_E = X_E \cup E_i$.

Thuật toán 1 trên có độ phức tạp là O(k).

Thuật toán 2. Thuật toán tìm phân hoạch U/B

Input Hệ tin $S = (U, A)$, $B \subseteq A$; $Card(B) = j$.

Output $U/B = \{E_1, E_2, ..., E_k\}$.

Algoritm

1. Coi mỗi đối tượng $o \subseteq U$ trên tập thuộc tính B là một véc tơ o[B] hoặc một từ trên tập V: $o[B] = (v_1, v_2, ..., v_j)$;

2. Sắp xếp U theo thứ tự từ điển trên B;

3. Đặt $E = \{E_1, E_2, ..., E_k\}$ là họ các nhóm sau khi sắp xếp ta được phân hoạch cần tìm.

Chú ý: Phép sắp xếp m từ có độ dài $Card(B)$ với độ phức tạp là $O(Card(B).m\log m)$. Nếu đặt $Card(U) = m$, $Card(B) = k$ ta có độ phức tạp của thuật toán 2 là $O(k.m\log m)$.

Thuật toán 3. Thuật toán tìm k(C, D)

Input Hệ quyết định $T = (U, C \cup D)$; $Card(U) = m$; $Card(C) = n$; $Card(D) = l$.

Ouput $k(C,D)$.

Algorithm

1. Tính *U/C* ta được $U/C = \{X_1, X_2, ..., X_k\} = E$

2. Tính *U/D* ta được $U/D = \{Y_1, Y_2, ..., Y_k\}$

3. POS = ϕ

4. for i = 1 to l thực hiện

Begin

Tính xấp xỉ dưới $(Y_i)_E$ của Y_i trong không gian $Apr = (U, U/C) = (U, E)$

$POS = POS \cup (Y_i)_E$

End

5. Tính $k(C,D) = POS/Card(U)$.

Chú ý: Theo nguyên lý cộng độ phức tạp của thuật toán 3 là $O(nm\log m)$ với $n = Card(C); m = Card(U)$.

Thuật toán 4. Tính rút gọn R dựa vào các tập thuộc tính bất khả phân biệt

Input Hệ quyết định $T = (U, C \cup D)$.

Output *R* là rút gọn của *C*

Algorithm

1. Tính nửa trên của ma trận phân biệt $M = (t_{ij})$ với $j > i$ và $t_{ij} = \{c \in C: o_i[c] = o_j[c]\}$.

2. Đặt Mmax là họ các tập cực đại của M (phần tử cực đại của M là phần tử không bị chứa trong phần tử khác của M).

3. Đặt $R = C$

4. for each $c \in R$ if $R \backslash \{c\}$ không là tập con của phần tử nào trong Mmax thì $R = R \backslash \{c\}$.

5. Kết thúc vòng lặp ta có một rút gọn của *C*.

Chú ý: Bước một thuật toán ta có thời gian tính ma trận là $O(m^2)$. Bước 2 có thời gian tính là $O(m)$. Ở bước 4 thời gian tính là $O(n)$. Theo nguyên lý cộng thời gian tính hay độ phức tạp của thuật toán 6 là $O(\max\{m^2, n\})$. Độ phức tạp phụ thuộc vào hệ quyết định có tập đối tượng lớn hay tập thuộc tính lớn. Đặt $l = \max\{m^2, n\}$ với $n = Card(C); m = Card(U)$; khi đó độ phức tạp của thuật toán 6 là $O(l)$.

4. Kết luận

Trong bài viết này chúng tôi đã giới thiệu một số nghiên cứu, tính chất có tính hệ thống, cơ bản của vùng dương, độ phụ thuộc, ràng buộc của các tập thuộc tính trong hệ tin, hệ quyết định, trên cơ sở đó làm nền để tính rút gọn. Đồng thời trong bài viết này chúng tôi cũng đã minh chứng hệ tin đa trị (a set-value information system) cũng có thể xét như một hệ tin đơn trị.

Tài liệu tham khảo

[1]. Guangming Lang, Quingguo Li, *Data Compression of Dynamic Set-valued Inforrmation Systems*, arXiv: 1209.6509v1 [cs.IT] 28 Sep 2012.
[2]. Acuna, E. (2003), *A Comparison of Filters and Wrappers for Feature Selection in Supervised Classification, R Package 1.0*. http://Math.uprm.edu/~edgar/dprep.html. Accessed on February 17, 2006.
[3]. Chen, D., Cui, D., Wang, C., Wang, Z. (2006), *A Rough Set-Based Hierarchical Clustering Algorithm for Categorical Data,* International Journal of Information Technology, Vol. 12, No.3, pp 149-159.
[4]. Deogun, J., Choubey, S., Raghavan, V. and Sever, H. (1998), *Feature Selection and Effective Classifers*, Journal of ASIS 49,5, pp 403-414.
[5]. Geng, L. and Hamilton, H. J. (2002). ESRS, *A Case Selection Algorithm Using Extended Semilarity-based Rough Sets*, Second IEEE International Coference on data Mining (ICDM'20).
[6]. Grochowski, M. and Jankowski, N. (2004), *Comparison of the Instance Selection Algorithms. II.Results and Comments*, In: ICAISC 2004, L. Rutkowski et al. (Eds.), LNAI 3070, pp 580-585, 2004.
[7]. Han, J., Hu,X., Lin, T.Y, *Feature Subset Selection Based on Relative Dependenccy between Attributes*, Rough Sets and Current Trends in Computing 2004, pp 176-185.
[8]. Hu, K., Lu, Y and Shi, C. (2003), *Feature ranking in Rough Set*, AI Communications, pp 41-50.
[9]. Jensen, R. and Shen, Q. (2003), *Finding Rough Set Reducts with Ant Colony Optimization*, Proceeding of the 2001 UK Workshop on computational intelligence, 69-74.
[10]. 1. Jensen, R. and Shen, Q. (2004), *Fuzzy-Rough Set Attribute Reduction with Application to Web Categorization*, Fuzzy Sets and System, vol.141, no 3, pp 469-485.
[11]. Shen, Q. and Chouchooulas, A. (2001), *Rough Set-based Dimensionality Reduction for Supervised and Unsupervised Learning*, Int. J. Applied Mathematics Computational. Science. Vo. 11, N. 3, pp 583-601.
[12]. Z. Pawlak (1991), *Rough Sets: Theoretical Aspect of Reasoning about Data.*

ALGORITHMS FINDING REDUCTIONS FOR A SINGLE AND A SET-VALUE INFORMATION SYSTEM USING THE CONCEPT OF POSITIVE REGION

Abstract:

This paper studies some concepts and properties of positive region of Pawlak's rough set. Consequently, we propose some constraints among attributes, especially among conditional attributes which are in decision systems. We then introduce algorithms to find the reductions for both a single and a set-value information system. Furthermore, we will prove that in the proposed algorithm a set-value information system can be considered as a single information system.

Keywords: *Rough set, positive region, information system, decision systems, data mining.*