ISSN: 1859-2171
e-ISSN: 2615-9562
TNU Journal of Science and Technology 225(06): 200 - 204

NGHIÊN CỨU CẢI TIẾN MỘT SỐ ĐỘ ĐO TRONG LÝ THUYẾT TẬP THÔ CHO BẢNG QUYẾT ĐỊNH KHÔNG ĐẦY ĐỦ

Nguyễn Anh Tuấn
Trường Cao đẳng Vĩnh Phúc

TÓM TẮT

Các phương pháp giảm bớt thuộc tính đều sử dụng các độ đo qua các lớp dung sai, như phương pháp sử dụng độ đo lượng thông tin (information quantity). Để giải quyết bài toán giảm bớt thuộc tính trực tiếp trên các bảng quyết định không đầy đủ và đánh giá sự thay đổi giá trị độ chắc chắn, độ nhất quán, độ hỗ trợ. Các độ đo này gặp khó khăn trong việc đánh giá tính hiệu quả (về khả năng phân lớp hay độ hỗ trợ của tập luật)... Trong bài báo này, tác giả cải tiến một số độ đo nhằm nâng cao hiệu năng trong lý thuyết tập thô cho bảng quyết định không đầy đủ và chứng minh tính đúng đắn của các độ đo đề xuất.

Từ khóa: *Lý thuyết tập thô, Độ đo, bảng quyết định không đầy đủ, nâng cao hiệu năng, Giảm bớt thuộc tính.*

Ngày nhận bài: 23/12/2019; Ngày hoàn thiện: 13/5/2020; Ngày đăng: 20/5/2020

RESEARCH ON IMPROVING SOME MEASUREMENTS IN CRUDE THEORETICAL OF INCOMPLETE DECISION TABLES

Nguyen Anh Tuan
Vinh Phuc College

ABSTRACT

Attribute reduction methods uses measurements by tolerance layers, such as the measurement of information quantity. To solve the problem of reducing attributes directly on the incomplete decision tables and assessing changes in the value of certainty, consistency, support. These measures have difficulty in assessing the effectiveness (in terms of the ability to classify or support the law set) ...In this paper, the author improved some measurements to improve performance in the crude theoretical for incomplete decision tables and proved the validity of the proposed measurements.

Keywords: *Tolerance Rough Set, measurements, incomplete decision tables, improve performance, attribute reduction.*

Received: 23/12/2019; Revised: 13/5/2020; Published: 20/5/2020

Email: tuanna573@gmail.com

1. Giới thiệu

Trong lý thuyết tập thô, lựa chọn một số tính năng quan trọng để quyết định có giảm bớt hay không là hết sức cần thiết. Giảm bớt các đối tượng là một quá trình để tìm tập hợp con tối ưu của các thuộc tính, giữ lại những thuộc tính quan trọng để đưa ra những quyết định chính xác nhất. Hầu hết các thuật toán giảm bớt các đối tượng đều dựa vào thông tin được thu thập từ miền dương [1]. Các phương pháp giảm bớt thuộc tính đều sử dụng các độ đo qua các lớp dung sai, như phương pháp sử dụng độ đo lượng thông tin [2]. Giảm bớt các thuộc tính dựa trên lý thuyết tập thô có chứa dữ liệu về các đối tượng đặc trưng bởi tập hợp các thuộc tính hữu hạn. Đối với một hệ thống thông tin, nếu các thuộc tính điều kiện và thuộc tính quyết định là khác nhau, thì nó được gọi là một hệ thống quyết định.

Mục đích của việc giảm bớt các thuộc tính là loại bỏ các thuộc tính dư thừa nhằm nâng cao tính hiệu quả của các thuật toán. J. 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 giảm bớt sử dụng lượng thông tin tăng thêm mờ. Thực nghiệm trên một số bộ dữ liệu mẫu cho thấy, độ chính xác phân lớp của các thuật toán FSCE, GAIN_RATION_AS_FRS cao hơn độ chính xác của các thuật toán sử dụng Entropy, lượng thông tin tăng thêm (gain ratio) theo tiếp cận thô truyền thống.

Trong thực tế, có nhiều cách giảm bớt cho một bảng quyết định, tuy nhiên mức giảm tối thiểu là NP-hard [4]. Do đó, có nhiều tác giả đề xuất những phương pháp khác nhau để làm giảm thuộc tính trong lý thuyết tập thô như: dựa trên miền dương, dựa trên ma trận, dựa trên độ đo entropy, tính toán hạt, dựa trên độ đo khoảng cách...

Trong bài báo này, tác giả nghiên cứu cải tiến cải tiến một số độ đo trong lý thuyết tập thô cho bảng quyết định không đầy đủ, nhằm đánh giá các phương pháp theo tiêu chuẩn khả năng phân lớp của tập rút gọn.

Cấu trúc bài báo như sau: Phần I Giới thiệu; Phần II: Cơ sở toán học; Phần III: Cải tiến độ đo để nâng cao hiệu năng trong bảng quyết định không đầy đủ. Phần IV: Kết luận và tài liệu tham khảo.

2. Cơ sở toán học

2.1.Định nghĩa hệ thông tin[5]

Hệ thông tin là $IS = (U, A)$ trong đó U là tập hữu hạn, khác rỗng các đối tượng; A là tập hữu hạn, khác rỗng các thuộc tính. Với mọi $u \in U$, $a \in A$ ký hiệu giá trị thuộc tính a tại đối tượng u là $a\ (u)$ thay vì (u, a).

Nếu $B = \{b_1, b_2, \dots, b_n\} \subseteq A$ là một tập con các thuộc tính thì ký hiệu bộ các giá trị $b_i(u)$ bởi $B(u)$.

Như vậy, nếu u và v là hai đối tượng, thì $B\ (u) = B\ (v)$ nếu $b_i(u) = b_i(v)$ với mọi $=1, \dots, n$.

Xét hệ thông tin $IS = (U, A)$. Mỗi tập con các thuộc tính $P \subseteq A$ xác định một quan hệ hai ngôi trên U, ký hiệu là $IND\ (P)$, xác định bởi:

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

$IND\ (P)$ là quan hệ P- không phân biệt được. Thấy rằng $IND\ (P)$ là một quan hệ tương đương trên U. Nếu $(u, v) \in IND\ (P)$ thì hai đối tượng u và v không phân biệt được bởi các thuộc tính trong P. Quan hệ tương đương $IND\ (P)$ xác định một phân hoạch trên U, ký hiệu là $U\ /IND\ (P)$ hay $U\ /P$. Ký hiệu lớp tương đương trong phân hoạch $U\ /P$ chứa đối tượng u là $[u]_p$, khi đó: $[u]_p = \{v \in U\ |(u, v) \in IND\ (P)\}$.

2.2 Đặt bài toán

Cho bảng quyết định không đầy đủ $IDS = \{(U, A \cup \{b\})\}$, với $U = \{x_1, x_2, \dots, x_n\}$.

Giả sử $\frac{U}{SIM\ (A)} = \{S_A(x_1), S_A(x_2), \dots S_A(x_n)\}$

và $\frac{U}{\{b\}} = \{(V_1), (V_2), \dots (V_m)\}$. với

$S_A(u_i) \in \frac{U}{SIM\ (A)}$;

$V_j \in \frac{U}{\{b\}}$ và $S_A(u_i) \cap V_j \neq \emptyset$.

Ký hiệu $des\ (S_A(u_i))$ và $des(V_j)$ lần lượt là các mô tả của lớp dung sai $S_A(u_i)$ và lớp tương đương V_j .

$$Z_{ij} : des\ (S_A(u_i)) \rightarrow des(V_j) \quad (1)$$

$$\mu\ Z_{ij} = S_A(u_i) \cap \frac{V_j}{S_A(u_i)} \quad (2)$$

$$s(Z_{ij}) = \frac{|S_A(u_i) \cap V_j|}{|U|} \quad (3)$$

$$\tau(Z_{ij}) = \frac{|S_A(u_i) \cap V_j|}{|V_j|} \quad (4)$$

Giả sử:

$$F = \frac{U}{\{b\}} = \{(V_1), (V_2), \dots (V_m)\} \quad (5)$$

là một phân lớp của U theo b, độ chính xác của phân lớp F theo A, ký hiệu là $\alpha_A(F)$, [6]:

$$\alpha_A(F) = \frac{\sum_{V_i \in \frac{U}{\{b\}}} |\underline{A}V_i|}{\sum_{V_i \in \frac{U}{\{b\}}} |\overline{A}V_i|} \quad (6)$$

Giả sử : IDS $= \{(U, A \cup \{b\})\}$ với $U = \{x_1, x_2, \dots, x_n\}$ và tập luật Red

$$Red = Z_{ij} \,|Z_{ij} : des\ (X_i) \rightarrow des(V_j) \text{ với } X_i \in MC_A; V_j \in \frac{U}{\{b\}}, i = 1 \dots n, j = 1 \dots m. \quad (7)$$

Khi đó:

Độ chắc chắn α của IDS:

$$\alpha(IDS) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{|X_i \cap V_j|}{|X_i|} \quad (8)$$

Độ nhất quán β của IDS:

$$\beta(IDS) = \frac{1}{m}\sum_{i=1}^{m}\left[1 - \frac{4}{|X_i|}\sum_{j=1}^{N_i}|X_i \cap V_j \,|\, \mu(Z_{ij})(1 - \mu(Z_{ij}))\right] \quad (9)$$

Độ hỗ trợ γ của IDS:

$$\gamma(IDS) = \sum_{j=1}^{n}\frac{|V_j|}{N_i|U|}\sum_{k=1}^{N_j}\frac{|X_k \cap V_j|}{|U|} \quad (10)$$

3. Cải tiến độ đo trong lý thuyết tập thô cho bảng quyết định không đầy đủ

Giả sử ta có : $IDS = \{(U, A \cup \{b\})\}$ với $U = \{x_1, x_2, \dots, x_n\}$ và tập luật Red

$$Red = Z_{ij} \,|Z_{ij} : des\ (X_i) \rightarrow des(V_j) \text{ với } X_i \in MC_A; V_j \in \frac{U}{\{b\}}, i = 1 \dots n, j = 1 \dots m.$$

3.1. Cải tiến các độ đo

Độ chắc chắn α của IDS:

$$\alpha(IDS) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{|S_A(u_i) \cap V_j|}{|S_A(u_i)|} \quad (11)$$

Độ nhất quán β của IDS:

$$\beta(IDS) = \frac{1}{n-1}\sum_{i=1}^{n}\left[\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{|S_A(u_i)\cap V_j|}{|S_A(u_i)|}\right]-\frac{1}{n-1} \quad (12)$$

Độ hỗ trợ γ của IDS:

$$\gamma(IDS) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|S_A(u_i)\cap V_j|}{n} \quad (13)$$

Ký hiệu N_i là số luật quyết định (số lớp quyết định) sinh bởi lớp dung sai $S_A(u_i)$.

Ta có:

$\alpha(IDS)$ đạt giá trị lớn nhất là 1 nếu $\mu(Z_{ij}) = 1$ với $\forall Z_{ij} \in Red$, nghĩa là IDS nhất quán, và $\alpha(IDS)$ nhỏ nhất là : $\frac{1}{n}$ nếu $N_i = n$, nghĩa là $m = U$ và $S_A(u_i) = U$ với mọi $u_i \in U$.

$\beta(IDS)$ đạt giá trị lớn nhất là 1 khi $\alpha(IDS$ đạt giá trị lớn nhất là 1 và $\beta(IDS$ nhỏ nhất là 0 khi $\alpha(IDS)$ đạt giá trị nhỏ nhất là $\frac{1}{n}$.

$\gamma(IDS)$ đạt giá trị lớn nhất là 1 nếu $S_A(u_i) = U$ với mọi $u_i \in U$. $\gamma(IDS)$ nhỏ nhất là $\frac{1}{n}$ nếu $S_A(u_i)$=$\{u_i\}$ với mọi $u_i \in U$.

3.2. Chứng minh tính đúng đắn độ đo đề xuất

Giả sử : $IDS = \{(U, A \cup \{b\})\}$, $IDS' = \{(U, B \cup \{b\})\}$, với $U = \{x_1, x_2, \dots, x_n\}$ và tập luật Red

$$Red = Z_{ij}\,|Z_{ij} : des\left(S_A(u_i)\right) \rightarrow des(V_j)$$

với $S_A(u_i) \in \frac{U}{SIM(A)}$; $V_j \in \frac{U}{\{b\}}, i = 1 \dots n, j = 1 \dots m.$

Nếu B $\subseteq$ A thì $\alpha(IDS) \geq \alpha(IDS')$; $\beta(IDS) \geq \beta(IDS')$; $\gamma(IDS) \leq \gamma(IDS')$

+ Độ chắc chắn α của IDS

Giả sử $N_i(A), N_i(B)$ tương ứng là số luật quyết định sinh bởi lớp dung sai: $S_A(u_i)$ và $S_B(u_i)$. Nếu B $\subseteq$ A thì $S_A(u_i) \subseteq S_B(u_i)$ với mọi $u_i \in U$.

Cho nên ta có thể suy ra: $N_i(A) \leq N_i(B)$.

Từ đó ta có:

$$\alpha(IDS) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{|S_A(u_i)\cap V_j|}{|S_A(u_i)|} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{N_i(A)} \geq \frac{1}{n}\sum_{i=1}^{n}\frac{1}{N_i(A)}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{N_i(B)}\sum_{j=1}^{N_i(B)}\frac{|S_B(u_i)\cap V_j|}{|S_B(u_i)|} = \alpha(IDS')$$

Do đó: $\alpha(IDS) \geq \alpha(IDS')$ (Điều phải chứng minh)

+ Độ nhất quán β của IDS:

Ta có:

$$\beta(IDS) = \frac{1}{n-1}\sum_{i=1}^{n}\left[\frac{1}{N_i}\sum_{j=1}^{N_i}\frac{|S_A(u_i)\cap V_j|}{|S_A(u_i)|}\right]-\frac{1}{n-1}$$

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

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}\frac{1}{N_i(B)}-\frac{1}{n-1}\right)\sum_{j=1}^{N_i(B)}\left(\frac{|S_B(u_i)\cap V_j|}{|S_B(u_i)|}-\frac{1}{n-1}\right)$$

$$= \beta\,(IDS')$$

Do đó: $\beta\,(IDS) \geq \beta\,(IDS')$ (Điều phải chứng minh)

+ Độ hỗ trợ γ của IDS:

$$\gamma(IDS) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|S_A(u_i)\cap V_j|}{n}$$

Nếu B $\subseteq$ A thì $S_A(u_i) \subseteq S_B(u_i)$ với mọi $u_i \in U$.

Ta có $S_A(u_i)\cap V_j \subseteq S_B(u_i)\cap V_j$ với mọi $u_i \in U,\ V_j \in \frac{U}{\{b\}} \Leftrightarrow |S_A(u_i)\cap V_j| \leq |S_B(u_i)\cap V_j|$

$$\Leftrightarrow \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|S_A(u_i)\cap V_j|}{n} \leq \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|S_B(u_i)\cap V_j|}{n}$$

$\Leftrightarrow \gamma\,(IDS) \leq \gamma(IDS')$ Điều phải chứng minh

4. Kết luận

Trong bài báo này, tác giả đã nghiên cứu cải tiến một số độ đo và chứng minh tính đúng đắn, từ đó có thể lựa chọn nhóm phương pháp cho phù hợp và tiến thành thử nghiệm đánh giá sự thay đổi của các độ đo cải tiến trên bảng quyết định không đầy đủ. Trên cơ sở đó, lựa chọn và đánh giá các phương pháp giảm bớt thuộc tính dựa trên tiêu chuẩn độ hỗ trợ của tập luật.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1]. Y. H. Qian, J. Y. Liang, W. Pedrycz, and C. Y. Dang, "Positive approximation: an accelerator for attribute reduction in rough set theory," *Artif. Intell*, vol. 174, pp. 597-618, 2010.

[2]. J. Y. Liang and Y. H. Qian, "Information granules and entropy theory in information systems," *Information Sciences*, vol. 51, pp. 1-18, 2008.

[3]. J. Dai, and Q. Xu, "Attribute selection base on information gain ratio in fuzzy rough set theory with application to tumor classification," *Applied Soft Computing*, vol. 13, no. 2013, pp. 211-211, 2013.

[4]. A. Skowron, and C. Rauszer, *The discernibility matrices and functions in information systems.* Intelligent Decision Support, 1992.

[5]. Y. Leung, and D. Y. Li, "Maximal consistent block technique for rule acquisition in incomplete information systems," *Information Sciences*, vol. 153, pp. 85-106, 2003.

[6]. J. Y. Liang and Y. H. Qian, "Information granules and entropy theory in information systems," *Information Sciences*, vol. 51, pp. 1-18, 2008.