Tạp chí Tin học và Điều khiển học, T.28, S.2 (2012), 129–140

MỘT PHƯƠNG PHÁP MỚI RÚT GỌN THUỘC TÍNH TRONG BẢNG QUYẾT ĐỊNH KHÔNG ĐẦY ĐỦ SỬ DỤNG METRIC*

NGUYỄN LONG GIANG, NGUYỄN THANH TÙNG, VŨ ĐỨC THI

Viện Công nghệ thông tin, Viện Khoa học và Công nghệ Việt Nam

Tóm tắt. Trong hệ thông tin không đầy đủ, mỗi tập thuộc tính đều sinh ra một phủ trên tập các đối tượng, trong đó mỗi phần tử của phủ là một lớp dung sai. Như vậy, khi một metric nào đó được định nghĩa trên họ các phủ thì cũng có nghĩa là một metric đã được xác lập trên tập các thuộc tính. Một khi đã có metric, ta có thể đánh giá độ gần nhau giữa các thuộc tính, xác định thuộc tính quan trọng... Nhờ đó, có thể xây dựng thuật toán hiệu quả để giải quyết bài toán rút gọn thuộc tính.

Bằng việc xây dựng một metric trên họ các phủ xuất phát từ Liang entropy mở rộng, bài báo đề xuất một phương pháp mới rút gọn thuộc tính trong bảng quyết định không đầy đủ. Bằng lý thuyết và thực nghiệm, bài báo chứng minh phương pháp sử dụng metric hiệu quả hơn các phương pháp sử dụng lượng thông tin và ma trận dung sai.

Abstract. In incomplete information systems, each subset of attributes determines a cover on the set of objects, in which each element is a tolerance class. Thus, a metric which is defined on the family of covers is established on the attribute sets. Once a metric is established, we can use the metric to measure attributes distance, cluster and discover important attributes. As a result, effective algorithms are constructed to solve attribute reduction in incomplete information systems.

With metric on the family of covers based on generalized Liang entropy, this paper proposes a new method for attribute reduction in incomplete decision table. The paper proves theoretically and experimentally that this metric method is more effective than other methods based on information quantity and tolerance matrix.

1. MỞ ĐẦU

Rút gọn thuộc tính là bài toán quan trọng nhất trong lý thuyết tập thô. Trong những năm gần đây, các phương pháp rút gọn thuộc tính đã thu hút sự chú ý và quan tâm của nhiều nhà nghiên cứu [16]. Đáng chú ý là phương pháp dựa trên miền dương, phương pháp sử dụng ma trận phân biệt, phương pháp sử dụng entropy thông tin, phương pháp sử dụng các độ đo trong tính toán hạt...Tuy nhiên, hầu hết các phương pháp này đều thực hiện trên các hệ thông tin đầy đủ.

Trong các bài toán thực tế, hệ thông tin thường thiếu giá trị trên các thuộc tính, gọi là hệ

*Nghiên cứu này được hoàn thành dưới sự hỗ trợ từ Quỹ phát triển khoa học và Công nghệ quốc gia (NAFOSTED) mã số 102.01-2010.09

thông tin không đầy đủ. Xuất phát từ mô hình tập thô dung sai trên hệ thông tin không đầy đủ do Kryszkiewicz [4] đề xuất, nhiều nhà khoa học trên thế giới đã quan tâm nghiên cứu nghiên cứu các độ đo không chắc chắn [7, 8, 12, 13] và đề xuất các thuật toán tìm tập rút gọn.

Trong hệ thông tin không đầy đủ: Liang và các cộng sự [9] đề xuất thuật toán tìm tập rút gọn sử dụng entropy thô với độ phức tạp $O\left(|A|^2|U|\right)$ nếu bỏ qua độ phức tạp của việc tính các lớp dung sai; Chin và các cộng sự [1] đề xuất thuật toán tìm tập rút gọn sử dụng lượng khác nhau giữa các lớp dung sai với độ phức tạp $O\left(|A|^3|U|^2\right)$; Li và các cộng sự [5] đề xuất thuật toán tìm tập rút gọn sử dụng phép kết hạt của tri thức với độ phức tạp $O\left(|A|^3|U|^2\right)$. Trong bảng quyết định không đầy đủ: Huang và các cộng sự [3] đề xuất thuật toán tìm tập rút gọn sử dụng độ đo lượng thông tin của tri thức với độ phức tạp $O\left(|C|^3|U|^2\right)$; Huang, Zhou và các cộng sự [2, 18] đề xuất thuật toán tìm tập rút gọn sử dụng ma trận dung sai với độ phức tạp $O\left(|C|^3|U|^2\right)$.

Kỹ thuật sử dụng metric đóng vai trò quan trọng trong khai phá dữ liệu. Trong mấy năm gần đây, kỹ thuật này được nhiều người quan tâm nghiên cứu và áp dụng vào việc giải quyết các bài toán trong khai phá dữ liệu như phân lớp, phân cụm, lựa chọn đặc trưng... Điểm khác biệt của metric so với các độ đo không chắc chắn trong lý thuyết tập thô là metric cho phép đánh giá độ gần nhau của tri thức. Với tính chất như vậy, metric có thể được sử dụng hiệu quả để giải quyết bài toán rút gọn thuộc tính trong lý thuyết tập thô. Trên thế giới, nhóm nghiên cứu của Yuhua Qian và các công sự [14, 15] đã xây dựng các khoảng cách tri thức trên hệ thông tin không đầy đủ và nghiên cứu một số tính chất của chúng. Tuy nhiên, các kết quả nghiên cứu về việc sử dụng metric rút gọn thuộc tính trong hệ thông tin không đầy đủ còn nhiều hạn chế.

Bằng việc xây dựng một metric trên hệ thông tin không đầy đủ dựa vào độ đo Liang entropy mở rộng, bài báo đề xuất một phương pháp mới tìm tập rút gọn của bảng quyết định không đầy đủ sử dụng metric. Bằng lý thuyết và thực nghiệm, bài báo chứng minh phương pháp mới hiệu quả hơn phương pháp sử dụng lượng thông tin của tri thức [3] và phương pháp sử dụng ma trận dung sai [2, 18].

2. CÁC KHÁI NIỆM CƠ BẢN

Phần này trình bày một số khái niệm cơ bản về mô hình tập thô dung sai trên hệ thông tin không đầy đủ do Kryszkiewicz [4] đề xuất.

Hệ thông tin là một bộ tứ $S = (U, A, V, f)$ trong đó U là tập khác rỗng, hữu hạn các đối tượng; A là tập khác rỗng, hữu hạn các thuộc tính; $V = \prod_{a \in C \cup D} V_a$ với V_a là tập giá trị của thuộc tính $a \in A$; $f : U \times (A) \to V$ là hàm thông tin, với $\forall a \in A, u \in U$ hàm f cho giá trị $f(u, a) \in V_a$. Nếu V_a chứa giá trị rỗng thì S được gọi là hệ thông tin không đầy đủ, ngược lại là hệ thông tin đầy đủ, giá trị rỗng được biểu diễn là '*'.

Xét hệ thông tin không đầy đủ $IS = (U, A, V, f)$. Với $P \subseteq A$ ta định nghĩa một quan hệ nhị phân trên U như sau:

$$SIM(P) = \{(u, v) \in U \times U \,|\forall a \in P, f(u, a) = f(v, a) \vee f(u, a) =' *' \ \vee f(v, a) =' *'\}$$

$SIM\,(P)$ được gọi là quan hệ dung sai (tolerance relation), hay quan hệ tương tự (similarity relation) trên U. Dễ thấy rằng $SIM\,(P) = \cap_{a\in P} SIM\,(\{a\})$. Quan hệ *SIM(P)* không phải là quan hệ tương đương vì chúng có tính phản xạ, đối xứng nhưng không có tính bắc cầu. Ký hiệu $U/SIM\,(P)$ biểu diễn các tập $\{S_P\,(u)\,|u \in U\}, S_P\,(u)$ là tập các đối tượng không có khả năng phân biệt được với u đối với 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 $S_P\,(u) \neq \emptyset$ với mọi $u \in U$ và $\cup_{u\in U} S_P\,(u) = U$. Ký hiệu tập tất cả các phủ của U sinh bởi các tập con thuộc tính $P \subseteq A$ là $COVER\,(U)$.

Trên $COVER\,(U)$, quan hệ thứ tự bộ phận $(COVER\,(U), \preceq)$ được định nghĩa như sau.

Định nghĩa 2.1. [10]. Cho hệ thông tin không đầy đủ $IS = (U, A, V, f)$ với $P, Q \subseteq A$. Ta nói:

1) Phủ $U/SIM\,(P)$ và phủ $U/SIM\,(Q)$ là như nhau (viết $U/SIM\,(P) = U/SIM\,(Q)$), khi và chỉ khi $\forall u \in U,\ \ S_P\,(u) = S_Q\,(u)$.

2) $U/SIM\,(P)$ mịn hơn $U/SIM\,(Q)$ (viết $U/SIM\,(P) \preceq U/SIM\,(Q)$) khi và chỉ khi $\forall u_i \in U,\ \ S_P\,(u_i) \subseteq S_Q\,(u_i)$.

Trên $(COVER\,(U), \preceq)$, phần tử nhỏ nhất gọi là phủ rời rạc $\omega = \{S_A\,(u) = \{u\}\,/u \in U\}$ và phần tử lớn nhất gọi là phủ một khối $\delta = \{S_A\,(u) = \{U\}\,/u \in U\}$.

Tính chất 2.1. [10]. Cho hệ thông tin không đầy đủ $IS = (U, A, V, f)$

1) Nếu $P \subseteq Q \subseteq A$ thì $U/SIM\,(Q) \preceq U/SIM\,(P)$.

2) Nếu $P, Q \subseteq A$ thì $S_{P\cup Q}\,(u) = S_P\,(u) \cap S_Q\,(u)$ với $\forall u \in U$.

Bảng quyết định là dạng đặc biệt của hệ thông tin, trong đó tập các thuộc tính A bao gồm hai tập con tách biệt nhau: tập các thuộc tính điều kiện C và tập các thuộc tính quyết định D. Như vậy, bảng quyết định là hệ thông tin $DS = (U, C \cup D, V, f)$ trong đó $C \cap D = \emptyset$. Với $a \in C$ nếu V_a chứa giá trị rỗng (biểu diễn là '*') thì DS được gọi là không đầy đủ, ngược lại DS được gọi là đầy đủ. Không mất tính chất tổng quát, giả thiết D chỉ gồm một thuộc tính quyết định duy nhất d và $* \notin V_d$.

3. LIANG ENTROPY MỞ RỘNG VÀ CÁC TÍNH CHẤT

3.1. Liang entropy mở rộng của tập thuộc tính

Định nghĩa 3.1. Cho hệ thông tin không đầy đủ $IIS = (U, A, V, f)$, $P \subseteq A$ và $U/SIM\,(P) = \left\{S_P\,(u_1), S_P\,(u_2), ..., S_P\left(u_{|U|}\right)\right\}$. Liang entropy mở rộng của P là đại lượng $IE\,(P)$, xác định như sau

$$IE(P) = \sum_{i=1}^{|U|} \frac{1}{|U|}\left(1 - \frac{|S_P(u_i)|}{|U|}\right)$$

trong đó $|S_P\,(u)|$ chỉ lực lượng tập $S_P\,(u)$. Nếu $U/SIM\,(P) = \omega$ thì $IE(P)$ đạt giá trị lớn nhất là $1 - 1/|U|$. Nếu $U/SIM\,(P) = \delta$ thì $IE(P)$ đạt giá trị nhỏ nhất là 0. Như vậy $0 \leq IE\,(P) \leq 1 - 1/|U|$.

Mệnh đề 3.1. *Cho hệ thông tin đầy đủ* $IS = (U, A, V, f)$, $P \subseteq A$ *với* $P \subseteq A$ *và* $U/P = \{P_1, P_2, ..., P_m\}$. *Ta có*

$$IE(P) = \sum_{i=1}^{|U|} \frac{1}{|U|}\left(1 - \frac{|S_P(u_i)|}{|U|}\right) = \sum_{i=1}^{m} \frac{|P_i|}{|U|}\left(1 - \frac{|P_i|}{|U|}\right) = E(P)$$

với $E(P)$ *là Liang entropy trong [6].*

Chứng minh. Giả sử $P_i = \{u_{i1}, u_{i2}, ..., u_{is_i}\}$ với $|P_i| = s_i$ và $\sum_{i=1}^{m} s_i = |U|$. Ta có:

$$P_i = S_P(u_{i1}) = S_P(u_{i2}) = ... = S_P(u_{is_i}),\ |P_i| = |S_P(u_{i1})| = |S_P(u_{i2})| = ... = |S_P(u_{is_i})| = s_i$$

,

$$\frac{|P_i|}{|U|}\left(1 - \frac{|P_i|}{|U|}\right) = \frac{1}{|U|}\left(|P_i| - \frac{|P_i|\,|P_i|}{|U|}\right)$$

$$= \frac{1}{|U|}\left(1 - \frac{|S_P(u_{i1})|}{|U|} + 1 - \frac{|S_P(u_{i2})|}{|U|} + ... + 1 - \frac{|S_P(u_{is_i})|}{|U|}\right)$$

$$E(P) = \sum_{i=1}^{m} \frac{|P_i|}{|U|}\left(1 - \frac{|P_i|}{|U|}\right) = \sum_{i=1}^{m}\sum_{k=1}^{s_i} \frac{1}{|U|}\left(1 - \frac{|S_P(u_{ik})|}{|U|}\right)$$

$$= \sum_{i=1}^{|U|} \frac{1}{|U|}\left(1 - \frac{|S_P(u_i)|}{|U|}\right) = IE(P)$$

■

Định nghĩa 3.2. Liang entropy mở rộng của $P \cup Q$ là đại lượng $IE(P \cup Q)$, xác định như sau

$$IE(P \cup Q) = \sum_{i=1}^{|U|} \frac{1}{|U|}\left(1 - \frac{|S_{P\cup Q}(u_i)|}{|U|}\right) = \sum_{i=1}^{|U|} \frac{1}{|U|}\left(1 - \frac{|S_P(u_i) \cap S_Q(u_i)|}{|U|}\right)$$

3.2. Liang entropy mở rộng có điều kiện

Định nghĩa 3.3. Cho hệ thông tin không đầy đủ $IIS = (U, A, V, f)$ và $P, Q \subseteq A$. Giả sử
$U/SIM(P) = \{S_P(u_1), S_P(u_2), ..., S_P(u_{|U|})\}$ và
$U/SIM(Q) = \{S_Q(u_1), S_Q(u_2), ..., S_Q(u_{|U|})\}$.

Liang entropy mở rộng có điều kiện của Q khi đã biết P được định nghĩa bởi

$$IE(Q|P) = \frac{1}{|U|}\sum_{i=1}^{|U|}\left(\frac{|S_P(u_i)| - |S_Q(u_i) \cap S_P(u_i)|}{|U|}\right)$$

Mệnh đề 3.2. *Cho hệ thông tin đầy đủ* $IS = (U, A, V, f)$, $P \subseteq A$ *với* $P, Q \subseteq A$. *Giả sử* $U/P = \{P_1, P_2, ..., P_m\}$, $U/Q = \{Q_1, Q_2, ..., Q_n\}$. *Ta có*

$$IE(Q\,|P) = \frac{1}{|U|}\sum_{i=1}^{|U|}\left(\frac{|S_P(u_i)| - |S_Q(u_i) \cap S_P(u_i)|}{|U|}\right) = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{|Q_i \cap P_j|}{|U|}\frac{\left|Q_i^c - P_j^c\right|}{|U|} = E(Q\,|P)$$

với $Q_i^c = U - Q_i$, $P_j^c = U - P_j$ *và* $E\,(Q\,|P)$ *là Liang entropy có điều kiện trong [6].*

Chứng minh. Giả sử $Q_i \cap P_j = \left\{u_{i1}, u_{i2}, ..., u_{is_j}\right\}$ với $|Q_i \cap P_j| = s_j$ và $|Q_i| = t_i$, khi đó $\sum\limits_{j=1}^{m} s_j = t_i$ và $\sum\limits_{i=1}^{n} t_i = |U|$. Ta có

$$Q_i \cap P_j = S_Q\left(u_{i1}\right) \cap S_P\left(u_{i1}\right) = S_Q\left(u_{i2}\right) \cap S_P\left(u_{i2}\right) = ... = S_Q\left(u_{is_j}\right) \cap S_P\left(u_{is_j}\right)$$

$$|Q_i \cap P_j| = |S_Q\left(u_{i1}\right) \cap S_P\left(u_{i1}\right)| = |S_Q\left(u_{i2}\right) \cap S_P\left(u_{i2}\right)| = ...$$

$$= \left|S_Q\left(u_{is_j}\right) \cap S_P\left(u_{is_j}\right)\right| = s_j$$

$$|Q_i \cap P_j|\left|Q_i^c - P_j^c\right| = |Q_i \cap P_j|\,|Q_i^c \cap P_j| = |Q_i \cap P_j|\,|P_j - (Q_i \cap P_j)|$$

$$= |S_P\left(u_{i1}\right) - (S_Q\left(u_{i1}\right) \cap S_P\left(u_{i1}\right))| + |S_P\left(u_{i2}\right) - (S_Q\left(u_{i2}\right) \cap S_P\left(u_{i2}\right))| + ... +$$

$$+ \left|S_P\left(u_{is_i}\right) - \left(S_Q\left(u_{is_j}\right) \cap S_P\left(u_{is_j}\right)\right)\right|$$

$$= \sum_{k=1}^{s_j} |S_P\left(u_{ik}\right) - (S_Q\left(u_{ik}\right) \cap S_P\left(u_{ik}\right))| = \sum_{k=1}^{s_j} |S_P\left(u_{ik}\right)| - |S_Q\left(u_{ik}\right) \cap S_P\left(u_{ik}\right)|.$$

Do đó

$$\sum_{j=1}^{m} |Q_i \cap P_j|\left|Q_i^c - P_j^c\right| = \sum_{j=1}^{m}\sum_{k=1}^{s_j} |S_P\left(u_{ik}\right)| - |S_Q\left(u_{ik}\right) \cap S_P\left(u_{ik}\right)|$$

$$= \sum_{k=1}^{t_i} |S_P\left(u_{ik}\right)| - |S_Q\left(u_{ik}\right) \cap S_P\left(u_{ik}\right)|$$

$$\Leftrightarrow \sum_{i=1}^{n}\sum_{j=1}^{m} |Q_i \cap P_j|\left|Q_i^c - P_j^c\right| = \sum_{i=1}^{n}\sum_{k=1}^{t_i} |S_P\left(u_{ik}\right)| - |S_Q\left(u_{ik}\right) \cap S_P\left(u_{ik}\right)|$$

$$= \sum_{i=1}^{|U|} |S_P\left(u_i\right)| - |S_Q\left(u_i\right) \cap S_P\left(u_i\right)|$$

$$\Leftrightarrow IE(Q\,|P) = \tfrac{1}{|U|}\sum_{i=1}^{|U|}\left(\tfrac{|S_P(u_i)| - |S_Q(u_i) \cap S_P(u_i)|}{|U|}\right) = \sum_{i=1}^{n}\sum_{j=1}^{m}\tfrac{|Q_i \cap P_j|}{|U|}\tfrac{\left|Q_i^c - P_j^c\right|}{|U|} = E(Q\,|P)$$

■

3.3. Một số tính chất của Liang entropy mở rộng

Mệnh đề 3.3. *Cho hệ thông tin đầy đủ* $IIS = (U, A, V, f)$ *với* $P, Q, R \subseteq A$.

a) *Nếu* $U/SIM\,(P) \preceq U/SIM\,(Q)$ *thì* $IE\,(P) \geq IE\,(Q)$ *và* $IE\,(P) = IE\,(Q)$ *khi và chỉ khi* $U/SIM\,(P) = U/SIM\,(Q)$.

b) *Nếu* $U/SIM\,(P) \preceq U/SIM\,(Q)$ *thì* $IE\,(P \cup Q) = IE\,(P)$.

c) $IE(P \cup Q) \geq IE(P)$ *và* $IE(P \cup Q) \geq IE(Q)$.

d) $IE(P \cup Q) = IE(P) + IE(Q|P) = IE(P) + IE(P|Q)$.

e) $0 \leq IE(Q|P) \leq 1 - 1/|U|$. $IE(Q|P) = 0$ *khi và chỉ khi* $U/SIM(P) \preceq U/SIM(Q)$, $IE(Q|P) = 1 - 1/|U|$ *khi và chỉ khi* $U/SIM(P) = \delta$ *và* $U/SIM(Q) = \omega$.

f) *Nếu* $U/SIM(P) \preceq U/SIM(Q)$ *thì* $IE(R|Q) \geq IE(R|P)$.

g) *Nếu* $U/SIM(P) \preceq U/SIM(Q)$ *thì* $IE(P|R) \geq IE(Q|R)$.

Chứng minh. a) Được suy ra từ Định nghĩa 3.1 và Định nghĩa 2.1.

b) Được suy ra từ Định nghĩa 3.1, Định nghĩa 3.2, Định nghĩa 2.1 và Tính chất 2.1.

c) Được suy ra từ a).

d) Từ Định nghĩa 3.1, Định nghĩa 3.2 và Định nghĩa 3.3 ta có

$$IE(Q|P) = \frac{1}{|U|}\sum_{i=1}^{|U|}\frac{|S_P(u_i)|-|S_P(u_i)\cap S_Q(u_i)|}{|U|} = 1-\frac{1}{|U|}\sum_{i=1}^{|U|}\frac{|S_P(u_i)\cap S_Q(u_i)|}{|U|}-1+\frac{1}{|U|}\sum_{i=1}^{|U|}\frac{|S_P(u_i)|}{|U|}$$

$$= \frac{1}{|U|}\sum_{i=1}^{|U|}1-\frac{|S_P(u_i)\cap S_Q(u_i)|}{|U|} - \frac{1}{|U|}\sum_{i=1}^{|U|}1-\frac{|S_P(u_i)|}{|U|} = IE(P\cup Q) - IE(P)$$

Do đó: $IE(P \cup Q) = IE(P) + IE(P|Q)$. Do tính chất đối xứng của $IE(P \cup Q)$ nên ta cũng có $IE(P \cup Q) = IE(Q) + IE(P|Q)$.

e) Hiển nhiên $IE(Q|P) \geq 0$. Theo phần d), $IE(Q|P) = IE(P \cup Q) - IE(P)$ nên $IE(Q|P) = 0 \Leftrightarrow IE(P \cup Q) = IE(P)$. Vì $U/SIM(P \cup Q) \preceq U/SIM(P)$ nên theo phần a) ta có:

$IE(P \cup Q) = IE(P) \Leftrightarrow U/SIM(P \cup Q) = U/SIM(P) \Leftrightarrow U/SIM(P) \preceq U/SIM(Q)$

Mặt khác, theo phần d) và Định nghĩa 3.1 ta có

$IE(Q|P) = IE(P \cup Q) - IE(P)$, $IE(P \cup Q) \leq 1 - 1/|U|$, $IE(P) \geq 0$ nên suy ra $IE(Q|P) \leq 1 - 1/|U|$.

Dấu ' = ' xảy ra khi và chỉ khi $IE(P) = 0 \wedge IE(P \cup Q) = 1 - 1/|U|$, nghĩa là $U/SIM(P) = \delta$ và $U/SIM(P \cup Q) = \omega$. Điều này tương đương với $U/SIM(P) = \delta$ và $U/SIM(Q) = \omega$.

f) Giả sử $U/SIM(R) = \{S_R(u_1), S_R(u_2), ..., S_R(u_{|U|})\}$. Từ $U/SIM(P) \preceq U/SIM(Q)$ ta có $S_P(u_i) \subseteq S_Q(u_i)$ với mọi $u_i \in U, i = 1..|U|$ và

$(S_Q(u_i) - S_P(u_i)) \cap S_R(u_i) \subseteq S_Q(u_i) - S_P(u_i)$

$\Leftrightarrow (S_Q(u_i) \cap S_R(u_i)) - (S_P(u_i) \cap S_R(u_i)) \subseteq S_Q(u_i) - S_P(u_i)$

$\Leftrightarrow |(S_Q(u_i) \cap S_R(u_i)) - (S_P(u_i) \cap S_R(u_i))| \leq |S_Q(u_i) - S_P(u_i)|$ (3.1)

Do $S_P(u_i) \subseteq S_Q(u_i)$ nên $S_P(u_i) \cap S_R(u_i) \subseteq S_Q(u_i) \cap S_R(u_i)$ và (3.1) tương đương:

$|S_Q(u_i) \cap S_R(u_i)| - |S_P(u_i) \cap S_R(u_i)| \leq |S_Q(u_i)| - |S_P(u_i)|$

$\Leftrightarrow |S_Q(u_i)| - |S_Q(u_i) \cap S_R(u_i)| \geq |S_P(u_i)| - |S_P(u_i) \cap S_R(u_i)|$

$\Leftrightarrow \frac{1}{|U|}\sum_{i=1}^{n} \frac{|S_Q(u_i)|-|S_Q(u_i)\cap S_R(u_i)|}{|U|} \geq \frac{1}{|U|}\sum_{i=1}^{n} \frac{|S_P(u_i)|-|S_P(u_i)\cap S_R(u_i)|}{|U|} \Leftrightarrow IE(R|Q) \geq IE(R|P)$

g) Từ $U/SIM(P) \preceq U/SIM(Q)$ nên với mọi $u_i \in U, i = 1..|U|$ ta có $S_P(u_i) \subseteq S_Q(u_i)$. Giả sử $U/SIM(R) = \{S_R(u_1), S_R(u_2), ..., S_R(u_{|U|})\}$, khi đó:

$S_P(u_i) \cap S_R(u_i) \subseteq S_Q(u_i) \cap S_R(u_i) \Leftrightarrow |S_P(u_i) \cap S_R(u_i)| \leq |S_Q(u_i) \cap S_R(u_i)|$

$\Leftrightarrow |S_R(u_i)| - |S_P(u_i) \cap S_R(u_i)| \geq |S_R(u_i)| - |S_Q(u_i) \cap S_R(u_i)|$

$\Leftrightarrow \frac{1}{|U|}\sum_{i=1}^{|U|} \frac{|S_R(u_i)|-|S_P(u_i)\cap S_R(u_i)|}{|U|} \geq \frac{1}{|U|}\sum_{i=1}^{|U|} \frac{|S_R(u_i)|-|S_Q(u_i)\cap S_R(u_i)|}{|U|} \Leftrightarrow IE(P|R) \geq IE(Q|R)$

■

4. METRIC TRÊN HỌ CÁC PHỦ VÀ CÁC TÍNH CHẤT

Một metric trên tập hợp U là một ánh xạ $d : U \times U \to [0, \infty)$ thỏa mãn các điều kiện sau với mọi $x, y, z \in U$.

P(1) $d(x, y) \geq 0$, điều kiện $d(x, y) = 0$ khi và chỉ khi $x = y$.

P(2) $d(x, y) = d(y, x)$.

P(3) $d(x, y) + d(y, z) \geq d(x, z)$.

Điều kiện $P(3)$ được gọi là tiên đề bất đẳng thức tam giác

Dựa trên kết quả trong [11], chúng tôi đề xuất một metric trên họ các phủ trong hệ thông tin không đầy đủ sử dụng Liang entropy mở rộng và nghiên cứu một số tính chất của chúng.

Bổ đề 4.1. *Cho hệ thông tin không đầy đủ $IIS = (U, A, V, f)$. Với mọi $P, Q, R \subseteq A$.*

a) $IE(P|R) + IE(Q|P \cup R) = IE(P \cup Q|R)$.

b) $IE(Q|P) + IE(P|R) \geq IE(Q|R)$.

Chứng minh. Giả sử:

$U/SIM(P) = \{S_P(u_1), S_P(u_2), ..., S_P(u_{|U|})\}$
$U/SIM(Q) = \{S_Q(u_1), S_Q(u_2), ..., S_Q(u_{|U|})\}$
$U/SIM(R) = \{S_R(u_1), S_R(u_2), ..., S_R(u_{|U|})\}$

a) Thật vậy $IE(P|R) + IE(Q|P \cup R) =$.

$$= \frac{1}{|U|}\sum_{i=1}^{|U|}\frac{|S_R(u_i)|-|S_P(u_i)\cap S_R(u_i)|+|S_{P\cup R}(u_i)|-|S_{P\cup R}(u_i)\cap S_Q(u_i)|}{|U|}$$

$$= \frac{1}{|U|}\sum_{i=1}^{|U|}\frac{|S_R(u_i)|-|S_{P\cup R}(u_i)|+|S_{P\cup R}(u_i)|-\left|S_{P\cup R}(u_i)\cap S_Q(u_i)\right|}{|U|}$$

$$= \frac{1}{|U|}\sum_{i=1}^{|U|}\frac{|S_R(u_i)|-\left|S_P(u_i)\cap S_Q(u_i)\cap S_R(u_i)\right|}{|U|}$$

$$= \frac{1}{|U|}\sum_{i=1}^{|U|}\frac{|S_R(u_i)|-\left|S_R(u_i)\cap S_{P\cup Q}(u_i)\right|}{|U|} = IE(P\cup Q|R)$$

b) Do $U/SIM(P\cup R) \preceq U/SIM(P)$, $U/SIM(P\cup Q) \preceq U/SIM(Q)$ nên áp dụng Mệnh đề 3.3 phần a) ta có $IE(Q|P) \geq IE(Q|P\cup R)$, $IE(P\cup Q|R) \geq IE(Q|R)$. Sử dụng kết quả ở phần a) ta có $IE(Q|P) + IE(P|R) \geq IE(Q|P\cup R) + IE(P|R) = IE(P\cup Q|R) \geq IE(Q|R)$.

■

Định lý 4.1. *Cho hệ thông tin không đầy đủ $IIS = (U, A, V, f)$. Với $P, Q \subseteq A$ giả sử $K(P) = U/SIM(P), K(Q) = U/SIM(Q)$. Khi đó với mọi $K(P), K(Q) \in COVER(U)$, ánh xạ $d_E : COVER(U)\times COVER(U) \to [0,\infty)$ xác định bởi*

$$d_E(K(P), K(Q)) = IE(P|Q) + IE(Q|P)$$

là một metric trên tập $COVER(U)$.

Chứng minh. (P1) Theo Mệnh đề 3.3 phần e), $d_E(K(P), K(Q)) \geq 0$ với mọi $K(P), K(Q) \in COVER(U)$, $d_E(K(P), K(Q)) = 0 \Leftrightarrow (IE(Q|P) = 0) \wedge (IE(P|Q) = 0)$.

$\Leftrightarrow (U/SIM(P) \preceq U/SIM(Q)) \wedge (U/SIM(Q) \preceq U/SIM(P)) \Leftrightarrow K(P) = K(Q)$

(P2) Từ định nghĩa của d_E suy ra $d_E(K(P), K(Q)) = d_E(K(Q), K(P))$ với mọi $K(P)$, $K(Q) \in COVER(U)$.

(P3) Với mọi $K(P), K(Q), K(R) \in COVER(U)$, áp dụng Bổ đề 4.1 phần b)

$IE(Q|P) + IE(P|R) \geq IE(Q|R)$ (4.1)

$IE(R|P) + IE(P|Q) \geq IE(R|Q)$ (4.2)

Cộng (4.1) với (4.2) theo vế với vế thu được:

$d_E(K(Q), K(P)) + d_E(K(P), K(R)) \geq d_E(K(Q), K(R))$

Từ $(P1), (P2), (P3)$ kết luận $d_E(K(P), K(Q))$ là một metric trên tập $COVER(U)$

■

Mệnh đề 4.1. *Cho hệ thông tin đầy đủ* $IIS = (U, A, V, f)$ *với* $P \subseteq A$. *Ta có*

$$d_E\left(K\left(P\right), K\left(A\right)\right) = \frac{1}{|U|}\sum_{i=1}^{|U|}\frac{|S_P(u_i)| - |S_A(u_i)|}{|U|}$$

Chứng minh. Từ giả thiết $P \subseteq A$ suy ra $U/SIM\left(A\right) \preceq U/SIM\left(P\right)$, theo Mệnh đề 3.3 phần e) ta có $IE\left(P\,|A\right) = 0$. Mặt khác, cũng từ $P \subseteq A$ ta có $S_A\left(u_i\right) \subseteq S_P\left(u_i\right)$, hay $S_P\left(u_i\right) \cap S_A\left(u_i\right) = S_A\left(u_i\right)$ với mọi $u_i \in U, i = 1..|U|$. Do đó

$d_E\left(K\left(P\right), K\left(A\right)\right) = IE\left(P\,|A\right) + IE\left(A\,|P\right) = IE\left(A\,|P\right)$

$= \frac{1}{|U|}\sum\limits_{i=1}^{|U|}\frac{|S_P(u_i)| - |S_P(u_i) \cap S_A(u_i)|}{|U|} = \frac{1}{|U|}\sum\limits_{i=1}^{|U|}\frac{|S_P(u_i)| - |S_A(u_i)|}{|U|}$

■

Mệnh đề 4.2. *Cho bảng quyết định không đầy đủ* $IDS = (U, C \cup D, V, f)$. *Nếu* $B \subseteq C$ *thì* $d_E\left(K\left(B\right), K\left(B \cup D\right)\right) \geq d_E\left(K\left(C\right), K\left(C \cup D\right)\right)$

Chứng minh. Xét bảng quyết định không đầy đủ $IDS = \left(U, C \cup D, V, f\right), U = \{u_1, u_2, ..., u_n\}$ và $B \subseteq C$. Với mọi $u_i \in U, i = 1..n$ ta có $S_C\left(u_i\right) \subseteq S_B\left(u_i\right)$, do đó:

$\left(S_B\left(u_i\right) - S_C\left(u_i\right)\right) \cap S_D\left(u_i\right) \subseteq S_B\left(u_i\right) - S_C\left(u_i\right)$

$\Leftrightarrow \left(S_B\left(u_i\right) \cap S_D\left(u_i\right)\right) - \left(S_C\left(u_i\right) \cap S_D\left(u_i\right)\right) \subseteq S_B\left(u_i\right) - S_C\left(u_i\right)$

$\Leftrightarrow \left|\left(S_B\left(u_i\right) \cap S_D\left(u_i\right)\right) - \left(S_C\left(u_i\right) \cap S_D\left(u_i\right)\right)\right| \leq \left|S_B\left(u_i\right) - S_C\left(u_i\right)\right|$ (4.3)

Do $S_C\left(u_i\right) \subseteq S_B\left(u_i\right)$ nên $S_C\left(u_i\right) \cap S_D\left(u_i\right) \subseteq S_B\left(u_i\right) \cap S_D\left(u_i\right)$ và (4.3) tương đương với:

$\left|S_B\left(u_i\right) \cap S_D\left(u_i\right)\right| - \left|S_C\left(u_i\right) \cap S_D\left(u_i\right)\right| \leq \left|S_B\left(u_i\right)\right| - \left|S_C\left(u_i\right)\right|$

$\Leftrightarrow \left|S_B\left(u_i\right)\right| - \left|S_B\left(u_i\right) \cap S_D\left(u_i\right)\right| \geq \left|S_C\left(u_i\right)\right| - \left|S_C\left(u_i\right) \cap S_D\left(u_i\right)\right|$ (4.4)

Do $S_B\left(u_i\right) \cap S_D\left(u_i\right) \subseteq S_B\left(u_i\right), S_C\left(u_i\right) \cap S_D\left(u_i\right) \subseteq S_C\left(u_i\right)$ nên (4.4) tương đương với:

$$\begin{array}{l}\left|S_B\left(u_i\right) \cup \left(S_B\left(u_i\right) \cap S_D\left(u_i\right)\right)\right| - \left|S_B\left(u_i\right) \cap \left(S_B\left(u_i\right) \cap S_D\left(u_i\right)\right)\right| \geq \\ \left|S_C\left(u_i\right) \cup \left(S_C\left(u_i\right) \cap S_D\left(u_i\right)\right)\right| - \left|S_C\left(u_i\right) \cap \left(S_C\left(u_i\right) \cap S_D\left(u_i\right)\right)\right|\end{array} \quad (4.5)$$

Do $S_{B \cup D}\left(u_i\right) = S_B\left(u_i\right) \cap S_D\left(u_i\right), S_{C \cup D}\left(u_i\right) = S_C\left(u_i\right) \cap S_D\left(u_i\right)$ nên (4.5) tương đương với:

$\Leftrightarrow \sum\limits_{i=1}^{n}\frac{|S_B(u_i)| - |S_{B \cup D}(u_i)|}{|U|^2} \geq \sum\limits_{i=1}^{n}\frac{|S_C(u_i)| - |S_{C \cup D}(u_i)|}{|U|^2}$ (4.6)

Do $B \subset B \cup D, C \subset C \cup D$ nên theo Mệnh đề 4.1, công thức (4.6) tương đương với

$d_E\left(K\left(B\right), K\left(B \cup D\right)\right) \geq d_E\left(K\left(C\right), K\left(C \cup D\right)\right)$

■

5. RÚT GỌN THUỘC TÍNH TRONG BẢNG QUYẾT ĐỊNH KHÔNG ĐẦY ĐỦ SỬ DỤNG METRIC

5.1. Tập rút gọn của bảng quyết định không đầy đủ dựa trên metric

Định nghĩa 5.1. Cho bảng quyết định không đầy đủ $IDS = (U, C \cup D, V, f)$. $R \subseteq C$ được gọi là một rút gọn của C dựa trên metric nếu thỏa mãn điều kiện:

(1) $d_E(K(R), K(R \cup D)) = d_E(K(C), K(C \cup D))$.

(2) $\forall r \in R,\ d_E(K(R - \{r\}), K(R - \{r\} \cup D)) \neq d_E(K(C), K(C \cup D))$.

5.2. Độ quan trọng của thuộc tính dựa trên metric

Định nghĩa 5.2. Cho bảng quyết định không đầy đủ $IDS = (U, C \cup D, V, f)$ với $B \subseteq C$. Độ quan trọng của thuộc tính $b \in C - B$ được định nghĩa

$$SIG_B(b) = d_E(K(B), K(B \cup D)) - d_E(K(B \cup \{b\}), K(B \cup \{b\} \cup D))$$

với giả thiết $S_\emptyset(u_i) = U$ với mọi $u_i \in U, i = 1..|U|$.

5.3. Thuật toán tìm tập rút gọn của bảng quyết định không đầy đủ sử dụng metric

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

Thuật toán 5.1. Tìm tập rút gọn của bảng quyết định không đầy đủ.

Đầu vào: Bảng quyết định không đầy đủ $IDS = (U, C \cup D, V, f)$.

Đầu ra: Tập rút gọn R.

1. $R = \emptyset$;

2. Tính $d_E(K(R), K(R \cup D))$;

3. Tính $d_E(K(C), K(C \cup D))$;

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

4. While $d_E(K(R), K(R \cup D)) \neq d_E(K(C), K(C \cup D))$ do

5. Begin

6. For each $a \in C - R$

7. Begin

8. Tính $d_E(K(R \cup \{a\}), K(R \cup \{a\} \cup D))$;

9. Tính $SIG_R(a) = d_E(K(R), K(R \cup D)) - d_E(K(R \cup \{a\}), K(R \cup \{a\} \cup D))$;

10. End;

11. Chọn $a_m \in C - R$ sao cho $SIG_R(a_m) = \underset{a \in C-R}{Max}\{SIG_R(a)\}$;

12. $R = R \cup \{a_m\}$;

13. Tính $d_E(K(R), K(R \cup D))$;

14. End;

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

15. For each $a \in R$

16. Begin

17. Tính $d_E(K(R - \{a\}), K(R - \{a\} \cup D))$;

Bảng 6.1. Kết quả thực hiện Thuật toán IQBAR và Thuật toán 5.1

STT	Bộ số liệu	$\|U\|$	$\|C\|$	Thuật toán $IQBAR$		Thuật toán 5.1	
				$\|R\|$	t	$\|R\|$	t
1	Hepatitis.data	155	19	4	1.296	4	0.89
2	Lung-cancer.data	32	56	4	0.187	4	0.171
3	Automobile.data	205	25	5	3	5	1.687
4	Anneal.data	798	38	9	179	9	86.921
5	Voting Records	435	16	15	25.562	15	16.734
6	Credit Approval	690	15	7	29.703	7	15.687

18. If $d_E\left(K\left(R-\{a\}\right),K\left(R-\{a\}\cup D\right)\right)=d_E\left(K\left(C\right),K\left(C\cup D\right)\right)$ then $R=R-\{a\}$;

19. End;

20. Return R;

Với bước *thêm dần vào R các thuộc tính có độ quan trọng lớn nhất*, tập thuộc tính R thu được từ câu lệnh từ 4 đến 14 thỏa mãn điều kiện bảo toàn khoảng cách $d_E\left(K\left(R\right),K\left(R\cup D\right)\right)=d_E\left(K\left(C\right),K\left(C\cup D\right)\right)$. Với bước *loại bỏ các thuộc tính dư thừa trong R nếu có*, câu lệnh từ 15 đến 19 đảm bảo tập R là tối thiểu, nghĩa là $\forall r\in R$, $d_E\left(K\left(R-\{r\}\right),K\left(\left(R-\{r\}\right)\cup D\right)\right)\neq d_E\left(K\left(C\right),K\left(C\cup D\right)\right)$. Theo Định nghĩa 5.1, tập R thu được là tập rút gọn của bảng quyết định dựa trên metric.

Để tính $SIG_R\left(a\right)$ ta chỉ cần tính $S_{R\cup\{a\}}(u_i)$, $S_{R\cup\{a\}\cup D}(u_i)$, vì $S_R(u_i), S_{R\cup D}(u_i)$ đã được tính ở vòng lặp trước. Độ phức tạp để tính $S_{R\cup\{a\}}(u_i)$ khi đã biết $S_R(u_i)$ với mọi $u_i\in U$ là $O\left(|U|^2\right)$, do đó giả sử $D=\{d\}$, độ phức tạp để tính tất cả các $SIG_E\left(a\right)$ là $\left(|C|+\left(|C|-1\right)+...+1\right)*|U|^2=\left(|C|*\left(|C|-1\right)/2\right)*|U|^2=O\left(|C|^2|U|^2\right)$. Độ phức tạp để chọn thuộc tính có độ quan trọng lớn nhất là $|C|+\left(|C|-1\right)+...+1=|C|*\left(|C|-1\right)/2=O\left(|C|^2\right)$. Do đó, độ phức tạp của thuật toán là $O\left(|C|^2|U|^2\right)$. Độ phức tạp này tốt hơn độ phức tạp của các thuật toán trong [2, 3, 18].

6. THỬ NGHIỆM THUẬT TOÁN

Cài đặt Thuật toán *IQBAR*[3] và Thuật toán 5.1 bằng ngôn ngữ $C++$.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 [17]. 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). Bảng 6.1 mô tả kết quả thực hiện của hai thuật toán.

Kết quả thử nghiệm cho thấy, tập rút gọn thu được khi thực hiện Thuật toán 5.1 và Thuật toán IQBAR trên 6 bộ số liệu là như nhau. Tuy nhiên, thời gian thực hiện Thuật toán 5.1 nhanh hơn Thuật toán IQBAR, do đó Thuật toán 5.1 hiệu quả hơn Thuật toán IQBAR.

7. KẾT LUẬN

Trên hệ thông tin không đầy đủ, bài báo đã thực hiện các nội dung nghiên cứu sau:

1) Đề xuất Liang entropy mở rộng và nghiên cứu một số tính chất của chúng.

2) Xây dựng một metric trên họ các phủ sử dụng Liang entropy mở rộng.

3) Đề xuất thuật toán heuristic tìm tập rút gọn của bảng quyết định không đầy đủ sử dụng metric được xây dựng với độ phức tạp $O\left(|C|^2|U|^2\right)$. Chứng minh bằng lý thuyết và thực nghiệm, thuật toán đề xuất hiệu quả hơn thuật toán trong [2, 3, 18].

TÀI LIỆU THAM KHẢO

[1] K.S. Chin, J.Y. Liang, C.Y. Dang, Rough set data analysis algorithms for incomplete information systems, *Proceedings of the 9th international conference on Rough sets, fuzzy sets, data mining, and granular computing, RSFDGrC'03*, (2003) 264–268.

[2] B. Huang, X. He, X.Z. Zhou, Rough computational methods based on tolerance matrix, *Automatica Sinica* **30** (2004) 363–370.

[3] B. Huang, H. X. Li, X. Z. Zhou, Attribute reduction based on information quantity under incomplete information systems, *Systems Application Theory & Practice* **34** (2005) 55–60.

[4] M. Kryszkiewicz, Rough set approach to incomplete information systems, *Information Science* **112** (1998) 39–49.

[5] J.H. Li, K.Q. Shi, A algorithm for attribute reduction based on knowledge granularity, *Computer Applications* **26** (6) (2006) 76–77.

[6] J.Y Liang, K.S. Chin, C.Y. Dang, C.M.YAM Richard, New method for measuring uncertainty and fuzziness in rough set theory, *International Journal of General Systems* **31** 331–342.

[7] J.Y. Liang, Y.H. Qian, Axiomatic approach of knowledge granulation in information system, *Lecture Notes in Artificial Intelligence* **4304** (2006) 1074–1078.

[8] J.Y. Liang, Y.H. Qian, Information granules and entropy theory in information systems, *Information Sciences*, **51** (2008) 1–18.

[9] J.Y. Liang, Z.B. Xu, The algorithm on knowledge reduction in incomplete information systems, International Journal of Uncertainty, *Fuzziness and Knowledge-Based Systems* **10** 1 (2002) 95–103.

[10] J.Y. Liang, Z.Z. Shi, D.Y. Li, M.J. Wierman, The information entropy, rough entropy and knowledge granulation in incomplete information system, *International Journal of General Systems* **35** (6) (2006) 641–654.

[11] Nguyễn Thanh Tùng, Về một metric trên họ các phân hoạch của một tập hợp hữu hạn, *Tạp chí Tin học và Điều khiển học* **26** (1) (2010) 73–75.

[12] Y.H. Qian, J.Y. Liang, Combination entropy and combination granulation in incomplete information system, *RSKT*, 2006 (184–190).

[13] Y.H. Qian, J.Y. Liang, New method for measuring uncertainty in incomplete information systems, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* (2008).

[14] Y.H. Qian, J.Y. Liang, C.Y. Dang, Knowledge structure, knowledge granulation and knowledge distance in a knowledge base, *International Journal of Approximate Reasoning* **50** (2009) 174–188.

[15] Y.H. Qian, J.Y. Liang, C.Y. Dang, F. Wang, W. Xu, Knowledge distance in information systems, *Journal of Systems Science and Systems Engineering* **16** (2007) 434–449.

[16] D. Shifei, D. Hao, Research and Development of Attribute Reduction Algorithm Based on Rough Set, *IEEE, CCDC,* 2010 (648–653).

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

[18] X.Z. Zhou, B. Huang, Rough set-based attribute reduction under incomplete Information Systems, *Journal of Nanjing University of Science and Technology* **27** (2003) 630–636.

Ngày nhận bài 20 - 12 - 2011