

Một phương pháp gia tăng để tính độ chính xác và độ phủ của các luật quyết định trên khối dữ liệu có tập đối tượng thay đổi

Đỗ Thị Lan Anh^{1,2}, Trịnh Đình Thắng¹

¹Viện Công nghệ Thông tin, Trường Đại học Sư phạm Hà Nội 2

²Học viện Khoa học Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

Tác giả liên hệ: Đỗ Thị Lan Anh, dothilananh@hpu2.edu.vn

Ngày nhận bài: 25/09/2018, ngày sửa chữa: 17/04/2019, ngày duyệt đăng: 22/04/2019

Xem sớm trực tuyến: 26/05/2019, định danh DOI: 10.32913/mic-ict-research-vn.v2019.n1.804

Biên tập lĩnh vực điều phối phân biện và quyết định nhận đăng: PGS.TS. Lê Hoàng Sơn

Tóm tắt: Bài báo đưa ra mô hình tăng hoặc giảm tập đối tượng của khối quyết định. Từ đó trình bày các thuật toán gia tăng để tính ma trận độ chính xác và ma trận độ phủ của các luật quyết định trên khối dữ liệu có tập đối tượng thay đổi. Đồng thời phát biểu và chứng minh độ phức tạp của các thuật toán này.

Từ khóa: Phương pháp gia tăng, ma trận độ chính xác, ma trận độ phủ, khối dữ liệu, khối quyết định.

Title: An incremental method for calculating accuracy and coverage of decision laws on data block having changed object set

Abstract: The paper gives a model of increasing or decreasing the object set of a decision block. From there, we present the incremental algorithms to calculate the precision matrix and the coverage matrix of the decision laws on the data block having the object set changed. The complexities of these algorithms have also been stated and proved.

Keywords: Incremental method, precision matrix, coverage matrix, data block, decision block.

I. GIỚI THIỆU

Việc nghiên cứu để tìm kiếm các luật quyết định trên bảng quyết định bằng cách đánh giá các độ đo của các luật quyết định cũng như các cách tiếp cận gia tăng, xác định luật quyết định, v.v. đã được nhiều nhóm tác giả nghiên cứu, chẳng hạn như trong [1–5]. Tuy nhiên, luật quyết định trên bảng quyết định chỉ mang tính chất thời điểm mà không áp dụng được cho cả một quá trình, một khoảng thời gian nào đó. Khi đó, để khắc phục nhược điểm này nhóm tác giả đã tập trung nghiên cứu và đề xuất một mô hình và thuật toán tương ứng để phát hiện các luật quyết định trên khối dữ liệu [6]. Trên khối quyết định, việc nghiên cứu các tính chất khi làm mịn hoặc làm thô các giá trị của thuộc tính chỉ số trên khối cũng đã được nhóm tác giả quan tâm nghiên cứu [7]. Nối tiếp theo hướng nghiên cứu trên, trong bài báo này nhóm tác giả đã đưa một phương pháp để tính toán gia tăng ma trận độ chính xác (Acc) và độ phủ (Cov) của các luật quyết định khi bổ sung, hay loại bỏ các đối tượng ra khỏi khối dữ liệu, đồng thời đánh giá độ phức tạp của các thuật toán của phương pháp này.

II. CÁC KHÁI NIỆM CƠ BẢN

1. Khối

Định nghĩa 1: Gọi $R = (id; A_1, A_2, \dots, A_n)$ là một bộ hữu hạn các phần tử, trong đó id là tập chỉ số hữu hạn khác rỗng, $\{A_i\}$ với $i = 1, \dots, n$ là các thuộc tính. Mỗi thuộc tính A_i có miền giá trị tương ứng là $dom(A_i)$. Một khối r trên R gồm một số hữu hạn phần tử mà mỗi phần tử là một họ các ánh xạ từ tập chỉ số id đến các miền trị của các thuộc tính $\{A_i\}$. Nói một cách khác,

$$t \in r(R) \Leftrightarrow t = \{t^i : id \rightarrow dom(A_i)\}_{i=1, \dots, n}.$$

Khối được ký hiệu là $r(R)$, hoặc $r(id; A_1, A_2, \dots, A_n)$, hoặc đơn giản là r .

2. Lát cắt của khối

Định nghĩa 2 ([8]): Cho $R = (id; A_1, A_2, \dots, A_n)$, và $r(R)$ là một khối trên R . Với mỗi $x \in id$ ta kí hiệu $r(R_x)$ là một khối với $R_x = (\{x\}; A_1, A_2, \dots, A_n)$ sao cho

$$t_x \in r(R_x) \Leftrightarrow t_x = \{t_x^i = t^i|_x\}_{i=1, \dots, n},$$

trong đó $t_x^i(x) = t^i(x)$. Khi đó $r(R_x)$ được gọi là một lát cắt trên khối $r(R)$ tại điểm x , kí hiệu là r_x . Sau đây, để cho đơn giản ta sử dụng các kí hiệu

$$x^{(i)} = (x; A_i) \text{ và } id^{(i)} = \{x^{(i)} \mid x \in id\}.$$

Ta gọi $x^{(i)}$ ($x \in id, i = 1, \dots, n$) là các thuộc tính chỉ số của lược đồ khối $R = (id; A_1, A_2, \dots, A_n)$.

3. Khối thông tin

Định nghĩa 3 ([6]): Cho $R = (id; A_1, A_2, \dots, A_n)$ và r là một khối trên R . Khi đó khối thông tin là một bộ bốn $IB = (U, A, V, f)$ với U là tập các đối tượng thuộc r gọi là không gian các đối tượng, A là tập các thuộc tính chỉ số của đối tượng, V là tập giá trị của các đối tượng ứng với thuộc tính chỉ số $x^{(i)}$ và chúng được xác định như sau:

$$A = \bigcup_{i=1}^n id^{(i)}, \quad V = \bigcup_{x^{(i)} \in A} V_{x^{(i)}}.$$

Cuối cùng, $f : U \times A \rightarrow V$ là hàm thông tin thỏa mãn rằng với mọi $u \in U$ và với mọi $x^{(i)} \in A$ ta có $f(u) \in V_{x^{(i)}}$. Khi đó, ta gọi $f(u, x^{(i)})$ là giá trị của đối tượng u tại thuộc tính chỉ số $x^{(i)}$.

Định nghĩa 4 ([6]): Cho $R = (id; A_1, A_2, \dots, A_n)$, r là một khối trên R , và r_x là lát cắt của khối r tại $x \in id$. Khi đó, lát cắt của khối thông tin tại x là một bộ bốn $IB_x = (U, A_x, V_x, f_x)$ với U là tập các đối tượng thuộc r gọi là không gian các đối tượng, A_x là tập các thuộc tính chỉ số của đối tượng trên lát cắt tại x và được xác định như sau:

$$A_x = \bigcup_{i=1}^n x^{(i)}.$$

Tập thứ ba trong bộ, V_x , được xác định như sau:

$$V_x = \bigcup_{x^{(i)} \in A_x} V_{x^{(i)}},$$

trong đó $V_{x^{(i)}}$ là tập giá trị của các đối tượng ứng với thuộc tính chỉ số $x^{(i)}$. Cuối cùng, $f_x : U_x \times A_x \rightarrow V_x$ là hàm thông tin thỏa mãn rằng với mọi $u \in U$ và với mọi $x^{(i)} \in A_x$ ta có $f(u, x^{(i)}) \in V_{x^{(i)}}$.

4. Khối quyết định

Định nghĩa 5 ([6]): Cho khối thông tin $IB = (U, A, V, f)$ với U là không gian các đối tượng và $A = \bigcup_{i=1}^n id^{(i)}$. Khi đó, nếu A được chia thành hai tập C và D sao cho $C = \bigcup_{i=1}^k x^{(i)}$, $D = \bigcup_{i=k+1}^n x^{(i)}$, $x \in id$, với $1 \leq k < n$, thì khối thông tin IB gọi là khối quyết định và kí hiệu là $DB = (U, C \cup D, V, f)$ với C là tập các thuộc tính chỉ số điều kiện và D là tập các thuộc tính chỉ số quyết định. Ta có thể kí hiệu khối quyết định một cách đơn giản là $DB = (U, C \cup D)$.

Định nghĩa 6 ([6]): Cho khối $DB = (U, C \cup D, V, f)$ với C là tập các thuộc tính chỉ số điều kiện và D là tập các thuộc tính chỉ số quyết định. Khi đó lát cắt của khối quyết định tại x , $x \in id$, là một bộ bốn $DB_x = (U, C^x \cup D^x, V_x, f_x)$ với U là tập các đối tượng thuộc r gọi là không gian các đối tượng, $C^x = \bigcup_{x^{(i)} \in A_x} V_{x^{(i)}}$, $D^x = \bigcup_{i=1}^k x^{(i)}$, $A_x = C^x \cup D^x$, $V_x = \bigcup_{i=k+1}^n x^{(i)}$, với $V_{x^{(i)}}$ là tập các giá trị của các đối tượng ứng với thuộc tính chỉ số $x^{(i)}$, $f_x : U_x \times A_x \rightarrow V_x$ là hàm thông tin thỏa mãn rằng với mọi $u \in U$ và với mọi $x^{(i)} \in A_x$ ta có $f(u, x^{(i)}) \in V_{x^{(i)}}$.

5. Luật quyết định trên khối và trên lát cắt

Định nghĩa 7 ([6]): Cho khối quyết định $DB = (U, C \cup D)$ với U là không gian các đối tượng, $C = \bigcup_{i=1}^k x^{(i)}$, $D = \bigcup_{i=k+1}^n x^{(i)}$, $C^x = \bigcup_{i=1}^k x^{(i)}$, $D^x = \bigcup_{i=k+1}^n x^{(i)}$, $x \in id$. Khi đó,

$$U/C = \{C_1, C_2, \dots, C_m\},$$

$$U/C^x = \{C_{x1}, C_{x2}, \dots, C_{xt_x}\},$$

$$U/D = \{D_1, D_2, \dots, D_h\},$$

$$U/D^x = \{D_{x1}, D_{x2}, \dots, D_{xs_x}\},$$

tương ứng là các phân hoạch được sinh ra bởi C , C^x , D , D^x và m , h , t_x , s_x lần lượt là số lớp tương đương của các phân hoạch U/C , U/C^x , U/D , U/D^x . Một luật quyết định trên khối có dạng

$$C_i \longrightarrow D_j, \quad i = 1, \dots, m \text{ và } j = 1, \dots, h$$

và trên lát cắt tại điểm x có dạng

$$C_{xi} \longrightarrow D_{xj}, \quad i = 1, \dots, t_x, \text{ và } j = 1, \dots, s_x.$$

Mệnh đề 1 ([6]): Cho khối quyết định $DB = (U, C \cup D)$ với U là không gian các đối tượng $C = \bigcup_{i=1}^k x^{(i)}$, $D = \bigcup_{i=k+1}^n x^{(i)}$, $C^x = \bigcup_{i=1}^k x^{(i)}$, $D^x = \bigcup_{i=k+1}^n x^{(i)}$, $x \in id$, $1 \leq k < n$, và các phân hoạch U/C , U/C^x , U/D , U/D^x là các phân hoạch được sinh ra bởi C , C^x , D , D^x , như trong định nghĩa 7. Khi đó, với mọi $C_i \in U/C$ và với mọi $D_j \in U/D$, $i = 1, \dots, m$, $j = 1, \dots, h$, ta có

$$C_i = \bigcap_{x \in id} C_{xp_x}, \quad D_j = \bigcap_{x \in id} D_{xq_x},$$

với $p_x \in \{1, 2, \dots, t_x\}$ và $q_x \in \{1, 2, \dots, s_x\}$.

Định nghĩa 8 ([6]): Cho khối quyết định $DB = (U, C \cup D)$, $C_i \in U/C$, $D_j \in U/D$, $C_{xp} \in U/C^x$, $D_{xq} \in U/D^x$, với $i = 1, \dots, m$, $j = 1, \dots, h$, $p \in \{1, \dots, t_x\}$, $q \in \{1, \dots, s_x\}$, $x \in id$.

Khi đó, độ hỗ trợ, độ chính xác và độ phủ của luật quyết định $C_i \longrightarrow D_j$ trên khối được cho tương ứng như sau:

$$\text{Sup}(C_i, D_j) = |C_i \cap D_j|,$$

$$\text{Acc}(C_i, D_j) = \frac{|C_i \cap D_j|}{|C_i|},$$

$$\text{Cov}(C_i, D_j) = \frac{|C_i \cap D_j|}{|D_j|}.$$

Còn độ hỗ trợ, độ chính xác và độ phủ của luật quyết định $C_{xp} \rightarrow D_{xq}$ trên lát cắt của khối tại điểm x được cho tương ứng như sau:

$$\begin{aligned} \text{Sup}(C_{xp}, D_{xq}) &= |C_{xp} \cap D_{xq}|, \\ \text{Acc}(C_{xp}, D_{xq}) &= \frac{|C_{xp} \cap D_{xq}|}{|C_{xp}|}, \\ \text{Cov}(C_{xp}, D_{xq}) &= \frac{|C_{xp} \cap D_{xq}|}{|D_{xq}|}. \end{aligned}$$

Từ định nghĩa trên, ta có các kết quả sau:

$$\begin{aligned} 0 \leq \text{Acc}(C_i, D_j) \leq 1, \quad 0 \leq \text{Acc}(C_{xp}, D_{xq}) \leq 1, \\ \sum_{j=1}^h \text{Acc}(C_i, D_j) = 1, \quad \sum_{q=1}^{s_x} \text{Acc}(C_{xp}, D_{xq}) = 1, \\ 0 \leq \text{Cov}(C_i, D_j) \leq 1, \quad 0 \leq \text{Cov}(C_{xp}, D_{xq}) \leq 1, \\ \sum_{j=1}^m \text{Cov}(C_i, D_j) = 1, \quad \sum_{p=1}^{t_x} \text{Cov}(C_{xp}, D_{xq}) = 1. \end{aligned}$$

Ta có thể biểu diễn độ đo của các luật quyết định trên khối dưới dạng các ma trận độ hỗ trợ, độ chính xác và độ phủ tương ứng như sau:

$$\begin{aligned} \text{Sup}(C, D) &= \begin{pmatrix} \text{Sup}(C_1, D_1) & \cdots & \text{Sup}(C_1, D_h) \\ \vdots & \ddots & \vdots \\ \text{Sup}(C_m, D_1) & \cdots & \text{Sup}(C_m, D_h) \end{pmatrix}, \\ \text{Acc}(C, D) &= \begin{pmatrix} \text{Acc}(C_1, D_1) & \cdots & \text{Acc}(C_1, D_h) \\ \vdots & \ddots & \vdots \\ \text{Acc}(C_m, D_1) & \cdots & \text{Acc}(C_m, D_h) \end{pmatrix}, \\ \text{Cov}(C, D) &= \begin{pmatrix} \text{Cov}(C_1, D_1) & \cdots & \text{Cov}(C_1, D_h) \\ \vdots & \ddots & \vdots \\ \text{Cov}(C_m, D_1) & \cdots & \text{Cov}(C_m, D_h) \end{pmatrix}. \end{aligned}$$

Với các luật quyết định trên các lát cắt của khối thì ta cũng có các ma trận độ hỗ trợ, độ chính xác và độ phủ tương tự.

Mệnh đề 2 ([1, 7]): Cho khối quyết định $DB = (U, C \cup D)$, với U là không gian các đối tượng, $C = \cup_{i=1}^k x^{(i)}$, $D = \cup_{i=k+1}^n x^{(i)}$, $x \in id$. Khi đó với mọi $C_i \in U/C$ và với mọi $D_j \in U/D$, $i \in \{1, \dots, m\}$ và $j \in \{1, \dots, h\}$, ta có

$$\begin{aligned} \text{Acc}(C_i, D_j) &= \frac{\text{Sup}(C_i, D_j)}{\sum_{q=1}^h \text{Sup}(C_i, D_q)}, \\ \text{Cov}(C_i, D_j) &= \frac{\text{Sup}(C_i, D_j)}{\sum_{p=1}^m \text{Sup}(C_p, D_j)}. \end{aligned}$$

III. KẾT QUẢ

1. Mô hình bổ sung và loại bỏ các đối tượng trên khối quyết định và trên lát cắt

Cho khối quyết định $DB = (U, C \cup D, V, f)$ như trong định nghĩa 7. Giả sử, ta cần bổ sung vào khối quyết định này N đối tượng, kí hiệu AN và loại bỏ cũng ở khối này M đối tượng, kí hiệu DM . Khi đó, ta cần tính các ma trận độ chính xác Acc và ma trận độ phủ Cov trên khối và trên lát cắt sau khi bổ sung và loại bỏ các đối tượng của khối quyết định. Các kết quả này sẽ giúp tìm ra các luật quyết định trên khối và trên lát cắt.

Giả sử khi bổ sung N đối tượng vào khối quyết định thì N đối tượng này sẽ sinh thêm p lớp tương đương điều kiện mới đối với tập U/C , p_x lớp tương đương điều kiện mới đối với tập U/C^x , q lớp tương đương quyết định mới đối với tập U/D và q_x lớp tương đương quyết định mới đối với tập U/D^x . Kí hiệu N_i là số đối tượng được bổ sung cho lớp $C_i \in U/C$ ($i = 1, \dots, m+p$), N_{xi} là số đối tượng được bổ sung cho lớp $C_{xi} \in U/C^x$ ($i = 1, \dots, t_x+p_x$) và trong N_i đối tượng này có N_{ij} đối tượng được bổ sung cho lớp $D_j \in U/D$ ($j = 1, \dots, h+q$), còn trên lát cắt tại x thì trong N_{xi} đối tượng này có N_{xij} đối tượng được bổ sung cho lớp $D_{xj} \in U/D^x$ ($j = 1, \dots, s_x+q_x$). Trong M đối tượng bị loại bỏ có M_i đối tượng bị loại khỏi lớp $C_i \in U/C$ ($i = 1, \dots, m$) có M_{ij} đối tượng bị loại khỏi lớp D_i ($j = 1, \dots, h$), còn trên lát cắt tại x thì trong M_{xi} ($j = 1, \dots, s_x$) đối tượng bị loại khỏi lớp $D_{xj} \in U/D^x$.

Từ mô hình này ta có

$$\begin{aligned} N_i &= \sum_{j=1}^{h+q} N_{ij}, & N &= \sum_{i=1}^{m+p} N_i = \sum_{i=1}^{m+p} \sum_{j=1}^{h+q} N_{ij}, \\ M_i &= \sum_{j=1}^h M_{ij}, & M &= \sum_{i=1}^m M_i = \sum_{i=1}^m \sum_{j=1}^h M_{ij}, \\ N_{xi} &= \sum_{j=1}^{s_x+q_x} N_{xij}, & N_x &= \sum_{i=1}^{t_x+p_x} N_{xi} = \sum_{i=1}^{t_x+p_x} \sum_{j=1}^{s_x+q_x} N_{xij}, \\ M_{xi} &= \sum_{j=1}^{s_x} M_{xij}, & M_x &= \sum_{i=1}^{t_x} M_{xi} = \sum_{i=1}^{t_x} \sum_{j=1}^{s_x} M_{xij}. \end{aligned}$$

Ta kí hiệu các lớp tương đương sau khi bổ sung và loại bỏ các đối tượng là

$$\begin{aligned} U/C &= \{C'_1, C'_2, \dots, C'_m, \dots\}, \\ U/C^x &= \{C'_{x1}, C'_{x2}, \dots, C'_{xt_x}, \dots\}, \\ U/D &= \{D'_1, D'_2, \dots, D'_h, \dots\}, \\ U/D^x &= \{D'_{x1}, D'_{x2}, \dots, D'_{xs_x}, \dots\}. \end{aligned}$$

Từ định nghĩa ta thấy C_i và C'_i ($i = 1, \dots, m$) chỉ khác nhau về số lượng phần tử khi bổ sung và loại bỏ các đối tượng, nghĩa là với mọi $a \in C$ ta có $f(C'_i, a) = f(C_i, a)$,

còn với các lớp từ $C'_{m+1}, C'_{m+2}, \dots, C'_{m+p}$ là các lớp tương đương điều kiện hoàn toàn mới. Trên lát cắt tại x thì C_{xi} và C'_{xi} ($i = 1, \dots, t_x$) cũng chỉ khác nhau về số lượng phần tử sau khi bổ sung và loại bỏ các đối tượng và với mọi $a \in C$ ta có $f(C'_{xi}, a) = f(C_{xi}, a)$, còn với các lớp từ C'_{xi} ($i = t_x + 1, \dots, t_x + p_x$) là các lớp tương đương điều kiện hoàn toàn mới.

Tương tự, D_i và D'_i ($i = 1, \dots, h$) cũng chỉ khác nhau về số lượng phần tử, nghĩa là với mọi $a \in D$ ta có $f(D'_i, a) = f(D_i, a)$, còn với các lớp từ $D'_{h+1}, D'_{h+2}, \dots, D'_{h+q}$ là các lớp tương đương quyết định hoàn toàn mới. Trên lát cắt tại x thì D_{xi} và D'_{xi} ($i = 1, \dots, s_x$) cũng chỉ khác nhau về số lượng phần tử sau khi bổ sung và loại bỏ các đối tượng, còn các lớp từ D'_{hj} ($j = s_x + 1, \dots, s_x + q_x$) là các lớp tương đương quyết định hoàn toàn mới.

Từ tính chất trên của các lớp tương đương điều kiện, các lớp tương đương quyết định trên khối và trên lát cắt ta có mối quan hệ giữa số lượng các phần tử của các lớp này như sau:

$$|C'_i| = \begin{cases} |C_i| + N_i - M_i, & i = 1, \dots, m, \\ N_i, & i = m + 1, \dots, m + p, \end{cases}$$

$$|C'_i| = \begin{cases} |C_i| + \sum_{j=1}^{h+q} N_{ij} - \sum_{j=1}^h M_{ij}, & i = 1, \dots, m, \\ \sum_{j=1}^{h+q} N_{ij}, & i = m + 1, \dots, m + p, \end{cases}$$

$$|D'_j| = \begin{cases} |D_j| + \sum_{i=1}^{m+p} N_{ij} - \sum_{i=1}^m M_{ij}, & j = 1, \dots, h, \\ \sum_{i=1}^{m+p} N_{ij}, & j = h + 1, \dots, h + q, \end{cases}$$

$$|C'_{xi}| = \begin{cases} |C_{xi}| + N_{xi} - M_{xi}, & i = 1, \dots, m, \\ N_{xi}, & i = m + 1, \dots, m + p, \end{cases}$$

$$|C'_{xi}| = \begin{cases} |C_{xi}| + \sum_{j=1}^{s_x+q_x} N_{xij} - \sum_{j=1}^{s_x} M_{xij}, & i = 1, \dots, t_x, \\ \sum_{j=1}^{s_x+q_x} N_{xij}, & i = t_x + 1, \dots, t_x + p_x, \end{cases}$$

$$|D'_{xj}| = \begin{cases} |D_{xj}| + \sum_{i=1}^{t_x+p_x} N_{xij} - \sum_{i=1}^{t_x} M_{xij}, & j = 1, \dots, s_x, \\ \sum_{i=1}^{t_x+p_x} N_{xij}, & j = s_x + 1, \dots, s_x + q_x. \end{cases}$$

Mệnh đề 3: Cho khối quyết định $DB = (U, C \cup D, V, f)$, AN và DM là tập các đối tượng bổ sung và loại bỏ tương ứng đối với khối quyết định DB . Khi đó ma trận độ chính xác trên khối quyết định sau khi bổ sung N và loại bỏ M đối tượng với $i \in \{1, \dots, m + p\}$ và $j \in \{1, \dots, h + q\}$ sẽ là

$$\text{Acc}(C', D') = \text{Acc}(C'_i, D'_j)_{ixj},$$

với $i \in \{1, \dots, m\}$ và $j \in \{1, \dots, h\}$ sẽ là

$$\text{Acc}(C'_i, D'_j) = \frac{|C_i \cap D_j| + N_{ij} - M_{ij}}{|C_i| + \sum_{j'=1}^{h+q} N_{ij'} - \sum_{j'=1}^h M_{ij'}},$$

với $i \in \{1, \dots, m\}$ và $j \in \{h + 1, \dots, h + q\}$ sẽ là

$$\text{Acc}(C'_i, D'_j) = \frac{N_{ij}}{|C_i| + \sum_{j'=1}^{h+q} N_{ij'} - \sum_{j'=1}^h M_{ij'}},$$

và với $i \in \{m + 1, \dots, m + p\}$ và $j \in \{1, \dots, h + q\}$ sẽ là

$$\text{Acc}(C'_i, D'_j) = \frac{N_{ij}}{\sum_{j'=1}^{h+q} N_{ij'}}.$$

Chứng minh: Ta lần lượt xác định giá trị của $\text{Acc}(C'_i, D'_j)$ theo ba trường hợp sau.

(i) $i = 1, \dots, m$ và $j = 1, \dots, h$: Theo định nghĩa, ta có

$$\text{Acc}(C'_i, D'_j) = \frac{|C'_i \cap D'_j|}{|C'_i|}.$$

Khi đó

$$|C'_i \cap D'_j| = |C_i \cap D_j| + N_{ij} - M_{ij} = A_1,$$

$$|C'_i| = |C_i| + N_i - M_i = |C_i| + \sum_{j'=1}^{h+q} N_{ij'} - \sum_{j'=1}^h M_{ij'} = B_1.$$

Kết quả này cho ta

$$\text{Acc}(C'_i, D'_j) = \frac{|C'_i \cap D'_j|}{|C'_i|} = \frac{A_1}{B_1}.$$

(ii) $i = 1, \dots, m$ và $j = h + 1, \dots, h + q$: Theo định nghĩa, ta có

$$\text{Acc}(C'_i, D'_j) = \frac{|C'_i \cap D'_j|}{|C'_i|}.$$

Khi đó

$$|C'_i \cap D'_j| = N_{ij} = A_2,$$

$$|C'_i| = |C_i| + N_i - M_i = |C_i| + \sum_{j=1}^{h+q} N_{ij} - \sum_{j=1}^h M_{ij} = B_2.$$

Suy ra

$$\text{Acc}(C'_i, D'_j) = \frac{|C'_i \cap D'_j|}{|C'_i|} = \frac{A_2}{B_2}.$$

(iii) $i = m + 1, \dots, m + p$ và $j = 1, \dots, h + q$: Theo định nghĩa, ta có

$$\text{Acc}(C'_i, D'_j) = \frac{|C'_i \cap D'_j|}{|C'_i|}.$$

Khi đó

$$|C'_i \cap D'_j| = N_{ij} = A_3,$$

$$|C'_i| = \sum_{j=1}^{h+q} N_{ij} = B_3.$$

Từ đó

$$\text{Acc}(C'_i, D'_j) = \frac{|C'_i \cap D'_j|}{|C'_i|} = \frac{A_3}{B_3}.$$

□

Mệnh đề 4: Cho khối quyết định $DB = (U, C \cup D, V, f)$, AN và DM là tập các đối tượng bổ sung và loại bỏ tương ứng đối với khối quyết định DB . Khi đó ma trận độ chính xác trên lát cắt x sau khi bổ sung N và loại bỏ M đối tượng với $i = 1, \dots, t_x + p_x$ và $j = 1, \dots, s_x + q_x$ sẽ là

$$\text{Acc}(C'_x, D'_x) = \text{Acc}(C'_{xi}, D'_{xj})_{ixj},$$

với $i \in \{1, \dots, t_x\}$ và $j \in \{1, \dots, s_x\}$ sẽ là

$$\text{Acc}(C'_{xi}, D'_{xj}) = \frac{|C_{xi} \cap D_{xj}| + N_{xij} - M_{xij}}{|C_{xi}| + \sum_{j'=1}^{s_x+q_x} N_{xij'} - \sum_{j'=1}^{s_x} M_{xij'}}$$

với $i \in \{1, \dots, t_x\}$ và $j \in \{s_x + 1, \dots, s_x + q_x\}$ sẽ là

$$\text{Acc}(C'_{xi}, D'_{xj}) = \frac{N_{xij}}{|C_{xi}| + \sum_{j'=1}^{s_x+q_x} N_{xij'} - \sum_{j'=1}^{s_x} M_{xij'}}$$

và với $i \in \{t_x + 1, \dots, t_x + p_x\}$, $j \in \{1, \dots, s_x + q_x\}$ sẽ là

$$\text{Acc}(C'_{xi}, D'_{xj}) = \frac{N_{xij}}{\sum_{j'=1}^{h+q} N_{xij'}}$$

Chứng minh: Ta cũng xác định giá trị của $\text{Acc}(C'_{xi}, D'_{xj})$ theo ba trường hợp sau.

(i) $i = 1, \dots, t_x$ và $j = 1, \dots, s_x$: Theo định nghĩa, ta có

$$\text{Acc}(C'_{xi}, D'_{xj}) = \frac{|C'_{xi} \cap D'_{xj}|}{|C'_{xi}|}$$

Khi đó

$$|C'_{xi} \cap D'_{xj}| = |C_{xi} \cap D_{xj}| + N_{xij} - M_{xij} = A_{x1},$$

$$|C'_{xi}| = |C_{xi}| + N_{xi} - M_{xi} = |C_{xi}| + \sum_{j'=1}^{s_x+q_x} N_{xij'} - \sum_{j'=1}^{s_x} M_{xij'} = B_{x1}.$$

Suy ra

$$\text{Acc}(C'_{xi}, D'_{xj}) = \frac{|C'_{xi} \cap D'_{xj}|}{|C'_{xi}|} = \frac{A_{x1}}{B_{x1}}.$$

(ii) $i = 1, \dots, t_x$ và $j = s_x + 1, \dots, s_x + q_x$: Theo định nghĩa, ta có

$$\text{Acc}(C'_{xi}, D'_{xj}) = \frac{|C'_{xi} \cap D'_{xj}|}{|C'_{xi}|}$$

Khi đó

$$|C'_{xi} \cap D'_{xj}| = N_{xij} = A_{x2},$$

$$|C'_{xi}| = |C_{xi}| + N_{xi} - M_{xi} = |C_{xi}| + \sum_{j'=1}^{s_x+q_x} N_{xij'} - \sum_{j'=1}^{s_x} M_{xij'} = B_{x2}.$$

Suy ra

$$\text{Acc}(C'_{xi}, D'_{xj}) = \frac{|C'_{xi} \cap D'_{xj}|}{|C'_{xi}|} = \frac{A_{x2}}{B_{x2}}.$$

(iii) $i = t_x + 1, \dots, t_x + p_x$ và $j = 1, \dots, s_x + q_x$: Theo định nghĩa, ta có

$$\text{Acc}(C'_{xi}, D'_{xj}) = \frac{|C'_{xi} \cap D'_{xj}|}{|C'_{xi}|}$$

Khi đó

$$|C'_{xi} \cap D'_{xj}| = N_{xij} = A_{x3},$$

$$|C'_{xi}| = N_{xi} = \sum_{j=1}^{s_x+q_x} N_{xij} = B_{x3}.$$

Suy ra

$$\text{Acc}(C'_{xi}, D'_{xj}) = \frac{|C'_{xi} \cap D'_{xj}|}{|C'_{xi}|} = \frac{A_{x3}}{B_{x3}}.$$

□

Mệnh đề 5: Cho khối quyết định $DB = (U, C \cup D, V, f)$, AN và DM là tập các đối tượng bổ sung và loại bỏ tương ứng đối với khối quyết định DB . Khi đó ma trận độ phủ trên khối quyết định sau khi bổ sung N và loại bỏ M đối tượng với $i \in \{1, \dots, m + p\}$ và $j \in \{1, \dots, h + q\}$ sẽ là

$$\text{Cov}(C', D') = \text{Cov}(C'_i, D'_j)_{ixj},$$

với $i \in \{1, \dots, m\}$ và $j \in \{1, \dots, h\}$ sẽ là

$$\text{Cov}(C'_i, D'_j) = \frac{|C_i \cap D_j| + N_{ij} - M_{ij}}{|D_j| + \sum_{i'=1}^{m+p} N_{i'j} - \sum_{i'=1}^m M_{i'j}}$$

với $i \in \{m + 1, \dots, m + p\}$ và $j \in \{1, \dots, h\}$ sẽ là

$$\text{Cov}(C'_i, D'_j) = \frac{N_{ij}}{|D_j| + \sum_{i'=1}^{m+p} N_{i'j} - \sum_{i'=1}^m M_{i'j}},$$

và với $i \in \{1, \dots, m + p\}$, và $j \in \{h + 1, \dots, h + q\}$ sẽ là

$$\text{Cov}(C'_i, D'_j) = \frac{N_{ij}}{\sum_{i'=1}^{m+p} N_{i'j}}.$$

Chứng minh: Ta cũng xác định giá trị của $\text{Cov}(C'_i, D'_j)$ theo ba trường hợp sau.

(i) $i = 1, \dots, m$ và $j = 1, \dots, h$: Theo định nghĩa, ta có

$$\text{Cov}(C'_i, D'_j) = \frac{|C'_i \cap D'_j|}{|D'_j|}$$

Khi đó

$$|C'_i \cap D'_j| = |C_i \cap D_j| + N_{ij} - M_{ij} = A'_1,$$

$$|D'_j| = |D_j| + N_j - M_j = |D_j| + \sum_{i'=1}^{m+p} N_{i'j} - \sum_{i'=1}^m M_{i'j} = B'_1.$$

Suy ra

$$\text{Cov}(C'_i, D'_j) = \frac{|C'_i \cap D'_j|}{|D'_j|} = \frac{A'_1}{B'_1}.$$

(ii) $i = m + 1, \dots, m + p$ và $j = 1, \dots, h$: Theo định nghĩa, ta có

$$\text{Cov}(C'_i, D'_j) = \frac{|C'_i \cap D'_j|}{|D'_j|}$$

Khi đó

$$|C'_i \cap D'_j| = N_{ij} = A'_2,$$

$$|D'_j| = |D_j| + \sum_{i'=1}^{m+p} N_{i'j} - \sum_{i'=1}^m M_{i'j} = B'_2.$$

Suy ra

$$\text{Cov}(C'_i, D'_j) = \frac{|C'_i \cap D'_j|}{|D'_j|} = \frac{A'_2}{B'_2}.$$

(iii) $i = 1, \dots, m + p$ và $j = h + 1, \dots, h + q$: Theo định nghĩa, ta có

$$\text{Cov}(C'_i, D'_j) = \frac{|C'_i \cap D'_j|}{|D'_j|}.$$

Khi đó

$$|C'_i \cap D'_j| = N_{ij} = A'_3,$$

$$|D'_j| = \sum_{i'=1}^{m+p} N_{i'j} = B'_3.$$

Suy ra

$$\text{Cov}(C'_i, D'_j) = \frac{|C'_i \cap D'_j|}{|D'_j|} = \frac{A'_3}{B'_3}.$$

□

Mệnh đề 6: Cho khối quyết định $DB = (U, C \cup D, V, f)$, AN và DM là tập các đối tượng bổ sung và loại bỏ tương ứng với khối quyết định DB . Khi đó ma trận độ phủ trên lát cắt tại x sau khi bổ sung N và loại bỏ M đối tượng với $i = 1, \dots, t_x + p_x$ và $j = 1, \dots, s_x + q_x$ sẽ là

$$\text{Cov}(C'_x, D'_x) = \text{Cov}(C'_{xi}, D'_{xj})_{ixj},$$

với $i \in \{1, \dots, t_x\}$ và $j \in \{1, \dots, s_x\}$ sẽ là

$$\text{Cov}(C'_{xi}, D'_{xj}) = \frac{|C_{xi} \cap D_{xj}| + N_{xij} - M_{xij}}{|D_{xj}| + \sum_{i'=1}^{m+p} N_{xi'j} - \sum_{i'=1}^m M_{xi'j}},$$

với $i \in \{t_x + 1, \dots, t_x + p_x\}$ và $j \in \{1, \dots, s_x\}$ sẽ là

$$\text{Cov}(C'_{xi}, D'_{xj}) = \frac{N_{xij}}{|D_{xj}| + \sum_{i'=1}^{t_x+p_x} N_{xi'j} - \sum_{i'=1}^{t_x} M_{xi'j}},$$

và với $i \in \{1, \dots, t_x + p_x\}$ và $j \in \{s_x + 1, \dots, s_x + q_x\}$ sẽ là

$$\text{Cov}(C'_{xi}, D'_{xj}) = \frac{N_{xij}}{\sum_{i'=1}^{t_x+p_x} N_{xi'j}}.$$

Chứng minh: Ta cũng xác định giá trị của $\text{Cov}(C'_{xi}, D'_{xj})$ theo ba trường hợp sau.

(i) $i = 1, \dots, m$ và $j = 1, \dots, h$: Theo định nghĩa, ta có

$$\text{Cov}(C'_{xi}, D'_{xj}) = \frac{|C'_{xi} \cap D'_{xj}|}{|D'_{xj}|}.$$

Khi đó

$$|C'_{xi} \cap D'_{xj}| = |C_{xi} \cap D_{xj}| + N_{xij} - M_{xij} = A'_{x1},$$

$$|D'_{xj}| = |D_{xj}| + N_{xj} - M_{xj} = |D_{xj}| + \sum_{i'=1}^{t_x+p_x} N_{xi'j} - \sum_{i'=1}^{t_x} M_{xi'j} = B'_{x1}.$$

Suy ra

$$\text{Cov}(C'_{xi}, D'_{xj}) = \frac{|C'_{xi} \cap D'_{xj}|}{|D'_{xj}|} = \frac{A'_{x1}}{B'_{x1}}.$$

(ii) $i = t_x + 1, \dots, t_x + p_x$ và $j = 1, \dots, s_x$: Theo định nghĩa, ta có

$$\text{Cov}(C'_{xi}, D'_{xj}) = \frac{|C'_{xi} \cap D'_{xj}|}{|D'_{xj}|}.$$

Khi đó

$$|C'_{xi} \cap D'_{xj}| = N_{xij} = A'_{x2},$$

$$|D'_{xj}| = |D_{xj}| + \sum_{i'=1}^{t_x+p_x} N_{xi'j} - \sum_{i'=1}^{t_x} M_{xi'j} = B_{x2}.$$

Suy ra

$$\text{Cov}(C'_{xi}, D'_{xj}) = \frac{|C'_{xi} \cap D'_{xj}|}{|D'_{xj}|} = \frac{A'_{x2}}{B'_{x2}}.$$

(iii) $i = 1, \dots, t_x + p_x$ và $j = s_x + 1, \dots, s_x + q_x$: Theo định nghĩa, ta có

$$\text{Cov}(C'_{xi}, D'_{xj}) = \frac{|C'_{xi} \cap D'_{xj}|}{|D'_{xj}|}.$$

Khi đó

$$|C'_{xi} \cap D'_{xj}| = N_{xij} = A'_{x3},$$

$$|D'_{xj}| = \sum_{i'=1}^{t_x+p_x} N_{xi'j} = B'_{x3}$$

Suy ra

$$\text{Cov}(C'_{xi}, D'_{xj}) = \frac{|C'_{xi} \cap D'_{xj}|}{|D'_{xj}|} = \frac{A'_{x3}}{B'_{x3}}.$$

□

2. Tính toán gia tăng Acc và Cov khi bổ sung và loại bỏ các đối tượng trên khối quyết định

1) *Bổ sung đối tượng x vào khối quyết định:* Chúng ta có bốn trường hợp sau.

(i) **Sinh lớp điều kiện mới và lớp quyết định mới:** Với trường hợp này ta có $x \notin C_i$ ($i = 1, \dots, m$) và $x \notin D_j$ ($j = 1, \dots, h$). Suy ra $x \in C'_{m+1}$ và $x \in D'_{h+1}$. Do đó $\text{Acc}(C'_{m+1}, D'_{h+1}) = 1$ và $\text{Cov}(C'_{m+1}, D'_{h+1}) = 1$. Với mọi $j = 1, \dots, h$, ta có $\text{Acc}(C'_{m+1}, D'_j) = \text{Cov}(C'_{m+1}, D'_j) = 0$. Với mọi $i = 1, \dots, m$, ta có $\text{Acc}(C'_i, D'_{h+1}) = \text{Cov}(C'_i, D'_{h+1}) = 0$. Mặt khác, với mọi $i = 1, \dots, m$ và $j = 1, \dots, h$, ta có $\text{Acc}(C'_i, D'_j) = \text{Cov}(C_i, D_j)$ và $\text{Cov}(C'_i, D'_j) = \text{Cov}(C_i, D_j)$.

(ii) **Chỉ sinh lớp điều kiện mới:** Trường hợp này ta có $x \notin C_i$ ($i = 1, \dots, m$) và tồn tại $j \in \{1, \dots, h\}$ sao cho $x \in D_{j^*}$. Suy ra $x \in C'_{m+1}$ và $x \in D'_{j^*}$. Do đó ta có $\text{Acc}(C'_{m+1}, D'_{j^*}) = 1$ và

$$\text{Cov}(C'_{m+1}, D'_{j^*}) = \frac{1}{|D_{j^*}| + 1}.$$

Nếu $k \neq j^*$ thì $\text{Acc}(C'_{m+1}, D'_k) = \text{Cov}(C'_{m+1}, D'_k) = 0$. Nếu $i \neq m + 1$ thì $\text{Acc}(C'_i, D'_{j^*}) = \text{Acc}(C_i, D_{j^*})$ và

$$\text{Cov}(C'_i, D'_{j^*}) = \frac{|C_i \cap D_{j^*}|}{|D_{j^*}| + 1}.$$

Mặt khác, với mọi $i \neq m + 1$ và mọi $j \neq j^*$, ta có $\text{Acc}(C'_i, D'_j) = \text{Acc}(C_i, D_j)$ và $\text{Cov}(C'_i, D'_j) = \text{Cov}(C_i, D_j)$.

(iii) **Chỉ sinh lớp quyết định mới:** Trường hợp này, $x \notin D_j$ ($j = 1, \dots, h$) và tồn tại $i^* \notin \{1, \dots, m\}$ sao cho $x \in C_{i^*}$. Suy ra $x \in D'_{h+1}$ và $x \in C'_{i^*}$. Do đó $\text{Cov}(C'_{i^*}, D'_{h+1}) = 1$ và

$$\text{Acc}(C'_{i^*}, D'_{h+1}) = \frac{1}{|C'_{i^*}| + 1}.$$

Nếu $i \neq i^*$ thì $\text{Acc}(C'_i, D'_{h+1}) = \text{Cov}(C'_i, D'_{h+1}) = 0$. Nếu $k \neq h + 1$ thì $\text{Cov}(C'_{i^*}, D'_k) = \text{Cov}(C_{i^*}, D_k)$ và

$$\text{Acc}(C'_{i^*}, D'_k) = \frac{|C_{i^*} \cap D_k|}{|C_{i^*}| + 1}.$$

Mặt khác, với mọi $i \neq i^*$ và mọi $j \neq h + 1$ ta có $\text{Acc}(C'_i, D'_j) = \text{Acc}(C_i, D_j)$, và $\text{Cov}(C'_i, D'_j) = \text{Cov}(C_i, D_j)$.

(iv) **Không sinh thêm lớp điều kiện mới hoặc lớp quyết định mới:** Với trường hợp này, tồn tại $i^* \in \{1, \dots, m\}$ sao cho $x \in C_{i^*}$ và tồn tại $j^* \in \{1, \dots, h\}$ sao cho $x \in D_{j^*}$. Như vậy, việc bổ sung phần tử x sẽ tăng thêm số phần tử của C_{i^*} và D_{j^*} . Khi đó

$$\text{Acc}(C'_{i^*}, D'_{j^*}) = \frac{|C_{i^*} \cap D_{j^*}| + 1}{|C_{i^*}| + 1},$$

$$\text{Cov}(C'_{i^*}, D'_{j^*}) = \frac{|C_{i^*} \cap D_{j^*}| + 1}{|D_{j^*}| + 1}.$$

Nếu $k \neq j^*$ thì $\text{Cov}(C'_{i^*}, D'_k) = \text{Cov}(C_{i^*}, D_k)$ và

$$\text{Acc}(C'_{i^*}, D'_k) = \frac{|C_{i^*} \cap D_k| + 1}{|C_{i^*}| + 1}.$$

Nếu $u \neq i^*$ thì $\text{Acc}(C'_u, D'_{j^*}) = \text{Acc}(C_u, D_{j^*})$ và

$$\text{Cov}(C'_u, D'_{j^*}) = \frac{|C_u \cap D_{j^*}| + 1}{|D_{j^*}| + 1}.$$

Nếu $i \neq i^*$ và $j \neq j^*$ thì $\text{Acc}(C'_i, D'_j) = \text{Acc}(C_i, D_j)$ và $\text{Cov}(C'_i, D'_j) = \text{Cov}(C_i, D_j)$.

2) **Loại bỏ phần tử x ra khỏi khối quyết định:** Với trường hợp này, tồn tại $i^* \in \{1, \dots, m\}$ sao cho $x \in C_{i^*}$ và tồn tại $j^* \in \{1, \dots, h\}$ sao cho $x \in D_{j^*}$. Như vậy, việc loại bỏ phần tử x sẽ làm giảm đi số phần tử của C_{i^*} và D_{j^*} . Khi đó

$$\text{Acc}(C'_{i^*}, D'_{j^*}) = \frac{|C_{i^*} \cap D_{j^*}| - 1}{|C_{i^*}| - 1},$$

$$\text{Cov}(C'_{i^*}, D'_{j^*}) = \frac{|C_{i^*} \cap D_{j^*}| - 1}{|D_{j^*}| - 1}.$$

Nếu $k \neq j^*$ thì $\text{Cov}(C'_{i^*}, D'_k) = \text{Cov}(C_{i^*}, D_k)$ và

$$\text{Acc}(C'_{i^*}, D'_k) = \frac{|C_{i^*} \cap D_k|}{|C_{i^*}| - 1}.$$

Thuật toán 1: Tính toán các ma trận Acc và Cov trước khi bổ sung và loại bỏ các đối tượng

Dữ liệu vào:

- Các lớp tương đương điều kiện C_i với $i = 1, \dots, m$
- Các lớp tương đương quyết định D_j với $i = 1, \dots, h$

Dữ liệu ra:

- Ma trận $\text{Acc}(C, D)$ và ma trận $\text{Cov}(C, D)$

// Tính đồng thời cả hai ma trận Acc và Cov

for $i = 1 : m$ **do**

for $j = 1 : h$ **do**

$$\text{Acc}(C_i, D_j) = \frac{|C_i \cap D_j|}{|C_i|};$$

$$\text{Cov}(C_i, D_j) = \frac{|C_i \cap D_j|}{|D_j|}.$$

end

end

Nếu $u \neq i^*$ thì $\text{Acc}(C'_u, D'_{j^*}) = \text{Acc}(C_u, D_{j^*})$ và

$$\text{Cov}(C'_u, D'_{j^*}) = \frac{|C_u \cap D_{j^*}|}{|D_{j^*}| - 1}.$$

Nếu $i \neq i^*$ và $j \neq j^*$ thì $\text{Acc}(C'_i, D'_j) = \text{Acc}(C_i, D_j)$ và $\text{Cov}(C'_i, D'_j) = \text{Cov}(C_i, D_j)$.

Nhận xét: Khi thực hiện thao tác loại bỏ phần tử ra khỏi khối quyết định có thể xảy ra trường hợp một lớp tương đương điều kiện hoặc một lớp tương đương quyết định có thể trở thành rỗng. Khi đó ta có hai trường hợp sau:

- Nếu $C'_i = \emptyset$ ($i = 1, \dots, m$) thì $C_i \cap D_j = \emptyset$. Suy ra $|C_i \cap D_j| = 0$, nghĩa là tất cả các phần tử ở dòng i của ma trận $\text{Acc}(C', D')$ và dòng i của ma trận $\text{Cov}(C', D')$ đều bằng 0.
- Nếu $D'_j = \emptyset$ ($j = 1, \dots, h$) thì $C_i \cap D_j = \emptyset$. Suy ra $|C_i \cap D_j| = 0$, nghĩa là tất cả các phần tử ở cột j của ma trận $\text{Acc}(C', D')$ và cột j của ma trận $\text{Cov}(C', D')$ đều bằng 0.

Do đó, trước khi tiến hành sinh các luật quyết định có ý nghĩa thì ta cần thực hiện việc loại bỏ các dòng/cột của ma trận Acc và ma trận Cov mà có toàn giá trị 0.

3. Các thuật toán tính gia tăng Acc và Cov sau khi bổ sung và loại bỏ các phần tử

Sau đây, phương pháp tính các ma trận Acc và Cov trước khi bổ sung và loại bỏ các đối tượng được trình bày trong Thuật toán 1. Phương pháp tính gia tăng các ma trận Acc và Cov sau khi bổ sung các phần tử được trình bày trong Thuật toán 2. Phương pháp tính gia tăng các ma trận Acc và Cov sau khi loại bỏ các phần tử được trình bày trong Thuật toán 3. Cuối cùng, phương pháp loại bỏ dòng/cột trong các ma trận Acc và Cov mà có toàn giá trị 0 được trình bày trong Thuật toán 4.

Thuật toán 2: Tính toán gia tăng các ma trận Acc và Cov sau khi bổ sung các phần tử

Dữ liệu vào:

- Các lớp tương đương điều kiện C_i với $i = 1, \dots, m$
- Các lớp tương đương quyết định D_j với $j = 1, \dots, h$
- Tập AN chứa N phần tử bổ sung

Dữ liệu ra:

- Ma trận $\text{Acc}(C', D')$ và ma trận $\text{Cov}(C', D')$

```

foreach  $x \in AN$  do
     $p = 0;$ 
    for  $i = 1 : m + p$  do
         $q = 0;$ 
        for  $j = 1 : h + q$  do
            Thực hiện trường hợp (i) của mục III.2.1;
             $p = p + 1$  và  $q = q + 1;$ 
        end
        if  $\forall x \notin C_i$  và  $\exists j^* : x \in D_{j^*}$  then
            Thực hiện trường hợp (ii) của mục III.2.1;
             $p = p + 1;$ 
        end
        if  $\exists i^* : x \in C_{i^*}$  và  $\forall j : x \notin D_j$  then
            Thực hiện trường hợp (iii) của mục III.2.1;
             $q = q + 1;$ 
        end
        if  $\exists i^* : x \in C_{i^*}$  và  $\exists j^* : x \in D_{j^*}$  then
            Thực hiện trường hợp (iv) của mục III.2.1;
        end
    end
end
    
```

Thuật toán 3: Tính toán gia tăng các ma trận Acc và Cov sau khi loại bỏ các phần tử

Dữ liệu vào:

- Các lớp tương đương điều kiện C_i với $i = 1, \dots, m$
- Các lớp tương đương quyết định D_j với $j = 1, \dots, h$
- Tập DM chứa M phần tử bị loại bỏ

Dữ liệu ra:

- Ma trận $\text{Acc}(C', D')$ và ma trận $\text{Cov}(C', D')$

```

foreach  $x \in DM$  do
    for  $i = 1 : m$  do
        for  $j = 1 : h$  do
            Thực hiện trường hợp ở mục III.2.2;
        end
    end
end
    
```

4. Độ phức tạp của các thuật toán tính gia tăng Acc và Cov sau khi bổ sung và loại bỏ các phần tử trên khối quyết định

Mệnh đề 7: Độ phức tạp thuật toán xác định Acc và Cov là $O(|U|^2)$.

Chứng minh: Ta có lực lượng trung bình của lớp tương đương điều kiện và lớp tương đương quyết định tương ứng là $|U|/m, |U|/h$. Vì có tất cả $m \times h$ phần tử, nên để tính độ chính xác và độ phủ của một luật số phép tính cần thực

Thuật toán 4: Loại bỏ dòng/cột trong các ma trận Acc và Cov mà có toàn giá trị 0

Dữ liệu vào:

- Ma trận $\text{Acc}(C', D')$
- Ma trận $\text{Cov}(C', D')$

Dữ liệu ra:

- Ma trận $\text{Acc}(C', D')$ và ma trận $\text{Cov}(C', D')$ sau khi đã loại bỏ dòng/cột toàn giá trị 0

```

// Thực hiện xóa dòng toàn giá trị 0
for  $i = 1 : m + p$  do
     $kt = 0;$ 
    for  $j = 1 : h + q$  do
        if  $\text{Acc}(C'_i, D'_j) \neq 0$  và  $\text{Cov}(C'_i, D'_j) \neq 0$  then
             $kt = 1;$ 
            break;
        end
    end
    if  $kt = 0$  then
        Xóa dòng  $i$ , loại bỏ  $C'_i$ ;
         $p = p - 1;$ 
         $i = i - 1;$ 
    end
end
// Thực hiện xóa cột toàn giá trị 0.
for  $j = 1 : h + q$  do
     $kt = 0;$ 
    for  $i = 1 : m + p$  do
        if  $\text{Acc}(C'_i, D'_j) \neq 0$  và  $\text{Cov}(C'_i, D'_j) \neq 0$  then
             $kt = 1;$ 
            break;
        end
    end
    if  $kt = 0$  then
        Xóa cột  $j$ , loại bỏ  $D'_j$ ;
         $q = q - 1;$ 
         $j = j - 1;$ 
    end
end
    
```

hiện sẽ là

$$\frac{|U|}{m} \times \frac{|U|}{h} \times m \times h = |U|^2.$$

Do đó độ phức tạp thời gian của thuật toán này là $O(|U|^2)$. □

Mệnh đề 8: Độ phức tạp thuật toán tính gia tăng Acc và Cov khi bổ sung N đối tượng là $O(N|U|^2)$.

Chứng minh: Khi bổ sung đối tượng x , để kiểm tra xem x thuộc lớp tương đương điều kiện nào ta cần thực hiện s_m phép so sánh, để kiểm tra x thuộc lớp tương đương quyết định nào ta cần s_h phép so sánh. Do đó, việc kiểm tra phần tử x mới bổ sung cần $s_m + s_h$ phép so sánh (s_m và s_h là số lớp tương đương điều kiện và số lớp tương đương quyết định tương ứng).

Cập nhật các ma trận Acc và Cov khi bổ sung x , ta có bốn trường hợp sau.

(i) Nếu hình thành một lớp điều kiện mới và một lớp quyết định mới thì cả hai ma trận Acc và Cov đều được bổ

sung thêm một dòng mới (kí hiệu là i^*) và một cột mới (kí hiệu là j^*). Như vậy ta có $\text{Acc}(C'_{i^*}, D'_{j^*}) = \text{Cov}(C'_{i^*}, D'_{j^*}) = 1$; $\text{Acc}(C'_{j^*}, D'_{j^*}) = \text{Cov}(C'_{j^*}, D'_{j^*}) = 0$, ($j = 1, \dots, s_h$); $\text{Acc}(C'_{i^*}, D'_{j^*}) = \text{Cov}(C'_{i^*}, D'_{j^*}) = 0$, ($i = 1, \dots, s_m$). Độ chính xác và độ phủ của các luật còn lại không thay đổi. Do vậy, trường hợp này chỉ cần $2(s_m + s_h + 1)$ phép gán.

(ii) Nếu chỉ hình thành lớp điều kiện mới, còn số lớp quyết định không đổi thì hai ma trận Acc và Cov được bổ sung thêm dòng mới i^* và x sẽ làm ảnh hưởng đến cột j^* của hai ma trận này. Như vậy, $x \in C'_{i^*}$ và $x \in D'_{j^*}$, suy ra $C'_{i^*} \cap D'_{j^*} = \{x\}$, còn $C'_u \cap D'_{j^*}$ không đổi nếu $u \neq i^*$, $C'_{i^*} \cap D'_k = \emptyset$ nếu $k \neq j^*$. Từ đó ta có $\text{Acc}(C'_{i^*}, D'_{j^*}) = \text{Cov}(C'_{i^*}, D'_{j^*}) = 1$. Nếu $u \neq i^*$ thì $\text{Acc}(C'_u, D'_{j^*})$ không đổi và

$$\text{Cov}(C'_u, D'_{j^*}) = \frac{|C'_u \cap D'_{j^*}|}{|D'_{j^*}| + 1}. \quad (1)$$

Nếu $k \neq j^*$ thì $\text{Acc}(C'_{i^*}, D'_k) = \text{Cov}(C'_{i^*}, D'_k) = 0$. Độ chính xác và độ phủ của các luật còn lại không thay đổi. Do đó, độ phức tạp trong trường hợp này phụ thuộc vào việc tính $\text{Cov}(C'_u, D'_{j^*})$. Khi đó, để xác định $\text{Cov}(C'_u, D'_{j^*})$ ta cần $s_m - 1$ phép cộng để tính tổng giá trị các phần tử của cột j^* , s_m phép tính $\text{Sup}(C'_u, D'_{j^*})$. Với mỗi phép tính $\text{Sup}(C'_u, D'_{j^*})$ cần $(|U|/s_m) \times (|U|/s_k)$ phép tính và s_m phép chia. Từ đó suy ra số phép tính cần thực hiện là

$$s_m - 1 + s_m \frac{|U|}{s_m} \frac{|U|}{s_k} + s_m = 2s_m + \frac{|U|^2}{s_k} - 1.$$

(iii) Nếu chỉ hình thành lớp quyết định mới, còn số lớp điều kiện không đổi thì tương tự như trường hợp (ii) thì ta chỉ cần thay đổi cột thành dòng: j^* thành i^* , tổng cột thành tổng dòng, Cov thành Acc. Như vậy, số phép tính trong trường hợp này là $2s_m + (|U|^2/s_m) - 1$.

(iv) Nếu không hình thành lớp điều kiện mới, cũng không hình thành lớp quyết định mới thì việc thực hiện trong trường hợp này tương tự với việc thực hiện đồng thời cả hai trường hợp 2 và 3. Từ đó số phép tính cần thực hiện sẽ là

$$2(s_m + s_h) + \frac{|U|^2}{s_m} + \frac{|U|^2}{s_k} - 2.$$

Do đó, với mỗi phần tử x được bổ sung thì trường hợp xấu nhất ta phải thực hiện số phép tính là

$$2(s_m + s_h) \left(s_m + s_h + \frac{|U|^2}{s_m} + \frac{|U|^2}{s_k} - 2 \right).$$

Như vậy, số phép tính cần thực hiện khi bổ sung N phần tử sẽ là

$$2N(s_m + s_h) \left(s_m + s_h + \frac{|U|^2}{s_m} + \frac{|U|^2}{s_k} - 2 \right).$$

Trong quá trình tính toán các ma trận Acc và Cov thì s_m dần tới $m+p$, s_h dần tới $h+q$, do đó $s_m \leq m+p$, $s_h \leq h+q$ nên ta có thể coi $s_m = m+p$, $s_h = h+q$. Từ đó suy ra số

phép tính cần thực hiện để cập nhật cho các ma trận Acc và Cov khi bổ sung N phần tử là

$$N(m+p+h+q) \left(2(m+p+h+q) + \frac{|U|^2}{m+p} + \frac{|U|^2}{h+q} - 2 \right).$$

Mặt khác, vì $m, h \leq |U|$, suy ra độ phức tạp của thuật toán là $O(N|U|^2)$. \square

Mệnh đề 9: Độ phức tạp thuật toán tính gia tăng Acc và Cov khi loại bỏ M đối tượng là $O(M|U|^2)$.

Chứng minh: Khi loại bỏ đối tượng x , để kiểm tra xem x thuộc lớp tương đương điều kiện nào ta cần thực hiện s_m phép so sánh, để kiểm tra x thuộc lớp tương đương quyết định nào ta cần s_h phép so sánh. Do đó, việc kiểm tra phần tử x để loại bỏ cần $s_m + s_h$ phép so sánh (s_m và s_h là số lớp tương đương điều kiện và số lớp tương đương quyết định tương ứng).

Cập nhật các ma trận Acc và Cov khi loại bỏ x . Khi loại bỏ x sẽ làm thay đổi đến dòng i^* và đến cột j^* của các ma trận Acc và Cov, do đó thực hiện này giống như việc thực hiện trường hợp (iv) trong chứng minh của mệnh đề 8. Từ đó suy ra số phép tính cần thực hiện khi xóa đi M đối tượng trên khối quyết định sẽ là

$$M(s_m + s_h) \left(2(s_m + s_h) + \frac{|U|^2}{s_m} + \frac{|U|^2}{s_k} - 2 \right).$$

Tương tự như chứng minh ở mệnh đề 8, ta xem $s_m = m+p$ và $s_h = h+q$, khi đó số phép tính cần thực hiện để cập nhật các ma trận Acc và Cov khi loại bỏ M đối tượng là

$$M(m+p+h+q) \left(2(m+p+h+q) + \frac{|U|^2}{m+p} + \frac{|U|^2}{h+q} - 2 \right).$$

Do $m, h \leq |U|$, suy ra độ phức tạp của thuật toán là $O(M|U|^2)$. \square

Mệnh đề 10: Độ phức tạp thuật toán xóa dòng/cột của ma trận Acc và Cov có toàn giá trị 0 là $O(|U|^2)$.

Chứng minh: Với mỗi dòng của ma trận Acc và Cov có toàn giá trị 0 thì ta cần thực hiện $h+q$ phép kiểm tra và $h+q$ phép xóa từng phần tử của dòng này. Tiếp đó thực hiện một phép xóa C_i ra khỏi U/C và hai phép gán. Như vậy ta cần thực hiện $2(h+q)+3$ phép tính để xóa một dòng. Trong trường hợp xấu nhất ta cần thực hiện $(m+p)(2(h+q)+3)$ phép tính để xóa hết tất cả $m+p$ dòng của các ma trận Acc và Cov.

Với mỗi cột có tất cả các phần tử bằng 0 trong các ma trận Acc và Cov thì tương tự như trên trong trường hợp xấu nhất ta cần thực hiện $(h+q)(2(m+p)+3)$ phép tính để xóa hết tất cả $h+q$ cột của các ma trận Acc và Cov. Do $m, h \leq |U|$, suy ra độ phức tạp của thuật toán là $O(|U|^2)$. \square

IV. KẾT LUẬN

Từ những thay đổi do việc bổ sung và loại bỏ các phần tử trên khối quyết định, bài báo đã phát biểu và chứng minh một số tính chất của các ma trận Acc và Cov, đề xuất một số thuật toán tính gia tăng của các ma trận độ chính xác và ma trận độ phủ trên khối quyết định và trên lát cắt. Từ các thuật toán đề xuất, độ phức tạp của chúng cũng đã được phát biểu và chứng minh. Những kết quả nói trên là cơ sở để giúp tính gia tăng ma trận độ chính xác và ma trận độ phủ trên khối và trên lát cắt, từ các kết quả đó tìm ra các luật quyết định có ý nghĩa trên khối và trên lát cắt khi tập đối tượng trên khối quyết định có thay đổi. Các kết quả này cũng góp phần làm phong phú thêm ứng dụng của lý thuyết thiết kế mô hình cơ sở dữ liệu dạng khối.

TÀI LIỆU THAM KHẢO

- [1] T. D. Thắng and D. T. L. Anh, "Một số thuật toán xác định ma trận độ hỗ trợ trên khối dữ liệu có giá trị thuộc tính thay đổi," in *Kỷ yếu Hội thảo quốc gia lần thứ XXI: Một số vấn đề chọn lọc của Công nghệ Thông tin và Truyền thông*, 2018, pp. 216–225.
- [2] C. C. Nghĩa, "Nghiên cứu các phương pháp rút gọn thuộc tính và sinh luật quyết định theo tiếp cận tập thô mờ," Ph.D. dissertation, Học viện Công nghệ Bưu chính Viễn thông, 2017.
- [3] D. Liu, T. Li, D. Ruan, and W. Zou, "An incremental approach for inducing knowledge from dynamic information systems," *Fundamenta Informaticae*, vol. 94, no. 2, pp. 245–260, 2009.
- [4] K. Shi and B. Yao, "Function s-rough sets and law identification," *Science in China Series F: Information Sciences*, vol. 51, no. 5, p. 499, 2008.
- [5] V. Putrenko, N. Pashvnska, and S. Nazarenko, "Data mining of network events with space-time cube application," in *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, 2018, pp. 79–83.
- [6] T. D. Thắng, T. M. Tuyền, and D. T. L. Anh, "Khai phá luật quyết định trên khối dữ liệu có giá trị thuộc tính thay đổi," in *Kỷ yếu Hội thảo quốc gia lần thứ XIX: Một số vấn đề chọn lọc của Công nghệ Thông tin và Truyền thông*, 2016, pp. 163–169.
- [7] T. D. Thắng, T. M. Tuyền, D. T. L. Anh, and N. T. Quyên, "Một số kết quả về khai phá luật quyết định trên khối dữ liệu có giá trị thuộc tính thay đổi," in *Kỷ yếu Hội nghị Khoa học công nghệ quốc gia lần thứ X (FAIR)*, 2017, pp. 623–632.
- [8] T. D. Thang and T. M. Tuyen, "Key and key attributes set, non-key attributes set with translation of block schemes," *International Journal of Advanced Research in Computer Science*, vol. 3, no. 3, pp. 335–339, 2012.



Đỗ Thị Lan Anh sinh năm 1988 tại Hà Nam. Tốt nghiệp Cử nhân Tin học tại Đại học Sư phạm Hà Nội 2 năm 2006. Tốt nghiệp Thạc sĩ Khoa học máy tính tại Đại học Sư phạm Hà Nội 2 năm 2013. Hiện tại công tác tại Viện Công nghệ Thông tin, Trường Đại học Sư phạm Hà Nội 2 và là nghiên cứu sinh tiến sĩ năm thứ tư tại Học viện Khoa học Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam. Lĩnh vực nghiên cứu bao gồm khai phá dữ liệu và cơ sở dữ liệu.



Trịnh Đình Thắng sinh năm 1955 tại Hà Nội. Tốt nghiệp Cử nhân Toán tại Đại học Sư phạm Hà Nội 1 năm 1977. Tốt nghiệp Thạc sĩ Toán học tại Đại học Sư phạm Hà Nội 1 năm 1986. Tốt nghiệp Tiến sĩ Công nghệ thông tin tại Viện Công nghệ Thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam. Hiện công tác tại Viện Công nghệ Thông tin, Trường Đại học Sư phạm Hà Nội 2. Lĩnh vực nghiên cứu bao gồm cơ sở dữ liệu, khai phá dữ liệu, mạng máy tính.