

MỞ RỘNG PHỤ THUỘC HÀM TRONG CƠ SỞ DỮ LIỆU BỊ NHIỄU

Hoàng Thị Lan Giao

Khoa Công nghệ Thông tin, Trường Đại học Khoa học – Đại học Huế

Email: hlgiao.it@gmail.com

TÓM TẮT

Trong một cơ sở dữ liệu quan hệ với phụ thuộc hàm cho trước $A \rightarrow b$ giá trị của thuộc tính b được xác định duy nhất bởi tập thuộc tính A . Tuy nhiên, qua quá trình truyền dữ liệu, dữ liệu nhận được đã bị nhiễu và có thể không còn đúng như nguyên bản. Rõ ràng, trên bộ dữ liệu mới này phụ thuộc hàm $A \rightarrow b$ không còn đúng nữa. Nói cách khác, nếu chỉ dựa vào dữ liệu trên tập thuộc tính A chúng ta không thể xác định được chính xác các đối tượng. Trong bài báo này, chúng tôi sẽ nghiên cứu việc mở rộng tập thuộc tính A thành một tập cực tiểu A' chứa A sao cho $A' \rightarrow b$ là phụ thuộc hàm đúng trên cơ sở dữ liệu mới.

Từ khóa: cơ sở dữ liệu quan hệ, phụ thuộc hàm, phụ thuộc hàm bị lỗi.

1. MỞ ĐẦU

Giả sử ta có một cơ sở dữ liệu thực gồm m đối tượng và Ω là tập n thuộc tính, được biểu diễn bằng một ma trận $M_{m \times n}$, khi đó mỗi dòng tương ứng là một bản ghi và không có hai dòng giống nhau. Gọi K là họ các khóa cực tiểu. Những dữ liệu này được chuyển dịch thông qua một kênh có lỗi, ta ký hiệu M^* là ma trận nhận được sau khi dịch chuyển. M và M^* khác nhau ít nhất là e giá trị trên mỗi dòng. Trong M không có hai dòng giống nhau nhưng trong M^* điều này không chắc chắn. Chẳng hạn cặp thuộc tính (ten, ho) là một khóa trong cơ sở dữ liệu M và ta có hai dòng tương ứng với cặp này là (Trần, Chi) và (Trần, Nhi). Tuy nhiên, khi chuyển dịch ta nhận được trong M^* hai cặp giá trị này đều là (Trần, Nhi). Khi đó đối tượng cần xác định trong M^* không duy nhất.

Có nhiều cách tiếp cận khác nhau, chẳng hạn thuật toán Tane [7] mở rộng các phụ thuộc hàm bằng cách đưa ra một sai số chấp nhận được dựa vào tỷ lệ giữa các bộ (đối tượng) đúng với phụ thuộc hàm đó và tổng số bộ trong cơ sở dữ liệu. Ý tưởng này cũng được chúng tôi mở rộng trên cả phụ thuộc đa trị xấp xỉ theo tiếp cận tập thô bằng cách xây dựng ma trận phụ thuộc [5,6].

Mục đích của nghiên cứu này theo hướng tiếp cận khác, tìm cách bổ sung thuộc tính vào khóa để chúng ta có thể xác định một cách duy nhất đối tượng cần tìm trong M^* . Vấn đề đặt ra là biết cấu trúc của M và chỉ nhận dữ liệu là M^* , chúng ta cần đưa ra kết luận dựa vào những

thông tin này. Bài toán này sẽ ứng dụng trong Khai phá dữ liệu, đặc biệt là trong Hỗ trợ quyết định, khi thuộc tính quyết định được xem như vế phải của phụ thuộc hàm.

Giả sử $A \rightarrow b$ ($A \subset \Omega$, $b \in \Omega$) đúng trong M nhưng không tất yếu $A \rightarrow a$ đúng trong M^* do dữ liệu bị biến dạng. Liệu chúng ta có thể bổ sung các thuộc tính vào A thành A' (trong M^*) để $A' \rightarrow b$, nếu có thì nên mở rộng như thế nào? Chẳng hạn, nếu số lỗi trên một dòng ít nhất là 1 ($e = 1$) và *gioi tinh* là một thuộc tính, khi đó hoặc *ten*, hoặc *gioi tinh* có thể đúng và như vậy bộ (Trần, Nhi, . . .) trong M có thể tìm thấy hai dòng khác nhau trong M^* :

(Trần, Nhi, Nữ, . . .)

(Trần, Nhi, Nam, . . .)

Như vậy nếu chúng ta bổ sung vào tập khóa thêm thuộc tính *gioitinh* thì chúng ta sẽ xác định được bộ dữ liệu chúng ta cần tìm kiếm.

2. XỬ LÝ PHỤ THUỘC HÀM BỊ NHIỄU

2.1. Khoảng cách Hamming

Cho M là ma trận m dòng và n cột. Mỗi dòng của M là một bộ n phần tử ($x_1^i, x_2^i, \dots, x_n^i$). Khoảng cách Hamming giữa hai dòng ($x_1^i, x_2^i, \dots, x_n^i$) và ($x_1^j, x_2^j, \dots, x_n^j$) được định nghĩa là

$$H(x^i, x^j) = \sum_{k=1}^n \text{sign}(x_k^i, x_k^j) \quad (1)$$

$$\text{trong đó hàm dấu sign được tính: } \text{sign}(a, b) = \begin{cases} 1 & \text{nếu } a \neq b \\ 0 & \text{nếu } a = b \end{cases} \quad (2)$$

Ví dụ 1. Cho bảng quyết định về thông tin 6 bệnh nhân bị cảm cúm dưới đây:

X	Đau đầu	Đau cơ	Thân nhiệt	Cảm cúm
X^1	Có	Có	Bình thường	Không
X^2	Có	Có	Cao	Có
X^3	Có	Có	Rất cao	Có
X^4	Không	Có	Bình thường	Không
X^5	Không	Không	Cao	Không
X^6	Không	Có	Rất cao	Có

ma trận M lưu trữ thông tin này gồm 6 dòng, 4 cột và khoảng cách $H(X^1, X^2) = 2$; $H(X^1, X^4) = 1$.

$$M = \begin{pmatrix} \text{Có} & \text{Có} & \text{Bình thường} & \text{Không} \\ \text{Có} & \text{Có} & \text{Cao} & \text{Có} \\ \text{Có} & \text{Có} & \text{Rất cao} & \text{Có} \\ \text{Không} & \text{Có} & \text{Bình thường} & \text{Không} \\ \text{Không} & \text{Không} & \text{Cao} & \text{Không} \\ \text{Không} & \text{Có} & \text{Rất cao} & \text{Có} \end{pmatrix} \quad (3)$$

2.2. Phụ thuộc hàm có lỗi

Cho A, B là tập con các thuộc tính của Ω , những giá trị trong những cột của A xác định những giá trị trong những cột của B duy nhất với ít nhất e lỗi xảy ra trong mỗi dòng (thực ra nó xác định trong M nhưng do dữ liệu trong M^* có thể sai). Khi đó ta ký hiệu $A \rightarrow \{e\}B$. Nói một cách khác, với một dòng bất kỳ r trong M^* , nếu có hai dòng s, t trong M , hai dòng này có khoảng cách Hamming với r không vượt quá e trong những cột của A , thì chúng sẽ bằng nhau trong những cột của B . Phụ thuộc hàm $A \rightarrow \{e\}B$ được gọi là phụ thuộc hàm e -lỗi. Ký hiệu $M(A)$ là ma trận dữ liệu tương ứng với phép chiếu M trên tập thuộc tính A . $M(A)$ là ma trận có kích thước $m \times |A|$.

Ví dụ 2. Trong Ví dụ 1, giả sử ma trận M^* nhận được là:

$$M^* = \begin{pmatrix} \text{Không} & \text{Có} & \text{Bình thường} & \text{Không} \\ \text{Có} & \text{Không} & \text{Cao} & \text{Có} \\ \text{Có} & \text{Có} & \text{Cao} & \text{Có} \\ \text{Không} & \text{Không} & \text{Cao} & \text{Không} \\ \text{Không} & \text{Không} & \text{Cao} & \text{Có} \\ \text{Không} & \text{Có} & \text{Rất cao} & \text{Có} \end{pmatrix} \quad (4)$$

lấy $A = \{\text{Đau đầu, Thân nhiệt}\}$; $B = \{\text{Cảm cúm}\}$, khi đó $A \rightarrow B$ trong M và $A \rightarrow \{1\}B$ trong M^*

Chúng ta dễ dàng nhận được kết quả sau:

Mệnh đề 1. Cho $A \subset \Omega, a \in \Omega$. Khi đó $A \rightarrow \{e\}a$ nếu và chỉ nếu khoảng cách Hamming của mỗi cặp trong $M(A)$, mà khác nhau tại a , ít nhất là $2e + 1$.

Rõ ràng nếu khoảng cách Hamming của từng cặp dòng trong $M(A)$, khác nhau tại a ít nhất là $2e$ thì tri thức nhận được trong $M^*(A)$ sẽ phát hiện lỗi (tức là có mặt lỗi trong M^*), nhưng không xác định dữ liệu trong a duy nhất. Có nghĩa là có thể có nhiều hơn một dòng của M có giá trị giống nhau trong $M^*(A)$. Từ đây chúng ta đưa ra định nghĩa tổng quát về phụ thuộc hàm d -khoảng cách.

Định nghĩa 1. Cho $A \subset \Omega, a \in \Omega$. Khi đó $A \rightarrow (d)a$ được gọi là phụ thuộc hàm d -khoảng cách nếu và chỉ nếu khoảng cách Hamming của mỗi cặp trong $M(A)$, mà khác nhau trong a , ít nhất là d .

Mục đích chính của nghiên cứu này là tìm mối liên hệ giữa phụ thuộc hàm và phụ thuộc hàm d -khoảng cách. Gọi F_a là họ các tập con cực tiểu F của Ω thỏa $F \rightarrow a$ (trong M).

Mệnh đề 2. Cho $A \subset \Omega, a \in \Omega$. Khi đó $A \rightarrow (d)a$ nếu và chỉ nếu với $d-1$ thuộc tính bất kỳ $a_1, a_2, \dots, a_{(d-1)}$ trong A , tồn tại $F \in F_a$ sao cho $F \subset A \setminus \{a_1, a_2, \dots, a_{(d-1)}\}$.

Chứng minh

- Điều kiện cần (" $\Rightarrow$ ")

Giả sử ngược lại, tồn tại một bộ $(d-1)$ thuộc tính $a_1, a_2, \dots, a_{(d-1)}$ trong A sao cho $A \setminus \{a_1, a_2, \dots, a_{(d-1)}\}$ không chứa phần tử nào của F_a , hay nói cách khác, không có phụ thuộc hàm $A \setminus \{a_1, a_2, \dots, a_{(d-1)}\} \rightarrow a$. Như vậy, tồn tại hai dòng của M bằng nhau trong $A \setminus \{a_1, a_2, \dots, a_{(d-1)}\}$

nhưng khác nhau trong a . Khi đó khoảng cách Hamming của hai dòng này trong $M(A)$ nhiều nhất là $d-1$ ($< d$). Điều này mâu thuẫn với giả thiết.

- Điều kiện đủ (“ $\Leftarrow$ ”)

Giả sử trong $M(A)$ chứa hai dòng có khoảng cách Hamming nhỏ hơn d và hai dòng này khác nhau trong a . Nếu chúng ta xóa đi những cột mà tại đó hai dòng này khác nhau. Khi đó $M(A \setminus \{a_1, a_2, \dots, a_{(d-1)}\})$ chứa hai dòng bằng nhau nhưng lại khác nhau trong a . Do đó $A \setminus \{a_1, a_2, \dots, a_{(d-1)}\} \rightarrow a$ không đúng trong M hay $A \setminus \{a_1, a_2, \dots, a_{(d-1)}\}$ không thể chứa phần tử nào của F_a .

Bây giờ chúng ta sẽ đưa ra sự mô tả tương đương hệ phụ thuộc hàm Armstrong bằng bao đóng [2]. Đặt $L(A) = \{a: a \in \Omega, A \rightarrow a\}$ ($A \subset \Omega$). Rõ ràng bao đóng này thỏa ba tính chất:

- i) $A \subset L(A)$,
- ii) $A \subset B \Rightarrow L(A) \subset L(B)$,
- iii) $L(L(A)) = L(A)$.

Thật vậy,

- i) $\forall a \in A, A \rightarrow a \Rightarrow A \subset L(A)$,
- ii) $\forall a \in L(A) \Rightarrow A \rightarrow a, A \subset B \Rightarrow B \rightarrow a \Rightarrow a \in L(B) \Rightarrow L(A) \subset L(B)$,
- iii) $A \subset L(A) \Rightarrow L(A) \subset L(L(A))$. Đặt $B = L(A) \Rightarrow \forall b \in B, A \rightarrow b$, do đó $A \rightarrow B$.

Hơn nữa, $\forall a \in L(L(A)) = L(B), B \rightarrow a$. Theo trên, ta có $A \rightarrow B$, như vậy $A \rightarrow a$ hay $a \in L(A)$.

Như vậy F_a gồm $\{a\}$ và họ bao hàm tự do F các tập con của $\Omega \setminus \{a\}$ (hệ bao hàm tự do của F có nghĩa là $F_{-1}, F_{-2} \in F, F_{-1} \neq F_{-2} \Rightarrow F_{-1} \not\subset F_{-2}$).

Bổ đề 1. [1,2] Cho họ bao hàm tự do F các tập con của $\Omega \setminus \{a\}$. Khi đó tồn tại một hệ phụ thuộc hàm (và vì vậy, tương ứng ma trận quan hệ M) sao cho nó xác định $F_a = F \cup \{a\}$ đối với $a \in \Omega$ nào đó.

Chứng minh. Cố định $a \in \Omega$ và định nghĩa $F_a = F \cup \{a\}$, đặt

$$L(A) = \begin{cases} A \cup \{a\} & \text{nếu } \exists F \in F_a \text{ và } F \subset A \\ A & \text{trong trường hợp ngược lại} \end{cases} \quad (5)$$

Dễ thấy $L(A)$ thỏa mãn các điều kiện i), ii), iii) của bao đóng.

Bổ đề trên khẳng định với một họ bao hàm thức tự do bất kỳ trên tập $n-1$ phần tử, tồn tại một cơ sở dữ liệu mà họ các tập cực tiểu $F \rightarrow a$ thỏa bằng đúng $F_a = F \cup \{a\}$.

Định nghĩa 2. Cho X là một tập n phần tử và F là một họ bao hàm tự do các tập con của nó, ký hiệu $F(d)$ là tập được xác định bởi:

$F(d) = \{ G \subset X \mid \forall x_1, x_2, \dots, x_{(d-1)} \in G, \exists F \in F, F \subset G \setminus \{x_1, x_2, \dots, x_{(d-1)}\} \text{ và } G \text{ là cực tiểu thỏa mãn tính chất này} \}$.

Chú ý rằng, $F(1) = F$. Đặc biệt trong phần sau chúng ta sẽ thấy đối với một họ bao hàm tự do của những tập F bao gồm về trái của những phụ thuộc hàm cực tiểu của một quan hệ thì $F(d)$ sẽ là về trái của những phụ thuộc hàm d -khoảng cách cực tiểu. Hơn nữa, $F(d)$ có thể rỗng trong khi F khác rỗng. Cố định $a \in X$ và k là số nguyên lớn hơn 2. Chúng ta xây dựng F là họ k phần tử của X chứa a . Thế thì với bất kỳ tập $G \subset X$, $G \setminus \{a\}$ không thể chứa bất kỳ phần tử của F và vì vậy $F(d)$ là rỗng với mọi $d \leq 2$. Mặt khác, nếu F gồm tất cả các tập con k phần tử của X thì tất cả tập $G \subset X$ với $(k+d-1)$ phần tử tạo thành $F(d)$. Chúng ta nói rằng họ F có thể được gắn bởi p phần tử nếu có $x_1, x_2, \dots, x_p$ sao cho mọi phần tử F của F , $F \cap \{x_1, x_2, \dots, x_p\} \neq \emptyset$. Nếu F có thể gắn bởi $(d-1)$ phần tử thì $F(d)$ là rỗng. Ngược lại $F(d)$ khác rỗng do X thỏa điều kiện đầu của phần tử trong $F(d)$ ($\forall x_1, x_2, \dots, x_{d-1}, X \cap \{x_1, x_2, \dots, x_{d-1}\} = \emptyset \Rightarrow X \subset X \setminus \{x_1, x_2, \dots, x_{d-1}\}$) và nếu chưa cực tiểu, chúng ta có thể rút gọn cho đến khi nhận được tập cực tiểu.

Định lý 1. [3] Cho $n \geq n_0(k, d)$ và F là một họ bao hàm tự do của những tập con có kích thước tối đa k của tập cho trước có n phần tử. Khi đó kích thước của những phần tử trong $F(d)$ tối đa là $c_1 k$. Mặt khác, tồn tại F như vậy mà tất cả các phần tử của nó có kích thước ít nhất là $c_2 k$, ở đây c_1, c_2 chỉ phụ thuộc vào d .

[3] đưa ra kết quả này để giới hạn việc mở rộng về trái các phụ thuộc hàm có lỗi là những phần tử của $F(d)$. Chúng tôi chứng minh định lý này để khẳng định tính đúng đắn của nó. Trước hết chúng ta sẽ chứng minh mệnh đề sau

Mệnh đề 3. $G \in F(d)$ nếu và chỉ nếu $\{F \mid F \in F, F \subset G\}$ không thể gắn bởi $d-1$ phần tử nhưng $\{F \mid F \in F, F \subset G \setminus \{a\}\}$ có thể gắn bởi $d-1$ phần tử đối với mỗi $a \in G$.

Chứng minh. Cho $G \in F(d) \Rightarrow \forall a_1, a_2, \dots, a_{d-1}, \exists F \in F$ sao cho $F \subset G \setminus \{a_1, a_2, \dots, a_{d-1}\}$ suy ra $F \cap \{a_1, a_2, \dots, a_{d-1}\} = \emptyset$ và $F \subset G$ do đó $\{F \mid F \in F, F \subset G\}$ không thể gắn bởi $d-1$ phần tử. Mặt khác, $G \in F(d)$ là tập cực tiểu thỏa tính chất: với mọi $a_1, a_2, \dots, a_{d-1}$, tồn tại $F \in F$ sao cho $F \subset G \setminus \{a_1, a_2, \dots, a_{d-1}\}$, vì vậy $\forall a \in G, G \setminus \{a\} \notin F(d) \Rightarrow$ tồn tại $a_1, a_2, \dots, a_{d-1}$, mà với mọi $F \in F, (F \subset G \setminus \{a\}), F \cap \{a_1, a_2, \dots, a_{d-1}\} \neq \emptyset \Rightarrow \{F \mid F \in F, F \subset G \setminus \{a\}\}$ có thể được gắn bởi $d-1$ phần tử.

Ngược lại, họ $F^* = \{F \mid F \subset G\}$ không thể gắn bởi $d-1$ phần tử suy ra tồn tại $a_1, a_2, \dots, a_{d-1}, \exists F \in F^*$ sao cho $F \cap \{a_1, a_2, \dots, a_{d-1}\} = \emptyset$. Mặt khác, $F \in F^*$, nên $F \subset G$, do đó $F \subset G \setminus \{a_1, a_2, \dots, a_{d-1}\}$. Xét tập con thực sự của G : $G' = G \setminus \{a\}$ ($a \in G$). Khi đó $F^* = \{F \mid F \subset G'\}$ có thể gắn bởi $d-1$ phần tử, vì vậy tồn tại $a_1, a_2, \dots, a_{d-1}$, sao cho $\forall F \in F^*, F \subset G' (= G \setminus \{a\}), F \cap \{a_1, a_2, \dots, a_{d-1}\} \neq \emptyset$. Khi đó, $F \not\subset G' \setminus \{a_1, a_2, \dots, a_{d-1}\}$, như vậy $G' \notin F(d)$, do đó G là tập cực tiểu thỏa $\forall a_1, a_2, \dots, a_{d-1}, \exists F, F \subset G \setminus \{a_1, a_2, \dots, a_{d-1}\}$. Hay nói cách khác $G \in F(d)$.

Bây giờ chúng ta sẽ chứng minh Định lý 1

Trước hết, chúng ta chứng minh giới hạn dưới: Cho họ bao hàm thức tự do F gồm các tập k phần tử ($k \geq 2$) sinh ra $F(d)$ chứa một phần tử có kích thước ít nhất là $c_2 k^d$. Cố định số nguyên $i: 1 \leq i \leq k$ và lấy tập con $A \subset X$ có kích thước $i+d-1$. Đặt $B_1, B_2, \dots, B_{C_i^{i+d-1}}$ là những

tập con i phần tử của A ; và $E_1, E_2, \dots$, là những tập con rời nhau của $X \setminus A$ với $|E_1| = |E_2| = \dots = k-i$. Điều này có thể thực hiện được nếu tổng các phần tử của những tập E_j không vượt quá số phần tử của $X \setminus A$. Tức là, bất đẳng thức sau phải đúng:

$$i + d - 1 + C_i^{i+d-1} \times (k - i) \leq n \quad (6)$$

Tiếp đến, đặt G^i là họ $\{B_1 \cup E_1, B_2 \cup E_2, \dots\}$, chúng ta sẽ chứng minh

$$G^i(d) = \bigcup_j E_j$$

Đặt $D = A \cup_j E_j$. Theo Mệnh đề 3, trước tiên ta chứng minh $\{F \in G^i \mid F \subset D\}$ không thể gắn bởi $d-1$ phần tử. Thật vậy, giả sử ngược lại, tức là, $\forall x_1, x_2, \dots, x_{d-1}, \exists F = B_j \cup E_j$ sao cho $F \cap \{x_1, x_2, \dots, x_{d-1}\} = \emptyset$. Giả sử tập $\{x_1, x_2, \dots, x_{d-1}\}$ có r phần tử thuộc A và $d-1-r$ phần tử thuộc phần bù của nó. Vì vậy, có tối đa $d-1-r$ tập E_j chứa các phần tử của $\{x_1, x_2, \dots, x_{d-1}\}$ (Mỗi E_j chứa một phần tử vì các E_j rời nhau). Chọn $B_j \subset A \setminus \{x_1, x_2, \dots, x_{d-1}\}$ có i phần tử. Khi đó số tập B_j được chọn lớn hơn $d-1-r$. Thật vậy, số tập B_j là

$$C_i^{i+d-1} = \frac{(d+i-1-r) \times (d+i-1-(r+1)) \times \dots \times (d-r)}{i \times ((i-1) \times \dots \times 1)} \geq \frac{d+i-1-r}{i} > d-1-r \geq 0$$

Điều này có nghĩa là, tồn tại tập $F = B_j \cup E_j$ sao cho $F \cap \{x_1, x_2, \dots, x_{d-1}\} = \emptyset$ (đpcm).

Phần còn lại, cần chứng minh $\{F \in G^i \mid F \subset D \setminus \{a\}\}$ có thể gắn bởi $d-1$ phần tử ($a \in D$). Xét hai trường hợp:

1. $a \in E_j$, chọn $\{x_1, x_2, \dots, x_{d-1}\} = A \setminus B_j$. Khi đó, $\forall F \subset D \setminus \{a\}$ và $F \in G^i$, ta có $F = B_l \cup E_l$. Do đó $l \neq j$ và $B_l \cap (A \setminus B_j) \neq \emptyset$
2. $a \in A$, xét Tập H có $d-1$ phần tử và $H \subset A \setminus \{a\}$, $\forall F \subset D \setminus \{a\}$, ta có $F = B_l \cup E_l$ với $B_l \subset A \setminus \{a\}$. Vì $|B_l| = i$ và $a \notin B_l$ suy ra $B_l \cap H \neq \emptyset$. Vậy $D \in G^i(d)$. Chọn $i = \left\lfloor k(1 - \frac{1}{d}) \right\rfloor$. Khi đó ta có các bất đẳng thức:

$$i \geq \frac{k(d-1)}{d} \leq i+1; \frac{k}{d} \leq k-i \quad (7)$$

Mặt khác, kích thước của D được xác định bởi vế trái của biểu thức (6):

$$\begin{aligned} |D| &= |A| + (k-i) \times C_i^{i+d-1} = i + d - 1 + (k-i) \times \frac{(i+1) \times \dots \times (i+d-1)}{(d-1)!} \\ &> (k-i) \times \frac{(i+1)^{d-1}}{(d-1)!} > \frac{k}{d} \times (k \times (1 - \frac{1}{d}))^{d-1} \times \frac{1}{(d-1)!} > k^d \times \frac{(d-1)^{d-1}}{d^d} \times \frac{1}{(d-1)!} = c_2 k^d \end{aligned}$$

Bây giờ chúng ta sẽ chứng minh giới hạn trên. Cho $G \in \mathcal{F}(d)$ với $\mathcal{F}$ là họ các tập con của X có số phần tử $\leq k$, chúng ta sẽ chỉ ra rằng $|G| \leq dk^d$. Có thể giả thiết mọi phần tử của $\mathcal{F}$ đều con của G . Từ Mệnh đề 3, suy ra rằng các tập con D có d phần tử của G gắn các phần tử của họ $\mathcal{F}$. Cụ thể, mỗi phần tử a của G có thể mở rộng thành tập D . Hơn nữa, Mệnh đề 3 còn chỉ ra rằng, hợp của những tập này bằng đúng G . Chúng ta ký hiệu họ các tập D là $\mathcal{D}$. Khi đó:

$$\bigcup_{\{D \in \mathcal{D}\}} D = G \quad (8)$$

$$\text{và} \quad D \cap F \neq \emptyset, \forall D \in \mathcal{D}, F \in \mathcal{F} \quad (9)$$

F không thể gắn bởi một tập ít hơn d phần tử. Cho $I \subset G$, I - bậc của D là số phần tử của D chứa I , ký hiệu $\deg_I(D) = |\{D \in \mathcal{D} : I \subset D\}|$.

Bổ đề 2. Nếu $|I| < d$ thì $\deg_I(D) \leq k^{d-|I|}$.

Bổ đề này sẽ được chứng minh bằng quy nạp theo $j = d - |I|$.

Với $j = d - |I| = 1 \Rightarrow |I| = d-1$, nếu tất cả các phần tử của F được gắn bởi $d-1$ phần tử, điều này vô lý. Vì vậy, tồn tại $F \in \mathcal{F} : F \cap I = \emptyset$. Từ (8), tập D thỏa mãn $I \subset D$ thì $D \cap F \neq \emptyset$. Do đó, số các tập D này bé hơn hoặc bằng số phần tử của F , hơn nữa $|F| \leq k = k^{d-|I|}$.

Giả sử mệnh đề đúng với $j = d - |I| \geq 1$, chúng ta chứng minh mệnh đề đúng với $j + 1 = d - |I|$. Đặt $|I^*| = d - j - 1$, khi đó tồn tại $F \in \mathcal{F} : F \cap I^* = \emptyset$, vì ngược lại thì F được gắn bởi ít hơn d phần tử, vô lý. Đặt $F = \{x_1, x_2, \dots, x_l\}$, với $l \leq k$. Theo (8), chúng ta có:

$$\{D \in \mathcal{D} : I^* \subset D\} = \bigcup_{i=1}^l \{D \in \mathcal{D} : I^* \cup \{x_i\} \subset D\}$$

Kích thước của tập ở vế phải là $\deg_{I^* \cup \{x_i\}}(D)$, theo giả thiết quy nạp với $j = d - |I^* \cup \{x_i\}| = \deg_{I^* \cup \{x_i\}}(D) = k^{d - |I^* \cup \{x_i\}| - 1} = k^j$. Theo (9): $\deg_{I^*}(D) \leq l \times k^{d - |I^*| - 1} \leq k^{d - |I^*|}$. Vậy, bổ đề được chứng minh với $j+1$.

Cuối cùng, xét bất kỳ $F = \{y_1, y_2, \dots, y_r\} \in \mathcal{F}$ với $r < k$. Từ (8) ta có họ $\{D \in \mathcal{D} : y_i \in D\}$ phủ $\mathcal{D}$. Áp dụng bổ đề vừa chứng minh với $I = \{y_i\} : |\{D \in \mathcal{D} : y_i \in D\}| \leq k^{d-1}$ suy ra $|\mathcal{D}| \leq k^d (= k \times k^{d-1})$ và $|\bigcup_{D \in \mathcal{D}} D| \leq |\mathcal{D}| \times d \quad (|D| = d) \leq dk^d$. Từ (7) suy ra $|G| \leq dk^d$. Định lý 1 đã được chứng minh hoàn toàn.

Như nhận xét trước đây, chúng ta xác định $F_a(d)$ như là tập vế trái của phụ thuộc hàm cực tiểu d -khoảng cách mô tả bởi Mệnh đề 2. Họ $\mathcal{F}(d)$ được xác định là hợp của họ $F_a(d)$ đối với mọi $a \in \Omega$.

Áp dụng Định lý 1 đối với phụ thuộc hàm và phụ thuộc hàm có lỗi chúng ta sẽ thu được kết quả ở định lý sau:

Định lý 2.[4] Cho $n \geq n_0(k, e)$ ($d = 2e+1$), giả sử tất cả những tập vế trái của phụ thuộc hàm cực tiểu có kích thước tối đa là k . Khi đó, kích thước các phần tử của $\mathcal{F}(2e+1)$ không thể lớn hơn $c_1 k$. Mặt khác, tồn tại một cơ sở dữ liệu với những tập vế trái của phụ thuộc hàm cực tiểu có kích thước tối đa là k , mà các phần tử của $\mathcal{F}(2e+1)$ có kích thước ít nhất là $c_2 k$, ở đây c_1, c_2 chỉ phụ thuộc vào e .

3. KẾT LUẬN

Họ các bao hàm thức tự do F_a được xác định bởi cận dưới tối ưu trong Định lý 1, biểu diễn trường hợp xấu nhất. Do đó chúng ta hy vọng có thể tìm được M^* trong họ này dựa vào việc sắp xếp chúng trên những đánh giá thực tế. Việc xây dựng tiêu chuẩn đánh giá hay sắp xếp các phần tử trong họ $\{M^*\}$ có thể là vấn đề nghiên cứu tiếp theo. Hơn nữa, với sai số $e = 1$, chúng ta có kích thước các phần tử của $\mathcal{F}(3)$ không thể lớn hơn $3k^3$ và tồn tại một cơ sở dữ liệu

với tập vế trái của phụ thuộc hàm cực tiểu với kích thước không vượt quá k có kích thước ít nhất là $c_2 k^3$, ở đây c_2 xấp xỉ $\frac{2}{27}$.

TÀI LIỆU THAM KHẢO

- [1]. Burosch, G., Demetrovics, J. and Katona G.O.H.(1987), *The Poset of Closures as a Model of Changing Databases*, Order 4, pp.127-142.
- [2]. Demetrovics, J. and Katona G.O.H., Miklós, D. (2000), *Error - correcting keys in relational databades*, in *Foundations of Information ang Knowledge Systems, FoIKS*, Lecture Notes in Computer Science, 1762, Springer, pp. 88-93.
- [3]. Demetrovics, J. and Katona G.O.H., Miklós, D.(2002), *Functional dependencies in presence of errors*, in *Foundations of Information and Knowledge Systems, FoIKS*, Lecture Notes in Computer Science, 2284, Springer, pp. 85-92.
- [4]. Demetrovics, J. and Katona G.O.H., Miklós, D.(2012), *Functional dependencies distorted by errors*, preprint submitted to Elsevier Science.
- [5]. Hồ Thuần, Hoàng Thị Lan Giao (2006), *Khám phá phụ thuộc đa trị dựa vào ma trận phụ thuộc*, Tạp chí Tin học và Điều khiển học.
- [6]. Hồ Thuần, Hoàng Thị Lan Giao (2007), *Mở rộng phụ thuộc hàm và phụ thuộc đa trị*, Tạp chí Tin Học và Điều khiển học.
- [7]. Yka Huhtala, Juha Karkkainen, Pasi Porkka and Hannu Toivonen (1999), *Tane: An Efficient Algorithm for Discovering Functional and Approximate Dependencies*, Computer Journal, Vol 42, No.2.

EXPANDING FUNCTIONAL DEPENDENCIES IN TROUBLE DATABASES

Hoang Thi Lan Giao

Department of Information Technology, Hue University College of Sciences

Email: hlgiao.it@gmail.com

ABSTRACT

In relational database with a given functional dependencies $A \rightarrow b$, values of attribute b will be uniquely determined by the ones of attributes A . However, during data transmission, the data received has been noised and may not be as the original. Clearly, on the new data the functional dependencies $A \rightarrow b$ will no longer true. In other words, It cannot be identified the objects based on the data of attributes A only. In this paper, we study how to expand the set of attributes A to a minimal set A' containing A so that $A' \rightarrow b$ is really a functional dependencies on the new data.

Keywords: *error functional dependencies, functional dependencies, relational database.*