

PHỤ THUỘC HÀM XẤP XỈ VÀ PHẦN TỬ NGOẠI LAI ĐỐI VỚI PHỤ THUỘC HÀM

VŨ ĐỨC THI¹, PHẠM HẠ THUY²

¹ Viện Công nghệ thông tin, Viện Khoa học và Công nghệ Việt Nam

² Trung tâm Khoa học và BDCB - Kiểm toán Nhà nước

Abstract. The aim of this paper is to present the concepts about Type 2 Approximate Functional Dependency - AFD 2 (relating to the correlation between attributes of the relational file) and the outliers with functional dependency. The paper also gives some properties of ADF 2 and algorithms for finding the outliers with functional dependency.

Tóm tắt. Bài báo giới thiệu về phụ thuộc hàm xấp xỉ loại 2 (loại phụ thuộc hàm xấp xỉ liên quan đến sự tương quan giữa các thuộc tính của một quan hệ) và phần tử ngoại lai đối với phụ thuộc hàm. Bài báo cũng trình bày một số tính chất của phụ thuộc hàm xấp xỉ loại 2 và thuật toán xác định phần tử ngoại lai đối với phụ thuộc hàm.

GIỚI THIỆU

Khái niệm phụ thuộc hàm xấp xỉ (Approximate functional dependency) và phương pháp phát hiện các phụ thuộc hàm xấp xỉ đã được nhiều tác giả đề cập đến và được ứng dụng trong nhiều bài toán phân lớp của data mining ([1, 2]). Những phụ thuộc như vậy biểu diễn sự phụ thuộc tự nhiên giữa các thuộc tính trong một quan hệ nhưng có chứa một vài hàng có sai sót hoặc không thỏa luật (gọi là phụ thuộc hàm xấp xỉ loại 1). Một trường hợp phụ thuộc xấp xỉ khác là có những nhóm thuộc tính mặc dù giữa chúng không có sự phụ thuộc hàm theo kiểu bằng nhau tuyệt đối (theo cách định nghĩa phụ thuộc hàm thông thường) mà có sự phụ thuộc xấp xỉ theo kiểu tương quan hàm số (tuyến tính hoặc phi tuyến), (gọi là phụ thuộc hàm xấp xỉ loại 2). Trường hợp này xảy ra khá nhiều và liên quan đến nhiều bài toán thực tế. Ví dụ trong bảng quan hệ ghi lại tình hình mua vào các loại hàng hóa của một doanh nghiệp ta thấy quan hệ giữa khối lượng hàng mua vào và chi phí để mua là phụ thuộc hàm xấp xỉ loại 2. Nếu sự chênh lệch giữa khối lượng hàng mua vào nhỏ hơn một mức nào đó thì cũng kéo theo sự chênh lệch của chi phí hai đợt mua cũng nhỏ (giá trị của hai bản ghi tại các thuộc tính này được phép có sự chênh lệch tương ứng). Cũng vậy đối với chỉ tiêu doanh thu và chi phí trong bảng kê về tình hình kinh doanh... Việc khảo sát loại phụ thuộc hàm xấp xỉ này có ứng dụng thực tiễn trong việc phát hiện những hiện tượng bất thường (ngoại lai) trong sản xuất kinh doanh, phục vụ cho hoạt động kiểm toán và quản lý kinh tế. Việc quyết định mức chênh lệch cho phép (vượt quá mức đó là hiện tượng bất thường) thường được xác định theo hướng hỗ trợ quyết định ([3]) có nghĩa là còn tùy thuộc thêm các yếu tố về yêu cầu và kinh nghiệm của các chuyên gia trong các lĩnh vực thực tế.

Dưới đây là khái niệm của phụ thuộc hàm xấp xỉ loại 2 và khảo sát một số tính chất của nó. Đồng thời, một số khái niệm, định nghĩa về phần tử ngoại lai theo phụ thuộc hàm và các mệnh đề làm cơ sở cho việc xây dựng thuật toán xác định phần tử ngoại lai đối với phụ thuộc hàm cũng được trình bày.

1. KHÁI NIỆM PHỤ THUỘC HÀM XẤP XỈ LOẠI 2

Cho r là một quan hệ trên tập thuộc tính $R = \{a_1, a_2, \dots, a_n\}$ trong đó các thuộc tính a_1, a_2, \dots, a_n có thể là các thuộc tính định danh (categorical), rời rạc hoặc liên tục (trường số). Đối với những thuộc tính định danh, ta tiến hành thực hiện ánh xạ tất cả các giá trị có thể tới một tập các số nguyên dương liên kề.

Định nghĩa 1.1. (khoảng cách giữa 2 bộ giá trị trên tập thuộc tính).

Với 2 bộ $t_1, t_2 \in r$, ta kí hiệu $\rho(t_1(X), t_2(X))$ là khoảng cách giữa t_1 và t_2 trên tập thuộc tính $X \subseteq R$, được xác định như sau:

$$\rho(t_1(X), t_2(X)) = \max_{a_i \in X} \frac{(|t_1(a_i) - t_2(a_i)|)}{\max(|t_1(a_i)|, |t_2(a_i)|)},$$

hàm $\max(x, y)$ là hàm chọn ra số lớn nhất trong 2 số x, y .

Trường hợp $\max(|t_1(a_i)|, |t_2(a_i)|) = 0$, ta qui ước:

$$|t_1(a_i) - t_2(a_i)| / \max(|t_1(a_i)|, |t_2(a_i)|) = 0.$$

Khoảng cách giữa 2 bộ giá trị trên tập thuộc tính có thể coi là hàm số của các đối số và là các bộ giá trị của quan hệ và tập các thuộc tính.

Một số tính chất của hàm khoảng cách $\rho(t_1(X), t_2(X))$

Dễ dàng chứng minh định nghĩa khoảng cách $\rho(t_1(X), t_2(X))$ thỏa mãn các tính chất của hàm khoảng cách $\forall t_1, t_2, t_3 \in r, \forall X, Y \subseteq R$, ta có:

- a1) $\rho(t_1(X), t_2(X)) \geq 0$
- a2) $\rho(t_1(X), t_2(X)) = 0 \leftrightarrow t_1(X) = t_2(X)$
- a3) $\rho(t_1(X), t_2(X)) \leq \rho(t_1(X), t_3(X)) + \rho(t_3(X), t_2(X))$

Và ngoài ra cũng dễ chứng minh các tính chất sau:

- a4) Nếu $X \subseteq Y$ thì $\rho(t_1(X), t_2(X)) \leq \rho(t_1(Y), t_2(Y))$
- a5) $\rho(t_1(XY), t_2(XY)) = \max(\rho(t_1(X), t_2(X)), \rho(t_1(Y), t_2(Y)))$

Định nghĩa 1.2. (phụ thuộc hàm xấp xỉ loại 2 - Type 2 Approximate Functional Dependency). Giả sử $X, Y \subseteq R$ và với một số δ cho trước, $0 \leq \delta < 1$, ta nói rằng X xác định hàm Y mức δ (hoặc giữa X, Y có phụ thuộc hàm xấp xỉ loại 2 mức δ), ký hiệu là $X \approx_{\delta} Y$ nếu với mọi cặp bộ $t_1, t_2 \in r$, mà $\rho(t_1(X), t_2(X)) \leq \delta$ thì ta cũng có $\rho(t_1(Y), t_2(Y)) \leq \delta$.

Ví dụ 1. Xét bảng quan hệ sau:

A	B	C	D	E
2.5	18	160	25	12
2.51	18	160.5	15	13
3.6	20	320	30	16
3.65	20	323	45	28
4.25	25	641	60	19
4.26	25	643	70	57

Ta thấy giữa các cột A, B có mối tương quan với cột C. Với $\delta = 0.05$ ta kiểm tra điều kiện phụ thuộc hàm xấp xỉ loại 2: $AB \rightarrow_{0.05} C$.

Với cặp hàng 1, 2, ta có

$$\begin{aligned} & \rho(t_1(AB), t_2(AB)) \\ &= \max(|t_1(A) - t_2(A)| / \max(|t_1(A)|, |t_2(A)|), |t_1(B) - t_2(B)| / \max(|t_1(B)|, |t_2(B)|)) \\ &= 0.00394 < 0.05. \end{aligned}$$

Ta cũng tính được: $\rho(t_1(C), t_2(C)) = 0.00311 < 0.05$.

Tương tự ta cũng sẽ kiểm tra dễ dàng với cặp hàng 3, 4 và cặp hàng 5, 6 cũng như các cặp hàng khác.

Vậy ta có: $AB \approx_{0.05} C$ (AB xác định hàm C mức 0.05).

2. MỘT SỐ TÍNH CHẤT CỦA PHỤ THUỘC HÀM XẤP XỈ LOẠI 2

Tính chất 1. Cho r là một quan hệ trên tập thuộc tính R . Một phụ thuộc hàm đúng trên r cũng là phụ thuộc hàm xấp xỉ loại 2 với mức δ tùy ý ($0 \leq \delta < 1$) đúng trên r .

Tính chất này dễ dàng suy theo Định nghĩa 1.2.

Tính chất 2. Cho r là một quan hệ trên R ; $X, Y \subseteq R$; δ_1, δ_2 là hai số sao cho $0 \leq \delta_1 < \delta_2 < 1$. Kí hiệu $X \approx_{\delta_1} Y$ và $X \approx_{\delta_2} Y$ là hai phụ thuộc hàm xấp xỉ loại 2 mức δ_1 và mức δ_2 giữa X và Y trên r , khi đó nếu $X \approx_{\delta_1} Y$ đúng trên r thì $X \approx_{\delta_2} Y$ cũng đúng trên r .

Hay viết một cách hình thức: $X \approx_{\delta_1} Y \Rightarrow X \approx_{\delta_2} Y$.

Thật vậy, đặt $e = \delta_2 - \delta_1$.

Giả sử $X \approx_{\delta_1} Y$ đúng trên r , tức là với mọi cặp $t_1, t_2 \in r$ nếu $\rho(t_1(X), t_2(X)) \leq \delta_1$ thì ta cũng có $\rho(t_1(Y), t_2(Y)) \leq \delta_1$.

Do vậy với mọi cặp bộ $t_1, t_2 \in r$, nếu $\rho(t_1(X), t_2(X)) \leq \delta_1 + e$ thì $\rho(t_1(Y), t_2(Y)) \leq \delta_1 + e$. Tức là nếu $\rho(t_1(Y), t_2(Y)) \leq \delta_2$ thì $\rho(t_1(Y), t_2(Y)) \leq \delta_2$.

Suy ra $X \approx_{\delta_2} Y$ đúng trên r .

Tính chất 3. (tính phản xạ) Nếu $Y \subseteq X$ khi đó $X \approx_{\delta} Y$ là phụ thuộc hàm xấp xỉ loại 2 với mức δ tùy ý ($0 \leq \delta < 1$).

Thật vậy, nếu $Y \subseteq X$ thì $X \rightarrow Y$ là phụ thuộc hàm đúng trên r thì theo Tính chất 1 nó cũng là phụ thuộc hàm xấp xỉ loại 2 với mức δ tùy ý ($0 \leq \delta < 1$).

Tính chất 4. (tính bắc cầu) Nếu $X \approx_{\delta} Y$ và $Y \approx_{\delta} Z$ thì $X \approx_{\delta} Z$.

Thật vậy:

Nếu $X \approx_{\delta} Y$ thì với mọi bộ $t_1, t_2 \in r$. Nếu $\rho(t_1(X), t_2(X)) \leq \delta$, ta có $\rho(t_1(Y), t_2(Y)) \leq \delta$.

Cũng do $Y \approx_{\delta} Z$ nên từ $\rho(t_1(Y), t_2(Y)) \leq \delta$ suy ra $\rho(t_1(Z), t_2(Z)) \leq \delta$, có nghĩa là $X \approx_{\delta} Z$. Điều phải chứng minh.

Tính chất 5. (tính gia tăng) Với mọi $X, Y, Z \subseteq R$ và mức δ nào đó, nếu $X \approx_{\delta} Y$ thì $XZ \approx_{\delta} YZ$.

Thật vậy, với mọi cặp bộ $t_1, t_2 \in r$ mà ta có $\rho(t_1(XZ), t_2(XZ)) \leq \delta$ thì do

$$\rho(t_1(X), t_2(X)) \leq \rho(t_1(XZ), t_2(XZ))$$

nên ta có

$$\rho(t_1(X), t_2(X)) \leq \delta.$$

Mặt khác, do $X \approx_{\delta} Y$ nên ta có $\rho(t_1(Y), t_2(Y)) \leq \delta$, và

$$\rho(t_1(Z), t_2(Z)) \leq \rho(t_1(XZ), t_2(XZ)) \leq \delta,$$

từ đó

$$\rho(t_1(YZ), t_2(YZ)) = \max(\rho(t_1(Z), t_2(Z)), \rho(t_1(Y), t_2(Y))) \leq \delta.$$

Suy ra điều phải chứng minh. \blacksquare

3. PHẦN TỬ NGOẠI LAI ĐỐI VỚI PHỤ THUỘC HÀM

Trong một quan hệ, phụ thuộc hàm mô tả sự phụ thuộc dữ liệu mà trong đó giá trị của một tập thuộc tính này xác định giá trị của tập thuộc tính kia. Ví dụ thu nhập của công chức trong một đơn vị đơn thuần chỉ phụ thuộc vào mức lương và phụ cấp. Trong bảng kê về thu nhập, ta có phụ thuộc hàm phản ánh nhóm thuộc tính hệ số lương, hệ số phụ cấp xác định thu nhập của từng người. Vì vậy nếu có những bản ghi có giá trị hệ số lương, hệ số phụ cấp như nhau nhưng giá trị thu nhập khác nhau thì là một hiện tượng bất thường, làm cho phụ thuộc hàm này không đúng. Đây là hiện tượng ngoại lai đối với phụ thuộc hàm. Những hiện tượng này thường xảy ra trong thực tế khi cập nhật hoặc sao chép những tập dữ liệu. Nguyên nhân có thể do sai sót hoặc gian lận. Việc phát hiện những hiện tượng nói trên cần thiết trong việc kiểm tra và làm sạch dữ liệu. Nội dung dưới đây chúng tôi trình bày khái niệm và phương pháp phát hiện trường hợp ngoại lai này.

Định nghĩa 3.1. (Phần tử ngoại lai đối với phụ thuộc hàm) Giả sử $X \rightarrow Y$ là một phụ thuộc hàm được giả thiết đúng trên quan hệ r . Khi đó cặp phần tử (t_1, t_2) với $t_1, t_2 \in r$ là cặp phần tử ngoại lai đối với phụ thuộc hàm $X \rightarrow Y$ nếu $t_1(X) = t_2(X)$ và $t_1(Y) \neq t_2(Y)$.

Trong nội dung dưới đây chúng tôi sử dụng khái niệm về hệ bằng nhau được trình bày trong [4].

Giả sử $r = \{t_1, t_2, \dots, t_m\}$ là bảng dữ liệu được giả thiết là một quan hệ. Kí hiệu E_r là hệ bằng nhau của r được xác định như sau:

$$E_r = \{E_{i,j} : 1 \leq i \leq j \leq m \text{ và } E_{i,j} = \{a \in R; t_i(a) = t_j(a)\}\}.$$

Định lý 3.1. (Nhận biết cặp ngoại lai đối với phụ thuộc hàm) Cho r là một bảng dữ liệu được giả thiết là một quan hệ trên tập thuộc tính R , E_r là hệ bằng nhau của r , $X \rightarrow Y$ là một phụ thuộc hàm được giả thiết đúng trên r . Cặp phần tử (t_i, t_j) với $t_i, t_j \in r$ là ngoại lai đối với phụ thuộc hàm $X \rightarrow Y$ khi và chỉ khi $E_{i,j} \in E_r$ mà $X \subseteq E_{i,j}$ nhưng $Y \not\subseteq E_{i,j}$.

Chứng minh. Thật vậy, giả sử (t_i, t_j) là cặp ngoại lai đối với phụ thuộc hàm $X \rightarrow Y$, có nghĩa là ta có $t_i(X) = t_j(X)$ nhưng $t_i(Y) \neq t_j(Y)$. Từ định nghĩa $E_{i,j}$ ta có $X \subseteq E_{i,j}$ và $Y \not\subseteq E_{i,j}$.

Ngược lại, nếu có $E_{i,j} \in E_r$ ($E_{i,j}$ xác định theo t_i, t_j) mà $X \subseteq E_{i,j}$ và $Y \not\subseteq E_{i,j}$ thì cũng theo cách xác định $E_{i,j}$ ta có $t_i(a) = t_j(a)$ với $a \in E_{i,j}$. Do $X \subseteq E_{i,j}$ nên $t_i(X) = t_j(X)$. Cũng do $Y \not\subseteq E_{i,j}$ nên tồn tại $b \in Y$ mà $t_i(b) \neq t_j(b)$ hay là $t_i(Y) \neq t_j(Y)$. Theo Định nghĩa 2.1 thì (t_i, t_j) là cặp ngoại lai. Điều phải chứng minh. \blacksquare

Từ định lý trên ta có thuật toán xác định các cặp ngoại lai đối với phụ thuộc hàm như sau.

Thuật toán 1. (Xác định các cặp ngoại lai đối với tập các phụ thuộc hàm)

Input: Tập thuộc tính $R = \{a_1, a_2, \dots, a_n\}$, bảng dữ liệu $r = \{t_1, t_2, \dots, t_m\}$ trên R

Tập các phụ thuộc hàm $F = \{X_1 \rightarrow Y_1; X_2 \rightarrow Y_2, \dots, X_m \rightarrow Y_s\}$

Output: OUTLI - tập các cặp ngoại lai đối với phụ thuộc hàm

Bắt đầu

Bước 1: Tính hệ bằng nhau của r :

$$E_r = \{E_{i,j} | 1 \leq i < j \leq m, E_{i,j} = \{a \in R; t_i(a) = t_j(a)\}\}.$$

Bước 2: Với mỗi phụ thuộc hàm $X_i \rightarrow Y_i \in F$ và mọi $E_{k,j} \in E_r$, kiểm tra điều kiện $X_i \subseteq E_{k,j}$ and $Y_i \subset E_{k,j}$. Nếu đúng, lưu cặp (t_k, t_j) vào tập OUTLI. Nếu không, kiểm tra tiếp các phụ thuộc hàm khác.

Kết thúc

Tập OUTLI là tập các cặp ngoại lai đối với phụ thuộc hàm của r .

Tính đúng đắn của Thuật toán 1 được suy ra từ Định lý 3.1.

4. PHẦN TỬ NGOẠI LAI ĐỐI VỚI KHÓA

Trong một quan hệ, khóa là tập thuộc tính xác định tính duy nhất của mỗi bản ghi (phần tử) của quan hệ. Do vậy trong quan hệ không thể có 2 bản ghi có giá trị trùng nhau trên khóa. Tuy nhiên trong thực tế, trong một tập dữ liệu, việc có những cặp bản ghi có giá trị trùng nhau trên khóa là có thể xảy ra (do cập nhật dữ liệu sai, hoặc có sự gian lận cố tình nhập nhiều lần cùng một bản ghi vào một bảng dữ liệu). Cặp bản ghi này gọi là ngoại lai đối với khóa và cần phát hiện để xử lý. Nội dung dưới đây sẽ trình bày khái niệm và phương pháp phát hiện ngoại lai đối với trường hợp này.

Định nghĩa 4.1. (Phần tử ngoại lai đối với khóa) Cho bảng dữ liệu r được giả thiết là một quan hệ trên tập thuộc tính R , B được giả thiết là tập các khóa tối thiểu của r . Cặp phần tử (t_i, t_j) với $t_i, t_j \in r$ ($i \neq j$) là một cặp ngoại lai đối với khóa nếu như với một khóa nào đó $K \in B$ mà ta có $t_i(K) = t_j(K)$.

Theo qui định, thì trong một quan hệ không có sự trùng nhau về giá trị của các bản ghi trên khóa. Do vậy sự trùng nhau về giá trị trên khóa là hiện tượng ngoại lai cần được xử lý.

Trường hợp nếu $t_i(K) = t_j(K)$ và $t_i(R \setminus K) = t_j(R \setminus K)$ tức chúng trùng nhau hoàn toàn về giá trị trên các thuộc tính, có nghĩa là ta có $t_i(R) = t_j(R)$ thì cặp (t_i, t_j) được gọi là cặp ngoại lai tầm thường (theo qui ước thì trong một quan hệ không được có 2 bộ trùng nhau hoàn toàn). Còn nếu $t_i(K) = t_j(K)$ nhưng $t_i(R \setminus K) \neq t_j(R \setminus K)$ thì ta có cặp ngoại lai theo phụ thuộc hàm $K \rightarrow R$.

Mệnh đề sau làm cơ sở cho việc phát hiện các cặp ngoại lai đối với khóa của quan hệ.

Mệnh đề 2.1. (Nhận biết cặp ngoại lai theo khóa) Cho bảng dữ liệu r được giả thiết là một quan hệ trên tập thuộc tính R , B là tập các khóa tối thiểu của r , E_r là hệ bằng nhau của r . Khi đó nếu ta có $E_{i,j} \in E_r$ chứa một khóa K nào đó (tức là $K \subseteq E_{i,j}$) thì cặp phần tử (t_i, t_j) với $t_i, t_j \in r$ (tương ứng với $E_{i,j}$) là một cặp ngoại lai đối với khóa.

Chứng minh. Theo định nghĩa khóa của một quan hệ, nếu K là một khóa thì ta có $K \rightarrow R$.

Giả sử $K \subseteq E_{i,j}$, theo cách xác định hệ bằng nhau, ta có $t_i(K) = t_j(K)$. Mặt khác do $K \rightarrow R$ nên ta phải có $t_i(R) = t_j(R)$. Vì vậy:

- Nếu $t_i(R) = t_j(R)$ ta có một cặp ngoại lai tầm thường (hai bản ghi trùng nhau hoàn toàn, vi phạm điều kiện r là một quan hệ) .

- Nếu $t_i(K) = t_j(K)$ và $t_i(R \setminus K) \neq t_j(R \setminus K)$ ta sẽ có một cặp ngoại lai đối với phụ thuộc hàm $K \rightarrow R$ (theo định nghĩa của khóa). Mệnh đề được chứng minh.

Ta có thuật toán sau để xác định các cặp ngoại lai theo khóa của một quan hệ.

Thuật toán 2. (Xác định phần tử ngoại lai đối với khóa của quan hệ)

Input: Bảng dữ liệu $r = \{t_1, t_2, \dots, t_m\}$ trên tập thuộc tính $R = (a_1, a_2, \dots, a_n)$.

B tập các khóa của r .

Output: tập các cặp ngoại lai theo khóa OUTLI2

Bắt đầu

Bước 1: Tính hệ bằng nhau $E_r = \{E_{i,j} | 1 \leq i < j \leq m \text{ và } E_{i,j} = \{a \in R; t_i(a) = t_j(a)\}\}$.

Bước 2: Với mỗi khóa $K \in B$ và với mỗi tập $E_{i,j} \in E_r$, ta kiểm tra điều kiện $K \subseteq E_{i,j}$.

Nếu đúng lưu cặp (t_i, t_j) vào OUTLI2.

Kết thúc

Tập OUTLI2 là tập các phần tử ngoại lai theo khóa.

Trên đây là một số kết quả nghiên cứu về phụ thuộc hàm xấp xỉ loại 2, phần tử ngoại lai đối với phụ thuộc hàm. Đây là những vấn đề mới đang được chúng tôi nghiên cứu. Các kết quả nghiên cứu khác về những vấn đề này có ý nghĩa trong việc giải quyết một số bài toán thực tế (ví dụ trong lĩnh vực kiểm toán) sẽ được chúng tôi giới thiệu trong các nội dung khác.

TÀI LIỆU THAM KHẢO

- [1] Yka Huhtala, Juha Kahkkainen, Pasi Porkka, Hannu Toivonen, An efficient algorithm for discovering functional and approximate dependencies, *Proc. 14th Int. Conf. on Data Engineering (ICDE '98)*, IEEE. Computer Society Press (1998) 392–401.
- [2] B. Liu, W. Hsu, and Y. Ma, Integrating classification and association mining, *Proc. Int. Conf. Knowledge Discovery and Data Mining (KDD'98)*, New York (1998) 80–86.
- [3] E. Turban, *Decision Support and Expert Systems Management Support Systems*, Prentice Hall, 1995.
- [4] Vũ Đức Thi, *Cơ sở dữ liệu- kiến thức và thực hành*, NXB Thống kê, Hà Nội, 1997.

Nhận bài ngày 3 - 11 - 2005

Nhận lại sau sửa ngày 1 - 3 - 2007