Tạp chí Tin học và Điều khiển học, T.23, S.2 (2007), 179–186

XÂY DỰNG CÂY QUYẾT ĐỊNH SỬ DỤNG PHỤ THUỘC HÀM XẤP XỈ

VŨ ĐỨC THI, TRẦN QUANG DIỆU

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

Abstract. In this paper, we introduce in brief on the concepts of Approximate Functional Dependency (AFD), of Approximate Functionally Cross-Characteristic Dependency known as type II AFD and describe an adoption of AFD to a constructing method of decisson tree for databases mining purposes.

Tóm tắt. Trong bài báo này, chúng tôi giới thiệu sơ lược về khái niệm phụ thuộc hàm xấp xỉ, phụ thuộc hàm xấp xỉ liên quan đến tương quan hàm số giữa các thuộc tính của một quan hệ (Phụ thuộc hàm xấp xỉ loại hai) và ứng dụng phụ thuộc hàm xấp xỉ nhằm xây dựng cây quyết định trong khai phá dữ liệu.

1. GIỚI THIỆU

Phụ thuộc hàm xấp xỉ (Approximate Functional Dependency AFD) và phương pháp phát hiện các phụ thuộc hàm xấp xỉ đã được nhiều tác giả đề cập và ứng dụng trong nhiều bài toán khai phá dữ liệu ([1,3]). Phụ thuộc hàm xấp xỉ là một phụ thuộc hàm có tính chất gần đúng đối với một quan hệ r và được định nghĩa như sau.

Định nghĩa 1. Phụ thuộc hàm xấp xỉ (Approximate Functional Dependency - AFD)

Cho ε, $0 \leqslant \varepsilon \leqslant 1$, $X \to Y$ là phụ thuộc hàm xấp xỉ nếu: $approx(X \to Y) \leqslant \varepsilon$, với $approx(X \to Y) = 1$ (max $\{|s|$, s là tập con của r và $X \to Y$ đúng trên $s\}/|r|$).

Ở đây, $|s|$, $|r|$ là số phần tử của s và r.

Trong các bài toán quản lý thực tế, thường xảy ra các nhóm thuộc tính có sự liên quan dưới dạng hàm số với nhau (tuyến tính hoặc phi tuyến). Đối với các trường hợp này, ta xét đến phụ thuộc hàm xấp xỉ loại hai - phụ thuộc hàm xấp xỉ liên quan đến tương quan hàm số giữa các thuộc tính của một quan hệ.

Việc xây dựng cây quyết định trong khai phá dữ liệu đã được nhiều tác giả đề cập đến trong các phương pháp khai phá dữ liệu nhằm kết xuất các thông tin hữu ích tiềm ẩn trong các cơ sở dữ liệu lớn. Cây quyết định là một kiểu mô hình dự báo (predictive model), nghĩa là một ánh xạ từ các quan sát về một sự vật/hiện tượng tới các kết luận về giá trị mục tiêu của sự vật/hiện tượng. Cây quyết định có cấu trúc hình cây và là một sự tượng trưng của một phương thức quyết định cho việc xác định lớp các sự kiện đã cho. Mỗi nút của cây chỉ ra một tên lớp hoặc một phép thử cụ thể, phép thử này chia không gian các dữ liệu tại nút đó thành các kết quả có thể đạt được của phép thử. Mỗi tập con được chia ra là không gian con của các dữ liệu được tương ứng với vấn đề con của sự phân lớp. Sự phân chia này

thông qua một cây con tương ứng. Quá trình xây dựng cây quyết định có thể xem như là một chiến thuật chia để trị cho sự phân lớp đối tượng ([4–6]). Một cây quyết định có thể mô tả bằng các khái niệm nút và đường nối các nút trong cây. Việc nghiên cứu xây dựng cây quyết định mang lại hiệu quả tốt cho việc làm trong sạch dữ liệu, phát hiện sai sót và đưa ra quyết định phù hợp trong từng hoàn cảnh và chiến lược cụ thể của bài toán quản lý.

Phụ thuộc hàm (FDs) đã được nghiên cứu rất nhiều trong khi phân tích, thiết kế cơ sở dữ liệu. Phụ thuộc hàm giữa các thuộc tính quan hệ cho phép xác định chính xác các mối quan hệ trong cơ sở dữ liệu ([2]). Các ràng buộc do phụ thuộc hàm quy định trong sơ đồ quan hệ tương đối độc lập với dữ liệu. Việc xây dựng cây quyết định sử dụng phụ thuộc hàm sẽ mang lại hiệu quả tốt do các tính chất ràng buộc chặt của phụ thuộc hàm.

Trong bài báo này, chúng tôi trình bày về khái niệm phụ thuộc hàm xấp xỉ, phụ thuộc hàm xấp xỉ loại hai, một số tính chất của phụ thuộc hàm xấp xỉ và ứng dụng vào việc xây dựng cây quyết định trong khai phá dữ liệu.

2. PHỤ THUỘC HÀM XẤP XỈ LOẠI HAI

Cho r là một quan hệ trên tập thuộc tính $R = \{A_1, A_2, ...A_n\}$ trong đó các thuộc tính A_1,A_2, ...A_n có thể là thuộc tính định danh, rời rạc hoặc liên tục. 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 liền kề.

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

Với 2 bộ $t_1, t_2 \in r$, ta ký 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

$$\rho(t_1(X), t_2(X)) = max(|t_1(A_i) - t_2(A_i)|)/max(|t_1(A_i), t_2(A_i)|), A_i \in X).$$

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

Trường hợp $max(|t_1(A_i)|, |t_2(A_i)|) = 0$, tức $t_1(A_i) = t_2(A_i) = 0$ thì 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 là các bộ giá trị của quan hệ và tập các thuộc tính.

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

Định nghĩa khoảng cách $\rho(t_1(X), t_2(X))$ nêu trên thỏa mãn các tính chất của hàm khoảng cách:

1. $\rho(t_1(X), t_2(X)) \geqslant 0$ với t_1, t_2, X tùy ý
2. $\rho(t_1(X), t_2(X)) = 0 \Leftrightarrow t_1(X) = t_2(X)$
3. $\rho(t_1(X), t_2(X)) \geqslant \rho(t_1(X), t_3(X)) + \rho(t_3(X), t_2(X))$
4. Nếu $X \subseteq Y$ thì $\rho(t_1(X), t_2(X)) \geqslant \rho(t_1(Y), t_2(Y))$
5. $\rho(t_1(XY), t_2(XY)) = max(\rho(t_1(X), t_2(X)), \rho(t_1(Y), t_2(Y)))$.

Định nghĩa 3. (Phụ thuộc hàm xấp xỉ loại hai - Type II Approximate Functional Dependency)

Giả sử $X, Y \subseteq R$ và với một số δ cho trước, $0 \leqslant \delta < 1$, ta nói rằng X xác định hàm Y mức δ (hoặc nói rằng X, Y có phụ thuộc hàm xấp xỉ loại hai mức δ), ký hiệu là $X \approx\!>_\delta Y$ nếu với mọi cặp bộ $t_1, t_2 \in r$, mà $\rho(t_1(X), t_2(X)) \leqslant \varepsilon$ thì ta cũng có $\rho(t_1(Y), t_2(Y)) \leqslant \varepsilon$.

Mệnh đề 1. (Điều kiện để phụ thuộc hàm xấp xỉ là phụ thuộc hàm xấp xỉ loại hai)

Giả sử $approx(X \to Y)$ là một phụ thuộc hàm xấp xỉ với độ xấp xỉ ε. Phụ thuộc hàm xấp xỉ $approx(X \to Y)$ là phụ thuộc hàm xấp xỉ loại hai mức $\delta = \varepsilon (X \approx\!>_\delta Y)$ khi và chỉ khi ứng với ε với mọi cặp bộ $t_1, t_2 \in r$, mà $\rho(t_1(X), t_2(X)) \leqslant \varepsilon$.

Ta thấy mệnh đề trên hoàn toàn đúng với theo định nghĩa và các tính chất của phụ thuộc hàm xấp xỉ loại hai.

Thuật toán 1. Kiểm tra phụ thuộc hàm xấp xỉ loại hai

Bước 1: Xây dựng hệ xấp xỉ mức δ của r: $E_{r\delta}$.

Bước 2: Với mỗi $E(\delta)_{i,j} \in E_{r\delta}$ (với $1 \leqslant i < j \leqslant m$) lần lượt kiểm tra điều kiện

$$X \subseteq E(\delta)_{i,j}) \Rightarrow (Y \subseteq E(\delta)_{i,j}).$$

+ Nếu có tồn tại $X \subseteq E(\delta)_{i,j}$ và $Y \nsubseteq E(\delta)_{i,j}$ thì dừng việc kiểm tra và kết luận r không thỏa phụ thuộc hàm xấp xỉ loại hai mức δ.

+ Nếu với mọi $E(\delta)_{i,j} \in E_{r\delta}$ thỏa mãn điều kiện $(X \subseteq E(\delta)_{i,j}) \Leftrightarrow (Y \subseteq E(\delta)_{i,j})$ thì kết luận r thỏa phụ thuộc hàm xấp xỉ loại hai mức δ.

Thuật toán 2. Xây dựng cây quyết định sử dụng phụ thuộc hàm xấp xỉ loại hai

Đầu vào: Mẫu thử (Examples) và xác định các phụ thuộc hàm xấp xỉ loại hai được tính trên tập dữ liệu mẫu.

Đầu ra: Cây quyết định dựa trên phụ thuộc hàm.

Bước 1. Xác định nhiễu của mẫu thử ($MajorityClass(Examples_i)$).

Bước 2. Chọn các phụ thuộc hàm xấp xỉ có mức độ nhiễu trong mức xấp xỉ δ,

$$approx(X \to Y) \leqslant \delta.$$

Bước 3. Kiểm tra phụ thuộc hàm xấp xỉ loại hai cho tập phụ thuộc hàm xấp xỉ tìm được ở Bước 2 ($X \approx\!>_\delta Y$).

+ Chọn phụ thuộc hàm xấp xỉ loại hai có mức xấp xỉ nhỏ nhất ($X \approx\!>_{min(\delta)} Y$)

+ Tạo cây quyết định với gốc là phụ thuộc hàm xấp xỉ vừa chọn

$DT = BuildDecisionTree(Example_i, X \approx\!>_{min(\delta)} Y, MajorityClass(examples))$.

+ Với mỗi giá trị v_i của tập phụ thuộc hàm xấp xỉ, tính

- Mẫu thử $Example_i = \{Ki \subseteq Example$ với $X \approx\!>_{min_\delta} Y = v_i\}$
- $Subtree = BuildDecisionTree(Example_i, X \approx\!>_{min_\delta} Y, MajorityClass(examples))$
- Thêm một nhánh Subtree vào cây với nhãn là v_i

Ví dụ 1. Giả sử ta có tập huấn luyện như sau (Bảng 1)

Chúng ta có thể xây dựng các phụ thuộc hàm xỉ loại hai với mức xấp xỉ $\delta = 0.05$ như sau.

Ta thấy giữa thuộc tính Tuổi, Hệ số lương có mối tương quan với chức danh. Với $\delta = 0.05$ ta kiểm tra điều kiện đối với phụ thuộc hàm xấp xỉ:

Bảng 1

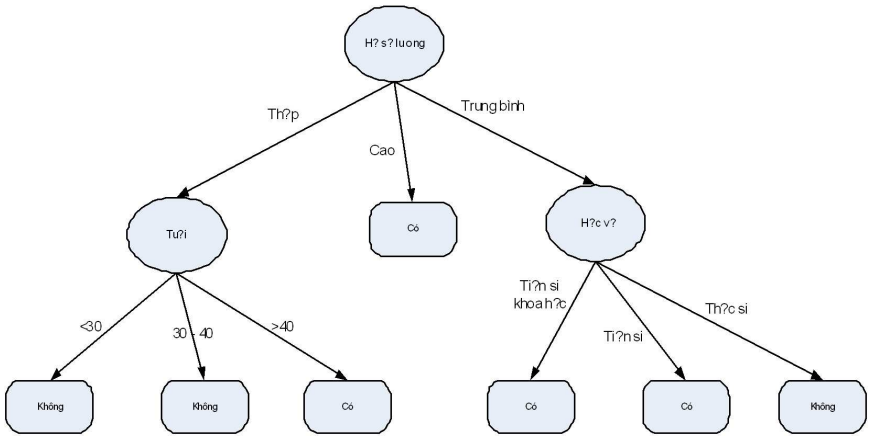
STT	Tuổi	HSL	Ngạch CC	HV	Hoàn thành công việc
1	>40	Cao	Nghiên cứu viên chính	Tiến sĩ khoa học	Có
2	>40	Cao	Nghiên cứu viên chính	Tiến sĩ	Có
3	>40	Trung bình	Nghiên cứu viên	Tiến sĩ	Có
4	>40	Trung bình	Nghiên cứu viên	Thạc sĩ	Không
5	30-40	Trung bình	Nghiên cứu viên chính	Tiến sĩ	Có
6	30-40	Thấp	Nghiên cứu viên	Thạc sĩ	Không
7	<30	Trung bình	Nghiên cứu viên	Tiến sĩ	Có
8	<30	Thấp	Nghiên cứu viên	Thạc sĩ	Không
9	30-40	Thấp	Nghiên cứu viên	Thạc sĩ	Không

Với cặp hàng 1, 2 ta có $\rho(t_1$ (Tuổi, Hệ số lương), t_2 (Tuổi, Hệ số lương)) $= 0 < 0.05$.

Ta cũng tính được $\rho(t_1$ (Có thể giao nhiệm vụ), t_2(Có thể giao nhiệm vụ)) $= 0 < 0.05$.

Tương tự ta cũng kiểm tra dễ dàng với các cột còn lại, vậy ta có phụ thuộc hàm Tuổi, hệ số lương $\approx >_{0.05}$ Có thể giao nhiệm vụ.

Sau khi tìm tất cả các phụ thuộc hàm xấp xỉ phù hợp với quá trình xây dựng cây quyết định, ta có thể xây dựng được cây quyết định như sau (Hình 1).



Hình 1. Cây quyết định

Thử nghiệm so sánh trên tập dữ liệu huấn luyện của hơn 5000 cán bộ công chức của Viện Khoa học và Công nghệ Việt Nam dùng một số phương pháp xây dựng cây quyết định như ID3, C4.5 thì thấy rằng sử dụng thuật toán xây dựng cây quyết định sử dụng phụ thuộc hàm xấp xỉ loại hai có thời gian xử lý nhanh hơn và độ chính xác tương ứng.

Vấn đề cơ sở dữ liệu đã được nghiên cứu từ rất sớm trong quá trình phát triển của công nghệ thông tin, các khái niệm, tính chất của cơ sở dữ liệu đặc biệt là phụ thuộc hàm trong cơ sở dữ liệu quan hệ đã được chứng minh một cách chặt chẽ. Khác với việc lựa chọn khá cảm tính trong các phương pháp lựa chọn thuộc tính để phát triển khác, tuy nhiên, với định nghĩa chặt như phụ thuộc hàm đã được nêu ở trên thì khi gặp một cơ sở dữ liệu lớn

và phức tạp, việc xác định các phụ thuộc hàm rất khó khăn, chính vì thế phương pháp xây dựng cây quyết đinh dựa trên phụ thuộc hàm xấp xỉ loại hai đã phần nào giải quyết các vấn đề trên.

Trên đây là một số kết quả nghiên cứu về phụ thuộc hàm xấp xỉ, phụ thuộc hàm xấp xỉ loại hai và ứng dụng trong xây dựng cây quyết định sử dụng phụ thuộc hàm trong khai phá dữ liệu. Các kết quả nghiên cứu về những vấn đề này có ý nghĩa trong việc giải quyết một số bài toán thực tế khi khai phá dữ liệu với cơ sở dữ liệu lớn nhằm đưa ra các quyết định phù hợp với bài toán thực tế như quản lý kinh tế, khai phá tri thức, hệ chuyên gia...

TÀI LIỆU THAM KHẢO

[1] B. Liu, W. Hsu, and Y. Ma, Integrating classification and association mining, *Proc. 1998 Int. Conf. Knowledge Discovery and Data mining (KDD'98)*, New York (80–86).

[2] Vũ Đức Thi, *Cơ sở dữ liệu- kiến thức và thực hành*, Nhà xuất bản Thống kê, 1997.

[3] 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 (392–401).

[4] Ho Tu Bao, Knowledge discovery and data mining techniques and practice, http://www.jaist.ac.jp/˜bao/

[5] P. E. Utgoff, Incremental induction of decision trees, www.cs.umass.edu/ũtgoff/papers/mlj-id5r.pdf 1989.

[6] Tutorial: Decision Tree: ID3, Monhash University, 2003, http://www.cs.bham.ac.uk/resources/courses/ai-intro/docs/dt/

[7] Ullas Nambiar, Subbarao Kambhampati, Mining approximate functional dependencies and concept similarities to answer imprecise queries, *Seventh International Workshop on the Web and Database*, Paris, France, June 17-18, 2004.

[8] Lopes Stepane, Pettit, Jean-Marc, and Lakhal, Lotfi, Efficient discovery of functional dependencies and armstrong relations, *Proceeding of ECDT 2000*, Lecture Notes in Computer Science, Vol 1777.

[9] N. Novelli, R. Ciccbetti, Fun: and efficient algorithm for mining functional and embedded dependencies, *Lecture Notes in Computer Science ICDT 2001* (189–203).

[10] Kwok-Wa Lam, Victor C. S. Lee, Building decision trees using functional dependencies, *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC 2004)* (470–473).

Nhận bài ngày 4 - 4 - 2007
Nhận lại sau sửa ngày 6 - 5 - 2007