Tạp chí Tin học và Điều khiển học, T.27, S.4 (2011), 341–352

LUẬT KẾT HỢP THEO CÁCH TIẾP CẬN ĐẠI SỐ GIA TỬ*

TRẦN THÁI SƠN[1], ĐỖ NAM TIẾN[1], PHẠM ĐÌNH PHONG[2]

[1] *Viện Công nghệ thông tin*
[2] *Công ty Prévoir Việt Nam*

Tóm tắt. Bài báo đề xuất phương pháp trích xuất luật kết hợp từ cơ sở dữ liệu theo cách tiếp cận của đại số gia tử. Cách tiếp cận này cho phép chuyển đổi giá trị trong cơ sở dữ liệu sang giá trị mờ một cách tự nhiên hơn và giảm thiểu được khối lượng thông tin cần tính toán.

Abstract. In this paper we propose a new method to quantitative association rule mining from relational database using Hedge Algebras approach. This approach allows to transform values in database into fuzzy values in more natural way and reduce a volume of computational information.

1. MỞ ĐẦU

Khai phá dữ liệu, cụ thể là trích xuất các luật kết hợp từ cơ sở dữ liệu, có xuất phát điểm từ bài toán nghiên cứu số liệu bán hàng trong siêu thị. Ở bài toán này, số liệu được biểu diễn dưới dạng bảng hai chiều, trong đó các cột thể hiện các loại mặt hàng (item), các hàng thể hiện các giao dịch (transactions) đã được tiến hành, số 1 cho thấy mặt hàng được mua, số 0 chỉ điều ngược lại. Từ bảng dữ liệu rất lớn này, người ta mong muốn rút ra được các quy luật giúp cho quản lý, kiểu như "Nếu một người đã mua bánh mỳ và bơ, khả năng người đó cũng mua giăm bông là rất cao". Luật có dạng như vậy gọi là luật kết hợp và là hướng nghiên cứu quan trọng trong lĩnh vực khai phá dữ liệu. Về sau, người ta thấy sẽ là rất không đầy đủ nếu chỉ xem xét các cơ sở dữ liệu chỉ bao gồm các phần tử 0 và 1. Chẳng hạn, trong CSDL nhân sự của một cơ quan có các mục như tuổi, thu nhập.. có giá trị trong miền số thực rất rộng. Để trích xuất ra các luật kết hợp, một phương pháp thường được sử dụng là chuyển số liệu trong CSDL đã cho về CSDL chỉ chứa các giá trị 0, 1 và áp dụng các kết quả đã có. Thí dụ, trong mục "tuổi", có thể chia ra các miền "trẻ", "trung niên" và "già" với các miền giá trị tương ứng là [0,35], [36,55], [56,80] và nếu một giá trị của CSDL ban đầu rơi vào miền giá trị nào thì ta ghi 1 cho vị trí tương ứng trong CSDL chuyển đổi, ngược lại gán giá trị 0. Phương pháp này đơn giản về mặt thực thi nhưng có thể gây băn khoăn do ranh giới cứng mà người ta đưa ra khi tiến hành chuyển đổi. Chẳng hạn hai người tuổi 35 và 36 tuy rất gần nhau về mặt tuổi tác nhưng lại thuộc hai lớp khác nhau là "trẻ" và "trung niên", dẫn tới việc đưa ra những luật

*Bài báo được thực hiện với sự hỗ trợ từ quỹ phát triển KHCNVN (Nafosted), mã số 102.01-2011.06

kết hợp có thể thiếu tính chính xác. Và người ta sử dụng cách tiếp cận mờ để khắc phục điều này, theo đó, một giá trị bất kỳ của CSDL ban đầu không chuyển đổi về giá trị 0 hoặc 1 như trên mà sẽ chuyển về một tập giá trị thực thuộc đoạn [0,1], là độ thuộc của giá trị đã cho vào các tập mờ được xác định trước. Thí dụ, người tuổi 35 trong ví dụ trên, ở CSDL đã chuyển đổi sẽ nhận tập giá trị (trẻ, 0,8), (trung niên, 0,6), (già, 0,1). Phương pháp này, tuy dẫn tới việc xử lý phức tạp hơn nhưng dễ chấp nhận hơn về mặt trực quan và hiện đang được nhiều nhà nghiên cứu quan tâm. Mặc dù vậy, phương pháp trích xuất luật kết hợp mờ vẫn có một số điểm yếu cần khắc phục. Đó là sự phụ thuộc chủ quan rất lớn vào việc lựa chọn các hàm thuộc cho các tập mờ dẫn đến việc xử lý vừa phức tạp vừa có thể thiếu chính xác. Bài báo sẽ đề xuất việc giải bài toán trích xuất luật kết hợp mờ theo cách tiếp cận của Đại số gia tử, ở đó các giá trị độ thuộc mờ sẽ nhận được thông qua các giá trị định lượng ngữ nghĩa, được xác định dựa trên các kết quả nghiên cứu lý thuyết về ĐSGT đã có từ trước.

Bài báo gồm 6 mục, ngoài mục mở đầu và mục kết luận thì mục 2 trình bày một số khái niệm cơ bản. Mục 3 là nội dung lý thuyết cách tiếp cận Đại số gia tử trong trích xuất luật kết hợp mờ. Mục 4 là thuật toán cụ thể. Mục 5 là ví dụ minh họa.

2. MỘT SỐ KHÁI NIỆM CƠ BẢN [11,13]

2.1. Luật kết hợp

Gọi $I = I_1, I_2, ..., I_n$ tập m thuộc tính riêng biệt, mỗi thuộc tính gọi là một mục. Gọi D là một cơ sở dữ liệu, trong đó mỗi bản ghi T là một giao dịch và chứa các tập mục, $T \subseteq I$.

Định nghĩa 2.1. Một luật kết hợp là một quan hệ có dạng $X \Rightarrow Y$, trong đó $X, Y \subset I$ là các tập mục gọi là *itemsets*, và $X \bigcap Y \neq \emptyset$. Ở đây, X được gọi là tiền đề, Y là mệnh đề kết quả. Hai thông số quan trọng của luật kết hợp là *độ hỗ trợ (s)* và *độ tin cậy (c)*.

Định nghĩa 2.2. Độ hỗ trợ (*support*) của luật kết hợp $X \Rightarrow Y$ tỷ lệ phần trăm các bản ghi $X \bigcup Y$ với tổng số các giao dịch có trong cơ sở dữ liệu.

Định nghĩa 2.3. Đối với một số giao dịch được đưa ra, độ tin cậy (confidence) là tỷ lệ của số giao dịch có chứa $X \bigcup Y$ với số giao dịch có chứa X. Đơn vị tính %.

Việc khai thác các luật kết hợp từ cơ sở dữ liệu chính là việc tìm tất cả các luật có độ hỗ trợ và độ tin cậy lớn hơn ngưỡng của độ hỗ trợ và độ tin cậy do người sử dụng xác định trước. Các ngưỡng của độ hỗ trợ và độ tin cậy được ký hiệu là $minsup$ và $mincof$. Tập mục $X \bigcup Y$ được gọi là tập mục lớn nếu $X \Rightarrow Y$ có độ hỗ trợ lớn hơn $minsup$. Một tính chất rất quan trọng đối với khai phá luật kết hợp là "Tập mục con bất kỳ của một tập mục lớn cũng là tập mục lớn". Về mặt cơ bản các thuật toán trích xuất luật kết hợp (Apriori) đều dựa trên ý tưởng: xuất phát từ các tập mục đơn ban đầu, kiểm tra xem mục nào không là tập mục lớn thì loại bỏ, còn lại chập vào nhau và lại tiếp tục kiểm tra cho đến tập mục lớn nhất có thể.

Với luật kết hợp mờ, như đã nói trong phần mở đầu, với mỗi mục có thể chia ra các miền mờ (như "trẻ", "trung niên", ...), thực chất là ta chia một mục ban đầu thành các mục con và giá trị của mỗi hàng tại mục đó sẽ nằm trong [0,1] chứ không chỉ là 0 hoặc 1. Khi đó, độ hỗ

trợ của một miền mờ s_i thuộc x_i được định nghĩa là.

$$FS(A_{s_i}^{x_i}) = \frac{1}{n}\sum_{j=1}^{n}\mu_{s_i}^{x_i}(d_j^{x_i}) \tag{2.1}$$

còn độ hỗ trợ của các miền mờ $s_1, s_2, .., s_k$ của các mục $x_1, x_2, .., x_k$ tương ứng sẽ là.

$$FS(A_{s_1}^{x_1}, A_{s_2}^{x_2}, ..., A_{s_k}^{x_k}) = \frac{1}{n}\sum_{j=1}^{n}[\mu_{s_1}^{x_1}(d_j^{x_1}) \circ \mu_{s_1}^{x_1}(d_j^{x_1}) \circ ... \circ \mu_{s_1}^{x_1}(d_j^{x_1})] \tag{2.2}$$

ở đó x_i là mục thứ i, s_i là miền mờ thuộc mục thứ i, n là số hàng trong CSDL, $\mu_{s_i}^{x_i}(d_j^{x_i})$ là độ thuộc của giá trị tại cột thứ i, hàng j vào tập mờ s_i.

2.2. Đại số gia tử

Ý tưởng sử dụng ĐSGT trong khai phá dữ liệu xuất phát từ cấu trúc khá tốt của các giá trị của một biến ngôn ngữ (như các biến "tuổi", "trình độ chuyên môn", ...). Trong ĐSGT (tuyến tính), các giá trị của biến ngôn ngữ được sắp xếp theo thứ tự (như các giá trị của biến ngôn ngữ "tuổi" là "rất trẻ" < "khá trẻ" < " trẻ" <..<"khá già" < "già" < "rất già"...) và các giá trị này được phân bố một cách có quy luật trên trục số giữa hai giá trị min, max của miền giá trị. Với một vài giả thiết hợp lý (đưa vào như tiên đề trong ĐSGT), mỗi một giá trị biến ngôn ngữ này được gắn với một khoảng lân cận trên trục số và các khoảng của tất cả các giá trị biến ngôn ngữ có cùng độ dài này tạo nên một phân hoạch của đoạn min, max nói trên và ta có thể lấy giá trị đại diện cho khoảng làm giá trị tính toán trong các ứng dụng của khai phá dữ liệu thay cho các giá trị hàm thuộc trong lý thuyết tập mờ của Zadeh. Cụ thể về ĐSGT có thể xem [1-2], [8-9].

Giả thiết đại số gia tử $\underline{AX^*} = (\underline{X^*}$, G, H, $\sigma, \phi, \leq)$ *tuyến tính và đầy đủ, trong đó* $\underline{X^*}$ là tập cơ sở, $G = (0, c^-, W, c^+, 1)$ là tập các phần tử sinh, H là tập các gia tử, $\leq$ là quan hệ thứ tự toàn phần trên $\underline{X^*}$, σ và ϕ là hai phép toán mở rộng sao cho $\forall x \in \underline{X^*}$ của tập $H(x)$ là tập tất cả các phần tử sinh ra từ x nhờ tác động của gia tử trong H.

Định nghĩa 2.4.

1. Gia tử h được gọi là dương nếu $hc^+ > c^+$(hay $hc^- < c^-$) và là âm nếu có điều ngược lại. Tập các gia tử dương được kí hiệu là H^+ còn tập các gia tử âm kí hiệu là H^-. $H = H^+ \bigcup H^-$.

2. Gia tử h được gọi là dương (âm) đối với gia tử k khi và chỉ khi $\exists x \in Dom(X)$ sao cho nếu $x < kx$ thì $kx < hkx (kx > hkx)$ hoặc $x > kx$ thì $kx > hkx (kx < hkx)$.

Tính chất 2.1. Tính chất dương (âm) của một gia tử này đối với một gia tử khác không phụ thuộc vào phần tử x mà chúng tác động.

Tính chất 2.2. Nếu $hx < kx$ thì $\forall$p,q$\in$H ta có $phx < qkx$ hay $H(hx) < H(kx)$

Định nghĩa 2.5. Một ánh xạ f được gọi là ánh xạ định lượng ngữ nghĩa của X nếu nó thỏa mãn các điều kiện sau:

Q1) f là song ánh;

Q2) f bảo toàn thứ tự trên $\underline{X^*}$, tức là $x < y \Rightarrow f(x) < f(y)$, và $f(0) = 0, f(1) = 1$;

Q3) Tính chất liên tục: $\forall x \in \underline{X^*}$, *f(ϕx) = infimumf(H(x))* và *f(σx) = supremumf(H(x))*.

Có thể hình dung $H(x)$ bao gồm các khái niệm mờ mà nó phản ánh ý nghĩa nào đó của khái niệm x. Vì vậy, kích thước của tập $H(x)$ có thể biểu diễn tính mờ của x. Nhờ ánh xạ ngữ nghĩa f, tập $H(x)$ (hay độ đo tính mờ của x) có thể mô phỏng định lượng bằng độ dài của tập $f(H(x))$ và kí hiệu là $fm(x)$. Với ý tưởng này độ đo tính mờ được tiên đề hoá qua định nghĩa 2.6.

Định nghĩa 2.6. Một hàm $fm : \underline{X^*} \to [0, 1]$, được gọi là một độ đo tính mờ của biến ngôn ngữ X, nếu nó có các tính chất sau:

F1) fm là một độ đo đầy đủ trên $\underline{X^*}$ nghĩa là $fm(c^-) + fm(c^+) = 1$ và, $\forall u \in \underline{X^*}$, $\Sigma_{h \in H} fm(hu) = fm(u)$;

F2) Nếu x là một khái niệm chính xác, tức là H(x) = x, thì fm(x) = 0. Đặc biệt ta có: $fm(0) = fm(W) = fm(1) = 0$;

F3) $\forall x, y \in \underline{X^*}$, $\forall h \in H$, ta có $\frac{fm(hx)}{fm(x)} = \frac{fm(hy)}{fm(y)}$, nghĩa là tỷ số này không phụ thuộc vào một phần tử cụ thể nào trong $\underline{X^*}$ mà chỉ phụ thuộc vào h, do đó ta có thể ký hiệu nó bằng $\mu(h)$ và được gọi là độ đo tính mờ của gia tử h.

Giả sử rằng $H^- = h_{-1}, ..., h_{-q}$, với $h_{-1} < h_{-2} < ... < h_{-q}$, và $H^+ = h_1, ..., h_p$, với $h_1 < ... < h_p$.

Từ Định nghĩa 2.5 ta thấy fm có các tính chất sau.

Mệnh đề 2.1.[8] Độ đo tính mờ fm của các khái niệm và $\mu(h)$ của các gia tử thỏa mãn các tính chất sau:

1) $fm(hx) = \mu(h) fm(x), \forall x \in \underline{X^*}$;

2) $fm(c^-) + fm(c^+) = 1$;

3) $\sum\limits_{i=-q, i\neq 0} fm(h_i c) = fm(c)$, với $c \in c^-, c^+$;

4) $\sum\limits_{-q \leq i \leq p, i \neq 0} fm(h_i x) = fm(x)$, $x \in \underline{X}$.

5) $\sum\limits_{i=-1}^{-q} \mu(h_i) = \alpha$ và $\sum\limits_{i=1}^{p} \mu(h_i) = \beta$, với $\alpha, \beta > 0$ và $\alpha + \beta = 1$.

Ánh xạ định lượng ngữ nghĩa được xây dựng dựa trên các tham số cho trước gồm các độ đo tính mờ của các phần tử sinh $fm(c^-)$, $fm(c^+)$ và độ đo tính mờ của các gia tử $\mu(h)$.

Định nghĩa 2.7. (Hàm Sign [8]). Hàm dấu $Sign : X \to -1, 0, 1$ là ánh xạ được định nghĩa đệ quy như sau, trong đó h và h' là các gia tử bất kỳ và $c \in \{c^-, c^+\}$:

a) $Sign(c^-) = -1, Sign(c^+) = +1$,

b) $Sign(h'hx) = -Sign(hx)$ nếu $h'hx \neq hx$ và h' là âm tính đối với h (hoặc đối với c, khi $h = I$ và $x = c$);

c) $Sign(h'hx) = Sign(hx)$ nếu $h'hx \neq hx$ và h' là dương tính đối với h (hoặc đối với c, khi $h = I$ và $x = c$);

d) $Sign(h'hx) = 0$ nếu $h'hx = hx$.

Hàm dấu Sign được đưa ra để sử dụng nhận biết khi nào gia tử tác động vào các từ làm tăng hay giảm ngữ nghĩa định lượng. Ta có khẳng định sau:

Bổ đề 2.1.[8] Với mọi h và x, nếu $Sign(hx) = +1$ thì $hx > x$, nếu $Sign(hx) = -1$ thì $hx < x$

Định nghĩa 2.8. Cho $\underline{AX^*}$ là đại số gia tử tuyến tính, đầy đủ và tự do, $fm(c^-)$ và $fm(c^+)$ là các độ đo tính mờ của phần tử sinh c^-, c^+ và $\mu(h)$ là độ đo tính mờ của các gia tử h trong H thỏa mãn các tính chất trong Mệnh đề 2.1. Ánh xạ định lượng ngữ nghĩa mờ là ánh xạ ν được xác định quy nạp như sau:

1) $\nu(W) = \theta = fm(c^-), \nu(c^-) = \theta - \alpha fm(c^-), \nu(c^+) = \theta + \alpha fm(c^+)$;

2) $\nu(h_j x) = \nu(x) + Sign(h_j x)(\sum_{i=1}^{j} fm(h_i x) - \omega(h_j x))$, với $1 \leq j \leq p$,

và $\nu(h_j x) = \nu(x) + Sign(h_j x)(\sum_{i=-1}^{j} fm(h_i x) - \omega(h_j x))$, với $-q \leq j \leq -1$,

Hai công thức này có thể viết thành một công thức chung, với $j \in [-q \wedge p] = \{j : -q \leq j \leq p \ \& \ j \neq 0\}$ là: $v(h_j x) = v(x) + Sign(h_j x)(\sum_{i=sign(j)}^{j} fm(h_i x) - \omega(h_j x) fm(h_j x))$, trong đó $fm(h_j x)$ được tính theo tính chất 1) Mệnh đề 2.1 và $\omega(h_j x) = \frac{1}{2}[1 + Sign(h_j x) Sign(h_p h_j x)(\beta - \alpha)] \in \{\alpha, \beta\}$.

3) $\nu(\phi c^-) = 0, \nu(\sigma c^-) = \theta = \nu(\phi c^+), \nu(\sigma c^+) = 1$, và với các phần tử dạng $h_j x$, $j \in [-q \wedge p]$, ta có:

$$\nu(\phi h_j x) = \nu(x) + Sign(h_j x)(\sum_{i=Sign(j)}^{j-1} fm(h_i x))$$

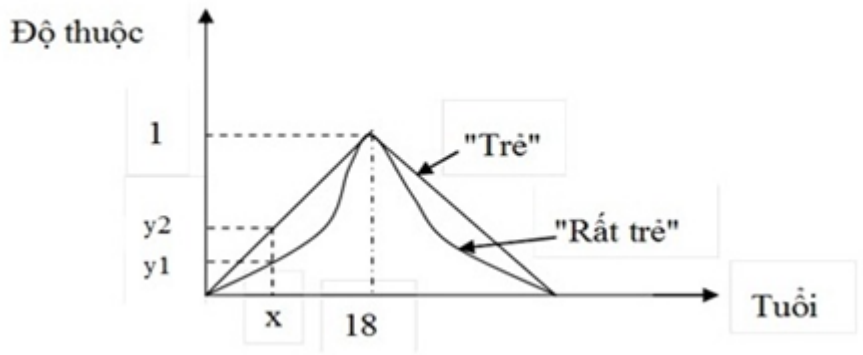
$$\nu(\sigma h_j x) = \nu(x) + Sign(h_j x)(\sum_{i=Sign(j)}^{j} fm(h_i x)).$$

Dễ dàng kiểm chứng: $\nu(c^-) = \beta fm(c^-)$ và $\nu(c^+) = 1 - \beta fm(c^+)$.

3. TIẾP CẬN ĐẠI SỐ GIA TỬ TRONG KHAI PHÁ DỮ LIỆU

Theo cách tiếp cận này, trước hết các mục cũng được phân chia thành các miền mờ, không phải bằng cách mang nhiều yếu tố chủ quan như trong lý thuyết tập mờ thông thường mà bằng cách ứng chúng với các ĐSGT. Chẳng hạn, trong mục "Tuổi", ta có thể xác định một ĐSGT với hai phần tử sinh là "trẻ" và "già" cùng các gia tử "rất", "khá" và "tương đối", từ đó có các miền mờ (là các phần tử của ĐSGT có độ dài 1) là "rất trẻ", "khá trẻ", "tương đối trẻ", "trung niên" (là giá trị trung gian W), "rất già", "khá già", "tương đối già". (nếu muốn, có thể xét các phần tử có độ dài 2 như "tương đối khá già", "rất rất trẻ", ...). Do các khoảng

mờ của các phần tử có cùng độ dài của ĐSGT tạo nên một tựa phân hoạch trên miền giá trị của ĐSGT nên các miền ở đây phủ kín miền giá trị của biến ngôn ngữ. Tiếp theo, thay vì việc xây dựng các hàm thuộc cho các miền mờ đã xác định, ta sử dụng giá trị định lượng ngữ nghĩa để xác định độ gần gũi (hay độ thuộc) của các giá trị tại hàng bất kỳ của mục đang xét đến các miền mờ mới xác định ở trên. Cụ thể, khoảng cách trên trục số giữa $(d_i^{x_i})$ và giá trị định lượng ngữ nghĩa của hai phần tử gần $(d_i^{x_i})$ nhất về hai phía có thể dùng để xác định độ gần gũi của $(d_i^{x_i})$ vào hai miền mờ (hai phần tử của ĐSGT) đó. Độ gần gũi giữa $(d_i^{x_i})$ với các phần tử khác của ĐSGT được xác định bằng 0. Để xác định độ thuộc cuối cùng, ta phải chuẩn hóa (chuyển về giá trị trong đoạn [0,1] rồi lấy nghịch đảo hoặc 1-khoảng cách đã chuẩn hóa đó). Ta sẽ có, ứng với mỗi giá trị $(d_i^{x_i})$ một cặp độ thuộc (thay vì có thể 2, 3 hay hơn giá trị độ thuộc trong cách tiếp cận tập mờ cổ điển) dùng để làm đầu vào trong thuật toán sẽ trình bày trong phần sau. Có thể thấy cách tính giá trị hàm thuộc như nêu trên đây là khá tự nhiên do ta đã có phân bố các giá trị định lượng ngữ nghĩa của các giá trị ngôn ngữ trên trục số theo một thứ tự xác định. Ngoài ra, thay vì có thể sử dụng đến ba hoặc nhiều hơn giá trị độ thuộc để tính toán như trong các thuật toán khai phá luật kết hợp mờ khác (xem [11,12]), ở đây, do phân bố thứ tự đã được xác định, ta thấy chỉ cần hai giá trị độ thuộc (vào các tập mờ gần nhất về hai phía) là đã phản ánh tốt thông tin về giá trị của tập mục đang xét (trên bản ghi hiện thời) và như vậy sẽ tiết kiệm đáng kể khối lượng tính toán cũng như bộ nhớ cần thiết. Nên nhớ là số lượng bản ghi tạo ra khi chuyển đổi số liệu là tăng theo cấp số nhân của số miền mờ tương ứng được tạo ra. Chẳng hạn, nếu có 8 mục, mỗi mục chia làm 3 miền mờ thì số lượng bản ghi mới tạo ra sẽ là 3^8, trong khi theo phương pháp ĐSGT, số bản ghi mới tạo ra chỉ là $<2^8$ (vì có những giá trị đầu mút chỉ nhận 1 giá trị độ thuộc). Có thể nói kỹ thêm một chút về việc dùng khoảng cách giữa các giá trị định lượng ngữ nghĩa để tính độ thuộc thay cho việc đưa ra hàm thuộc như lý thuyết tập mờ của Zadeh. Theo [14], nếu tập mờ "trẻ" được đặc trưng bởi hàm thuộc, thí dụ, trong hình vẽ sau, đồ thị của hàm thuộc $\mu_{'tre'}$ là tam giác cân, đỉnh là (18,1), hai đáy là (0,0) và (0,36) thì hàm thuộc của rất trẻ sẽ là $\mu^2_{'tre'}$ thể hiện bằng hàm bậc hai trong hình vẽ, cũng có đỉnh tọa độ là (18,1)



Khi đó, với một giá trị tuổi x bất kỳ (nhỏ hơn 18), ta sẽ có giá trị độ thuộc của x vào hai tập mờ "trẻ" và "rất trẻ" tương ứng sẽ là $y2$ và $y1$, trong đó $y1 < y2$. Điều đó có nghĩa độ thuộc của x vào tập "trẻ" luôn lớn hơn độ thuộc của x vào tập "rất trẻ", hay nói cách khác, x luôn được coi là "trẻ" nhiều hơn là "rất trẻ" dù x có ít tuổi bao nhiêu đi nữa, mâu thuẫn với suy nghĩ thông thường của con người. Mặt khác, nếu không sử dụng ý tưởng biến ngôn ngữ của Zadeh, thì thứ nhất, mỗi giá trị biến ngôn ngữ buộc ta phải tạo ra một hàm thuộc mới (có thể là một quá trình phức tạp) mà nhiều khi chẳng liên quan gì đến hàm thuộc đã có mặc dù về mặt ngữ nghĩa chúng có thể liên quan chặt chẽ với nhau (như "trẻ" và "rất trẻ" hoặc "tương đối khá trẻ").

Vì thế, ta có thể đi đến thuật toán trích xuất luật kết hợp cụ thể sau.

4. THUẬT TOÁN TRÍCH XUẤT LUẬT KẾT HỢP TỪ CƠ SỞ DỮ LIỆU

Ký hiệu các tham số của thuật toán như sau:

n: Tổng số giao dịch trong cơ sở dữ liệu (được sinh ra sau quá trình chuyển dữ liệu thô (n') thành nhãn gia tử tương ứng)

m: Tổng số các thuộc tính (số thuộc tính của dữ liệu thô (m')* số nhãn gia tử)

A_j: Thuộc tính thứ j, $1 \leq j \leq m, A_j$

$D(i)$ dữ liệu giao dịch thứ i, $1 \leq i \leq n$

$v_j(i)$: Giá trị định lượng của A_j trong $D(i)$;

$f_{jk}(i)$ giá trị độ thuộc của $v_j(i)$ với nhãn gia tử $R_{jk}, 1 \leq f_{jk}(i) \leq m$;

$Sup(A_{jk})$: Độ hỗ trợ của A_{jk}

Sup: Giá trị hỗ trợ của mỗi tập mục lớn;

$Conf$: Độ tin cậy của mỗi tập mục lớn

$Minsup$: Giá trị hỗ trợ tối thiểu cho trước

$Minconf$: Giá trị tin cậy cho trước

C_r: Tập các tập mục có khả năng với r thuộc tính (tập mục), $1 \leq r \leq m$;

L_r: Tập các tập mục lớn thỏa mãn với r nhãn gia tử (tập mục) $1 \leq r \leq m$;

Thuật toán khai phá dữ liệu dựa trên đại số gia tử cho các giá trị định lượng được thực hiện như sau:

Input: m', n', *theta* **(tỉ lệ giữa 2 phần tử sinh),** *minsup* **và** *minconf*

Output: luật kết hợp

Bước 1: Chuyển các giá trị định lượng $v_j(i)$ của mỗi giao dịch $D(i)$, i từ 1 tới n, với mỗi thuộc tính A_j, nếu A_j nằm ở ngoài 1 trong 2 đầu mút (2 nhãn gia tử cực đại và cực tiểu) thì A_j chỉ có 1 nhãn gia tử ứng với đầu mút đó, nếu không thì A_j được biểu diễn bởi 2 nhãn gia tử liên tiếp có đoạn giá trị nhỏ nhất trên trường giá trị của A_j, mỗi nhãn ứng với 1 giá

trị biểu diễn độ thuộc $f_{jk}(i)(j = 1, 2)$ của A_j với nhãn gia tử đó. Độ thuộc này được tính là khoảng cách của A_j tới giá trị biểu diễn cho nhãn gia tử tương ứng.

Bước 2: Tính giá trị hỗ trợ

$$Sup(R_{jk}) = \frac{\sum_{j=1}^{n} f_{jk}}{n} \tag{4.3}$$

Bước 3: Nếu $Sup(R_{jk}) \geq minsup$ thì đưa R_{jk} vào L_1

Bước 4: Nếu L_1 không rỗng, tiếp tục bước sau, nếu rỗng thoát chương trình.

Bước 5: Thuật toán xây dựng tập mục lớn mức r từ các tập mục lớn mức $r-1$ bằng cách chọn 2 tập mục lớn mức $r-1$ chỉ khác nhau duy nhất một mục, hợp 2 tập mục này ta được tập mục ứng viên C_r, nếu tập mục này xuất hiện trong cơ sở dữ liệu và có giá trị hỗ trợ thỏa mãn thì nó được đưa vào danh sách các tập mục lớn mức r.

Bước 6: Thực hiện theo các bước con sau đây lặp lại cho các tập mục lớn mức lớn hơn được sinh ra tiếp theo dạng $(r+1)$ tập mục lớn S với mục $(s_1, s_2, ..., s_t, ..., s_{r+1})$ trong C_{r+1}, $1 \leq t \leq r+1$

(a)Tính giá trị hỗ trợ $sup(S)$ của S trong giao dịch

$$Sup(S) = \frac{\sum_{j=1}^{k} f_{jk}}{n} \tag{4.4}$$

(b)Nếu $Sup(S) \geq minsup$, thì đưa S vào L_{r+1}.

Bước 7: Nếu L_{r+1} là rỗng, thì thực hiện bước tiếp theo, ngược lại, đặt $r = r+1$, thực hiện lại bước 5 và 6.

Bước 8: Thu thập các tập mục lớn mức lớn hơn nếu có.

Bước 9: Đưa ra các luật kết hợp từ các tập mục lớn vừa thu thập theo cách sau:

(a)Với mỗi luật kết hợp khả thi sau đây: $s_1 \bigcap ... \bigcap s_x \bigcap s_y \bigcap ... \bigcap s_q \rightarrow s_k (k = \overline{1,q}, x = k-1, y = k+1)$

(b)Tính độ tin cậy của luật: $Conf(s_1 \bigcap ... \bigcap s_x \bigcap s_y \bigcap ... \bigcap s_q \rightarrow s_k) = \frac{Sup(S/s_k)}{Sup(S)}$

5. VÍ DỤ MINH HỌA

Ta có kết quả thu được của việc áp dụng thuật toán trên với số liệu lấy từ CSDLFAM95, số liệu điều tra dân số Mỹ năm 1995 (http://www.stat.ucla.edu /data/fpp). Ở đây chỉ liệt kê một số luật tiêu biểu để so sánh với các kết quả trong [12].

Bảng a

Conf.	Rules	Supp.

0,983	$oldage \rightarrow fewchildren$	0,234
0,967	$fewpersons \rightarrow fewchildren$	0,520
0,897	$feweducation \rightarrow lowhincome$	0,143
0,884	$lowincome \rightarrow lowhincome$	0,500
0,850	$loweducation \rightarrow lowincome$	0,136
0,837	$lowhincome \rightarrow lowincome$	0,500
0,808	$loweducation \rightarrow fewchildren$	0,129
0,801	$manychildren \rightarrow manypersons$	0,060
0,798	$oldage \rightarrow lowhincome$	0,190
0,790	$oldage \rightarrow fewpersons$	0,188
0,784	$lowincome \rightarrow fewchildren$	0,444
0,768	$lowhincome \rightarrow fewchildren$	0,459
0,766	$highhincome \rightarrow highincome$	0,050
0,756	$oldage \rightarrow lowincome$	0,180
0,743	$averchildren \rightarrow averepersons$	0,151
0,728	$fewpersons \rightarrow lowincome$	0,391
0,724	$higheducation \rightarrow fewchildren$	0,197
0,720	$fewchildren \rightarrow fewpersons$	0,520
0,719	$medhincome \rightarrow medincome$	0,243
0,719	$mededucation \rightarrow fewchildren$	0,409

Bảng b1

Supp.	Rules	Conf.
0,219	$low\_EdLelow\_InFa \rightarrow low\_InHe$	91,77
0,149	$high\_Age \rightarrow low\_NuKi$	90,26
0,219	$low\_EdLelow\_InHe \rightarrow low\_InFa$	90,19
0,242	$low\_EdLe \rightarrow low\_InHe$	89,51
0,096	$high\_Agelow\_FaPe \rightarrow low\_NuKi$	89,5
0,083	$low\_Agemedium\_EdLe \rightarrow low\_InHe$	88,86
0,058	$high\_Agemedium\_EdLe \rightarrow low\_NuKi$	88,77
0,083	$low\_Agemedium\_EdLe \rightarrow low\_InFa$	88,0
0,238	$low\_EdLe \rightarrow low\_InFa$	87,97
0,117	$high\_Agelow\_InFa \rightarrow low\_NuKi$	87,54
0,119	$high\_Agelow\_InHe \rightarrow low\_NuKi$	87,43
0,18	$low\_Agelow\_InFa \rightarrow low\_InHe$	87,15
0,661	$low\_InFa \rightarrow low\_InHe$	86,93

Vì tôn trọng nguyên bản, ta giữ lại ở đây những từ ngữ mà các tác giả trong [12] dùng ở Bảng (a) (trong đó med là viết tắt của medium; aver là viết tắt của average). Còn ở Bảng (b1) các từ viết tắt là ve: very; qu:quite; hi: high; lo: low; Nuki: number of Kids; FaPe: Persons in a Family; InFa: Income of a family; InHe: Income of family's head; EdLe: Level of Education. Kết quả này có được khi ta sử dụng ĐSGT có hai phần tử sinh (là high và low)

cùng phần tử trung hòa medium với phân bố đều trên miền giá trị ($fm(c^-)$=$fm(c^+)$=0,5) và không sử dụng gia tử. Ta thấy, khi so sánh hai bảng:

1. Các luật cơ bản của (a) đều có trong (b1) với độ tin cậy xấp xỉ. Chẳng hạn, luật "old age → few children" với độ tin cậy 0,983 ở Bảng (a) ứng với luật $high\_Age \rightarrow low\_NuKi$ với độ tin cậy 90,26 (tức 0,902)ở Bảng (b1) hay luật $fewpersons \rightarrow fewchildren$ với độ tin cậy 0,967 ứng với luật $low\_FaPe \rightarrow low\_NuK$ với độ tin cậy 78,05, luật $feweducation \rightarrow lowhincome$ với độ tin cậy 0,897 ứng với luật $low\_EdLe \rightarrow low\_InHe$ với độ tin cậy 89,51,...

2. Độ tin cậy ở hai bảng có khác nhau do sử dụng về thực chất các hàm thuộc khác nhau nhưng đều ở trong ngưỡng cao giống nhau. Trong Bảng (b1) có một số luật mà Bảng (a) không có, thí dụ luật $low\_EdLelow_InFa \rightarrow low_InHe$ do các tác giả trong [12] chỉ xét các luật vế trái có một mục.

3. Ta có thể trích xuất các luật chi tiết hơn một cách dễ dàng như trong Bảng (b2) sau vì theo tiếp cận của DSGT, việc sinh các phần tử ngôn ngữ có thể tính toán dễ dàng (theo như Mệnh đề 2.1 và Định nghĩa 2.8 đã nêu) chứ không cần phải sinh ra các hàm thuộc mới theo lý thuyết tập mờ cổ điển (dễ dẫn đến sai sót như thí dụ về hàm thuộc "trẻ" và "rất trẻ" đã nói đến trong Mục 3 của bài báo). Ở đây $fm(c^-) = fm(c^+) = 0.5$ và $\mu(very)$=0,875, $\mu(quite)$=0,125. (Bảng b2)

Bảng b2

Supp.	Rules	Conf.
78,03	$qu\_hi\_Age \rightarrow ve\_lo\_NuKi$	0,13105318
78,1	$ve\_hi\_Age \rightarrow ve\_lo\_NuKi$	0,082249284
69,13	$ve\_lo\_InFa \rightarrow ve\_lo\_InHe$	0,43760595
68,99	$ve\_lo\_InHe \rightarrow ve\_lo\_InFa$	0,43760595
68,68	$ve\_lo\_Age \rightarrow ve\_lo\_InFa$	0,09446891
68,76	$ve\_hi\_Age \rightarrow ve\_lo\_InHe$	0,064412944
66,98	$ve\_hi\_Age \rightarrow ve\_lo\_InFa$	0,06274383
66,49	$ve\_lo\_FaPe \rightarrow ve\_lo\_InFa$	0,2789387
63,97	$ve\_lo\_NuKi \rightarrow ve\_lo\_InFa$	0,39723718
63,86	$qu\_lo\_Age \rightarrow ve\_lo\_InFa$	0,18667413
64,17	$ve\_lo\_NuKi \rightarrow ve\_lo\_InHe$	0,39849037
65,04	$ve\_lo\_FaPe \rightarrow ve\_lo\_InHe$	0,27287552
67,76	$ve\_lo\_Age \rightarrow ve\_lo\_InHe$	0,093199834
62,9	$middle\_NuKi \rightarrow ve\_lo\_InFa$	0,010170757
62,75	$ve\_lo\_InFa \rightarrow ve\_lo\_NuKi$	0,39723718
62,82	$ve\_lo\_InHe \rightarrow ve\_lo\_NuKi$	0,39849037
62,54	$qu\_lo\_FaPe \rightarrow ve\_lo\_InHe$	0,23233615
63,0	$qu\_lo\_Age \rightarrow ve\_lo\_InHe$	0,18472897
62,37	$qu\_hi\_EdLe \rightarrow ve\_lo\_NuKi$	0,18516514
61,92	$middle\_NuKi \rightarrow ve\_lo\_InHe$	0,010012479

62,17	$qu\_hi\_Age \rightarrow ve\_lo\_InHe$	0,104422025
61,87	$qu\_lo\_NuKi \rightarrow ve\_lo\_InFa$	0,13953367

6. KẾT LUẬN

Bài báo đã đưa ra cách tiếp cận ĐSGT cho bài toán trích xuất luật kết hợp mờ từ CSDL mà theo quan điểm của chúng tôi có 2 ưu điểm nổi trội:

1. Sử dụng ĐSGT có thể xác định giá trị thuộc của phần tử trong CSDL một cách tự nhiên và đơn giản hơn so với cách tiếp cận của lý thuyết tập mờ cổ điển.

2. Khối lượng tính toán sẽ giảm một cách đáng kể trong khi vẫn đạt được kết quả tương đương, chưa kể nếu cần ta có thể trích xuất ra các luật mang tính chi tiết hơn.

TÀI LIỆU THAM KHẢO

[1] N. Cat Ho, Fuzziness in Structure of Linguistic Truth Values: A Foundation for Development of Fuzzy Reasoning, *Proc. of ISMVL 87*, Boston, USA, (IEEE Computer Society Press, New York), 1987 (326–335).

[2] N. Cat Ho and W. Wechler, Hedge algebras: an algebraic approach to structure of sets of linguistic truth values, *Fuzzy Sets and Systems* **35** (1990) 281–293.

[3] N. Cat Ho and W. Wechler, Extended hedge algebras and their application to Fuzzy logic, *Fuzzy Sets and Systems* **52** (1992) 259–281.

[4] Nguyễn Cát Hồ, Trần Thái Sơn, Về khoảng cách giữa các giá trị của biến ngôn ngữ trong Đại số gia tử và bài toán sắp xếp mờ, *Tạp chí Tin học và Điều khiển học* **11** (1) (1995) 10–20

[5] Nguyễn Cát Hồ, Trần Thái Sơn, Logic mờ và quyết định mờ dựa trên cấu trúc thứ tự của giá trị ngôn ngữ, *Tạp chí Tin học và Điều khiển học* **9** (4) (1993) 1–9.

[6] N. Cát Hồ, H.Văn Nam, T.D Khang and L.H. Chau, Hedge Algebras, Linguistic- valued Logic and their Application to Fuzzy Reasoning, Inter. J. of Uncertainty, *Fuzziness and Knowledge-Based System* **7** (4) (1999) 347–361.

[7] Trần Thái Sơn, Lập luận xấp xỉ với giá trị của biến ngôn ngữ, *Tạp chí Tin học và Điều khiển học* **15** (2) (1999) 6-10.

[8] Nguyen Cat Ho, Tran Thai Son, Tran Dinh Khang, Le Xuan Viet, Fuzziness Measure, Quantified Semantic Mapping And Interpolative Method of Approximate Reasoning in Medical Expert Systems, *Tạp chí tin học và điều khiển* **18** (3)(2002) 237-252.

[9] N.C. Hồ, N.V. Long, Đại số gia tử đầy đủ tuyến tính, *Tạp chí Tin học và Điều khiển học* **19** (3)(2003) 274-280.

[10] [10] N.C. Hồ, N.V. Long, Cơ sở toán học của độ đo tính mờ của thông tin ngôn ngữ, *Tạp chí Tin học và Điều khiển học* **20** (1) 64-72.

[11] R.Srikant and R.Agrawal, Mining quantitative association rules in larse relational tables, *The 1996 ACM SIGMOD International Conference on Managment of Data Montreal Canada*, June 1996 (1-12).

[12] Hannes Verlinda, Martine De Cock and Raymond Buote, Fuzzy Versus Quantiative Association Rules: A Fair Data-Driven Comparison, *IEEE Transactions on SMC*, **36** (3) (June 2006) 679-684.

[13] David L. Olson and Yanhong Li, Mining Fuzzy Weighted Association Rules, *Proccedings of the 40th Hawai International Conference on System Sciences*, 2007 (1-9).

[14] L. A. Zadeh, The concept of linguistic variable and its application to approximate reasoning., *Inform. Sci.* **8** (I) (1975) 199–249; **8** (II) (1975) 310–357; **9** (III) (1975) 43–80.

Nhận bài ngày 25 - 02 - 2011

Nhận lại sau sửa ngày 25 - 4 - 2011