Tạp chí Tin học và Điều khiển học, T.22, S.3 (2006), 244–256

PHÁT HIỆN LUẬT KẾT HỢP MỜ CÓ ĐỘ HỖ TRỢ CỰC TIỂU KHÔNG GIỐNG NHAU

ĐỖ VĂN THÀNH

Bộ Kế hoạch và Đầu tư

Abstract. Mining Association Rules from transaction databases with unequal minimum supports is a problem proposed and reseached by the author [3]. The algorithm for mining closed frequent itemsets with unequal minimum supports of each item in transaction databases was called CHARM-NEW. This algorithm was indeed improved and developed from the CHARM which is one of the most efficient algorithms for mining closed frequent itemsets with the same minimum support from transaction databases.

The goal of this paper is to propose and to find out measures for mining fuzzy association rules from quantitative databases with unequal minimum supports. The paper will concentrate on developing an algorithm for mining closed fuzzy frequent itemsets with unequal minimum supports of each attribute in quantitative databases.

Tóm tắt. Phát hiện luật kết hợp từ các cơ sở dữ liệu tác vụ với độ hỗ trợ cực tiểu không giống nhau là vấn đề được tác giả đề xuất và nghiên cứu ([3]). Thuật toán phát hiện các tập phổ biến đóng với độ hỗ trợ cực tiểu không giống nhau của mỗi tập mục dữ liệu trong các cơ sở dữ liệu tác vụ được gọi là CHARM-NEW. Thật ra thuật toán này được cải tiến và phát triển từ thuật toán CHARM, đó là một trong những thuật toán hiệu quả nhất để tìm tập phổ biến đóng với độ hỗ trợ cực tiểu như nhau từ các cơ sở dữ liệu tác vụ.

Mục đích của bài báo này là đề xuất và tìm kiếm giải pháp để phát hiện các luật kết hợp mờ từ các cơ sở dữ liệu định lượng với độ hỗ trợ cực tiểu không giống nhau. Bài báo sẽ tập trung phát triển thuật toán phát hiện tập phổ biến mờ đóng với độ hỗ trợ cực tiểu không giống nhau của mỗi tập mục dữ liệu trong các cơ sở dữ liệu định lượng.

1. GIỚI THIỆU

Quá trình phát hiện luật kết hợp được chia thành hai giai đoạn. Mục đích của giai đoạn đầu là tìm các tập phổ biến có độ hỗ trợ lớn hơn hoặc bằng một giá trị chung nào đó (gọi là độ hỗ trợ cực tiểu, ký hiệu là minSupp), còn của giai đoạn 2 là tìm các luật kết hợp từ các tập tìm được ở giai đoạn 1 và có độ tin cậy lớn hơn hoặc bằng một giá trị chung khác (gọi là độ tin cậy cực tiểu, ký hiệu minConf). Trong quá trình đó, giai đoạn tìm các tập phổ biến là phức tạp và tốn nhiều chi phí nhất.

Những năm qua người ta đã tập trung nghiên cứu và đề xuất được nhiều thuật toán tìm tập phổ biến hiệu quả từ các cơ sở dữ liệu (CSDL) tác vụ (hay nhị phân) theo nhiều cách tiếp cận khác nhau [1, 9, 15]. Những thuật toán mới và hiệu quả nhất về vấn đề đó cho đến nay là những thuật toán chỉ cần tìm các tập phổ biến đóng [3, 9, 14, 15] nhờ chứng minh được

rằng các luật kết hợp được sinh ra từ các tập phổ biến đóng và từ các tập phổ biến là như nhau, trong khi không gian các tập phổ biến đóng là nhỏ hơn rất nhiều so với không gian các tập phổ biến.

Tuy nhiên các thuật toán trên đều được xây dựng dựa trên thừa nhận minSupp của các tập phổ biến là như nhau. Một số hạn chế của luật kết hợp được tìm từ các tập phổ biến có độ hỗ trợ cực tiểu như nhau đã được chỉ ra trong [2–8, 11–13]. Hiện có bốn cách tiếp cận để khắc phục những hạn chế của việc tìm tập phổ biến có độ hỗ trợ cực tiểu chung giống nhau [2–8, 11–13].

Thứ nhất là: tìm các tập phổ biến trong mối quan hệ có sự ràng buộc về độ hỗ trợ ([11, 12]) bằng cách đề xuất mô hình biểu diễn ràng buộc độ hỗ trợ của các tập mục dữ liệu (gọi là cách tiếp cận *ràng buộc độ hỗ trợ*). Cách tiếp cận này có nhược điểm là tạo ra nhiều phức tạp với người sử dụng, đó là đòi hỏi họ phải có kiến thức cơ sở nhất định trong lĩnh vực ứng dụng ([3]).

Thứ hai là: gắn trọng số vào mỗi mục dữ liệu để đo vai trò quan trọng của nó và áp dụng có cải tiến một trong các thuật toán tìm tập phổ biến đã có để tìm các tập phổ biến có gắn trọng số [2, 13] (gọi là cách tiếp cận trọng số). Nhược điểm lớn nhất của cách tiếp cận này là không đảm bảo được tính chất *tập con của tập phổ biến là tập phổ biến* ([1]) mà trong nhiều trường hợp ứng dụng, tính chất này gần như là một đòi hỏi tất nhiên, chẳng hạn trong lĩnh vực thương mại, nếu một nhóm mặt hàng đã được nhiều người mua thì một số mặt hàng thuộc nhóm đó cũng phải được xem là như vậy.

Thứ ba là: tìm các tập phổ biến theo độ hỗ trợ cực tiểu khác nhau tuỳ thuộc vào từng mức khái niệm của các tập mục dữ liệu ([5, 8]) (gọi là cách tiếp cận nhiều mức hoặc phân bậc). Cách tiếp cận này khá thích hợp với những thuật toán tìm tập phổ biến theo chiều rộng của đồ thị biểu diễn không gian tìm kiếm của các tập mục dữ liệu theo kiểu như thuật toán Apriori [1, 2], đó là những thuật toán tìm k- tập phổ biến bằng cách kết nối 2 tập $(k-1)$-tập phổ biến ở mức trên đó. Cách tiếp cận này có nhược điểm chính, khó vượt qua, là bằng cách nào để xác định được *một cách hợp lý* độ hỗ trợ cực tiểu *cho từng mức.*

Thứ tư là: cách tiếp cận được đề xuất trong [3] (gọi là cách tiếp cận độ hỗ trợ). Ở đó vai trò quan trọng của các tập mục dữ liệu được đo bằng độ hỗ trợ cực tiểu, xem các tập mục dữ liệu khác nhau là có độ hỗ trợ cực tiểu khác nhau. Các cách tiếp cận độ hỗ trợ và theo trọng số được đề xuất trong [2, 13] có vẻ giống nhau vì trước tiên chúng cùng đo tầm quan trọng của mục dữ liệu bằng độ hỗ trợ cực tiểu hoặc bằng trọng số nhưng về bản chất chúng khác nhau do theo cách tiếp cận độ hỗ trợ thì các tập phổ biến được tìm theo độ hỗ trợ cực tiểu không giống nhau đối với mỗi tập mục dữ liệu và quan trọng hơn là tính chất Apriori của các tập phổ biến như *tập con của tập phổ biến là tập phổ biến đã được bảo toàn* do đó quá trình tìm tập phổ biến sẽ được thực hiện hiệu quả hơn nhiều. Trong [3] đã đề xuất thuật toán CHARM-NEW trên cơ sở cải tiến thuật toán CHARM [15] để tìm các tập phổ biến đóng cực đại từ cơ sở dữ liệu tác vụ (hay nhị phân) với điều kiện về độ hỗ trợ cực tiểu như vậy.

Thực tế việc phát hiện các luật kết hợp thực sự trở nên có ý nghĩa ứng dụng to lớn khi giải quyết được vấn đề phát hiện luật kết hợp từ các CSDL định lượng ([10]) Để giải quyết vấn đề vừa nêu người ta đã đề xuất ứng dụng lý thuyết tập mờ để chuyển đổi CSDL định lượng thành CSDL mới (tạm gọi là CSDL "mờ"), và từ đó vấn đề phát hiện luật kết hợp

mờ được ra đời ([2, 4). Vấn đề này đang được quan tâm nghiên cứu, phát triển mạnh.

Bài báo tập trung phát triển một số khái niệm liên quan đến luật kết hợp mờ, thuật toán tổng quát phát hiện luật kết hợp mờ, đặc biệt là *thuật toán tìm tập phổ biến mờ đóng cực đại với các tập mục dữ liệu mờ có độ hỗ trợ cực tiểu không giống nhau.*

Phần còn lại của bài báo được cấu trúc như sau: Mục 2 sẽ cung cấp một số khái niệm cơ bản tối thiểu cần thiết có tính chất chuẩn bị để giải quyết vấn đề do bài báo đặt ra. Các khái niệm đó một số được đề xuất mới, một số là kế thừa hoặc được phát triển tiếp từ các khái niệm tương tự của một số nghiên cứu trước đó. Mục 3 sẽ trình bày những vấn đề then chốt nhất của thuật toán phát hiện luật kết hợp mờ có độ hỗ trợ cực tiểu không giống nhau. Mục 4 và Mục 5 giới thiệu một số ví dụ minh họa, một số kết luận và hướng nghiên cứu tiếp theo của bài báo.

2. KIẾN THỨC CHUẨN BỊ

Ký hiệu $\boldsymbol{I} = \{i_1, i_2, ..., i_m\}$ là tập các mục dữ liệu định lượng, là mục dữ liệu số hoặc mục dữ liệu phân loại; tập $X \subset \boldsymbol{I}$ được gọi là tập thuộc tính; $\boldsymbol{O} = \{t_1, t_2, ..., t_m\}$ là tập định danh của các tác vụ. Quan hệ nhị phân $\boldsymbol{D} \subset \boldsymbol{I} \times \boldsymbol{O}$ được gọi là cơ sở dữ liệu định lượng. Giả sử mỗi mục dữ liệu i_k $(k = 1, ..., m)$ có một số tập mờ tương ứng với nó. Ký hiệu $F_{i_k} = \{\chi_{i_k}^1, \chi_{i_k}^2, ..., \chi_{i_k}^h\}$ là tập các tập mờ tương ứng với mục dữ liệu i_k và $\chi_{i_k}^j$ là tập mờ thứ j trong F_{i_k} ([2, 3, 4, 7]).

Một luật kết hợp mờ có dạng $r = X \in A \to Y \in B$ (còn có thể được diễn giải: X là $A \to Y$ là B) với $X = \{x_1, x_2, ..., x_p\}, Y = \{y_1, y_2, ..., y_q\}$ là các tập thuộc tính, $X \cap Y = \emptyset$; $A = \{\chi_{x_1}, \chi_{x_2}, ..., \chi_{x_p}\}$, $B = \{\chi_{y_1}, \chi_{y_2}, ..., \chi_{y_q}\}$ là tập các tập mờ liên kết với các mục dữ liệu trong tập X và Y tương ứng, chẳng hạn mục dữ liệu x_k trong X sẽ có tập mờ χ_{x_k} trong A ([2, 3, 4, 7]). Cặp $\langle X, A\rangle$ với X là tập thuộc tính, A là tập gồm một số tập mờ nào đó tương ứng liên kết với các mục dữ liệu trong X được gọi là tập mục dữ liệu mờ. $\langle X, A\rangle$ được gọi là k tập mục dữ liệu mờ nếu tập X chứa k thuộc tính.

Giả sử $\{\text{minSup}i_1, \text{minSup}i_2, ..., \text{minSup}i_m / \text{minSup}i_j \in [0, 1]\}$ với mọi $j = 1, ..., m$ là tập các độ hỗ trợ cực tiểu của các mục dữ liệu trong $\boldsymbol{I} = \{i_1, i_2, ..., i_m\}$ tương ứng, nói cách khác $\text{minSup}i_j$ được gọi là độ hỗ trợ cực tiểu của thuộc tính i_j.

Định nghĩa 1. [3] Độ hỗ trợ cực tiểu của tập mục dữ liệu X ký hiệu là $\text{minSup}X = \max\{\text{minSup}i_j\}$ với mọi mục dữ liệu $i_j \in X$.

Ta dễ dàng thấy nếu $X \supseteq Y$ thì $\text{minSup}X \geqslant \text{minSup}Y$.

Định nghĩa 2. [2, 4] Độ hỗ trợ của tập mục dữ liệu mờ $\langle X, A\rangle$ đối với cơ sở dữ liệu $\boldsymbol{D}$ ký hiệu là $\text{Sup}\langle X, A\rangle$ được xác định như sau:

$$\text{Sup}\langle X, A\rangle = \frac{\sum\limits_{t_i \in O} \Pi_{x_j \in X}\{\int_{\chi x_j}(t_i[x_j])\}}{\|\boldsymbol{O}\|}$$

trong đó,

$$\int_{\chi x_j}(t_i[x_j]) = \begin{cases} m_{\chi x_j}(t_i[x_j]) \text{ nếu } m_{\chi x_j} \geqslant \omega_j \\ 0 \text{ nếu ngược lại} \end{cases}$$

$\chi_{x_j} \in A, \boldsymbol{\Pi}$ là toán tử nhân (tổng quát $\boldsymbol{\Pi}$ có thể là hoán tử T-norm) $t_i[x_j]$ là giá trị của mục

dữ liệu x_j trong tác vụ (hay định danh) thứ i là t_i, của $\boldsymbol{O}$, $m_{\chi x_j}$ là hàm thành viên của tập mờ χ_{x_j} liên kết với mục dữ liệu x_j tương ứng, $\omega_j \in [0,1]$ được gọi là ngưỡng cực tiểu của tập mờ χ_{x_j}.

Độ hỗ trợ của luật kết hợp mờ $X \in A \to Y \in B$ là $\mathrm{Sup}\langle Z, C\rangle$ với $Z = \{X, Y\}, C = \{A, B\}$ và độ tin cậy của luật đó ký hiệu là $\mathrm{Conf}\langle Z, C\rangle$ được xác định bởi

$$\mathrm{Conf}\langle Z, C\rangle = \mathrm{Sup}\langle Z, C\rangle / \mathrm{Sup}\langle X, A\rangle$$

Định nghĩa 3. Tập mục dữ liệu mờ $\langle Y, B\rangle$ được gọi là tập con của $\langle X, A\rangle$ nếu $Y \subseteq X$ và $B \subseteq A$.

Định nghĩa 4. Độ hỗ trợ cực tiểu của tập mục dữ liệu mờ $\langle X, A\rangle$, kí hiệu $\mathrm{minSup}\langle X, A\rangle = \mathrm{minSup}X$. Tập $\langle X, A\rangle$ được gọi là tập phổ biến mờ nếu $\mathrm{Sup}\langle X, A\rangle \geqslant \mathrm{minSup}\langle X, A\rangle$; tập $\langle X, A\rangle$ được gọi là tập phổ biến mờ cực đại nếu nó là tập phổ biến mờ và không tồn tại bất kỳ tập phổ biến mờ $\langle Y, B\rangle$ nào chứa nó như là một tập con thực sự.

Tính chất 1. *Tập phổ biến mờ có tính chất Apriori, tức là nếu $\langle X, A\rangle$ là tập phổ biến mờ và $\langle Y, B\rangle$ là tập con của $\langle X, A\rangle$ thì $\langle Y, B\rangle$ cũng là tập phổ biến mờ.*

Chứng minh: Dựa vào nhận xét rằng do $Y \subseteq X$ và $B \subseteq A$ nên

$$\sum_{t_i \in O} \Pi_{y_j \in Y} \{ \int_{\chi x_j} (t_i[y_j]) \} \geqslant \sum_{t_i} \Pi_{x_j \in X} \{ \int_{\chi x_j} (t_i[x_j]) \},$$

ta dễ dàng nhận được: $\mathrm{Sup}\langle Y, B\rangle \geqslant \mathrm{Sup}\langle X, A\rangle$.

Mặt khác ta lại có $\mathrm{Sup}\langle X, A\rangle \geqslant \mathrm{minSup}X \geqslant \mathrm{minSup}Y$ do $\langle X, A\rangle$ là tập phổ biến mờ và $Y \subseteq X$. Vì vậy $\mathrm{Sup}\langle Y, B\rangle \geqslant \mathrm{minSup}Y$ hay $\langle Y, B\rangle$ cũng là tập phổ biến mờ. ■

Định nghĩa 5. Luật kết hợp mờ $X \in A \to Y \in B$ xác định từ CSDL $\boldsymbol{D}$ được gọi là luật tin cậy nếu $\langle Z, C\rangle$ với $Z = \{X, Y\}$ và $C = \{A, B\}$ là tập phổ biến mờ và độ tin cậy của luật này không nhỏ hơn độ tin cậy cực tiểu minConf cho trước, tức là $\mathrm{Sup}\langle Z, C\rangle \geqslant \mathrm{minSup}Z$ và $\mathrm{Conf}\langle Z, C\rangle \geqslant \mathrm{minConf}$.

Định nghĩa 6. Ta gọi ngữ *cảnh dữ liệu mờ* (Data Fuzzy Context) là bộ ba $\boldsymbol{\mathcal{DFC}} = (\boldsymbol{O}, \boldsymbol{I}, \boldsymbol{F_I})$, trong đó $\boldsymbol{O}$ là tập hữu hạn các đối tượng (object), $\boldsymbol{I}$ là tập tất cả các mục dữ liệu và $\boldsymbol{F_I}$ là tập tất cả các tập mờ liên kết với các mục dữ liệu trong $\boldsymbol{I}$.

Ký hiệu $\boldsymbol{M}$ là tập một số tập mờ nào đó ứng với các mục dữ liệu trong $\boldsymbol{I}$ sao cho ứng với mỗi $i \in \boldsymbol{I}$ chỉ có một tập mờ trong $\boldsymbol{M}$.

Định nghĩa 7. Ta gọi ngữ *cảnh phát hiện dữ liệu mờ* (Data Fuzzy mining context) là bộ ba $\boldsymbol{\mathcal{DMC}} = (\boldsymbol{O}, \boldsymbol{I}, \boldsymbol{M})$.

Nhận xét:

- Giả sử λ_k là số các tập mờ liên kết với mục dữ liệu i_k trong tập I gồm n phần tử, thế thì mỗi ngữ cảnh dữ liệu mờ sẽ tương ứng với $\lambda_1.\lambda_2...\lambda_n$ ngữ cảnh phát hiện dữ liệu mờ. Việc phát hiện các luật kết hợp mờ hiện nay ([5, 6, 10]) mới chỉ được thực hiện đối với mỗi ngữ cảnh phát hiện dữ liệu mờ.

- Khái niệm ngữ cảnh dữ liệu và ngữ cảnh phát hiện dữ liệu mờ được phát triển và có sự khác biệt so với khái niệm tương ứng trong [9].

Các khái niệm như kết nối Galoa và tập mục dữ liệu mờ đóng bây giờ có thể được phát triển từ các khái niệm có liên quan như sau ([9, 15]):

Định nghĩa 8. (Kết nối Galois) Cho $\boldsymbol{\mathcal{DFC}} = (\boldsymbol{O}, \boldsymbol{I}, \boldsymbol{M})$ là một *ngữ cảnh phát hiện dữ liệu mờ*. Kết nối Galois của nó là tập các ánh xạ được xác định như sau:

Với $C \subseteq \boldsymbol{O}$ và $\langle X, A\rangle \subseteq \langle \boldsymbol{I}, \boldsymbol{M}\rangle$

$\overline{f} : 2^{\boldsymbol{O}} \to 2^{\boldsymbol{I}}$,

$\overline{f}(C) = \langle X, A\rangle$, ở đây $X = \{i \in \boldsymbol{I} | \forall o \in C, m_{\chi_j}(o[i]) \geqslant \omega_{\chi_i}\} \geqslant \omega_{\chi_i})$ là ngưỡng cực tiểu của tập mờ χ_i liên kết với các mục dữ liệu i trong X, $\chi_i \in A \subseteq \omega_{x_i}$.

$$\overline{g} : 2^{\boldsymbol{I}} \to 2^{\boldsymbol{O}}$$

$$\overline{g}(\langle X, A\rangle) = \{o \in \boldsymbol{O} | \forall i \in X, m_{\chi_i}(o[i]) \geqslant \omega_{\chi_i}\}.$$

$$\overline{h} : 2^{\boldsymbol{O}} \to 2^{\boldsymbol{O}} \text{ sao cho } \overline{h} = \overline{f}.\overline{g}.$$

Định nghĩa 9. Tập mục dữ liệu mờ $\langle X, A\rangle$ được gọi là đóng nếu $\overline{h}(\langle X, A\rangle) = \langle X, A\rangle$.

Nhận xét:

- Các ánh xạ $\overline{h}, \overline{f}, \overline{g}$ được phát triển tiếp từ các ánh xạ h, f, g tương ứng ([9]) cho trường hợp ngữ cảnh phát hiện dữ liệu mờ.

- Trong trường hợp CSDL ban đầu là CSDL nhị phân (mục dữ liệu nhận giá trị nhị phân), tập mục dữ liệu mờ $\langle X, A\rangle$ là đóng khi và chỉ khi X là tập đóng, tức là $h(X) = X$ với ánh xạ h được xác định như trong [9]. Việc chứng minh nó là rất đơn giản.

- Trường hợp CSDL ban đầu là CSDL định lượng thì nói chung không xảy ra mối quan hệ về tính đóng giữa tập mục dữ liệu mờ $\langle X, A\rangle$ và tập mục dữ liệu X. Mối quan hệ này sẽ được trình bày trong một bài báo khác.

Giả sử $\langle X, A\rangle$ là tập mục dữ liệu mờ, ký hiệu:

$$|\overline{g}\langle X, A\rangle| = \sum_{o \in \boldsymbol{O}} \Pi_{x_j \in X} \{\int_{\chi_{x_j}} (o[x_j])\}.$$

Tính chất sau đây được phát triển từ tính chất liên quan trong [15], là cơ sở để xây dựng thuật toán tìm tập phổ biến mờ đóng.

Tính chất 2.

a) *Giả sử* $\langle X, A\rangle, \langle Y, B\rangle$ *là hai tập mục dữ liệu mờ bất kỳ, nếu* $\text{minSup}X > |\overline{g}(\langle Y, B\rangle)|/\|\boldsymbol{O}\|$ *hoặc* $\text{minSup}Y > |\overline{g}(\langle X, A\rangle)|/\|\boldsymbol{O}\|$ *thì* $\langle X \cup Y, A \cup B\rangle$ *không là tập phổ biến mờ.*

b) *Nếu* $\overline{g}\langle X, A\rangle \subset \overline{g}(\langle Y, B\rangle)$ $\langle X, A\rangle$ *là tập phổ biến mờ, và* $\text{minSup}X \geqslant \text{minSup}Y$ *hoặc* $|\overline{g}(\langle X, A\rangle)|/\|\boldsymbol{O}\| \geqslant \text{minSup}Y$ *thì* $\langle X \cup Y, A \cup B\rangle$ *cũng là tập phổ biến mờ.*

c) *Nếu* $\overline{g}(\langle X, A\rangle) = \overline{g}(\langle Y, B\rangle)$ *và* $\langle X, A\rangle$ *hoặc* $\langle Y, B\rangle$ *là tập phổ biến mờ thì* $\langle X \cup Y, A \cup B\rangle$ *cũng là tập phổ biến mờ.*

Chứng minh:

a) Theo định nghĩa của ký hiệu $|*|$, ta thấy $|\overline{g}(\langle X, A\rangle)|/\|\boldsymbol{O}\| = \text{Sup}\langle X, A\rangle$.

Xét trường hợp $\text{minSup}Y > |\overline{g}(\langle X, A\rangle)|/\|\boldsymbol{O}\|$

Giả sử $\langle X \cup Y, A \cup B\rangle$ là tập phổ biến mờ thế thì $\text{Sup}\langle X \cup Y, A \cup B\rangle \geqslant \text{minSup}(X \cup Y) \geqslant \text{minSup}Y\rangle\text{Sup}\langle X, A\rangle$, điều này là vô lý do $\langle X, A\rangle$ là tập con của $\langle X \cup Y, A \cup B\rangle$ nên $\text{Sup}\langle X, A\rangle \geqslant \text{Sup}\langle X \cup Y, A \cup B\rangle$.

Nhận xét: Theo tính chất a), cho dù $\langle X, A\rangle, \langle Y, B\rangle$ đều là những tập phổ biến mờ nhưng $\langle X \cup Y, A \cup B\rangle$ chưa chắc có tính chất như vậy, vì thế nó thường được áp dụng để loại bỏ hoặc tìm kiếm những tập phổ biến mờ cực đại trong trường hợp cả hai tập $\langle X, A\rangle, \langle Y, B\rangle$ đều đã là tập phổ biến mờ.

b) Từ $\overline{g}(\langle X, A\rangle) \subset \overline{g}(\langle Y, B\rangle)$ và định nghĩa của $\overline{g}$, ta nhận được $\overline{g}\langle X \cup Y, A \cup B\rangle = \overline{g}(\langle X, A\rangle)$ nên suy ra

$$\mathrm{Sup}\langle X \cup Y, A \cup B\rangle = \mathrm{Sup}\langle X, A\rangle. \qquad (*)$$

Mặt khác do $\langle X, A\rangle$ là tập phổ biến mờ và theo nhận xét của Định nghĩa 1 ta có $\mathrm{Sup}\langle X, A\rangle \geqslant \mathrm{minSup}\langle X, A\rangle = \mathrm{minSup}X$.

- Nếu $\mathrm{minSup}X \geqslant \mathrm{minSup}Y$ thì

$$\mathrm{minSup}X = \mathrm{minSup}(X \cup Y). \qquad (**)$$

Từ (*) và (**) suy ra $\mathrm{Sup}\langle X \cup Y, A \cup B\rangle \geqslant \mathrm{minSup}(X \cup Y)$ hay $\langle X \cup Y, A \cup B\rangle$ là tập phổ biến mờ.

- Nếu $|\overline{g}(\langle X, A\rangle)|/\|\boldsymbol{O}\| \geqslant \mathrm{minSup}Y$ hay $\mathrm{Sup}\langle X, A\rangle \geqslant \mathrm{minSup}Y$ suy ra $\mathrm{Sup}\langle X, A\rangle \geqslant \max(\mathrm{minSup}X, \mathrm{minSup}Y) = \mathrm{minSup}(X \cup Y)$, $\langle X \cup Y, A \cup B\rangle$ là tập phổ biến mờ.

c) Được suy ra trực tiếp từ chứng minh b).

Theo [3], có thể nói quá trình phát hiện các luật kết hợp mờ với các tập thuộc tính có độ hỗ trợ cực tiểu không giống nhau từ một CSDL định lượng bất kỳ cũng gồm 3 giai đoạn chủ yếu là:

- *Giai đoạn 1*: Chuyển CSDL định lượng thành *ngữ cảnh dữ liệu mờ* (hoặc CSDL mờ): trong giai đoạn này các khái niệm mờ ứng với từng thuộc tính, các hàm thành viên của các khái niệm mờ, các độ hỗ trợ cực tiểu cho từng mục dữ liệu sẽ được xác định trước tiên bởi người sử dụng, và từ đó người sử dụng quyết định lựa chọn một *ngữ cảnh phát hiện luật kết hợp mờ* trong ngữ cảnh dữ liệu mờ đã được xác định trước đó.

- *Giai đoạn 2*: Tìm các tập phổ biến mờ có dạng $\langle Z, C\rangle$ sao cho $\mathrm{Sup}\langle Z, C\rangle \geqslant \mathrm{minSup}Z = \mathrm{minSup}(Z, C)$ là độ hỗ trợ cực tiểu của tập mục dữ liệu $\langle Z, C\rangle$.

- *Giai đoạn 3*: Từ các tập phổ biến đóng $\langle Z, C\rangle$ tìm được ở giai đoạn 2, sinh ra các luật kết hợp mờ dạng: $\langle X, A\rangle \to \langle Z - X, C - A\rangle$, ở đây $X \subset Z$ và $A \subset C$. Giai đoạn này là đơn giản.

Phần tiếp theo của bài báo chỉ tập trung vào giai đoạn 2, cụ thể là xây dựng thuật toán tìm *tập phổ biến mờ đóng cực đại* với các mục dữ liệu có độ hỗ trợ cực tiểu không giống nhau, bằng cách phát triển tiếp thuật toán CHARM-NEW ([3]) cho *trường hợp cơ sở dữ liệu định lượng* với việc ứng dụng lý thuyết tập mờ.

3. THUẬT TOÁN PHÁT HIỆN TẬP PHỔ BIẾN MỜ ĐÓNG VỚI ĐỘ HỖ TRỢ CỰC TIỂU KHÔNG GIỐNG NHAU

3.1. Ý tưởng chính của thuật toán

Thuật toán được đề xuất theo cách như sau: Để tìm các tập phổ biến mờ đóng cực đại, tương tự như các thuật toán CHARM [15] và CHARM-NEW [3], thuật toán sử dụng phương pháp duyệt theo chiều sâu trên không gian dàn các tập thuộc tính của $\langle \boldsymbol{I}, \boldsymbol{M}\rangle$. Tương tự

CHARM-NEW mỗi đỉnh của đồ thị biểu diễn không gian tìm kiếm các tập phổ biến đóng là bộ ba $\{\langle X, A\rangle, \text{minSup}X, \overline{g}(\langle X, A\rangle)\}$. Thuật toán sắp xếp các nút ở mức 1 của cây đồ thị không gian các tập phổ biến mờ theo thứ tự tăng dần của độ hỗ trợ cực tiểu của nó từ trái qua phải. Với cách sắp xếp đó các tập k-mục dữ liệu mờ ($k > 1$) được sinh ra theo phương pháp duyệt theo chiều sâu từng nhánh của cây đồ thị vẫn được sắp xếp theo thứ tự tăng dần của độ hỗ trợ cực tiểu của chúng theo thứ tự từ trái sang phải, tập sinh ra trước có độ hỗ trợ cực tiểu nhỏ hơn độ hỗ trợ cực tiểu của tập sinh ra sau, các nút thuộc nhánh bên trái đều có độ hỗ trợ cực tiểu nhỏ hơn các nút ở nhánh phải. Cơ chế hoạt động của thuật toán tìm tập phổ biến mờ cũng khá tương tự như CHARM-NEW. Cụ thể, giả sử đang xử lý nhánh có nút gốc là $\{\langle X, A\rangle, \text{minSup}X, \overline{g}(\langle X, A\rangle)\}$ ta muốn kết hợp nó với nút $\{\langle Y, B\rangle, \text{minSup}B, \overline{g}(\langle Y, B\rangle)\}$ để sinh ra nút con mới, trong đó $\langle Y, B\rangle$ được sắp thứ tự sau $\langle X, A\rangle$.

Khi đó xảy ra các trường hợp sau:

1. Khi $\overline{g}(\langle X, A\rangle) = \overline{g}(\langle Y, B\rangle)$ nếu $\langle X, A\rangle$ và $\langle Y, B\rangle$ là các tập phổ biến mờ thì $\langle X \cup Y, A \cup B\rangle$ cũng là tập phổ biến mờ (Tính chất 2c), do đó ta có thể thay thế mọi sự xuất hiện của $\langle X, A\rangle$ bởi $\langle X \cup Y, A \cup B\rangle$ và không cần xem xét các nhánh của tập $\langle Y, B\rangle$ trong các bước tìm kiếm tiếp theo;

2. Khi $\overline{g}(\langle X, A\rangle) \supset \overline{g}(\langle Y, B\rangle)$ nếu $\langle X, A\rangle, \langle Y, B\rangle$ là các tập phổ biến mờ và do các nút của đồ thị được sắp theo thứ tự tăng dần của độ hỗ trợ cực tiểu của tập mục dữ liệu trong nút nên $\text{minSup}X \leqslant \text{minSup}Y$ do đó $\langle X \cup Y, A \cup B\rangle$ cũng là tập phổ biến mờ (Tính chất 2b), nên ta có thể loại bỏ nhánh có nút gốc là $\{\langle Y, B\rangle, \text{minSup}Y, \overline{g}(\langle Y, B\rangle)\}$ và bổ sung nút $\{\langle X \cup Y, A \cup B\rangle, \text{minSup}X \cup Y, \overline{g}(\langle X \cup Y, A \cup B\rangle)\}$ vào tập các nút.

3. Khi $\overline{g}(\langle X, A\rangle) \subset \overline{g}(\langle Y, B\rangle)$ và $\langle X, A\rangle, \langle Y, B\rangle$ là các tập phổ biến mờ ta chưa thể kết luận được $\langle X \cup Y, A \cup B\rangle$ có phải là tập phổ biến mờ hay không, nói cách khác từ các nút gốc $\{\langle X, A\rangle, \text{minSup}X, \overline{g}(\langle X, A\rangle)\}$ và $\{\langle Y, B\rangle, \text{minSup}Y, \overline{g}(\langle Y, B\rangle)\}$ vẫn có tiềm năng sinh ra các tập phổ biến mờ khác nên ta không thể loại bỏ hay thay thế chúng bằng nút khác được, tuy nhiên nếu thêm điều kiện $|\overline{g}(\langle X, A\rangle)|/\|O\| \geqslant \text{minSup}Y$ hoặc $\text{minSup}X \geqslant \text{minSup}Y$ thì $\langle X \cup Y, A \cup B\rangle$ là tập phổ biến mờ nên có thể bổ sung nút $\{\langle X \cup Y, A \cup B\rangle, \text{minSup}X \cup Y, \overline{g}(\langle X \cup Y, A \cup B\rangle)\}$ vào tập các nút.

4. Khi $\overline{g}(\langle X, A\rangle) \neq \overline{g}(\langle Y, B\rangle)$ sẽ xảy ra tình huống tương tự như trường hợp 3, tức là chưa thể kết luận được $\langle X \cup Y, A \cup B\rangle$ có phải là tập phổ biến mờ hay không, và từ các nhánh có nút gốc $\{\langle X, A\rangle, \text{minSup}X, \overline{g}(\langle X, A\rangle)\}, \{\langle Y, B\rangle, \text{minSup}Y, \overline{g}(\langle Y, B\rangle)\}$ đều có thể phát sinh ra những tập phổ biến mờ mới.

Dưới đây chỉ giới thiệu phần cốt lõi nhất của thuật toán tìm tập phổ biến mờ đóng cực đại được cải tiến từ CHARM [15] và được phát triển từ CHARM-NEW [3] gọi là FUZZY-CHARM-NEW. Các thủ tục và hàm FUZZY-CHARM-EXTENDED-NEW, FUZZY-CHARM-PROPERTY-NEW có ý nghĩa và vai trò như CHARM-EXTENDED, CHARM-PROPERTY như trong thuật toán CHARM [15].

Ký hiệu $\boldsymbol{\Omega}$ là tập tất cả các tập phổ biến mờ đóng, $h(i)$ là cách đánh số tự nhiên của các thuộc tính $i \in \boldsymbol{I}$, và quy ước với mục dữ liệu i_n thì $h(i_n) = n$. Ta nói $i < j$ nếu $h(i) < h(j)$ và $j = i + 1$ nếu $h(j) = h(i) + 1$. Giả sử $\boldsymbol{I}$ là tập gồm m thuộc tính.

3.2. Thuật toán FUZZY-CHARM-NEW

FUZZY-CHARM-NEW $(\{\langle i_1, \chi_{i_k}\rangle, \text{minSup}i_1\}, \{\langle i_2, \chi_{i_2}\rangle, \text{minSup}i_2\}, ..., \{\langle i_m, \chi_{i_m}\rangle, \text{minSup}i_m\})$,

Nodes= $(\{\langle i, \chi_i\rangle, \text{minSup}i, \overline{g}(\langle i, \chi_i\rangle\}/i \in I,\ \text{Sup}\langle i, \chi_i\rangle \geqslant \text{minSup}i)$. Các đỉnh này được sắp xếp từ trái sang phải theo thứ tự tăng dần của thành phần thứ hai minSup$i$;

FUZZY-CHARM-EXTENDED-NEW (Nodes, $\mathbf{\Omega}$);

FUZZY-CHARM-EXTENDED-NEW (Nodes, Ω)

for each $\{\langle X_i, A_i\rangle, \text{minSup}X_i, \overline{g}(\langle X_i, A_i\rangle)\}$ in Nodes {

$NewN := \emptyset; X := X_i; h(j) := h(i) + 1; A := A_i$

While ($h(j) \leqslant m$ and $\{\langle X_j, A_j\rangle, \text{minSup}X_j, \overline{g}(\langle X_j, A_j\rangle\}$ in Nodes) {

$X := X \cup X_j; A := A \cup A_j$ và $Y := \overline{g}(\langle X_i, A_i\rangle) \cap \overline{g}(\langle X_j, A_j\rangle); B = A_i \cap A_j;$

FUZZY-CHARM-PROPERTY-NEW (Nodes, $NewN$)

j ++ }

If $NewN \neq \emptyset$ then FUZZY-CHARM-EXTEND ($NewN, \mathbf{\Omega}$)

$\mathbf{\Omega} := \mathbf{\Omega} \cup \langle X, A\rangle\}$

FUZZY-CHARM-PROPERTY-NEW (Nodes, $NewN$)

if ($|Y|/\|\mathcal{O}\| \geqslant \text{minSup}\mathcal{X}$) then

if ($\overline{g}(\langle X_i, A_i\rangle) = \overline{g}(\langle X_j, A_j\rangle)$) then

Loai $\{\langle X_j, A_j\rangle, \text{minSup}X_j, \overline{g}(\langle X_j, A_j\rangle)\}$ ra khoi Nodes

Thay the tat ca $\langle X_i, A_i\rangle$ boi $\langle X, A\rangle$

else if ($\overline{g}(\langle X_i, A_i\rangle) \supset (\langle X_j, A_j\rangle)$) then

Bổ sung $\{\langle X, A\rangle, \text{minSup}X, \overline{g}(\langle X, A\rangle)\}$ vào Notes

Loai $\{\langle X_j, A_j\rangle, \text{minSup}X_j, \overline{g}(\langle X_j, A_j\rangle)\}$ ra khoi Nodes

else if ($\overline{g}(\langle X_i, A_i\rangle) \subset \overline{g}(\langle X_j, A_j\rangle)$) and

(minSup$X_j \leqslant |\overline{g}(\langle X_i, A_i\rangle)|/\|\mathcal{O}\|$)) then

Thay the tat ca $\langle X_i, A_i\rangle$ boi $\langle X, A\rangle$

else if (($\overline{g}(\langle X_i, A_i\rangle \neq \overline{g}(\langle X_j, A_j\rangle)$) and

(minSup$X_j \leqslant |\overline{g}(\langle X_i, A_i\rangle)|/\|\mathcal{O}\|$)

and (minSup$X_i \leqslant |\overline{g}(\langle X_j, A_j\rangle)|/\|\mathcal{O}\|$) then

Bổ sung $\{\langle X, A\rangle, \text{minSup}X, \overline{g}(\langle X, A\rangle)\}$ vào $NewN$;

3.3. Nhận xét và đánh giá thuật toán

- Thuật toán FUZZY CHARM-NEW cho phép tìm các tập mục dữ liệu mờ đóng cực đại có độ hỗ trợ lớn hơn độ hỗ trợ cực tiểu không giống nhau ứng với từng tập mục dữ liệu từ CSDL *định lượng bất kỳ*. Thuật toán này được phát triển tiếp từ thuật toán CHARM-NEW [3] tìm tập phổ biến đóng cực đại có độ hỗ trợ cực tiểu không giống nhau *từ các CSDL nhị phân* (hay tác vụ).

- Trong [3] đã chỉ ra rằng khi độ hỗ trợ cực tiểu là chung như nhau cho tất cả các tập phổ biến thì CHARM-NEW sẽ trở thành CHARM, là thuật toán tìm các tập phổ biến đóng cực đại với độ hỗ trợ cực tiểu chung từ CSDL nhị phân hiệu quả nhất cho đến nay [15].

- Đối với FUZZY CHARM-NEW, hình thức khá giống CHARM-NEW [3]; FUZZY CHARM-NEW cũng sẽ trở thành thuật toán CHARM-NEW khi CSDL định lượng suy

biến thành cơ sở dữ liệu nhị phân.

Thật vậy, các mục dữ liệu của CSDL nhị phân do chỉ nhận một trong 2 giá trị là: 1 hoặc 0 hay "có" hoặc "không"... khi đó liên kết một cách tự nhiên hợp lý với mỗi mục dữ liệu nhị phân $x \in X$ cũng chỉ có thể có các khái niệm mờ *có* và *không* với các hàm thành viên chỉ nhận 2 giá trị 1 và 0.

Với hàm thành viên xác định như vậy dễ dàng suy ra: $\overline{g}(\langle X, A\rangle = g(X)$ và $\mathrm{Sup}\langle X, A\rangle = \mathrm{Sup}X$, Tính chất 2 ở trên trở thành tính chất để xây dựng thuật toán CHARM-NEW và FUZZY CHARM-NEW trở thành CHARM-NEW [3].

- Trong [14, 15] đã chỉ ra độ phức tạp của các thuật toán phát hiện luật kết hợp, nói chung là NP khó, trong đó thuật toán CHARM là ít phức tạp hơn nhiều so với các thuật toán phát hiện luật kết hợp khác. Trong [3] cũng đã ra chỉ ra rằng độ phức tạp của CHARM-NEW là ít hơn CHARM trong trường hợp độ hỗ trợ của các tập phổ biến là thực sự khác nhau. So với CHARM-NEW, thuật toán FUZZY CHARM-NEW là phức tạp hơn và chủ yếu ở việc tính các $|\overline{g}(\langle X_j, A_j\rangle)|$ trong các quá trình tìm kiếm và tỉa bớt các tập không phải là tập phổ biến mờ. Ước lượng chính xác độ phức tạp của thuật toán này đang được nghiên cứu làm rõ.

- Tư tưởng tìm kiếm và tỉa bớt các tập không là phổ biến của FUZZY CHARM-NEW là giống CHARM và CHARM-NEW, chúng chỉ khác nhau ở các biểu thức điều kiện trong thuật toán và đã được chứng minh trong Tính chất 2. Nói cách khác tính đúng đắn của FUZZY CHARM-NEW được khẳng định thông qua Tính chất 2 ở trên và tính đúng đắn của thuật toán CHARM.

4. VÍ DỤ MINH HỌA

CSDL trong Bảng 1 dưới đây thừa nhận rằng độ hỗ trợ cực tiểu đối với các mục dữ liệu Tuổi, Số xe máy, Thu nhập, Có gia đình tương ứng là: 0,15; 0,1; 0,05; 0,2;

Bảng 1. Cơ sở dữ liệu định lượng mẫu ban đầu

Định danh	Tuổi	Số xe máy	Thu nhập (triệu đồng)	Có Gia đình
t_1	60	0	0,6	không
t_2	40	3	6,0	có
t_3	30	0	1,5	có
t_4	25	1	3,0	không
t_5	70	2	0	có
t_6	57	4	4,0	có

Đối với mục dữ liệu **Tuổi** ta có khái niệm mờ: a) trẻ, b) trung niên, c) già; đối với **Số xe máy** ta có các khái niệm mờ: d) nhiều, e) ít; **Thu nhập** có các khái niệm mờ f) cao, g) trung bình, h) thấp; **Có gia đình** có các khái niệm mờ: i) có, j) không. Qui ước sử dụng các chữ cái a, b, c, d, e, f, g, h, i, j để biểu thị gọn tương ứng cho các khái niệm mờ: trẻ, trung niên, già, nhiều, ít, cao, trung bình, thấp, có, không.

Giả sử các hàm thành viên tương ứng của các khái niệm mờ trên được chọn thích hợp, chẳng hạn:

$$m_b(t) = \begin{cases} 0 \text{ nếu } t \geqslant 60 \text{ hoặc } t \leqslant 20 \\ (t-20)(60-t)/400 \text{ nếu } 20 < t < 60. \end{cases}$$

$$m_d(t) = \begin{cases} 1 \text{ nếu } t \geqslant 5 \\ (5-t)/5 \text{ nếu } t < 5. \end{cases}$$

$$m_g(t) = \begin{cases} t/(3\ \textit{triệu}) \text{ nếu } t \leqslant 3.\ \textit{triệu} \\ 1 \text{ nếu } 3.\ \textit{triệu} < t \leqslant 4.\ \textit{triệu} \\ \dfrac{5.\ \textit{triệu} - t}{1\ \textit{triệu}} \text{ nếu } 4.\ \textit{triệu} < t \leqslant 5.\ \textit{triệu} \\ 0 \text{ nếu } t \geqslant 5.\ \textit{triệu} \end{cases}$$

$$m_i(t) = \begin{cases} 1 \text{ nếu } t = \text{“}\textit{co}\text{”} \\ 0 \text{ nếu } t = \text{“}\textit{khong}\text{”}. \end{cases}$$

Khi đó CSDL định lượng đã cho được chuyển thành ngữ cảnh dữ liệu mờ được mô tả trong Bảng 2.

Bảng 2. Ngữ cảnh dữ liệu mờ của CSDL định lượng trong Bảng 1

Định danh	Tuổi	a	b	c	Số XM	d	e	Thu nhập	f	g	h	Có GĐ	i	j
t_1	60	0,0	0,0	1,0	0	0,0	1,0	0,6	0,12	0,2	1,0	k	0,0	1,0
t_2	40	0,5	1,0	0,5	3	0,6	0,4	6,0	1,0	0,0	0,0	c	1,0	0,0
t_3	30	0,75	0,75	0,25	0	0,0	1,0	1,5	0,3	0,5	1,0	c	1,0	0,0
t_4	25	0,87	0,44	0,12	1	0,2	0,8	3,0	0,5	1	0,66	k	0,0	1,0
t_5	70	0,0	0,0	1,0	2	0,4	0,6	0,0	0,0	0,0	1,0	c	1,0	0,0
t_6	57	0,08	0,28	0,92	4	0,8	0,2	4,0	0,8	1	0,33	c	1,0	0,0

Giả sử ngữ cảnh phát hiện dữ liệu mờ: với mục dữ liệu Tuổi liên kết với khái niệm mờ: b) trung niên. Số xe máy liên kết với d) nhiều. Thu nhập liên kết với g) trung bình, và Có gia đình liên kết với i) có. Giả sử ngưỡng cực tiểu tương ứng đối với 4 khái niệm mờ trên là: 0,3; 0,1; 0,15; 0,5.

Khi đó ngữ cảnh phát hiện dữ liệu mờ tương ứng được xác định trong Bảng 3, ở đây $\mathcal{O}$ là ký hiệu tập định danh:

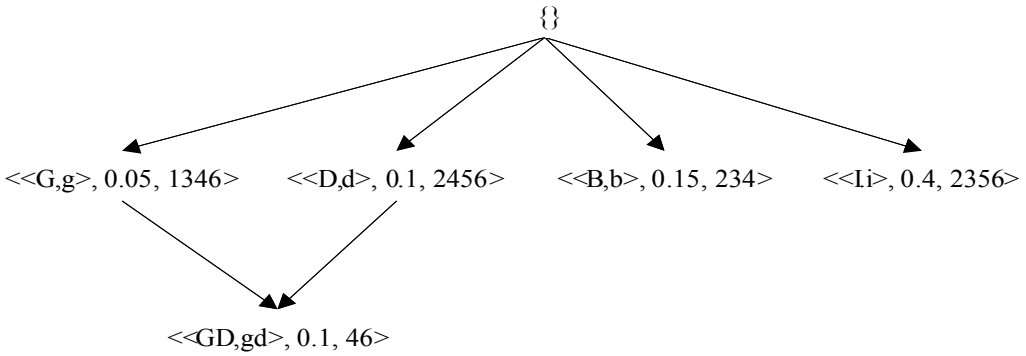
Bảng 3. Một ngữ cảnh phát hiện dữ liệu mờ

$\mathcal{O}$	Tuổi 0,15	b 0,3	Số XM 0,1	d 0,15	Thu nhập 0,05	g 0,1	Có GĐ 0,4	i 0,6
t_1	60	0,0	0	0,0	0,6	0,2	k	0,0
t_2	40	1,0	3	0,6	6,0	0,0	c	1,0
t_3	30	0,75	0	0,0	1,5	0,5	c	1,0
t_4	25	0,44	1	0,2	3,0	1	k	0,0
t_5	70	0,0	2	0,4	0,0	0,0	c	1,0
t_6	57	0,28	4	0,8	4,0	1	c	1,0

Ký hiệu B, D, G, I tương ứng là các mục dữ liệu Tuổi, Số xe máy, Thu nhập, Có gia đình; số k để biểu diễn cho giao dịch t_k trong tập định danh. Cây đồ thị biểu diễn không gian tìm kiếm tập phổ biến mờ đóng cực đại theo thuật toán FUZZY CHARM-NEW được mô tả trong Hình 1.

Mức 1 trong cây đồ thị này là tập các đỉnh có dạng $\{\langle A, a\rangle, \text{minSup}A, \overline{g}\langle A, a\rangle\}$, ở đây A là một trong các mục dữ liệu $\{G, B, D, I\}$, a là khái niệm mờ ứng với mục dữ liệu A trong ngữ cảnh phát hiện dữ liệu mờ nói trên; $\overline{g}(\langle A, a\rangle)$ được xác định theo Định nghĩa 8, chẳng hạn $\overline{g}(\langle B, b\rangle) = 234$ do ngưỡng cực tiểu của khái niệm mờ b là 0,3 cho nên các giao dịch thứ 1, 5, 6 không thuộc tập định danh $\overline{g}(\langle B, b\rangle)$.

Do $\text{Sup}(\langle G, g\rangle) = (m_g(t_1[G]) + m_g(t_3[G]) + m_g(t_4[G]) + m_g(t_6[G]))/6 = (0,2 + 0,5 + 1,0 + 1,0)/6 = 0,45 > 0,05 = \text{minSup}G$; tương tự $\text{Sup}(\langle D, d\rangle) = 0,33 > 0,1 = \text{minSup}D$; $\text{Sup}(\langle B, b\rangle) = 0,37 > 0,15 = \text{minSup}B$ và $\text{Sup}(\langle I, i\rangle) = 0,66 > 0,4 = \text{minSup}I$ cho nên tất cả 4 đỉnh thuộc mức 1 đều là những tập phổ biến mờ.



Hình 1. Không gian tìm kiếm tập phổ biến mờ cực đại theo FUZZY CHARM - NEW

Các nút thuộc mức 1 được sắp theo thứ tự tăng dần của độ hỗ trợ cực tiểu của các mục dữ liệu trong CSDL. Việc tìm tập phổ biến mờ đóng cực đại được thực hiện theo chiến lược tìm kiếm theo chiều sâu trong không gian tìm kiếm theo thứ tự từ trái sang phải.

Bắt đầu từ nút $\{\langle G, g\rangle$,0,05, 1346$\}$ khi ta kết hợp với nút $\{\langle D, d\rangle$,0,1, 2456$\}$. Ta có $\overline{g}(\langle G, g\rangle) \cap \overline{g}(\langle D, d\rangle) = 46$ nên $|\overline{g}(\langle G, g\rangle) \cap \overline{g}(\langle D, d\rangle)|/\|\boldsymbol{O}\| = |\overline{g}(\langle G, g\rangle \cup \langle D, d\rangle)|/\|\boldsymbol{O}\| = (m_g(t_4[G]).m_d(t_4[D]) + m_g(t_6[G]).m_d(t_6[D])/6 = (1.0,2 + 1.0,8)/6 = 0,16 = \text{Sup}(\langle GD, gd\rangle) > \text{minSup}GD = 0,1$. Mặt khác do $\overline{g}(\langle G, g\rangle) = 1346 \neq \overline{g}(\langle D, d\rangle) = 2456$ và $\text{minSup}D < |\overline{g}(\langle G, b\rangle)|/\|\boldsymbol{O}\|, \text{minSup}G < |\overline{g}(\langle D, d\rangle)|/\|\boldsymbol{O}\|$ nên có thể bổ sung $\{\langle GD, gd\rangle, 0,1, 46\}$ vào nút của đồ thị.

Kết hợp $\{\langle G, g\rangle, 0,05,\ 1346\}$ với $\{\langle B, b\rangle, 0,15,\ 234\}$, do $\text{Sup}(\langle GB, gb\rangle) = 0,09 < 0,15 = \text{minSup}GB$ nên kết hợp này không được thực hiện. Kết hợp $\{\langle G, g\rangle, 0,05,\ 1346\}$ với $\{\langle I, i\rangle, 0,4,\ 2356\}$, do $\text{Sup}(\langle GI, gi\rangle) = 0,25 < 0,4 = \text{minSup}GI$ nên kết hợp này cũng không thực hiện được. Như vậy nhánh với nút gốc $\{\langle G, g\rangle, 0,05, 1346\}$ không phát triển được nữa và $\langle GD, gd\rangle$ là tập phổ biến mờ đóng cực đại.

Tiếp tục với nhánh có nút gốc là $\{\langle D, d\rangle, 0,1,\ 2456\}$, nhận xét thấy $\text{Sup}(\langle DB, db\rangle) = (m_d(t_2[D]).m_b(t_2[B]) + m_d(t_4[D]).m_b(t_4[B])/6 = (0,6 \times 1,0 + 0,2 \times 0.44)/6 = 0,11 < 0,15 = \text{minSup}DB$ nên kết hợp nút $\{\langle D, d\rangle, 0,1,\ 2456\}$ với nút $\{\langle B, b\rangle, 0,15, 234\}$ không được thực hiện. Tương tự do $\text{Sup}(\langle DI, di\rangle) = (0,6 + 0,4 + 0,8)/6 = 0,3 < 0,4 = \text{minSup}DI$ nên không

kết hợp được nút $\{\langle D, d\rangle, 0{,}1,\ 2456\}$ với nút $\{\langle I, i\rangle, 0{,}4, 2356\}$. Nói cách các tập phổ biến mờ không được phát triển từ nhánh có nút gốc $\{\langle D, d\rangle, 0{,}1,\ 2456\}$.

Thực hiện tương tự thuật toán FUZZY CHARM-NEW cho các nhánh còn lại. Kết quả cuối cùng nhận được:

- $\langle GD, gd\rangle$ là tập phổ biến mờ đóng cực đại với độ hỗ trợ là $\text{Sup}(\langle GD, gd\rangle) = 0{,}16$ (độ hỗ trợ cực tiểu của nó là 0,1);

- $\langle B, b\rangle$ là tập phổ biến mờ đóng cực đại với độ hỗ trợ là $\text{Sup}(\langle B, b\rangle) = 0{,}37$ (độ hỗ trợ cực tiểu của nó là 0,15);

- $\langle I, i\rangle$ là tập phổ biến cực đại với độ hỗ trợ là $\text{Sup}(\langle I, i\rangle) = 0{,}66$ (độ hỗ trợ cực tiểu của nó là 0,4);

5. KẾT LUẬN

Bài báo đã đề xuất bài toán tìm các luật kết hợp mờ có độ hỗ trợ cực tiểu của các tập mục dữ liệu không giống nhau từ các CSDL định lượng. Để giải quyết bài toán đặt ra, một số khái niệm mới đã được đề xuất, phát triển trên cơ sở tôn trọng và kế thừa một số khái niệm của những nghiên cứu trước đó. Phương pháp giải quyết vấn đề ở đây là tiếp tục phát triển thuật toán CHARM-NEW do tác giả đề xuất. Bài báo cũng chỉ ra các tính chất đặc trưng cơ bản để xây dựng thuật toán FUZZY CHARM-NEW (Tính chất 2).

Những vấn đề cần tiếp tục nghiên cứu sau bài báo này:

- Đánh giá ý nghĩa các luật kết hợp mờ với độ hỗ trợ cực tiểu không giống nhau với một số cách tiếp cận khác;

- So sánh, đánh giá độ phức tạp của các thuật toán phát hiện luật kết hợp mờ nói chung và thuật toán phát hiện luật kết hợp mờ có độ hỗ trợ cực tiểu không giống nhau.

- Nghiên cứu xem các tập mục dữ liệu mờ có tạo thành cấu trúc dàn không? nếu có thì tìm hiểu tính chất của dàn này.

- Xây dựng chương trình thử nghiệm, đề xuất các ngưỡng cực tiểu hợp lý của độ hỗ trợ cực tiểu và của khái niệm mờ trong ứng dụng thực tế.

Lời cảm ơn. Tác giả xin chân thành cảm ơn những ý kiến và bình luận xác đáng góp phần hoàn thiện hơn cho bài báo.

TÀI LIỆU THAM KHẢO

[1] R. Agrawal, T. Imielinski, and A. Swami, Mining association rules between sets of items in large databases, *Proceedings of the ACM SIGMOD Int'l Conference on Management of Data*, May 1993 (207–216)

[2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, Fast Discovery of Association Rules. Advances in Knowledge discovery and DataMining, edited by U.M. fayyad, G.Platstsky-Shapiro,P.Smyth, and Uthurusamy, AAAI Press/The MIT Press,1996, pp.306-328.

[3] Bayardo R.J.; Efficiently mining long patterns from Databases. In ACM SIGMOD Conf. Management of Data, June 1998.

[4] Cai C.H.; Mining Association Rules with Weighted Items, Thesis, Chinese University of HongKong, 8/1998.

[5] Đỗ Văn Thành. Phát hiện các luật kết hợp có độ hỗ trợ cực tiểu không giống nhau Khoa học và Công nghệ. T42, N1, 2004, 79-90.

[6] Gyenesei A.; A Fuzzy Aproach for Mining Quantitative Association Rules. Turku Centre for Computer Sciences, TUCS Technical Report, No 336, 2000.

[7] Han J., and Fu, Y.; Attribute-Oriented Induction in Data Mining. Advances in Knowledge discovery and DataMining, Edited by U.M. fayyad, G.Platstsky-Shapiro,P.Smyth, and Uthurusamy, AAAI Press/The MIT Press,1996, pp. 399-421.

[8] Han J., Kamber M.; Data mining: Concepts and Techniques. Morgan Kaufman Publishers, 2001, 550 pages.

[9] Koh S. Y. and Rountree N.; Finding Sporadic Rules Using Appriori-Inverse, Proceeding of 9th Pacific - Asia Conference, PAKDD 2005, Ha Noi, Vietnam, 18-20, 2005, pp 97-106.

[10] Kuod M., Ada P.; Mining Fuzzy Association Rules. In SIGMOD Record, 27(1), 1998.

[11] Lin D.I. and Kedem Z.. M.; Pincer Search: A new algorithms for discovering the maximum frequent set. In 6th Int. Conf. On Database Theory, January 1997.

[12] Lin D.I. and Dunham M.H.; Mining Association Rules: Anti-Skew algorithms. In 14th Int. Conf. On Data Engineering, February 1998.

[13] Mannina H.; Association rules, Handbook of Data Mining and Knowledge Discovery, Edited by Klosgen W. and Zytkow J. M.; Oxford University Press, 2002, pp 344-348

[14] Murai T., and Sato Y.; Association Rules from a point of view of Modal logic and Rough Sets. The fourth Asian Systems Symposium, 2000, Japan, pp 427-432

[15] Pasquier N., Bastide Y., Taouil R., and Lakhal L.; Efficient Mining of Association Rules Using Closed Itemset Latics. Information Systems, Vol 24, No. 1, pp. 20-46, 1999.

[16] Thanh D.V., Phuong N.T.T., and Xuan V.M.; Relationship between Association Rules and Sequence Mining, Hội nghị khoa học kỷ niệm 2 năm thành lập Khoa Công nghệ, ĐHQG Hà nội, tháng 2/2002.

[17] Wang K., He Y., Han J.; Mining Frequent Itemset Using Support Constraints. Proceedings of the 26th VLDB Conference, Cairo, Egypt, 2000.

[18] Wang K., He Y., Han J.; Pushing support constraints into frequent itemset mining. School of Computing, National Univer. Of Singapore, 2000.

[19] Westphal C.; Data mining solutions, methods and tools for solving real-world problems. Robert Ipsen 1998.

[20] Zaki M. J. and Ogihara M.; Theoretical Foundation of Association rules. In 3rd ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, June 1998.

[21] Zaki, M. J. and Ching-Jui Hsiao. CHARM: an efficient algorithm for closed association rule mining, 2000. In Htttp//www.cs.rpi.edu/ zaki

Nhận bài ngày 20 - 8 - 2004
Nhận lại sau sửa ngày 17 - 5 - 2006