Tạp chí Tin học và Điều khiển học, T.27, S.2 (2011), 142–153

PHÁT HIỆN TẬP MỤC SPORADIC KHÔNG TUYỆT ĐỐI HAI NGƯỠNG MỜ

CÙ THU THỦY[1] , HÀ QUANG THỤY[2]

[1]*Khoa Hệ thống Thông tin Kinh tế, Học viện Tài chính*
[2]*Khoa Công nghệ Thông tin, Đại học Công nghệ, Đại học Quốc gia Hà Nội*

Tóm tắt. Luật sporadic là luật kết hợp hiếm. Luật sporadic được chia thành hai loại là: luật sporadic tuyệt đối và luật sporadic không tuyệt đối. Vấn đề phát hiện luật sporadic tuyệt đối về cơ bản đã được giải quyết trên cả cơ sở dữ liệu (CSDL) tác vụ và định lượng. Tuy nhiên, cho đến nay người ta mới nghiên cứu đề xuất một số phương pháp tìm các luật sporadic không tuyệt đối trên CSDL tác vụ. Vấn đề phát hiện các luật sporadic không tuyệt đối trên CSDL định lượng vẫn chưa được nghiên cứu đề xuất. Bài báo này sẽ giải quyết một phần vấn đề đó. Giới thiệu thuật toán MFISI (Mining Fuzzy Imperfectly Sporadic Itemsets) dựa trên thuật toán MCISI để tìm các tập mục dữ liệu sporadic không tuyệt đối hai ngưỡng mờ. Một số kết quả thực nghiệm tìm tập mục dữ liệu sporadic không tuyệt đối hai ngưỡng mờ trên cơ sở dữ liệu thực Census Income đã được chỉ ra.

Abstract. A sporadic rule is a kind of rare association rules (which have low support but high confidence). Sporadic rules are divided into two types: perfectly and imperfectly. The problem of mining perfectly sporadic rules has been completely solved by both transactional and quantitative databases. However, the problem of mining imperfectly sporadic rules on quantitative database has not been studied so far. The aim of this paper is to give a solution to this problem. Firstly, we define the problem of mining fuzzy imperfectly sporadic rules with two thresholds. Secondly, the MFISI (Mining Fuzzy Imperfectly Sporadic Itemsets) algorithm based on our MCISI algorithm is proposed to find fuzzy imperfectly sporadic itemsets with two thresholds is. Some experiment results for finding fuzzy imperfectly sporadic itemsets with two thresholds on the Census Income database are given.

1. GIỚI THIỆU CHUNG

Trong những năm gần đây khai phá luật kết hợp hiếm (các luật có độ hỗ trợ thấp nhưng có độ tin cậy cao) nhận được nhiều quan tâm của các nhà nghiên cứu. Các luật như vậy dù ít khi xảy ra nhưng trong nhiều trường hợp chúng lại là các luật rất có giá trị. Các tác giả trong [4–7, 12–20] khẳng định việc sử dụng các thuật toán tìm tập phổ biến truyền thống để tìm các tập không phổ biến (cho các luật kết hợp hiếm) là không hiệu quả. Khi sử dụng các thuật toán truyền thống để tìm các tập không phổ biến thì phải đặt ngưỡng độ hỗ trợ cực tiểu rất nhỏ cho nên số lượng các tập tìm được sẽ khá lớn (mà chỉ có một phần trong các tập này là tập không phổ biến theo ngưỡng độ hỗ trợ cực tiểu *minSup*); chi phí cho việc tìm kiếm sẽ tăng lên. Nhằm khắc phục những khó khăn này, các nhà nghiên cứu đã phát triển các thuật toán riêng để phát hiện các luật kết hợp hiếm. Các hướng nghiên cứu chính tìm các

luật kết hợp hiếm được giới thiệu trong [24]. Theo hướng tiếp cận sử dụng đường biên phân chia giữa các tập, các tác giả trong [4, 5] đưa ra khái niệm luật sporadic (luật kết hợp hiếm hay còn gọi là luật hiếm) và phân chia luật sporadic thành hai loại là: luật sporadic tuyệt đối và luật sporadic không tuyệt đối. Vấn đề tìm các tập sporadic tuyệt đối cho luật kết hợp sporadic tuyệt đối về cơ bản đã được giải quyết trên cả CSDL tác vụ và CSDL định lượng. Tuy nhiên, với các luật sporadic không tuyệt đối thì các phương pháp đề xuất mới chỉ nhằm tìm luật này trên CSDL tác vụ. Vấn đề phát hiện luật sporadic không tuyệt đối trên CSDL định lượng hiện vẫn chưa được giải quyết triệt để. Mục 2 sẽ giới thiệu cụ thể về các phương pháp tìm luật sporadic.

Bài báo đề xuất giải pháp nhằm tìm các luật sporadic không tuyệt đối trên CSDL định lượng bằng cách áp dụng lý thuyết tập mờ. Cụ thể, bài báo sẽ giới thiệu phương pháp tìm các tập sporadic không tuyệt đối mờ bằng cách đề xuất bài toán tìm các luật kết hợp mờ có dạng $r = X$ is $A \to B$ is B sao cho:

$$\begin{cases} conf(r) \geq minConf, \\ minSup \leq sup(< X \cup Y, A \cup B >) < maxSup, \\ \exists x \in < X \cup Y, A \cup B >, sup(x) \geq maxSup, \end{cases}$$

trong đó, $minConf, minSup, maxSup$ là những giá trị do người sử dụng đưa vào trong quá trình thực hiện phát hiện luật, và chúng tương ứng được gọi là độ tin cậy cực tiểu, độ hỗ trợ cận dưới và độ hỗ trợ cận trên ($minSup < maxSup$) của luật. Bài báo gọi các luật dạng này là luật sporadic không tuyệt đối hai ngưỡng mờ và bài toán trên cũng được gọi là bài toán phát hiện luật sporadic không tuyệt đối *hai ngưỡng* mờ. Độ hỗ trợ cận dưới $minSup$ được đưa vào xuất phát từ nhận xét rằng một tập sporadic không tuyệt đối mờ dù có độ hỗ trợ nhỏ đến đâu cũng phải dương.

Phần còn lại của bài báo được cấu trúc như sau. Mục 2 giới thiệu về các công trình nghiên cứu có liên quan. Mục 3 giới thiệu về luật kết hợp mờ và tập sporadic không tuyệt đối hai ngưỡng mờ làm cơ sở cho việc đề xuất thuật toán tìm các tập sporadic không tuyệt đối hai ngưỡng mờ. Mục 4 trình bày thuật toán MFISI và các ví dụ minh họa các bước thực hiện của thuật toán. Mục 5 trình bày các thực nghiệm thuật toán trên cơ sở dữ liệu Census Income. Và cuối cùng đưa ra một số kết luận và hướng nghiên cứu tiếp theo.

2. CÁC CÔNG TRÌNH LIÊN QUAN

Koh Y. S. và cộng sự lần đầu tiên giới thiệu về luật sporadic và phân chia luật sporadic thành hai loại là: luật sporadic tuyệt đối và luật sporadic không tuyệt đối [4]. Luật sporadic tuyệt đối với độ hỗ trợ cực tiểu $maxSup$ và độ tin cậy cực tiểu $minConf$ là các luật kết hợp dạng $X \to Y$ sao cho

$$\begin{cases} conf(X \to Y) \geq minConf, \\ sup(X \cup Y) < maxSup, \\ \forall x \in X \cup Y, sup(x) < maxSup, \end{cases}$$

Như vậy, luật sporadic tuyệt đối sẽ bao gồm cả phần tiền đề và phần hệ quả là các tập hiếm. Cũng trong nghiên cứu này, các tác giả đề xuất thuật toán Apriori-Inverse để tìm các tập sporadic tuyệt đối trên CSDL tác vụ. Thuật toán Apriori-Inverse được phát triển dựa trên

tư tưởng của thuật toán Apriori là tìm kiếm theo từng mức (level-wise search). Nhằm hạn chế các tập mục có độ hỗ trợ quá nhỏ, trong thuật toán còn sử dụng ngưỡng $minAS$ (minimum absolute support). Thuật toán Apriori-Inverse cho kết quả là tập các tập mục có độ hỗ trợ nhỏ hơn $maxSup$ nhưng lớn hơn ngưỡng $minAS$. Tuy nhiên hiệu quả của thuật toán chưa cao do thuật toán được phát triển dựa trên thuật toán Apriori (là thuật toán có độ phức tạp trung bình so với các thuật toán khác tìm tập phổ biến).

Zhou A. và cộng sự đã đề xuất hai thuật toán MBS và HMS nhằm tìm luật kết hợp hiếm giữa các tập mục dữ liệu hiếm trên CSDL tác vụ trong [7]. Quá trình phát hiện luật được chia làm hai giai đoạn:

Giai đoạn 1: Tìm tất cả các tập không phổ biến tiềm năng,tức là các tập dữ liệu I thỏa mãn điều kiện:

$\exists A, B : A \cap B = \emptyset, A \cup B = I, \forall i_k \in A, j_k \in B, sup(i_k) \leq minSup, sup(j_k) \leq minSup, interest(X, Y) \geq minInterest$

trong đó $minSup$ là độ hỗ trợ cực tiểu, $Interest(X, Y) = |sup(X \cup Y) - sup(X)supY|$

minInteres là tham số xác định tập dữ liệu tiềm năng do người sử dụng đưa ra.

Giai đoạn 2: Sinh luật có ý nghĩa từ các tập không phổ biến tiềm năng, tức là sẽ sinh các luật $A \to B$ thỏa mãn điều kiện: $sup(A) \leq minSup, sup(B) \leq minSup, interest(A, B) \geq minInterest, correlation(A, B) > 1$ và $CPIR(A, B) \geq minConf$.

Hai thuật toán chỉ cần duyệt qua CSDL hai lần. Sử dụng hàm $interest(X, Y)$ để giảm không gian tìm kiếm và sử dụng 2 chỉ số $correlation(X, Y)$ và $CPIR(X, Y)$ nhằm rút ra các luật có giá trị. Hạn chế của hai thuật toán là giới hạn về độ dài của luật tìm được do phải chi phí nhiều về bộ nhớ. Cũng trong nghiên cứu này các tác giả giới thiệu cách áp dụng hai thuật toán trên trong việc sinh các luật hiếm định lượng bằng cách gắn số lượng đi cùng các mục dữ liệu và coi các mục dữ liệu với số lượng khác nhau là khác nhau. Các luật sinh ra sẽ có dạng như $\{A = q1\} \to \{B = q2\}; \{A = q1\} \to \{B \geq q2\}$ hay {A="số lượng lớn"} → {B ="số lượng nhỏ"}.

Tuy nhiên phương pháp rời rạc hoá CSDL định lượng như trên có một số nhược điểm chính như sau [3, 21]: (i) Khi rời rạc hoá CSDL định lượng, số thuộc tính có thể sẽ tăng lên nhiều và dẫn đến phình to CSDL nhị phân. (ii) Nếu một thuộc tính định lượng được chia thành nhiều khoảng khi đó độ hỗ trợ của thuộc tính khoảng đơn phân chia có thể là rất nhỏ. (iii) Tại các điểm "biên gãy" của các thuộc tính được rời rạc hoá thường là thiếu tính tự nhiên do những giá trị rất gần nhau (hoặc tương tự nhau) của một thuộc tính lại nằm ở hai khoảng chia khác nhau. Để khắc phục nhược điểm trên, người ta sử dụng lý thuyết tập mờ trong quá trình chuyển đổi CSDL định lượng thành CSDL mới tựa như cơ sở dữ liệu nhị phân (có thể gọi là CSDL mờ), và từ đó vấn đề phát hiện luật kết hợp mờ được ra đời.

Trong [22], các tác giả đã đề xuất bài toán tổng quát hơn là phát hiện luật sporadic tuyệt đối hai ngưỡng và giới thiệu thuật toán MCPSI để tìm các tập sporadic tuyệt đối hai ngưỡng đóng. Khác với cách tiếp cận trong [4], thuật toán tìm tập sporadic tuyệt đối hai ngưỡng MCPSI được phát triển theo cách tiếp cận của thuật toán CHARM [9]. CHARM là một trong những thuật toán tìm tập phổ biến từ cơ sở dữ liệu tác vụ hiệu quả nhất. Thuật toán CHARM được xây dựng dựa trên tính chất cấu trúc dàn Galois của các tập mục dữ liệu đóng. Thuật toán này tìm các tập phổ biến đóng theo chiều sâu của không gian tìm kiếm nên tập phổ biến đóng tìm được thực chất cũng gồm cả tập phổ biến đóng cực đại. Giống như trường hợp thuật toán CHARM, không gian tìm kiếm các tập sporadic tuyệt đối hai ngưỡng đóng của thuật toán MCPSI đã được thu hẹp, đồng thời do số lượng các tập sporadic tuyệt đối hai ngưỡng

đóng giảm đi dẫn đến việc loại bỏ được nhiều luật sporadic tuyệt đối hai ngưỡng dư thừa.

Trong [24], đã trình bày bài toán tìm các tập sporadic tuyệt đối hai ngưỡng trên CSDL định lượng. Ở đây, bài toán tìm các luật sporadic tuyệt đối hai ngưỡng mờ có dạng $r = X$ is $A \to Y$ is B sao cho

$$\begin{cases} conf(r) \geq minConf, \\ \text{minSup} \leq sup(< X \cup Y, A \cup B >) < maxSup, \\ \forall x \in < X \cup Y, A \cup B >, minSup \leq sup(x) < maxSup. \end{cases}$$

Trong nghiên cứu này, thuật toán MFPSI đã được giới thiệu nhằm tìm các tập sporadic tuyệt đối hai ngưỡng mờ cho các luật sporadic tuyệt đối mờ trên CSDL định lượng. Thuật toán MFPSI được phát triển dựa trên tư tưởng của thuật toán Apriori.

Đề xuất đầu tiên về tìm các luật sporadic không tuyệt đối được Koh Y. S. và cộng sự giới thiệu trong [5]. Luật sporadic không tuyệt đối với độ hỗ trợ cực tiểu $maxSup$ và độ tin cậy cực tiểu $minConf$ là các luật kết hợp dạng $X \to Y$ sao cho

$$\begin{cases} conf(X \to Y) \geq minConf, \\ sup(X \cup Y) < maxSup, \\ \exists x \in X \cup Y, sup(x) \geq maxSup. \end{cases}$$

Các tác giả phân chia luật kết hợp sporadic không tuyệt đối thành 4 dạng cơ bản là: (1) Các luật có sự xuất hiện đồng thời của tập phổ biến và không phổ biến trong cả hai phần tiền đề và hệ quả; (2) Các luật chỉ có các tập phổ biến trong cả hai phần tiền đề và hệ quả nhưng hợp của các tập này lại là tập không phổ biến; (3) Các luật chỉ có các tập phổ biến ở phần tiền đề và chỉ có các tập không phổ biến ở phần hệ quả; (4) Các luật chỉ có tập không phổ biến ở phần tiền đề và chỉ có tập phổ biến ở phần hệ quả. Cũng trong nghiên cứu này thuật toán MIISR đã được đề xuất nhưng chỉ để tìm luật kết hợp sporadic không tuyệt đối ở dạng thứ 3 trên CSDL tác vụ. Trong [23], các tác giả đã phát triển tiếp hướng nghiên cứu này bằng việc đề xuất bài toán phát hiện luật kết hợp sporadic không tuyệt đối hai ngưỡng. Thuật toán MCISI đã được giới thiệu nhằm tìm tập sporadic không tuyệt đối hai ngưỡng đóng cho các luật sporadic không tuyệt đối trên CSDL tác vụ. Thuật toán MCISI cũng được phát triển dựa trên tư tưởng của thuật toán CHARM. Như vậy việc khai phá luật sporadic không tuyệt đối về cơ bản đã được giải quyết trên CSDL tác vụ. Tuy nhiên, vấn đề phát hiện các luật sporadic không tuyệt đối trên CSDL định lượng vẫn chưa được nghiên cứu đề xuất. Bài báo này sẽ đề xuất bài toán phát hiện luật sporadic không tuyệt đối hai ngưỡng mờ và giới thiệu thuật toán MFISI nhằm tìm các tập mục dữ liệu sporadic (tập sporadic) không tuyệt đối hai ngưỡng mờ cho các luật này. Thuật toán MFISI được phát triển dựa trên ý tưởng của thuật toán MCISI tìm tập sporadic không tuyệt đối hai ngưỡng từ CSDL tác vụ.

3. TẬP SPORADIC KHÔNG TUYỆT ĐỐI HAI NGƯỠNG MỜ

3.1. Luật kết hợp mờ

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

ứng với thuộc tính i_k và $\chi_{i_k}^j$ là tập mờ thứ j trong F_{i_k}. CSDL **D** có các thuộc tính gắn với tập mờ được gọi là CSDL mờ.

Luật kết hợp mờ có dạng: $r = X$ is $A \to Y$ is B, hoặc $r = X \in A \to Y \in B$ (nghĩa là nếu X thuộc về tập mờ A, khi đó Y sẽ thuộc về tập mờ B) với $X = \{x_1, x_2, \ldots, x_p\}, Y = \{y_1, y_2, \ldots, y_q\}$ là các tập thuộc tính, $X \cap Y = \emptyset$; $A = \{\chi_{x_1}, \chi_{x_2}, \ldots, \chi_{x_p}\}$, $B = \{\chi_{y_1}, \chi_{y_2}, \ldots, \chi_{y_q}\}$ là những tập mờ liên kết với các thuộc tính trong tập X và Y tương ứng, chẳng hạn thuộc tính x_k trong X sẽ có tập mờ χ_{x_k} trong A với điều kiện χ_{x_k} cũng phải thuộc F_{x_k}. Cặp $< X, A >$ với X là tập thuộc tính, A là tập gồm một số tập mờ nào đó tương ứng liên kết với các thuộc tính trong X được gọi là tập k mục dữ liệu (k-Itemset) nếu tập X chứa k thuộc tính.

Độ hỗ trợ của tập mục dữ liệu mờ $< X, A >$ đối với cơ sở dữ liệu **D** ký hiệu là $sup < X, A >$ được xác định như sau

$$sup < X, A >= \frac{\sum_{t_i \in O} \otimes_{x_j \in X} \left\{ \int_{X_{x_j}} (t_i[x_j]) \right\}}{||O||}$$

trong đó:

$\otimes$ là toán tử T-norm, được lựa chọn phép tích đại số cho nghiên cứu này.

$t_i[x_j]$ là giá trị của thuộc tính x_j trong bản ghi thứ i của **O**,

$$\int_{X_{x_j}} (t_i[x_j]) = \begin{cases} m_{X_{x_j}}(t_i[x_i]) \ \textit{Nếu} \ m_{\chi_{x_j}}(t_i[x_j]) \geq \omega_{\chi_{x_j}} \\ 0 \ \textit{Nếu Ngược lại}, \end{cases}$$

với $m_{\chi_{x_j}}(t_i[x_j])$ là hàm thành viên của thuộc tính x_j ứng với tập mờ (hay khái niệm mờ) χ_{x_j} và $\omega_{\chi_{x_j}} \in [0, 1]$ là ngưỡng (xác định bởi người dùng) của hàm thuộc.

Độ hỗ trợ của luật kết hợp mờ X is $A \to Y$ is B là $sup(< Z, C >)$ với $Z = \{X, Y\}, C = \{A, B\}$ và độ tin cậy của luật ký hiệu là $conf(< Z, C >)$ được xác định bởi $conf(< Z, C >) = Sup(< Z, C >)/Sup(< X, A >)$.

Luật kết hợp mờ $X \in A \to Y \in B$ được gọi là luật tin cậy nếu độ hỗ trợ và độ tin cậy của nó tương ứng lớn hơn hoặc bằng các ngưỡng độ hỗ trợ cực tiểu và độ tin cậy cực tiểu được xác định trước bởi người sử dụng.

3.2. Tập sporadic không tuyệt đối hai ngưỡng mờ

Định nghĩa 1. Tập $< X, A >$ được gọi là tập sporadic không tuyệt đối hai ngưỡng mờ nếu

$$minSup \leq sup(< X, A >) < maxSup, \text{ và } \quad \exists x \in < X, A >, sup(x) \geq maxSup.$$

Định nghĩa 2. Tập sporadic không tuyệt đối hai ngưỡng mờ $< Y, B >$ được gọi là tập con của $< X, A >$ nếu $Y \subseteq X$ và $B \subseteq A$.

Dễ dàng nhận thấy rằng, các tập sporadic không tuyệt đối hai ngưỡng mờ không có tính chất Apriori, tức là tập con của tập sporadic không tuyệt đối hai ngưỡng mờ chưa chắc là tập sporadic không tuyệt đối hai ngưỡng mờ.

4. TẬP SPORADIC KHÔNG TUYỆT ĐỐI HAI NGƯỠNG MỜ

4.1. Ý tưởng xây dựng thuật toán

Quá trình tìm tập sporadic không tuyệt đối hai ngưỡng mờ được tiến hành tương tự như việc tìm các tập phổ biến mờ nói chung và bao gồm các bước cơ bản sau:

i) Xây dựng tập mờ cho các thuộc tính phân loại và thuộc tính số của CSDL.

ii) Chuyển CSDL ban đầu thành CSDL mờ.

iii) Tìm các tập sporadic không tuyệt đối hai ngưỡng mờ.

Cụ thể từng bước sẽ được thực hiện như sau:

a. Xây dựng tập mờ cho các thuộc tính phân loại và thuộc tính số của CSDL

Để xây dựng tập mờ cho các thuộc tính phân loại và thuộc tính số nên để người sử dụng lựa chọn một trong hai cách là:

- Người sử dụng tự đưa ra tập mờ cho từng thuộc tính dựa trên kinh nghiệm hay quan niệm của người sử dụng về thuộc tính đó.
- Chương trình sẽ đưa ra tập mờ bằng cách ứng dụng các kỹ thuật phân cụm để phát hiện các tập mờ.

Dù áp dụng hình thức nào thì việc xây dựng tập mờ cho các thuộc tính phải đảm bảo tính rời rạc của các tập và phải bao phủ tập thuộc tính đó.

b. Chuyển CSDL ban đầu thành CSDL mờ

Sau khi xây dựng được các tập mờ cho các thuộc tính phân loại và thuộc tính số sẽ chuyển CSDL ban đầu thành CSDL mới cho việc phát hiện luật sporadic không tuyệt đối hai ngưỡng mờ. Trong giai đoạn này cần điền giá trị cho các cột mới bằng cách sử dụng hàm thành viên. Bài báo sử dụng phương pháp phân hoạch và cách xây dựng hàm thành viên trong nghiên cứu [3].

c. Tìm các tập sporadic không tuyệt đối hai ngưỡng mờ

Mục 4.2 sẽ trình bày về thuật toán đề xuất, thuật toán MFISI (Mining Fuzzy Imperfectly Sporadic Itemsets with Two Thresholds) nhằm tìm các tập sporadic không tuyệt đối hai ngưỡng mờ.

4.2. Thuật toán MFISI tìm tập sporadic không tuyệt đối hai ngưỡng mờ

Input: CSDL **D**, *minSup*, *maxSup*.
Output: Tập các tập sporadic không tuyệt đối hai ngưỡng mờ (The Set of Fuzzy Imperfectly Sporadic Itemsets with Two Thresholds).

(1) Chuyển CSDL **D(I,O)** ban đầu thành CSDL mờ $\mathbf{D}_F(\mathbf{I}_F,\mathbf{O}_F)$ Bước này sử dụng cách chia khoảng và hàm thành viên như mô tả trong [24], trong đó, $\mathbf{I}_F$ là tập các thuộc tính trong $\mathbf{D}_F$, mỗi thuộc tính x_j của $\mathbf{I}_F$ đều được gắn với một tập mờ χ_{x_j}. Mỗi tập mờ χ_{x_j} đều có một ngưỡng $\omega_{\chi_{x_j}}$.

(2) Từ tập thuộc tính ban đầu tách thành hai tập:

$FI = \{< X_i, A_i >, sup(X_i, A_i) \geq maxSup; < X_i, A_i > \in \mathbf{I}_F\}$

\# FI là tập các thuộc tính phổ biến theo *maxSup*.

$IFI = \{< X_j, A_j >, minSup \leq sup(< X_j, A_j >) < maxSup; < X_j, A_j > \in \mathbf{I}_F\}$

\# IFI là tập các thuộc tính không phổ biến theo *maxSup* nhưng có độ hỗ trợ lớn hơn hoặc bằng *minSup*.

Bảng 1. Bảng CSDL định lượng

Định danh	Tuổi	Số xe máy	Thu nhập (triệu đồng)	Có gia đình
t_1	20	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	0,0	Có
t_6	57	4	4,0	Có

(3) Tìm các tập sporadic không tuyệt đối hai ngưỡng mờ.

(3.1) Với mỗi thuộc tính trong FI khởi tạo không gian tìm kiếm như sau:

Kết hợp mỗi thuộc tính trong FI với các thuộc tính khác bên phải thuộc tính đang xét trong FI và với tất cả các thuộc tính trong IFI. Loại bỏ các tập có độ hỗ trợ nhỏ hơn $minSup$ để tạo không gian tìm kiếm.

For each $< Xi, Ai >$ in FI

$Nodes = \{\{< X_i, A_i >, < Y_i, B_i >\}, (< Y_i, B_i > \in FI \setminus < X_i, A_i >$ hoặc $< Y_i, B_i > \in IFI) \wedge sup(< X_i, A_i >, < Y_i, B_i >) \geq minSup\}$.

MFISI-EXTEND($Nodes, C$) # Hàm này thực hiện tìm các tập sporadic không tuyệt đối hai ngưỡng mờ trên không gian tìm kiếm khởi tạo ở trên.

(3.2) $FIS = FIS \cup C$.

Hàm MFISI-EXTEND nhằm tìm các tập sporadic không tuyệt đối hai ngưỡng mờ trên không gian tìm kiếm.

MFISI-EXTEND($Nodes, C$):

For each $< X_i, A_i >$ in $Nodes$

$NewN = \emptyset$ and $X =< X_i, A_i >$

$Foreach < X_J, A_J >$ in $Nodes$

$X = X \cup < X_J, A_J >$

If $NewN \neq \emptyset$ then MFISI-EXTEND($NewN$)

If $sup(X) < maxSup$.

$C = C \cup X$ # if X is not subsumed

4.3. Ví dụ minh họa

CSDL được mô tả trong Bảng 1 dưới đây gồm các thuộc tính Tuổi, Số xe máy, Thu nhập, Có gia đình.

Thực hiện xây dựng tương tự như trong [24] sẽ nhận được CSDL mờ như Bảng 1:

Do hàm thuộc của mỗi tập mờ χ_{x_j} có một ngưỡng $\omega_{\chi_{x_j}}$ nên chỉ những giá trị nào vượt ngưỡng mới được tính đến, ngược lại những giá trị không vượt ngưỡng được xem bằng 0. Ngưỡng $\omega_{\chi_{x_j}}$ phụ thuộc vào mỗi hàm thuộc và từng thuộc tính. Giả thiết các thuộc tính trong CSDL trên đều lấy $\omega_{\chi_{x_j}} = 0.4$.

Bảng 2. Bảng CSDL mờ

Định danh	Tuổi	1	2	3	Số XM	4	5	Thu nhập	6	7	8	Có GĐ	9	10
t_1	20	1	0	0	0	1	0	0,6	1	0	0	k	0,0	1,0
t_2	40	0	1	0	3	0,5	0,5	6,0	0	0	1	c	1,0	0,0
t_3	30	0,5	0,5	0	0	1	0	1,5	1	0	0	c	1,0	0,0
t_4	25	1	0	0	1	1	0	3,0	0,5	0,5	0	k	0,0	1,0
t_5	70	0	0	1	2	1	0	0,0	1	0	0	c	1,0	0,0
t_6	57	0	0,83	0,17	4	0	1	4,0	0	1	0	c	1,0	0,0

Bảng 3. Các thuộc tính và độ hỗ trợ của các thuộc tính

Tập thuộc tính	Độ hỗ trợ
{[Tuổi, Tuổi_trẻ]}(1)	0,4
{[Tuổi, Tuổi_trung niên]}(2)	0,39
{[Tuổi, Tuổi_già]}(3)	0,17
{[Số xe máy, Số xe máy_ít]}(4)	0,75
{[Số xe máy, Số xe máy_nhiều]}(5)	0,25
{[Thu nhập, Thu nhập_thấp]}(6)	0,58
{[Thu nhập, Thu nhập_trung bình]}(7)	0,25
{[Thu nhập, Thu nhập_cao]}(8)	0,17
{[Gia đình, Gia đình_có]}(9)	0,67
{[Gia đình, Gia đình_không]}(10)	0,33

Nếu chọn độ hỗ trợ $minSup = 0.2$ và $maxSup = 0.4$, ta có bảng kết quả tính độ hỗ trợ đối với từng thuộc tính như Bảng 1.

Như vậy $FI = \{\{1\}, \{4\}, \{6\}, \{9\}\}; IFI = \{\{2\}, \{5\}, \{7\}, \{10\}\}$. Xét phần tử thứ nhất $\{1\}$ của tập FI, sẽ đi ghép cặp để tạo không gian tìm kiếm $\{\{1,4\}, \{1,6\}, \{1,9\}, \{1,5\}, \{1,7\}, \{1,10\}\}$ (Bảng 1).

Lưu ý: Khi ghép các thuộc tính để tạo tập ứng cử viên không được ghép các thuộc tính có cùng nguồn gốc ban đầu với nhau. Chẳng hạn, không được ghép {[Tuổi, Tuổi_trẻ]} với {[Tuổi, Tuổi_già]} vì có cùng gốc ban đầu là [Tuổi].

Như vậy Nodes $=\{\{1,4\}, \{1,6\}, \{1,10\}\}$ Từ không gian tìm kiếm trên, thực hiện hàm MFISI-EXTEND$(Nodes, C)$ ta tìm được tập các tập sporadic không tuyệt đối hai ngưỡng mờ là: $\{\{1,6\}, \{1,10\}, \{1,4,6\}, \{1,4,10\}, \{1,6,10\}, \{1,4,6,10\}\}$ (Bảng 1).

5. KẾT QUẢ THỰC NGHIỆM

Để đánh giá hiệu quả thực hiện của thuật toán MFPSI, ta tiến hành thử nghiệm thuật toán này trên CSDL thực Census Income từ nguồn [11]. Phần thử nghiệm thực hiện trên máy tính Lenovo-IBM Codual 2.0ghz, 2GB bộ nhớ, cài đặt hệ điều hành Windows Vista. Thuật

Bảng 4. Các tập 2-mục dữ liệu và độ hỗ trợ của các tập mục

Tập thuộc tính	Độ hỗ trợ
$\{1,4\}$	0,41
$\{1,6\}$	0,33
$\{1,9\}$	0,09
$\{1,5\}$	0,0
$\{1,7\}$	0,09
$\{1,10\}$	0,33

Bảng 5. Các tập sporadic không tuyệt đối hai ngưỡng mờ tìm được ở *Nodes* thứ nhất

Tập thuộc tính	Độ hỗ trợ
$\{1,6\}$	0,33
$\{1,10\}$	0,33
$\{1,4,6\}$	0,33
$\{1,4,10\}$	0,33
$\{1,6,10\}$	0,25
$\{1,4,6,10\}$	0,25

toán MFPSI được lập trình trên ngôn ngữ C++. CSDL Census Income ban đầu gồm 14 thuộc tính và 48842 bản ghi. Các phần dữ liệu thiếu được loại bỏ trước khi thử nghiệm. Các thuộc tính được chọn dành cho việc thử nghiệm thuật toán gồm:
(1) age: continuous.
(2) sex: Female, Male.
(3) workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
(4) occupation: Tech_support, Craft_repair, Other_service, Sales, Exec_managerial,

Prof_specialty, Handlers_cleaners, Machine_op_inspct, Adm_clerical, Farming_fishing, Transport_moving, Priv_house_serv, Protective_serv, Armed_Forces.
(5) capital-gain: continuous.
(6) capital-loss: continuous.
(7) hours-per-week: continuous.

Thuộc tính (1) có khái niệm mờ: T-trẻ (17,35) , 2) T-trung niên [35,55) , 3) T-già [55,80) Thuộc tính (5), (6), (7) chia thành 3 phần tương ứng với giá trị: Thấp, Trung bình, Cao. Cách phân chia chúng tôi thực hiện dựa trên việc đếm số giá trị của thuộc tính và chia đều các giá trị này trên 3 khoảng.
Trường hợp 1: chọn tham số chồng lấp là 10%, hệ số $minSup = 0{,}1$, hệ số $maxSup$ thay đổi có kết quả như Bảng 5
Trường hợp 2: chọn tham số chồng lấp là 40%, hệ số $minSup = 0{,}1$, hệ số $maxSup$ thay đổi có kết quả như Bảng 5

Bảng 6. Kết quả thử nghiệm Trường hợp 1

maxSup	0,15	0,2	0,3	0,4	0,5
Số tập tìm được	6	6	8	8	12

Bảng 7. Kết quả thử nghiệm Trường hợp 2

maxSup	0,15	0,2	0,3	0,4	0,5
Số tập tìm được	6	8	9	9	11

Trường hợp 3: chọn tham số chồng lấp là 10%, hệ số $maxSup = 0{,}5$, hệ số *minSup* thay đổi có kết quả như Bảng 5

Trường hợp 4: chọn tham số chồng lấp là 40%, hệ số $maxSup = 0{,}5$, hệ số *minSup* thay đổi có kết quả như Bảng 5

Kết quả thử nghiệm cho thấy, khi cố định hệ số *minSup*, lựa chọn giá trị hệ số *maxSup* tăng dần thì số tập sporadic không tuyệt đối hai ngưỡng mờ tìm được cũng tăng dần (trường hợp 1 và 2). Ngược lại, khi cố định hệ số *maxSup*, lựa chọn giá trị hệ số *minSup* tăng dần thì số tập sporadic không tuyệt đối hai ngưỡng mờ tìm được giảm dần (trường hợp 3 và 4). Điều này hoàn toàn phù hợp với quy luật chung trong khai phá luật kết hợp.

Trường hợp 5: cố định hệ số $minSup = 0{,}1$, hệ số *maxSup* và tham số chồng lấp thay đổi có kết quả như Bảng 5.

Số tập sporadic không tuyệt đối hai ngưỡng mờ tìm được cũng khác nhau khi chọn cùng ngưỡng *minSup* và *maxSup* nhưng thay đổi giá trị của tham số chồng lấp (trường hợp 5).

6. KẾT LUẬN

Bài báo đã góp phần giải quyết vấn đề khai phá luật kết hợp hiếm trên CSDL định lượng bằng cách đề xuất bài toán phát hiện luật kết hợp sporadic không tuyệt đối hai ngưỡng mờ. Bài báo đã xây dựng thuật toán MFISI nhằm tìm các tập sporadic không tuyệt đối hai ngưỡng mờ. MFISI sử dụng phương pháp phân hoạch và cách xây dựng hàm thành viên do Gyenesei A. và Teuhola J. giới thiệu trong [3], được phát triển dựa trên tư tưởng của thuật toán MCISI [23]. Tiến hành thử nghiệm kết quả nghiên cứu trên CSDL Census Income [11]. Nghiên cứu tiếp theo sẽ là tìm giải pháp nhằm sinh các luật sporadic không tuyệt đối mờ có giá trị từ các tập sporadic không tuyệt đối hai ngưỡng mờ tìm được.

Bảng 8. Kết quả thử nghiệm Trường hợp 3

maxSup	0,1	0,15	0,2	0,3	0,4
Số tập tìm được	12	6	6	4	4

Bảng 9. Kết quả thử nghiệm Trường hợp 4

maxSup	0,1	0,15	0,2	0,3	0,4
Số tập tìm được	11	5	3	2	2

Bảng 10. Kết quả thử nghiệm Trường hợp 5

		Tham số chồng lấp				
minSup	*maxSup*	10%	20%	30%	40%	50%
0,1	0,15	6	7	6	6	7
0,1	0,2	6	7	6	8	9
0,1	0,3	8	9	9	9	9
0,1	0,5	12	12	11	11	11

TÀI LIỆU THAM KHẢO

[1] R. Agrawal, and R,rikant, Fast algorithms for mining association rules, *Proc. Very Large Database International Conference*, Santiago, 1994 (487–498).

[2] R.Agrawal, H.Mannila, R. Srikant, H. Toivonen, and A. Inkeri Verkamo, Fast discovery of association rules, *Advances in Knowledge discovery and DataMining*, The MIT Press, 1996 (307–328).

[3] A. Gyenesei, and J. Teuhola, Multidimensional fuzzy partitioning of attribute ranges for mining quantitative data, *International Journal of Intelligent Systems* **19** (2004) 1111–1126.

[4] Y. S. Koh, and N. Rountree, Finding sporadic rules using apriori-inverse, *PAKDD 2005*, LNAI 3518, 2005 (97–106).

[5] Y. S. Koh, and N. Rountree, and O Keefe R. A, Mining interesting imperfectly sporadic rules, *Knowledge and Information System* **14** (2) (2008) 179–196.

[6] R. U. Kiran, and P. K. Reddy, An improved multiple minimum support based approach to mine rare association rules, http://www.iiit.net/techreports/2009_24.pdf.

[7] Ling Zhou, and Stephen Yau, Association rule and quantitative association rule mining among infrequent items, *MDM 07*, San Jose, California, USA, August, 2007.

[8] N Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, Efficient mining of association rules using closed itemset latics, *Information Systems* **24** (1) (1999) 20–46.

[9] M. J. Zaki, and C. Hsiao, CHARM: An efficient algorithm for closed association rule mining, *Proceedings, SIAM-02 International Conference on Data Mining*, 2002.

[10] M. J. Zaki, Mining non-redundant association rules, *Data Mining and Knowledge Discovery* **9** (2004) 223–248.

[11] UCI-Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets.html

[12] B. Liu, W. Hsu, and Y. Ma, Mining association rules with multiple minimum supports, *procedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-99)*, San Diego, CA, USA, August 15-18, 1999 (337–341).

[13] F. Tao, F. Murtagh, and M. Farid, Weightes association rule mining using weighted support and significance framework, *Procedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-03)*, Washington, DC, USA, August 24-27, 2003 (661–666).

[14] E. Cohen, M. Datac, S. Fujiwaca, A. GioMis, P. Indyk, R. Motwani, J.D. Ullman, and C. Yang, Finding interesting association rules without support pruning, *IEEE Transactions on Knowledge and Data Engineering* **13** (2001) 64–78 (doi:10.1109/69.908981).

[15] K. Wang, Y. He, and D.W. Cheung, Mining confident rules without support requirement, *Proceedings of the Tenth International Conference on Information and Knowledge Management*, New York, USA, November 5-10, 2001 (89–96).

[16] M. Hahsler, A model-based frenquency constraint for mining associations from transaction data, *Data Mining and Knowledge Discovery* **13** (2) (September 2006) 137–166.

[17] J. Li, X. Zhang, G. Dong, K. Ramanohanaraol, and Q. Sun, Efficient mining of high confidence association rules without support threshold, *Proceedings of the 3rd European Conference on Principle and Practice of Knowledge Discovery in Databases, PKDD* 1999 (406–411).

[18] R. J. Bayardo, R. Agrawal, and D. Gunopulos, Constraint-based rule mining in large, dense databases, *Data Mining and Knowledge Discovery* **4** (2/3) 217–240 (doi:10.1023/A:1009895914772).

[19] L. Szathmary, A. Napoli, and P. Valtchev, Towards rare itemsets mining, *Proceedings of the 19th IEEE International Confedence on Tools with Artificial Intelligence*, Washington DC, USA, October 29-31, 2007 305–312).

[20] L. Troiano, G. Scibelli, C. Birtolo, A fast algorithm for mining rare itemsets, *Proceeding of the Ninth International Conference on Intelligent Systems Design and Applications*, 2009 (1149–1155) (doi: 10.1109/ISDA.2009.55).

[21] Chan Man Kuok, Ada Fu, and Man Hon Wong, "Mining Fuzzy Association Rules in Databases", Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong.

[22] Cu Thu Thuy, Do Van Thanh, Mining perfectly sporadic rules with two thresholds, *Proceedings of MASS 2010*, Wuhan, China, Aug. 22-24, 2010.

[23] Cu Thu Thuy, Do Van Thanh, Mining imperfectly sporadic rules with two thresholds, *International Journal of Computer Theory and Engineering* **2** (5) (October, 2010) 1793–8201.

[24] Cù Thu Thủy, Hà Quang Thụy, Phát hiện luật kết hợp tuyệt đối hai ngưỡng mờ, *Hội thảo quốc gia về CNTT và Truyền thông - lần thứ 13*, Hưng Yên, 19-20/8/2010.

Ngày nhận bài 23 - 3 - 2011

Nhận lại sau sửa 7 - 6 - 2011