ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

CÙ THU THỦY

NGHIÊN CỨU PHÁT HIỆN LUẬT KẾT HỢP HIẾM VÀ ỨNG DỤNG

Chuyên ngành: Hệ thống thông tin
Mã số: 62 48 05 01

TÓM TẮT LUẬN ÁN TIẾN SỸ CÔNG NGHỆ THÔNG TIN

HÀ NỘI - 2013

Công trình được hoàn thành tại: Trường Đại học Công nghệ - ĐH Quốc gia Hà nội.

NGƯỜI HƯỚNG DẪN KHOA HỌC:

1. PGS.TS. Đỗ Văn Thành
2. PGS.TS. Hà Quang Thụy

Phản biện 1: PGS.TS. Nguyễn Đình Hóa

Phản biện 2: PGS.TS. Ngô Quốc Tạo

Phản biện 3: PGS.TS. Đỗ Trung Tuấn

Luận án sẽ được bảo vệ trước Hội đồng Đại học Quốc gia chấm luận án tiến sĩ họp tại: Trường Đại học Công Nghệ - ĐHQG Hà Nội

Vào: giờ ngày tháng năm 2013

Có thể tìm hiểu luận án tại thư viện:

- Thư viện Quốc gia Việt nam
- Trung tâm Thông tin – Thư viện, Đại học Quốc gia Hà nội

DANH MỤC CÔNG TRÌNH CỦA TÁC GIẢ CÓ LIÊN QUAN ĐẾN LUẬN ÁN

1. Cù Thu Thủy, Đỗ Văn Thành (2008), "Một giải pháp mới về phân tích thị trường chứng khoán Việt Nam", *Tạp chí Tin học và Điều khiển học* Tập 24 (2), tr. 107-118.

2. Cù Thu Thủy, Đỗ Văn Thành (2009), "Phát hiện luật kết hợp với ràng buộc mục dữ liệu âm", *Tạp chí Tin học và Điều khiển học* Tập 25 (4), tr. 345-354.

3. Cu Thu Thuy, Do Van Thanh (2010), "Mining Perfectly Sporadic Rules with Two Thresholds", *In Proceedings of MASS2010*, Wuhan, China.

4. Cu Thu Thuy, Do Van Thanh (2010), "Mining Imperfectly Sporadic Rules with Two Thresholds", *International Journal of Computer Theory and Engineering* Vol. 2 (5), pp. 1793-8201.

5. Cù Thu Thủy, Hà Quang Thụy (2010), "Phát hiện luật kết hợp Sporadic tuyệt đối hai ngưỡng mờ", *Kỷ yếu Hội thảo quốc gia lần thứ XIII Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông*, Hưng Yên, tr. 263-275.

6. Cù Thu Thủy, Hà Quang Thụy (2011), "Phát hiện tập mục Sporadic không tuyệt đối hai ngưỡng mờ", *Tạp chí Tin học và Điều khiển học* Tập 27 (2), tr. 142-153.

7. Do Van Thanh, Cu Thu Thuy, Pham Thi Thu Trang (2010), "Building CPI Forecasting Model by Combining the Smooth Transition Regression Model and Mining Association Rules.", *Journal on Information Technologies and Communications* Vol E-1 (7), pp.16-27.

8. Đỗ Văn Thành, Phạm Thị Thu Trang, Cù Thu Thủy (2009), " Xây dựng mô hình dự báo giá bằng kết hợp mô hình hồi quy chuyển tiếp trơn và kỹ thuật phát hiện luật kết hợp", *Kỷ yếu Hội thảo lần thứ hai trong khuôn khổ Nghị định thư Việt Nam - Thái Lan*, Đại học Kinh tế Quốc dân, tr. 308-322.

MỞ ĐẦU

1. Lý do chọn đề tài

Trong lĩnh vực khai phá dữ liệu (data mining), luật kết hợp (association rule) được dùng để chỉ mối quan hệ kiểu "điều kiện $\rightarrow$ hệ quả" giữa các phần tử dữ liệu (chẳng hạn, sự xuất hiện của tập mặt hàng này "kéo theo" sự xuất hiện của tập mặt hàng khác) trong một tập bao gồm nhiều đối tượng dữ liệu (chẳng hạn, các giao dịch mua hàng). Phát hiện luật kết hợp là phát hiện các mối quan hệ đó trong phạm vi của một tập dữ liệu đã cho. Lý thuyết luật kết hợp được Rakesh Agrawal và cộng sự giới thiệu lần đầu tiên vào năm 1993 [13] và nhanh chóng trở thành một trong những hướng nghiên cứu khai phá dữ liệu quan trọng, đặc biệt trong những năm gần đây. Phát hiện luật kết hợp đã được ứng dụng thành công trong nhiều lĩnh vực kinh tế-xã hội khác nhau như thương mại, y tế, sinh học, tài chính-ngân hàng,...[18, 23, 25, 44, 69, 86, 87]. Hiện tại, nhiều khuynh hướng nghiên cứu và ứng dụng liên quan đến phát hiện luật kết hợp đã và đang tiếp tục được hình thành.

Một trong những vấn đề về phát hiện luật kết hợp hiện đang nhận được nhiều quan tâm của các nhà nghiên cứu là phát hiện luật kết hợp hiếm [26, 47, 49, 50, 53, 58, 66, 68, 80]. Luật kết hợp hiếm (còn được gọi là *luật hiếm*) là những luật kết hợp ít xảy ra. Mặc dù tần suất xảy ra thấp, nhưng trong nhiều trường hợp, các luật này lại rất có giá trị.

Phần lớn các thuật toán phát hiện luật kết hợp hiện nay thường thực hiện tìm các luật có độ hỗ trợ và độ tin cậy cao. Việc ứng dụng các thuật toán này để tìm các luật kết hợp hiếm (có độ hỗ trợ thấp, độ tin cậy cao) là không hiệu quả do phải đặt ngưỡng độ hỗ trợ cực tiểu rất nhỏ, nên số lượng các tập phổ biến tìm được sẽ khá lớn (trong khi chỉ có một phần trong các tập tìm được có độ hỗ trợ nhỏ hơn ngưỡng độ hỗ trợ cực tiểu *minSup*) và như vậy chi phí cho việc tìm kiếm sẽ tăng lên. Nhằm khắc phục những khó khăn này, các thuật toán phát hiện luật kết hợp hiếm được phát triển. Hai khuynh hướng phát hiện luật kết hợp hiếm được quan tâm nhiều nhất là:

(i) Sử dụng ràng buộc phần hệ quả của luật. Các phương pháp này đưa ra danh sách các mục dữ liệu sẽ xuất hiện trong một phần của luật và được sử dụng làm điều kiện khi sinh luật. Tuy nhiên, cách tiếp cận này chỉ hiệu quả khi biết trước thông tin về các mục dữ liệu, chẳng hạn phải xác định trước được mục dữ liệu nào sẽ xuất hiện trong phần hệ quả của luật [22, 56, 66].

(ii) Sử dụng đường ranh giới để phân chia tập không phổ biến với tập phổ biến và chỉ phát hiện luật hiếm từ những tập (được gọi là tập hiếm) thuộc không gian các tập không phổ biến [49, 50, 58, 75, 76, 80]. Tuy đạt được những kết quả nhất định nhưng hướng nghiên cứu này vẫn còn nhiều hạn chế như: do phải sinh ra tất cả các tập không phổ biến nên chi phí cho không gian nhớ là rất cao, và xảy ra tình trạng dư thừa nhiều luật kết hợp được sinh ra từ các tập hiếm tìm được.

Cả hai hướng nghiên cứu nói trên tập trung chủ yếu vào vấn đề phát hiện luật kết hợp hiếm trên CSDL tác vụ và vẫn chưa được giải quyết triệt để.

Vấn đề phát hiện luật kết hợp hiếm trên CSDL định lượng mới chỉ được đề cập lần đầu trong [58] và cũng chỉ nhằm phát hiện luật kết hợp hiếm từ các tập chỉ chứa các mục dữ liệu không phổ biến. Tuy nhiên, tập hiếm không chỉ gồm các mục dữ liệu

không phổ biến mà còn là sự kết hợp giữa một số mục dữ liệu không phổ biến với mục dữ liệu phổ biến hay sự kết hợp giữa những mục dữ liệu phổ biến. Như vậy, vấn đề phát hiện luật kết hợp hiếm trên CSDL định lượng hiện cũng chưa được giải quyết đầy đủ.

Luận án này sẽ tiếp nối những nghiên cứu trước đó nhằm giải quyết những hạn chế được nêu ra ở trên.

2. Mục tiêu cụ thể và phạm vi nghiên cứu

Mục tiêu cụ thể của luận án là phát triển vấn đề và đề xuất thuật toán phát hiện luật kết hợp hiếm trên cả hai loại CSDL tác vụ và định lượng, đồng thời ứng dụng ban đầu một phần kết quả nghiên cứu lý thuyết đạt được trong xây dựng mô hình phân tích và dự báo một số vấn đề cụ thể do thực tiễn đặt ra.

Phát hiện luật kết hợp hiếm có phạm vi rất rộng vì vậy nghiên cứu sinh tập trung giải quyết giai đoạn 1 của bài toán phát hiện luật hiếm, đó là đề xuất các giải pháp hiệu quả tìm tập hiếm cho cả CSDL tác vụ và định lượng.

3. Những đóng góp của luận án

Về nghiên cứu lý thuyết, luận án tập trung xác định một số dạng luật kết hợp hiếm Sporadic trên cả CSDL tác vụ và CSDL định lượng, đồng thời phát triển các thuật toán tương ứng phát hiện các tập mục dữ liệu hiếm cho các dạng luật hiếm này.

Đối với bài toán phát hiện luật hiếm trên CSDL tác vụ, luận án theo hướng tiếp cận đi tìm các tập không phổ biến đóng cho các luật hiếm thay vì việc đi tìm tất cả các tập không phổ biến như các nghiên cứu về luật hiếm trước đây. Hướng tiếp cận này của luận án là được phát triển dựa theo tư tưởng của thuật toán CHARM [94]; việc chỉ phải tìm tập hiếm đóng không những hạn chế được chi phí mà còn hạn chế được các luật hiếm dư thừa. Luận án phát triển ba thuật toán tìm các tập hiếm cho ba dạng luật kết hợp hiếm trên CSDL tác vụ là: thuật toán MCPSI phát hiện tập Sporadic tuyệt đối hai ngưỡng [32], thuật toán MCISI phát hiện tập Sporadic không tuyệt đối hai ngưỡng [33] và thuật toán NC-CHARM phát hiện tập dữ liệu với ràng buộc mục dữ liệu âm [2].

Đối với bài toán phát hiện luật hiếm trên CSDL định lượng, luận án theo hướng tiếp cận sử dụng lý thuyết tập mờ để chuyển CSDL định lượng về CSDL mờ và thực hiện phát hiện luật hiếm trên CSDL mờ này. Luận án đề xuất hai dạng luật kết hợp Sporadic cho CSDL định lượng (luật kết hợp Sporadic tuyệt đối hai ngưỡng mờ [3], luật kết hợp Sporadic không tuyệt đối hai ngưỡng mờ [4]) và phát triển hai thuật toán tìm tập hiếm cho hai dạng luật này. Thuật toán MFPSI phát hiện tập Sporadic tuyệt đối hai ngưỡng mờ [3] được phát triển theo tư tưởng của thuật toán Apriori [16], còn thuật toán MFISI phát hiện tập Sporadic không tuyệt đối hai ngưỡng mờ [4] được phát triển theo tư tưởng của thuật toán tìm tập hiếm cho luật Sporadic không tuyệt đối trên CSDL tác vụ do tác giả luận án đề xuất [33].

Về triển khai ứng dụng, luận án đề xuất kết hợp phát hiện luật kết hợp mẫu âm và mô hình hồi quy chuyển tiếp trơn phi tuyến để xây dựng mô hình phân tích và dự báo chỉ số CPI và chỉ số chứng khoán Việt Nam. Kết quả dự báo kiểm định theo mô hình được xây dựng cho thấy chất lượng dự báo được cải thiện rõ rệt, độ chính xác của kết quả dự báo so với thực tiễn là khá cao [1, 7, 36].

2. Góp phần giải quyết bài toán phát hiện luật kết hợp hiếm trên CSDL định lượng:

- Đề xuất bài toán phát luật kết hợp Sporadic tuyệt đối hai ngưỡng mờ và giới thiệu thuật toán MFPSI (được phát triển từ tư tưởng của thuật toán Apriori) nhằm tìm các tập mục cho các luật này.

- Đề xuất bài toán phát hiện luật kết hợp Sporadic không tuyệt đối hai ngưỡng mờ và giới thiệu thuật toán MFISI (được phát triển từ thuật toán MCISI của chúng tôi) nhằm tìm các tập mục cho các luật này.

Đóng góp của luận án là phát triển khuynh hướng ứng dụng tập mờ trong việc phát hiện luật kết hợp hiếm trên CSDL định lượng và đã phát triển thuật toán riêng để tìm các tập mục mờ cho luật kết hợp hiếm.

3. Góp phần nghiên cứu ứng dụng luật kết hợp trong phân tích và dự báo kinh tế, luận án đã đề xuất sử dụng luật kết hợp mẫu âm và mô hình hồi quy chuyển tiếp trơn trong việc xây dựng mô hình phân tích và dự báo chỉ số chứng khoán, giá cả và chỉ số giá tiêu dùng CPI của Việt Nam. Dự báo kiểm định các mô hình dự báo được xây dựng cho thấy kết quả dự báo là khá sát với giá trị thực tế thống kê.

4. Một hạn chế trong phần ứng dụng là luận án chưa tiến hành triển khai phát hiện luật kết hợp hiếm Sporadic trong các lĩnh vực chứng khoán cũng như giá hàng hóa và chỉ số CPI.

Hướng nghiên cứu trong tương lai

Như trong phần Phát hiện luật kết hợp với ràng buộc mục dữ liệu âm đã chỉ ra không phải CSDL tác vụ có mục dữ liệu âm nào cũng đều chuyển được về tập các mục dữ liệu dương với ràng buộc mục dữ liệu âm. Nghiên cứu tiếp theo của chúng tôi sẽ là tìm các điều kiện cần và đủ để có thể thực hiện được việc chuyển đổi biểu diễn đó.

Cả năm thuật toán được trình bầy trong luận án đều chỉ nhằm tìm các tập phổ biến cho các luật kết hợp hiếm trên cả hai loại CSDL tác vụ và CSDL định lượng. Cũng giống như vấn đề phát hiện luật kết hợp, nhiệm vụ nghiên cứu tiếp theo của chúng tôi là phải sinh được các luật hiếm có giá trị từ các tập hiếm tìm được. Đây cũng là hướng nghiên cứu hay và không dễ vì các luật kết hợp hiếm có những tính chất riêng.

Tiếp tục triển khai ứng dụng luật kết hợp với các phương pháp khác trong xây dựng mô hình phân tích và dự báo kinh tế.

Dự báo kiểm định chấp nhận mô hình dự báo chỉ số CPI:

Dữ liệu về chỉ số CPI và NB1 từ tuần thứ 95 đến tuần 103 trong tệp dữ liệu thứ hai được dùng để đánh giá mô hình dự báo. Dựa trên mô hình dự báo đã xây dựng cho chỉ số CPI_d1 tính CPI_d1(t) với t=95 đến t=103 và chỉ số CPI(t) được tính tương ứng theo CPI-d1(t). Bảng 4.1 thể hiện kết quả chỉ số CPI được tính theo mô hình đã xây dựng và chỉ số CPI theo thống kê thực tế.

Bảng 4.1: Chỉ số CPI được tính theo mô hình xây dựng và thống kê

Tháng	Tuần	Chỉ số CPI theo tuần			Chỉ số CPI theo tháng		
		CPI theo mô hình dự báo	*CPI theo kết quả thống kê*	*% sai lệch*	*CPI theo mô hình dự báo*	*CPI theo kết quả thống kê*	*% sai lệch*
11. 2009	95	100,47	100,48	0,0112%	100,51	100,55	0,04 %
	96	100,62	100,68	0,0640%			
	97	100,50	100,57	0,0678%			
	98	100,45	100,47	0,0196%			
12.2009	99	100,50	100,62	0,1221%	101,342	101,380	0,039 %
	100	100,88	100,98	0,1011%			
	101	101,60	101,46	0,1370%			
	102	101,80	101,87	0,0645%			
	103	101,93	101,97	0,0405%			

Theo bảng này ta thấy độ chính xác của kết quả dự báo là rất cao. Hơn nữa đây là mô hình dự báo không điều kiện, cụ thể CPI trong tương lai hoàn toàn có thể được tính từ các trễ của NB1.

KẾT LUẬN

Các kết quả chính của luận án

Luận án tập trung nghiên cứu, phát triển cả về lý thuyết và ứng dụng vấn đề phát hiện luật kết hợp hiếm. Qua phân tích kết quả đạt được cũng như hạn chế được nêu trong các nghiên cứu trước đây về luật kết hợp hiếm, luận án đề xuất một số vấn đề về luật kết hợp hiếm Sporadic và đã đạt được một số kết quả:

1. Góp phần giải quyết bài toán phát hiện luật kết hợp hiếm trên CSDL tác vụ:

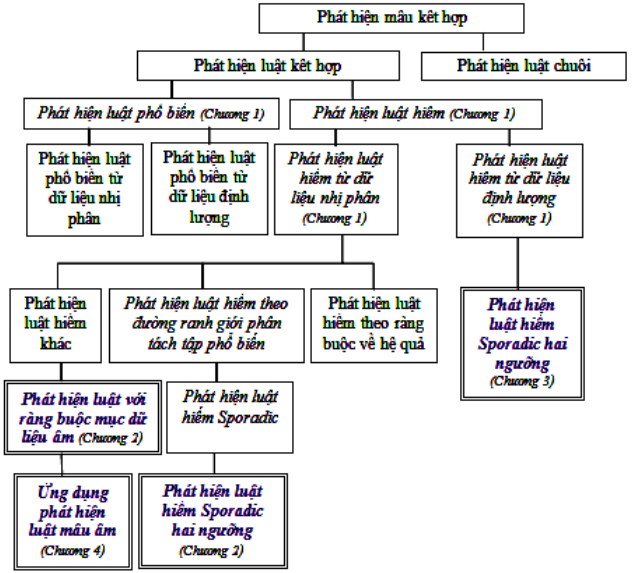
- Mở rộng bài toán phát hiện luật kết hợp Sporadic tuyệt đối hai ngưỡng và luật kết hợp Sporadic không tuyệt đối hai ngưỡng. Đề xuất hai thuật toán MCPSI và MCISI tìm các tập mục cho hai luật kết hợp hiếm này.

- Đề xuất bài toán phát hiện luật kết hợp với ràng buộc mục dữ liệu âm và giới thiệu thuật toán NC-CHARM nhằm tìm các tập phổ biến cho các luật hiếm này.

Luận án đã sử dụng chiến lược đi tìm các tập hiếm đóng thay vì đi tìm tất cả các tập hiếm cho các luật hiếm vì vậy đã tiết kiệm được chi phí và hạn chế được các luật dư thừa. Cả ba thuật toán MCPSI, MCISI và NC-CHARM đều được phát triển từ thuật toán CHARM [94] là một trong những thuật toán phát hiện luật kết hợp hiệu quả nhất trên CSDL tác vụ.

4.Tổ chức luận án

Ngoài phần mở đầu và kết luận, nội dung chính của luận án được bố cục thành 4 chương. Hình 0.1. trình bày phân bố các chủ đề phát hiện luật kết hợp được đề cập trong 4 chương nội dung của luận án. Các chủ đề nghiên cứu trong các hình chữ nhật với đường biên kép là các kết quả đóng góp chính của luận án.



Hình 0.1. Phân bố các chủ đề phát hiện luật kết hợp trong luận án

Chương 1. PHÁT HIỆN LUẬT KẾT HỢP VÀ LUẬT KẾT HỢP HIẾM

1.1. Luật kết hợp và phương pháp chung phát hiện luật kết hợp

1.1.1. Bài toán phát hiện luật kết hợp

Mục đích của bài toán phát hiện luật kết hợp là tìm ra mối quan hệ giữa các tập mục dữ liệu trong các CSDL lớn. Khái niệm luật kết hợp và phát hiện luật kết hợp được R. Agrawal và cộng sự đề xuất lần đầu tiên vào năm 1993 nhằm phát hiện các mẫu có giá trị trong CSDL tác vụ tại siêu thị [13, 14, 16].

Kí hiệu $\mathbf{I} = \{i_1, i_2, ..., i_n\}$ là tập các thuộc tính nhị phân (mỗi thuộc tính biểu diễn một mặt hàng trong siêu thị và được gọi là một mục dữ liệu, như vậy, **I** là tập tất cả các mặt hàng có trong siêu thị); tập $X \subset \mathbf{I}$ được gọi là tập mục dữ liệu hoặc tập mục (itemset); và $\mathbf{O} = \{t_1, t_2, ..., t_m\}$ là tập định danh của các tác vụ (mỗi vụ mua hàng được xem là một tác vụ). Quan hệ $\mathbf{D} \subseteq \mathbf{I} \times \mathbf{O}$ được gọi là CSDL tác vụ. Mỗi tác vụ t được biểu diễn như một véc tơ nhị phân, trong đó $t[k] = 1$ nếu mặt hàng i_k xuất hiện trong t và ngược lại $t[k] = 0$.

Cho một tập mục $X \subseteq \mathbf{I}$, độ hỗ trợ của tập X, kí hiệu là sup(X), được định nghĩa là số (hoặc phần trăm) tác vụ trong **D** chứa X.

Luật kết hợp (association rule) được định nghĩa hình thức là biểu diễn dạng $X \rightarrow Y$, trong đó $X \subseteq \mathbf{I}$, $Y \subseteq \mathbf{I}$, $X \cap Y = \emptyset$. X được gọi là phần tiền đề (antecedent) và Y được gọi là phần hệ quả (consequent) của luật.

Độ hỗ trợ (support) của luật $X \rightarrow Y$, kí hiệu là $sup(X \rightarrow Y)$ được định nghĩa là số (hoặc phần trăm) tác vụ trong **D** chứa $X \cup Y$.

Theo R. Agrawal và cộng sự [13], luật kết hợp được phát hiện cần đáp ứng ràng buộc độ hỗ trợ, theo đó, độ hỗ trợ của tập mục $W = X \cup Y$ phải vượt qua (không nhỏ thua) một ngưỡng hỗ trợ tối thiểu do người dùng đưa vào. Mọi tập W có tính chất nói trên được gọi là tập phổ biến hay tập mục lớn.

Độ tin cậy (confidence) của luật $X \rightarrow Y$, kí hiệu là $conf(X \rightarrow Y)$, được định nghĩa là số (hoặc phần trăm) tác vụ trong **D** chứa X cũng chứa Y.

Luật kết hợp được phát hiện cần có tính tin cậy, theo đó nó cần có độ tin cậy không nhỏ thua một ngưỡng tin cậy tối thiểu do người dùng đưa vào. Luật kết hợp có độ hỗ trợ và độ tin cậy tương ứng không nhỏ thua ngưỡng hỗ trợ tối thiểu và ngưỡng tin cậy tối thiểu được gọi là luật mạnh.

1.1.2. Quy trình hai bước phát hiện luật kết hợp

Phần lớn các thuật toán phát hiện luật kết hợp đều được chia thành hai giai đoạn như sau: (1) Tìm tất cả các tập phổ biến trong CSDL **D**. (2) Với mỗi tập phổ biến I_1 tìm được ở giai đoạn 1, sinh ra tất cả các luật mạnh có dạng $I_2 \rightarrow I_1 - I_2$, $I_2 \subset I_1$. Trong hai giai đoạn trên, giai đoạn 1 là khó khăn, phức tạp và tốn nhiều chi phí nhất.

4.3.1. Dữ liệu phục vụ xây dựng mô hình dự báo chỉ số CPI

Giá của các mặt hàng được thu thập hàng tuần trong năm 2008 và 2009. CPI là chỉ số được sử dụng để đánh giá mức độ lạm phát ở nước ta. Song chỉ số này chỉ được thu thập theo tháng, trong khi các mặt hàng khác lại thu thập theo tuần. Giải pháp khắc phục được đề xuất là sử dụng chỉ số giá tiêu dùng của tháng để xác định chỉ số giá tiêu dùng cho 4 tuần trong tháng.

4.3.2. Phát hiện mối quan hệ giữa giá hàng hóa và chỉ số CPI

Chọn độ hỗ trợ cực tiểu minSup = 10% và độ tin cậy cực tiểu minConf = 90% đã phát hiện được 214 luật trong đó có 12 luật chỉ có chỉ số CPI ở phần hệ quả. Trong 12 luật ở trên có 9 luật là chỉ số CPI tăng và 3 luật chỉ số CPI giảm. Tất cả các luật kết hợp này đều là luật kết hợp mẫu âm và rất khó để có thể giải thích mối quan hệ thể hiện trong luật bằng các lý thuyết kinh tế.

4.3.3. Xây dựng mô hình dự báo chỉ số CPI

Xây dựng mô hình dự báo chỉ số CPI: Các luật kết hợp ở trên cho biết tương quan về biến động giữa giá của một số mặt hàng với chỉ số CPI, nhưng chưa cho biết nó sẽ ảnh hưởng đến mức độ nào. Việc xây dựng mô hình dự báo chỉ số CPI trên các quan hệ này sẽ giúp trả lời câu hỏi đó.

Giả sử cần xây dựng mô hình dự báo chỉ số CPI dựa trên luật Rule 93:

XB41; XA81;NB12 → CPI1 (13,725% 92,86% 14 13 12,745%)

Luật 93 thể hiện mối quan hệ giữa chỉ số CPI và giá nhập khẩu của mặt hàng cotton Mỹ loại 1 (NB1), giá xuất khẩu cao su SVR loại 1 (XA8), giá xuất khẩu tôm loại 20-30 con/1kg (XB4). Luật cho biết có 14 trong số 103 tuần (chiếm 13,725%) của năm 2008 và 2009 trong đó giá của NB1 giảm nhưng giá của XA8 và XB4 tăng. Chỉ có 13 trong 103 tuần (chiếm 12,7455 %) ở đó giá nhập khẩu NB1 giảm nhưng giá xuất khẩu mặt hàng XA8, XB4 và chỉ số CPI lại tăng. Như vậy độ hỗ trợ của luật 93 là 12,745% và độ tin cậy là 92,96%. Độ tin cậy của luật chỉ ra rằng khi giá của NB1 giảm, giá XA8 và XB4 tăng thì chỉ số CPI tăng với độ tin cậy là 92,86%.

Để xây dựng mô hình dự báo chỉ số CPI từ giá của NB1, XA8 và XB4 thì dữ liệu về chỉ số CPI và giá của NB1, XA8, XB4 được chia thành 2 tập. Tập thứ 1 bao gồm 94 tuần của năm 2008 và 2009 được dùng để xây dựng mô hình dự báo chỉ số CPI. Tập thứ 2 gồm 9 tuần của tháng 11 và tháng 12 năm 2009 được dùng để kiểm định mô hình.

Ứng dụng quy trình 3 bước để xây dựng mô hình hồi quy chuyển tiếp trơn logistic trên tập thứ 1 bằng việc sử dụng phần mềm JMULTI, ta nhận được mô hình dự báo chỉ số CPI như sau:

$$CPI\_d1(t) = \begin{Bmatrix} -5{,}997 - 7{,}096\, CPI\_d1(t-1) + 7{,}347\, CPI\_d1(t-2) \\ -6{,}267\, CPI\_d1(t-3) - NB1\_d1(t-4) \end{Bmatrix} +$$

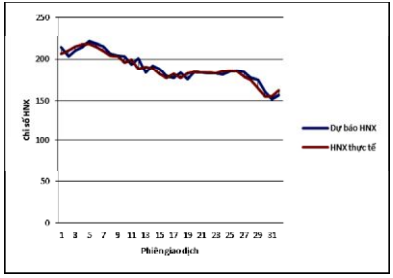
$$+ \frac{\begin{Bmatrix} 6{,}04 + 7{,}46\, CPI\_d1(t-1) - 7{,}132\, CPI\_d1(t-2) \\ +5{,}582\, CPI\_d1(t-3) + 0{,}018\, NB1\_d1(t-4) \end{Bmatrix}}{1 + \exp\{-2{,}86\,(CPI\_d1(t-3) + 0{,}803)\}}$$

Phân tích mô hình dự báo chỉ số HNX

Mô hình này cho phép nghiên cứu, phân tích và dự báo chỉ số HNX thông qua việc nghiên cứu, phân tích và dự báo các mã cổ phiếu ACB và PVI.

Dự báo kiểm nghiệm chấp nhận mô hình:

Sử dụng mô hình dự báo được xây dựng để dự báo giá trị chỉ số HNX từ ngày 16/10/2009 đến hết ngày 31/11/2009, gồm 32 phiên giao dịch và đối chiếu với giá trị thống kê thực tế của chỉ số này trong tập thứ hai, ta thấy có 17 trong 32 phiên giao dịch (bằng 53,2%) có phần trăm sai số tuyệt đối của kết quả dự báo so với giá trị thực tế của chỉ số HNX không vượt quá 0,025%, có 20 phiên giao dịch (xấp xỉ 67%) có phần trăm sai số tuyệt đối không vượt quá 0,03%,.... Như vậy độ chính xác của dự báo là khá cao (hình 4.1).



Hình 4.1: Chỉ số HNX được tính theo mô hình xây dựng và thực tế

Dự báo tiên nghiệm chỉ số chứng khoán HNX:

Việc dự báo tiên nghiệm chỉ số HNX được thực hiện thông qua dự báo giá của các cổ phiếu ACB và PVI. Cụ thể việc dự báo chỉ số HNX tại thời điểm t nào đó có thể được tính thông qua giá trị dự báo của sai phân cấp 1 của chỉ số HNX tại thời điểm này là HNX_d1(t).

4.3. Ứng dụng luật kết hợp mẫu âm và mô hình chuyển tiếp trơn trong phân tích dữ liệu giá và dự báo chỉ số CPI

Chúng tôi đề xuất cách kết hợp kỹ thuật phát hiện luật kết hợp để tìm ra mối quan hệ giữa chỉ số CPI và giá cả của các mặt hàng thiết yếu của đời sống dân sinh cũng như những mặt hàng xuất nhập khẩu chủ đạo của nền kinh tế; tiếp sau đó sẽ ứng dụng mô hình hồi quy chuyển tiếp trơn để xây dựng mô hình dự báo chỉ số CPI dựa trên mối quan hệ giữa CPI và một số mặt hàng nào được phát hiện.

Quy trình xây dựng mô hình dự báo chỉ số CPI cũng được thực hiện qua 2 giai đoạn như nêu ở mục 4.2. Giai đoạn 1 nhằm phát hiện các luật kết hợp biểu diễn mối tương quan giữa chỉ số CPI với giá của các mặt hàng. Giai đoạn 2 nhằm xây dựng các mô hình dự báo chỉ số CPI dựa trên mô hình hồi quy chuyển tiếp trơn phi tuyến và một số quan hệ được phát hiện ở giai đoạn 1.

1.2. Phát hiện luật kết hợp từ CSDL tác vụ

1.2.1. Phát hiện luật kết hợp với một ngưỡng độ hỗ trợ

Trong giai đoạn đầu tiên, bài toán phát hiện luật kết hợp đề cập tới một ngưỡng độ hỗ trợ chung (độ hỗ trợ cực tiểu) do người sử dụng đưa vào. Việc tìm các tập phổ biến được giải quyết theo 3 cách tiếp cận:

- Tìm tất cả các tập phổ biến.
- Tìm tất cả các tập phổ biến đóng.
- Tìm tất cả các tập phổ biến cực đại.

1.2.2. Phát hiện luật kết hợp với độ hỗ trợ khác nhau

Vai trò quan trọng khác nhau của các mục dữ liệu cho thấy việc sử dụng một ngưỡng độ hỗ trợ chung là không phù hợp. Các nhà nghiên cứu đã đề xuất các hướng phát hiện luật kết hợp sau:

- Phát hiện luật kết hợp có ràng buộc mục dữ liệu.
- Phát hiện luật kết hợp với độ hỗ trợ nhiều mức.
- Phát hiện luật kết hợp có trọng số.
- Phát hiện luật kết hợp có ràng buộc độ hỗ trợ.
- Phát hiện luật kết hợp không sử dụng độ hỗ trợ cực tiểu.

1.3. Phát hiện luật kết hợp từ CSDL định lượng

Hầu hết các CSDL là CSDL định lượng mà không phải là CSDL tác vụ. Phát hiện luật kết hợp từ các CSDL định lượng (thuộc tính nhận giá trị số hoặc phân loại) có ý nghĩa ứng dụng lớn hơn nhiều so với CSDL tác vụ. Năm 1996, R. Srikant và R. Agrawal [73] lần đầu đề cập tới bài toán này. Giải pháp của các tác giả rất đơn giản: đầu tiên, rời rạc hoá các thuộc tính định lượng để chuyển CSDL đã cho thành CSDL tác vụ, và sau đó, áp dụng một thuật toán phát hiện luật kết hợp từ CSDL tác vụ đã biết.

Phương pháp rời rạc hoá CSDL định lượng như trên có một số nhược điểm chính như sau [2]:

(i) Khi rời rạc hoá CSDL định lượng, số thuộc tính có thể sẽ tăng lên nhiều và dẫn đến phình to CSDL tác vụ.

(ii) Nếu một thuộc tính định lượng được chia thành nhiều khoảng khi đó độ hỗ trợ của thuộc tính khoảng đơn trong 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 khi những giá trị rất gần nhau (hoặc tương tự nhau) của một thuộc tính lại nằm ở hai khoảng chia khác nhau.

Để giải quyết những hạn chế này, người ta đã đề xuất ứng dụng lý thuyết tập mờ để chuyển đổi CSDL định lượng ban đầu thành CSDL mờ và thực hiện phát hiện luật kết hợp trên CSDL này. Từ đó hướng nghiên cứu phát hiện luật kết hợp mờ ra đời và phát triển.

1.4. Phát hiện luật kết hợp hiếm

1.4.1. Giới thiệu chung về phát hiện luật kết hợp hiếm

Phần lớn các thuật toán phát hiện luật kết hợp hiện nay thường chỉ tìm các luật có độ hỗ trợ và độ tin cậy cao. Việc ứng dụng các thuật toán này để tìm tập hiếm là không hiệu quả vì khi đó phải đặt ngưỡng độ hỗ trợ cực tiểu rất nhỏ nên số lượng các tập tìm được sẽ khá lớn (trong khi chỉ có một phần trong các tập tìm được là tập không phổ biến theo ngưỡng độ hỗ trợ cực tiểu này), chi phí cho việc tìm kiếm sẽ tăng lên. Nhằm khắc phục những khó khăn này, các thuật toán riêng để tìm các tập hiếm đã được phát triển.

1.4.2. Một số hướng nghiên cứu chính phát hiện luật kết hợp hiếm

- Sử dụng ràng buộc phần hệ quả của luật.

- Thiết lập đường biên phân chia giữa các tập phổ biến và không phổ biến.

1.4.3. Luật hiếm Sporadic

Theo hướng tiếp cận đường biên phân chia giữa tập phổ biến và tập không phổ biến, luật hiếm Sporadic do Y. S. Koh và cộng sự đề xuất [49, 50] là một dạng luật hiếm thú vị được luận án này tập trung nghiên cứu.

Các tác giả chia luật Sporadic thành hai loại là: luật Sporadic tuyệt đối và luật Sporadic không tuyệt đối.

Luật Sporadic tuyệt đối $X \rightarrow Y$ với độ hỗ trợ cực tiểu maxSup và độ tin cậy cực tiểu minConf là các luật kết hợp thỏa mãn:

$$\begin{cases} \text{conf}(X \rightarrow Y) \geq \text{minConf}, \\ \text{sup}(X \cup Y) < \text{maxSup}, \\ \forall x \in X \cup Y, \text{sup}(x) < \text{maxSup}. \end{cases} \quad (1.1)$$

Độ hỗ trợ của luật Sporadic tuyệt đối nhỏ hơn maxSup (tính hiếm) và mọi mục dữ liệu trong tập $X \cup Y$ đều có độ hỗ trợ nhỏ thua maxSup (tính hiếm "tuyệt đối"). Dựa theo ý tưởng của thuật toán Apriori, Y. S. Koh và N. Rountree phát triển thuật toán Apriori-Inverse [49] để tìm các tập Sporadic tuyệt đối.

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 $X \rightarrow Y$ sao cho:

$$\begin{cases} \text{conf}(X \rightarrow Y) \geq \text{minConf}, \\ \text{sup}(X \cup Y) < \text{maxSup}, \\ \exists x \in X \cup Y, \text{sup}(x) \geq \text{maxSup}. \end{cases} \quad (1.2)$$

Khác với luật Sporadic tuyệt đối, luật Sporadic không tuyệt đối vẫn đảm bảo tính hiếm nhưng không đòi hỏi tính hiếm "tuyệt đối" (tồn tại mục dữ liệu trong tập $X \cup Y$ có độ hỗ trợ không nhỏ thua maxSup). Các tác giả chia luật kết hợp Sporadic không tuyệt đối thành 4 dạng và giới thiệu kỹ thuật để tìm các luật Sporadic không tuyệt đối "thú vị". Đó là các luật có các mục dữ liệu ở phần tiền đề có độ hỗ trợ cao hơn maxSup nhưng giao của các tập này có độ hỗ trợ nhỏ hơn maxSup và phần hệ quả của luật có độ hỗ trợ nhỏ hơn maxSup. Đây chính là các luật thuộc dạng thứ ba trong phân loại ở trên. Thuật toán MIISR đã được đề xuất nhằm tìm phần tiền đề cho các luật dạng này [50].

vào bên phải của mã chỉ số chứng khoán hay mã cổ phiếu đó; thêm chữ số "2 " nếu chỉ số chứng khoán hoặc giá cổ phiếu giảm so với phiên trước.

4.2.2. Phát hiện mối quan hệ giữa chỉ số chứng khoán và các cổ phiếu

Với độ hỗ trợ là 35% và độ tin cậy là 90%, thực hiện phát hiện luật kết hợp trên CSDL tác vụ có mẫu âm, chúng tôi đã thu được 99 luật kết hợp.

Để xây dựng mô hình dự báo các chỉ số chứng khoán HNX và HOSE bằng mô hình hồi quy chuyển tiếp trơn phi tuyến chúng ta cần lựa chọn các luật kết hợp chỉ có mục dữ liệu liên quan đến HNX hoặc HOSE ở phần kết quả của luật. Trong trường hợp này, tất cả các luật kết hợp phát hiện được mà phần kết quả có chứa chỉ số HNX hoặc HOSE thì cũng đều chỉ chứa riêng mỗi chỉ số đó.

4.2.3. Xây dựng mô hình dự báo chỉ số chứng khoán

Về nguyên tắc, mỗi luật kết hợp chỉ có chỉ số HNX (hoặc chỉ số HOSE) ở phần kết quả sẽ cho phép ta xây dựng được một mô hình dự báo cho chỉ số này.

Chẳng hạn xét luật: PVI1; ACB1→HNX1 (38,037% 94,35% 124 117 35,890%)

Luật này cho biết: trong tổng số 350 ngày có 124 ngày chiếm hơn 38,07% trong tổng số là những ngày giá cổ phiếu của Tổng công ty cổ phần Bảo hiểm Dầu khí Việt Nam (PVI) và Ngân hàng thương mại cổ phần Á Châu (ACB) tăng giá trong đó có 117 ngày bằng 35,89% trong tổng số ngày giá cổ phiếu PVI, ACB và HNX-index cùng tăng giá, nói cách khác độ hỗ trợ của luật là 35,89%. Luật này có độ tin cậy là 94,35% và cũng cho biết có đến 94,35% những ngày khi mà PVI và ACB tăng giá thì HNX cũng tăng điểm. Có thể nói tín hiệu để nhận biết HNX tăng điểm dựa vào sự tăng giá của PVI và ACB là khá cao.

Xây dựng mô hình dự báo chỉ số HNX:

Xây dựng mô hình dự báo chỉ số HNX

Để xây dựng mô hình dự báo chỉ số HNX dựa trên luật kết hợp, dữ liệu về chỉ số chứng khoán HNX và giá của các mã cổ phiếu ACB, PVI thu thập theo các phiên giao dịch được chia thành hai tập. Tập thứ nhất bao gồm dữ liệu của các phiên giao dịch từ ngày 2/6/2008 đến hết ngày 15/10/2009 và tập thứ hai bao gồm dữ liệu các phiên giao dịch từ ngày 16/10/2009 đến ngày 31/11/2009. Tập thứ nhất được sử dụng để xây dựng mô hình, tập thứ hai được sử dụng để kiểm định chấp nhận mô hình.

Ứng dụng phần mềm JMULTI [99] trên tập thứ nhất để kiểm định tính chất tuyến tính, lựa chọn mô hình, lựa chọn biến chuyển tiếp và giá trị ban đầu của mô hình sau đó ước lượng tham số của mô hình.

Từ bảng ước lượng sẽ xây dựng được mô hình dự báo dạng:

$$HNX\_dl(t) = \begin{pmatrix} 18{,}87 + 13{,}44 HNX\_dl(t-1) + 0{,}44 ACB\_dl(t) \\ -29{,}40 PVI\_dl(t) - 5{,}0 PVI\_dl(t-3) \end{pmatrix} + \begin{pmatrix} -18{,}84 - 13{,}53 HNX\_dl(t-1) + 1{,}5 ACB\_dl(t) \\ +29{,}38 PVI\_dl(t) + 5{,}1 PVI\_dl(t-3) \end{pmatrix} * \frac{1}{1+\exp(-4{,}06*[ACB\_dl(t)+5{,}24])}$$

Bảng 3.2: Kết quả thử nghiệm thuật toán MFISI

		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	0,4	8	10	9	9	9
0,1	0,5	12	12	11	11	11

Kết quả thử nghiệm cho thấy số tập Sporadic không tuyệt đối hai ngưỡng mờ tìm được là khác nhau khi chọn cùng ngưỡng minSup và maxSup nhưng thay đổi giá trị của tham số chồng lấp.

Chương 4 - ỨNG DỤNG LUẬT KẾT HỢP MẪU ÂM VÀ MÔ HÌNH HỒI QUY CHUYỂN TIẾP TRƠN TRONG PHÂN TÍCH VÀ DỰ BÁO KINH TẾ

4.1. Mô hình hồi quy chuyển tiếp trơn

4.1.1. Phân tích hồi quy

4.1.2. Mô hình hồi quy chuyển tiếp trơn logistic

4.1.3. Xây dựng mô hình hồi quy chuyển tiếp trơn logistic

- Chỉ định mô hình

- Ước lượng tham số mô hình

- Đánh giá- Kiểm định sai lầm trong chỉ định mô hình

4.2. Ứng dụng luật kết hợp mẫu âm và mô hình hồi quy chuyển tiếp trơn trong phân tích dữ liệu chứng khoán

Nội dung phần này sẽ nghiên cứu ứng dụng luật kết hợp và mô hình hồi quy chuyển tiếp trơn logistic để xây dựng mô hình dự báo các chỉ số HNX hoặc HOSE theo một số mã cổ phiếu blue chip của thị trường chứng khoán Việt Nam.

Quy trình xây dựng mô hình dự báo chỉ số chứng khoán được thực hiện qua 2 giai đoạn. Giai đoạn 1 nhằm phát hiện các luật kết hợp biểu diễn mối tương quan giữa mỗi chỉ số chứng khoán của Việt Nam với giá của các cổ phiếu blue chip trên hai sàn giao dịch Hà Nội và Thành phố Hồ Chí Minh. Giai đoạn 2 nhằm xây dựng các mô hình dự báo chỉ số chứng khoán dựa trên mô hình hồi quy chuyển tiếp trơn phi tuyến và một số quan hệ được phát hiện ở Giai đoạn 1.

4.2.1. Dữ liệu phục vụ xây dựng mô hình

Dữ liệu phục vụ việc phát hiện luật kết hợp chứng khoán và xây dựng mô hình dự báo được thu thập theo các phiên giao dịch trên hai sàn chứng khoán Hà Nội và Thành phố Hồ Chí Minh kể từ ngày 2/6/2008 đến ngày 31/11/2009 bao gồm các thông tin sau: ngày giao dịch, giá trị của hai chỉ số HNX, HOSE và giá của các cổ phiếu Blue chip. Các luật kết hợp phục vụ việc xây dựng mô hình dự báo chỉ số chứng khoán được phát hiện từ CSDL tác vụ có mẫu âm. Tập dữ liệu này được xây dựng như sau: xuất phát từ tập dữ liệu về biến động của các chỉ số chứng khoán và biến động giá của các mã cổ phiếu blue chip, nếu chỉ số chứng khoán hoặc giá của một cổ phiếu blue chip nào đó tăng giá so với phiên trước đó thì ta thêm chữ số ”1”

1.4.4. Khuynh hướng nghiên cứu về luật hiếm

Việc sinh ra tất cả các luật hiếm hữu ích vẫn là một vấn đề khó. Quá trình này vẫn bị giới hạn bởi tính chất tự nhiên của dữ liệu.Việc phát triển các kỹ thuật tương ứng dành cho phát hiện luật kết hợp hiếm hiện vẫn là vấn đề mở theo một vài hướng tiếp cận có ý nghĩa khác nhau.

- Hướng thứ nhất là tìm ra cách phù hợp nhằm phát hiện ra các tập hiếm.

- Hướng tiếp cận thứ hai là chỉ đi tìm các luật hiếm cụ thể.

- Hướng thứ ba dựa trên việc phát triển các thuật toán tiền xử lý, tức là dựa trên các độ đo giá trị để xác định các luật hiếm.

Chương 2 - PHÁT HIỆN LUẬT KẾT HỢP HIẾM TRÊN CƠ SỞ DỮ LIỆU TÁC VỤ

2.1. Luật kết hợp Sporadic tuyệt đối hai ngưỡng

2.1.1. Giới thiệu về luật Sporadic tuyệt đối hai ngưỡng

Chúng tôi phát triển giải pháp hiệu quả hơn trong việc phát hiện luật Sporadic tuyệt đối bằng cách đề xuất mở rộng bài toán phát hiện các luật kết hợp $A \rightarrow B$:

$$\begin{cases} \text{conf}\,(A \rightarrow B) \geq m\text{inConf}, \\ \text{minSup} \leq \text{sup}(A \cup B) < \text{maxSup}, \\ \forall x \in A \cup B,\ \text{sup}\,(x) < \text{max Sup}. \end{cases} \quad (2.1)$$

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. Các luật đó được gọi là luật Sporadic tuyệt đối *hai ngưỡng* và bài toán trên cũng được gọi là bài toán phát hiện luật kết hợp Sporadic tuyệt đối hai ngưỡng.

Khác với cách tiếp cận trong [49], thuật toán tìm tập Sporadic tuyệt đối hai ngưỡng MCPSI trong nghiên cứu của chúng tôi được phát triển theo cách tiếp cận của thuật toán CHARM [94]. Thuật toán được xây dựng dựa trên tính chất cấu trúc dàn Galois của các tập mục dữ liệu đóng. 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 loại bỏ được nhiều luật Sporadic tuyệt đối hai ngưỡng dư thừa.

2.1.2. Tập Sporadic tuyệt đối hai ngưỡng

Định nghĩa 2.1: Tập X được gọi là tập Sporadic tuyệt đối hai ngưỡng nếu:

$$\text{minSup} \leq \text{sup(X)} < \text{maxSup}, \text{ và}$$

$$\forall x \in X, \text{sup(x)} < \text{maxSup}.$$

Tập Sporadictuyệt đối hai ngưỡng X được gọi là tập Sporadic tuyệt đối hai ngưỡng cực đại nếu không tồn tại tập Sporadic tuyệt đối hai ngưỡng nào chứa nó thực sự.

Định nghĩa 2.2: Ngữ cảnh khai phá dữ liệu là bộ ba $\hat{D}$ = (**O**, **INF**, **R**), trong đó **O** là tập các tác vụ, **INF** là tập tất cả các mục dữ liệu không phổ biến theo maxSup

nhưng phổ biến theo minSup và $\mathbf{R} \subseteq \mathbf{INF} \times \mathbf{O}$ là quan hệ nhị phân. Mỗi cặp (t,i) $\in \mathbf{R}$ ký hiệu cho sự kiện đối tượng t$\in$**O** quan hệ với mục dữ liệu i $\in$**INF**.

Định nghĩa 2.3: (Kết nối Galois) Cho $\hat{D}$ = (**O**, **INF**, **R**) là ngữ cảnh phát hiện dữ liệu. Với O $\subseteq$ **O** và I $\subseteq$ **INF**, xác định:

f: $2^{\mathbf{O}} \rightarrow 2^{\mathbf{INF}}$ g: $2^{\mathbf{INF}} \rightarrow 2^{\mathbf{O}}$

f(O) = {I |i$\in$I; $\forall$t$\in$O; (t,i)$\in$**R**} g(I) = {t |t$\in$O; $\forall$i$\in$I; (t,i)$\in$**R**}

f(O) là tập mục dữ liệu chung cho tất cả các đối tượng của O và g(I) là tập các đối tượng quan hệ với tất cả các mục dữ liệu trong I. Cặp ánh xạ (f,g) gọi là kết nối Galois giữa tập các tập con của **O** và tập các tập con của **INF**.

Toán tử h = $f_o g$ và h' = $g_o f$ được gọi là toán tử đóng Galois.

Định nghĩa 2.4: X là tập Sporadic tuyệt đối hai ngưỡng, X được gọi là đóng nếu h(X) = X, ở đây h là phép kết nối Galois được xác định như trên.

Nhận xét 2.1: Khi ngưỡng minSup = $\frac{1}{|O|}$, với $|O|$ là tổng số tất cả các tác vụ trong $\hat{D}$ thì bài toán phát hiện luật Sporadic tuyệt đối hai ngưỡng trở thành bài toán phát hiện luật Sporadic tuyệt đối được đề xuất trong [49]. Còn khi minSup = minAS, là ngưỡng được xác định trong thuật toán Apriori-Inverse thì bài toán phát hiện luật Sporadic tuyệt đối hai ngưỡng trở thành bài toán phát hiện luật Sporadic tuyệt đối theo cách tiếp cận được đề xuất trong Apriori-Inverse.

Tính chất 2.1: *Các tập Sporadic tuyệt đối hai ngưỡng có tính chất Apriori, tức là tập con của tập Sporadic tuyệt đối hai ngưỡng là tập Sporadic tuyệt đối hai ngưỡng.*

Tính chất đối ngẫu của tính chất này là mọi tập chứa tập con không phải là tập Sporadic tuyệt đối hai ngưỡng cũng không là tập Sporadic tuyệt đối hai ngưỡng.

Tính chất 2.2: *Độ hỗ trợ của tập Sporadic tuyệt đối hai ngưỡng X cũng bằng độ hỗ trợ bao đóng của nó, tức là sup(X) = sup(h(X)).*

Tính chất 2.3: *Nếu X là tập Sporadic tuyệt đối hai ngưỡng cực đại thì X là tập đóng.*

Tính chất 2.4: *Các luật kết hợp được sinh ra từ các tập Sporadic tuyệt đối hai ngưỡng và từ các tập Sporadic tuyệt đối hai ngưỡng cực đại là như nhau.*

2.1.3. Thuật toán tìm tập Sporadic tuyệt đối hai ngưỡng đóng

Thuật toán MCPSI được đề xuất nhằm tìm các tập Sporadic tuyệt đối hai ngưỡng đóng. Thuật toán MCPSI phát triển dựa trên tư tưởng của thuật toán CHARM. Hình 2.1. minh họa giả ngôn ngữ của thuật toán.

Độ phức tạp của thuật toán MCPSI: Độ phức tạp của thuật toán MCPSI là O(l.|**C**|) với l là độ dài trung bình của các định danh và **C** là tập Sporadic tuyệt đối hai ngưỡng đóng.

Mệnh đề 2.1*: Thuật toán MCPSI là đúng đắn và đầy đủ.*

Kết quả thử nghiệm: Để đánh giá hiệu quả thực hiện của thuật toán MCPSI, chúng tôi tiến hành thử nghiệm thuật toán này và thuật toán Apriori-Inverse trong [49] để tìm các tập Sporadic tuyệt đối trên các CSDL giả định và một số CSDL thực

3.3.3. Thuật toán tìm tập Sporadic không tuyệt đối hai ngưỡng mờ

Thuật toán MFISI được đề xuất nhằm tìm các tập Sporadic không tuyệt đối hai ngưỡng mờ. Thuật toán MFISI được phát triển từ ý tưởng của thuật toán MCISI tìm các tập Sporadic không tuyệt đối hai ngưỡng trên CSDL tác vụ.

Đầu vào: CSDL **D**, minSup, maxSup
Kết quả: Tập các tập Sporadic không tuyệt đối hai ngưỡng mờ **FIS**
Bước 1: Chuyển CSDL **D** $\subseteq$ **I** × **O** ban đầu thành CSDL mờ $\mathbf{D}_F \subseteq \mathbf{I}_F \times \mathbf{O}_F$
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. Mỗi tập mờ có một ngưỡng $\omega_{x_{ij}}$
Bước 2: Từ tập thuộc tính ban đầu tách thành hai tập:
1. FI = {<X_i,A_i>, sup(<X_i,A_i>) ≥ maxSup; <X_i,A_i> $\in \mathbf{I}_F$}
   //FI là tập các thuộc tính phổ biến theo maxSup
2. IFI = {<X_j,A_j>, minSup ≤ 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
Bước 3: Tìm các tập Sporadic không tuyệt đối hai ngưỡng mờ
// 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.
3. for each <X_i,A_i> in FI do begin
4. Nodes={{<X_i,A_i>,<Y_i,B_i>},<Y_i,B_i>∈ FI\<X_i,A_i> hoặc <Y_i,B_i>∈IFI) Λ sup(<X_i,A_i>,<Y_i,B_i>) ≥ minSup}
5. 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.
6. **FIS** = **FIS** ∪ **C**
7. end

MFISI-EXTEND(Nodes, **C**):
8. for each <X_i,A_i> in Nodes do begin
9. NewN = ∅ ; X = <X_i,A_i>
10. for each <X_j,A_j> in Nodes do
11. X = X ∪ <X_j,A_j>
12. if NewN ≠ ∅ then MFISI-EXTEND(NewN, **C**)
13. if sup(X) < maxSup then
14. **C** = **C** ∪ X // if X is not subsumed
15. end

Hình 3.2: Thuật toán MFISI

Kết quả thử nghiệm:

Để đánh giá hiệu quả thực hiện của thuật toán MFISI, chúng tôi tiến hành thử nghiệm trên CSDLthực Census Income từ nguồn [100].

Bảng 3.1: Kết quả thực hiện MFPSI với tham số chồng lấp và độ hỗ trợ minSup, maxSup khác nhau

		Tham số chồng lấp			
minSup	maxSup	20%	30%	40%	50%
0,1	0,3	10	9	9	9
0,1	0,4	13	9	9	9
0,1	0,5	17	13	13	12
0,2	0,3	2	3	1	0
0,2	0,4	3	3	1	0
0,2	0,5	6	5	3	2

Khi cố định độ hỗ trợ cận dưới minSup = 0,1 và thay đổi độ hỗ trợ cận trên maxSup lần lượt là 0,3, 0,4 và 0,5 thì nhận được số tập Sporadic tuyệt đối hai ngưỡng mờ lần lượt là 10, 13 và 17 (với tham số chồng lấp là 20%).

Nếu chọn độ hỗ trợ cận dưới minSup = 0,2 và thay đổi độ hỗ trợ cận trên maxSup lần lượt là 0,3, 0,4 và 0,5 thì nhận được số tập Sporadic tuyệt đối hai ngưỡng mờ lần lượt là 2, 3 và 6 (với tham số chồng lấp là 20%).

Như vậy, khi cố định ngưỡng minSup và lựa chọn tham số maxSup có giá trị tăng dần thì số tập Sporadic tuyệt đối hai ngưỡng mờ cũng tăng, điều này là hoàn toàn phù hợp với quy luật phát hiện luật kết hợp. Số tập Sporadic tuyệt đối hai ngưỡng mờ tìm được cũng sẽ thay đổi khi chọn hai ngưỡng độ hỗ trợ minSup và maxSup như nhau nhưng thay đổi tham số chồng lấp.

3.3. Luật kết hợp Sporadic không tuyệt đối hai ngưỡng mờ

3.3.1. Giới thiệu về luật Sporadic không tuyệt đối hai ngưỡng mờ

Chúng tôi đề xuất vấn đề tìm các luật kết hợp mờ có dạng r ≡ X is A → Y is B sao cho:

$$\begin{cases} \text{conf}(r) \geq \textit{m}\text{inConf}, \\ \text{minSup} \leq \sup(< X \cup Y, A \cup B >) < \text{maxSup}, \\ \exists x \in < X \cup Y, A \cup B >,\ \sup(x) \geq \text{maxSup}. \end{cases} \quad (3.2)$$

Các luật dạng này được gọi là luật Sporadic không tuyệt đối *hai ngưỡng* mờ và bài toán trên được gọi là bài toán phát hiện luật Sporadic không tuyệt đối hai ngưỡng mờ. Phần này sẽ trình bầy giải pháp tìm các tập Sporadic không tuyệt đối mờ cho các luật Sporadic không tuyệt đối mờ từ CSDL định lượng nào đó.

3.3.2. Tập Sporadic không tuyệt đối hai ngưỡng mờ

Định nghĩa 3.3: Tập <X,A> được gọi là tập Sporadic không tuyệt đối hai ngưỡng mờ nếu:

$$\text{minSup} \leq \sup(<X,A>) < \text{maxSup}, \text{ và}$$

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

Định nghĩa 3.4: Tập Sporadic không tuyệt đối hai ngưỡng mờ <Y,B> được gọi là tập con của <X,A> nếu $Y \subseteq X$ và $B \subseteq A$.

từ nguồn dữ liệu [100]. Phần thử nghiệm thực hiện trên máy tính Lenovo-IBM Codual 2.0ghz, 2GB bộ nhớ, cài đặt hệ điều hành Windows Vista. Thuật toán MCPSI và phần mô phỏng thuật toán Apriori-Inverse cùng được lập trình trên ngôn ngữ C++.

Đầu vào: CSDL **D**, minSup, maxSup
Kết quả: Tập các tập Sporadic tuyệt đối hai ngưỡng đóng **C**
MCPSI ALGORITHM(**D**, minSup, maxSup):
1. Nodes = $\{I_j \times g(I_j) : I_j \in \mathbf{I} \wedge |g(I_j)| < \text{maxSup} \wedge |g(I_j)| \geq \text{minSup}\}$
2. MCPSI-EXTEND(Nodes, **C**)

MCPSI-EXTEND(Nodes, **C**):
3. for each $X_i \times g(X_i)$ in Nodes do begin
4. NewN = ∅ ; X = X_i
5. for each $X_j \times g(X_j)$ in Nodes, with $k(j) > k(i)$ do begin
6. X = X∪X_j ; Y = $g(X_i) \cap g(X_j)$
7. CHARM-PROPERTY(Nodes, NewN)
8. end
9. if NewN ≠ ∅ then MCPSI-EXTEND(NewN, **C**)
10. **C** = **C**∪X // if X is not subsumed
11. end

Hàm CHARM-PROPERTY được xây dựng như trong [94].

Hình 2.1: Thuật toán MCPSI

Thử nghiệm trên CSDL giả định: Bảng 2.1 là kết quả thử nghiệm thuật toán MCPSI nhằm tìm các tập Sporadic tuyệt đối hai ngưỡng đóng và thuật toán Apriori-Inverse nhằm tìm các tập Sporadic tuyệt đối trên cùng tập dữ liệu với hai ngưỡng minSup và maxSup, trong đó minSup được chọn bằng minAS. Như đã biết khi minSup = minAS thì việc tìm tập Sporadic tuyệt đối hai ngưỡng trở thành việc tìm tập Sporadic tuyệt đối theo cách tiếp cận của Apriori-Inverse.

Bảng 2.1: Kết quả thực hiện MCPSI và Apriori-Inverse trên CSDL giả định

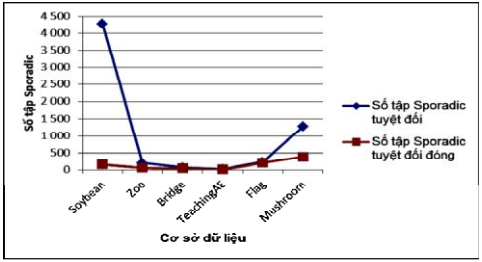
TT	Tên CSDL	minSup	maxSup	Thuật toán Apriori-Inverse		Thuật toán MCPSI	
				Số tập Sporadic tuyệt đối	Thời gian (sec)	Số tập Sporadic tuyệt đối đóng	Thời gian (sec)
1	T05I1000D10K	0,0005	0,01	3 588	67,695	1 767	62,015
2	T10I1000D10K	0,0005	0,01	1 696	38,691	1 272	37,928
3	T15I1000D10K	0,0005	0,01	955	23,917	846	22,681
4	T20I1000D10K	0,0005	0,01	610	15,614	576	14,890
5	T25I1000D10K	0,0005	0,01	416	10,463	397	9,688
6	T30I1000D10K	0,0005	0,01	347	8,048	334	7,627

Kết quả thực hiện hai thuật toán trong bảng 2.1 cho thấy thuật toán MCPSI hiệu quả hơn thuật toán Apriori-Inverse không chỉ ở số lượng tập Sproradic tuyệt đối hai

ngưỡng đóng tìm được ít hơn so với tập Sporadic tuyệt đối mà còn ở thời gian thực hiện của thuật toán.

Thử nghiệm trên CSDL thực:

Dữ liệu thử nghiệm thuật toán là 6 tệp dữ liệu lấy từ nguồn [100]. Tệp ban đầu được chuyển sang dạng CSDL tác vụ. Thông tin về các CSDL, kết quả thực hiện thuật toán MCPSI và thuật toán Apriori-Inverse được mô tả trong hình 2.2.



Hình 2.2: Số tập Sporadic tuyệt đối và Sporadic tuyệt đối hai ngưỡng đóng trên các CSDL thực

2.2. Luật kết hợp Sporadic không tuyệt đối hai ngưỡng

2.2.1. Giới thiệu về luật kết hợp Sporadic không tuyệt đối hai ngưỡng

Trong phần này, chúng tôi phát triển giải pháp hiệu quả cho việc tìm các luật Sporadic không tuyệt đối được đề xuất trong [50]. Cụ thể sẽ nghiên cứu xây dựng thuật toán tìm các tập Sporadic không tuyệt đối cho các luật kết hợp $A \rightarrow B$ sao cho:

$$\begin{cases} \text{conf}(A \rightarrow B) \geq \text{minConf}, \\ \text{minSup} \leq \text{sup}(A \cup B) < \text{maxSup}, \\ \exists x \in A \cup B,\ \text{sup}(x) \geq \text{max Sup}. \end{cases} \quad (2.2)$$

ở đây minSup, maxSup (minSup < maxSup) tương ứng được gọi là độ hỗ trợ cận dưới, cận trên và minConf là độ tin cậy cực tiểu của luật. Các luật kết hợp trong trường hợp này được gọi là luật Sporadic không tuyệt đối *hai ngưỡng*. Các tập Sporadic của các luật đó cũng được gọi là tập Sporadic không tuyệt đối hai ngưỡng.

2.2.2. Tập Sporadic không tuyệt đối hai ngưỡng

Định nghĩa 2.5: Tập X được gọi là tập Sporadic không tuyệt đối hai ngưỡng nếu:

$$\text{minSup} \leq \text{sup}(X) < \text{maxSup}, \text{ và}$$

$$\exists x \in X,\ \text{sup}(x) \geq \text{maxSup}$$

Định nghĩa 2.6: X là tập Sporadic không tuyệt đối hai ngưỡng, X được gọi là tập Sporadic không tuyệt đối hai ngưỡng đóng nếu nó là tập đóng, tức là $h(X) = X$.

Tính chất 2.5: *Độ hỗ trợ của tập Sporadic không tuyệt đối hai ngưỡng bằng độ hỗ trợ bao đóng của nó, tức là sup(X) = sup(h(X)).*

Định nghĩa 3.2: Tập Sporadic tuyệt đối hai ngưỡng mờ <Y,B> được gọi là tập con của <X,A> nếu $Y \subseteq X$ và $B \subseteq A$.

Tính chất 3.1: *Các tập Sporadic tuyệt đối hai ngưỡng mờ có tính chất Apriori, tức là tập con của tập Sporadic tuyệt đối hai ngưỡng mờ là tập Sporadic tuyệt đối hai ngưỡng mờ.*

3.2.3. Thuật toán tìm tập Sporadic tuyệt đối hai ngưỡng mờ

Ý tưởng của thuật toán:

Quá trình tìm tập Sporadic 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:

(a) Xây dựng tập mờ cho các thuộc tính phân loại và thuộc tính số.

(b) Chuyển CSDL ban đầu thành CSDL mờ.

(c) Tìm các tập Sporadic tuyệt đối hai ngưỡng mờ.

Thuật toán tìm tập Sporadic tuyệt đối hai ngưỡng mờ:

Đầu vào: CSDL **D**, minSup, maxSup
Kết quả: Tập các tập Sporadic tuyệt đối hai ngưỡng mờ.
Bước 1: Chuyển CSDL $\mathbf{D} \subseteq \mathbf{I} \times \mathbf{O}$ ban đầu thành CSDL mờ $\mathbf{D}_F \subseteq \mathbf{I}_F \times \mathbf{O}_F$
Bước này sử dụng cách chia khoảng và hàm thành viên như mô tả trong phần 1.3.3. Trong đó: $\mathbf{I}_F$ là tập các thuộc tính của $\mathbf{D}_F$, mỗi thuộc tính x_j của $\mathbf{I}_F$ được gắn với một tập mờ. Mỗi tập mờ có một ngưỡng ω_{x_j}.
Bước 2: Tìm các tập Sporadic tuyệt đối hai ngưỡng mờ có kích cỡ là 1:

```
1.  S1 = ∅
2.  for each item xj ∈ IF do begin
3.      if sup(xj) < maxSup and sup(xj) ≥ minSup
4.          then S1 = S1∪xj
5   end
```

Bước 3: Tìm tập Sporadic tuyệt đối hai ngưỡng mờ có kích cỡ là k (k ≥ 2):

```
6.  for (k = 2; Sk-1 ≠ ∅; k++) do begin
7.      Sk = ∅
8.      for each xj ∈ Ck (Ck là tập ứng cử viên sinh ra từ Sk-1) do begin
9.          if sup(xj) ≥ minSup
10.             then Sk = Sk∪xj
11.     end
12. end
13. return ⋃k Sk
```

Hình 3.1: Thuật toán MFPSI

Kết quả thử nghiệm: Để đánh giá hiệu quả thực hiện của thuật toán MFPSI, chúng tôi tiến hành thực nghiệm đối với CSDL thực Census Income từ nguồn [100].

Thuật toán tìm các tập phổ biến đóng với ràng buộc mục dữ liệu âm được gọi là NC-CHARM. Hình 2.4 thể hiện giả ngôn ngữ của thuật toán.

Độ phức tạp của thuật toán NC-CHARM: Độ phức tạp của NC-CHARM là O(l.|$\mathfrak{I}$|).|C|) với l là độ dài trung bình của các định danh, **C** là tập mục phổ biến đóng và |$\mathfrak{I}$| là số phần tử trong tập ràng buộc mục dữ liệu âm.

Kết quả thử nghiệm: Thuật toán NC-CHARM được thử nghiệm trên các CSDL giả định với ngưỡng minSup = 0,01. Tập ràng buộc âm được sinh ngẫu nhiên, bao gồm 100 điều kiện ràng buộc. Mỗi điều kiện ràng buộc có số mục dữ liệu được chọn ngẫu nhiên và không quá 5 mục dữ liệu. Kết quả của việc tìm các tập phổ biến thỏa mãn điều kiện ràng buộc âm được thể hiện ở bảng 2.4.

Bảng 2.4: Bảng kết quả thử nghiệm thuật toán NC-CHARM

TT	Tên CSDL	Số tập phổ biến tìm được	Thời gian (sec)
1	T05I1000D10K	4	4,210
2	T10I1000D10K	5	33,670
3	T15I1000D10K	8	82,340
4	T20I1000D10K	11	145,910
5	T25I1000D10K	13	163,650
6	T30I1000D10K	13	335,970

Chương 3 - PHÁT HIỆN LUẬT KẾT HỢP HIẾM TRÊN CƠ SỞ DỮ LIỆU ĐỊNH LƯỢNG

3.1. Giới thiệu về phát hiện luật kết hợp hiếm trên CSDL định lượng

3.2. Luật kết hợp Sporadic tuyệt đối hai ngưỡng mờ

3.2.1. Giới thiệu về luật Sporadic tuyệt đối hai ngưỡng mờ

Chúng tôi đề xuất bài toán phát hiện luật kết hợp mờ dạng r ≡ X is A → Y is B sao cho:

$$\begin{cases} \text{conf}(r) \geq m\text{inConf}, \\ \text{minSup} \leq \sup(< X \cup Y, A \cup B >) < \text{maxSup}, \\ \forall x \in < X \cup Y, A \cup B >, \text{minSup} \leq \sup(x) < \max \text{Sup}. \end{cases} \quad (3.1)$$

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. Các luật dạng này được gọi là luật Sporadic tuyệt đối *hai ngưỡng mờ* và bài toán trên được gọi là bài toán phát hiện luật Sporadic tuyệt đối hai ngưỡng mờ.

Luận án đã nghiên cứu đề xuất giải pháp tìm các tập Sporadic tuyệt đối mờ cho các luật Sporadic tuyệt đối mờ.

3.2.2. Tập Sporadic tuyệt đối hai ngưỡng mờ

Định nghĩa 3.1:Tập <X,A> được gọi là tập Sporadictuyệt đối hai ngưỡng mờ nếu:

minSup ≤ sup(<X,A>) < maxSup, và
∀x ∈<X,A>, sup(x) < maxSup.

Tính chất 2.6: *Tập các tập Sporadic không tuyệt đối hai ngưỡng cực đại và tập các tập Sporadic không tuyệt đối hai ngưỡng đóng cực đại là trùng nhau.*

Tính chất 2.7: *Các luật kết hợp được sinh ra từ các tập Sporadic không tuyệt đối hai ngưỡng và từ các tập Sporadic không tuyệt đối hai ngưỡng cực đại là như nhau.*

Các tính chất 2.6, 2.7 là cơ sở để đề xuất thuật toán tìm tập Sporadic không tuyệt đối hai ngưỡng dưới đây.

2.2.3. Thuật toán tìm tập Sporadic không tuyệt đối hai ngưỡng đóng

Đầu vào: CSDL **D**, minSup, maxSup
Kết quả: Tập các tập Sporadic không tuyệt đối hai ngưỡng đóng **CS**
MCISI ALGORITHM (**D**, minSup, maxSup):
1. FI = {$I_j \times g(I_j) : I_j \in \mathbf{I} \ \Lambda \ |g(I_j)| \geq \text{maxSup}$}
2. IFI = {$K_j \times g(K_j) : K_j \in \mathbf{I} \ \Lambda \ |g(K_j)| < \text{maxSup} \ \Lambda \ |g(K_j)| \geq \text{minSup}$}
3. for each $I_j \times g(I_j)$ in FI do begin
4. Nodes={$P_j \times g(P_j)$: $P_j = I_j \cup M_j$, $g(P_j) = g(I_j) \cap g(M_j)$, $M_j \in \text{FI} \setminus \{I_1,...,I_j\}$ or $M_j \in \text{IFI} \ \Lambda \ |g(P_j)| \geq \text{minSup}$}
/Kết hợp I_j với các mục dữ liệu còn lại ở bên phải mục đang xét trong FI và các mục dữ liệu trong IFI
5. MCISI-EXTEND(Nodes,**C**)
6. **CS** = **CS** ∪ **C**
7. end

MCISI-EXTEND(Nodes, **C**):
8. for each $X_i \times g(X_i)$ in Nodes do begin
9. NewN = ∅ ; X = X_i
10. for each $X_j \times g(X_j)$ in Nodes, with $k(j) > k(i)$ do begin
11. X = X ∪ X_j ; Y = $g(X_i) \cap g(X_j)$
12. CHARM-PROPERTY(Nodes, NewN)
13. end
14. if NewN ≠ ∅ then MCISI-EXTEND(NewN, **C**)
15. if sup(X) < maxSup then
16. **C** = **C** ∪ X // if X is not subsumed
17. end

Hình 2.3: Thuật toán MCISI

Độ phức tạp của thuật toán MCISI: Độ phức tạp của thuật toán MCISI là O(|FI|.l.|$\overline{C}$|), FI là tập các mục dữ liệu phổ biến theo maxSup, l là độ dài trung bình của các định danh và |$\overline{C}$| là kích thước trung bình của các tập Sporadic không tuyệt đối hai ngưỡng đóng tìm được.

Mệnh đề 2.2 *Thuật toán* MCISI *là đúng đắn và đầy đủ.*

Kết quả thử nghiệm:

a. Thử nghiệm trên tập dữ liệu giả định

Kết quả thử nghiệm thuật toán MCISI trên các CSDL với hai ngưỡng minSup và maxSup được chọn phù hợp trong việc tìm các tập hiếm thể hiện ở bảng 2.2.

Bảng 2.2: Bảng kết quả thử nghiệm trên CSDL giả định

TT	Tên CSDL	minSup	maxSup	Số tập Sporadic	Thời gian (giây)
1	T5I1000D10K	0,005	0,05	0	0,122
2	T10I1000D10K	0,005	0,05	5	1,652
3	T15I1000D10K	0,005	0,05	211	14,396
4	T20I1000D10K	0,005	0,05	1 841	52,020
5	T25I1000D10K	0,005	0,05	6 715	142,087
6	T30I1000D10K	0,005	0,05	15 593	315,711

Bảng 2.2 là kết quả thử nghiệm thuật toán MCISI trên các CSDL giả định với độ hỗ trợ cận dưới minSup = 0,005 và độ hỗ trợ cận trên maxSup = 0,05. Kết quả trong bảng 2.2 cho thấy thuật toán đã thực hiện được trên các tệp dữ liệu lớn với thời gian là thực hiện nhỏ.

b. Thử nghiệm trên CSDL thực

Bảng 2.3: Thông tin về CSDL thực và kết quả thử nghiệm

TT	Tên CSDL	Số mục dữ liệu	Số bản ghi	minSup	maxSup	Số tập Sporadic không tuyệt đối hai ngưỡng đóng	Thời gian thực hiện (giây)
1	Soybean	76	47	0,1	0,5	2 987	0,452
2	Mushroom	118	8 124	0,1	0,5	6 365	279
3	Zoo	43	101	0,1	0,5	3 125	0,515
4	Bridge	220	108	0,1	0,5	398	0,062
5	Teaching AE	104	151	0,1	0,5	5	0,027

2.3. Luật kết hợp với ràng buộc mục dữ liệu âm

2.3.1. Giới thiệu về luật kết hợp với ràng buộc mục dữ liệu âm

Giả sử $\mathbf{I} = \{i_1, i_2, \ldots, i_j, \ldots, i_n\}$ là tập các mục dữ liệu và được gọi là tập các mục dữ liệu dương. Ký hiệu $-i_j$ là ký hiệu mục dữ liệu âm của mục dữ liệu i_j và $\overline{I} = \{-i_1, -i_2, \ldots, -i_j, \ldots, -i_n\}$ được gọi là tập các mục dữ liệu âm của **I**, tập $\overline{B} \subseteq \overline{I}$ là ký hiệu tập mục dữ liệu âm của tập $B \subseteq \mathbf{I}$.

Luật kết hợp mẫu âm đã được quan tâm trong một số công trình nghiên cứu và nó có dạng tổng quát như sau: $A_1 \cup \overline{A}_2 \rightarrow B_1 \cup \overline{B}_2$, ở đây A_1, $B_1 \subseteq \mathbf{I}$, và $\overline{A}_2, \overline{B}_2 \subseteq \overline{I}$. Chẳng hạn luật $A \rightarrow \overline{B}$ có nghĩa là tập mục dữ liệu A xuất hiện trong tác vụ t thì các mục dữ liệu trong B sẽ không xuất hiện trong tác vụ này và do vậy $\sup(A \rightarrow \overline{B}) = \sup(A\overline{B}) = \sup(A) - \sup(AB)$.

Trong phần dưới đây sẽ trình bầy một dạng đặc biệt của luật kết hợp mẫu âm, đó là luật kết hợp với ràng buộc mục dữ liệu âm. Cụ thể luận án đã nghiên cứu giải quyết bài toán sau đây:

Phát hiện các luật kết hợp A→B với:

$\sup(A \cup B) \geq$ minSup; $\text{conf}(A \rightarrow B) \geq$ minConf *và trong điều kiện tồn tại một số ràng buộc mục dữ liệu âm.*

2.3.2. Tập phổ biến có ràng buộc mục dữ liệu âm

Ta gọi cặp $(A, \overline{B})$, trong đó $A \subseteq \mathbf{I}$ và $\overline{B} \subseteq \overline{I}$ là cặp ràng buộc mục dữ liệu âm nếu mỗi khi các mục dữ liệu trong A xuất hiện trong những tác vụ nào đó thì các mục dữ liệu trong B, với $A \cap B = \emptyset$, là không thể xuất hiện trong các tác vụ này.

Giả sử $\mathbf{D} \subseteq \mathbf{I} \times \mathbf{O}$ là CSDL tác vụ gồm các mục dữ liệu dương. Ký hiệu $\mathfrak{I} = \{(A_i, \overline{B_i}), i = 1,2, \ldots, k\}$ là tập tất cả các cặp ràng buộc mục dữ liệu âm cho trước.

Giả sử X là tập con bất kỳ của **I**, ký hiệu $Y = \{x \in \mathbf{I} \cup \overline{I} \text{ / nếu } x \in \mathbf{I} \text{ thì } x \in X$ hoặc nếu $x \in \overline{I}$ thì tồn tại cặp $(A_i, \overline{B_i}) \in \mathfrak{I}$ sao cho $x \in \overline{B_i}$ và $A_i \subseteq X\}$.

Mệnh đề 2.3. *Tập các tác vụ hỗ trợ X và Y xuất hiện là như nhau.*

Mệnh đề 2.4. *Bài toán tìm tập phổ biến từ CSDL **D** với tập điều kiện ràng buộc mục dữ liệu âm $\mathfrak{I}$ cho trước có thể được đưa về bài toán tìm tập phổ biến từ CSDL tác vụ có mục dữ liệu âm thích hợp. Ngược lại chưa chắc đúng.*

Mệnh đề 2.5. *Giả sử X, Y được xác định như trong Mệnh đề 2.3. Nếu X là tập phổ biến đóng cực đại trong CSDL tác vụ **D** và thoả mãn tập ràng buộc mục dữ liệu âm $\mathfrak{I}$ thì Y cũng là tập phổ biến đóng cực đại trong CSDL tác vụ có mục dữ liệu âm $\overline{D}$.*

2.3.3. Thuật toán tìm tập phổ biến với ràng buộc mục dữ liệu âm

Đầu vào: CSDL **D**, minSup, tập ràng buộc $\mathfrak{I}$
Kết quả: Tập các tập phổ biến đóng với ràng buộc mục dữ liệu âm **C**
NC-CHARM ALGORITHM(**D**, minSup, $\mathfrak{I}$):
1. Nodes = $\{I_j \times g(I_j) : I_j \in I \wedge |g(I_j)| \geq \text{minSup}\}$.
2. NC-CHARM-EXTEND(Nodes, $\mathfrak{I}$, **C**)

NC-CHARM-EXTEND(Nodes, $\mathfrak{I}$, **C**):
3. for each $X_i \times g(X_i)$ in Nodes do begin
4. NewN = $\emptyset$; X = X_i
5. for each $X_j \times g(X_j)$ in Nodes, with k(j) > k(i) do begin
6. X = X $\cup$ X_j ; Y = $g(X_i) \cap g(X_j)$
7. CHARM-PROPERTY(Nodes, NewN)
8. end
9. if NewN $\neq \emptyset$ then NC-CHARM-EXTEND(NewN, $\mathfrak{I}$, **C**)
10. temp = X
11. for each $(A_i, \overline{B_i}) \in \mathfrak{I}$ do
12. if $A_i \subseteq X$ then X = X $\cup$ $\overline{B_i}$
13. if X = temp then remove X × g(X) from Nodes
14. **C** = **C** $\cup$ X // if X is not subsumed
15. end

Hình 2.4: Thuật toán NC-CHARM