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

LUẬN ÁN TIẾN SĨ CÔNG NGHỆ THÔNG TIN

HÀ NỘI - 2013

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**

LUẬN ÁN TIẾN SĨ CÔNG NGHỆ THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC:

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

HÀ NỘI - 2013

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. Các kết quả được viết chung với các tác giả khác đều được sự đồng ý của các đồng tác giả trước khi đưa vào luận án. Các kết quả nêu trong luận án là trung thực và chưa từng được công bố trong các công trình nào khác.

Tác giả

Cù Thu Thủy

LỜI CẢM ƠN

Luận án được thực hiện tại Bộ môn Hệ thống thông tin - Khoa Công nghệ thông tin - Trường Đại học Công nghệ - Đại học Quốc gia Hà Nội, dưới sự hướng dẫn khoa học của PGS.TS. Đỗ Văn Thành và PGS.TS. Hà Quang Thụy.

Trước tiên tôi xin bày tỏ lòng biết ơn sâu sắc tới thầy Đỗ Văn Thành và thầy Hà Quang Thụy, những người đã đưa tôi đến với lĩnh vực nghiên cứu này. Các thầy đã tận tình giảng dạy, hướng dẫn giúp tôi tiếp cận và đạt được thành công trong công việc nghiên cứu của mình. Các thầy đã luôn tận tâm động viên, khuyến khích và chỉ dẫn giúp tôi hoàn thành được bản luận án này.

Tôi xin bày tỏ lòng biết ơn tới các Thầy Cô thuộc Khoa Công nghệ thông tin và cán bộ Phòng Đào tạo - Trường Đại học Công nghệ, đã tạo mọi điều kiện thuận lợi giúp đỡ tôi trong quá trình học tập và nghiên cứu tại trường.

Tôi xin cảm ơn TS. Yun Sing Koh và GS. TSKH. Marzena Kryszkiewicz đã chia sẻ những tài liệu và kinh nghiệm nghiên cứu.

Tôi xin chân thành cảm ơn PGS.TS. Hồ Thuần, PGS.TSKH. Nguyễn Xuân Huy, PGS.TS Đoàn Văn Ban, GS.TS Vũ Đức Thi, PGS.TS Lương Chi Mai, PGS.TS Đỗ Trung Tuấn, PGS.TS. Nguyễn Hà Nam đã đóng góp ý kiến quý báu giúp tôi hoàn thiện bản luận án.

Tôi xin cảm ơn tập thể cán bộ, giảng viên Khoa Hệ thống thông tin kinh tế, Ban Giám đốc Học viện Tài chính đã nhiệt tình ủng hộ, hết lòng tạo điều kiện giúp đỡ tôi trong suốt thời gian học tập và nghiên cứu.

Sự động viên, cổ vũ của bạn bè là nguồn động lực quan trọng để tôi hoàn thành luận án. Tôi xin bày tỏ lòng biết ơn sâu sắc tới gia đình, chồng và các con tôi đã tạo điểm tựa vững chắc cho tôi có được thành công như ngày hôm nay.

Tác giả

Cù Thu Thủy

MỤC LỤC

DANH MỤC CÁC KÍ HIỆU VÀ CHỮ VIẾT TẮT

Kí hiệu	Tiếng Anh	Tiếng Việt
CSDL	Database	Cơ sở dữ liệu
CPI	Consumer Price Index	Chỉ số giá tiêu dùng
GDP	Gross Domestic Product	Tổng sản phẩm quốc nội
CHARM	Closed Association Rules Mining	Phát hiện luật kết hợp đóng
conf	Confidence	Độ tin cậy
NC-CHARM	Negative Constrains - Closed Association Rules Mining	Phát hiện luật kết hợp đóng với ràng buộc mục dữ liệu âm.
minAS	Minimum absolute support	Độ hỗ trợ cận dưới
minConf	Minimum confidence	Độ tin cậy cực tiểu
minSup	Minimum support	Độ hỗ trợ cực tiểu. Trong luật kết hợp Sporadic hai ngưỡng sẽ được coi là độ hỗ trợ cận dưới.
maxSup	Maximum support	Độ hỗ trợ cận trên
MCISI	Mining Closed Imperfectly Sporadic Itemsets	Phát hiện tập mục Sporadic tuyệt đối đóng
MCPSI	Mining Closed Perfectly Sporadic Itemsets	Phát hiện tập mục Sporadic không tuyệt đối đóng
MFISI	Mining Fuzzy Imperfectly Sporadic Itemsets	Phát hiện tập mục Sporadic tuyệt đối mờ
MFPSI	Mining Fuzzy Perfectly Sporadic Itemsets	Phát hiện tập mục Sporadic không tuyệt đối mờ.
PPI	Producer Price Index	Chỉ số giá của người sản xuất
STR	Smooth Transition Regression	Hồi quy chuyển tiếp trơn
sup	Support	Độ hỗ trợ
WPI	Wholesale Price Index	Chỉ số giá bán buôn

DANH MỤC CÁC BẢNG

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

MỞ ĐẦU

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 → 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. Trong [49], Y. S. Koh và N. Rountree trình bầy khái quát về ứng dụng của khai phá luật hiếm, trong đó giới thiệu ví dụ luật kết hợp hiếm "máy pha cà phê" → "máy xay cà phê" có độ hỗ trợ rất thấp là 0,8% song có độ tin cậy khá cao tới 80% và giá trị bán hai mặt hàng này rất đáng kể. L. Szathmary và cộng sự [76] giới thiệu luật kết hợp hiếm "ăn chay" → "bệnh tim mạch" trong CSDL điều trị bệnh nhân Stanislas ở Pháp và luật kết hợp hiếm "thuốc hạ lipid trong máu Cerivastatin" → "tác động xấu khi điều 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) 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 kết hợp 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.

Mục tiêu cụ thể và phạm vi nghiên cứu của luận án

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.

Bài toán phát hiện luật kết hợp hiếm cũng được chia làm hai giai đoạn:

Giai đoạn 1: Tìm tất cả các tập mục dữ liệu để sinh ra các luật kết hợp hiếm. Các tập mục dữ liệu này được gọi là tập mục dữ liệu hiếm (hay tập hiếm).

Giai đoạn 2: Với mỗi tập hiếm tìm được ở giai đoạn 1, sinh ra tất cả các luật hiếm có độ tin cậy lớn hơn hoặc bằng độ tin cậy cực tiểu đã được xác định trước.

Trong hai giai đoạn trên thì giai đoạn 1 là khó khăn, phức tạp và tốn nhiều chi phí nhất. Giai đoạn thứ 2 có thể giải quyết đơn giản hơn khi tìm được tất cả các tập hiếm và độ hỗ trợ của chúng.

Tương tự như phát hiện luật kết hợp phổ biến, việc phát hiện luật kết hợp hiếm cũng có một phạm vi rất rộng. Trong luận án này, nghiên cứu sinh tập trung chủ yếu giải quyết giai đoạn 1 của bài toán phát hiện luật kết hợp hiếm. Cụ thể luận án phát triển giải pháp hiệu quả để tìm tập hiếm trên cả CSDL tác vụ và định lượng. Ở Việt Nam, đã có một số luận án tiến sĩ nghiên cứu về luật kết hợp [9, 10, 12] nhưng chưa có một luận án nào nghiên cứu về phát hiện luật kết hợp hiếm.

Ý nghĩa khoa học và thực tiễn của luận án

Về mặt khoa học, luận án đề xuất hướng tiếp cận phát hiện luật kết hợp hiếm trên CSDL tác vụ dựa trên không gian tập dữ liệu hiếm đóng. Nhờ đó, đã nâng cao hiệu quả của việc phát hiện luật kết hợp hiếm vì không gian các tập dữ liệu hiếm và đóng là nhỏ hơn không gian các tập dữ liệu hiếm. Luận án sử dụng lý thuyết tập mờ trong vấn đề phát hiện luật kết hợp hiếm trên CSDL định lượng.

Luận án có tính thực tiễn vì đã đề cập việc ứng dụng luật kết hợp cùng với mô hình hồi quy chuyển tiếp trơn để xây dựng mô hình phân tích và dự báo kinh tế.

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 phát hiện các tập dữ liệu hiếm tương ứng cho các dạng luật hiếm này.

Đối với bài toán phát hiện luật kết hợp hiếm trên CSDL tác vụ, luận án theo hướng tiếp cận đi tìm các tập không phổ biến đóng cho các luật kết hợp hiếm thay vì việc đi tìm tất cả các tập không phổ biến như các nghiên cứu về luật hiếm trước đây. Cơ sở của hướng tiếp cận này của luận án dựa trên các tính chất sau đây: (1) Tập tất cả các tập hiếm cực đại và tập tất cả các tập hiếm đóng cực đại là bằng nhau; (2) Các luật kết hợp hiếm được sinh ra từ các tập hiếm và từ các tập hiếm cực đại là như nhau. Tiếp cận nói trên là tương đồng với tư tưởng của thuật toán CHARM [94], là một trong những thuật toán hiệu quả nhất để phát hiện luật kết hợp mạnh trên CSDL tác vụ. Tập các tập không phổ biến đóng là nhỏ hơn tập các tập không phổ biến, vì vậy, 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 mục hiếm cho ba dạng luật kết hợp hiếm trên CSDL tác vụ là: thuật toán MCPSI (Mining Closed Perfectly Sporadic Itemsets) phát hiện tập mục Sporadic tuyệt đối hai ngưỡng [32], thuật toán MCISI (Mining Closed Imperfectly Sporadic Itemsets) phát hiện tập mục Sporadic không tuyệt đối hai ngưỡng [33] và thuật toán NC-CHARM (Negative Constrains - CHARM) phát hiện tập dữ liệu với ràng buộc mục âm [2]. Cả ba thuật toán trên đây được phát triển theo hướng bổ sung, phát triển các giải pháp cho phát hiện luật kết hợp Sporadic dựa theo cách tiếp cận và ý tưởng của thuật toán CHARM.

Đối với bài toán phát hiện luật kết hợp hiếm trên CSDL định lượng, luận án theo hướng tiếp cận tương tự như phát hiện luật kết hợp mạnh trên CSDL định lượng là 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. Tương tự như đối với luật kết hợp mạnh, việc ứng dụng tập mờ sẽ giúp biểu diễn luật kết hợp hiếm tự nhiên hơn, gần gũi hơn với người sử dụng và nhất là khắc phục được vấn đề "điểm biên gãy" trong

phân khoảng các thuộc tính định lượng. Hai dạng luật kết hợp Sporadic cho CSDL định lượng đã được luận án đề xuất là luật kết hợp Sporadic tuyệt đối hai ngưỡng mờ [3] và luật kết hợp Sporadic không tuyệt đối hai ngưỡng mờ [4]. Luận án đã phát triển hai thuật toán tìm tập hiếm cho hai dạng luật này. Thuật toán MFPSI (Mining Fuzzy Perfectly Sporadic Itemsets) phát hiện tập mục Sporadic tuyệt đối hai ngưỡng mờ [3] được phát triển theo tư tưởng của thuật toán Apriori [16], còn thuật toán MFISI (Mining Fuzzy Imperfectly Sporadic Itemsets) phát hiện tập mục Sporadic không tuyệt đối hai ngưỡng mờ [4] được phát triển theo tư tưởng của thuật toán của chúng tôi tìm tập hiếm cho luật Sporadic không tuyệt đối trên CSDL tác vụ [33].

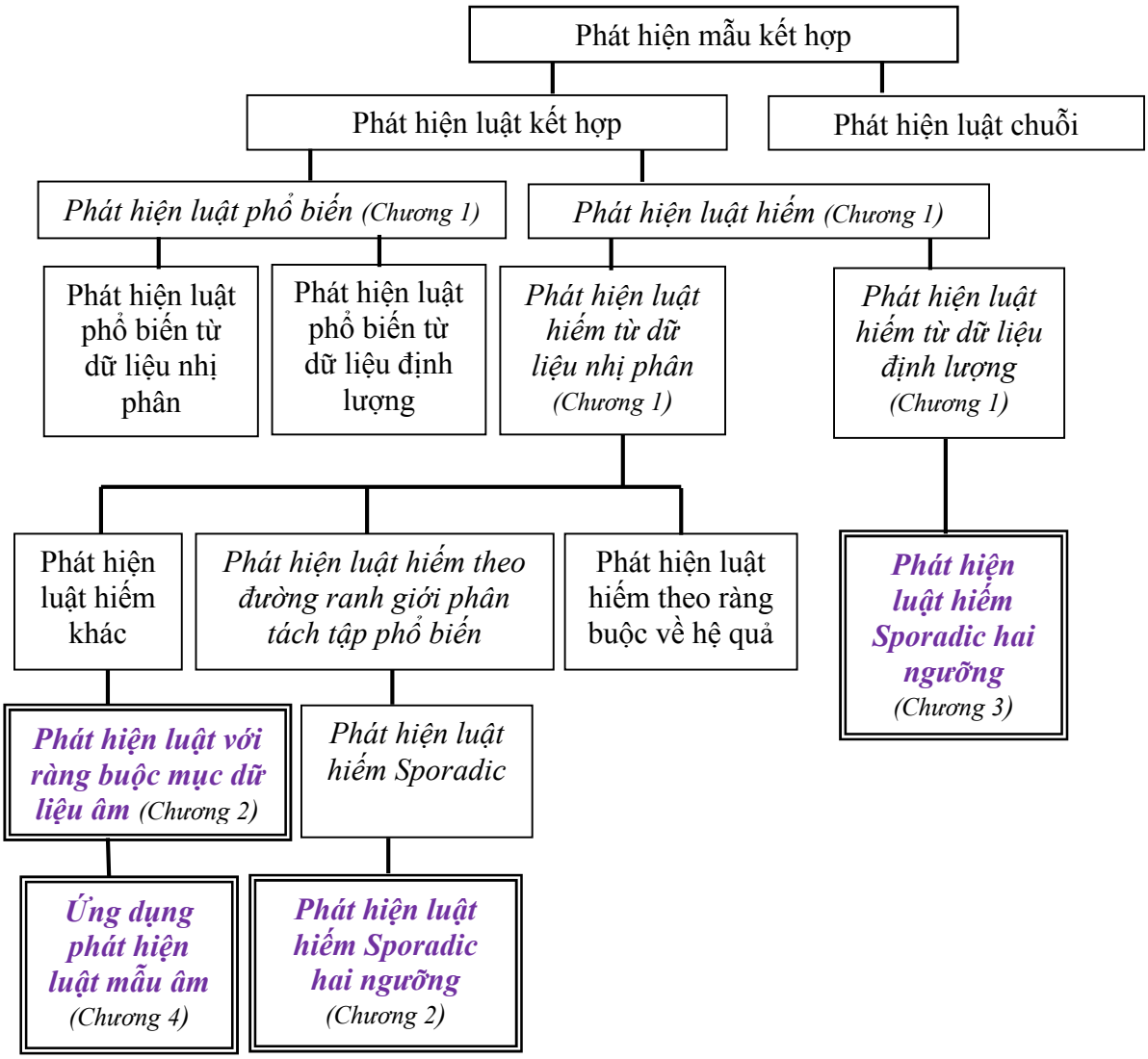
Về triển khai ứng dụng, luận án đã đề xuất kết hợp vấn đề phát hiện luật kết hợp mẫu âm trong công nghệ thông tin và mô hình hồi quy chuyển tiếp trơn phi tuyến trong kinh tế lượng để xây dựng mô hình phân tích và dự báo chỉ số giá tiêu dùng CPI và chỉ số chứng khoán Việt Nam. Kết quả dự báo kiểm định theo mô hình được xây dựng theo cách tiếp cận này 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].

Cấu trúc của luận án

Tiếp nối phần mở đầu này, nội dung chính của luận án được bố cục thành 4 chương và phần kết luận. Hình 0.1 trình bày phân bố các chủ đề phát hiện luật kết hợp được đề cập trong bốn 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. Các chương luận án là tổng hợp nội dung các bài báo công bố các kết quả nghiên cứu được thực hiện trong luận án (chương 2 với [2, 32-33], chương 3 với [3-4], chương 4 với [1, 7, 36]).

Phần kết luận tổng hợp các kết quả đạt được cũng như nêu lên một số hạn chế của luận án, và đồng thời trình bầy một số định hướng nghiên cứu trong tương lai.



Hình 0.1: Phân bố các chủ đề phát hiện luật kết hợp trong nội dung của luận án

Về khái niệm cơ sở dữ liệu tác vụ và cơ sở dữ liệu định lượng

Để phù hợp với nhiều công trình nghiên cứu về luật kết hợp, luận án sử dụng hai khái niệm *cơ sở dữ liệu tác vụ* và *cơ sở dữ liệu định lượng*. Hai khái niệm này mang nội dung như được giới thiệu dưới đây và phạm vi tác động của chúng được hạn chế trong luận án.

Trong công trình nghiên cứu khởi thủy về luật kết hợp, R. Agrawal và cộng sự (1993) đã giới thiệu bài toán phát hiện luật kết hợp trong CSDL tác vụ (a database of transactions) **D** [13], ở đó, mỗi tác vụ (transaction) t của CSDL được biểu diễn

bằng một dòng chứa một số mục dữ liệu. Do mỗi dòng này thực chất tương ứng với một vector nhị phân, nhận giá trị 1 hoặc 0, tuỳ thuộc mục dữ liệu có thuộc dòng hay không nên CSDL tác vụ còn được gọi là CSDL nhị phân (mỗi thuộc tính của CSDL nhận giá trị 1 hoặc 0). Giống như hầu hết các công trình nghiên cứu khác trước đó về luật kết hợp, luận án đã sử dụng khái niệm CSDL tác vụ (hay CSDL nhị phân) do R. Agrawal và cộng sự đề xuất trong [13].

Luận án cũng sử dụng khái niệm CSDL định lượng do R. Srikant và R. Agrawal (1996) đề xuất lần đầu trong [73] và cũng đã được hầu hết các nhà nghiên cứu về luật kết hợp sử dụng. Theo đó, cơ sở dữ liệu định lượng là CSDL có các thuộc tính nhận giá trị số hoặc giá trị phân loại (quantitative or categorical) [73].

Về ví dụ được sử dụng trong luận án

Hai CSDL trong hai ví dụ 0.1 và ví dụ 0.2 dưới đây được sử dụng xuyên suốt các chương của luận án (ngoại trừ các trường hợp chỉ rõ sử dụng CSDL khác).

Ví dụ 0.1: Bảng 0.1 biểu diễn một CSDL tác vụ ở đây: A, B, C, D, E, F,... được gọi là các mục dữ liệu (hay thuộc tính đối với CSDL nhị phân), t_i, i=1, 2,... được gọi là các tác vụ. Trong luận án này đã sử dụng ký hiệu **I** để biểu diễn tập các mục dữ liệu, ký hiệu **O** để biểu diễn tập các tác vụ và ký hiệu **D** để biểu diễn CSDL tác vụ. Trường hợp ví dụ 0.1, **I** = {A, B, C, D, E, F, G, H, J}, **O** =$\{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8\}$ và $\mathbf{D} \subseteq \mathbf{I} \times \mathbf{O}$.

Bảng 0.1: CSDL tác vụ

Tác vụ	Mục dữ liệu
t_1	A B C D H J
t_2	A E
t_3	A G J
t_4	A B C E F H J
t_5	E
t_6	A D E H
t_7	A C F J
t_8	E J

Ví dụ 0.2: Bảng 0.2 biểu diễn một CSDL định lượng với các thuộc tính Tuổi, Số xe máy, Thu nhập, Có gia đình.

Bảng 0.2: 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	có
t_6	57	4	4,0	có

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

Đầu tiên, chương này giới thiệu tổng quan về luật kết hợp: khái niệm luật kết hợp, bài toán phát hiện luật kết hợp, phương pháp chung phát hiện luật kết hợp, phát hiện luật kết hợp với độ hỗ trợ cực tiểu không giống nhau. Tiếp theo, vấn đề phát hiện luật kết hợp từ CSDL định lượng được trình bày. Phần cuối của chương sẽ trình bày về vấn đề phát hiện luật kết hợp hiếm: giới thiệu chung về luật kết hợp hiếm, một số hướng nghiên cứu chính và khuynh hướng nghiên cứu về 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 và các mối quan hệ này là có ích trong hỗ trợ quyết định. Trong CSDL siêu thị, việc phát hiện được quan hệ "78% số khách hàng mua sữa và đường cũng mua bơ" sẽ rất có ích cho quyết định kinh doanh, chẳng hạn, quyết định về số lượng nhập các mặt hàng này hoặc bố trí chúng tại các ngăn hàng liền kề nhau. Trong CSDL dân số, quan hệ "60% số người lao động ở độ tuổi trung niên có thu nhập thấp hơn mức thu nhập bình quân" sẽ rất có ích cho việc điều chỉnh chính sách thu nhập [13, 14, 16].

Khái niệm luật kết hợp (Association Rule) và phát hiện luật kết hợp (Association Rule Mining) được Rakesh Agrawal và cộng sự đề xuất lần đầu tiên vào năm 1993 nhằm phát hiện các mẫu có giá trị trong CSDL tác vụ (transaction database) tại siêu thị [10]. Bài toán này được phát biểu hình thức như dưới đây.

Kí hiệu $\mathbf{I} = \{i_1, i_2,..., i_n\}$ là tập các mục dữ liệu (mỗi mặt hàng trong siêu thị chính là một mục dữ liệu, và cũng có thể xem nó là một thuộc tính nhận giá trị nhị phân, khi đó **I** là các thuộc tính của CSDL); 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 dữ liệu $X \subseteq \mathbf{I}$, 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 mối quan hệ giữa hai tập mục dưới dạng $X \rightarrow Y$, trong đó $X \subseteq \mathbf{I}$, $Y \subseteq \mathbf{I}$, $X \cap Y = \varnothing$. 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 $X \cup Y$.

$$\sup(X \rightarrow Y) = \frac{|X \cup Y|}{|D|} \qquad (1.1)$$

Theo Agrawal R. và cộng sự [13], luật kết hợp được phát hiện cần đáp ứng ràng buộc độ hỗ trợ (support constraint), theo đó, độ hỗ trợ của tập mục $W = X \cup Y$ (hợp tập tiền đề và tập hệ quả của luật) 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 (frequent itemset) và còn được gọi là tập mục lớn (large itemset).

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

$$\text{conf}(X \rightarrow Y) = \frac{\sup(X \cup Y)}{\sup(X)} \qquad (1.2)$$

Luật kết hợp được phát hiện cần có tính tin cậy, theo đó nó cần có độ tin cậy vượt qua (không nhỏ thua) một ngưỡng tin cậy tối thiểu do người dùng đưa vào. Luật đáp ứng ràng buộc độ hỗ trợ và có tính tin cậy được gọi là luật mạnh (strong association rule).

1.1.2. Quy trình hai bước phát hiện luật kết hợp

Mục đích của bài toán phát hiện luật kết hợp trong CSDL tác vụ **D** là đi tìm tất cả các luật kết hợp mạnh (độ hỗ trợ cực tiểu và độ tin cậy cực tiểu do người sử dụng

đưa ra trong quá trình phát hiện luật). Rất nhiều giải pháp phát hiện luật kết hợp đã được đề xuất, chẳng hạn, theo thống kê của MicroSoft [101], đã có 2671 tác giả công bố 1526 công trình khoa học có giá trị (với 10224 lần được chỉ dẫn) về phát hiện luật kết hợp. Phần lớn các thuật toán phát hiện luật kết hợp chia quá trình giải bài toán này thành hai giai đoạn như sau:

(1) Giai đoạn 1: Tìm tất cả các tập phổ biến trong CSDL **D**.

(2) Giai đoạn 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. Bài toán tìm tập phổ biến trong không gian các tập con của tập mục **I** có độ phức tạp tính toán là $O(2^{|\mathbf{I}|})$. Giai đoạn 2 được giải quyết đơn giản hơn khi đã có các tập phổ biến và độ hỗ trợ của chúng.

Các phần tiếp theo sẽ trình bày một cách cơ bản, tóm lược về tiến trình phát triển nghiên cứu về luật kết hợp. Ban đầu là nghiên cứu phát hiện luật kết hợp trong các CSDL tác vụ, có độ hỗ trợ cực tiểu chung như nhau và chúng đều là các luật mạnh,..., tiếp theo được mở rộng sang CSDL định lượng, và/hoặc độ hỗ trợ cực tiểu của các luật kết hợp là không giống nhau và/hoặc các luật kết hợp là luật hiếm,... Nói cách khác nghiên cứu phát hiện luật kết hợp càng càng được phát triển để thích ứng với nhu cầu đa dạng của thực tiễn.

1.2. Phát hiện luật kết hợp từ CSDL tác vụ

Phát hiện luật kết hợp trong CSDL tác vụ được khởi đầu từ phát hiện luật kết hợp với một ngưỡng độ hỗ trợ, và sau đó, tới phát hiện luật kết hợp với độ hỗ trợ khác nhau cho các mục dữ liệu.

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 phát hiện luật kết hợp tuân thủ theo quy trình chung hai bước, chủ yếu tập trung vào bước tìm ra tập các tập phổ biến, với ba hướng giải quyết:

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

1.2.1.1. Phát hiện luật kết hợp từ tất cả các tập phổ biến

Đây là cách tiếp cận nguyên thủy [13]. Các phương pháp thuộc cách tiếp cận này được chia thành các phương pháp duyệt không gian tìm kiếm và các phương pháp xác định trước độ hỗ trợ. Bỏ qua độ phức tạp vào – ra và tính toán khi duyệt CSDL, các thuật toán này đều thực hiện tìm kiếm trên cây các tập con của tập mục **I** vì vậy độ phức tạp tính toán là $O(2^{|\mathbf{I}|})$.

Phương pháp duyệt không gian tìm kiếm được chia thành hai nhóm tương ứng khi duyệt cây các tập mục: duyệt theo chiều rộng (Breadth First Search - BFS) và duyệt theo chiều sâu (Depth First Search - DFS).

Duyệt theo chiều rộng là duyệt theo kích thước k của các tập mục ứng viên lần lượt từ kích thước 1, 2.... Một số thuật toán phổ biến theo cách tiếp cận này là: Apriori [16], Partition [70],..., mà theo [88], thuật toán Apriori (hình 1.1, kí hiệu diễn giải ở trong bảng 1.1) được xếp vào tốp 10 thuật toán khai phá dữ liệu điển hình nhất.

Duyệt theo chiều sâu là duyệt xong các tập ứng viên liên quan với một tập mục phổ biến mới chuyển sang xem xét đối với tập phổ biến cùng kích thước khác. Các thuật toán điển hình theo cách tiếp cận này là: FP-Growth [42], ECLAT [96],...

Bảng 1.1: Bảng diễn giải các kí hiệu sử dụng trong thuật toán Apriori

Kí hiệu	Ý nghĩa
k-itemsets	Tập k – mục dữ liệu.
L_k	Tập các k - tập dữ liệu phổ biến. Mỗi một phần tử của tập này có 2 trường: i) tập dữ liệu và ii) độ hỗ trợ
C_k	Tập các k - tập dữ liệu ứng cử viên (tiềm năng là tập phổ biến). Mỗi một phần tử của tập này có 2 trường: i) tập dữ liệu và ii) độ hỗ trợ

Đầu vào: CSDL **D**, độ hỗ trợ cực tiểu minSup
Kết quả: Tập các tập phổ biến
Thuật toán Apriori tìm các tập phổ biến:

```
1. L1 ← {1- tập mục dữ liệu phổ biến }
2. for (k ← 2; Lk-1 ≠ ∅; k++) do begin
3.       Ck ← Apriori-gen(Lk-1);       // sinh các ứng cử viên k-mục dữ liệu
4.       forall transactions  t∈D do begin
5.                Ct ←  subset(Ck, t);  // ứng cử viên được chứa trong t
6.                       forall candidates c∈ Ct do
7.                                c.count++
8.       end
9.       Lk ←{c∈ Ck | c.count ≥ minSup}
10. end
11. Answer  ← ∪k Lk
```

Hàm Apriori-Gen sinh ra ứng cử viên:

```
Procedure apriori-gen(Lk-1)
insert into Ck                                  //bước kết nối
select p.item1, p.item2, ..., p.itemk-1, q.itemk-1
from Lk-1 p, Lk-1 q
where p.item1=q.item1,..., p.itemk-2=q.itemk-2, p.itemk-1<q.itemk-1;

forall itemsets c ∈ Ck do                       //bước tỉa
      forall (k-1)-subsets s of c do
        if (s ∉ Lk-1) then
            delete c from Ck;
```

Hình 1.1: Thuật toán Apriori

Đối với các CSDL lớn, độ phức tạp vào-ra và tính toán khi duyệt CSDL là rất lớn, cho nên các phương pháp xác định trước độ hỗ trợ được phát triển nhằm làm giảm số lần truy cập CSDL. Một số phương pháp cho phép xây dựng các cấu trúc dữ liệu trung gian thích hợp để phát hiện các tập phổ biến chỉ cần hai lần duyệt CSDL [70, 71].

1.2.1.2. Phát hiện luật kết hợp từ các tập phổ biến đóng

Như đã biết, bài toán tìm tập phổ biến nói chung có độ phức tạp tính toán $O(2^{|\mathbf{I}|})$. Một trong các hướng giảm độ phức tạp tính toán là phát triển các phương pháp giảm số lượng tập mục phải duyệt.

M. J. Zaki và C. Hsiao [94] định nghĩa kết nối Galois và tập mục dữ liệu đóng, xây dựng dàn tập mục dữ liệu đóng để tìm tập phổ biến đóng cho phép giảm thiểu độ phức tạp tính toán do số lượng tập phổ biến đóng nhỏ hơn số lượng tập phổ biến. Về lý thuyết, kích cỡ của dàn tập mục đóng là $|L_C| = 2^K|\mathbf{D}|$ với K là độ dài của tập đóng cực đại. Kết quả thực nghiệm cho thấy tốc độ phát triển trung bình không gian tìm kiếm nhỏ hơn 2^K.

Một số thuật toán tìm tập phổ biến đóng thông dụng là: CHARM [94], CLOSE [64], CLOSET+ [65],... Thuật toán CHARM được đánh giá là thuật toán hiệu quả nhất trong việc tìm các tập phổ biến đóng. Phần dưới đây sẽ trình bày về kết nối Galois và thuật toán CHARM [64, 94].

<u>*Kết nối Galois*</u>

Định nghĩa 1.1 (Ngữ cảnh khai phá dữ liệu): Ngữ cảnh khai phá dữ liệu là bộ ba $\hat{D}$ = (**O**, **I**, **R**), trong đó **O** là tập các tác vụ, **I** là tập các mục dữ liệu phổ biến theo minSup và $\mathbf{R} \subseteq \mathbf{I}\times\mathbf{O}$ là quan hệ nhị phân. Mỗi cặp $(i,t)\in\mathbf{R}$ ký hiệu cho sự kiện tác vụ $t\in\mathbf{O}$ quan hệ với mục dữ liệu $i\in\mathbf{I}$.

Định nghĩa 1.2 (Kết nối Galois): Cho $\hat{D}$ = (**O**, **I**, **R**) là ngữ cảnh phát hiện dữ liệu. Với $O \subseteq \mathbf{O}$ và $I \subseteq \mathbf{I}$, xác định:

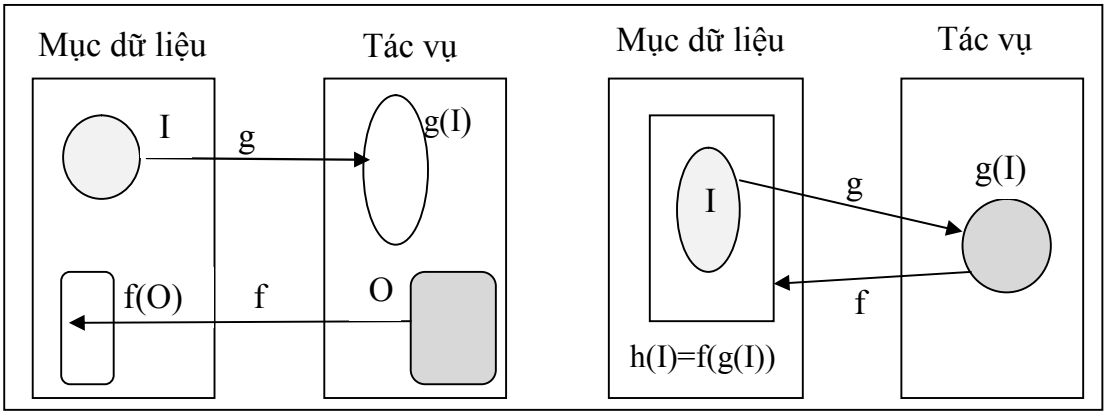
f: $2^{\mathbf{O}} \to 2^{\mathbf{I}}$ g: $2^{\mathbf{I}} \to 2^{\mathbf{O}}$

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

f(O) là tập mục dữ liệu chung cho tất cả các tác vụ của O và g(I) là tập các tác vụ 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 **I** (hình 1.2). Kí hiệu tập mục dữ liệu I và tập các tác vụ g(I) tương ứng với nó là I×g(I) và được gọi là cặp Tập mục dữ liệu × Tập định danh (IT-pair).

Toán tử $h = f_o g$ trong $2^{\mathbf{I}}$ và $h' = g_o f$ trong $2^{\mathbf{O}}$ gọi là toán tử đóng Galois.



Hình 1.2: Kết nối Galois và toán tử đóng Galois

Tính chất của kết nối Galois và toán tử đóng

Tính chất 1.1: Với I, I_1, $I_2 \subseteq$ **I** và O, O_1, $O_2 \subseteq$ **O**, ta có :

(1) $I_1 \subseteq I_2 \Rightarrow g(I_1) \supseteq g(I_2)$ (1') $O_1 \subseteq O_2 \Rightarrow f(O_1) \supseteq f(O_2)$

(2) $I \subseteq h(I)$ (2') $O \subseteq h'(O)$ (tính mở rộng)

(3) $h(h(I)) = h(I)$ (3') $h'(h'(O) = h'(O)$ (tính luỹ đẳng)

(4) $I_1 \subseteq I_2 \Rightarrow h(I_1) \subseteq h(I_2)$ (4') $O_1 \subseteq O_2 \Rightarrow h'(O_1) \subseteq h'(O_2)$ (tính đơn điệu)

(5) $h'(g(I)) = g(I)$ (5') $h(f(O)) = f(O)$

(6) $O \subseteq g(I) \Leftrightarrow I \subseteq f(O)$

Định nghĩa 1.3: (Tập mục dữ liệu đóng) Tập mục dữ liệu $X \subseteq$ **I** được gọi là tập đóng nếu X = h(X).

Tập X vừa là tập phổ biến vừa là tập đóng được gọi là tập phổ biến đóng.

Ví dụ 1.1: Xét CSDL trong ví dụ 0.1.

Với tập mục dữ liệu AJ, ta có: h(AJ) = f(g(AJ)) = f(1347) = AJ. Vậy AJ là tập mục dữ liệu đóng.

Với tập mục dữ liệu AC, ta có: h(AC) = f(g(AC)) = f(147) = ACJ. Vậy AC không là tập mục dữ liệu đóng.

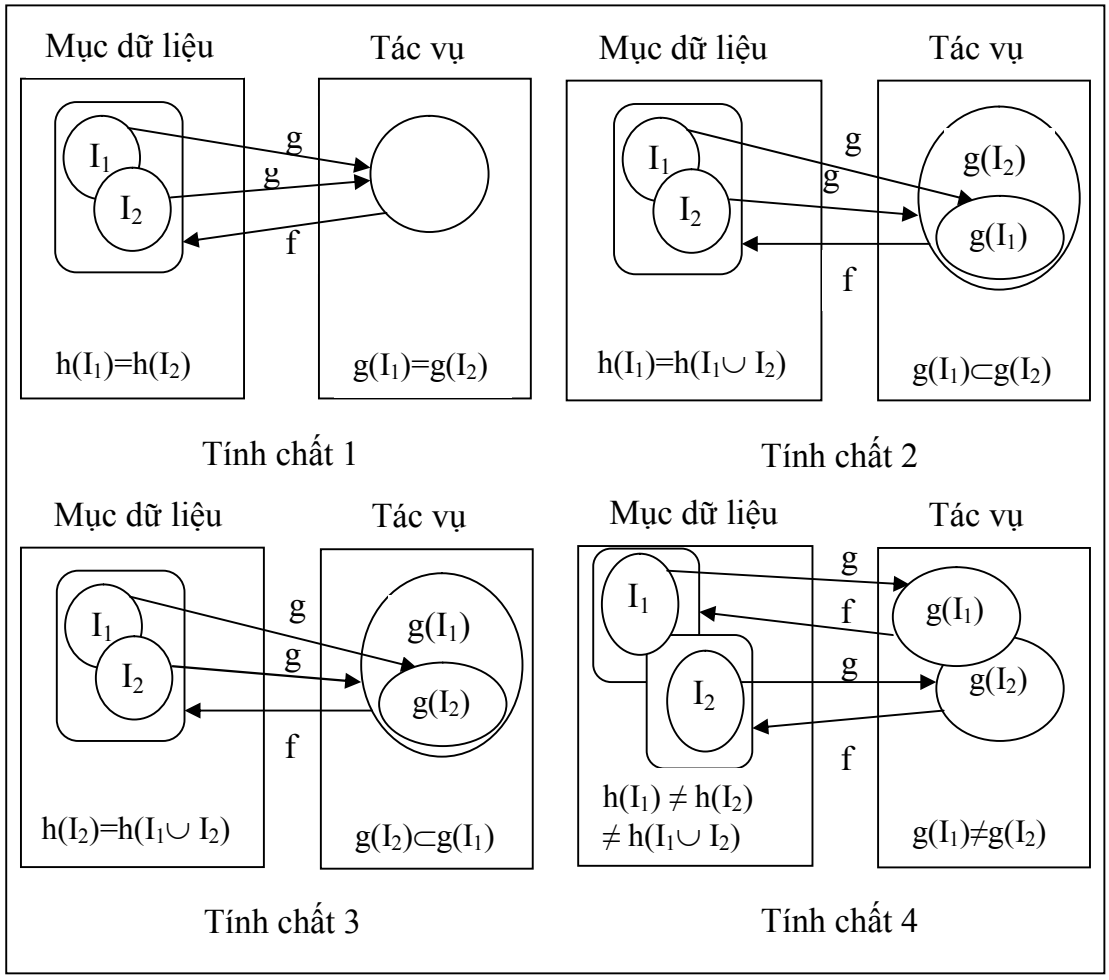
Nếu chọn ngưỡng độ hỗ trợ cực tiểu là 0,4 thì tập mục AJ là tập phổ biến đóng theo định nghĩa 1.3.

Tính chất 1.2: Độ hỗ trợ của tập mục dữ liệu I bằng độ hỗ trợ bao đóng của nó, tức là sup(I) = sup(h(I)).

Tính chất của các cặp Tập mục dữ liệu x Tập định danh:

Giả sử có ánh xạ k: $2^{\mathbf{I}} \rightarrow N$. Xét hai tập mục dữ liệu $I_1, I_2 \subseteq \mathbf{I}$, ta có $I_1 \leq I_2$ nếu và chỉ nếu $k(I_1) \leq k(I_2)$. Như vậy k là trật tự sắp xếp các mục dữ liệu (chẳng hạn, k có thể là sắp xếp theo trình tự từ điển của các mục dữ liệu hoặc sắp xếp theo trình tự tăng dần của độ hỗ trợ).

Giả sử ta đang thực hiện trên nhánh $I_1 \times g(I_1)$ của không gian tìm kiếm và muốn kết hợp nó với nhánh khác cùng mức $I_2 \times g(I_2)$ (với $I_1 \leq I_2$ theo trật tự k). Khi đó có bốn trường hợp xảy ra như sau (hình 1.3):



Hình 1.3: Tính chất của các cặp Tập mục dữ liệu × Tập định danh

(1) Nếu $g(I_1) = g(I_2)$ thì $g(I_1 \cup I_2) = g(I_1) \cap g(I_2) = g(I_1) = g(I_2)$. Do vậy ta có thể thay xuất hiện của I_1 bằng $I_1 \cup I_2$, loại bỏ I_2 trong các tập sẽ xét sau này vì I_1, I_2 cùng thuộc về tập đóng $I_1 \cup I_2$. $g(I_1)$ được thay thế bằng $g(I_1 \cup I_2)$.

(2) Nếu $g(I_1) \subset g(I_2)$ thì $g(I_1 \cup I_2) = g(I_1) \cap g(I_2) = g(I_1) \neq g(I_2)$. Khi đó mỗi xuất hiện của I_1 được thay thế bởi $I_1 \cup I_2$ vì nếu I_1 xuất hiện trong các tác vụ nào thì I_2 cũng xuất hiện trong các tác vụ đó. Nhưng do $g(I_1) \neq g(I_2)$ nên không thể loại bỏ I_2, nó sẽ sinh ra tập đóng khác.

(3) Nếu $g(I_1) \supset g(I_2)$ thì $g(I_1 \cup I_2) = g(I_1) \cap g(I_2) = g(I_2) \neq g(I_1)$. Trong trường hợp này mỗi xuất hiện của I_2 sẽ được thay thế bởi $I_1 \cup I_2$. Tuy nhiên I_1 vẫn được giữ lại vì sẽ sinh ra tập đóng khác.

(4) Nếu $g(I_1) \neq g(I_2)$ thì $g(I_1 \cup I_2) = g(I_1) \cap g(I_2) \neq g(I_2) \neq g(I_1)$. Trong trường hợp này không khử đi tập nào; cả I_1 và I_2 đều sinh ra các tập đóng khác.

Thuật toán CHARM:

Thuật toán CHARM được xây dựng dựa trên bốn tính chất của các cặp Tập mục dữ liệu × Tập định danh. Thuật toán cho kết quả là tập **C** gồm tất cả các tập phổ biến đóng. Mô tả thuật toán CHARM được thể hiện trong hình 1.4.

1.2.1.3. Phát hiện luật kết hợp từ các tập phổ biến cực đại

Với những CSDL quá dầy (mọi giao dịch đều có số lượng lớn các mặt hàng) thì số lượng tập phổ biến đóng cũng rất lớn và phương pháp chỉ tìm các tập phổ biến cực đại được đề xuất để khắc phục tình huống này.

Tập phổ biến X là cực đại nếu không có tập phổ biến khác chứa nó. Như vậy không gian tập phổ biến cực đại là nhỏ hơn không gian tập phổ biến đóng. Từ các tập phổ biến cực đại cho phép sinh ra được tất cả các tập phổ biến nhưng có hạn chế là không ghi được độ hỗ trợ của chúng [27, 37]. Một số thuật toán tìm tập phổ biến cực đại điển hình là Max-Miner [21], MAFIA [27], GENMAX [37]...

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.

Đầu vào: CSDL **D**, độ hỗ trợ cực tiểu minSup
Kết quả: Tập các tập phổ biến đóng **C**
CHARM($\mathbf{D} \subseteq \mathbf{I} \times \mathbf{O}$)
1. Nodes = $\{I_j \times g(I_j) : I_j \in \mathbf{I} \wedge |g(I_j)| \geq \text{minSup}\}$
2. CHARM-EXTEND(Nodes, **C**)

CHARM-EXTEND(Nodes, **C**):
3. for each $X_i \times g(X_i)$ in Nodes
4. NewN = $\emptyset$ and $X = X_i$
5. for each $X_j \times g(X_j)$ in Nodes, with $j > i$
6. $X = X \cup X_j$ and $Y = g(X_i) \cap g(X_j)$
7. CHARM-PROPERTY(Nodes, NewN)
8. if NewN $\neq \emptyset$ then CHARM-EXTEND(NewN)
9. $\mathbf{C} = \mathbf{C} \cup X$ // if X is not subsumed

CHARM-PROPERTY(Nodes, NewN):
10. if ($|Y| \geq$ minSup then
11. if $g(X_i) = g(X_j)$ then //Property 1
12. Remove X_j from Nodes
13. Replace all X_i with X
14. else if $g(X_i) \subset g(X_j)$ then //Property 2
15. Replace all X_i with X
16. else if $g(X_i) \supset g(X_j)$ then //Property 3
17. Remove X_j from Nodes
18. Add $X \times Y$ to NewN
19. else if $g(X_i) \neq g(X_j)$ then //Property 4
20. Add $X \times Y$ to NewN

Hình 1.4: Thuật toán CHARM

1.2.2.1. Phát hiện luật kết hợp có ràng buộc mục dữ liệu

Phát hiện luật kết hợp trong CSDL sinh ra rất nhiều luật trong khi người sử dụng lại chỉ quan tâm đến một phần trong các luật được phát hiện, chẳng hạn, chỉ quan tâm đến các luật có chứa một mục dữ liệu cụ thể, vì vậy, các nghiên cứu phát hiện luật kết hợp theo ràng buộc mục dữ liệu ra đời.

R. Srikant và cộng sự [74] đề xuất ràng buộc mục dữ liệu dạng biểu thức nhị phân thể hiện việc có xuất hiện hay không các mục ở trong luật. Các tác giả giới thiệu ba cách kết hợp thuật toán Apriori với bước tiền xử lý để phát hiện luật dạng này và chỉ ra rằng việc kết hợp tiền xử lý làm giảm đáng kể thời gian thực hiện.

Trong [81], S.V. Tseng đề xuất bài toán tìm luật kết hợp thỏa mãn điều kiện IS ($IS \subset \mathbf{I}$), theo đó thuật toán đi tìm luật từ các tập dữ liệu chứa IS thỏa mãn độ hỗ trợ và độ tin cậy cực tiểu. Sau khi tiền xử lý để nhận được thông tin hữu ích về CSDL (như ước lượng số lượng luật ứng với điều kiện ràng buộc), áp dụng thuật toán Apriori để phát hiện hiệu quả các luật kết hợp.

1.2.2.2. Phát hiện luật kết hợp với độ hỗ trợ nhiều mức

Thực tiễn cho thấy, với cùng một CSDL, có thể có nhiều tập mục có tần suất xuất hiện rất cao nhưng nhiều tập mục khác lại có tần suất xuất hiện rất thấp và việc sử dụng một ngưỡng độ hỗ trợ (tương ứng với giả thiết tần suất xuất hiện của các mục là như nhau) là không hợp lý [5]. Hướng tiếp cận phát hiện luật kết hợp với độ hỗ trợ nhiều mức được đưa ra nhằm khắc phục điều bất hợp lý này, theo đó, người dùng có thể đưa ra ngưỡng độ hỗ trợ cực tiểu khác nhau cho từng mục dữ liệu. Bằng việc đặt độ hỗ trợ cực tiểu thấp cho các mục dữ liệu tần số thấp cho phép người sử dụng sẽ tìm được các luật kết hợp đa dạng hơn.

B. Lui và cộng sự [59] giới thiệu thuật toán MSApriori phát hiện luật kết hợp với độ hỗ trợ cực tiểu cho từng mục dữ liệu (mục dữ liệu i có độ hỗ trợ cực tiểu MIS(i); MIS - minimum item supports). Độ hỗ trợ của tập mục X = min{MIS (i): $\forall i \in X$}. Tương tự cho định nghĩa độ hỗ trợ cực tiểu của một luật. Như vậy, luật r: $a_1a_2\ldots a_k \rightarrow a_{k+1}\ldots a_r$ thỏa mãn độ hỗ trợ cực tiểu nếu có độ hỗ trợ lớn hơn hoặc bằng min(MIS(a_1), MIS(a_2), …, MIS(a_r)). Trong thực nghiệm, các tác giả xây dựng công thức thiết lập giá trị MIS cho từng mục dữ liệu dựa trên các tham số: (1) LS là ngưỡng độ hỗ trợ nhỏ nhất, (2) tham số β ($0 \leq \beta \leq 1.0$) và (3) tần số xuất hiện (tính theo %) của các mục dữ liệu trong CSDL. Hai tham số LS và β được người sử dụng xác định. Mặc dù thuật toán MSApriori có tính đến tần suất xuất hiện của các mục

dữ liệu trong CSDL nhưng chỉ tiêu để phát hiện luật phụ thuộc chủ yếu vào giá trị của tham số β chứ không phải là tần suất xuất hiện của từng mục dữ liệu.

H. Yun và cộng sự [93] chỉ ra rằng việc xác định giá trị tham số β trong thuật toán MSApriori phù hợp nhất sẽ gặp rất nhiều khó khăn, đặc biệt trong các CSDL có nhiều mục dữ liệu. Khắc phục hạn chế này, các tác giả đề xuất thuật toán RSAA (Relative Support Apriori Algorithm) nhằm sinh các luật trong đó có mục dữ liệu hiếm mà không sử dụng thêm tham số nào khác do người sử dụng đưa vào. Thuật toán sử dụng các tham số: (1) Độ hỗ trợ thứ nhất: là giá trị do người sử dụng đưa vào (giống như độ hỗ trợ dùng trong phát hiện luật kết hợp phổ biến); (2) Độ hỗ trợ thứ hai: là giá trị do người sử dụng đưa vào dùng để phát hiện luật kết hợp hiếm (độ hỗ trợ này luôn nhỏ hơn độ hỗ trợ thứ nhất); và (3) Độ hỗ trợ quan hệ (relative support) theo công thức do các tác giả đề xuất. Tương tự MSApriori, thuật toán RSAA cũng cần đặt ngưỡng độ hỗ trợ cực tiểu riêng cho mỗi mục dữ liệu. Việc sử dụng ngưỡng độ hỗ trợ cực tiểu khác nhau dẫn đến vấn đề cần phải có cách lựa chọn ngưỡng độ hỗ trợ cực tiểu phù hợp cho từng mục dữ liệu.

Z. Chunjiang và cộng sự [29] cũng sử dụng cách đặt ngưỡng độ hỗ trợ như trong [59]. Tuy nhiên thuật toán sử dụng cách tiếp cận FP-Tree nên thực hiện nhanh hơn thuật toán MSApriori.

R.U. Kiran và P.K. Reddy [46] chỉ ra một số hạn chế của thuật toán MSApriori. Để khắc phục hạn chế đó, các tác giả sử dụng tham số SD (support difference) để xác định độ hỗ trợ cực tiểu cho các mục dữ liệu. Tham số SD liên quan đến tần suất xuất hiện của mục dữ liệu vì vậy tập mục dữ liệu chứa mục dữ liệu này được xem như là tập phổ biến. Thuật toán IMSApriori (Improved Multiple Support Apriori Algorithm) được đề xuất để tìm các tập phổ biến [46].

Phát triển nghiên cứu này, hai tác giả trên [47] sử dụng mô hình ràng buộc cực tiểu (minimum constraint model) áp dụng cho các CSDL rất thưa và đề xuất thuật toán sử dụng tiếp cận của thuật toán FP-growth thay cho thuật toán Apriori.

1.2.2.3. Phát hiện luật kết hợp có trọng số

F. Tao và cộng sự [77] đưa ra phương pháp sử dụng độ hỗ trợ có trọng số. Mỗi mục/tập mục được gắn trọng số theo độ thú vị (interestingness) của nó. Tập mục càng thú vị thì cần được gắn trọng số lớn. Phát hiện luật kết hợp có trọng số đã dựa trên tính chất đóng của trọng số (weighted downward closure property). Thứ tự ưu tiên chọn các tập mục dựa vào tầm thú vị của nó thay vì tần suất xuất hiện. Như vậy, các luật kết hợp sinh ra theo hướng này phụ thuộc vào việc sử dụng trọng số. Để có được các luật kết hợp hữu ích thì phải xác định được cách thức gắn trọng số phù hợp cho các mục dữ liệu. Các tác giả đã đề xuất hai loại trọng số là: trọng số mục dữ liệu (item weight) và trọng số tập mục dữ liệu (itemset weight). Trọng số mục dữ liệu w(i) được gắn với một mục dữ liệu và thể hiện tầm quan trọng của nó (chẳng hạn trong CSDL siêu thị có thể chọn dựa vào lợi nhuận của mặt hàng). Trọng số tập mục dữ liệu được tính dựa vào trọng số của mục dữ liệu. Cách đơn giản nhất để xác định trọng số của tập mục dữ liệu là dựa vào giá trị trung bình cộng của trọng số mục dữ liệu.

Rahman và cộng sự [67] ứng dụng kỹ thuật xử lý song song phân chia CSDL ra nhiều bộ xử lý nhằm tìm song song các tập phổ biến. Hơn nữa các tác giả sử dụng kỹ thuật chỉ tìm các tập phổ biến đóng theo trọng số thay vì tìm tất cả các tập phổ biến nên đã cải thiện đáng kể chi phí phát hiện luật.

1.2.2.4. Phát hiện luật kết hợp có ràng buộc độ hỗ trợ

Sử dụng ràng buộc độ hỗ trợ giảm dần theo độ dài của tập mục (length-decreasing support constraint), M. Seno và G. Karypis đề xuất thuật toán LPMiner [72]. Thuật toán hướng tới việc tìm các tập mục dữ liệu thỏa mãn điều kiện f(l) với l là độ dài của tập mục dữ liệu, cụ thể $f(l_a) \geq f(l_b)$ với mọi l_a và l_b thỏa mãn điều kiện $l_a < l_b$. Các tác giả đưa ra ngưỡng độ hỗ trợ mà theo đó sẽ giảm dần theo chiều dài của tập mục dữ liệu. Một tập mục được coi là phổ biến nếu thỏa mãn ràng buộc độ hỗ trợ giảm dần theo độ dài của nó. Trái với cách tiếp cận truyền thống, tập mục được coi là phổ biến ngay cả khi tập con của nó là không phổ biến. Như vậy tính

chất đóng về độ hỗ trợ theo thuật toán Apriori đã không còn đúng. Để khắc phục vấn đề này, các tác giả đã phát triển tính chất giá trị nhỏ nhất (SVE - smallest valid extension). Cách tiếp cận này đề cao các tập mục nhỏ; tuy nhiên tập mục dài có thể rất hữu ích, ngay cả khi chúng ít phổ biến hơn. Thuật toán tìm ra các tập dài mà không cần phải sinh một số lượng lớn các tập ngắn tránh được sự bùng nổ số lượng lớn các tập mục nhỏ.

K. Wang và cộng sự [84] đề xuất ràng buộc độ hỗ trợ như là cách để xác định ràng buộc trong độ hỗ trợ cực tiểu. Ràng buộc độ hỗ trợ có dạng $SC_i(B_1,B_2,...,B_s) \geq \theta_i$, với $s \geq 0$, sẽ xác định tập mục nào thỏa mãn độ hỗ trợ cực tiểu. Mỗi B_j được gọi là một thùng (bin), là tập của các mục dữ liệu mà không cần phân biệt bằng việc chỉ rõ độ hỗ trợ cực tiểu. θ_i là độ hỗ trợ cực tiểu có giá trị trong đoạn [0, 1], hoặc là hàm xác định độ hỗ trợ cực tiểu. Bất cứ tập mục nào chứa ít nhất một mục dữ liệu trong B_j sẽ có độ hỗ trợ cực tiểu là θ_i. Tư tưởng chủ đạo của cách tiếp cận này là đưa ràng buộc độ hỗ trợ để cắt tỉa trong quá trình sinh tập mục. Nếu có nhiều hơn một ràng buộc được áp dụng cho một tập mục thì giá trị nhỏ nhất sẽ được chọn. Chẳng hạn, giả sử có 4 ràng buộc độ hỗ trợ $SC_1(B1, B3) \geq 0.2$, $SC_2(B_3) \geq 0.4$, $SC_3(B_2) \geq 0.6$, và $SC_0() \geq 0.8$. Nếu ta có tập mục chứa $\{B_1,B_2,B_3\}$ thì độ hỗ trợ cực tiểu là 0.2. Tuy nhiên nếu tập mục chứa $\{B_2,B_3\}$ thì độ hỗ trợ cực tiểu là 0.4. Độ hỗ trợ 0.8 sẽ được chọn cho tất cả các tập mục không chứa các ràng buộc trên (đây chính là ngưỡng độ hỗ trợ cực tiểu mặc định).

1.2.2.5. Phát hiện luật kết hợp không sử dụng độ hỗ trợ cực tiểu

E. Cohen và cộng sự [30] giới thiệu kỹ thuật tìm luật hỗ trợ có độ tin cậy cao và bỏ qua ràng buộc theo ngưỡng độ hỗ trợ. Các tác giả xem CSDL như một ma trận kích thước n×m (n: số lượng giao dịch, m: số lượng mục dữ liệu) gồm các phần tử có giá trị 0/1. Ma trận được giả định là "thưa" cho nên số lượng giá trị 1 trên một dòng (một giao dịch) có cỡ r (r<m). Độ tương tự của hai cột (mục dữ liệu) được tính bằng thương của số lượng hàng có giá trị 1 ở cả hai cột chia cho số lượng hàng chứa giá trị 1 hoặc ở một cột hoặc ở cả hai cột. Theo kỹ thuật này, đầu tiên, mọi cặp hai cột có độ tương tự vượt qua ngưỡng được xác định, và sau đó, mọi cặp hai cột có độ

tin cậy cao được xác định nhờ áp dụng giải pháp tỉa. Để kỹ thuật nói trên đáp ứng được với CSDL lớn (n cỡ 10^9, m cỡ 10^6, và r cỡ 10^2), các tác giả đề xuất cách tiếp cận ba giai đoạn: tính toán chữ ký băm cho các cột, sinh ứng viên và cắt tỉa. Việc loại bỏ ràng buộc về độ hỗ trợ là một giải pháp hay, nhưng lại có nhược điểm là chi phí xử lý cao.

K. Wang và cộng sự [83] chỉ ra rằng các phương pháp phát hiện luật kết hợp truyền thống là không hiệu quả với trường hợp phát hiện luật có độ hỗ trợ rất nhỏ hay không sử dụng độ hỗ trợ. Vì vậy, các tác giả đề xuất phương pháp tìm tất cả các luật thỏa mãn điều kiện độ tin cậy cực tiểu mà không xét đến ngưỡng độ hỗ trợ cực tiểu. Các luật thỏa mãn điều kiện này được gọi là "luật tin cậy". Khác với phương pháp dựa trên độ hỗ trợ, luật tin cậy không thỏa mãn tính chất đóng (vì luật r_1: Tuổi >35 Λ Giới tính = Nam $\rightarrow$ Lương = Thấp có độ tin cậy nhỏ hơn các luật: r_2:Giới tính = Nam $\rightarrow$ Lương = Thấp hay luật r_3: Tuổi >35 $\rightarrow$ Lương = Thấp).

Trong nghiên cứu này, các tác giả đã đề xuất phương pháp cắt tỉa dựa vào độ tin cậy để sinh luật. Giả thiết có 3 luật r_1, r_2 và r_3 mô tả như trên. Các luật r_2 và r_3 là hai trường hợp đặc biệt của luật r_1. Độ tin cậy của luật r_2 và r_3 phải lớn hơn hoặc bằng độ tin cậy của r_1. Vì vậy, có thể loại bỏ r_1 khi r_2 hoặc r_3 là không tin cậy. Từ nhận xét này các tác giả đưa ra quy tắc: Với mỗi thuộc tính a_i không xuất hiện ở trong luật $x \rightarrow c$ thì: (i) các luật có được bằng cách bổ sung thêm thuộc tính a_i vào phần tiền đề của luật có độ hỗ trợ ít nhất là bằng luật $x \rightarrow c$; (ii) Nếu luật $x \rightarrow c$ là luật tin cậy thì luật có được bằng cách bổ sung thêm thuộc tính a_i vào phần tiền đề cũng là luật tin cậy. Tính chất này còn được gọi là tính chất đóng không gian (universal-existential upward closure). Các tác giả sử dụng tính chất này để sinh các luật mà không sử dụng ràng buộc về độ hỗ trợ. Tuy nhiên cũng giống như trường hợp phát hiện luật dựa trên độ hỗ trợ, phương pháp này cũng yêu cầu nhiều bộ nhớ cho việc tìm các ứng cử viên trong quá trình thực hiện.

H. Xiong và cộng sự [90] nghiên cứu các tập mục dữ liệu trong đó chứa các mục với độ hỗ trợ ở các mức khác nhau. Các tác giả giới thiệu độ đo H-độ tin cậy để khai phá các mẫu có bó cụm cao (hyperclique). Các mẫu bó cụm cao là một dạng của luật kết hợp có chứa các đối tượng có liên kết cao với nhau, tức là, mỗi cặp các

đối tượng trong một mẫu bó cụm cao có đặc điểm giống nhau (hệ số tương quan) ở trên một ngưỡng xác định. H-độ tin cậy có đặc tính rất hữu ích trong việc loại bỏ các tập ứng cử viên có các mục dữ liệu có độ hỗ trợ khác nhau. H-độ tin cậy có tính chất anti-monotone (tức là nếu $P \subseteq P'$ thì $hconf(P) \geq hconf(P')$). Một mẫu bó cụm cao P là mẫu kết hợp có liên kết mạnh vì mỗi mục dữ liệu bất kỳ $x \in P$ trong một tác vụ hàm ý thể hiện $P\backslash\{x\}$ trong cùng tác vụ. Độ đo H-độ tin cậy được thiết lập nhằm lưu giữ những mối liên kết cao dạng này. Mặc dù đã có các mẫu bó cụm cao trong quá trình sinh luật chúng ta vẫn có thể bỏ qua các luật giá trị. Ví dụ, tập dữ liệu {A,B,C} tạo ra các luật có độ tin cậy thấp $A \rightarrow BC$, $B \rightarrow AC$ và $C \rightarrow AB$, nhưng luật có độ tin cậy cao $AB \rightarrow C$ có thể bị bỏ qua.

1.3. Phát hiện luật kết hợp từ CSDL định lượng

1.3.1. Phát hiện luật kết hợp 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 (số, 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 đã biết từ CSDL tác vụ (kiểu như thuật toán Apriori).

Phương pháp rời rạc hoá CSDL định lượng như sau:

Nếu A là thuộc tính định lượng rời rạc có tập giá trị $\{v_1, v_2, \ldots, v_k\}$ và k đủ bé thì biến đổi thuộc tính này thành k thuộc tính A_v_1, A_v_2, … A_v_k. Giá trị của bản ghi tại trường A_v_k bằng True (Yes hoặc 1) nếu giá trị thuộc tính A ban đầu là v_k, ngược lại nó sẽ nhận giá trị False (No hoặc 0) như bảng 1.2.

Bảng 1.2: Rời rạc hoá thuộc tính định lượng có số giá trị nhỏ

Thu nhập	rời rạc hoá →	Thu nhập: cao	Thu nhập: thấp
cao		1	0
thấp		0	1

Nếu A là thuộc tính số liên tục hoặc có giá trị rời rạc $\{v_1, v_2, \ldots, v_p\}$ với p lớn, thì ta ánh xạ thành q thuộc tính nhị phân <A: $start_1..end_1$>, <A: $start_2..end_2$>,…, <A: $start_q..end_q$>. Giá trị của bản ghi tại trường <A: $start_i..end_i$> sẽ bằng True (Yes hoặc 1) nếu giá trị ban đầu của nó tại trường A thuộc khoảng [$start_i..end_i$], ngược lại sẽ bằng False (No hoặc 0) như minh họa trong bảng 1.3.

Bảng 1.3: Rời rạc hoá thuộc tính định lượng có giá trị số

Tuổi	⟶	<Tuổi: 1..29>	<Tuổi: 30..59>	<Tuổi: 60..80>
70	rời rạc hoá	0	0	1
45		0	1	0
22		1	0	0
17		1	0	0

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

(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 do những giá trị rất gần nhau (hoặc tương tự nhau) của một thuộc tính lại nằm ở hai khoảng chia khác nhau, chẳng hạn khi rời rạc hoá thuộc tính tuổi ở trên, 59 tuổi được coi là "*trung niên*" trong khi 60 tuổi được xem là "*già*".

Để giải quyết tốt nhất vấn đề 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 [34, 38-41, 44, 45, 54, 55, 57, 61, 63, 82, 98].

1.3.2. Phát hiện 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 $X \subseteq \mathbf{I}$ 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 $\mathbf{D} \subset \mathbf{I} \times \mathbf{O}$ được gọi là CSDL định lượng. Giả sử mỗi thuộc tính i_k (k=1,…,m) có một 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 mờ tương ứng với thuộc tính i_k và $\chi_{i_k}^j$ là khái niệm mờ thứ j trong F_{i_k}. CSDL **D** có các thuộc tính gắn với tập mờ được gọi là CSDL mờ [54].

Theo C. M. Kuok và cộng sự [54], luật kết hợp mờ có dạng: X is A $\rightarrow$ Y is B với $X = \{x_1, x_2, ..., x_p\}$, $Y = \{y_1, y_2, ..., y_q\}$ là các tập thuộc tính, X$\cap$Y = $\varnothing$; A = $\{\chi_{x_1}, \chi_{x_2}, ..., \chi_{x_p}\}$, B = $\{\chi_{y_1}, \chi_{y_2} ..., \chi_{y_q}\}$ là một số 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 dữ liệu mờ <X, A> đối với CSDL **D** ký hiệu là sup(<X, A>) được xác định như sau:

$$\sup(<X, A>) = \frac{\sum_{t_i \in O} \otimes_{x_j \in X} \left\{ \int_{\chi_{x_j}} (t_i[x_j]) \right\}}{|O|} \quad (1.3)$$

trong đó: $\otimes$ là toán tử T-norm, $t_i[x_j]$ là giá trị của thuộc tính x_j trong bản ghi thứ i của **O**

$$\int_{\chi_{x_j}} (t_i[x_j]) = \begin{Bmatrix} m_{\chi_{x_j}}(t_i[x_j]) \; neu \; m_{\chi_{x_j}}(t_i[x_j]) \geq \omega_{\chi_{x_j}} \\ 0 \; neu \, nguoc \, lai \end{Bmatrix} \quad (1.4)$$

với $m_{\chi_{x_j}}(t_i[x_j])$ là hàm thành viên của thuộc tính x_j ứng với tập 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 $\rightarrow$ 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 công thức:

$$\text{conf}(<Z,C>) = \sup(<Z,C>)/\sup(<X,A>) \quad (1.5)$$

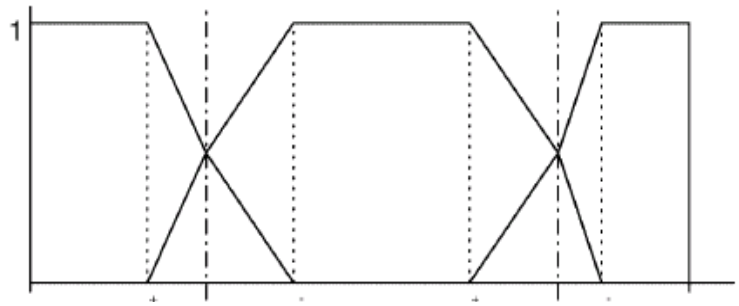
Luật kết hợp mờ X is A $\rightarrow$ Y is B được gọi là luật tin cậy nếu độ hỗ trợ và độ tin cậy của nó tương ứng lớn hơn hoặc bằng các ngưỡng độ hỗ trợ cực tiểu và độ tin cậy cực tiểu được xác định trước bởi người sử dụng.

1.3.3. Phân hoạch mờ

A. Gyenesei và J. Teuhola [41] đề xuất phương pháp phân hoạch mờ nhiều chiều nhằm phát hiện mẫu phổ biến mờ và luật kết hợp mờ. Phương pháp này sử dụng hướng tiếp cận từ trên xuống (top-down), trong đó sẽ lặp đi lặp lại việc đưa thêm các điểm phân chia mới cho các khoảng dựa trên việc đo ý nghĩa của nhiều biến. Ưu điểm của phương pháp là không cần tạo giả thiết về phân bố dữ liệu và về các khoảng nhỏ nhất.

Khi phân hoạch khoảng giá trị của mỗi thuộc tính thành một số khoảng mờ (hình 1.5) và chuyển các giá trị sang tương ứng trong CSDL với thuộc tính mờ (gọi tắt CSDL mờ) cần xây dựng hàm thành viên. Các tác giả đưa ra cách xây dựng hàm thành viên như sau:

Bước 1: Định nghĩa các đường biên của khoảng mờ



Hình 1.5: Minh họa về các phân hoạch mờ

Đường biên ở trên 1 (upper 1 - bound) được kí hiệu là d_i^+ cho một khoảng mờ i được tính như sau:

$$d_i^+ = s_i - 0.5(s_i - s_{i-1})p/100 \qquad (1.6)$$

trong đó: p là tham số chồng lấp (tính theo phần trăm)

s_{i-1} (s_i) là điểm chia bên trái (bên phải) của khoảng mờ i (i={1,2,…, m})

d_i^+ cũng là đường biên ở dưới 0 của khoảng mờ i+1

Đường biên ở dưới 1 (lower 1 – bound) được kí hiệu là d_j^- cho khoảng mờ j được tính như sau:

$$d_j^- = s_{j-1} + 0.5(s_j - s_{j-1})p/100 \qquad (1.7)$$

trong đó: p là tham số chồng lấp (tính theo phần trăm)

s_{j-1} (s_j) là điểm chia bên trái (bên phải) của khoảng mờ j (j={2, 3,…, m+1})

d_j^- cũng là đường biên ở trên 0 của khoảng mờ j-1

Bước 2: Xây dựng hàm thành viên cho mỗi khoảng mờ có giá trị thuộc [0,1] bằng cách sử dụng các đường biên định nghĩa ở bước 1. Các hàm thành viên thoả mãn điều kiện: với mỗi thuộc tính, tổng giá trị của các hàm thành viên là 1.

$$f(x) = \begin{cases} 0 & nếu \quad x \le d_{i-1}^+ \\ \dfrac{x - d_{i-1}^+}{2(s_{i-1} - d_{i-1}^+)} & nếu \quad d_{i-1}^+ < x \le s_{i-1} \\ \dfrac{x - s_{i-1}}{2(d^-_i - s_{i-1})} + \dfrac{1}{2} & nếu \quad s_{i-1} < x \le d_i^- \\ 1 & nếu \quad d_i^- < x \le d_i^+ \\ \dfrac{s_i - x}{2(s_i - d_i^+)} + \dfrac{1}{2} & nếu \quad d_i^+ < x \le s_i \\ \dfrac{d_{i+1}^- - x}{2(d_{i+1}^- - s_i)} & nếu \quad s_i < x \le d_{i+1}^- \\ 0 & nếu \quad d_{i+1}^- < x \end{cases} \qquad (1.8)$$

Hàm thành viên với các khoảng mờ i= 2…m

$$f(x) = \begin{cases} 1 & nếu \quad x \le d_1^+ \\ \dfrac{s_1 - x}{2(s_1 - d_1^+)} + \dfrac{1}{2} & nếu \quad d_1^+ < x \le s_1 \\ \dfrac{d_2^- - x}{2(d_2^- - s_i)} & nếu \quad s_1 < x \le d_2^- \\ 0 & nếu \quad d_2^- < x \end{cases} \qquad (1.9)$$

Hàm thành viên với khoảng mờ 1

$$f(x) = \begin{cases} 0 & nếu \quad x \le d_m^+ \\ \frac{x - d_{i-1}^+}{2(s_{i-1} - d_{i-1}^+)} & nếu \quad d_m^+ < x \le s_m \\ \frac{x - s_{i-1}}{2(d^-_i - s_{i-1})} + \frac{1}{2} & nếu \quad s_m < x \le d_{m+1}^- \\ 1 & nếu \quad d_{m+1}^- < x \end{cases} \qquad (1.10)$$

Hàm thành viên với khoảng mờ m+1

1.4. Phát hiện luật kết hợp hiếm

1.4.1. Giới thiệu chung về luật kết hợp hiếm

Luật kết hợp hiếm hàm ý chỉ các luật kết hợp không xảy ra thường xuyên trong các CSDL. Mặc dù ít khi xảy ra, nhưng trong nhiều trường hợp chúng lại là các luật rất có giá trị.

Phát hiện luật kết hợp hiếm là một phần của bài toán phát hiện luật kết hợp và hiện đang nhận được nhiều sự quan tâm của các nhà nghiên cứu. Luật kết hợp hiếm được ứng dụng ở nhiều các lĩnh vực khác nhau. Các luật hiếm sẽ giúp cho việc học phát âm từ, xác định ảnh hưởng của các hoạt động trong việc học trực tuyến đến kết quả đánh giá cuối cùng của sinh viên, xác định được các bệnh hiếm gặp trong y khoa, dự báo việc hỏng thiết bị truyền thông, phát hiện dấu hiệu tràn dầu trên hình ảnh vệ tinh, hay giúp xác định được các mặt hàng tuy ít xảy ra trong các giao dịch mua bán nhưng lại có giá trị lớn hoặc mang lại lợi nhuận cao trong kinh tế [21, 26, 46, 47, 49, 50, 53, 56, 58, 59, 66, 68, 72, 75, 76, 80, 83-85, 90, 93]. Như đã được giới thiệu, L. Szathmary và cộng sự [76] trình bày hai ứng dụng nổi tiếng của luật kết hợp hiếm, đó là luật kết hợp hiếm "ăn chay" $\rightarrow$ "bệnh tim mạch" trong CSDL điều trị bệnh nhân Stanislas ở Pháp và luật kết hợp hiếm giữa các loại thuốc hạ lipid trong máu Cerivastatin ảnh hưởng tới một số bệnh nhân (dẫn tới quyết định thu hồi loại thuốc này trên thị trường dược phẩm).

Phần lớn các thuật toán phát hiện luật kết hợp hiện nay thường chỉ thực hiện tìm các tập phổ biến cho các luật có độ hỗ trợ và độ tin cậy cao. Việc ứng dụng các thuật toán này, chẳng hạn như thuật toán Apriori để tìm tập hiếm (có độ hỗ trợ nhỏ

hơn một giá trị chung minSup nào đó cho trước) tương ứng với các luật 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 minSup), chi phí cho việc tìm kiếm sẽ tăng lên. Nhằm khắc phục những khó khăn này, các thuật toán riêng tìm các luật hiếm đã được phát triển theo cách tiếp cận chính được trình bày trong phần 1.4.2.

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

1.4.2.1. Sử dụng ràng buộc phần hệ quả của luật

Các phương pháp này đưa ra danh sách các mục sẽ xuất hiện trong một phần của luật và được sử dụng như là điều kiện trong quá trình sinh luật. Cách tiếp cận này chỉ hiệu quả khi biết trước được 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.

Phương pháp phát hiện luật kết hợp hiếm bằng cách cố định phần hệ quả được I. Rahal và cộng sự giới thiệu vào năm 2004 [66]. Các tác giả sử dụng kỹ thuật SE-tree và P-tree nhằm tìm các luật tin cậy nhỏ nhất sử dụng phần hệ quả cố định (fixed-consequent) mà không cần xác định ngưỡng độ hỗ trợ.

Giả sử có hai luật R_1 và R_2, với độ tin cậy lớn hơn độ tin cậy cực tiểu: R_1: A $\rightarrow$ C và R_2: AB $\rightarrow$ C, R_1 được cho là hay hơn vì phần tiền đề của luật R_1 là tập con của phần tiền đề của luật R_2. Độ tin cậy của luật R_1 là lớn hơn hoặc bằng độ tin cậy của luật R_2. R_1 được coi là luật nhỏ và R_2 được coi là luật không nhỏ (hay phức hợp).

J. Li và cộng sự [56], giới thiệu hướng tiếp cận khác là tìm các luật có độ tin cậy cao (100%) bằng cách sử dụng kỹ thuật phân hoạch CSDL và đường biên. Theo hướng này, các tác giả chỉ dùng ngưỡng độ tin cậy cực tiểu mà không dùng ngưỡng độ hỗ trợ cực tiểu. Tuy nhiên phần hệ quả của luật phải được xác định trước. Bằng cách thực hiện tương tự, phương thức để tìm các luật có độ tin cậy cao (chẳng hạn 90%) hay các luật có độ tin cậy bằng không cũng được giới thiệu. Phương pháp này còn được gọi là phương pháp EP (Emerging Pattern).

Trong [22], R.J. Bayardo và cộng sự chỉ ra rằng các tập phổ biến ứng viên tìm được để phát hiện luật là rất nhiều, nhất là trong các CSDL dầy. Nhằm hạn chế nhược điểm này, các tác giả đưa ra phương pháp tìm kiếm luật dựa trên ràng buộc phần hệ quả (consequent constraint) C trong quá trình phát hiện luật. Ràng buộc phần hệ quả được xác định bởi người sử dụng.

Các tác giả đã đưa ra một độ đo mới, được gọi là hệ số cải tiến (improvement). Tư tưởng chính của các tác giả là nhằm phát hiện các luật có độ tin cậy lớn hơn giá trị hệ số cải tiến cực tiểu.

Hệ số cải tiến của luật $A \rightarrow C$ được định nghĩa như sau:

$$Imp(A \rightarrow C) = \min\{conf(A \rightarrow C) - conf(A' \rightarrow C)\} \text{ với tất cả } A' \subset A \qquad (1.11)$$

Nếu hệ số cải tiến của một luật lớn hơn 0 thì loại bỏ các kết hợp không rỗng của các mục dữ liệu từ phần tiền đề của luật sẽ làm giảm độ tin cậy ít nhất là bằng hệ số cải tiến. Vì vậy, tất cả các mục dữ liệu và kết hợp của các mục dữ liệu trong phần tiền đề của luật với hệ số cải tiến lớn sẽ góp phần quan trọng trong việc dự báo. Ngược lại, với các luật có hệ số cải tiến âm được cho là các luật không mong muốn.

Các tác giả phát triển thuật toán Dense-Miner nhằm tìm tất cả các luật có phần hệ quả của luật là C và thỏa mãn 3 tham số do người sử dụng xác định là: độ hỗ trợ cực tiểu, độ tin cậy cực tiểu và hệ số cải tiến.

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

Theo hướng tiếp cận đường biên phân chia giữa tập phổ biến và tập không phổ biến, luật hiếm Sporadic tuyệt đối và không tuyệt đối do Y. S. Koh và cộng sự đề xuất [49, 50, 51] là một dạng luật hiếm thú vị được luận án này tập trung nghiên cứu sẽ được trình bày tại mục nội dung tiếp theo (mục 1.4.3).

Cũng theo hướng này trong [75, 76], L. Szathmary và cộng sự tiến hành phát hiện luật hiếm với độ hỗ trợ cực tiểu. Trong [75], các tác giả đưa ra phương pháp tìm tất cả các tập hiếm qua thi hành hai bước: (i) Tìm tất cả các tập hiếm cực tiểu;

Các tập này được coi như những bộ sinh cực tiểu để đi tìm các tập hiếm. (ii) Tìm tất cả các tập hiếm dựa trên tập hiếm cực tiểu.

Không gian tập hiếm được chia làm hai phần: tập hiếm có độ hỗ trợ "bằng không" và tập hiếm có độ hỗ trợ "khác không". Như vậy, toàn bộ không gian được chia làm 3 vùng. Đường biên phân chia giữa các vùng phụ thuộc vào giá trị của minSup. Mỗi vùng được phân định bởi hai tập là: tập các phần tử cực đại và tập các phần tử cực tiểu.

Phương pháp tìm các tập hiếm theo hướng tiếp cận bắt đầu từ dưới đi lên của không gian tìm kiếm, tức là bắt đầu từ vùng các tập phổ biến [75]. Đưa ra khái niệm đường biên âm (negative border) và đường biên dương (positive border) của các tập phổ biến; tương ứng là khái niệm đường biên dưới âm (negative lower border) và đường biên dưới dương (positive lower border) của các tập hiếm.

Hai thuật toán Apriori-Rare và MRG-Exp được đề xuất trong [75]. Thuật toán MRG-Exp được đánh giá hiệu quả hơn vì không cần duyệt tất cả các tập phổ biến mà chỉ tìm các tập sinh phổ biến. Đồng thời, các tác giả giới thiệu thuật toán ARIMA để tìm tất cả các tập hiếm có độ hỗ trợ khác không từ tập các tập hiếm cực tiểu. Thuật toán ARIMA cũng thực hiện tìm kiếm theo chiều rộng.

L. Szathmary và cộng sự chỉ ra một số hạn chế của nghiên cứu này là:

- Vì sinh ra tất cả các tập hiếm nên chi phí cho không gian nhớ là rất cao.

- Nếu trong CSDL chỉ có ít tập hiếm thì các tập này sẽ nằm ở phía trên của không gian vì vậy cách tìm kiếm từ dưới lên sẽ không hiệu quả.

- Để tính độ hỗ trợ của các tập mục thuật toán đã phải quét CSDL ở mỗi mức.

- Việc sinh các luật hiếm từ tất cả các tập hiếm sẽ tạo ra tập luật rất lớn.

Trong [76], L. Szathmary và cộng sự mở rộng một số nội dung nhằm khắc phục các hạn chế [75]. Các tác giả đã đạt được một số kết quả: (i) Sinh các luật hiếm có ý nghĩa một cách hiệu quả (ii) Các tập con của luật hiếm có thể tính toán được trực tiếp giống như với các luật phổ biến (iii) Thuật toán dễ thực hiện.

Quá trình phát hiện luật hiếm có giá trị được chia thành 3 giai đoạn:

(i) Thực hiện tìm tập các tập hiếm cực tiểu. Giai đoạn này sẽ sử dụng thuật toán MRG-Exp. Ban đầu thuật toán sẽ đi tìm các tập phổ biến sinh, sau đó tìm các tập hiếm sinh cực tiểu (mRGs). Thuật toán MRG-Exp sẽ giữ lại các tập mục này. Tập các tập hiếm cực tiểu sẽ giúp xác định tập các tập hiếm sinh cực tiểu.

(ii) Tìm các tập đóng của các tập hiếm sinh cực tiểu tìm được ở giai đoạn trước và vì vậy sẽ có được một lớp tương đương tương ứng.

(iii) Từ lớp tương đương hiếm tìm được sẽ sinh các luật hiếm giống như cách tìm các luật kết hợp không dư thừa cực tiểu. Các tác giả gọi các luật này là luật "mRG" vì phần tiền đề của luật là tập sinh hiếm cực tiểu.

Như vậy, L. Szathmary và cộng sự đã giới thiệu khá toàn diện phương pháp tìm luật hiếm có giá trị và được gọi là luật mRG. Các luật này có hai ưu điểm: (1) Chúng có thông tin cực đại (maximally informative) theo nghĩa đây là các luật có phần tiền đề là tập dữ liệu sinh và nếu bổ sung thêm phần hệ quả của luật vào thì sẽ tạo thành tập dữ liệu đóng. (2) Số lượng luật được sinh là tối thiểu, tức là các luật mRG là thể hiện rút gọn của tất cả các luật có độ tin cậy cao có thể sinh từ các tập hiếm cực tiểu.

L. Zhou và cộng sự [58] giới thiệu hai phương pháp tìm các luật kết hợp giữa các mục dữ liệu không phổ biến trên cả CSDL tác vụ và định lượng. Các tác giả sử dụng tham số interest(X,Y), hệ số tương quan correlation(X,Y), và tham số CPIR(Y\X) trong quá trình phát hiện luật. Định nghĩa luật có ý nghĩa giữa các tập không phổ biến: Giả sử **I** là tập các mục dữ liệu của CSDL **D**, $J = A \cup B$, $A \cap B = \varnothing$, $sup(A) \neq 0$, $sup(B) \neq 0$, các hệ số minSup, minConf, min-interest >0 do người sử dụng xác định. Nếu $sup(A) \leq minSup$, $sup(B) \leq minSup$, $interest(A,B) \geq$ min-interest, correlation(A,B)>1 và $CPIR(A,B) \geq minConf$ thì $A \rightarrow B$ là luật hiếm có ý nghĩa. Thuật toán MBS và Thuật toán HBS để thực hiện phát hiện luật hiếm trên CSDL tác vụ được đề xuất trong [58].

Gần đây, Troiano và cộng sự [80] giới thiệu thuật toán Rarity tăng tốc độ tìm ra tất cả các tập hiếm. Cũng sử dụng đường biên phân chia giữa các tập phổ biến và tập không phổ biến giống như trong thuật toán ARIMA, tuy nhiên, thuật toán Rarity

lại thực hiện chiến lược tìm kiếm bằng cách khác: bắt đầu từ các tập dữ liệu hiếm dài nhất ở đỉnh của không gian và tìm kiếm dần xuống. Trong quá trình duyệt không gian sẽ cắt tỉa các tập phổ biến và chỉ giữ lại các tập hiếm. Như đã biết, tập con của tập phổ biến là tập phổ biến. Tuy nhiên, tập con của tập không phổ biến chưa chắc là tập không phổ biến, vì vậy khác với các thuật toán khác, thuật toán Rarity thực hiện chiến lược tìm kiếm từ trên xuống trong không gian các tập mục mà ở đó các tập hiếm thường xuất hiện ở đỉnh của không gian. Để đánh giá hiệu quả của thuật toán Rarity, các tác giả đã tiến hành so sánh với thuật toán ARIMA. Kết quả thực nghiệm cho thấy thuật toán Rarity thực hiện nhanh hơn thuật toán ARIMA ở phần lớn các trường hợp nhưng lại yêu cầu nhiều bộ nhớ hơn. Khi độ hỗ trợ được thiết lập rất nhỏ so với kích cỡ của CSDL thì không thể so sánh được hiệu quả thực hiện của hai thuật toán. Nguyên nhân là do có quá nhiều tập phổ biến tìm được ở mỗi mức và sẽ có rất nhiều ứng cử viên tìm được ở các mức tiếp theo. Từ đó dẫn đến cần giảm số lượng tập dữ liệu con khi tính toán. Lựa chọn này dựa trên điều kiện là một tập mục dữ liệu sẽ là tập phổ biến nếu nó là tập con của tập phổ biến.

1.4.2.3. Phát hiện luật kết hợp hiếm từ các CSDL định lượng

Nhằm phát hiện luật kết hợp định lượng hiếm, cũng trong [58], L. Zhou và cộng sự đưa ra định nghĩa luật kết hợp định lượng có ý nghĩa.

Luật đơn giản (simple rule): Nếu tập mục định lượng X={(A=q1),(B=q2)} thỏa mãn QminSup, tức là sup(X) ≥ QminSup thì luật {A=q1} → {B=q2} là luật định lượng có ý nghĩa.

Luật chung (general rule): Nếu tập mục định lượng Y={(A=q1),(B≥q2)} thỏa mãn QminSup, tức là sup(Y) ≥ QminSup thì luật {A=q1} → {B ≥q2} là luật định lượng có ý nghĩa.

Luật ngữ nghĩa (semantic rule): Người sử dụng có thể sử dụng các cụm từ chỉ số lượng như: số lượng lớn, số lượng trung bình, số lượng nhỏ. Khi đó ta cũng có thể định nghĩa các luật định lượng dựa trên các thuật ngữ chỉ số lượng này, chẳng hạn luật {A=”Số lượng lớn”} → {B =”Số lượng nhỏ”}.

Bằng việc gắn số lượng đi cùng các mục dữ liệu và coi các mục dữ liệu với số lượng khác nhau là khác nhau, các tác giả có thể áp dụng thuật toán MBS (hoặc HBS) để sinh các luật hiếm định lượng.

Hai thuật toán MBS và HBS phát hiện luật kết hợp giữa các mục không phổ biến cũng có thể được dùng để tìm luật kết hợp giữa các mục phổ biến nhưng chỉ giới hạn với độ dài mục nhất định. Cả 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 hai 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 chi phí về bộ nhớ. Theo các tác giả, sử dụng ràng buộc nhằm giảm kích cỡ của các tập dữ liệu sinh là một định hướng nghiên cứu tiếp theo.

1.4.3. Luật hiếm Sporadic

Y.S. Koh và N. Rountree [49] đề cập bài toán phát hiện luật Sporadic, một kiểu luật kết hợp hiếm. 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:

$$\begin{cases} \text{conf}(X \rightarrow Y) \geq m\text{inConf}, \\ \text{sup}(X \cup Y) < \text{maxSup}, \\ \forall x \in X \cup Y, \text{sup}(x) < \text{max Sup}. \end{cases} \quad (1.12)$$

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").

Theo tư tưởng của thuật toán Apriori, Y.S. Koh và N. Rountree phát triển thuật toán Apriori-Inverse (hình 1.6) [49] tìm kiếm theo chiều rộng để tìm các tập Sporadic tuyệt đối. Nhằm loại bỏ các tập mục có độ hỗ trợ quá nhỏ, thuật toán dùng ngưỡng minAS (minimum absolute support) và kết quả của thuật toán là tập các tập mục có độ hỗ trợ nhỏ hơn maxSup nhưng lớn hơn minAS.

Đầu vào: CSDL **D**, ngưỡng maxsup
Kết quả: Tập các tập Sporadic tuyệt đối
(1) Generate inverted index I of (item, [TID-list]) from **D**
(2) Generate sporadic itemsets of size 1:
    $S_1 = Ø$
    for each item i ∈ I do begin
        if count(I,i)/|**D**| < maximum support and
          count(I,i) > minimum absolute support
        then $S_1 = S_1 \cup \{i\}$
    end
(3) Find S_k, the set of sporadic k-itemsets where k ≥ 2:
    for (k=2; $S_{k-1} \neq Ø$; k++) do begin
        $S_k = Ø$
        for each i ∈ {itemsets that are extens of S_{k-1}} do begin
            if all subsets of i of size k-1 ∈ S_{k-1}
              and count(I,i) > minimum absolute support
            then $S_k = S_k \cup \{i\}$
        end
    end
    return $\bigcup_k S_k$

Hình 1.6: Thuật toán Apriori-Inverse

Mở rộng luật Sporadic tuyệt đối, Y.S. Koh và cộng sự, 2008 [50] giới thiệu luật Sporadic không tuyệt đối và thuật toán phát hiện luật loại này.

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:

$$\begin{cases} \text{conf}(X \rightarrow Y) \geq m\text{inConf}, \\ \text{sup}(X \cup Y) < \text{maxSup}, \\ \exists x \in X \cup Y, \text{sup}(x) \geq \text{max Sup}. \end{cases} \quad (1.13)$$

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:

(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.

Đầu vào: CSDL **D**, tập mục **I**, maxsup, minconf, tham số θ
Kết quả: Tập các tập tiền đề của luật Sporadic không tuyệt đối
(1) Generate inverted index Idx of (item, [TID-list]) from **D**
    $N \leftarrow |\mathbf{D}|$
    $Idx \leftarrow$ invert (**D**,**I**)
(2) Generate candidate consequent items
    $C \leftarrow \{\{i\} | i \in \text{dom } Idx, \text{count}(i,Idx) < N.\text{maxsup}\}$
(3) Generate candidate antecedent itemsets
    $A \leftarrow \emptyset$
    for all items $i \in C$ do begin
        $k \leftarrow 1$
        $a \leftarrow \text{count}(\{i\}, Idx)$
        $A_{i,k} \leftarrow \emptyset$
        $U \leftarrow (\mathbf{U}_{i \in Idx(i)} D(tid)) \backslash \{i\}$
        $A_{i,k} \leftarrow \{j \mid j \in U, \text{count}(\{i,j\},Idx) > \text{minabssup}(N,a,b,\theta)\}$
        while $A_{i,k} \neq \emptyset$
            $k \leftarrow k + 1$
            $A_{i,k} \leftarrow \emptyset$
            $C_k \leftarrow \{x \cup y | x,y \in A_{i,k-1}, |x \cap y| = k - 2\}$
            $A_{i,k} \leftarrow \{j | j \in U, \text{count}(\{j\} \cup \{i\},Idx) > \text{minabssup}(N,a,b,\theta)\}$
        end
        $A \leftarrow A \cup \{\text{lhs} \rightarrow i | \text{lhs} \in \mathbf{U}_{m=1}^{k-1} A_{i,m}\}$
    end
    return A

Hình 1.7: Thuật toán MIISR

Các tác giả đã giới thiệu kỹ thuật để tìm các luật Sporadic không tuyệt đối "thú vị" (interestingness). 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 (hình 1.7) đã được đề xuất nhằm tìm phần tiền đề cho các luật dạng này.

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. Các luật hiếm thường là sự kết hợp của: (1) các mục dữ liệu hiếm; (2) các mục dữ liệu hiếm và các mục dữ liệu phổ biến; (3) các mục dữ liệu phổ biến, có độ hỗ trợ cao khi xét riêng từng mục dữ liệu, nhưng khi kết hợp lại tạo thành các tập mục có độ hỗ trợ nhỏ. Chính vì vậy không thể dùng các kỹ thuật phát hiện tập phổ biến thông thường để phát hiện các luật kết hợp hiếm. Độ hỗ trợ thấp của các tập mục gây trở ngại lớn cho quá trình phát hiện luật hiếm.

Trong [51], Y.S. Koh và cộng sự đã chỉ ra rằng: Phát hiện luật kết hợp hiếm yêu cầu kỹ thuật tiền xử lý khác so với việc phát hiện luật phổ biến. Mặc dù bắt đầu trong cùng vùng dữ liệu, nhưng tính chất của các luật là khác nhau. Các kỹ thuật tiền xử lý hiện nay nhằm hỗ trợ cho việc phát hiện luật phổ biến được thiết kế chỉ phù hợp với đặc tính của các luật phổ biến. 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. Theo cách này, kỹ thuật phát hiện luật kết hợp hiếm hiện tại đã sử dụng nhiều ngưỡng tuỳ ý (arbitrary thresholds) nhằm tìm ra các tập hiếm. Tuy nhiên kỹ thuật này lại không đưa ra được cách phát hiện nhiễu. Một vấn đề quan trọng trong tìm các tập hiếm là phân biệt được các tập mục giá trị từ các tập nhiễu. Cũng giống như phát hiện tập phổ biến có thể đưa vào các hình thức như: tập đóng, tập sinh,...

Hướng tiếp cận thứ hai là chỉ đi tìm các luật hiếm cụ thể. Phương pháp này trở thành cách tiếp cận phổ biến, đặc biệt với các CSDL y tế, ở đó các mục dữ liệu có thể là phổ biến khi xét độc lập nhưng là không phổ biến khi xét kết hợp cùng nhau. Chẳng hạn, hai triệu chứng thông thường kết hợp lại có thể tạo căn bệnh hiếm. Trong tình huống này, có một số luật hiếm nhưng chỉ có một luật được sinh. Những kỹ thuật gần đây chỉ cho phép chúng ta sinh ra luật con của những luật này. Tuy nhiên, không phải tất cả các luật hiếm đều có giá trị. Hiện vẫn chưa có kỹ thuật cho phép sinh ra tất cả các luật hiếm có giá trị. Một trở ngại trong việc phát hiện luật kết hợp hiếm là sẽ sinh ra rất nhiều luật và trong đó lại có nhiều luật không có ý nghĩa. CSDL thực chứa nhiều nhiễu. Một phần rất tự nhiên của các luật hiếm là chúng dễ bị che khuất bởi nhiễu, hoặc có thể chúng ta sẽ coi những luật nhiễu như là những luật có giá trị.

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. Các độ đo giá trị hiện tại áp dụng cho các luật phổ biến [20] thường không phù hợp khi xét kết hợp với những luật có độ hỗ trợ thấp (tức là các luật hiếm). Các kỹ thuật hiện tại được thiết kế dùng trong phát hiện luật kết hợp phổ biến không phù hợp khi áp dụng phát hiện các luật kết hợp hiếm.

Kết luận chương 1:

Nội dung chương 1 đã giới thiệu tổng quan các hướng nghiên cứu về phát hiện luật kết hợp từ CSDL tác vụ, phát hiện luật kết hợp từ CSDL định lượng và phát hiện luật kết hợp hiếm. Các kết quả nghiên cứu tiêu biểu của các tác giả liên quan trong từng phần nội dung đã được trình bày một cách tóm tắt. Từ phần nghiên cứu tổng quan này đã giúp nghiên cứu sinh có kiến thức và căn cứ cơ sở để lựa chọn và thực hiện hướng nghiên cứu của mình.

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

Chương 2 trình bày một số kết quả nghiên cứu nhằm phát hiện luật kết hợp hiếm trên CSDL tác vụ (thuộc tính nhận giá trị nhị phân): luật kết hợp Sporadic tuyệt đối hai ngưỡng, luật kết hợp Sporadic không tuyệt đối hai ngưỡng và luật kết hợp với ràng buộc mục dữ liệu âm. Kết quả nghiên cứu lần lượt đã được đăng trên kỷ yếu hội nghị quốc tế Management and Service Science - MASS 2010 [32], tạp chí International Journal of Computer Theory and Engineering [33] và tạp chí Tin học và Điều khiển học [2].

2.1. Luật kết hợp Sporadic tuyệt đối hai ngưỡng

2.1.1. Giới thiệu về luật Sporadic tuyệt đối hai ngưỡng

Luật Sporadic là một dạng luật kết hợp hiếm đã được giới thiệu trong phần 1.4.3 ở chương 1. Trong [49], Y.S. Koh đã đề xuất thuật toán Apriori-Inverse được phát triển từ thuật toán Apriori để tìm các tập phổ biến (được gọi là tập Sporadic tuyệt đối) cho các luật Sporadic tuyệt đối. Apriori là thuật toán có độ phức tạp trung bình so với các thuật toán khác tìm tập phổ biến cho các luật kết hợp bởi vậy Apriori-Inverse có khả năng chưa phải là thuật toán hiệu quả để tìm tập Sporadic tuyệt đối. Chúng tôi phát triển giải pháp hiệu quả hơn trong việc tìm các tập như vậy bằng cách đề xuất mở rộng bài toán phát hiện các luật kết hợp $A \rightarrow B$ sao cho:

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

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.

Độ hỗ trợ cận dưới minSup được đưa vào trước hết xuất phát từ nhận xét rằng một tập Sporadic tuyệt đối dù có độ hỗ trợ nhỏ đến đâu cũng phải dương.

Nhằm hạn chế các tập có độ hỗ trợ quá nhỏ không mong muốn, thuật toán Apriori-Inverse trong [49] chỉ đi tìm các tập Sporadic tuyệt đối ở đó độ hỗ trợ của mọi phần tử của nó không nhỏ hơn minAS. Đó là giá trị phụ thuộc vào từng CSDL cụ thể và được mặc định trong thuật toán này. Với việc bổ sung độ hỗ trợ cận dưới minSup, bài toán tìm tập Sporadic tuyệt đối trở thành một trường hợp đặc biệt của bài toán tìm tập Sporadic tuyệt đối hai ngưỡng khi độ hỗ trợ cận dưới bằng minAS.

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 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 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 loại bỏ được nhiều luật Sporadic tuyệt đối hai ngưỡng dư thừa.

Phần thử nghiệm cũng khẳng định lại điều đó. Việc tiến hành thử nghiệm cả hai thuật toán Apriori-Inverse và MCPSI trên cùng một số CSDL giả định và CSDL thực cho thấy trong mọi trường hợp thử nghiệm, thời gian thực hiện thuật toán MCPSI ít hơn thời gian thực hiện thuật toán Apriori-Inverse, đồng thời số lượng các tập Sporadic tuyệt đối hai ngưỡng đóng tìm được theo MCPSI cũng ít hơn số lượng các tập Sporadic tuyệt đối tìm được theo Apriori-Inverse.

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:

$minSup \leq sup(X) < maxSup$, và

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

Tập Sporadic tuyệt 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 **R** $\subseteq$ **INF**×**O** là quan hệ nhị phân. Mỗi cặp (i,t) $\in$ **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∈I ; ∀t∈O; (i,t)∈**R**} g(I) = {t | t∈O ; ∀i∈I; (i,t)∈**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 = $f_o g$ trong $2^{\mathbf{INF}}$ và h' = $g_o f$ trong $2^{\mathbf{O}}$ 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.

Tập Sporadic tuyệt đối hai ngưỡng đóng X được gọi là tập Sporadic tuyệt đối hai ngưỡng đóng cực đại nếu nó không phải là tập con thực sự của bất kỳ tập Sporadic không tuyệt đối hai ngưỡng đóng nào.

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.

- Theo định nghĩa 2.1, tập Sporadic tuyệt đối hai ngưỡng là tập không phổ biến theo ngưỡng maxSup nhưng là tập phổ biến theo ngưỡng minSup. Theo định

nghĩa 2.4, tập Sporadic tuyệt đối hai ngưỡng đóng cũng là tập phổ biến đóng theo độ hỗ trợ minSup.

Tính chất 2.1: *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.*

Chứng minh: Giả sử X là tập Sporadic tuyệt đối hai ngưỡng nào đó và tập X'⊆X, ta cần chứng minh X' cũng là tập Sporadic tuyệt đối hai ngưỡng.

Thật vậy do X'⊆X nên minSup≤ sup(X)< sup(X'). Mặt khác với mọi x∈X' thì x∈X nên sup(x)<maxSup và do đó sup(X')≤ sup(x)< maxSup. Từ đó suy ra X' 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 bằng độ hỗ trợ bao đóng của nó, tức là sup(X) = sup(h(X)).*

Chứng minh: Theo định nghĩa 2.3 thì sup(X) = |g(X)| và sup(h(X)) = |g(h(X))|. Vậy chỉ cần chứng minh g(X) = g(h(X)).

(i) Từ nhận xét 2.1, X là tập phổ biến theo ngưỡng minSup nên theo tính chất 1.1 – mục (2') ta có g(X) ⊆ h'(g(X)) = g(f(g(X)) = g(h(X)). Vậy g(X) ⊆ g(h(X)).

(ii) Theo tính chất 1.1 – mục (2) thì X ⊆ h(X) nên g(h(X)) ⊆ g(X) (tính chất 1.1 – mục (1)).

Từ (i) và (ii) có g(X) = g(h(X))■

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

Chứng minh: Giả sử X là tập Sporadic tuyệt đối hai ngưỡng cực đại bất kỳ. Theo tính chất 1.1- mục (2) ta có X⊆h(X).

(i) Theo tính chất 2.3 và do X là tập Sporadic tuyệt đối hai ngưỡng nên minSup ≤ sup(h(X)) = sup(X).

(ii) Mặt khác với mọi x∈h(X), sup(x)< maxSup là hiển nhiên vì h(X)⊆**INF** và theo định nghĩa của **INF**.

Từ (i) và (ii) suy ra h(X) là tập Sporadic tuyệt đối hai ngưỡng chứa X. Do X là tập Sporadic tuyệt đối hai ngưỡng cực đại nên suy ra X = h(X)■

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.*

Chứng minh: Ta chỉ cần chứng minh rằng mọi luật Sporadic tuyệt đối hai ngưỡng đều có thể được sinh ra từ các tập Sporadic tuyệt đối hai ngưỡng cực đại.

Giả sử $A \rightarrow B$ là luật như vậy, nên $A \cup B$ là tập Sporadic tuyệt đối hai ngưỡng và $A \rightarrow B$ là luật kết hợp theo độ hỗ trợ cực tiểu minSup và độ tin cậy cực tiểu minConf. Từ [64] suy ra rằng $A \rightarrow B$ cũng được sinh ra từ tập phổ biến cực đại với độ hỗ trợ cực tiểu là minSup.

Không giảm tính tổng quát ta có thể coi rằng $A \cup B$ là tập phổ biến cực đại theo độ hỗ trợ cực tiểu minSup và ta sẽ chứng minh $A \cup B$ là tập Sporadic tuyệt đối hai ngưỡng cực đại.

Giả sử ngược lại $\exists C$: $C \supset A \cup B$ sao cho $minSup \leq sup(C) < sup(A \cup B) < maxSup$, như vậy có nghĩa C là tập phổ biến theo độ hỗ trợ cực tiểu minSup thực sự chứa $A \cup B$. Điều này mâu thuẫn với giả thiết về $A \cup B$■

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

2.1.3.1. Ý tưởng của thuật toán

Thuật toán MCPSI được phát triển theo cách tiếp cận của thuật toán CHARM [94]. Thuật toán MCPSI tìm các tập Sporadic tuyệt đối hai ngưỡng đóng bằng cách: Xuất phát từ tập các mục dữ liệu không phổ biến theo maxSup nhưng phổ biến theo minSup, thuật toán sẽ duyệt theo chiều sâu của không gian tìm kiếm các tập phổ biến đóng theo ngưỡng minSup theo đúng tinh thần của thuật toán CHARM. Thuật toán sẽ duyệt và tìm kiếm các tập Sporadic tuyệt đối hai ngưỡng đóng trên cây tìm kiếm bao gồm cả tập mục dữ liệu và tập định danh của chúng. Tất cả các tập không phải là tập Sporadic tuyệt đối hai ngưỡng cũng như các nhánh không phải là tập đóng đều bị tỉa. Chiến lược tỉa được thực hiện trên cơ sở dựa vào 4 tính chất của các cặp tập mục dữ liệu và tập định danh được trình bày trong phần 1.2.1.2.

Thuật toán MCPSI có thể được tóm tắt như sau:

- Thuật toán bắt đầu bằng việc khởi tạo tập các mục dữ liệu và tập định danh của ngữ cảnh khai phá dữ liệu $\hat{D}$. Tập các mục dữ liệu này là không phổ biến theo maxSup nhưng là phổ biến theo minSup.

- Hàm MCPSI-EXTEND cho kết quả là tập các tập Sporadic tuyệt đối hai ngưỡng đóng **C**.

- Hàm CHARM-PROPERTY kiểm tra ràng buộc về giá trị cận dưới minSup và kiểm tra xem nút có thoả mãn các tính chất của cặp tập mục dữ liệu và tập định danh theo kết nối Galois hay không.

2.1.3.2. Thuật toán MCPSI

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
//k is a function for sorting items in Nodes
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

Độ phức tạp của thuật toán MCPSI: Thuật toán MCPSI dựa trên thuật toán CHARM để tìm kiếm các tập Sporadic tuyệt đối hai ngưỡng đóng và sự khác biệt nằm ở bước khởi tạo tập Nodes ban đầu. Bắt đầu từ tập các mục dữ liệu đơn cùng các định danh tương ứng, thuật toán thực hiện việc xử lý trên một nhánh sẽ có 4

trường hợp xảy ra. Kết thúc việc thực hiện, mỗi nút trên cây sẽ thể hiện tập dữ liệu đóng. Vậy thuật toán sẽ thực hiện O(|**C**|) phép giao, với |**C**| là cỡ của tập các tập Sporadic tuyệt đối hai ngưỡng đóng.

Nếu mỗi định danh có chiều dài trung bình là l, thì chi phí cho phép giao là 2.l. Vậy độ phức tạp của thuật toán MCPSI là O(2.l.|**C**|) hay O(l.|**C**|).

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

Tính đúng đắn:

Chứng minh: Cần chỉ ra rằng những tập tìm được bởi thuật toán MCPSI là tập Sporadic tuyệt đối hai ngưỡng đóng.

Thật vậy, thuật toán MCPSI gồm 2 giai đoạn chính.

Giai đoạn thứ nhất, dòng lệnh 1 khởi tạo không gian tìm kiếm tập phổ biến đóng theo độ hỗ trợ cận dưới minSup và độ hỗ trợ cận trên maxSup. Các mục dữ liệu được sắp xếp theo một trình tự nhất định.

Giai đoạn thứ hai, dòng lệnh 2 thực hiện hàm MCPSI-EXTEND. Hàm này tiến hành tìm các tập phổ biến đóng theo minSup nhưng không phổ biến theo maxSup. Cách thực hiện của hàm này là tương tự như hàm CHARM-EXTEND trong [94]. Hàm CHARM-PROPERTY sẽ kiểm tra ràng buộc về độ hỗ trợ theo ngưỡng minSup và kiểm tra xem nút có thoả mãn bốn tính chất về cặp tập mục dữ liệu và định danh hay không? Như vậy, kết thúc hàm MCPSI-EXTEND cho kết quả là tập các tập phổ biến đóng theo minSup và do nó chỉ bao gồm các mục dữ liệu có độ hỗ trợ nhỏ hơn maxSup nên độ hỗ trợ của tập này cũng nhỏ hơn maxSup. Tập này chính là tập các tập Sporadic tuyệt đối hai ngưỡng đóng theo định nghĩa 2.4 ở trên.

Tính đầy đủ:

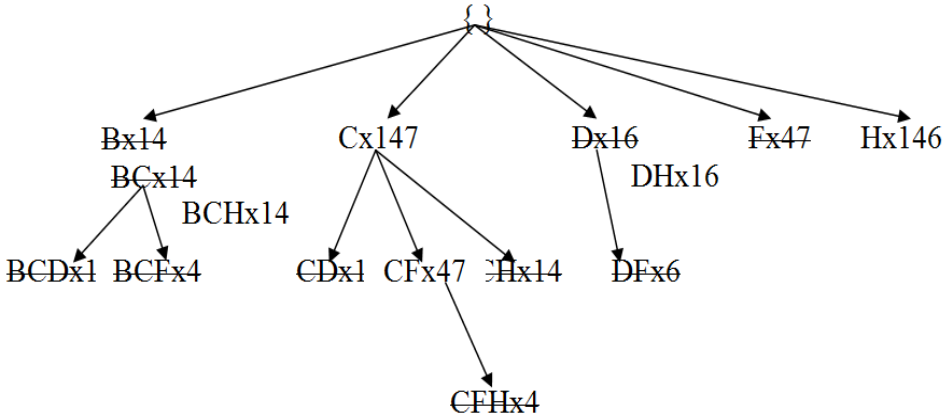
Chứng minh: Cần chỉ ra rằng mọi luật Sporadic tuyệt đối hai ngưỡng đều được sinh ra bởi một trong các tập Sporadic tuyệt đối hai ngưỡng được tìm bởi thuật toán MCPSI.

Thật vậy, theo tính chất 2.4 mọi luật Sporadic tuyệt đối hai ngưỡng đều được sinh ra bởi tập Sporadic tuyệt đối hai ngưỡng cực đại và theo tính chất 2.3 tập này cũng là tập Sporadic tuyệt đối hai ngưỡng đóng. Thuật toán MCPSI tìm các tập như vậy■

Ví dụ 2.1: Xét CSDL **D** được xác định như trong ví dụ 0.1.

Giả thiết xét với minSup = 0,25 và maxSup = 0,5.

Áp dụng thuật toán MCPSI với các mục dữ liệu được sắp xếp theo thứ tự tăng dần của từ vựng. Ban đầu khởi tạo tập Nodes = {B×14, C×147, D×16, F×47, H×146} (dòng 1)



Hình 2.2: Không gian tìm kiếm tập Sporadic tuyệt đối hai ngưỡng

Vì xét theo thứ tự tăng dần của từ vựng nên dòng 3 tiến hành tại nhánh B×14. Đặt X = B (dòng 4). Tiếp theo nút này sẽ được kết hợp với các nút lân cận ở bên phải (dòng 5). Khi kết hợp B với C vì g(B) ⊂ g(C) nên thay B bằng BC (X = BC). Khi kết hợp với D được tập BCD nhưng tập này có độ hỗ trợ nhỏ hơn minSup nên bị loại. Khi kết hợp với F được tập BCF cũng có độ hỗ trợ nhỏ hơn minSup nên cũng bị loại. Kết hợp với H, g(BC) ⊂ g(H) nên thay BC bằng BCH (X = BCH), tập này có độ hỗ trợ không nhỏ hơn minSup. Kết thúc trên nhánh B chỉ tìm được tập BCH. Tập mục dữ liệu BCH có sup(BCH) = 0,25 thỏa mãn điều kiện minSup ≤ sup(BCH) < maxSup và h(BCH) = f(g(BCH)) = f(14) = BCH. Theo định nghĩa 2.4 thì BCH là tập Sporadic tuyệt đối hai ngưỡng đóng.

Tiến hành tương tự như trên với các nhánh C×147, D×16, F×46 và H×146.

Kết thúc, ta được kết quả: **C** = {BCH×14, CF×47, C×147, DH×16, H×146} là tập các tập Sporadic tuyệt đối hai ngưỡng đóng của ngữ cảnh phát hiện dữ liệu $\hat{D}$. Hình 2.2 minh họa việc tìm kiếm các tập Sporadic tuyệt đối hai ngưỡng đóng.

2.1.3.4. 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 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++.

a. Thử nghiệm trên CSDL giả định

Mục đích của việc thử nghiệm này là kiểm tra hiệu quả của thuật toán MCPSI với dữ liệu lớn và có đặc điểm khác nhau. Dữ liệu giả định được thiết lập dựa trên nguyên tắc sinh dữ liệu do R. Agrawal và R. Srikant đề xuất trong [13, 16].

Tệp dữ liệu giả định mô phỏng hoạt động giao dịch mua hàng với các tham số được xác định trước. Để sinh dữ liệu thử nghiệm, chúng tôi đã sử dụng các tham số: |**O**| là số lượng giao dịch, |T| là độ dài trung bình của các giao dịch, |L| là số lượng các tập phổ biến, và |**I**| là số lượng các mục dữ liệu. Bước đầu, sinh ra kích thước của một giao dịch theo phân bố xác suất Poisson với kỳ vọng là kích thước trung bình của giao dịch. Mỗi giao dịch sẽ được điền đầy bởi các mục dữ liệu bằng cách trong mỗi giao dịch xác định một chuỗi các tập phổ biến tiềm năng. Mô tả đầy đủ của thuật toán sinh dữ liệu có thể tìm được trong [13, 16]. Thông tin về các tập dữ liệu giả định được mô tả trong bảng 2.1.

Bảng 2.1: Thông tin về các CSDL giả định

TT	Tên CSDL	Số mục dữ liệu	Số giao dịch	Độ dài trung bình của một giao dịch
1	T05I1000D10K	1 000	10 000	5
2	T10I1000D10K	1 000	10 000	10
3	T15I1000D10K	1 000	10 000	15
4	T20I1000D10K	1 000	10 000	20
5	T25I1000D10K	1 000	10 000	25
6	T30I1000D10K	1 000	10 000	30

Để so sánh hiệu quả thực hiện thuật toán MCPSI với thuật toán Apriori-Inverse, chúng tôi đã xây dựng chương trình theo hai thuật toán này. Bảng 2.2 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.

Do tính chất của các tập dữ liệu giả định là rất thưa nên trong quá trình thử nghiệm chúng tôi đã lựa chọn hai ngưỡng độ hỗ trợ là nhỏ, cụ thể minSup = 0,0005 và maxSup = 0,01.

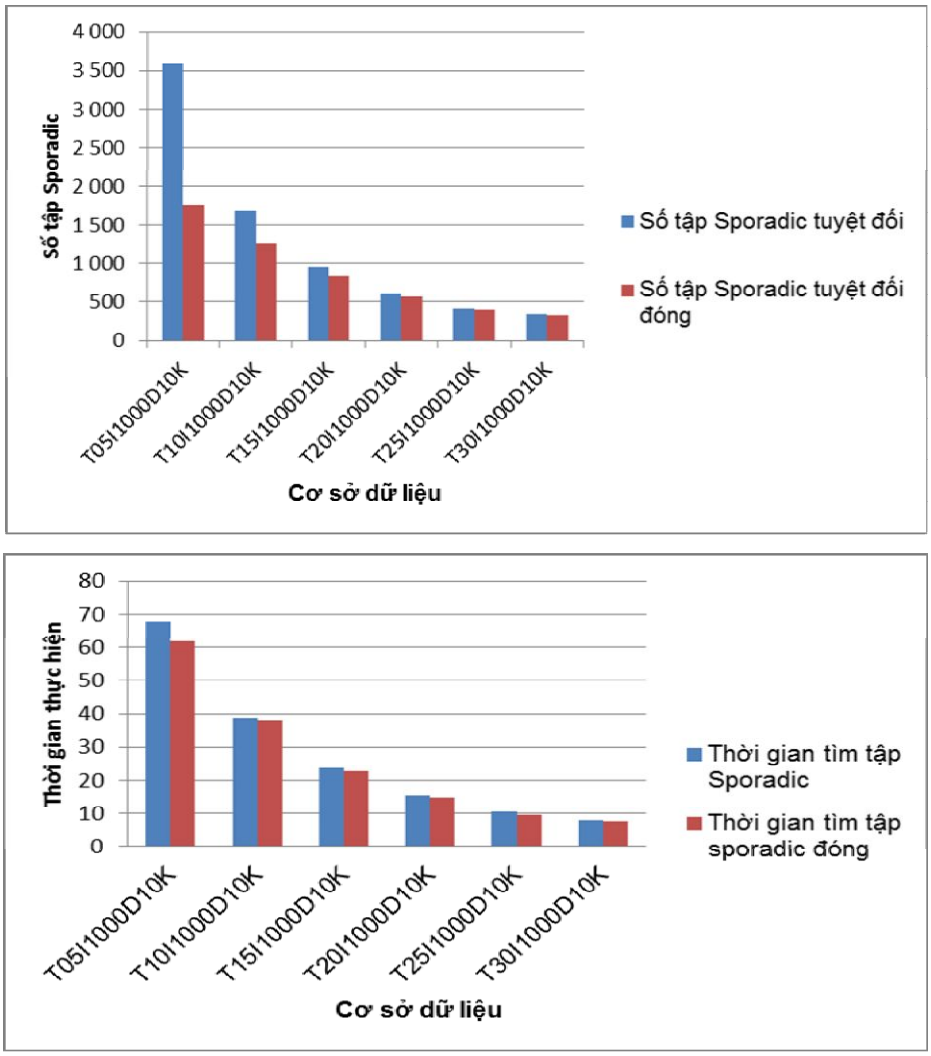
Bảng 2.2: Kết quả thực hiện MCPSI và Apriori-Inverse trên CSDL giả định

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 2.2 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.

Để thấy rõ mức độ giảm của tập Sporadic tuyệt đối đóng so với tập Sporadic tuyệt đối trên cùng CSDL giả định, có thể quan sát trên hình 2.3.

Xét cụ thể hơn với tệp dữ liệu T05I1000D10K và cho giá trị cận dưới thay đổi (cũng được áp dụng đối với thuật toán Apriori-Inverse), nhận được kết quả như trong bảng 2.3.



Hình 2.3: Biểu đồ so sánh kết quả thực hiện MCPSI và Apriori-Inverse trên các CSDL giả định

Bảng 2.3: Kết quả thực hiện MCPSI và Apriori-Inverse trên T5I1000D10K

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)
0,0005	0,01	3 588	67,695	1 767	62,015
0,001	0,01	757	50,388	714	47,899
0,005	0,01	224	6,865	224	6,585
0,001	0,1	1 702	78,553	1 438	75,256
0,005	0,1	374	20,492	374	19,787
0,01	0,1	152	4,435	152	4,321

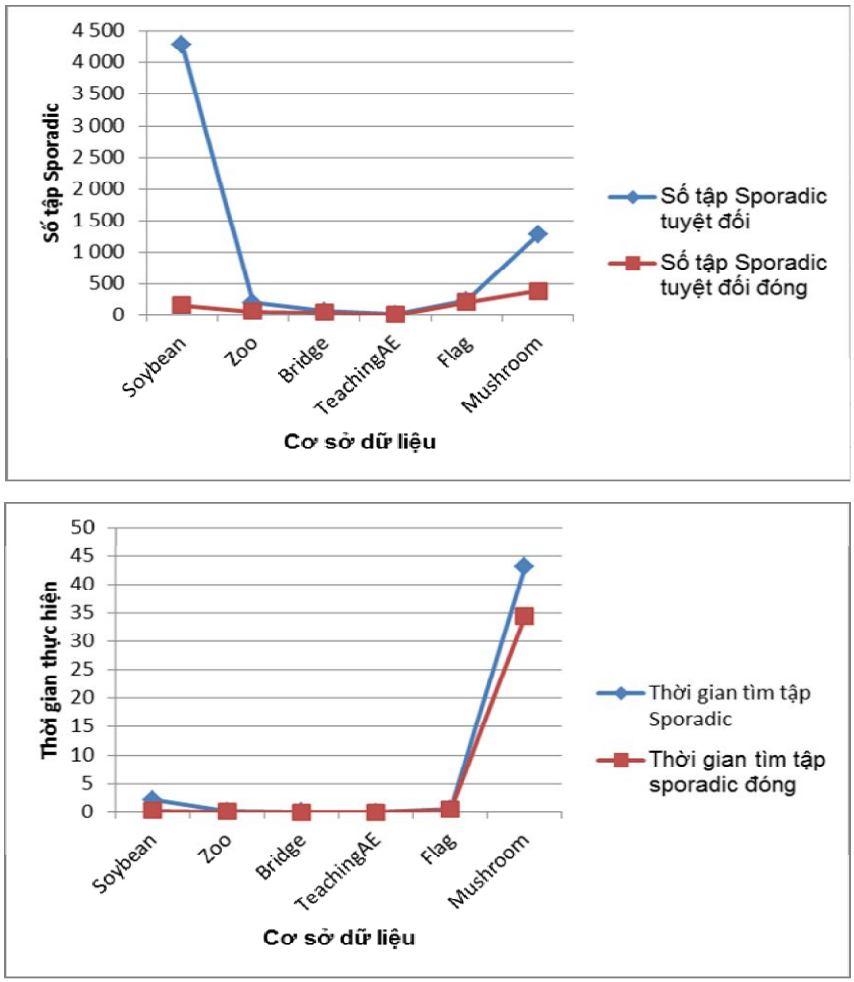
Khi thử nghiệm hai thuật toán nhiều lần với ngưỡng minSup và maxSup khác nhau trên CSDL T5I1000D10K đã xuất hiện những trường hợp xấu nhất là số tập Sporadic tuyệt đối bằng với tập Sporadic tuyệt đối hai ngưỡng đóng (khi minSup = 0,005; maxSup = 0,01 và minSup = 0,01; maxSup = 0,1), nhưng ngay với những trường hợp đó, thời gian thực hiện của thuật toán MCPSI vẫn ít hơn so với thời gian thực hiện của thuật toán Apriori-Inverse.

b. 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 bảng 2.4 và hình 2.4. Trên các CSDL thực dữ liệu không thưa như trên các CSDL giả định nên chúng tôi chọn hai ngưỡng có giá trị lớn hơn, cụ thể cùng chọn minSup = 0,1 và maxSup = 0,5 cho tất cả các trường hợp. Kết quả thực hiện hai thuật toán trong bảng 2.4 cho thấy số lượng tập Sproradic tuyệt đối hai ngưỡng đóng tìm được cũng ít hơn so với tập Sporadic tuyệt đối. Như vậy, kết quả thử nghiệm trên các CSDL thực là tương tự như kết quả thử nghiệm trên các CSDL giả định.

Bảng 2.4: Kết quả thực hiện MCPSI và Apriori-Inverse trên CSDL thực

Tên CSDL	Số mục dữ liệu (I)	Số tác vụ (D)	Min Sup	Max Sup	Thuật toán Apriori-Inverse		Thuật toán MCPSI	
					Số tập Sporadic tuyệt đối	Thời gian (sec)	Số tập Sporadic tuyệt đối đóng	Thời gian (sec)
Soybean	76	47	0,1	0,5	4 275	2,246	154	0,312
Zoo	43	101	0,1	0,5	203	0,187	56	0,094
Bridge	220	108	0,1	0,5	63	0,074	42	0,014
TeachingAE	104	151	0,1	0,5	13	0,031	12	0,015
Flag	310	194	0,1	0,5	235	0,515	210	0,443
Mushroom	118	8 124	0,1	0,5	1 273	43,028	387	34,336



Hình 2.4: Đồ thị so sánh kết quả thực hiện MCPSI và Apriori-Inverse 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

Vấn đề phát hiện luật kết hợp Sporadic không tuyệt đối cho đến nay vẫn chưa được giải quyết triệt để. Trong [50] các tác giả đã phân chia luật kết hợp Sporadic không tuyệt đối thành 4 dạng và đã đề xuất thuật toán MIISR để chỉ tìm các luật Sporadic không tuyệt đối ở dạng thứ 3 trong phân loại này.

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→B sao cho:

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

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.

Khi đó vấn đề phát hiện luật Sporadic không tuyệt đối trong [50] thực chất là một trường hợp riêng của việc phát hiện luật Sporadic không tuyệt đối hai ngưỡng, ở đó minSup có giá trị bằng $\frac{1}{|O|}$, với $|O|$ là tổng số các tác vụ của tệp dữ liệu.

Khác với nghiên cứu của các tác giả trong [50], chúng tôi áp dụng hướng tiếp cận phát hiện tập mục dữ liệu đóng trong việc tìm các tập Sporadic không tuyệt đối hai ngưỡng vì thế sẽ cho phép thu hẹp không gian tìm kiếm và loại bỏ được nhiều luật dư thừa. Thuật toán tìm các tập Sporadic không tuyệt đối hai ngưỡng cũng được phát triển dựa trên tư tưởng của thuật toán CHARM [94].

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:

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

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

Tập Sporadic không tuyệt đối hai ngưỡng X được gọi là tập Sporadic không tuyệt đối hai ngưỡng cực đại nếu nó không là tập con thực sự của bất kỳ tập Sporadic không tuyệt đối hai ngưỡng nào.

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.

Tập Sporadic không tuyệt đối hai ngưỡng đóng X được gọi là tập Sporadic không tuyệt đối hai ngưỡng đóng cực đại nếu nó không phải là tập con thực sự của bất kỳ tập Sporadic không tuyệt đối hai ngưỡng đóng nào.

Nhận xét 2.2: Theo định nghĩa 2.5, tập Sporadic không tuyệt đối hai ngưỡng là tập không phổ biến theo ngưỡng maxSup nhưng là tập phổ biến theo ngưỡng minSup.

Tính chất Apriori của các tập Sporadic không tuyệt đối hai ngưỡng là không được bảo toàn, tức là tập con của tập Sporadic không tuyệt đối hai ngưỡng chưa chắc là tập có tính chất như vậy.

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

Việc chứng minh tính chất này là tương tự như chứng minh tính chất 2.3.

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.*

Chứng minh: Ta chỉ cần chứng minh rằng mọi tập Sporadic không tuyệt đối hai ngưỡng cực đại cũng là tập đóng.

Giả sử X là tập Sporadic cực đại hai ngưỡng nào đó, trước hết ta chứng minh rằng X là tập phổ biến cực đại theo minSup.

Thật vậy X là tập phổ biến theo minSup là hiển nhiên theo định nghĩa 2.5. Giả sử ngược lại X không phải là cực đại theo minSup thì tồn tại X' là tập phổ biến theo minSup và $X \subset X'$. Theo tính chất Apriori thì $sup(X') \leq sup(X) < maxSup$. Mặt khác vì X là tập Sporadic không tuyệt đối hai ngưỡng nên tồn tại $x \in X \subset X'$ sao cho $sup(x) \geq maxSup$. Từ đó suy ra X' là tập Sporadic không tuyệt đối hai ngưỡng chứa X. Điều này mâu thuẫn với giả thiết X là tập Sporadic không tuyệt đối hai ngưỡng cực đại.

Mặt khác, theo tính chất của phép kết nối Galois luôn có $X \subseteq h(X)$ và do $sup(h(X)) = sup(X) \geq minSup$ nên h(X) cũng là tập phổ biến theo minSup nên khi X là tập phổ biến cực đại theo minSup thì h(X)=X hay X là tập Sporadic không tuyệt đối hai ngưỡng đóng cực đại■

Nhận xét 2.3: Giả sử X là tập Sporadic không tuyệt đối hai ngưỡng, X là tập phổ biến cực đại theo độ hỗ trợ cực tiểu minSup thì X cũng là tập Sporadic không tuyệt đối hai ngưỡng cực đại.

Việc chứng minh nhận xét này được suy trực tiếp từ cách chứng minh tính chất 2.6 ở trên.

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.*

Chứng minh: Ta chỉ cần chứng minh mọi luật Sporadic không tuyệt đối hai ngưỡng đều được sinh ra từ các tập Sporadic không tuyệt đối hai ngưỡng cực đại.

Giả sử $A \rightarrow B$ là luật như vậy, nên $A \cup B$ là tập Sporadic không tuyệt đối hai ngưỡng và $A \rightarrow B$ là luật kết hợp theo độ hỗ trợ cực tiểu minSup và độ tin cậy cực tiểu minConf. Từ [64] suy ra rằng $A \rightarrow B$ cũng được sinh ra từ tập phổ biến cực đại với độ hỗ trợ cực tiểu là minSup.

Không giảm tính tổng quát ta có thể coi rằng $A \cup B$ là tập phổ biến cực đại theo độ hỗ trợ cực tiểu minSup và ta sẽ chứng minh $A \cup B$ là tập Sporadic không tuyệt đối hai ngưỡng cực đại.

Giả sử ngược lại $\exists C$: $C \supset A \cup B$ sao cho $minSup \leq sup(C) < sup(A \cup B) < maxSup$, như vậy có nghĩa C là tập phổ biến cực đại theo độ hỗ trợ cực tiểu minSup thực sự chứa $A \cup B$. Điều này mâu thuẫn với giả thiết về $A \cup B$■

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

2.2.3.1. Ý tưởng của thuật toán

Thuật toán MCISI tìm các tập Sporadic không tuyệt đối hai ngưỡng đóng cực đại bằng cách:

- Xuất phát từ tập các mục dữ liệu ban đầu của tập dữ liệu, thiết lập hai tập: (1) Tập các mục dữ liệu phổ biến theo maxSup. (2) Tập các mục dữ liệu không phổ

biến theo maxSup nhưng phổ biến theo minSup. Các phần tử trong hai tập này được sắp xếp theo một trình tự nhất định (chẳng hạn, xếp theo thứ tự từ vựng).

- Tiếp theo sẽ kết hợp từng mục dữ liệu trong tập (1) với các mục dữ liệu còn lại bên phải của tập này và với tất cả các mục dữ liệu khác trong (2) để khởi tạo không gian tìm kiếm các tập Sporadic không tuyệt đối hai ngưỡng đóng. Kết quả của việc kết hợp này sẽ là tập các tập hai mục dữ liệu có chứa ít nhất một mục dữ liệu phổ biến theo maxSup. Độ hỗ trợ của các tập hai mục dữ liệu này là nhỏ hơn độ hỗ trợ maxSup nhưng không nhỏ hơn độ hỗ trợ minSup. Nói cách khác, các tập mục dữ liệu này không phổ biến theo độ hỗ trợ cận trên maxSup nhưng phổ biến theo độ hỗ trợ cận dưới minSup.

- Thực hiện tìm các tập Sporadic không tuyệt đối hai ngưỡng đóng trên không gian khởi tạo theo tinh thần thuật toán CHARM [94]. Thuật toán sẽ duyệt và tìm kiếm theo chiều sâu của không gian tìm kiếm các tập Sporadic không tuyệt đối hai ngưỡng đóng. Tất cả các tập không phải là tập Sporadic không tuyệt đối hai ngưỡng cũng như các nhánh không phải là tập đóng đều bị tỉa. Chiến lược tỉa cũng trên cơ sở dựa vào bốn tính chất của các cặp tập mục dữ liệu và tập định danh.

2.2.3.2. Thuật toán MCISI

Độ phức tạp của thuật toán MCISI: Thuật toán được xây dựng dựa trên thuật toán CHARM, với quá trình khởi tạo tập Nodes ban đầu thỏa mãn không chỉ là các tập phổ biến mà còn có các tập hiếm theo maxSup. Tuy nhiên, số phần tử của tập Nodes ban đầu không ảnh hưởng đến độ phức tạp của thuật toán.

Tại câu lệnh 3, một vòng lặp được thực hiện với kích thước của các phần tử phổ biến FI. Mỗi lần lặp tương ứng với một lần thực hiện thuật toán CHARM, do đó, độ phức tạp tương đương với thuật toán CHARM là O(l.|C|), với l là độ dài trung bình của các tidset và **C** là tập Sporadic không tuyệt đối hai ngưỡng đóng. Gọi $|\bar{C}|$ là kích thước trung bình của các tập Sporadic không tuyệt đối hai ngưỡng đóng tìm được, và giả thiết độ dài trung bình của các tidset là như nhau, thuật toán MCISI sẽ có độ phức tạp là O(|FI|.l.$|\bar{C}|$).

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 = $\{I_j \times g(I_j) : I_j \in \mathbf{I} \ \Lambda \mid g(I_j) \mid \geq \text{maxSup}\}$
2. IFI = $\{K_j \times g(K_j) : K_j \in \mathbf{I} \ \Lambda \mid g(K_j) \mid < \text{maxSup} \ \Lambda \mid g(K_j) \mid \geq \text{minSup}\}$
3. for each $I_j \times g(I_j)$ in FI do begin
4. Nodes=$\{P_j \times g(P_j)$: $P_j = I_j \cup M_j$, $g(P_j) = g(I_j) \cap g(M_j)$, $M_j \in FI \setminus \{I_1,...,I_j\}$ or $M_j \in$ IFI $\Lambda \mid g(P_j) \mid \geq$ 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** $\cup$ **C**
7. end

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

Hình 2.5: Thuật toán MCISI

Ở đây g là một phép kết nối Galois. Hàm CHARM-PROPERTY được xây dựng như trong [94].

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

Tính đúng đắn

Trước hết sẽ chứng minh rằng những tập tìm được bởi MCISI là tập Sporadic không tuyệt đối hai ngưỡng đóng cực đại.

Thật vậy, thuật toán MCISI gồm 3 giai đoạn chính.

Giai đoạn thứ nhất (dòng lệnh 1, 2) khởi tạo tập FI gồm các mục dữ liệu phổ biến theo độ hỗ trợ maxSup và tập IFI gồm các mục dữ liệu không phổ biến theo độ hỗ trợ maxSup nhưng phổ biến theo minSup. Các mục dữ liệu trong hai tập này được sắp thứ tự.

Giai đoạn thứ 2, các dòng lệnh 3, 4, 5 sẽ thực hiện kết hợp từng mục dữ liệu trong FI với các mục dữ liệu còn lại bên phải mục dữ liệu đang xét trong FI và với tất cả các mục dữ liệu khác trong IFI để tạo không gian tìm kiếm Nodes. Tiếp theo sẽ thực hiện hàm MCISI-EXTEND(Nodes,**C**) trên không gian mới khởi tạo. Hàm này sẽ đi tìm các tập phổ biến đóng theo minSup trên không gian Nodes theo đúng tinh thần của thuật toán CHARM trong [94]. Tập phổ biến đóng theo minSup cuối cùng theo mỗi nhánh của cây không gian tìm kiếm cũng là tập phổ biến đóng cực đại. Tiếp theo đó dòng lệnh 15 sẽ kiểm tra điều kiện để loại đi các tập có độ hỗ trợ lớn hơn hoặc bằng maxSup. Như vậy, kết thúc hàm MCISI-EXTEND cho kết quả tập **C** là tập các tập phổ biến đóng theo minSup, nhưng không phổ biến theo maxSup và chứa ít nhất một mục dữ liệu phổ biến theo maxSup. Theo định nghĩa 2.6 tập này sẽ là tập Sporadic không tuyệt đối hai ngưỡng đóng.

Giai đoạn thứ 3: dòng lệnh 7 sẽ kết hợp tất cả các tập tìm được từ các không gian khác nhau khởi tạo từ các tập mục dữ liệu trong FI. Tập này chính là tập các tập Sporadic không tuyệt đối hai ngưỡng đóng.

Tính đầy đủ

Cần chỉ ra rằng mọi luật Sporadic không tuyệt đối hai ngưỡng đều được sinh ra bởi một trong các tập Sporadic được tìm bởi thuật toán MCISI.

Thật vậy: theo tính chất 2.7 mọi luật Sporadic không tuyệt đối hai ngưỡng đều được sinh ra bởi tập Sporadic không tuyệt đối hai ngưỡng cực đại và theo tính chất 2.6 tập này cũng là tập Sporadic không tuyệt đối hai ngưỡng đóng cực đại■

Ví dụ 2.2: Xét CSDL **D** được xác định như trong ví dụ 0.1.

Giả thiết xét với minSup = 0,25 và maxSup = 0,5.

Ban đầu ta có hai tập:

FI = {A×123467, E×24568, J×13478} là tập các mục dữ liệu phổ biến theo độ hỗ trợ cận trên maxSup. (dòng 1)

IFI = {B×14, C×147, D×16, F×47, H×146} là tập các mục dữ liệu không phổ biến theo độ hỗ trợ cận trên maxSup, nhưng phổ biến theo độ hỗ trợ cận dưới minSup. (dòng 2)

Dòng 3, xét với mục dữ liệu đầu tiên A×123467 của tập FI.

Dòng 4 có Nodes = {AB×14, AC×147, AD×16, AE×246, AF×47, AH×146, AJ×1347}, các mục dữ liệu được sắp xếp theo thứ tự tăng dần của từ vựng.

Dòng 5 sẽ thực hiện hàm MCISI-EXTEND(Nodes,**C**) trên Nodes được thiết lập ở dòng 4 như sau:

Ban đầu xét AB×14, X = AB (dòng 8). Tiếp theo nút này sẽ được kết hợp với các nút lân cận ở bên phải (dòng 9). Khi kết hợp với AC vì g(AB) ⊂ g(AC) nên thay AB bằng ABC (X = ABC). Khi kết hợp với AD được ABCD nhưng tập này có độ hỗ trợ nhỏ hơn minSup nên bị loại. Kết hợp với AE được ABCE có độ hỗ trợ nhỏ hơn minSup nên bị loại. Kết hợp với AF được ABCF cũng bị loại do có độ hỗ trợ nhỏ hơn minSup. Khi kết hợp với AH vì g(ABC) ⊂ g(AH) nên thay ABC bằng ABCH (X = ABCH). Khi kết hợp với AJ vì g(ABCH) ⊂ g(AJ) nên thay ABCH bằng ABCHJ (X = ABCHJ). Kiểm tra sup(ABCHJ) = 0,25 có độ hỗ trợ nhỏ hơn maxSup nên bổ sung ABCHJ×14 vào **C**. Tập mục dữ liệu ABCHJ thỏa mãn điều kiện minSup ≤ sup(ABCHJ) < maxSup, h(ABCHJ) = f(g(ABCHJ)) =f(14)= ABCHJ và có chứa A, J là mục dữ liệu phổ biến theo maxSup. Vậy ABCHJ là tập Sporadic không tuyệt đối hai ngưỡng đóng theo định nghĩa 2.6.

Tương tự tiến hành với các nút còn lại trên Nodes, cuối cùng sẽ có kết quả **C** = {ABCHJ×14, ACFJ×47, ACJ×147, ADH×16, AEH×46, AE×246, AH×146}. Vậy **SC** = {ABCHJ×14, ACFJ×47, ACJ×147, ADH×16, AEH×46, AE×246, AH×146} (dòng 6).

Tiếp tục thực hiện như trên với các mục dữ liệu còn lại của FI để tạo các Nodes khác. Cuối cùng kết hợp các tập tìm được trên các Nodes sẽ có **SC** là tập các tập Sporadic không tuyệt đối hai ngưỡng đóng của CSDL **D**.

2.2.3.4. Kết quả thử nghiệm

Để đánh giá hiệu quả thực hiện của thuật toán MCISI, chúng tôi tiến hành thử nghiệm trên các CSDL giả định và một số CSDL trong [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 MCISI được lập trình trên ngôn ngữ C++.

a. Thử nghiệm trên CSDL giả định

Thông tin về các CSDL giả định được mô tả trong bảng 2.1. Kết quả thử nghiệm thuật toán MCISI trên CSDL T5I1000D10K khi chọn cố định độ hỗ trợ cận dưới minSup = 0,001 và maxSup thay đổi được thể hiện ở bảng 2.5. Do dữ liệu trong các CSDL giả định là rất thưa nên cần chọn hai hệ số độ hỗ trợ nhỏ. Khi độ hỗ trợ cận trên maxSup thay đổi giảm dần từ 0,1 đến 0,01 thì số tập Sporadic không tuyệt đối hai ngưỡng đóng tìm được đã tăng từ 0 lên 599 tập mục dữ liệu. Thời gian thực hiện cũng tăng lên do số tập Sporadic không tuyệt đối hai ngưỡng tìm được tăng.

Bảng 2.5: Bảng kết quả thử nghiệm trên CSDL T5I1000D10K

TT	Tên CSDL	minSup	maxSup	Số tập Sporadic	Thời gian (giây)
1	T5I1000D10K	0,001	0,1	0	0,215
2	T5I1000D10K	0,001	0,04	0	0,207
3	T5I1000D10K	0,001	0,02	242	2,542
4	T5I1000D10K	0,001	0,01	599	17,516

Bảng 2.6 là kết quả thử nghiệm thuật toán MCISI trên sáu 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 2.6 cho thấy thuật toán đã thực hiện được trên các tệp dữ liệu lớn với thời gian thực hiện nhỏ.

Bảng 2.6: 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

b. Thử nghiệm trên CSDL thực

Dữ liệu thử nghiệm thuật toán là năm CSDL lấy từ nguồn [100]. CSDL ban đầu được chuyển sang dạng tác vụ. Chọn độ hỗ trợ cận dưới minSup = 0,1 và độ hỗ trợ cận trên maxSup = 0,5. Thông tin về các CSDL và kết quả thực hiện thuật toán MCISI mô tả trong bảng 2.7.

Bảng 2.7: 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

Khi minSup $= \frac{1}{|O|}$, với $|O|$ là tổng số các tác vụ trong CSDL thì thuật toán MCISI sẽ tìm các tập Sporadic không tuyệt đối đóng cho các luật Sporadic không tuyệt đối trong [50]. Thực hiện thuật toán MCISI trên các CSDL với minSup được lựa chọn phù hợp đối với mỗi CSDL nhận được kết quả là bảng 2.8 về các tập Sporadic không tuyệt đối đóng.

Bảng 2.8: Kết quả tìm các tập Sporadic không tuyệt đối trên CSDL thực

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	1/47	0,5	8 853	15,273
2	Zoo	43	101	1/101	0,5	5 253	9,126
3	Bridge	220	108	1/108	0,5	1 253	2,605
4	Teaching AE	104	151	1/151	0,5	7	0,34

Trong các CSDL thực thử nghiệm thì CSDL Mushroom có nhiều tác vụ nhất nên chúng tôi đã tiến thành thử nghiệm riêng trên CSDL này. Thực hiện thuật toán MCISI trên tệp dữ liệu Mushroom với minSup = 0,1, maxSup thay đổi từ 0,2 đến 0,5 nhận được kết quả trong bảng 2.9.

Bảng 2.9: Kết quả thử nghiệm trên tệp dữ liệu Mushroom với minSup = 0,1

minSup	maxSup	Số tập Sporadic không tuyệt đối hai ngưỡng đóng	Thời gian (giây)
0,1	0,5	6365	279
0,1	0,4	6174	220
0,1	0,3	5717	181
0,1	0,2	4773	163

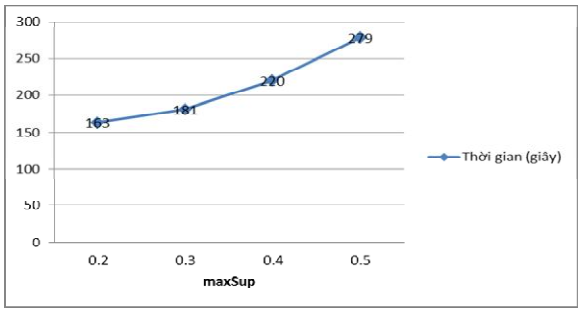
Thực hiện thuật toán MCISI trên tệp dữ liệu Mushroom với maxSup = 0,5, minSup thay đổi từ 0,1 đến 0,4 nhận được kết quả trong bảng 2.10.

Bảng 2.10: Kết quả thử nghiệm trên tệp dữ liệu Mushroom với maxSup = 0,5

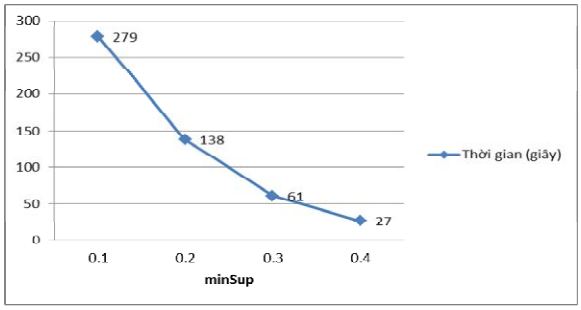
minSup	maxSup	Số tập Sporadic không tuyệt đối hai ngưỡng đóng	Thời gian (giây)
0,1	0,5	6365	279
0,2	0,5	1367	138
0,3	0,5	440	61
0,4	0,5	106	27

Để thấy rõ hơn mối quan hệ về sự tương quan giữa giá trị minSup, maxSup, số tập Sporadic không tuyệt đối hai ngưỡng đóng tìm được và thời gian thực hiện thuật

toán, số liệu trên bảng 2.9 và bảng 2.10 được chuyển sang dạng đồ thị như hình 2.6 và hình 2.7.



Hình 2.6: Kết quả thử nghiệm trên tệp dữ liệu Mushroom với minSup = 0,1



Hình 2.7: Kết quả thử nghiệm trên tệp dữ liệu Mushroom với maxSup = 0,5

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 **I** = $\{i_1, i_2,\ldots, i_j, ..., i_n\}$ 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 $\subseteq$ **I**.

Luật kết hợp mẫu âm đã được quan tâm trong một số công trình nghiên cứu [28, 68, 86, 89] và nó có dạng tổng quát là: $A_1 \cup \overline{A}_2 \rightarrow B_1 \cup \overline{B}_2$, ở đây $A_1, B_1 \subseteq$ **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).

Thực tế, bài toán tìm các tập phổ biến có mục dữ liệu âm từ CSDL tác vụ thông qua các tập phổ biến chỉ có các mục dữ liệu dương đã được một số tác giả

quan tâm nghiên cứu [17, 31, 52]. Giải pháp hiện được xem là thành công nhất về vấn đề này được giới thiệu trong [52]. Tác giả bài báo này đã đề xuất biểu diễn các tập phổ biến có mục dữ liệu âm thành 3 thành phần chỉ gồm các mục dữ liệu dương, từ đó giúp tính được độ hỗ trợ của các tập có mục dữ liệu âm và tìm tập phổ biến có mục dữ liệu âm bằng cách dựa vào cải tiến phát triển thuật toán Apriori. Tuy nhiên thuật toán tìm các tập phổ biến có mục dữ liệu âm theo cách tiếp cận này còn khá phức tạp, chưa hiệu quả và cần được nghiên cứu phát triển và hoàn thiện tiếp. Việc nghiên cứu đề xuất thuật toán phát hiện các luật như vậy thực tế đang được nhiều nhà nghiên cứu quan tâm.

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.

Thực tế cho thấy rằng giữa các mục dữ liệu tồn tại nhiều kiểu ràng buộc khác nhau. Chẳng hạn có thể xẩy ra trường hợp có một số nhóm mục dữ liệu không bao giờ xuất hiện đồng thời trong cùng một tác vụ, nói cách khác nếu một nhóm mục dữ liệu đã xuất hiện trong một tác vụ nào đó thì có thể có nhóm mục dữ liệu khác không thể xuất hiện trong tác vụ này. Ví dụ thực tiễn công tác điều hành các hoạt động thương mại cho thấy trong rất nhiều trường hợp nhà nước cho phép nhập khẩu nhóm mặt hàng này, thì đồng thời phải cấm nhập khẩu nhóm mặt hàng khác; hoặc khi xây dựng các dòng thuế cho các nhóm ngành hàng, vẫn thường xẩy ra trường hợp việc cho phép tăng, giảm thuế một số mặt hàng trong nhóm phải được gắn liền với việc không cho phép tăng, giảm thuế của một số mặt hàng khác; đặc biệt trong y học thì những tình huống như vậy là khá phổ biến, chẳng hạn khi người bệnh có một số triệu chứng biểu hiện của một căn bệnh nào đó thì chắc chắn người này không thể có một số triệu chứng biểu hiện cho một số căn bệnh khác,… Từ thực tiễn này, vấn đề tìm tập phổ biến và các luật kết hợp có ràng buộc mục dữ liệu âm đã được nẩy sinh. Luật kết hợp với ràng buộc mục dữ liệu âm không chỉ là một dạng của luật kết hợp mẫu âm mà nó còn là một dạng luật kết hợp được tìm theo cách tiếp cận phát hiện luật kết hợp hiếm.

Một cách hình thức, phần dưới đây sẽ nghiên cứu giải quyết bài toán sau:

Phát hiện các luật kết hợp $A \rightarrow B$ với:

$conf(A \rightarrow B) \geq minConf$, $sup(A \cup B) \geq minSup$ *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 $\subseteq$ **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 = $\varnothing$, là không thể xuất hiện trong các tác vụ này.

Nhận xét 2.4: Dễ dàng thấy rằng nói chung không tồn tại mối quan hệ tập hợp giữa các cặp ràng buộc mục dữ liệu âm, cụ thể là giả sử (A_i, $\overline{B_i}$), i=1,2 là hai cặp ràng buộc mục dữ liệu âm, từ $A_1 \subseteq A_2$, không thể rút ra được quan hệ tập hợp giữa các tập $\overline{B_i}$ tương ứng và ngược lại.

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

Giả sử X là tập con bất kỳ của **I**, ký hiệu Y = {x $\in$**I** $\cup$ $\overline{I}$ / nếu x $\in$**I** 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.

Chứng minh: Giả sử tác vụ $t_i \in$ **O** hỗ trợ tập X, khi đó với mọi y $\in$ Y nếu y $\in$ X thì hiển nhiên t_i chứa y, nếu không phải như vậy thì tồn tại cặp ràng buộc mục dữ liệu âm (A_i, $\overline{B_i}$) sao cho y $\in \overline{B_i}$ và $A_i \subseteq$ X. Do t_i hỗ trợ A_i và theo định nghĩa của cặp ràng buộc mục dữ liệu âm, t_i hỗ trợ $A_i \cup \overline{B_i}$, từ đó suy ra t_i hỗ trợ y hay nói cách khác t_i hỗ trợ Y.

Ngược lại với mỗi $t_i \in$ **O** hỗ trợ tập Y, với mọi x $\in$ X, do x $\in$ Y nên t_i hỗ trợ x và vì vậy t_i hỗ trợ tập X ■

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.

Chứng minh: Ký hiệu $\overline{D} \subseteq$ (**I** $\cup$ $\overline{I}$) $\times$ **O** là tập dữ liệu có mục dữ liệu âm. $\overline{D}$ được xây dựng từ **D** như sau:

Duyệt theo các phần tử trong **O**, với mỗi t $\in$ **O**, giả sử t hỗ trợ tập mục dữ liệu A $\subseteq$ **I**, duyệt theo tất cả các phần tử trong $\mathfrak{I}$, nếu $\exists$(A_i, $\overline{B_i}$)$\in \mathfrak{I}$ sao cho $A_i \subseteq$ A thì ta bổ sung $\overline{B_i}$ vào A.

Theo mệnh đề 2.3, giả sử X là tập phổ biến tìm được từ CSDL **D** với tập ràng buộc $\mathfrak{I}$ thì Y được xác định như nêu trên sẽ là tập phổ biến đối với tập dữ liệu có mục dữ liệu âm $\overline{D}$.

Ngược lại chưa chắc đúng và sẽ được chứng minh trong ví dụ 2.4■

Ví dụ 2.3: Xét CSDL **D** được xác định như trong ví dụ 0.1. $\overline{I}$ = {-A, -B, -C, -D, -E, -F, -G, -H, -J} là tập các mục dữ liệu âm. Tập các ràng buộc mục dữ liệu âm $\mathfrak{I}$ = {(AE,-G), (EF,-D-G), (AC,-G), (DE,-J)}.

Theo cách xây dựng $\overline{D}$ trong mệnh đề 2.3, ta nhận được CSDL tác vụ có mục dữ liệu âm như trong bảng 2.11.

Bảng 2.11: Bảng dữ liệu với các mục dữ liệu âm của ví dụ 2.3

Tác vụ	Mục dữ liệu
t_1	A B C D H J -G
t_2	A E -G
t_3	A G J
t_4	A B C E F H J -D -G
t_5	E
t_6	A D E H -J
t_7	A C F J -G
t_8	E J

Ví dụ 2.4: Xét CSDL tác vụ có mục dữ liệu âm $\overline{D} \subseteq (\mathbf{I} \cup \overline{I}) \times \mathbf{O}$, ở đây **I** = {A, B, C} và $\overline{I}$ = {-A, -B, -C}, như bảng 2.12.

Bảng 2.12: Bảng dữ liệu minh họa cho ví dụ 2.4

Các tác vụ	**Các mục dữ liệu**
t_1	A B -C
t_2	A -B C
t_3	-A B C
t_4	A B C

Bắt đầu từ tác vụ t_1, ta thấy có thể xảy ra một trong 3 cặp ràng buộc mục dữ liệu âm sau: (A, -C); (B, -C) và (AB, -C). Cặp đầu không thể xảy ra vì ở tác vụ t_2, A và C đồng thời xuất hiện; tương tự các cặp (B, -C) và (AB, -C) cũng không được chấp nhận bởi các tác vụ t_3, t_4 một cách tương ứng.

Lập luận hoàn toàn tương tự cho các tác vụ còn lại. Nói cách khác trong trường hợp này không thể xây dựng được các cặp ràng buộc mục dữ liệu âm từ tập dữ liệu có mục dữ liệu âm.

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 có mục dữ liệu âm $\overline{D}$.

Chứng minh:

- Theo mệnh đề 2.3 nếu X là tập phổ biến trong tập dữ liệu **D** và thoả mãn tập ràng buộc mục dữ liệu âm $\mathfrak{I}$ thì Y cũng là tập phổ biến trong $\overline{D}$.

- Nếu X là đóng trong tập dữ liệu **D** theo các phép kết nối Galois f, g, h như được xác định trong phần 1.2.1.2 thì dễ dàng thấy rằng Y cũng là đóng theo các phép kết nối này trong tập dữ liệu $\overline{D}$.

- Nếu tập X còn là tập cực đại trong tập dữ liệu **D** thì tập Y cũng có tính chất đó. Thật vậy giả sử $Y \cup \{y\}$ với $y \notin Y$ là tập phổ biến, khi đó với y có 2 khả năng: nếu $y \in \mathbf{I}$ thì $y \notin X$ và $X \cup \{y\}$ là tập phổ biến, điều này là mâu thuẫn với tính chất phổ biến cực đại của X; nếu $y \in \overline{I}$ thì điều đó mâu thuẫn với cách xây dựng $\overline{D}$ đó là tất cả các mục dữ liệu âm đã được xác định bởi $\mathfrak{I}$ và được bổ sung tối đa vào các tác vụ■

Nhận xét 2.5*:*

Mệnh đề 2.5 cho biết để tìm các tập phổ biến từ CSDL tác vụ chỉ có các mục dữ liệu dương nào đó trong điều kiện có ràng buộc mục dữ liệu âm, ta có thể biểu diễn CSDL tác vụ này dưới dạng CSDL tác vụ có mục dữ liệu âm, và tập phổ biến tìm được sẽ là tập có một số mục dữ liệu âm và khi đó luật kết hợp được sinh từ các tập phổ biến này sẽ là luật có thể có mục dữ liệu âm ở một hoặc cả 2 phần tiền đề và hệ quả của luật kết hợp. Người ta gọi những luật kết hợp như vậy là luật kết hợp có mục dữ liệu âm hay luật kết hợp có mẫu âm [17, 31, 52].

Nếu tập các mục dữ liệu dương không quá lớn, thì việc tìm các tập phổ biến từ CSDL tác vụ có mục dữ liệu âm có thể được thực hiện theo các thuật toán tìm tập phổ biến thông dụng như Apriori [16],... bằng cách coi mỗi mục dữ liệu âm là một mục dữ liệu mới và khi đó số lượng các mục dữ liệu sau khi được bổ sung có thể lớn gấp 2 lần số lượng các mục dữ liệu ban đầu.

Khi số mục dữ liệu dương là khá lớn thì giải pháp này là không khả thi vì như đã biết độ phức tạp của thuật toán tìm các tập phổ biến là hàm mũ của số các mục dữ liệu và số các tác vụ trong CSDL [64, 94].

Các mệnh đề 2.3, 2.5 đã gợi ý rằng việc tìm các tập phổ biến đóng cực đại từ CSDL tác vụ với mục dữ liệu dương **D** và thoả mãn tập ràng buộc $\mathfrak{I}$ thực chất có thể qui được về việc tìm tập phổ biến đóng cực đại từ CSDL có mục dữ liệu âm $\overline{D}$. Và việc tìm các tập phổ biến đóng cực đại có mục dữ liệu âm từ $\overline{D}$ có thể được thực hiện bằng cách chỉ cần thông qua việc duyệt trên CSDL tác vụ với các mẫu dương **D** trên cơ sở dựa vào việc cải tiến và phát triển thuật toán CHARM.

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

2.3.3.1.Ý tưởng thuật toán

Thuật toán được tiến hành theo hai bước. Trước hết sử dụng thuật toán CHARM để tìm tập phổ biến đóng cực đại với các mục dữ liệu dương từ CSDL tác vụ **D**. Mỗi khi tìm được tập phổ biến đóng cực đại X trong tệp dữ liệu này thì khởi tạo và thực hiện bước thứ 2 bằng cách duyệt các cặp ràng buộc mục dữ liệu âm, nếu tập thứ nhất của cặp này nằm trong tập X thì bổ sung tập thứ hai của cặp ràng buộc vào một tập mà sau này sẽ trở thành tập phổ biến đóng cực đại có mục dữ liệu âm trong tập dữ liệu $\overline{D}$.

2.3.3.2. Thuật toán NC-CHARM

Thuật toán tìm các tập phổ biến đóng với ràng buộc mục dữ liệu âm được gọi là thuật toán NC-CHARM (Negative Constrains – CHARM). Giả ngôn ngữ của thuật toán được thể hiện trong hình 2.8.

Độ phức tạp của thuật toán NC-CHARM: So với thuật toán CHARM thì thuật toán này khác ở phần từ lệnh 10 đến lệnh 13. Đây là phần thực hiện tiếp theo phần thực hiện phép giao. Câu lệnh này được thực hiện bằng số phần tử trong tập ràng buộc mục dữ liệu âm (|$\mathfrak{I}$|) đối với mỗi tập đóng tìm được, do đó tổng chi phí thực hiện phép so sánh là (|$\mathfrak{I}$|.|**C**|). Kết hợp với độ phức tạp của thuật toán CHARM ta có độ phức tạp của NC-CHARM là O(l. |$\mathfrak{I}$|).|**C**|) với l là độ dài trung bình của các định danh.

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

NC-CHARM-EXTEND(Nodes, $\Im$, **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, $\Im$, **C**)
10. temp = X
11. for each $(A_i, \overline{B_i}) \in \Im$ do
12. if $A_i \subseteq X$ then $X = X \cup \overline{B_i}$
13. if X = temp then remove $X \times g(X)$ from Nodes
14. **C** = **C** $\cup$ X // if X is not subsumed
15. end

Hình 2.8: Thuật toán NC-CHARM

Ở đây g là một phép kết nối Galois, k là một phép sắp thứ tự (theo thứ tự từ vựng hoặc theo độ hỗ trợ) cho các mục dữ liệu. Hàm CHARM-PROPERTY được xây dựng như trong [94].

Tính đúng đắn của thuật toán

Thuật toán NC-CHARM được xây dựng dựa trên việc phát triển thuật toán CHARM. Bước thứ nhất của thuật toán NC-CHARM sử dụng những nội dung cơ bản nhất của thuật toán CHARM để tìm tập phổ biến đóng cực đại từ CSDL tác vụ các mục dữ liệu dương. Tính đúng đắn và hiệu quả của thuật toán này đã được minh chứng trong [94].

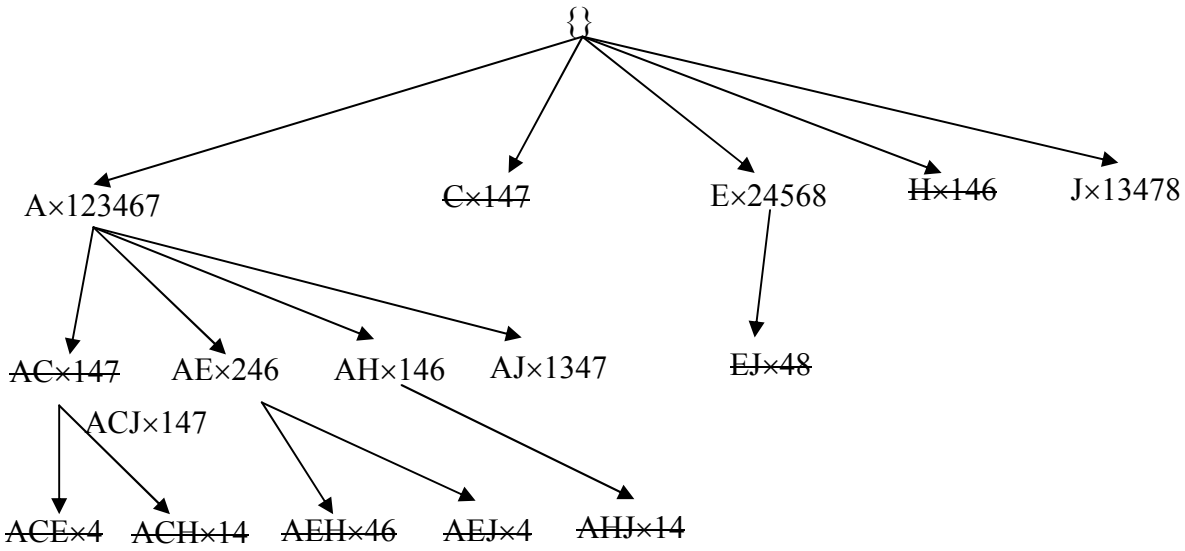
Sau khi tìm được tập phổ biến đóng cực đại X trong tập dữ liệu các mục dữ liệu dương **D**, thuật toán chuyển sang bước thứ hai. Bước này, được thể hiện từ lệnh 10 đến lệnh 13 trong thuật toán NC-CHARM, sẽ duyệt và kiểm tra các cặp ràng

buộc mục dữ liệu âm xem có những mục dữ liệu âm nào cần được bổ sung tiếp vào X để tập này vẫn còn là tập phổ biến đóng cực đại trong tập dữ liệu có mục dữ liệu âm $\overline{D}$ hay không. Câu lệnh if-then cuối cùng sẽ bổ sung hoặc loại bỏ tập X khỏi nút của cây biểu diễn không gian tìm kiếm [94] tuỳ thuộc vào việc có hay không mục dữ liệu âm được bổ sung vào X. Tập các nút của cây, biểu diễn không gian tìm kiếm của các mục dữ liệu, **C** chính là tập các tập phổ biến đóng cực đại trong CSDL tác vụ có mục dữ liệu âm $\overline{D}$.

Ví dụ 2.5: Xét CSDL **D** như trong ví dụ 2.3. Giả sử độ hỗ trợ cực tiểu minSup=3/8. Các nút của cây tìm kiếm được sắp theo thứ tự tăng dần của từ vựng.

Ban đầu khởi tạo tập Nodes = {A×123467, C×147, E×24568, H×146, J×13478} (dòng 1).

Thuật toán được bắt đầu ở nút A×123467. Gán X = A và kết hợp nút này với các nút lân cận phải của nó. Khi kết hợp A với C vì g(A) ⊃ g(C) nên loại bỏ C và NewN = {AC}. Khi kết hợp A với E được tập mục AE, NewN = {AC, AE}. Khi kết hợp A với H, vì g(A) ⊃ g(H), do vậy nhánh H sẽ bị loại bỏ, nút con AH sẽ thay thế cho H và NewN= {AC, AE, AH}. Kết hợp A với J được tập AJ và NewN= {AC, AE, AH, AJ}



Hình 2.9: Cây tìm kiếm tập phổ biến với ràng buộc mục dữ liệu âm

Do NewN ≠ ∅ nên thuật toán sẽ gọi NC-CHARM-EXTEND cho tập này. Đặt X = AC, sau đó kết hợp AC với AE được tập ACE không phổ biến sẽ loại bỏ. Kết hợp AC với AH được tập ACH không phổ biến sẽ loại bỏ. Kết hợp AC với AJ vì g(AC) ⊂ g(AJ) nên thay AC bằng ACJ và NewN = {ACJ}. Do NewN chỉ có một phần tử nên sẽ dừng lại. Tiếp theo gán temp = {ACJ} và duyệt các cặp ràng buộc âm và nhận thấy có tập ràng buộc (AC, -G) thoả mãn điều kiện có thành phần thứ nhất là con của tập ACJ vì vậy thành phần thứ hai sẽ được kết hợp vào tập {ACJ} thành tập mới là {ACJ-G}. Dòng lệnh tiếp theo kiểm tra thấy X = {ACJ-G} khác với temp nên bổ sung vào tập **C**. Tập ACJ có sup(ACJ) ≥ minSup, h(ACJ) = f(g(ACJ)) = f(146) = ACJ và thỏa mãn cặp ràng buộc mục dữ liệu âm (AC,-G). Như vậy tập (ACJ-G) là tập mục dữ liệu đóng thỏa mãn ràng buộc mục dữ liệu âm của CSDL **D** (hình 2.9).

Tiến hành tương tự với các nhánh B×12346, C×1356 và F×1256. Kết thúc, ta được kết quả là **C** = {ACJ-G, AE-G} là tập phổ biến đóng cực đại với ràng buộc mục dữ liệu âm.

2.3.3.4. Kết quả thử nghiệm

Để đánh giá hiệu quả thực hiện của thuật toán NC-CHARM, chúng tôi tiến hành thử nghiệm trên các CSDL giả định. 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 NC-CHARM được lập trình trên ngôn ngữ C++.

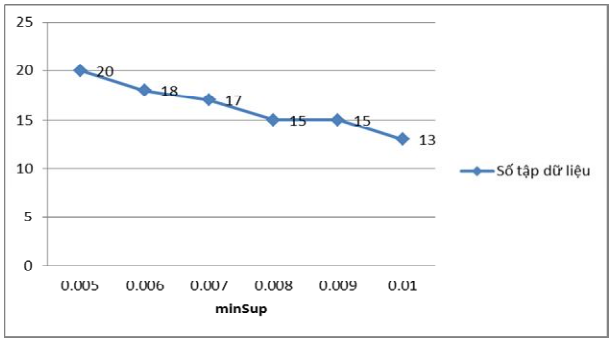
Bảng 2.13: Bảng kết quả thử nghiệm thuật toán NC-CHARM

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

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 2.13.

Thực hiện thử nghiệm thuật toán trên tệp dữ liệu T30I1000D10K (là CSDL có độ dài trung bình của các giao dịch là lớn nhất trong số các CSDL giả định), với ngưỡng minSup thay đổi từ 0,005 đến 0,01 có kết quả về số tập dữ liệu tìm được như trên hình 2.10.



Hình 2.10: Kết quả thử nghiệm NC-CHARM trên tệp dữ liệu T30I1000D10K

Do tập ràng buộc mục dữ liệu âm được sinh ngẫu nhiên trong phần thử nghiệm của chúng tôi và số lượng điều kiện ràng buộc âm là tương đối lớn (100 điều kiện) nên số tập phổ biến đóng thỏa mãn điều kiện là không nhiều. Khi ngưỡng minSup được chọn tăng dần từ 0,005 đến 0,01 thì số lượng tập phổ biến đóng thỏa mãn điều kiện ràng buộc mục dữ liệu âm trong tập kết quả là giảm dần. Kết quả này là hoàn toàn phù hợp với lý thuyết phát hiện luật kết hợp (hình 2.10).

Kết luận chương 2:

Trong chương 2, chúng tôi đã đề xuất mở rộng bài toán phát hiện luật kết hợp Sporadic tuyệt đối, không tuyệt đối hai ngưỡng và luật kết hợp với ràng buộc mục dữ liệu âm trên CSDL tác vụ. Ba thuật toán lần lượt được đề xuất là MCPSI, MCISI và NC-CHARM tương ứng nhằm tìm các tập phổ biến cho các luật kết hợp hiếm này. Khác với các nghiên cứu trước đây, cả ba thuật toán đều đi tìm tập phổ biến đóng cho các luật hiếm vì vậy đã tiết kiệm được chi phí và hạn chế được các luật dư thừa. Các thuật toán này được phát triển theo tư tưởng của thuật toán CHARM [94], 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. Phần thực nghiệm cũng đã chứng tỏ hiệu quả của các thuật toán do chúng tôi đề xuất.

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

Phát hiện luật kết hợp Sporadic trên CSDL tác vụ về cơ bản đã được giải quyết và được trình bày trong chương 2. Nội dung của chương 3 bàn về vấn đề phát hiện luật kết hợp hiếm trên CSDL định lượng do chúng tôi đề xuất đó là: luật kết hợp Sporadic tuyệt đối hai ngưỡng mờ và luật kết hợp Sporadic không tuyệt đối hai ngưỡng mờ. Kết quả nghiên cứu lần lượt đã được đăng trên Hội thảo quốc gia về Công nghệ thông tin và Truyền thông - lần thứ 13 [3] và Tạp chí Tin học và Điều khiển học [4].

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

Vấn đề phát hiện luật kết hợp mờ đã nhận được nhiều sự quan tâm của các nhà nghiên cứu [34, 38-41, 44, 45, 54, 61, 63, 82, 98]. Hiện có một số thuật toán nhằm phát hiện luật kết hợp phổ biến mờ. Tuy nhiên nếu áp dụng các thuật toán này cho việc phát hiện luật hiếm mờ cũng sẽ gặp những khó khăn tương tự như với trường hợp tìm các luật hiếm trên CSDL tác vụ. Chính vì vậy, chúng tôi đã nghiên cứu và đề xuất bài toán phát hiện luật kết hợp Sporadic hai ngưỡng mờ. Luật kết hợp Sporadic mờ cũng được chia thành hai loại giống như trên CSDL tác vụ là: luật kết hợp Sporadic tuyệt đối hai ngưỡng mờ và luật kết hợp Sporadic không tuyệt đối hai ngưỡng mờ. Hai thuật toán nhằm tìm các tập Sporadic phổ biến hai ngưỡng mờ tương ứng cho hai loại luật trên cũng đã được đề xuất.

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 nghiên cứu đề xuất giải pháp nhằm tìm các tập Sporadic tuyệt đối mờ cho các luật Sporadic tuyệt đối mờ bằng cách đề xuất bài toán phát hiện luật kết hợp mờ có dạng $r \equiv X \text{ is } A \rightarrow Y \text{ is } B$ sao cho:

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

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 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.

3.2.2. Tập Sporadic tuyệt đối hai ngưỡng mờ

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

minSup ≤ sup(<X,A>) < maxSup, và

∀x ∈ <X,A>, sup(x) < maxSup.

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

Chứng minh: Giả sử <X,A> là tập Sporadic tuyệt đối hai ngưỡng mờ nào đó và tập <X',A'> ⊆ <X,A>, ta cần chứng minh <X',A'> cũng là tập Sporadic tuyệt đối hai ngưỡng mờ.

Thật vậy do X' ⊆ X và A' ⊆ A nên:

$$\sum_{t_i \in O} \prod_{x_j \in X'} \left\{ \int_{\chi_{x_j}} (t_i[x_j] \right\} \geq \sum_{t_i \in O} \prod_{x_j \in X} \left\{ \int_{\chi_{x_j}} (t_i[x_j] \right\}$$

trong đó $\int_{\chi_{x_j}} (t_i[x_j])$ được xác định như trong công thức (1.4) (mục 1.3.2)

(i) Ta có: minSup ≤ sup(<X,A>) ≤ sup(<X',A'>)

(ii) Mặt khác với mọi x ∈ <X',A'> thì x ∈ <X,A> nên sup(x) < maxSup và vì vậy sup(<X',A'>) < maxSup

Từ (i) và (ii) suy ra <X',A'> là tập Sporadic tuyệt đối hai ngưỡng mờ■

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

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

3.2.3.1. 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 của tệp dữ liệu.

(b) Chuyển CSDL ban đầu thành CSDL mờ.

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

Cụ thể từng bước sẽ được thực hiện như sau:

a. Xây dựng tập mờ cho các thuộc tính

Để xây dựng tập mờ cho các thuộc tính phân loại và thuộc tính số có thể 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 lớp 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 giá trị của 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 tuyệt đối hai ngưỡng mờ. Trong giai đoạn này cần điền giá trị cho các thuộc tính mới bằng cách sử dụng hàm thành viên. Chúng tôi sử dụng phương pháp phân hoạch và

cách xây dựng hàm thành viên giới thiệu trong [41] và đã được tổng kết ở phần 1.3.3 của chương 1.

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

Thuật toán MFPSI (Mining Fuzzy Perfectly Sporadic Itemsets) được đề xuất nhằm tìm các tập Sporadic tuyệt đối hai ngưỡng mờ. Thuật toán MFPSI được phát triển theo tư tưởng của thuật toán Apriori [16].

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 $\omega_{\chi_{x_j}}$.
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

Ví dụ 3.1: CSDL được mô tả trong bảng 0.2 gồm các thuộc tính Tuổi, Số xe máy, Thu nhập, Có gia đình.

Xét thuộc tính Tuổi với rời rạc hóa: 1) Tuổi-trẻ (0,29), 2) Tuổi-trung niên [30,59), 3) Tuổi-già [60,110)

Khi đó $S_1 = 30$, $S_2 = 60$. Giả thiết p = 30%, các đường biên có giá trị là:

$d_1^+ = 30 - 0,5\ (30 - 0)\ 30\% = 30 - 4,5 = 25,5$

$S_1 = 30$

$d_2^- = 30 + 0,5\ (60 - 30)\ 30\% = 30 + 4,5 = 34,5$

$d_2^+ = 60 - 0,5\ (60 - 30)\ 30\% = 60 - 4,5 = 55,5$

$S_2 = 60$

$d_3^- = 60 + 0,5\ (110 - 60)\ 30\% = 60 + 7,5 = 67,5$

Xét thuộc tính Số xe máy với khái niệm mờ: 4) Số xe máy-ít (0,2], 5) Số xe máy-nhiều [3, 7)

Khi đó $S_1 = 3$. Giả thiết p = 30%, các đường biên có giá trị là:

$d_1^+ = 3 - 0,5\ (3 - 0)\ 30\% = 3 - 0,45 = 2,55$

$S_1 = 3$

$d_2^- = 3 + 0,5\ (6 - 3)\ 30\% = 3 + 0,45 = 3,45$

Xét thuộc tính Thu nhập với khái niệm mờ: 6) Thu nhập-thấp (0,3), 7) Thu nhập-trung bình [3,5), 8) Thu nhập-cao [5,10)

Khi đó $S_1 = 3$, $S_2 = 5$. Giả thiết p = 30%, các đường biên có giá trị là:

$d_1^+ = 3 - 0,5\ (3 - 0)\ 30\% = 3 - 0,45 = 2,55$;

$S_1 = 3$

$d_2^- = 3 + 0,5\ (5 - 3)\ 30\% = 3 + 0,3 = 3,3$

$d_2^+ = 5 - 0,5\ (5 - 3)\ 30\% = 5 - 0,3 = 5,3$

$S_2 = 5$

$d_3^- = 5 + 0,5\ (10 - 5)\ 30\% = 5 + 0,75 = 5,75$

Xét thuộc tính Có gia đình sẽ có khái niệm : 9) Gia đình-có, 10) Gia đình-không

Kết quả sẽ có tập dữ liệu mờ như trong bảng 3.1.

Bảng 3.1: 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	1
t_2	*40*	0	1	0	*3*	0,5	0.5	*6,0*	0	0	1	*c*	1	0
t_3	*30*	0,5	0,5	0	*0*	1	0	*1,5*	1	0	0	*c*	1	0
t_4	*25*	1	0	0	*1*	1	0	*3,0*	0,5	0,5	0	*k*	0	1
t_5	*70*	0	0	1	*2*	1	0	*0,0*	1	0	0	*c*	1	0
t_6	*57*	0	0,83	0,17	*4*	0	1	*4,0*	0	1	0	*c*	1	0

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 $\omega_{\chi_{x_j}}$ mới được tính đến, 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 tập dữ liệu trên lấy $\omega_{\chi_{x_j}}$ bằng 0,4.

Chọn độ hỗ trợ minSup = 0,2 và maxSup = 0,4, ta có bảng 3.2 biểu diễn kết quả tính độ hỗ trợ đối với từng thuộc tính.

Bảng 3.2: Các thuộc tính và độ hỗ trợ của các thuộc tính

Tập thuộc tính	Độ hỗ trợ	Là tập Sporadic tuyệt đối?
Tuổi-trẻ (1)	0,4	Không
Tuổi-trung niên (2)	0,39	Có
Tuổi-già (3)	0,17	Không
Số xe máy-ít (4)	0,75	Không
Số xe máy-nhiều (5)	0,25	Có
Thu nhập-thấp (6)	0,58	Không
Thu nhập-trung bình (7)	0,25	Có
Thu nhập-cao (8)	0,17	Không
Gia đình-có]} (9)	0,67	Không
Gia đình-không (10)	0,33	Có

Như vậy $\mathbf{IF_1}$ = {{2}, {5}, {7}, {10}}

Tập ứng viên sẽ là: {{2,5}, {2,7}, {2,10}, {5,7}, {5,10}, {7,10}}. Bảng 3.3 biểu diễn kết quả tính độ hỗ trợ đối với từng thuộc tính.

Như vậy $\mathbf{IF_2}$ = {{2,5}}. Tập các tập Sporadic tuyệt đối hai ngưỡng mờ là: {{2}, {5}, {7}, {10}, {2,5}}.

Bảng 3.3: Các tập 2-thuộc tính và độ hỗ trợ của các tập dữ liệu

Tập thuộc tính	Độ hỗ trợ	Là tập Sporadic tuyệt đối?
{2,5}	0,22	Có
{2,7}	0,14	Không
{2,10}	0	Không
{5,7}	0,17	Không
{5,10}	0	Không
{7,10}	0,08	Không

Lưu ý: Khi ghép các thuộc tính để tạo tập ứng cử viên không được ghép các thuộc tính có cùng nguồn gốc với nhau. Chẳng hạn, không được ghép thuộc tính Tuổi-trẻ với Tuổi-già vì có cùng gốc ban đầu là Tuổi.

3.2.3.3. 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]. Phần thực nghiệm thi hành trên máy tính Lenovo-IBM Codual 2.0ghz, 2GB bộ nhớ, cài đặt hệ điều hành Windows Vista. Thuật toán MFPSI được lập trình trên ngôn ngữ C++.

CSDL 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.

Bảng 3.4: Kết quả thực hiện thử nghiệm thuật toán MFPSI

		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

Bảng 3.4 là kết quả thực hiện thử nghiệm thuật toán MFPSI. 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

Trong phần này chúng tôi 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 sẽ giới

thiệu phương pháp tìm các tập Sporadic không tuyệt đối mờ cho các luật Sporadic không tuyệt đối mờ bằng cách đề xuất bài toán tìm các luật kết hợp mờ có dạng $r \equiv X \text{ is } A \rightarrow Y \text{ is } B$ sao cho:

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

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 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.

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:

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

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

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 $B \subseteq A$.

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.

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

3.3.3.1. Ý tưởng của 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:

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

(b) Chuyển CSDL ban đầu thành CSDL mờ.

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

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

Đầu vào: CSDL **D**, minSup, maxSup
Kết quả: Tập các tập Sporadic không tuyệt đối hai ngưỡng mờ **FIS**
Bước 1: Chuyển CSDL $\mathbf{D} \subseteq \mathbf{I} \times \mathbf{O}$ ban đầu thành CSDL mờ $\mathbf{D}_F \subseteq \mathbf{I}_F \times \mathbf{O}_F$ 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 $\omega_{\chi_{x_j}}$

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> ∈$\mathbf{I}_F$}
   //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> ∈$\mathbf{I}_F$}
   //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> or <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
8. NewN = ∅ ; X = <X_i,A_i>
9. for each <X_j,A_j> in Nodes do
10. X = X ∪ <X_j,A_j>
11. if NewN ≠ ∅ then MFISI-EXTEND(NewN, **C**)
12. if sup(X) < maxSup then
13. **C** = **C** ∪ X // if X is not subsumed
14. end

Hình 3.2: Thuật toán MFISI

Thuật toán MFISI (Mining Fuzzy Imperfectly Sporadic Itemsets) được đề xuất nhằm tìm các tập Sporadic không tuyệt đối hai ngưỡng mờ. Thuật toán MFISI (hình 3.2) được phát triển từ tư tưởng của thuật toán MCISI [33] tìm các tập Sporadic không tuyệt đối hai ngưỡng trên CSDL tác vụ.

Ví dụ 3.2: Xét trên CSDL mô tả trong bảng 0.2 và CSDL mờ như trong bảng 3.1. Nếu chọn độ hỗ trợ minSup = 0,2 và maxSup = 0,5, ta có bảng 3.5 là kết quả tính độ hỗ trợ đối với từng thuộc tính.

Bảng 3.5: Các thuộc tính và độ hỗ trợ của các thuộc tính

Tập thuộc tính	Độ hỗ trợ
Tuổi-trẻ (1)	0,4
Tuổi-trung niên (2)	0,39
Tuổi-già (3)	0,17
Số xe máy-ít (4)	0,75
Số xe máy-nhiều (5)	0,25
Thu nhập-thấp (6)	0,58
Thu nhập-trung bình (7)	0,25
Thu nhập-cao (8)	0,17
Gia đình-có (9)	0,67
Gia đình-không (10)	0,33

Ban đầu có hai tập **FI** = {{1}, {4}, {6}, {9}} và **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 3.6).

Bảng 3.6: Các tập 2-thuộc tính và độ hỗ trợ của các tập dữ liệu

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

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: {{1,4}, {1,6}, {1,10}, {1,4,6}, {1,4,10}, {1,6,10}, {1,4,6,10}} (bảng 3.7).

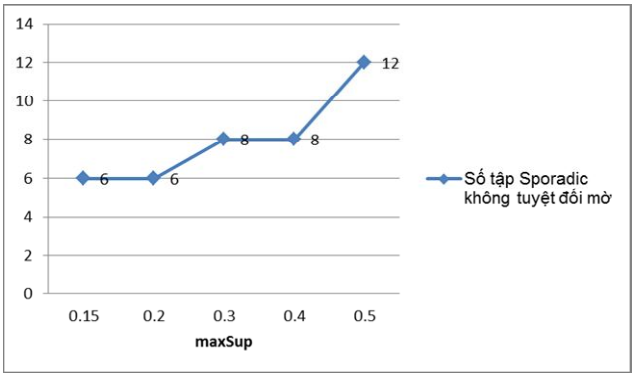
Bảng 3.7: Tập Sporadic không tuyệt đối mờ tìm được ở Nodes thứ nhất

Tập thuộc tính	Độ hỗ trợ
{1,4}	0,41
{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

3.3.3.3. Kết quả thử nghiệm

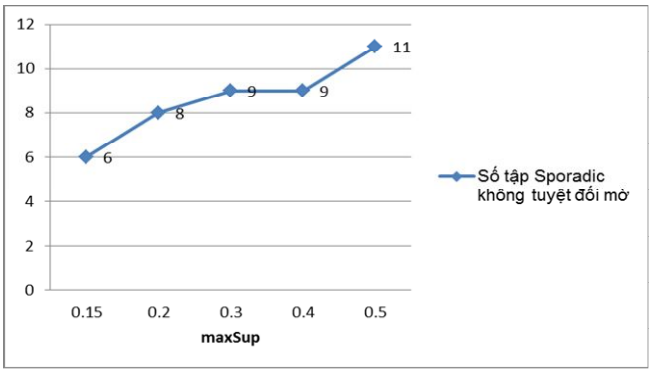
Để đánh giá hiệu quả thực hiện của thuật toán MFPSI, chúng tôi tiến hành thử nghiệm trên CSDL thực Census Income từ nguồn [100] (mô tả về CSDL này được trình bày trong phần 3.2.3.3). Phần thử nghiệm thực hiện trên máy tính Lenovo-IBM Codual 2.0ghz, 2GB bộ nhớ, cài đặt hệ điều hành Windows Vista. Thuật toán MFPSI được lập trình trên ngôn ngữ C++.

Trường hợp 1: chọn tham số chồng lấp là 10%, hệ số minSup = 0,1, hệ số maxSup thay đổi có kết quả về số tập Sporadic không tuyệt đối hai ngưỡng mờ như trong hình 3.3.



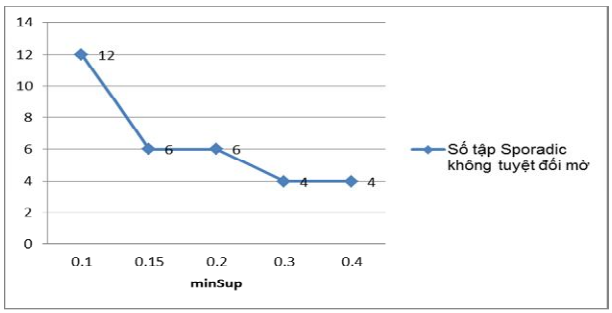
Hình 3.3: Kết quả thử nghiệm ở trường hợp 1

Trường hợp 2: chọn tham số chồng lấp là 40%, hệ số minSup = 0,1, hệ số maxSup thay đổi có kết quả như trong hình 3.4.



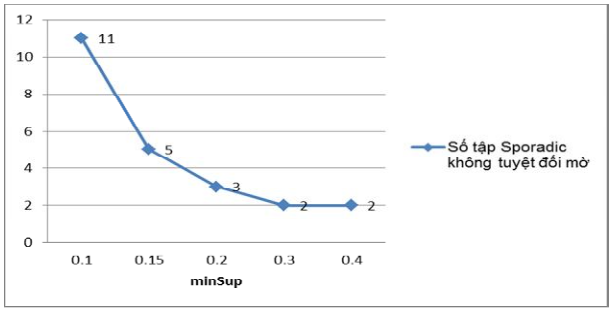
Hình 3.4: Kết quả thử nghiệm ở trường hợp 2

Trường hợp 3: chọn tham số chồng lấp là 10%, hệ số maxSup = 0,5, hệ số minSup thay đổi có kết quả như trong hình 3.5.



Hình 3.5: Kết quả thử nghiệm ở trường hợp 3

Trường hợp 4: chọn tham số chồng lấp là 40%, hệ số maxSup = 0,5, hệ số minSup thay đổi có kết quả như trong hình 3.6.



Hình 3.6: Kết quả thử nghiệm ở trường hợp 4

Kết quả thử nghiệm trong cả bốn trường hợp 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 phát hiện luật kết hợp.

Trường hợp 5: cố định hệ số minSup = 0,1, hệ số maxSup thay đổi từ 0,15 đến 0,5 và tham số chồng lấp thay đổi lần lượt là 10%, 20%, 30%, 40% và 50% có kết quả như trong bảng 3.8.

Bảng 3.8: Kết quả thử nghiệm ở 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	0,4	8	10	9	9	9
0,1	0,5	12	12	11	11	11

Kết quả thử nghiệm ở trường hợp 5 cho thấy 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.

Cũng giống như vấn đề tìm tập Sporadic tuyệt đối hai ngưỡng mờ (mục 3.1, chương 3), các hệ số minSup, maxSup và tham số chồng lấp có ảnh hưởng đến số tập Sporadic không tuyệt đối hai ngưỡng mờ tìm được. Vấn đề lựa chọn các giá trị ngưỡng phù hợp với từng CSDL chúng tôi chưa thực hiện trong phạm vi nghiên cứu của luận án này. Đây là định hướng nghiên cứu tiếp theo của chúng tôi trong tương lai.

Kết luận chương 3:

Trong chương 3, chúng tôi đã đề xuất bài toán phát hiện luật kết hợp Sporadic tuyệt đối hai ngưỡng mờ và luật kết hợp Sporadic không tuyệt đối hai ngưỡng mờ trên CSDL định lượng. Hai thuật toán lần lượt được phát triển là MFPSI và MFISI tương ứng nhằm tìm các tập phổ biến cho các luật kết hợp hiếm mờ. Thuật toán MFPSI được phát triển từ tư tưởng của thuật toán Apriori. Thuật toán MFISI được phát triển từ thuật toán MCISI đã được trình bày ở chương 2. Phần thử nghiệm cũng đã chứng tỏ hiệu quả của thuật toán do chúng tôi phát triển.

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

Nội dung của chương 4 bàn về vấn đề ứng dụng luật kết hợp và mô hình hồi quy chuyển tiếp trơn trong xây dựng mô hình phân tích và dự báo kinh tế. Hai lĩnh vực được chúng tôi lựa chọn là dự báo chỉ số chứng khoán, dự báo giá hàng hóa và chỉ số giá tiêu dùng CPI. Kết quả nghiên cứu lần lượt đã được đăng trên tạp chí Tin học và Điều khiển học [1], tạp chí Journal on Information Technologies and Communications [36] và kỷ yếu Hội thảo lần thứ hai trong khuôn khổ Nghị định thư Việt Nam - Thái Lan [7].

4.1. Mô hình hồi quy chuyển tiếp trơn

4.1.1. Phân tích hồi quy

Phân tích hồi quy là phương pháp nghiên cứu các mối quan hệ kinh tế xã hội có tính chất tương đối, không phải là quan hệ hàm số chặt chẽ. Phân tích hồi quy nghiên cứu mối liên hệ phụ thuộc của một biến (gọi là biến phụ thuộc, biến được giải thích) với một hay nhiều biến khác (biến độc lập, biến giải thích), trong đó ước lượng giá trị trung bình của biến phụ thuộc theo các giá trị xác định của biến độc lập [11]. Trong chương này chúng tôi ứng dụng mô hình hồi quy phi tuyến để xây dựng mô hình phân tích và dự báo kinh tế.

Các mô hình phi tuyến được ứng dụng khá phổ biến và nói chung cho chất lượng dự báo cao hơn các mô hình tuyến tính trong dự báo như: dự báo các biến kinh tế; dự báo lưu lượng nước sông; dự báo các hiện tượng thuộc ngành khí tượng học như nhiệt độ trái đất, nhiệt độ nước biển, mức độ bao phủ của mây, vệt đen của mặt trời,...; dự báo cỡ của các quần thể động vật, các quá trình hóa sinh;... [48].

Các mô hình phi tuyến được phân thành hai nhóm. Nhóm thứ nhất gồm các mô hình, ở đó mô hình tuyến tính không phải là dạng đặc biệt của nó. Nhóm thứ hai gắn với một số mô hình phi tuyến bao trùm cả các mô hình tuyến tính. Mô hình hồi

quy chuyển tiếp trơn phi tuyến thuộc nhóm thứ hai. Mô hình này lấy mô hình tuyến tính làm xuất phát điểm và sau đó xem xét, mở rộng dạng phi tuyến nếu thấy chúng cần thiết [62].

4.1.2. Mô hình hồi quy chuyển tiếp trơn logistic

Mô hình hồi quy chuyển tiếp trơn (STR) do Bacon và Watts giới thiệu lần đầu tiên năm 1971 [19]. Các tác giả đã sử dụng hàm hyperbol để mô tả tính chuyển tiếp. Năm 1977, Maddala đã sử dụng hàm logistic làm hàm chuyển tiếp [60]. Năm 1996, Teräsvirta giới thiệu mô hình này trong [78] và từ đó đã trở thành dạng chuẩn.

Mô hình STR chuẩn tổng quát có dạng như sau [79]:

$$\begin{aligned} y_t &= \phi' z_t + \theta' z_t G(\gamma, c, s_t) + u_t \\ &= \{\phi + \theta G(\gamma, c, s_t)\}' z_t + u_t \qquad t = 1,2,...T \end{aligned} \tag{4.1}$$

trong đó $z_t = (w_t', x_t')$' là một véctơ các biến giải thích, $w_t' = (1, y_{t-1},\ldots, y_{t-p})$', và $x_t = (x_{1t}, \ldots, x_{kt})$' là một véctơ các biến ngoại sinh. Ngoài ra, $\phi = (\phi_0, \phi_1, \ldots,\phi_m)$' và $\theta = (\theta_0, \theta_1, \ldots, \theta_m)$' là véctơ tham số ($(m+1)\times 1$) và $u_t \sim iid(0, \sigma^2)$. Hàm chuyển tiếp $G(\gamma, c, s_t)$ là một hàm của biến chuyển tiếp liên tục s_t bị chặn, nó liên tục tại mọi vị trí trong không gian tham số với mọi giá trị của s_t, γ là tham số độ dốc, và $c = (c_1, \ldots, c_k)$' là véctơ các tham số vị trí, $c_1 \le \ldots \le c_k$.

Giả định rằng hàm chuyển tiếp là hàm logistic tổng quát như sau:

$$G(\gamma, c, s_t) = \left(1 + \exp\left\{-\gamma \prod_{k=1}^{K} (s_t - c_k)\right\}\right)^{-1}, \qquad \gamma > 0 \tag{4.2}$$

trong đó $\gamma > 0$ là tham số.

Các phương trình (4.1) và (4.2) xác định mô hình STR logistic (LSTR). Các lựa chọn phổ biến nhất của K là K = 1 hoặc K = 2. Đối với K = 1, các tham số $\phi + \theta G(\gamma, c, s_t)$ thay đổi đơn điệu và là một hàm của s_t từ ϕ tới $\phi + \theta$. Đối với K = 2, chúng thay đổi đơn điệu xung quanh điểm giữa $(c_1 + c_2)/2$, tại đó hàm logistic đạt giá trị cực tiểu. Giá trị cực tiểu nằm giữa 0 và 1/2. Nó đạt giá trị 0 khi $\gamma \to \infty$ và

bằng 1/2 khi $c_1 = c_2$ và $\gamma < \infty$. Tham số độ dốc γ sẽ kiểm soát độ dốc, c_1 và c_2 là các tham số vị trí của hàm chuyển tiếp.

Mô hình LSTR với K = 1 (LSTR1) có thể mô hình hóa hành vi bất đối xứng. Ví dụ s_t đo lường các giai đoạn trong chu kỳ kinh doanh, khi đó mô hình LSTR1 có thể mô tả những giai đoạn mà tính chất động của chúng trong miền tăng trưởng khác với tính chất động trong miền suy thoái và chuyển tiếp từ thái cực này sang thái cực kia là trơn. Mặt khác, mô hình LSTR2 (K = 2) phù hợp trong những trường hợp mà tính chất động cục bộ của quá trình tương tự nhau ứng với giá trị lớn và nhỏ của s_t nhưng lại khác khi nó nhận giá trị trung bình ở giữa.

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

Phần này sẽ xem xét việc mô hình hóa các quan hệ phi tuyến bằng cách sử dụng mô hình STR (4.1) với hàm chuyển tiếp (4.2). Quá trình mô hình hóa bao gồm các giai đoạn: chỉ định mô hình, ước lượng các tham số, và đánh giá mô hình.

4.1.3.1.Chỉ định mô hình

Giai đoạn chỉ định gồm hai bước. Đầu tiên, thực hiện kiểm định tính chất tuyến tính của mô hình. Bước tiếp theo là lựa chọn mô hình. Nếu mô hình không phải là tuyến tính thì mô hình STR (dạng LSTR1 hoặc LSTR2) sẽ được lựa chọn.

a. Kiểm định tuyến tính

Vấn đề kiểm định tính tuyến tính của mô hình được thực hiện bằng cách xấp xỉ hàm (4.1) với hàm chuyển tiếp (4.2) bằng khai triển Taylor xung quanh giả thuyết gốc $\gamma = 0$. Thông thường ban đầu người ta thường giả định K = 1 trong (4.2) và sử dụng xấp xỉ Taylor đến bậc ba. Sau đó ta kiểm định theo cách tương tự như vậy với K = 2 trong hàm chuyển tiếp (4.2).

b. Lựa chọn dạng mô hình

Khi tính chất tuyến tính bị bác bỏ và chọn xong một biến chuyển tiếp (thường do các phần mềm ứng dụng lựa chọn trên cơ sở tính toán tối ưu giữa các biến trong mô hình) thì chuyển sang bước tiếp theo là chọn mô hình, tức là lựa chọn mô hình STR với K = 1 hoặc K = 2 trong hàm chuyển tiếp (4.2).

4.1.3.2. Ước lượng tham số mô hình

a. *Chọn giá trị ban đầu*

Các tham số của mô hình STR được ước lượng bằng cách sử dụng phương pháp hợp lý cực đại có điều kiện. Khi đó việc tìm ra giá trị ban đầu phù hợp là rất quan trọng. Trong thực hành, có những phần mềm ứng dụng giúp ta lựa chọn giá trị ban đầu của mô hình dựa trên dạng mô hình được lựa chọn và tập số liệu thực tế được sử dụng để ước lượng các tham số của mô hình.

b. Ước lượng tham số

Khi đã tìm được giá trị đầu, các tham số chưa biết có thể ước lượng được bằng sử dụng thuật toán Newton-Raphson để cực đại hàm hợp lý có điều kiện.

c. Phân tích kiểm định thống kê của các tham số

Dựa vào bảng ước lượng các tham số của mô hình, phân tích ý nghĩa thống kê của các tham số trong mô hình như thống kê T, F, độ đo sự phù hợp của mô hình R^2 và R^2 được hiệu chỉnh; thống kê D-W đo tương quan phần dư, sai số chuẩn của mỗi tham số ước lượng,...

Một mô hình được coi là được chấp nhận về mặt thống kê nếu các kiểm định đều có ý nghĩa thống kê.

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

Các bước đánh giá – kiểm định sai lầm trong chỉ định mô hình gồm:

- Kiểm định không có tự tương quan phần dư (trong trường hợp chuỗi thời gian);
- Kiểm định không có thành phần phi tuyến bị bỏ sót;
- Kiểm định tính hội tụ của tham số, tức là khi tăng mẫu quan sát thì mỗi tham số của mô hình sẽ hội tụ đến một giá trị nào đó.
- Ngoài ra còn thực hiện một số kiểm định khác như: kiểm định giả thuyết gốc là không xẩy ra hiện tượng phương sai thay đổi tự hồi quy (ARCH),...

4.2. Ứng dụng luật kết hợp mẫu âm và mô hình hồi quy chuyển tiếp trơn trong xây dựng mô hình phân tích và dự báo chỉ số chứng khoán

Đã có nhiều nghiên cứu và nhiều phương pháp được đề xuất để phân tích và dự báo diễn biến của các chỉ số chứng khoán. Những phương pháp phân tích kỹ thuật chỉ số chứng khoán được ưa chuộng hiện nay thường được dựa trên việc trực quan hoá và phân tích số liệu thống kê, trong khi việc dự báo chỉ số chứng khoán thường được xây dựng dựa trên mô hình mạng nơtron hoặc mô hình phân tích, dự báo chuỗi thời gian [8, 48, 62, 79].

Như đã biết, mức độ tin cậy của dự báo phụ thuộc vào căn cứ khoa học được ứng dụng để xây dựng dự báo tốt đến mức độ nào? Tuy nhiên ngay cả khi dự báo được dựa trên những cách tiếp cận khoa học thì dường như vẫn là chưa đủ. Tương lai quá bất định là khó khăn chủ yếu khi thực hiện dự báo bởi vì rất khó đoán định tương lai của những thứ mà chính chúng ta cũng không biết rằng chúng ta không biết. Chúng ta chỉ có thể dự báo được, hoặc ít nhất là có thể tưởng tượng được dựa trên những gì chúng ta đã biết.

Các mô hình dự báo *không điều kiện* nói chung [35], là các mô hình dự báo được xây dựng dựa vào mạng nơtron hoặc mô hình phân tích, dự báo chuỗi thời gian đều có giả định rằng tương lai được diễn ra giống hoặc ít nhất là gần giống với hiện tại và quá khứ. Bởi lẽ vậy khi tương lai được tiên lượng có những biến động bất thường thì việc sử dụng các mô hình dự báo không điều kiện sẽ cho kết quả dự báo nói chung cũng có độ chính xác không cao.

Mặt khác như đã biết nhược điểm lớn nhất của các mô hình dự báo được xây dựng dựa vào mạng nơtron hoặc mô hình phân tích, dự báo chuỗi thời gian là ở chỗ nó không hỗ trợ cho các hoạt động phân tích, tìm ra nguyên nhân, xác định được những yếu tố chủ yếu tác động đến sự biến động của biến cần được dự báo vì thế chúng có rất ít khả năng hỗ trợ xây dựng giải pháp vượt qua thách thức.

Nhằm ứng phó với sự bất định của tương lai và sự hạn chế của các mô hình dự báo không điều kiện, khác với các cách tiếp cận trước đó về dự báo chỉ số chứng

khoán, chúng tôi đề xuất xây dựng mô hình dự báo chỉ số chứng khoán của Việt Nam theo *mô hình dự báo có điều kiện* [35], theo đó có thể hình thành nhiều kịch bản dự báo dựa trên những giả định khác nhau về các biến độc lập (hay biến ngoại sinh) tham gia trong mô hình dự báo. Cụ thể mô hình dự báo chỉ số chứng khoán của Việt Nam được xây dựng dựa vào mô hình hồi quy chuyển tiếp trơn phi tuyến như trình bày trong phần 4.1 và dựa trên mối tương quan giữa chỉ số chứng khoán đó với những mã cổ phiếu blue chip trên hai sàn giao dịch Hà Nội và thành phố Hồ Chí Minh.

Phân tích các thị trường chứng khoán đều cho thấy trong mỗi phiên giao dịch thường xẩy ra một số mã cổ phiếu tăng giá, một số mã cổ phiếu giữ nguyên giá trong khi một số mã cổ phiếu khác lại giảm giá. Như vậy các luật kết hợp được phát hiện từ CSDL phản ánh biến động của các chỉ số chứng khoán Việt Nam và của giá các cổ phiếu blue chip là các luật kết hợp mẫu âm. Tuy nhiên các luật kết hợp như vậy chỉ cho biết mối quan hệ tiền đề - kết quả về biến động giá giữa chỉ số chứng khoán với một số cổ phiếu blue chip mà không định lượng được mối tương quan đó. Mô hình phân tích và dự báo kinh tế hứa hẹn có thể giúp giải quyết được vấn đề này.

Như đã biết mối tương quan giữa các biến kinh tế nói chung, giữa các chỉ số chứng khoán và cổ phiếu blue chip Việt Nam nói riêng phần lớn không phải là quan hệ tuyến tính mà là quan hệ phi tuyến. Vấn đề xác định mô hình hồi quy phi tuyến giữa các biến kinh tế tuy sớm được quan tâm nghiên cứu, nhưng việc xây dựng được những mô hình như vậy là rất khó khăn. Hiện nay người ta cũng đã xây dựng được phần mềm ứng dụng hỗ trợ việc xây dựng mô hình hồi quy chuyển tiếp trơn phi tuyến ở đó hàm chuyển tiếp trơn có dạng hàm mũ hoặc dạng logicstic [99].

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.

Mặt khác, như đã biết quá trình thực hiện dự báo bằng mô hình định lượng cần phải được tiến hành theo nguyên tắc 3 bước [35]. Bước thứ nhất (được gọi là bước

dự báo trong mẫu hay *dự báo hậu nghiệm*) nhằm xây dựng mô hình dự báo đối với vấn đề đặt ra. Bước tiếp theo (được gọi là bước *dự báo kiểm nghiệm*) nhằm đánh giá độ chính xác của kết quả dự báo so với thực tiễn, nếu độ chính xác của dự báo đáp ứng yêu cầu đề ra thì mô hình dự báo được chấp nhận để dự báo tương lai. Bước thứ ba - cuối cùng (được gọi là bước *dự báo tiên nghiệm*) sẽ ứng dụng mô hình được xây dựng ở Bước thứ nhất để dự báo tương lai của vấn đề được đặt ra.

Để thực hiện nguyên tắc này, ta phải chia tập dữ liệu thu thập được thành hai tập thành phần với hai mục đích sử dụng khác nhau. Tập thứ nhất dùng để xây dựng mô hình dự báo được chấp nhận về mặt thống kê (bước thứ nhất) và tập thứ hai được sử dụng để dự báo kiểm nghiệm (bước thứ hai).

Về bản chất tập thứ hai thực tế là ta đã biết, nhưng nó không được sử dụng để xây dựng mô hình, nó được dùng để đối chiếu, so sánh với kết quả dự báo theo mô hình được xây dựng dựa trên tập dữ liệu thứ nhất. Kết quả so sánh giá trị dự báo và giá trị thực tế là nhỏ có thể chấp nhận được như yêu cầu của người làm dự báo (ví dụ tổng trung bình bình phương sai số không vượt quá ngưỡng nào đó hoặc phần trăm sai số tuyệt đối của kết quả dự báo so với giá trị thống kê thực tế của nó không vượt quá mức ngưỡng nào đó như mức 1%, 5%, hay 10%,…) thì có thể sử dụng mô hình này để dự báo giá trị tương lai của các biến trong mô hình. Nguyên tắc này sẽ được tuân thủ một cách đầy đủ khi xây dựng mô hình dự báo chỉ số chứng khoán Việt Nam.

Quy trình xây dựng mô hình dự báo chỉ số chứng khoán: quy trình này được thực hiện qua 2 giai đoạn. Giai đoạn 1 nhằm phát hiện các luật kết hợp biểu diễn mối quan hệ 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” 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.

Ví dụ: ACB là mã cổ phiếu của Ngân hàng Thương mại Á châu, ACB1 là ký hiệu mã cổ phiếu này tăng giá so với phiên trước đó, ACB2 là ký hiệu mã cổ phiếu này giảm giá và nó chính là mục dữ liệu mẫu âm.

Theo cách này ta nhận được CSDL tác vụ có mẫu âm, một phần của nó được thể hiện ở dạng như trong hình 4.1.

```
PVS2, BVS2, PVI2, NTP2, KLS2, VSP2, BCC2, BTS2, REE2, SAM2, SJS2, HOSE2, HNX2, DJI1,...
BVS2, PVI2, NTP2, KLS2, VSP2, BCC2, BTS2, REE2, SAM2, SJS2, HOSE2, HNX2, DJI2,...
NTP1, KLS2, VSP2, BTS2, REE2, SAM2, SJS2, HOSE2,HNX2, DJI1,...
ACB2, PVS2, BVS2, PVI2, NTP1, KLS2, VSP2, BCC1, BTS1, SAM2, SJS2, HOSE2, HNX2, DJI1,...
ACB1, PVS1, BVS2, PVI1, NTP1, KLS2, VSP1, BCC1, BTS1, REE2, SAM2, HOSE2, HNX1, DJI2,...
```

Hình 4.1: Tập dữ liệu về chứng khoán

Nhận xét: Bài toán phát hiện luật kết hợp có độ phức tạp hàm mũ đối với số các mục dữ liệu trong CSDL nên về mặt lý thuyết ta khó có thể phát hiện được các luật này khi số mục dữ liệu là khá lớn. Tuy nhiên trong thực tiễn vẫn phát hiện được các luật kết hợp ngay cả trong trường hợp số các mục dữ liệu là rất lớn. Nguyên nhân của hiện tượng này là dữ liệu trong CSDL tác vụ nói chung là thưa. Tình trạng thưa có thể sẽ mất đi nếu CSDL còn chứa nhiều mục dữ liệu âm.

Như đã biết việc xây dựng thuật toán hiệu quả, khả thi để phát hiện luật kết hợp mẫu âm cho đến nay vẫn là vấn đề mở tuy rằng đã có một số kết quả nghiên cứu quan trọng về cơ sở lý thuyết của các luật này [52]. Với nhận xét rằng bằng việc chuyển đổi biểu diễn CSDL tác vụ theo cách vừa được giới thiệu ở trên, ta có thể đưa bài toán phát hiện luật kết hợp mẫu âm về bài toán phát hiện luật kết hợp từ CSDL tác vụ thông thường (tức là chỉ gồm mục dữ liệu mẫu dương).

Trong trường hợp bài toán dự báo chỉ số chứng khoán Việt Nam do số lượng các cổ phiếu blue chip và các chỉ số chứng khoán là không lớn (31 cổ phiếu blue chip, 2 chỉ số chứng khoán) nên có thể biểu diễn CSDL tác vụ mẫu âm theo cách ở trên và khi đó nhiều luật kết hợp phát hiện được từ CSDL này thực chất là luật kết hợp mẫu âm. Nói cách khác trong nhiều trường hợp ta có thể phát hiện luật kết hợp mẫu âm theo cách phát hiện luật kết hợp từ CSDL tác vụ thông thường.

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. Phân tích các luật này cho thấy:

15 luật có độ hỗ trợ cao nhất (các luật từ Rule 1 đến Rule 15) đều là các luật chỉ chứa các mẫu âm hay các luật liên quan đến các mã cổ phiếu giảm giá. Các luật này không chỉ cho biết những mã cổ phiếu nào có tỷ lệ các phiên giảm giá so với tổng các phiên giao dịch cao nhất và vượt mức 35% của tổng số phiên mà còn cho biết những tín hiệu giảm giá của mã cổ phiếu đó. 84 luật kết hợp còn lại đều chỉ chứa các mẫu dương, đó là các luật chỉ chứa các mã cổ phiếu tăng giá. Như vậy có thể nói trong 350 phiên giao dịch được chọn thì xu thế tăng giá của các mã cổ phiếu và của các chỉ số chứng khoán vẫn là chủ yếu.

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 phải lựa chọn các luật kết hợp chỉ có mục dữ liệu liên quan đến HNX hoặc HOSE ở phần kết quả của luật. Trong

tập dữ liệu này tất cả các luật kết hợp mà phần kết quả có chứa chỉ số HNX hoặc HOSE thì cũng đều chỉ chứa riêng các chỉ số đó.

Cụ thể có 7 luật chứa HNX ở phần kết quả là:

Rule 20: PVS1; ACB1 → HNX1 (39,264% 90,63% 128 116 35,583%)

Rule 21: PVI1; ACB1 → HNX1 (38,037% 94,35% 124 117 35,890%)

Rule 23: SD91; ACB1 → HNX1 (39,264% 91,41% 128 117 35,890%)

Rule 24: VN1; ACB1 → HNX1 (39,877% 92,31% 130 120 36,810%)

Rule 27: KLS1; ACB1 → HNX1 (40,184% 90,08% 131 118 36,196%)

Rule 58: HPC1; VN1 → HNX1 (39,877% 90,00% 130 117 35,890%)

Rule 61: SD71; VN1 → HNX1 (39,877% 90,00% 130 117 35,890%)

và có 15 luật chứa HOSE ở phần kết quả là:

Rule 22: PVI1; ACB1 → HOSE1 (38,037% 93,55% 124 116 35,583%)

Rule 25: HNX1; ACB1 → HOSE1 (40,491% 90,91% 132 120 36,810%)

Rule 29: SD91; ACB1 → HOSE1 (39,264% 92,19% 128 118 36,196%)

Rule 33: KLS1; ACB1 → HOSE1 (40,184% 92,37% 131 121 37,117%)

Rule 38: PVI1; PVS1 → HOSE1 (40,798% 90,23% 133 120 36,810%)

Rule 39: HNX1; PVS1 → HOSE1 (41,411% 91,85% 135 124 38,037%)

Rule 45: HNX1; PVI1 → HOSE1 (40,491% 92,42% 132 122 37,423%)

Rule 50: SD91; HNX1 → HOSE1 (40,798% 91,73% 133 122 37,423%)

Rule 55: VCS1; HNX1 → HOSE1 (41,104% 90,30% 134 121 37,117%)

Rule 56: SDT1; HNX1 → HOSE1 (40,184% 93,13% 131 122 37,423%)

Rule 57: KLS1; HNX1 → HOSE1 (40,184% 93,13% 131 122 37,423%)

Rule 59: HPC1; HNX1 → HOSE1 (38,650% 92,86% 126 117 35,890%)

Rule 60: BVS1; HNX1 → HOSE1 (38,344% 93,60% 125 117 35,890%)

Rule 62: SD71; HNX1 → HOSE1 (38,957% 92,13% 127 117 35,890%)

Rule 84: SDT1; VCS1 → HOSE1 (39,264% 90,63% 128 116 35,583%)

Điểm đáng lưu ý là tất cả các luật có HNX hoặc HOSE ở phần kết quả đều cho

thấy HNX, HOSE tăng điểm trong khi các mã chứng khoán blue chip khác đều tăng giá. Như vậy xu thế tăng điểm của các chỉ số chứng khoán Việt Nam trong 350 phiên giao dịch được chọn vẫn là chủ đạo, tỷ lệ các phiên có chỉ số HNX hoặc HOSE giảm điểm là không quá 35%. Điều đó là phù hợp với thực tiễn dù rằng cuối năm 2008 và đầu năm 2009, các mã cổ phiếu và hầu hết các chỉ số chứng khoán đều giảm và giảm rất sâu với tốc độ rất nhanh, việc hồi phục tăng điểm thì diễn ra từ từ và chậm chạp hơn nhiều.

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. Phương pháp xây dựng mô hình dự báo chỉ số chứng khoán dựa trên mô hình hồi quy chuyển tiếp trơn phi tuyến và dựa trên các luật kết hợp được phát hiện như vậy là như nhau nên dưới đây chỉ trình bầy việc xây dựng mô hình dự báo chỉ số HNX dựa trên một luật kết hợp cụ thể, việc xây dựng mô hình dự báo chỉ số HNX hoặc HOSE dựa trên các luật kết hợp khác được tiến hành tương tự.

Xét luật **Rule 21**:

PVI1; ACB1 $\rightarrow$ 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.

4.2.3.1. 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 Rule 21, 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.

Ứng dụng phần mềm JMULTI [99] 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, ta nhận được kết quả ước lượng tham số của mô hình trong hình 4.2.

variable	start	estimate	SD	t-stat	p-value
----- **linear part** ------					
CONST	17.46413	18.87363	6.9187	2.7279	0.0067
HNX_d1(t-1)	12.94189	13.43648	3.1220	4.3038	0.0000
ACB_d1(t)	0.40047	0.44047	0.2080	2.1173	0.0350
PVI_d1(t)	-28.38974	-29.39717	6.3765	-4.6102	0.0000
PVI_d1(t-3)	-4.72523	-5.00314	1.7301	-2.8919	0.0041
---- **nonlinear part** ----					
CONST	-17.42497	-18.84424	6.9393	-2.7156	0.0070
HNX_d1(t-1)	-13.03255	-13.53159	3.1281	-4.3258	0.0000
ACB_d1(t)	1.55070	1.51878	0.2119	7.1660	0.0000
PVI_d1(t)	28.35787	29.37854	6.3928	4.5956	0.0000
PVI_d1(t-3)	4.81641	5.10029	1.7455	2.9219	0.0037
Gamma	4.85242	4.06359	1.8016	2.2555	0.0248
C1	-4.87586	-5.24045	1.0120	-5.1782	0.0000

Hình 4.2: Ước lượng các tham số của mô hình dự báo chứng khoán

Theo bảng ước lượng này ta thấy xác suất của thống kê T (p-value) của các thành phần tuyến tính và phi tuyến đều có ý nghĩa thống kê nên mô hình biểu diễn mối quan hệ giữa giá trị của chỉ số chứng khoán HNX và giá của các mã cổ phiếu PVI và ACB sẽ gồm hai phần tuyến tính và phi tuyến. Cụ thể mô hình có dạng:

$$HNX\_d1(t) = \begin{pmatrix} 18{,}87 + 13{,}44 HNX\_d1(t-1) + 0{,}44 ACB\_d1(t) \\ - 29{,}40 PVI\_d1(t) - 5{,}0 PVI\_d1(t-3) \end{pmatrix} +$$

$$\begin{pmatrix} -18{,}84 - 13{,}53 HNX\_d1(t-1) + 1{,}5 ACB\_d1(t) \\ + 29{,}38 PVI\_d1(t) + 5{,}1 PVI\_d1(t-3) \end{pmatrix} * \frac{1}{1 + \exp(-4{,}06 * [ACB\_d1(t) + 5{,}24])}$$

ở đây HNX_d1, ACB_d1, PVI_d1 tương ứng là ký hiệu sai phân bậc 1 của HNX, ACB và PVI.

Việc kiểm định sai lầm của chỉ định mô hình như kiểm định không có tự tương quan phần dư, kiểm định thành phần phi tuyến bị bỏ sót, kiểm định tính hội tụ của các tham số, kiểm định phương sai thay đổi điều kiện tự hồi quy (ARCH),... cho thấy không có sai lầm trong chỉ định mô hình. Điều đó có nghĩa là mô hình dự báo chỉ số chứng khoán HNX được xác định ở trên là được chấp nhận về mặt kiểm định thống kê.

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.

Phần tuyến tính của mô hình phi tuyến trên cho thấy giá trị sai phân bậc 1 của chỉ số chứng khoán HNX biến đổi cùng chiều với sai phân bậc 1 của nó sau 01 trễ (hay sau một phiên giao dịch), biến đổi cùng chiều với sai phân bậc 1 của mã cổ phiếu ACB và biến đổi trái chiều với với sai phân bậc 1 của mã cổ phiếu PVI trong cả hai trường hợp không có trễ và sau 3 trễ.

Phần phi tuyến của mô hình bao gồm tích hai thành phần. Thành phần thứ nhất là thành phần tự hồi quy và có dạng tương tự như phần tuyến tính trong khi thành phần thứ hai là hàm logicstic với hàm chuyển tiếp trơn là sai phân bậc 1 của mã chứng khoán ACB với tham số vị trí là c1 = -5,24 và tham số độ dốc là 4,06. Thành phần thứ hai cho biết sự biến động của giá trị sai phân bậc 1 của chỉ số chứng khoán HNX trong miền tăng trưởng khác với tính chất biến động của nó trong miền suy thoái và việc chuyển tiếp từ thái cực này sang thái cực kia là trơn.

4.2.3.2. 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, ta có kết quả trong bảng 4.1.

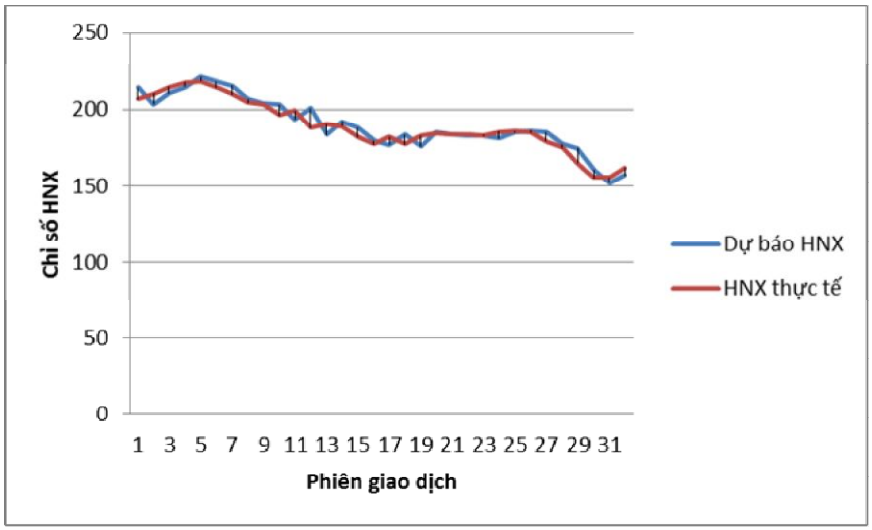
Bảng 4.1: Chỉ số HNX được tính theo mô hình xây dựng và thực tế

TT	Phiên giao dịch (ngày)	Dự báo HNX	Thực tế	Chênh lệch điểm	% sai số tuyệt đối
1	16/10/2009	214,38	206,82	-7,56	0,0366
2	19/10/2009	203,37	210,06	6,69	0,0318
3	20/10/2009	210,41	214,87	4,46	0,0208
4	21/10/2009	214,68	217,74	3,06	0,0141
5	22/10/2009	221,2	218,38	-2,82	0,0129
6	23/10/2009	218	214,27	-3,73	0,0174
7	26/10/2009	215,38	209,77	-5,61	0,0267
8	27/10/2009	206,85	204,44	-2,41	0,0118
9	28/10/2009	203,79	203,51	-0,28	0,0014
10	29/10/2009	202,93	196,14	-6,79	0,0346
11	30/10/2009	193,72	199,38	5,66	0,0284
12	02/11/2009	200,82	188,02	-12,8	0,0681
13	03/11/2009	183,33	190,27	6,94	0,0365
14	04/11/2009	191,86	189,14	-2,72	0,0144
15	05/11/2009	187,96	181,81	-6,15	0,0338
16	06/11/2009	179,53	177,34	-2,19	0,0123
17	09/11/2009	176,85	181,81	4,96	0,0273
18	10/11/2009	183,3	177,34	-5,96	0,0336
19	11/11/2009	175,7	182,59	6,89	0,0377
20	12/11/2009	184,68	184,37	-0,31	0,0017
21	13/11/2009	183,78	183,64	-0,14	0,0008
22	16/11/2009	182,85	183,17	0,32	0,0017
23	17/11/2009	182,84	182,67	-0,17	0,0009
24	18/11/2009	181,13	185,17	4,04	0,0218
25	19/11/2009	184,79	185,68	0,89	0,0048
26	20/11/2009	185,84	184,79	-1,05	0,0057
27	23/11/2009	184,71	179,13	-5,58	0,0312
28	24/11/2009	177,48	175,2	-2,28	0,0130
29	25/11/2009	174,49	164,87	-9,62	0,0583
30	26/11/2009	160,68	155,19	-5,49	0,0354
31	27/11/2009	152,01	155,41	3,4	0,0219
32	30/11/2009	156,79	161,86	5,07	0,0313

Do biên độ biến động của chỉ số chứng khoán HNX được qui định ở mức 5% nên khi dự báo chỉ số này nếu kết quả dự báo có phần trăm sai số tuyệt đối không vượt quá 0,025% thì dự báo có thể được xem là chính xác cao. Kết quả thử nghiệm

về dự báo chỉ số chứng khoán HNX theo mô hình được xác định trong 32 phiên cho thấy có 17 phiên bằng 53,2% cho kết quả dự báo là chính xác cao. Nếu xem ngưỡng của phần trăm sai số tuyệt đối của kết quả dự báo chỉ số chứng khoán HNX theo mô hình và thực tiễn là 0,03% thì sẽ có 20 phiên xấp xỉ bằng 67% cho kết quả dự báo không vượt quá ngưỡng này.

Biểu diễn trực quan chuỗi giá trị chỉ số HNX thực tế và dự báo, ta nhận được đồ thị ở hình 4.3.



Hình 4.3: Chỉ số HNX được tính theo mô hình xây dựng và thực tế

Kết quả dự báo kiểm nghiệm cho thấy ta có thể chấp nhận và ứng dụng mô hình dự báo trên để phân tích và dự báo chỉ số chứng khoán HNX. Mô hình này cho thấy kiểu phụ thuộc của chỉ số chứng khoán HNX vào giá của các cổ phiếu ACB và PVI được thể hiện thông qua các sai phân cấp 1 của nó và giải thích khá tốt quan hệ về sự biến động của chỉ số HNX và giá các cổ phiếu ACB và PVI.

4.2.3.3. 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).

Theo mô hình trên, HNX_d1 là biến nội sinh, các biến ACB_d1 và PVI_d1 là các biến ngoại sinh, và để dự báo HNX_d1(t) trước hết ta phải dự báo ACB_d1(t), PVI_d1(t) và tính các giá trị HNX_d1(t-1), PVI_d1(t-3).

Trong mô hình dự báo có điều kiện, mức độ chính xác của dự báo ngoài việc phụ thuộc vào chất lượng của mô hình đó, còn phụ thuộc vào kết quả dự báo các biến ngoại sinh (hay biến độc lập). Để dự báo các biến ngoại sinh trong bối cảnh tương lai có nhiều bất ổn khó lường người ta thường phải kết hợp phương pháp dự báo định tính với phương pháp dự báo định lượng. Trong dự báo định lượng các biến ngoại sinh, các mô hình dự báo chuỗi thời gian, nhất là mô hình ARIMA hoặc GARCH thường được sử dụng nhất [62].

Trong trường hợp của mô hình dự báo chỉ số chứng khoán vừa được xây dựng, phương pháp dự báo định lượng ACB_d1(t) và PVI_d1(t) là rất tương tự như phương pháp dự báo được trình bầy trong [6]. Phương pháp dự báo định tính các biến ngoại sinh ACB_d1(t) và PVI_d1(t) được thực hiện trên cơ sở phân tích và lượng hóa hiệu quả hoạt động sản xuất kinh doanh, bối cảnh của thị trường liên quan đến lĩnh vực hoạt động của ACB và PVI, tâm lý và kỳ vọng nhà đầu tư về hai công ty này,…

Để lượng hóa các yếu tố đã được phân tích cần hình thành một số tập giả định khác nhau về các biến ngoại sinh ACB_d1(t) và PVI_d1(t). Với mỗi tập giả định đó, kết hợp với kết quả dự báo định lượng của ACB_d1(t) và PVI_d1(t) tương ứng sẽ xác định được một bộ giá trị dự báo của các biến ngoại sinh này. Và ứng với mỗi bộ giá trị dự báo của các biến ngoại sinh ta tính toán được tương ứng một giá trị của chỉ số chứng khoán HNX. Nói cách khác giá trị của chỉ số chứng khoán Việt Nam được dự báo theo các kịch bản khác nhau về giá cổ phiếu của hai công ty ACB và PCI.

4.3. Ứng dụng luật kết hợp mẫu âm và mô hình hồi quy chuyển tiếp trơn trong xây dựng mô hình dự báo chỉ số giá tiêu dùng (CPI)

Năm 2008 là năm lạm phát ở Việt Nam tăng cao kỷ lục, giá cả hàng hóa biến động bất thường. Dù tỷ lệ lạm phát đã được kiềm chế trong năm 2009 nhưng lại tăng lên trong 2010 và đặc biệt tăng cao trong nửa đầu năm 2011. Để ổn định kinh tế vĩ mô, việc bình ổn giá và kiềm chế lạm phát càng trở nên cần thiết và cấp bách.

Chỉ số đo lường lạm phát của một nền kinh tế thường là chỉ số giảm phát GDP dựa trên tính toán của tổng sản phẩm quốc nội, nhưng cũng có thể là chỉ số giá hàng hóa (CPI), hay chỉ số giá hàng sản xuất (PPI), hay chỉ số giá hàng bán (WPI),... Mô hình dự báo lạm phát của các chỉ số này trong các quốc gia khác nhau là rất khác nhau ngay cả khi chúng cùng được xây dựng bởi một phương pháp.

Các nước có nền kinh tế phát triển sử dụng mô hình chuyển tiếp trơn để xây dựng mô hình dự báo lạm phát cho chỉ số CPI, ở đây chỉ số CPI được xem như một chỉ số kinh tế có mối quan hệ với các chỉ số khác như tỷ lệ tăng GDP, tỷ lệ thất nghiệp, tỷ giá quy đổi tiền tệ, chỉ số giá xuất nhập khẩu,.... Phân tích chuyển tiếp trơn được sử dụng để xác định đường chuyển tiếp trong xu hướng của chuỗi giá, đặc biệt là tốc độ chuyển tiếp và điểm trung vị của quá trình động này, giữa hai chế độ chính sách tiền tệ.

Ở Việt Nam chỉ số lạm phát được tính dựa trên chỉ số giá tiêu dùng CPI. Vì vậy mô hình dự báo chỉ số CPI cũng sẽ là mô hình dự báo lạm phát. Biến động giá cả của các sản phẩm hàng hoá và dịch vụ là những hiện tượng kinh tế - xã hội khá phổ biến. Người ta đã nghiên cứu và xác định được các yếu tố kinh tế, xã hội chủ yếu có tác động đến việc hình thành và biến động giá cả. Tuy nhiên những câu hỏi như "sự biến động giá của nhóm mặt hàng này có tác động, ảnh hưởng thế nào đến sự biến động giá của nhóm mặt hàng khác", "sự biến động giá cả của những mặt hàng nào ảnh hưởng nhiều nhất đến chỉ số CPI" vẫn là *những câu hỏi mở*, chưa được chú trọng và kết quả nghiên cứu còn hết sức hạn chế. Mục đích nghiên cứu phần này của chúng tôi nhằm góp phần trả lời cho câu hỏi trên. Cụ thể 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, tiếp sau đó sẽ ứng dụng mô hình hồi quy chuyển tiếp trơn phân tích mối quan hệ về biến động giá cả của một số sản phẩm hàng hóa với chỉ số CPI và xây dựng mô hình dự báo chỉ số CPI.

Quy trình xây dựng mô hình dự báo chỉ số CPI cũng được thực hiện qua 2 giai đoạn. 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.

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. Giá cả các sản phẩm xuất, nhập khẩu chủ yếu được thu thập từ Tổng cục Hải quan và tính trung bình theo tuần, trong khi giá cả của các sản phẩm thiết yếu của đời sống dân sinh được thu thập từ 3/1/2008 đến hết ngày 31/12/2009 ở địa bàn Hà Nội vào thứ hai, thứ tư, thứ sáu và giá trung bình của 3 ngày này được lấy làm giá cả của sản phẩm đó trong tuần.

Khi phân tích dữ liệu thu thập chúng tôi nhận thấy biên độ giao động của giá cả một số mặt hàng rất nhỏ hoặc thay đổi vài tháng một lần (bao gồm 14 mặt hàng Chính phủ thực hiện bình ổn giá). Chúng tôi đã loại bỏ những mặt hàng này ra khỏi phạm vi nghiên cứu. Cuối cùng dữ liệu thu thập được đưa vào nghiên cứu giá của các mặt hàng còn lại trong 103 tuần.

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 theo cách CPI trung bình cả 4 tuần là CPI của tháng và theo các trường hợp sau:

- Khi CPI của tháng này tăng hơn so với tháng trước và thấp hơn so với tháng sau đó thì CPI của 4 tuần được chọn theo thứ tự tăng dần tuy nó khác nhau rất nhỏ và vẫn đảm bảo CPI trung bình của 4 tuần bằng CPI của tháng.

- Nếu CPI của tháng đó giảm so với tháng trước và tháng sau lại giảm hơn so với tháng đó thì CPI của 4 tuần trong tháng được chọn theo hướng giảm dần.

- Nếu CPI của tháng đó giảm (cao) so với tháng trước và giảm (cao) hơn so với tháng sau thì CPI của 4 tuần trong tháng được chọn sao cho 2 tuần đầu giảm (tăng) dần và 2 tuần sau tăng (giảm) dần.

Với mỗi mặt hàng chúng tôi đều gắn mã để tiện cho việc nghiên cứu. Kết quả thu được tập dữ liệu về giá của 121 mặt hàng (CPI cũng được xem như là một mặt hàng). Trong đó có 13 mặt hàng xuất khẩu (có mã từ XA1 đến XA9, và XB1 đến XB4); 16 mặt hàng nhập khẩu (có mã từ NA1 đến NA9 và NB1 đến NB7); 80 mặt hàng thiết yếu của đời sống dân sinh (có mã từ DA1 đến DA9, từ DB1 đến BD9,..., từ DK1 đến DK9); và chỉ số CPI.

Các luật kết hợp được phát hiện trong nghiên cứu này cũng là luật kết hợp nhị phân biểu diễn mối quan hệ về việc tăng, giảm giá so với tuần ngay trước đó của nhóm mặt hàng này với nhóm mặt hàng khác. Để làm được điều đó cần có CSDL tác vụ. CSDL tác vụ được tạo ra từ tập dữ liệu gốc bằng cách: Nếu giá một hàng hoá của tuần này cao hơn tuần trước đó (giá tăng) thì thêm chữ số ”1” vào bên phải của mã giá hàng hoá và thêm chữ số ”2 ” nếu giá tuần này thấp hơn (giá giảm).

Theo quy tắc này, thì tập dữ liệu về giá của các mặt hàng có thể được biểu diễn như ở hình 4.4 và được gọi là CSDL tác vụ với mục dữ liệu âm về biến động giá.

```
XA52, XA61, XA71, XA82, XA91, XB21, XB31, XB42, CPI2
WA32, WA41, WA61, WA72, WA82, WA91, WB12, WB21, CPI1
NB62, XA12, XA21, XA42, XA52, XA61, XA71, XA81, XA92, XB21, XB31, XB41, CPI1
NB32, NB42, NB51, NB71, XA32, XA41, XA52, XA61, XA71, XA81, XA91, XB11, XB22, XB32,
NB51, XA12, XA21, XA32, XA42, XA51, XA62, XA71, XA81, XA92, XB22, XB31, XB41, CPI1
XA42, XB32, XB41, CPI1
```

Hình 4.4: CSDL về giá của các mặt hàng

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. Đó là các luật:

Rule 92: XB41; XA81; NA31; NB12 → CPI1 (11,765% 91,67% 12 11 10,784%)

Rule 93: XB41; XA81; NB12 → CPI1 (13,725% 92,86% 14 13 12,745%)

Rule 102: XA92; XA71; NB62 → CPI1 (11,765% 91,67% 12 11 10,784%)

Rule 118: DB12; XA21; XA32 → CPI2 (11,765% 91,67% 12 11 10,784%)

Rule 124: XA62; XA82; XA52 → CPI2 (11,765% 91,67% 12 11 10,784%)

Rule 165: XA92; XA81; XA21; XA71 → CPI1 (12,745% 92,31% 13 12 11,765%)

Rule 169: NB31; XA21; XA71 → CPI1 (13,725% 92,86% 14 13 12,745%)

Rule 174: XA62; XA91 → CPI2 (11,765% 91,67% 12 11 10,784%)

Rule 181: XA92; XA81; XA21; XB21 → CPI1 (11,765% 91,67% 12 11 10,784%)

Rule 195: NB31; XA51; XA11 → CPI1 (11,765% 91,67% 12 11 10,784%)

Rule 203: DK61; XA41; NB21 → CPI1 (11,765% 91,67% 12 11 10,784%)

Rule 205: XB41; XA81; XA21 → CPI1 (12,745% 92,31% 13 12 11,765%).

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.

Chúng ta có thể phát hiện dấu hiệu thay đổi của chỉ số CPI từ dấu hiệu thay đổi về giá của các mặt hàng trong nhiều nhóm gồm các mặt hàng nhập, xuất khẩu hay các mặt hàng dân sinh. Có nhóm thì các mặt hàng thay đổi theo chiều hướng tăng nhưng ở nhóm khác lại thay đổi theo chiều hướng giảm.

4.3.3. Xây dựng mô hình dự báo chỉ số CPI

4.3.3.1. 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 các 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 $\rightarrow$ 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 CSDL về chỉ số CPI và giá của NB1, XA8, XB4 được chia thành 2 phần. Phần 1 bao gồm 94 tuần của năm 2008 và 2009 được dùng để xây dựng mô hình dự báo chỉ số CPI. Phần thứ 2 gồm 9 tuần của tháng 11 và tháng 12 năm 2009 được dùng để kiểm định mô hình.

Giai đoạn 1: Áp dụng phần mềm JMULTI [99] với phần CSDL thứ nhất để thực hiện kiểm định chuỗi thời gian với CPI, XA8, XB4 và NB1. Chúng tôi thấy rằng CPI, XA8 và NB1 là chuỗi không dừng nhưng XB4 và các chuỗi sai phân bậc 1 của các chuỗi đó là dừng. Vì vậy, chúng ta xây dựng mô hình dự báo cho chuỗi sai phân bậc 1 của CPI (kí hiệu là CPI_d1) từ các chuỗi sai phân bậc 1 của XA8, XB4 và NB1 (kí hiệu tương ứng là XA8_d1, XB4_d1, NB1_d1). Kết quả kiểm định tính chất tuyến tính cho CPI_d1 chỉ ra rằng mô hình là LSTR1, biến chuyển tiếp trơn là CPI_d1(t-3) và giá trị lớn nhất của biến phụ thuộc CPI_d1 và các biến độc lập XA8_d1, XB4_d1, NB1_d1 là cùng bằng 4.

variable	start	estimate	SD	t-stat	p-value
----- linear part ------					
CONST	-13.86256	-5.99704	3.2616	-1.8387	0.0698
CPI_dl(t-1)	-5.78085	-7.09577	4.1723	-1.7007	0.0930
CPI_dl(t-2)	5.48318	7.34688	4.0032	1.8353	0.0703
CPI_dl(t-3)	-10.31479	-6.26734	3.2103	-1.9522	0.0545
NBl_dl(t-4)	-0.01966	-0.01908	0.0093	-2.0548	0.0433
---- nonlinear part ----					
CONST	14.35961	6.04024	3.2552	1.8555	0.0673
CPI_dl(t-1)	6.29465	7.45941	4.1772	1.7858	0.0781
CPI_dl(t-2)	-5.39052	-7.13244	4.0042	-1.7812	0.0788
CPI_dl(t-3)	9.15263	5.58218	3.2195	1.7338	0.0869
NBl_dl(t-4)	0.01947	0.01840	0.0093	1.9862	0.0506
Gamma	0.92928	2.85916	0.0000	0.0000	0.0009
C1	-1.34000	-0.80295	0.0000	-0.0000	0.0000

Hình 4.5: Ước lượng các tham số của mô hình dự báo CPI

Giai đoạn 2: Kết quả ước lượng các tham số thể hiện trong hình 4.5. Từ kết quả này rút ra được các nhận xét sau:

- Tất cả các giá trị p-value của các biến độc lập đều nhỏ hơn 1. Điều đó có nghĩa tất cả các biến tuyến tính và phi tuyến của mô hình có ý nghĩa ở mức trên 90%.
- Các biến XA8_d1(t), XB4_d1(t), XA8_d1(t-1), XA8_d1(t-2), XA8_d1(t-3), XA8_d1(t-4),…không ảnh hưởng đến sự thay đổi của CPI_d1(t).
- Các biến NB1_d1(t-4), CPI_d1(t-1), CPI_d1(t-2), CPI_d1(t-3) ảnh hưởng trực tiếp đến CPI_d1(t).
- Hệ số xác định R2 = 4,9696e-01 và hệ số điều chỉnh R2 = 0,5026 cho thấy các biến độc lập giải thích 50% sự thay đổi của biến phụ thuộc CPI_d1(t).

Từ hình 4.5 rút ra được mô hình dự báo chỉ số CPI_d1 như sau:

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

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

Phần tuyến tính của mô hình cho biết chỉ số CPI_d1(t) thay đổi cùng chiều với CPI_d1(t-2) nhưng thay đổi ngược chiều với CPI_d1(t-1), CPI_d1(t-3), CPI_d1(t-4) và NB1_d1(t-4).

Phần phi tuyến gồm hai phần. Phần thứ nhất là thành phần tự hồi quy. Phần này giống với phần tuyến tính nhưng dấu của các hệ số độc lập là ngược lại. Phần thứ hai là hàm logicstic với hàm chuyển tiếp trơn là sai phân bậc 1 của PCI_d1(t-3) với tham số vị trí là -0,803 và tham số độ dốc là 2,86. Thành phần thứ hai cho biết sự biến động của giá trị sai phân bậc 1 chỉ số CPI trong miền tăng trưởng khác với tính chất biến động của nó trong miền suy thoái và việc chuyển tiếp từ thái cực này sang thái cực kia là trơn.

Giai đoạn 3: Thực hiện kiểm định mô hình. Các kiểm định cho thấy mô hình dự báo chỉ số CPI không có tự tương quan phần dư, không có thành phần tuyến tính bị bỏ sót và không có sự thay đổi của các tham số.

4.3.3.2. Đánh giá 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 4.2 thể hiện kết quả chỉ số CPI được tính theo mô hình đã xây dựng và chỉ số CPI theo thống kê.

Kết quả bảng 4.2 cho thấy tỷ lệ % sai lệch cho cả trường hợp theo tuần và theo tháng là rất nhỏ. Như vậy mô hình xây dựng có thể dùng để dự báo chỉ số CPI của Việt Nam.

Trong mô hình dự báo ở trên, tất cả các biến độc lập đều là trễ của CPI-d1 và NB1-d1. Như vậy để dự báo chỉ số CPI không cần phải dự báo các biến độc lập khác trong mô hình. Để dự báo chỉ số CPI(t) chỉ cần tính CPI_d1(t) từ các giá trị CPI_d1(t-1), CPI_d1(t-2), CPI_d1(t-3), CPI_d1(t-4) và NB1_d1(t-4).

Bảng 4.2: 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%			

Kết luận chương 4 :

Chương 4 đã trình bày kết quả ứng dụng luật kết hợp và mô hình hồi quy chuyển tiếp trơn phi tuyến trong xây dựng mô hình phân tích và dự báo chỉ số chứng khoán và chỉ số giá tiêu dùng Việt Nam.

Mô hình dự báo chỉ số chứng khoán là **mô hình dự báo có điều kiện**, cụ thể việc dự báo chỉ số này một mặt phụ thuộc vào mô hình dự báo được xây dựng và mặt khác phụ thuộc vào dự báo hai biến độc lập khác trong mô hình là ACB và PVI. Dự báo có điều kiện là phương pháp dự báo có sự kết hợp giữa phương pháp định lượng với phương pháp định tính, nó được sử dụng để dự báo một sự kiện mà trong tương lai có thể nó phải chịu những tác động khó lường của nhiều yếu tố khác. Mô hình dự báo có điều kiện chỉ số chứng khoán HNX cho thấy có thể quy việc dự báo chỉ số này về việc dự báo giá của một vài cổ phiếu khác bằng phương pháp định lượng và định tính.

Do các biến độc lập trong mô hình dự báo CPI đều là các biến trễ của biến giá một số mặt hàng khác nên mô hình dự báo CPI là **mô hình dự báo không điều kiện**, tức là có thể dự báo được CPI theo phương pháp kinh tế lượng mà không cần bất kỳ điều kiện nào khác.

Kiểm định kết quả dự báo theo mô hình so với thực tế của cả hai mô hình trên cho thấy sai số dự báo đều khá nhỏ, nói cách khác độ chính xác của dự báo là khá cao điều đó cho thấy triển vọng của cách tiếp cận kết hợp luật kết hợp trong công

nghệ thông tin và mô hình hồi quy chuyển tiếp trơn trong kinh tế trong việc xây dựng mô hình phân tích và dự báo của nhiều hiện tượng kinh tế - xã hội.

Về nguyên tắc ứng với mỗi luật kết hợp ta xây dựng được một mô hình phân tích và dự báo dựa trên mô hình LSTR. Như thế có nghĩa là ta có thể xây dựng được nhiều mô hình dự báo khác nhau về các chỉ số HNX và CPI theo cùng một cách tiếp cận. Vấn đề đặt ra khi đó cần chọn mô hình dự báo nào được sử dụng chính thức. Để trả lời câu hỏi này ta có thể ứng dụng kỹ thuật kiểm định bao và kết hợp dự báo. Trả lời câu hỏi này cần tiến hành hai nội dung sau:

Thứ nhất: sử dụng phương pháp kiểm định bao dự báo để xác định xem dự báo này có bao quát được tất cả các thông tin hữu ích của một sự báo khác hay không?

Nếu một dự báo bị một dự báo khác bao thì ta sẽ loại bỏ dự báo bị bao đó ra khỏi phạm vi xem xét. Nếu không có dự báo nào bị bao bởi dự báo kia thì cả hai mô hình đều có chứa những thông tin bổ sung thêm và ta nên giữ lại cả hai mô hình dự báo này để phục vụ cho việc xây dựng dự báo kết hợp, nhằm có thể khai thác những thông tin hữu ích của cả hai dự báo đó. Quá trình trên được thực hiện đối với mọi cặp dự báo. Nếu như tất cả các dự báo bị bao được loại bỏ thì dự báo kết hợp sẽ được xây dựng theo một cách nào đó cho tất cả các dự báo được giữ lại.

Thứ hai: tiến hành kết hợp nhiều kết quả dự báo thành một kết quả dự báo mới có độ chính xác cao hơn so với mỗi kết quả dự báo thành phần.

Kết hợp dự báo là việc kết hợp hai hoặc nhiều hơn các mô hình dự báo về một hiện tượng kinh tế - xã hội nào đó thành một mô hình dự báo. Điều đó có nghĩa là nó cho phép kết hợp nhiều kết quả dự báo cá biệt thành một kết quả dự báo duy nhất (gọi là dự báo kết hợp). Người ta đã chỉ ra rằng độ chính xác so với thực tiễn của dự báo kết hợp là cao hơn so với mỗi dự báo thành phần.

Kiểm định bao và kết hợp dự báo hiện đang được nhiều nhà nghiên cứu kinh tế hàng đầu thế giới quan tâm và có rất nhiều triển vọng trở thành một phương pháp dự báo mới. Trong luận án này chúng tôi không trình bầy các kỹ thuật này.

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, và đặc biệt nghiên cứu sâu hơn về phát hiện luật kết hợp hiếm. Từ việc phân tích kết quả đạt được cũng như hạn chế của các nghiên cứu trước về luật kết hợp hiếm, luận án đã 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. Cụ thể như sau:

- Đề xuất mở rộng bài toán phát hiện luật kết hợp Sporadic tuyệt đối hai ngưỡng và luật kết hợp Sporadic không tuyệt đối hai ngưỡng. Hai thuật toán lần lượt được giới thiệu là MCPSI và MCISI tương ứng nhằm tìm các tập phổ biến cho các 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.

Đóng góp của chúng tôi là 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.

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

- Đề xuất bài toán phát hiện 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 phổ biến 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 phổ biến cho các luật này.

Đóng góp của chúng tôi là phát triển khuynh hướng áp 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 phổ biến mờ cho luật kết hợp hiếm.

3. Góp phần nghiên cứu ứng dụng luật kết hợp trong phân tích và dự báo kinh tế, chúng tôi 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. Kết quả dự báo kiểm định các mô hình dự báo đó cho thấy độ chính xác của kết quả dự báo là khá sát với số liệu 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 trong lĩnh vực giá cả, lạm phát.

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 dữ liệu 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 do chúng tôi đề xuất 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 tiếp theo của chúng tôi là phải sinh được các luật hiếm có giá trị từ các tập phổ biến tìm được. Đây cũng là hướng nghiên cứu hay và không dễ vì các luật kết hợp hiếm có những tính chất riêng.

Áp dụng hướng phát hiện song song luật hiếm như cách tiếp cận khai phá song song luật kết hợp như trong [15, 28, 43, 67, 97].

Tiếp tục triển khai ứng dụng luật kết hợp với các phương pháp khác để phân tích và dự báo dữ liệu kinh tế.

DANH MỤC CÔNG TRÌNH KHOA HỌC 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.

TÀI LIỆU THAM KHẢO

Tiếng Việt

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. 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.
4. 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.
5. Đỗ Văn Thành (2004), "Phát hiện các luật kết hợp có độ hỗ trợ cực tiểu không giống nhau", *Khoa học và Công nghệ,* tập 42 (1), tr. 79-90.
6. Đỗ Văn Thành (2007), "Giải pháp dự báo ngắn hạn về tăng trưởng kinh tế Việt Nam", *Tạp chí Tin học và Điều khiển học*, tập 23 (4), tr. 374-386.
7. 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.
8. Lê Thị Mai Linh (2003), *Phân tích và đầu tư chứng khoán*, Nhà xuất bản Chính trị Quốc gia, Hà Nội.
9. Nguyễn Đình Thuân (2005), *Một số vấn đề về phụ thuộc dữ liệu và luật kết hợp trong cơ sở dữ liệu có yếu tố thời gian*, Luận án Tiến sĩ, Viện Công nghệ thông tin, Hà Nội.
10. Nguyễn Hữu Trọng (2008), *Phát triển một số thuật toán khai thác luật kết hợp trên cơ sở dữ liệu gia tăng*, Luận án Tiến sĩ, Viện Công nghệ thông tin, Hà Nội.

11. Phạm Thị Thắng (2010), *Kinh tế lượng trong lĩnh vực Tài chính ngân hàng*, Nhà xuất bản Tài chính, Hà Nội.
12. Võ Đình Bảy (2011), *Nâng cao hiệu quả của các thuật toán khai thác luật kết hợp dựa trên dàn*, Luận án Tiến sĩ, Đại học Khoa học Tự nhiên (Đại học Quốc gia Thành phố Hồ Chí Minh), TP Hồ Chí Minh.

Tiếng Anh

13. Agrawal R., Imielinski T., and Swami A. (1993), "Mining Association Rules between Sets of Items in Large Databases", *Proc. of ACM SIGMOD Conf. Management of Data*, pp. 207-216.
14. Agrawal R., Mannila H., Srikant R., Toivonen H., and Inkeri Verkamo A. (1996), "Fast Discovery of Association Rules", *Advances in Knowledge discovery and Data Mining*, pp. 307-328.
15. Agrawal R., and Shafer J. (1996), "Parallel Mining of Association Rules", *IEEE Transactions in Knowledge and Data Engineering,* Vol. 8 (6), pp. 962-969.
16. Agrawal R., and Srikant R. (1994), "Fast Algorithms for Mining Association Rules", *Proc. of the Very Large Database International Conference*, Santiago, pp. 487-498.
17. Antonic M. L., Zaiane O. R. (2004), "Mining Positive and Negative Rules: An Approach for Confined Rules", *Proc. of the Intl. Conf on Principles and Practice of Knowledge Discovery in Database*, Italy, pp. 27-38.
18. Antonie M. L., and Zaıane O. R. (2004), "An Associative Classifier based on Positive and Negative Rules", *Proc. of DMKD'04,* Paris, France, pp. 64-69.
19. Bacon D. W., and Watts D. G. (1971), "Estimating the Transition between Two Intersecting Straight Lines", *Biometrika,* Vol. 58 (3), pp. 525-534.
20. Bal J., Balcázar L. (2009), "Confidence Width: An Objective Measure for Association Rule Novelty", *Proc. of QIMIE'09/ PAKDD'09*, pp. 5-16.
21. Bayardo R. J. (1998), "Efficiently Mining Long Patterns From Databases", *Proc. of SIGMOD'98*, Seattle, Washington, pp. 85-93.
22. Bayardo R. J., Agrawal R., and Gunopulos D. (1999), "Constraint-based Rule Mining in Large, Dense Databases", *Proc. of ICDE.1999,* pp. 188-197.

23. Besemann C., Denton A., and Yekkirala A., “Differential Association Rule Mining for the Study of Protein-Protein Interaction Networks”, *Proc. of BIOKDD04: 4th Workshop on Data Mining in Bioinformatics*, pp. 72-81.
24. Bonchi F., Lucchese C. (2004), “On Closed Constrained Frequent Pattern Mining”, *In ICDM IEEE Computer Society*, pp. 35-42.
25. Brijs T., Swinnen G., Vanhoof K., and Wets, G. (1999), “The Use of Association Rules for Product Assortment Decisions: A Case Study”, *In Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, pp. 254-260.
26. Bucila C., Gehrke J. E., Kifer D., and White. W. (2003), “Dualminer: A Dual-pruning Algorithm for Itemsets with Constraints”, *Data Mining and Knowledge Discovery,* Vol. 7 (3), pp. 241-272.
27. Burdick D., Calimlim M., and Gehrke J. (2001), “Mafia: A Maximal Frequent Itemset Algorithm for Transactional Databases”, *Proceedings 17th International Conference on Data Engineering*, pp. 443-452.
28. Cheung D. W., and Xiao Y. (1999), “Effect of Data Distribution in Parallel Mining of Associations”, *Data Mining and Knowledge Discovery,* Vol. 3 (3), pp. 291-314.
29. Chunjiang Z. , Huarui W. , Xiang S., and Baozhu Y. (2007), “Algorithm for Mining Association Rules with Multiple Minimum Supports based on FP-Tree”, *New Zealand Journal of Agricultural Research*, Vol. 50, pp. 1375-1381.
30. Cohen E., Datar M., Fujiwara S., Gionis A., Indyk P., Motwani R., Ullman J.D., Yang C. (2000), “Finding Interesting Association Rules Without Support Pruning”, *Proc. of 16th International Conference on Data Engineering (ICDE'00)*, pp. 64-78.
31. Cornelis C., Yan P., Kang X., Chen G. (2006), “Mining Positive and Negative Association Rules from Large Databases”, *IEEE Computer Society*, pp. 613-618.
32. Cu Thu Thuy, Do Van Thanh (2010), “Mining Perfectly Sporadic Rules with Two Thresholds”, *In Proceedings of MASS 2010*, Wuhan, China.

33. 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.
34. Delgado M., Marín N., Sánchez D., and Vila M. A. (2003), "Fuzzy Association Rules: General Model and Applications", *IEEE Transactions on Fuzzy Systems*, Vol. 11 (2), pp. 214-225.
35. Diebold F. X. (2007), *Elements of Forecasting*, Fourth Edition. Thomson: South-Western.
36. 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 (3), pp. 16-27.
37. Gouda K., and Zaki M.J. (2005), "GenMax: An Efficient Algorithm for Mining Maximal Frequent Itemsets", *Data Mining and Knowledge Discovery*, Vol. 11 (3), pp. 1-20.
38. Gupta M., and Joshi R. C. (2009), "Privacy Preserving Fuzzy Association Rules Hiding in Quantitative Data", *International Journal of Computer Theory and Engineering*, Vol. 1 (4), pp. 1793-8201.
39. Gyenesei A. (2000), "A Fuzzy Approach for Mining Quantitative Association Rules", *Turku Centre for Computer Science,* TUCS Technical Reports, No336.
40. Gyenesei A. (2000), "Mining Weighted Association Rules for Fuzzy Quantitative Items", *Proc. of PKDD Conference*, pp. 416-423.
41. Gyenesei A., and Teuhola J. (2004), "Multidimensional Fuzzy Partitioning of Attribute Ranges for Mining Quantitative Data", *International Journal of Intelligent System*, Vol. 19 (11*)*, pp. 1111-1126.
42. Han J., Pei J., Yin J., and Mao R. (2004), "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach", *Data Mining and Knowledge Discovery*, Vol. 8, pp. 53-87.

43. Han E-H., Karypis G., and Kumar V. (1997), “Scalable Parallel Data Mining for Association Rules”, *IEEE trasaction on Knowledge and Data Engineering*, Vol. 12 (3), pp. 337-352.
44. He Y., Tang Y., Zhang Y., and Sunderraman R. (2006), “Adaptive Fuzzy Association Rule Mining for Effective Decision Support in Biomedical Applications”, *Int. J. Data Mining and Bioinformatics*, Vol. 1 (1), pp. 3-18.
45. Hong T.P., Lin K.Y., and Wang S.L. (2003), “Fuzzy Data Mining for Interesting Generalized Association Rules”, *Fuzzy Sets and Systems,* Vol. 138 (2), pp. 255-269.
46. Kiran R. U., and Reddy P. K. (2009), “An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules”, *Proc. of CIDM 2009*, pp. 340-347.
47. Kiran R. U. and Reddy P. K. (2010), “Mining Rare Association Rules in the Datasets with Widely Varying Items’ Frequencies”, *Proc. of the 15th International Conference on Database Systems for Advanced Applications Tsukuba*, Japan, pp. 49-62.
48. Kock A. B. and Teräsvirta T. (2010), "Forecasting with Nonlinear Time Series Models", CREATES Research Papers 2010-01, School of Economics and Management, University of Aarhus.
49. Koh Y. S., Rountree N. (2005), “Finding Sporadic Rules Using Apriori-Inverse”, *Proc. of PAKDD2005*, pp. 97-106.
50. Koh Y. S., Rountree N., O’Keefe R. A. (2008), “Mining Interesting Imperfectly Sporadic Rules”, *Knowledge and Information System*, Vol. 14 (2), pp. 179-196.
51. Koh Y. S. and Rountree N. (2010), *Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection,* Information Science Reference (Imprint of: IGI Publishing), America, pp. 1-14.
52. Kryszkiewicz M. (2005),” Generalized Disjunction-Free Representation of Frequent Patterns with Negation”, *Journal of Experimental & Theoretical Artificial Intelligence*, Vol. 17 (1-2), pp. 63-82.

53. Kubat M., Holte R. C., and Matwin S.(1998), “Machine Learning for The Detection of Oil Spills in Satellite Radar Images”, *Journal of Machine Learning* Vol. 30 (2-3), pp. 195-215.
54. Kuok C. M., Fu A., and Wong M. H. (1998), “Mining Fuzzy Association Rules in Databases”, ACM SIGMOD Record, Vol. 27 (1), pp. 41-46.
55. Latiri C. C., Elloumi S., Chevallety J.P., and Jaouay A. (2003), “Extension of Fuzzy Galois Connection for Information Retrieval Using a Fuzzy Quantifier”, *IEEE International Conference on Computer Systems and Applications*, pp.84.
56. Li J., Zhang X., Dong G., Ramamohanarao K., and Sun Q. (1999), “Efficient Mining of High Confidence Association Rules without Support Threshold”, P*roc. of the 3rd European Conference on Principle and Practice of Knowledge Discovery in Databases*, pp. 406 - 411.
57. Lin N.P., and Chueh. (2007), “Fuzzy Correlation Rules Mining”, *Proc. of the 6th WSEAS International Conference on Applied Computer Science*, pp.13-18.
58. Ling Zhou, and Stephen Yau (2007), “Association Rule and Quantitative Association Rule Mining among Infrequent Items”, *Proc. of the 8th international workshop on Multimedia data mining*, New York, USA.
59. Liu B., Hsu W., and Ma Y. (1999), “Mining Association Rules with Multiple Minimum Supports”, *Proc. of KDD 1999*, pp. 337-341.
60. Maddala D. S. (1977), *Econometrics*, McGraw-Hill, New York, USA.
61. Muyeba M., Khan M. S., and Coenen F. (2008),”Fuzzy Weighted Association Rule Mining with Weighted Support and Confidence Framework”, *In PAKDD Workshop 2008*, pp. 49-61.
62. Nguyen Khac Minh (2009), *Theoretical Foundation of Nonlinearn Time Series and Application for Building Inflation Models of Viet Nam*, In Time Series models and application for analyzing inflation, Lectute Document of EU Technical Assistant Program for Viet Nam, Hà Nội, Việt Nam.
63. Olson D. L., and Li Y. (2007), “Mining Fuzzy Weighted Association Rules”, *Proc. of the 40th Hawaii International Conference on System Sciences*, Hawaii, USA.

64. Pasquier N., Bastide Y., Taouil R., Lakhal L. (1999), “Efficient Mining of Association Rules Using Closed Itemset Latics”, *Journal Information Systems*, Vol. 24 (1), pp.25-46.
65. Pei J., Han J., and Mao R. (2000), "CLOSET: An Efficient Algorithm for Mining Frequent Closed Itemsets", *Proc. of Workshop on Research Issues in Data Mining and Knowledge Discovery*, pp. 21-30.
66. Rahal I., Ren D., Wu W., and Perrizo, W. (2004), “Mining Confident Minimal Rules with Fixed Consequents”, *Proc. of the 16th IEEE International Conference on Tools with Artificial Intelligence*, pp. 6 -13.
67. Rahman A.M., and Balasubramanie P. (2009), “Weighted Support Association Rule Mining using Closed Itemset Lattices in Parallel”, *International Journal of Computer Science and Network Security*, Vol. 9 (3), pp. 247-253.
68. Romero C., Romero J. R., Luna J. M., and Ventura S. (2010), “Mining Rare Association Rules from e-Learning Data”, *Proc. of the Third International Conference on Education Data Mining*, pp. 171-180.
69. Romero C., Ventura S., Vasilyeva E., and Pechenizkiy M. (2010), “Class Association Rule Mining from Students’ Test Data”, *Proc. of the Third International Conference on Education Data Mining*, pp. 137-138.
70. Savasere A., Omiecinski E., and Navathe S. (1995), An Efficient Algorithm for Mining Association Rules in Large Databases, *Proc. of the 21st International Conference on Very Large Data Bases*, pp. 432-444.
71. Savasere A., Omiecinski E., and Navathe S. (1998), ”Mining for Strong Negative Associations in a Large Database of Customer Transactions”, *Proc. of Intl. Conf. on Data Engineering*, pp. 494-502.
72. Seno M., and Karypis G. (2001), “LPMINER: An Algorithm for Finding Frequent Itemsets Using Length-decreasing Support Constraint”, *Proc. of the 2001 IEEE International Conference on Data Mining ICDM*, pp. 505-512.
73. Srikant R., and Agrawal R. (1996), ” Mining Quantitative Association Rules in Large Relational Table”, *Proc. of ACM SIGMOD Conference on Management of Data* , pp. 1-12.

74. Srikant R., Vu Q., and Agrawal R. (1997), "Mining Association Rules with Item Constraints", *Proc. of the Third International Conference on Knowledge Discovery and Data Mining (KDD'97)*, pp. 67-73.
75. Szathmary L., Napoli A., Valtchev P. (2007), "Towards Rare Itemset Mining", *Proc. of the 19th IEEE International Conference on Tools with Artificial Intelligence*, pp. 305-312.
76. Szathmary L., Valtchev P., and Napoli A. (2010), "Generating Rare Association Rules Using Minimal Rare Itemsets Family", *International Journal of Software and Informatics*, Vol. 4 (3), pp. 219-238.
77. Tao F., Murtagh F., Farid M. (2003), "Weighted Association Rule Mining Using Weighted Support and Significance Framework", *Proc. of KDD 2003*, pp. 661-666.
78. Teräsvirta T. (1996), *Modelling Economic Relationships with Smooth Transition Regressions*, Working Paper Series in Economics and Finance 131, Stockholm School of Economics.
79. Teräsvirta T. (2005), *Forecasting Economic Variables with Nonlinear Models*, Working Paper Series in Economics and Finance 598, Stockholm School of Economics 2005.
80. Troiano L., Scibelli G., Birtolo C. (2009), "A Fast Algorithm for Mining Rare Itemsets", *Proc. of ISDA 2009*, pp.1149-1155.
81. Tseng S. V. (1998), "An Efficient Method for Mining Association Rules with Item Constraints", *Discovery Science - First International Conference*, pp. 423-424.
82. Tseng V. S., Chen Y., Chen C. H., and Shin J. W. (2006), "Mining Fuzzy Association Patterns in Gene Expression Databases", *International Journal of Fuzzy Systems*, Vol. 8 (2), pp. 87-93.
83. Wang K., He Y., and Cheung D. W. (2001), "Mining Confident Rules without Support Requirement", *Proc. of the Tenth International Conference on Information and Knowledge Management*, pp. 89-96.

84. Wang K., He Y., and Han, J. (2003), "Pushing Support Constraints into Association Rules Mining", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15(3), pp. 642-658.
85. Weiss G. M., and Hirsh H. (1998), "Learning to Predict Rare Events in Event Sequences", *Proc. of the Fourth International Conference on Knowledge Discovery and Data Mining*, pp. 359-363.
86. Wong P. C., Whitney P., and Thomas J. (1999), *"Visualizing Association Rules for Text Mining", Proc. of INFOVIS1999*, pp. 120-123.
87. Wong C., Shiu S., and Pal S. (2001), "Mining Fuzzy Association Rules for Web Access Case Adaptation", *Proc. of Soft Computing in Case-Based Reasoning Workshop*, *in conjunction with the 4th International Conference in Case-Based Reasoning*, pp. 213-220.
88. Wu X., Kumar V., Quinlan J. R., Ghosh J., Yang Q., Motoda H., Geoffrey J. McLachlan, Angus Ng, Liu B., Yu P. S., Zhou Z. H., Steinbach M., Hand D. J., Steinberg D. (2007), "Top 10 Algorithms in Data Mining", *Knowledge and Information Systems*, Vol. 14 (1), pp. 1-37.
89. Wu X., Zhang C., and Zhang S. (2004), "Efficient Mining of Both Positive and Negative Association Rules", *ACM Transactions on Information Systems*, Vol. 22(3), pp. 381-405.
90. Xiong H., Tan P., and Kumar V. (2003), "Mining Strong Affinity Association Patterns in Data Sets with Skewed Support Distribution", *Proc. of the Third IEEE International Conference on Data Mining*, pp. 387-394.
91. Yan P., Chen G., Cornelis C., Cock M. D. and Kerre E.E. (2004), "Mining Positive and Negative Fuzzy Association Rules", *Proc. of KSE2004*, pp. 270-276.
92. Yuan X., Buckles B.P., Yuan Z. and Zhang J.(2002), "Mining Negative Association Rules", *Proc. of Seventh Intl. Symposium on Computers and Communication,* pp. 623-629.
93. Yun H., Ha D., Hwang B., Ryu K. H. (2003), "Mining Association Rules on Significant Rare Data Using Relative Support", *The Journal of Systems and Software* 67 (2003), pp. 181-191.

94. Zaki M. J., Hsiao C. (1999), *CHARM: An Efficient Algorithm for Closed Association Rule Mining*, Technical Report 99-10, Computer Science Department, Rensselaer Polytechnic Institute, Troy NY 12180, pp. 1-20.
95. Zaki M. J. (2004), "Mining Non-Redundant Association Rules", *Data Min. Knowl. Discov*, Vol. 9 (3), pp. 223-248.
96. Zaki M. J., Parthasarathy S., Ogihara M., and Li W. (1997), "New Algorithms for Fast Discovery of Association Rules", *Proc. of KDD 1997*, pp. 283-286.
97. Zaki M., Ogihara M., Parthasarathy S., Li M. (1996), "Parallel Data Mining for Association Rules on Shared-memory Multi-processors", *Proc. of the 1996 ACM/IEEE conference on Supercomputing* (CDROM).
98. Zhang L., Shi Y., and Yang X. (2005), "A Fuzzy Mining Algorithm for Association-Rule Knowledge Discovery", *Proc. of the Eleventh Americas Conference on Information Systems*, pp. 1487-1496.
99. http://www.jmulti.de/: phần JMULTI Open – Source Software.
100. http://archive.ics.uci.edu/ml/datasets.html: UCI-Machine Learning Repository.
101. http://academic.research.microsoft.com/Keyword/2246/association-rule-mining: Truy nhập ngày 18/11/2011.