BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Phạm Việt Anh

CÁC MÔ HÌNH TIÊN TIẾN CHO RÚT GỌN THUỘC TÍNH GIA TĂNG DỰA TRÊN TẬP MỜ TRỰC CẢM VÀ TẬP THÔ LÂN CẬN TRỌNG SỐ

LUẬN ÁN TIẾN SĨ NGÀNH MÁY TÍNH

Hà Nội - Năm 2026

BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Phạm Việt Anh

CÁC MÔ HÌNH TIÊN TIẾN CHO RÚT GỌN THUỘC TÍNH GIA TĂNG DỰA TRÊN TẬP MỜ TRỰC CẢM VÀ TẬP THÔ LÂN CẬN TRỌNG SỐ

LUẬN ÁN TIẾN SĨ NGÀNH MÁY TÍNH

Ngành: Hệ thống thông tin
Mã số: 9 48 01 04

Xác nhận của Học viện	**Người hướng dẫn 1**	**Người hướng dẫn 2**
Khoa học và Công nghệ	*(Ký, ghi rõ họ tên)*	*(Ký, ghi rõ họ tên)*

Hà Nội - Năm 2026

LỜI CAM ĐOAN

Tôi xin cam đoan luận án: "Các mô hình tiên tiến cho rút gọn thuộc tính gia tăng dựa trên tập mờ trực cảm và tập thô lân cận trọng số" là công trình nghiên cứu của chính mình dưới sự hướng dẫn khoa học của tập thể hướng dẫn. Luận án sử dụng thông tin trích dẫn từ nhiều nguồn tham khảo khác nhau và các thông tin trích dẫn được ghi rõ nguồn gốc. Các kết quả nghiên cứu của tôi được công bố chung với các tác giả khác đã được sự nhất trí của đồng tác giả khi đưa vào luận án. Các số liệu, kết quả được trình bày trong luận án là hoàn toàn trung thực và chưa từng được công bố trong bất kỳ một công trình nào khác ngoài các công trình công bố của tác giả. Luận án được hoàn thành trong thời gian tôi làm nghiên cứu sinh tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Hà Nội, ngày 10 tháng 02 năm 2026

Tác giả luận án

Phạm Việt Anh

LỜI CẢM ƠN

Luận án này được hoàn thành nhờ vào sự nỗ lực không ngừng của tác giả, cùng sự hỗ trợ tận tình từ các thầy cô hướng dẫn, sự đóng góp quý báu từ các chuyên gia, sự động viên, khích lệ tinh thần từ gia đình, bạn bè, đồng nghiệp, sinh viên trong suốt hành trình học tập và nghiên cứu.

Trước hết, tác giả xin bày tỏ lòng biết ơn sâu sắc đối với PGS.TS Nguyễn Long Giang và TS Nguyễn Ngọc Thủy, những người Thầy đã dành sự quan tâm đặc biệt, chỉ dẫn tận tình và động viên tác giả hoàn thành luận án đúng tiến độ và đạt được những mục tiêu đề ra. Sự hỗ trợ chuyên môn, những lời khuyên quý báu của các Thầy đã giúp tác giả hoàn thiện bản thân, nâng cao chất lượng nghiên cứu, củng cố niềm tin học thuật trong suốt quá trình thực hiện đề tài.

Tác giả cũng xin bày tỏ lòng biết ơn đến Ban Lãnh đạo, các Cán bộ Phòng Đào tạo và các phòng chức năng của Học viện Khoa học và Công nghệ vì đã dành sự quan tâm, sự hỗ trợ nhiệt tình trong suốt quá trình nghiên cứu của tôi.

Cuối cùng, tác giả xin gửi lời tri ân sâu sắc tới gia đình và người thân, những người đã luôn hy sinh vô điều kiện, là nguồn động viên mạnh mẽ về tinh thần và vật chất, mang đến môi trường tốt nhất để tác giả hoàn thành luận án. Sự hy sinh, tình yêu thương và niềm tin vô hạn của gia đình chính là động lực to lớn giúp tác giả vượt qua mọi khó khăn, thử thách trong suốt hành trình học tập và nghiên cứu.

Hà Nội, ngày 10 tháng 02 năm 2026

Tác giả luận án

Phạm Việt Anh

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ VIẾT TẮT

TT	**Ký hiệu**	**Tiếng Anh**	**Tiếng Việt**
1	IS	Information system	Hệ thông tin/Bảng quyết định
2	RS	Rough set	Tập thô
3	FRS	Fuzzy rough set	Tập thô mờ
4	NRS	Neighborhood rough set	Tập thô lân cận
5	WNRS	Weighted neighborhood rough set	Tập thô lân cận trọng số
6	KNNRS	k-nearest neighbor rough set	Tập thô k-lân cận gần nhất
7	WKNRS	Weighted k-nearest neighborhood rough set	Tập thô lân cận k láng giềng trọng số
8	IFS	Intuitionistic fuzzy set	Tập mờ trực cảm
9	IFRS	Intuitionistic fuzzy rough set	Tập thô mờ trực cảm
10	α, β-IFS	α, β-level intuitionistic fuzzy set	Tập mờ trực cảm mức α, β
11	IE	Information entropy	Entroypy thông tin
12	IFPOS	Intuitionistic fuzzy positive region	Miền dương mờ trực cảm
13	IFWNRS	Intuitionistic fuzzy weighted neighborhood rough set	Tập thô lân cận mờ trực cảm có trọng số
14	KNN	k-nearest neighbors	k-láng giềng gần nhất
15	ARPD	Attribute reduction based on the α, β-level intuitionistic fuzzy partition distance	Rút gọn thuộc tính dựa trên khoảng cách phân hoạch mờ trực cảm mức α, β
16	ARIFW	Attribute reduction based on the intuitionistic fuzzy weighted neighborhood rough set	Rút gọn thuộc tính dựa trên tập thô lân cận mờ trực cảm có trọng số

DANH MỤC BẢNG

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

MỞ ĐẦU

1. Tính cấp thiết của luận án

Trong thời đại ngày nay, dữ liệu lớn đã trở thành một trong những xu thế công nghệ nổi bật và thu hút được được rất nhiều sự quan tâm. Trước sự bùng nổ mạnh mẽ về số lượng dữ liệu, các công nghệ hiện nay đã gặp phải nhiều khó khăn trong quá trình lưu trữ cũng như khai phá tri thức. Cùng với đó, chất lượng dữ liệu cũng bị suy giảm khi có quá nhiều thông tin nhiễu, làm giảm hiệu quả hoạt động của các mô hình học máy. Chính vì vậy, nhiều giải pháp trên thế giới đã được đề xuất, trong đó rút gọn thuộc tính đã nổi lên như một hướng nghiên cứu quan trọng. Rút gọn thuộc tính là bài toán quan trọng trong bước tiền xử lý dữ liệu với mục tiêu chính là loại bỏ các thuộc tính dư thừa, dễ gây nhầm lẫn để tăng hiệu quả về mặt thời gian, độ chính xác cũng như tính đơn giản trong quá trình xây dựng các mô hình phân lớp hay các mô hình luật kết hợp [1, 2, 3]. Trên thế giới, các nghiên cứu về rút gọn thuộc tính hiện nay đang trở nên rất sôi động và tập trung chủ yếu vào quá trình xử lý trên các bảng quyết định.

Như một công cụ hữu hiệu, *mô hình tập thô* đã đặt một cơ sở vững chắc trong việc hình thành các thuật toán rút gọn thuộc tính trên bảng quyết định [4]. Bằng việc sử dụng mối quan hệ không phân biệt được giữa các đối tượng trong tập vũ trụ để định nghĩa một rút gọn, các thuật toán theo tiếp cận mô hình tập thô thường xử lý rất tốt trên dữ liệu có chứa thuộc tính rời rạc, điển hình là các kỹ thuật dựa trên ma trận phân biệt [5, 6, 7] và các phương pháp heuristic [8, 9, 10, 11, 12]. Tuy nhiên, khi xử lý trên các dữ liệu chứa các thuộc tính có miền giá trị số liên tục, các phương pháp dựa trên mô hình này phải trải qua bước rời rạc hóa dữ liệu. Cụ thể, với mỗi miền giá trị của thuộc tính, các phương pháp sẽ gán với một giá trị rời rạc. Do đó, quá trình này sẽ ảnh hưởng rất lớn tới việc bảo toàn thông tin và làm suy giảm hiệu quả của rút gọn thu được. Từ những khó khăn này, một số mở rộng của mô hình tập thô sau đó đã được phát triển để xử lý trực tiếp dữ liệu gốc, trong đó các mô hình mở rộng phổ biến và hiệu quả nhất được phát triển theo hai nhánh chính bao gồm *mô hình tập thô lân cận* và *mô hình tập thô mờ*.

Mô hình tập thô lân cận là một trong nhánh nghiên cứu quan trọng, đánh dấu bước phát triển đáng kể của lý thuyết tập thô. Mô hình này sử dụng mối quan hệ lân cận thay vì mối quan hệ không thể phân biệt được trong mô hình cổ điển. Theo đó, mỗi đối tượng trong vũ trụ sẽ được đặc trưng bởi một *hạt thông tin lân cận* chứa các đối tượng

có quan hệ lân cận với đối tượng đã cho trong bán kính δ. Dựa trên mô hình này, rất nhiều phương pháp rút gọn thuộc tính đã được đề xuất và xử lý trên nhiều trường hợp khác nhau của bảng quyết định. Điển hình là một số phương pháp rút gọn thuộc tính trên các bảng quyết định hỗn hợp sử dụng độ phụ thuộc [13, 14], tỉ lệ lỗi quyết định lân cận [15], chỉ số phân biệt lân cận [16] và entropy kết hợp [18]. Nhìn chung, các phương pháp theo hướng tiếp cận tập thô lân cận có khả năng xử lý rất tốt trên các bảng quyết định số hoặc hỗn hợp bởi khả năng mô tả đặc trưng của một đối tượng là đầy đủ hơn so với lý thuyết tập thô truyền thống. Tuy nhiên, các phương pháp này chỉ tập trung vào số lượng đối tượng trong một hạt thông tin. Điều này có nghĩa là tất cả các đối tượng trong một hạt thông tin đều được gán mức độ quan trọng như nhau đối với một quyết định nhất định. Mặc dù trên thực tế, dữ liệu luôn được phân bổ đa dạng, nghĩa là mỗi đối tượng trong một vùng lân cận đóng một vai trò khác nhau.

Dựa trên sự kết hợp giữa lý thuyết tập thô truyền thống và lý thuyết tập mờ, lý thuyết tập thô mờ đã được đề xuất bởi Dübois và Prade [34] được xem là nhánh nghiên cứu thứ hai nhằm xử lý cho dữ liệu liên tục. Trong mô hình này, mỗi đối tượng cho trước được biểu diễn bởi một hạt thông tin mờ, trong đó các đối tượng còn lại từ tập vũ trụ sẽ thuộc về lớp này dựa trên một độ thuộc. Trên cơ sở của mô hình tập thô mờ các thuật toán rút gọn thuộc tính được phát triển với nhiều độ đo khác nhau, điển hình là khoảng cách mờ [42, 43], miền dương mờ [44, 45, 46], thông tin tương hỗ mờ [47], entropy mờ [48, 49] và hạt thông tin mờ [50]. Kết quả thực nghiệm cho thấy các thuật toán rút gọn thuộc tính theo phương pháp tập thô mờ hiệu quả hơn các thuật toán truyền thống đối với các bảng quyết định có thuộc tính liên tục và thuộc tính số. Tuy nhiên, một số nghiên cứu đã chỉ ra rằng phương pháp rút gọn thuộc tính dựa trên tập thô mờ kém hiệu quả hơn khi xử lý các tập dữ liệu nhiễu có độ chính xác phân loại thấp.

Trong những năm gần đây, các phương pháp rút gọn thuộc tính đã mở rộng sang hướng tiếp cận *tập mờ trực cảm*, đặc biệt thông qua việc kết hợp với lý thuyết tập thô để hình thành *mô hình tập thô mờ trực cảm*. Ưu điểm của mô hình này xuất phát từ chính đặc trưng của tập mờ trực cảm, trong đó thành phần hàm không thuộc đóng vai trò quan trọng giúp điều chỉnh tốt các thông tin từ một số đối tượng nhiễu trong dữ liệu về đúng phân lớp [52]. Do đó, mô hình tập thô mờ trực cảm có khả năng phân loại các đối tượng tốt hơn so với tập mờ cổ điển, đặc biệt là trên các bộ dữ liệu nhiễu hoặc có độ nhất quán thấp. Theo hướng tiếp cận từ mô hình tập mờ trực cảm và tập thô mờ trực cảm, một số công trình đã mở rộng các độ đo truyền thống như miền dương mờ

trực cảm [53, 54, 55, 56], entropy thông tin mờ trực cảm [57, 58] và khoảng cách mờ trực cảm [59]. Một số kết quả thực nghiệm đã chứng minh được hiệu quả vượt trội của các thuật toán theo mô hình tập thô mờ trực cảm so với các thuật toán theo mô hình tập thô mờ. Tuy nhiên, mô hình tập thô mờ trực cảm vẫn còn tồn tại một số hạn chế mà đến nay nhiều nghiên cứu chưa thể khắc phục triệt để. Thứ nhất, việc bổ sung thành phần độ không thuộc làm cho các thuật toán theo hướng tiếp cận này đòi hỏi không gian lưu trữ lớn hơn và có độ phức tạp tính toán cao hơn so với các phương pháp tập thô mờ truyền thống, vốn chỉ xét đến thành phần độ thuộc. Rõ ràng, đối với các tập dữ liệu có số chiều lớn, số lượng thành phần độ không thuộc sẽ gia tăng đáng kể, gây khó khăn cho quá trình thực thi và làm giảm hiệu quả tính toán của các thuật toán. Thứ hai, các đối tượng có sự phân bố khác biệt so với phần lớn các đối tượng trong tập vũ trụ thường tạo ra nhiều phần tử trong các hạt thông tin mờ trực cảm có độ thuộc nhỏ và độ không thuộc lớn. Những đối tượng này có thể được sinh ra bởi nhiễu và làm suy giảm hiệu quả của các mô hình phân lớp. Khi đó, một số độ đo đánh giá thuộc tính thực hiện trên các giá trị này có thể dẫn đến sự suy giảm về chất lượng của các tập rút gọn thu được.

Bên cạnh đó, dữ liệu ngày nay luôn có sự gia tăng và thay đổi theo thời gian khiến cho các bảng quyết định có kích thước vô cùng lớn. Để giải quyết vấn đề này, các phương pháp rút gọn thuộc tính theo hướng tiếp cận gia tăng đã trở thành một hướng nghiên cứu mở rộng và đầy tiềm năng. Đối với trường hợp bảng quyết định có sự thay đổi về tập đối tượng, các thuật toán gia tăng chỉ tính toán trên phần thay đổi của dữ liệu chứ không xử lý trên toàn bộ dữ liệu. Do đó, thời gian thực thi được giảm thiểu đi rất nhiều. Trong trường hợp bảng quyết định có kích thước lớn, các thuật toán gia tăng tìm rút gọn trên mỗi thành phần của bảng quyết định bị chia nhỏ, sau đó thực hiện cập nhật lại chúng khi bổ sung các thành phần còn lại. Dựa trên hướng tiếp cận này, các phương pháp gia tăng cũng được áp dụng trên nhiều mô hình khác nhau. Trên mô hình tập thô, một số nghiên cứu đã đề xuất các phương pháp rút gọn thuộc tính gia tăng sử dụng ma trận phân biệt [60, 61, 62, 63, 64], miền dương [65, 66, 67, 68], entropy thông tin [69, 70] và hạt thông tin [71, 72, 73, 74, 75].

Đối với mô hình tập thô mờ, rất nhiều nghiên cứu cũng được triển khai để xử lý trên các bảng quyết định động chứa thuộc tính số, liên tục. Trong trường hợp bảng quyết định bổ sung tập đối tượng, nghiên cứu [76] đã xây dựng công thức tính độ phụ thuộc mờ và đề xuất thuật toán FIAT nhằm tìm kiếm một tập con thuộc tính xấp xỉ. Cũng

trong trường hợp này, nghiên cứu [77] cũng đề xuất hai thuật toán gia tăng sử dụng quan hệ phân biệt để tìm kiếm các rút gọn. Bằng việc mở rộng sang quan hệ tương đương mờ, các tác giả trong [78] đã đã thiết kế thuật toán gia tăng IARM nhằm tìm kiếm một rút gọn hiệu quả trên bảng quyết định có sự bổ sung thêm đối tượng. Tiếp theo, Zhang và các cộng sự trong [79] đã thiết kế một thuật toán gia tăng dựa trên độ đo entropy thông tin mờ. Trên cơ sở của khái niệm tập đối tượng then chốt, Ni và các cộng sự trong [80] đã trình bày hai thuật toán gia tăng sử dụng hàm thuộc mờ và miền dương mờ. Để đáp ứng các kịch bản thực tế của bảng quyết định khi có sự bổ sung và loại bỏ tập đối tượng, các tác giả trong [81, 82] đã định nghĩa lại rút gọn từ độ đo khoảng cách mờ và thiết kế các thuật toán gia tăng. Thông qua một số kết quả thực nghiệm, các nghiên cứu đều chứng minh được hiệu suất của các thuật toán gia tăng là vượt trội so với các thuật toán rút gọn thuộc tính trên bảng quyết định cố định. Tuy nhiên, khi áp dụng trên các bảng quyết định nhiễu, không nhất quán, các thuật toán này vẫn còn gặp nhiều khó khăn khi tập rút gọn chưa thể cải thiện hiệu quả về độ chính xác phân lớp. Do đó, đây được xem là một trong những khoảng trống nghiên cứu quan trọng, tạo động lực cho việc nghiên cứu và phát triển các thuật toán gia tăng dựa trên sự mở rộng của mô hình tập mờ trực cảm.

2. Mục tiêu nghiên cứu

Xuất phát từ những hạn chế còn tồn tại theo hai hướng nghiên cứu mở rộng của mô hình tập thô, cùng với những rào cản trong việc khai thác ưu điểm của mô hình tập mờ trực cảm khi thiết kế các thuật toán rút gọn thuộc tính trên các bảng quyết định có số chiều lớn và có sự thay đổi tập đối tượng, luận án tập trung vào hai mục tiêu nghiên cứu quan trọng sau:

1) *Đề xuất một số mô hình được mở rộng từ mô hình tập mờ trực cảm*: Trên cơ sở đã trình bày, vấn đề đầu tiên của luận án là xây dựng một số mô hình có khả năng kế thừa và tận dụng lợi thế của mô hình tập mờ trực cảm nhằm khắc phục những hạn chế của hai nhánh nghiên cứu mở rộng từ mô hình tập thô truyền thống. Thêm vào đó, các mô hình đề xuất có khả năng cải thiện thời gian thực thi và giảm thiểu ảnh hưởng của các đối tượng nhiễu trong dữ liệu so với mô hình tập thô mờ trực cảm.

2) *Thiết kế các thuật toán gia tăng dựa trên các mô hình được đề xuất*: Từ một số tính chất quan trọng của các mô hình đề xuất, vấn đề thứ hai của luận án là việc thiết kế các thuật toán gia tăng để xử lý trong các kịch bản thực tế của dữ liệu khi có sự bổ sung và loại bỏ tập đối tượng. Các thuật toán này không chỉ thích ứng với dữ liệu thay đổi

đổi mà còn có khả năng xử lý tốt dữ liệu nhiễu, nhiều chiều qua đó nâng cao hiệu quả phân lớp và chất lượng rút gọn thu được so với các mô hình phát triển trên hai nhánh mở rộng của mô hình tập thô.

3. Nội dung nghiên cứu

Nội dung của nghiên cứu tập trung vào các khái niệm nền tảng liên quan đến bảng quyết định, tập rút gọn và các độ đo đánh giá thuộc tính. Trên cơ sở đó, luận án cũng phân tích ưu điểm từ các mô hình được mở rộng từ lý thuyết tập thô như tập thô lân cận, tập thô mờ và tập thô mờ trực cảm nhằm xây dựng một số mô hình mới và phát triển các phương pháp rút gọn thuộc tính. Các phương pháp này được thiết kế để áp dụng hiệu quả cho các bảng quyết định cố định cũng như bảng quyết định có sự thay đổi về tập đối tượng.

4. Phạm vi nghiên cứu

Phạm vi nghiên cứu của luận án bao gồm:

1) Nghiên cứu các mô hình được mở rộng từ các nhánh của lý thuyết tập thô và các mô hình tiên tiến có sự kế thừa các ưu điểm vượt trội của tập mờ trực cảm trong việc xử lý trên các tập dữ liệu nhiễu.

2) Trên cơ sở các mô hình đã đề xuất, nghiên cứu tập trung vào các thuật toán rút gọn thuộc tính áp dụng cho bảng quyết định cố định và bảng quyết định thay đổi tập đối tượng, nhằm tìm kiếm các rút gọn hiệu quả trong phân lớp, đồng thời giảm thiểu thời gian thực thi của các thuật toán theo tiếp cận tập mờ trực cảm và tập thô mờ trực cảm hiện có.

5. Cơ sở khoa học và thực tiễn của đề tài

Phương pháp nghiên cứu của luận án được trình bày dựa trên sự kết hợp giữa cơ sở lý thuyết toán học và quá trình đánh giá thực nghiệm.

1) *Về lý thuyết:* Nghiên cứu và chứng minh một số tính chất quan trọng của các mô hình đề xuất, nghiên cứu các thuật toán Heuristic để tìm kiếm rút gọn của bảng quyết định cố định, bảng quyết định có sự bổ sung và loại bỏ tập đối tượng dựa trên các độ đo trong không gian của mô hình đề xuất.

2) *Về thực nghiệm:* Thử nghiệm, so sánh, đánh giá các thuật toán đề xuất với các thuật toán đã công bố trên các bộ dữ liệu tiêu chuẩn được thu thập từ kho dữ liệu UCI[1] và OpenML[2] để đánh giá hiệu quả của các thuật toán đề xuất theo các mục tiêu đặt ra.

[1] `https://archive.ics.uci.edu/datasets`

[2] `https://openml.org/search?type=data&status=active&sort=runs`.

6. Những đóng góp mới của luận án

Luận án tập trung làm rõ những đóng góp mới theo ba nhóm chính sau đây:

1) Đóng góp về mô hình và cơ sở lý thuyết:

- Đề xuất mô hình tập mờ trực cảm mức α, β và trình bày một số tính chất then chốt của mô hình, làm cơ sở lý thuyết cho việc xây dựng các phương pháp rút gọn thuộc tính trong không gian tập mờ trực cảm.

- Đề xuất mô hình tập thô lân cận mờ trực cảm có trọng số, cho phép đánh giá mức độ ảnh hưởng của các thuộc tính điều kiện đến quyết định của từng đối tượng, đồng thời đặc trưng hóa vai trò chi tiết của các đối tượng trong mỗi hạt thông tin.

2) Đóng góp về phương pháp rút gọn thuộc tính:

- Xây dựng độ đo khoảng cách phân hoạch mờ trực cảm mức α, β làm cơ sở để định nghĩa lại một rút gọn mới và xây dựng độ quan trọng của thuộc tính để lựa chọn được các thuộc tính có ý nghĩa cao. Qua đó, đề xuất một thuật toán rút gọn thuộc tính trên bảng quyết định cố định.

- Xây dựng độ đo khoảng cách giữa hai họ lân cận mờ trực cảm có trọng số, làm nền tảng cho việc định nghĩa một rút gọn mới và độ quan trọng của thuộc tính.

3) Đóng góp về tính mở rộng và khả năng ứng dụng

- Mở rộng công thức khoảng cách phân hoạch mờ trực cảm mức α, β nhằm tính toán nhanh trên các bảng quyết định có sự thay đổi tập đối tượng và hướng tới việc xây dựng các thuật toán gia tăng ứng dụng cho các kịch bản thực tế của dữ liệu.

- Mở rộng công thức tính khoảng cách giữa hai họ lân cận mờ trực cảm có trọng số để xử lý cho các trường hợp bảng quyết định có sự bổ sung và loại bỏ tập đối tượng hướng tới việc đề xuất hai thuật toán gia tăng trên bảng quyết định thay đổi tập đối tượng.

7. Cấu trúc luận án

Ngoài phần mở đầu và kết luận, luận án có 03 chương nội dung nghiên cứu:

Chương 1. Luận án ban đầu giới thiệu bài toán rút gọn thuộc tính thông qua một số hướng tiếp cận chính và các khái niệm cơ bản về bảng quyết định. Qua đó, luận án sẽ trình bày tổng quan về lý thuyết tập mờ trực cảm, tập thô mờ trực cảm, làm cơ sở cho việc hình thành một số mô hình mở rộng và đề xuất các thuật toán rút gọn thuộc tính. Các đóng góp chính của luận án được trình bày chi tiết trong Chương 2 và Chương 3.

Chương 2. Luận án trình bày về các thuật toán rút gọn thuộc tính trên bảng quyết định cố định và bảng quyết định có sự bổ sung và loại bỏ tập đối tượng dựa trên mô

hình tập mờ trực cảm mức α, β bao gồm các bước chính như sau:

1) Xây dựng mô hình tập mờ trực cảm mức α, β và trình bày một số tính chất quan trọng của mô hình.

2) Xây dựng độ đo khoảng cách giữa hai phân hoạch mờ trực cảm mức α, β và định nghĩa một rút gọn hiệu quả làm cơ sở trong việc thiết kế một thuật toán rút gọn thuộc tính trên bảng quyết định cố định.

3) Phát triển hai công thức tính toán khoảng cách làm cơ sở thiết kế hai thuật toán gia tăng rút gọn thuộc tính trên bảng quyết định có sự bổ sung và loại bỏ tập đối tượng.

Dựa trên các phương pháp được đề xuất, luận án sẽ trình bày một số thực nghiệm để chứng minh hiệu quả của phương pháp đạt được so với các phương pháp điển hình dựa trên các mô hình tập thô mờ, tập thô mờ trực cảm.

Chương 3. Luận án trình bày về các thuật toán rút gọn thuộc tính trên bảng quyết định có sự bổ sung và loại bỏ tập đối tượng theo tiếp cận mô hình tập thô lân cận mờ trực cảm có trọng số bao gồm các bước chính như sau:

1) Trình bày khái quát một số hạn chế từ các phướng pháp rút gọn thuộc tính trong việc đánh giá vai trò của mỗi thuộc tính lên khả năng phân lớp của mỗi đối tượng và đề xuất mô hình tập thô lân cận mờ trực cảm có trọng số.

2) Xây dựng độ đo khoảng cách giữa hai họ lân cận mờ trực cảm có trọng số và định nghĩa lại một rút gọn hiệu quả có tính tổng quát hóa tốt hơn so với các rút gọn khác để thiết kế một thuật toán rút gọn thuộc tính bảng quyết định cố định.

3) Phát triển hai công thức tính toán khoảng cách làm cơ sở thiết kế hai thuật toán gia tăng rút gọn thuộc tính trên bảng quyết định có sự bổ sung và loại bỏ tập đối tượng.

Thông qua các phương pháp đề xuất, luận án cũng trình bày một số thực nghiệm để chứng minh ưu điểm của phương pháp trong việc khắc phục các hạn chế từ một số phương pháp điển hình khác theo tiếp cận của mô hình tập thô mờ, tập thô lân cận trọng số và tập thô mờ trực cảm. Cuối cùng, phần kết luận sẽ trình bày những kết quả đã đạt được của luận án, các kiến nghị và hướng phát triển trong tương lai.

CHƯƠNG 1
TỔNG QUAN VỀ BÀI TOÁN RÚT GỌN THUỘC TÍNH TRÊN BẢNG QUYẾT ĐỊNH

1.1 Mở đầu

Trước thời đại phát triển mạnh mẽ của dữ liệu lớn, các cơ sở dữ liệu vẫn không ngừng gia tăng cả về số lượng các bản ghi và số lượng các thuộc tính. Điều này đã mang đến rất nhiều khó khăn cho việc triển khai các thuật toán trong lĩnh vực khai phá dữ liệu. Vấn đề này đã đặt ra một thách thức không hề nhỏ trong việc tìm cách loại bỏ đi số lượng các thuộc tính mà vẫn bảo toàn được những thông tin có giá trị quan trọng của dữ liệu. Do đó, rút gọn thuộc tính đã trở thành đề tài thu hút sự quan tâm của nhiều nhà nghiên cứu trong khoảng thời gian gần đây. Để có cái nhìn chi tiết về rút gọn thuộc tính, Chương 1 sẽ trình bày một số đóng góp chính như sau:

Thứ nhất, giới thiệu tổng quan về bài toán rút gọn thuộc tính và số hướng tiếp cận điển hình.

Thứ hai, trình bày tổng quan một số mô hình áp dụng cho bài toán rút gọn thuộc tính dựa trên hai nhánh mở rộng từ lý thuyết tập thô. Qua đó, phân tích những ưu điểm và nhược điểm của mỗi mô hình để rút ra những động lực trong nghiên cứu.

Thứ ba, trình bày các khái niệm cơ bản về bảng quyết định, mô hình tập thô mờ trực cảm, làm cơ sở đề xuất một số mô hình mở rộng đạt hiệu quả cao trong việc áp dụng các thuật toán rút gọn thuộc tính.

Kết quả nghiên cứu của chương này được công bố trong các công trình [CT3] thuộc phần Danh mục các công trình nghiên cứu của luận án.

1.2 Tổng quan về rút gọn thuộc tính

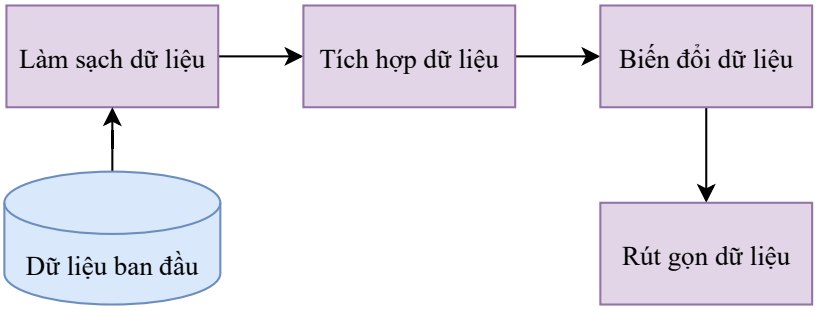
Rút gọn thuộc tính là một trong các hướng nghiên cứu then chốt được sử dụng phổ biến trong nhiều lĩnh vực khác nhau như học máy [83], thị giác máy tính [84], chẩn đoán y tế [85]. Ví dụ, trong lĩnh vực khai phá dữ liệu và học máy, rút gọn thuộc tính cho phép chuyển đổi dữ liệu thô thành các đặc trưng rõ ràng và có ý nghĩa, phản ánh được những thông tin cốt lõi, qua đó hỗ trợ hiệu quả cho việc nhận diện các lớp dữ liệu hoặc trích xuất những đặc trưng giàu thông tin. Thông qua việc nâng cao hiệu năng mô hình và cải thiện khả năng khái quát hóa, các phương pháp rút gọn thuộc tính không chỉ góp phần giảm chiều dữ liệu mà còn hạn chế tính dư thừa, khắc phục hiện tượng quá khớp

và nâng cao độ chính xác trong các bài toán phân loại hoặc hồi quy [86]. Trong chẩn đoán y tế, việc lựa chọn thuộc tính giúp loại bỏ các gen không liên quan và dư thừa không tham gia vào biểu hiện của các bệnh cụ thể, điều này rất quan trọng cho việc phát hiện khối u sớm và chẩn đoán ung thư [87].

Tại Việt Nam, đã có nhiều công trình nghiên cứu về các phương pháp rút gọn thuộc tính trên bảng quyết định. Tiêu biểu là luận án của tác giả Nguyễn Văn Thiện [1], trong đó luận án đã đề xuất độ phụ thuộc mờ và khoảng cách mờ để xây dựng các thuật toán tìm tập rút gọn trên các bảng quyết định số. Để giải quyết các kịch bản thực tế của dữ liệu, luận án của tác giả Hồ Thị Phượng [2] đã đề xuất một số thuật toán gia tăng nhằm tìm tập rút gọn trên các bảng quyết định thay đổi. Gần đây, luận án của tác giả Trần Thanh Đại [3] đã đưa ra phương pháp rút gọn thuộc tính theo hướng tiếp cận lai ghép dựa trên khoảng cách phân hoạch mờ trực cảm. Kết quả thực nghiệm cho thấy phương pháp này xử lý hiệu quả đối với các tập dữ liệu nhiễu trong việc cải thiện độ chính xác phân lớp từ tập thuộc tính thu được.

1.2.1 Định nghĩa về bài toán rút gọn thuộc tính

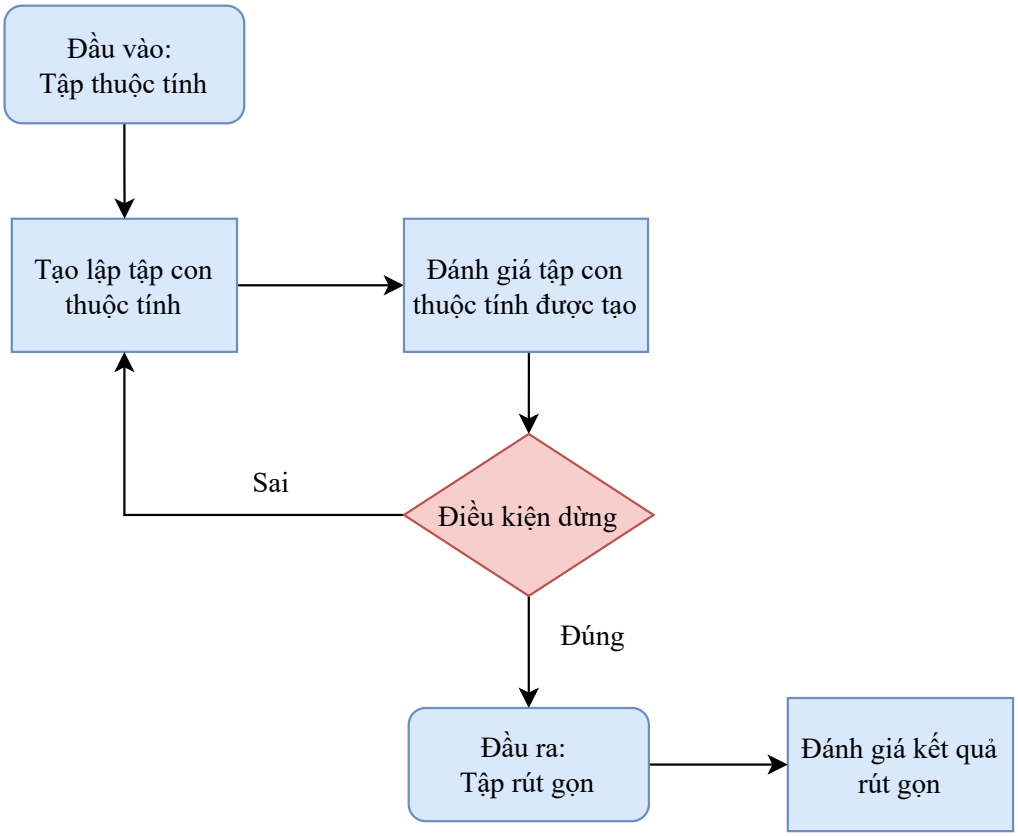
Rút gọn thuộc tính là một bài toán quan trọng được triển khai tại bước tiền xử lý dữ liệu trong quá trình khai phá dữ liệu. Bên cạnh các nhiệm vụ cơ bản của bước tiền xử lý dữ liệu được trình bày trong Hình 1.1 như làm sạch dữ liệu, tích hợp dữ liệu, chuyển đổi dữ liệu và rút gọn dữ liệu thì rút gọn thuộc tính có chức năng chính là giảm chiều dữ liệu nhưng vẫn bảo toàn thông tin so với tập dữ liệu ban đầu, trong khi làm tăng hiệu quả của các thuật toán khai thác dữ liệu như: tăng tính đơn giản và dễ hiểu của luật, tăng hiệu năng của các thuật toán nhờ loại bỏ đi các thuộc tính dư thừa, tăng độ chính xác cho các mô hình nhờ loại bỏ đi các thuộc tính nhiễu.



Hình 1.1: Các bước trong giai đoạn tiền xử lý dữ liệu

Cho tới nay, các công trình nghiên cứu về rút gọn thuộc tính thường tập trung vào

nghiên cứu các kỹ thuật lựa chọn thuộc tính. Lựa chọn thuộc tính là quá trình lựa chọn một tập con gồm $|B|$ thuộc tính từ tập gồm $|C|$ thuộc tính với $B \subseteq C$ sao cho không gian thuộc tính được thu gọn lại một cách tối ưu theo một tiêu chuẩn nhất định. Tuy nhiên, việc tìm ra một tập con thuộc tính tốt nhất thường rất khó thực hiện và một số bài toán liên quan tới vấn đề này thuộc lớp bài toán NP-khó [2]. Nhìn chung, một thuật toán lựa chọn thuộc tính thường bao gồm bốn bước cơ bản được mô tả chi tiết trong Hình 1.2.



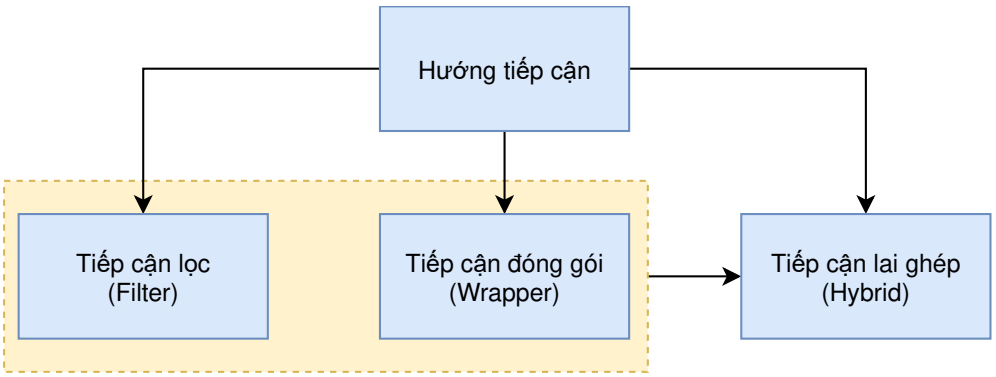
Hình 1.2: Các bước cơ bản trong lựa chọn thuộc tính

1) *Tạo lập tập con:* Đây là quá trình tìm kiếm liên tiếp nhằm tạo ra các tập con thuộc tính. Rõ ràng, quá trình này là rất quan trọng và góp phần rất lớn vào việc thu được một tập con thuộc tính cuối cùng. Giả sử rằng, tập dữ liệu ban đầu có $|C|$ thuộc tính, khi đó dựa trên tính chất của tập hợp, có thể thu được $2^{|C|}$ tập con thuộc tính. Như vậy, sẽ rất khó khăn khi tìm được một tập con được xem là tốt nhất từ toàn bộ các tập con này. Ngoài ra, khi $|C|$ là một số lớn, rõ ràng việc tạo ra các tập con sẽ là bất khả thi. Do đó, việc tạo lập tập con phải thông qua một số tiêu chí để lựa chọn được một tập con ý nghĩa. Quá trình tạo lập các tập con có thể xem như một giai đoạn tạo ra các ứng tuyển phục vụ cho quá trình đánh giá tiếp theo.

2) *Đánh giá tập con:* Từ quá trình tạo lập các tập con, quá trình đánh giá tập con là việc so sánh các tập con dựa theo một tiêu chuẩn được định nghĩa. Nói cách khác, mỗi tập con sinh ra sẽ được so sánh với tập con tốt nhất trước đó. Nếu tập con này tốt hơn, nó sẽ thay thế tập cũ. Quá trình này sẽ được dừng lại khi điều kiện dừng xảy ra.
3) *Kiểm tra điều kiện dừng:* Điều kiện dừng là việc kết thúc quá trình tìm kiếm tập con thuộc tính tối ưu thông qua một số ràng buộc đã được quy ước. Các ràng buộc này có thể dựa vào số lượng thuộc tính thu được theo quy định, số bước lặp cho quá trình lựa chọn, việc thêm hay loại bớt một thuộc tính nào đó trong tập con thuộc tính không làm cho tập con trở nên tốt hơn hay tập con tốt nhất đã đảm bảo các tiêu chuẩn đánh giá.
4) *Kiểm chứng kết quả:* Tập con tốt nhất cuối cùng phải được kiểm chứng thông qua việc tiến hành các phép kiểm định, so sánh các kết quả khai phá so với tập thuộc tính ban đầu trên các tập dữ liệu khác nhau.

1.2.2 Các hướng tiếp cận trong rút gọn thuộc tính

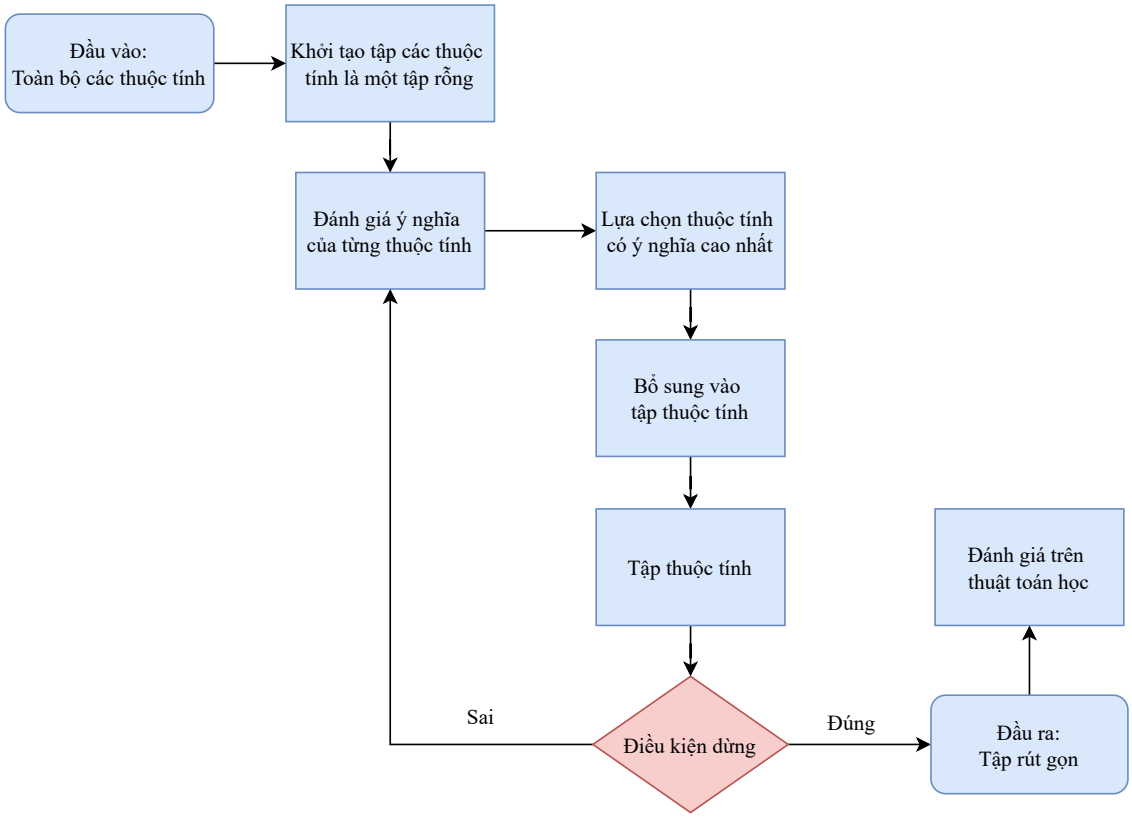
Hiện nay, có rất nhiều hướng tiếp cận trong bài toán rút gọn thuộc tính. Tuy nhiên, ba cách tiếp cận thường được sử dụng là: Lọc (Filter), đóng gói (Wrapper) và lai ghép (Hybrid). Các cách tiếp cận này được biểu diễn trong Hình 1.3 và đều có những mục tiêu riêng trong việc giảm số lượng thuộc tính.



Hình 1.3: Phân loại các tiếp cận rút gọn thuộc tính

1) *Tiếp cận lọc:* Hướng tiếp cận lọc được xem là một hướng tiếp cận phổ biến và được sử dụng rộng rãi trong các bài toán rút gọn thuộc tính. Hiện nay, các phương pháp theo hướng tiếp cận lọc thường sử dụng một độ đo làm cơ sở để đánh giá ý nghĩa cho một tập con thuộc tính thu được. Thông qua các tính chất của độ đo, phương pháp này sẽ định nghĩa một rút gọn là một tập con các thuộc tính bảo toàn lượng thông tin của dữ liệu so với toàn bộ tập thuộc tính ban đầu. Qua đó, các thuộc tính có ý nghĩa cao nhất từ độ đo sẽ được lựa chọn lần lượt cho tới khi điều kiện dừng xảy ra hoặc tập con thuộc

tính thu được thỏa mãn các tính chất của rút gọn. Cuối cùng, phương pháp sẽ sử dụng rút gọn thu được để đánh giá hiệu năng trên các thuật toán học. Các bước của hướng tiếp cận lọc được trình bày chi tiết trong Hình 1.4 dưới đây.



Hình 1.4: Hướng tiếp cận lọc trong rút gọn thuộc tính

2) *Tiếp cận đóng gói:* Đây là cách tiếp cận thường được gắn với một mô hình cụ thể để đánh giá tập con thuộc tính tốt nhất, do đó cách tiếp cận này có chi phí tính toán rất lớn, đặc biệt khi xử lý trên các bộ dữ liệu nhiều chiều. Bên cạnh đó, rút gọn thu được theo cách tiếp cận đóng gói chỉ phù hợp với một mô hình học máy cụ thể, nếu sử dụng tập rút gọn đó sang mô hình học máy khác có thể sẽ không hiệu quả. Do đó, cho tới, nay tiếp cận lọc vẫn được sử dụng phổ biến hơn do mục tiêu đánh giá tập thuộc tính rút gọn được khái quát theo tiêu chí bảo toàn thông tin so với tập dữ liệu gốc.

3) *Tiếp cận lai ghép:* Để khai thác những ưu điểm của tiếp cận lọc và tiếp cận đóng gói, nhiều nhà nghiên cứu đề suất kỹ thuật *lai ghép lọc-đóng gói* (Filter-Wrapper) để chọn lọc tập con thuộc tính có độ chính xác phân lớp cao nhất. Ưu điểm của phương pháp này đó là thời gian tìm kiếm rút gọn đạt hiệu quả cao nhất trên một mô hình phân lớp cụ thể nhanh hơn rất nhiều so với tiếp cận lai ghép truyền thống do không gian các tập rút gọn ứng viên đã được giới hạn tại giai đoạn lọc của thuật toán. Cụ thể, trong giai đoạn lọc, một tập chứa các rút gọn ứng viên sẽ được hình thành và đưa đến giai đoạn

đóng gói với một mô hình phân lớp đã được định nghĩa trước để đánh giá xem rút gọn ứng viên nào đạt độ chính xác cao nhất. Rõ ràng, phương pháp này cũng mang theo những nhược điểm của phương pháp đóng gói truyền thống khi việc đánh giá trên một mô hình không thể đại diện hiệu năng phân lớp cho các mô hình khác. Do đó, tại bước này, có thể sử dụng với nhiều mô hình phân lớp khác nhau để kiểm chứng về kết quả rút gọn thu được.

1.2.3 Bảng quyết định và một số mô hình trong rút gọn thuộc tính

Đầu tiên, *bảng quyết định* là một trường hợp đặc biệt của hệ thông tin quyết định và được biểu diễn bởi một cặp $IS = (U, C \cup D)$, trong đó U là một tập hữu hạn khác rỗng các đối tượng, C và D là các tập hữu hạn khác rỗng các thuộc tính thỏa mãn $C \cap D = \emptyset$. Mỗi thuộc tính $c \in C \cup D$ xác định một ánh xạ $c : U \to V_c$ là một giá trị của thuộc tính c. Khi đó, cho $u \in U$ và $c \in C \cup D$, giá trị của thuộc tính c với đối tượng u được ký hiệu là $c(u)$. Ở đây, C được gọi là tập các thuộc tính điều kiện và D là tập các thuộc tính quyết định. Trong trường hợp D có nhiều thuộc tính quyết định thì bằng một phép chuyển đổi hoàn toàn có thể biểu diễn D dưới dạng một thuộc tính quyết định [88]. Do đó, luận án chỉ xét bảng quyết định $IS = (U, C \cup D)$.

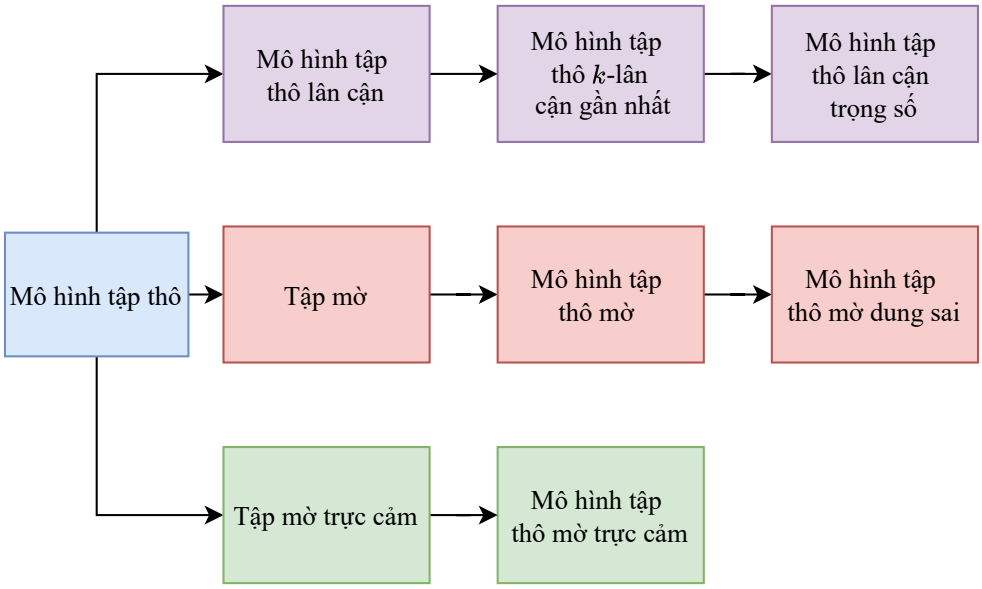
Ví dụ 1.1 Cho bảng quyết định $IS = (U, C \cup D)$, trong đó $U = \{u_1, u_2, ...u_5\}$ và $C = \{a_1, a_2, ...a_5\}$ như trong Bảng 1.1.

Bảng 1.1: Một ví dụ về bảng quyết định

U	a_1	a_2	a_3	a_4	a_5	D
u_1	0.72	0.93	0.65	0.52	0.69	Yes
u_2	0.43	0.92	0.52	0.44	0.57	No
u_3	0.21	0.64	0.85	0.62	0.21	Yes
u_4	0.95	0.82	0.21	0.81	0.94	No
u_5	0.45	0.72	0.28	0.87	0.18	Yes

Theo các khái niệm về bảng quyết định, rút gọn thuộc tính là việc loại bỏ đi các thuộc tính dư thừa trong C để trích lọc được một tập con thuộc tính điều kiện $A \subseteq C$ bảo toàn thông tin của bảng quyết định dựa trên một tiêu chuẩn được định nghĩa. Lý thuyết tập thô do Pawlak đề xuất [4] là một công cụ toán học hữu ích để xử lý thông tin thiếu hụt và không chắc chắn trong phân tích dữ liệu. Cốt lõi của lý thuyết này là xấp xỉ tập đối tượng dựa trên quan hệ không phân biệt trong bảng thông tin. Một vấn

đề quan trọng của lý thuyết tập thô là rút gọn thuộc tính, nhằm giữ lại các thuộc tính cần thiết và loại bỏ các thuộc tính dư thừa mà vẫn đảm bảo độ chính xác trong phân loại và dự đoán. Từ nền tảng này, nhiều phương pháp đã được phát triển để giải quyết hiệu quả bài toán rút gọn thuộc tính, tiêu biểu là các kỹ thuật dựa trên ma trận phân biệt [5, 6, 7] và các phương pháp heuristic [8, 9, 10, 11, 12]. Tuy nhiên, mô hình tập thô ban đầu chỉ phù hợp với dữ liệu chứa giá trị rời rạc hoặc định danh và còn hạn chế khi áp dụng cho dữ liệu liên tục hoặc số. Để khắc phục hạn chế trên, nhiều mô hình đã kế thừa nền tảng lý thuyết của Pawlak nhưng thay thế quan hệ không phân biệt bằng các dạng quan hệ khác. Từ đó hình thành hai nhánh nghiên cứu chính bao gồm mô hình tập thô lân cận và mô hình tập thô mờ, như minh họa ở Hình 1.5.



Hình 1.5: Các nhánh mở rộng từ mô hình tập thô

Nhằm tránh bước rời rạc hóa dữ liệu và mở rộng khả năng áp dụng cho các tập dữ liệu chứa thuộc tính số hoặc liên tục, một số nghiên cứu ban đầu đã đề xuất sử dụng quan hệ lân cận để thay thế cho quan hệ không phân biệt được trong lý thuyết tập thô truyền thống. Cụ thể, tập thô lân cận (NRS) và tập thô k-lân cận gần nhất (KNNRS) được Hu và cộng sự lần đầu giới thiệu vào năm 2008 [13]. Trong nghiên cứu này, các tác giả đã định nghĩa quan hệ lân cận thông qua khoảng cách giữa hai đối tượng $u, v \in U$ theo một tập con thuộc tính điều kiện trên bảng quyết định. Cụ thể, cho một tập con thuộc tính $A \subseteq C$, khoảng cách giữa hai đối tượng u và v tương ứng với tập thuộc tính A, ký hiệu là $\Delta_A(u, v)$, được xác định như sau:

$$\Delta_A(u, v) = \sqrt[p]{\sum_{a \in A} |a(u) - a(v)|^p} \tag{1.1}$$

trong đó, $\Delta_A(u,v)$ được gọi là khoảng cách Manhattan nếu $p=1$, khoảng cách Euclidean nếu $p=2$ và khoảng cách Chebyshev nếu $p=\infty$.

Giả sử rằng, δ là một bán kính lân cận có giá trị nằm trong khoảng $[0,1]$, một quan hệ lân cận $\mathcal{R}_A^\delta$ trên U từ tập thuộc tính A, được xác định như sau [14]:

$$\mathcal{R}_A^\delta = \{(u,v) \in U \times U : \Delta_A(u,v) \leq \delta\} \tag{1.2}$$

Khi đó, nếu $(u,v) \in R_A^\delta$, chúng ta nói rằng u và v có quan hệ lân cận với nhau theo tập thuộc tính A. Dựa trên quan hệ này, với một đối tượng $u \in U$, $[u]_A^\delta = \left\{v \in U : (u,v) \in \mathcal{R}_A^\delta\right\}$ được gọi là một hạt thông tin lân cận của đối tượng u được tạo ra bởi quan hệ $\mathcal{R}_A^\delta$ và họ của các hạt thông tin lân cận $U/\mathcal{R}_A^\delta = \left\{[u]_A^\delta : u \in U\right\}$ được gọi là một phủ lân cận của tập con thuộc tính A.

Dựa trên khái niệm về hạt thông tin lân cận, xét một tập đối tượng $X \subseteq U$, các định nghĩa về xấp xỉ trên $\overline{N_A}(X)$ và xấp xỉ dưới $\underline{N_A}(X)$ của X được trình bày như sau

$$\overline{N_A}(X) = \left\{u \in U : [u]_A^\delta \cap X \neq \emptyset\right\} \tag{1.3}$$

và

$$\underline{N_A}(X) = \left\{u \in U : [u]_A^\delta \subseteq X\right\} \tag{1.4}$$

Từ cơ sở này, nhiều biến thể của mô hình tập thô lân cận đã được phát triển để nâng cao hiệu quả trong việc rút gọn thuộc tính. Bằng việc sử dụng độ phụ thuộc lân cận, Hu và cộng sự [14] đã đề xuất một thuật toán rút gọn với hiệu năng vượt trội so với các thuật toán tập thô truyền thống. Năm 2018, Wang và cộng sự [16] tiếp tục đưa ra một thuật toán hiệu quả dựa trên chỉ số phân biệt lân cận để xác định tập con đặc trưng thích hợp. Vài năm sau đó, Wang và cộng sự [17] giới thiệu độ đo thông tin lân cận tương đối, xét đến cả xấp xỉ dưới và xấp xỉ trên để nhận diện tập thuộc tính quan trọng. Từ góc nhìn về đại số và lý thuyết thông tin, Sun và cộng sự [19] đề xuất độ đo entropy đa hạt lân cận mờ và áp dụng vào bài toán chọn đặc trưng tối ưu. Zhang và cộng sự [20] xây dựng độ đo entropy thông tin hỗn hợp lân cận có điều kiện nhằm xử lý dữ liệu đa dạng trong chọn đặc trưng. Yang và cộng sự [21] lại tiếp cận theo hướng học metric khoảng cách (distance metric learning), tối ưu hóa cấu trúc hạt thông tin và đề xuất hai thuật toán trích chọn đặc trưng liên quan.

Mở rộng từ mô hình tập thô k-lân cận gần nhất, Chen và các cộng sự [22] đề xuất một cải tiến dựa trên việc định nghĩa độ bất thường của hạt. Xu và Zhu [23] sau đó đề xuất một thuật toán rút gọn thuộc tính dựa trên entropy phân phối lân cận có trọng

số để thu được đầy đủ thông tin không chắc chắn do phân phối nhãn gây ra. Ngoài ra, nhiều biến thể khác của mô hình tập thô lân cận cũng đã được phát triển, tiêu biểu như tập thô lân cận phủ [24], tập thô lân cận địa phương [25], tập thô quyết định cực đại [26], hay tập thô lân cận mềm [27, 28].

Ưu điểm của mô hình tập thô lân cận là khả năng lựa chọn trực tiếp các thuộc tính từ các bảng quyết định số, loại bỏ sự cần thiết của quá trình rời rạc hóa dữ liệu trong khi vẫn đảm bảo hiệu quả phân lớp. Bên cạnh đó, quan hệ lân cận chỉ tập trung vào các đối tượng thuộc vào lân cận của một đối tượng cho trước. Do đó, mô hình tập thô lân cận giúp thu hẹp phạm vi tính toán và tăng khả năng xử lý cho các thuật toán rút gọn thuộc tính. Tuy nhiên, mô hình tập thô lân cận không xét tới ảnh hưởng của từng thuộc tính cho mỗi quyết định của các đối tượng. Nói cách khác, mô hình này giả định rằng trọng số của mỗi thuộc tính điều kiện là như nhau. Điều này có thể dẫn đến việc mô tả sai về mối quan hệ giữa các thuộc tính điều kiện và thuộc tính quyết định. Do đó, một số thuộc tính có mối quan hệ chặt chẽ với quyết định có thể không được đại diện đầy đủ để phản ánh tầm quan trọng thực sự của chúng. Kết quả này dẫn đến việc bỏ qua những thuộc tính có ý nghĩa trong quá trình rút gọn. Để giải quyết vấn đề này, Hu và các cộng sự [29] đã đề xuất mô hình tập thô lân cận trọng số (WNRS) sử dụng khoảng cách dựa trên trọng số các thuộc tính:

$$\Delta_A^{\omega}(u,v) = \sqrt{\sum_{a\in A} \omega^2(a).(a(u)-a(v))^2} \tag{1.5}$$

trong đó, $\omega(a)$ là trọng số của thuộc tính $a \in A$, được xác định như sau. Cho bảng quyết định $IS = (U, C \cup D)$, $U = \{u_1, u_2, \ldots, u_n\}$ là tập các đối tượng, $C = \{a_1, a_2, \ldots, a_m\}$ là tập các thuộc tính điều kiện và D là tập thuộc tính quyết định. Hệ số phân hoạch của các thuộc tính là một vector $\lambda = (\lambda(a_1), \lambda(a_2), \ldots, \lambda(a_m))^{\mathrm{T}}$ được xác định dựa trên một bài toán tìm nghiệm tối ưu $\lambda^* = \text{argmin}\|\mathbf{A}\lambda - \mathbf{Y}\|^2$ từ hàm $\mathcal{L}(\lambda) = \|\mathbf{A}\lambda - \mathbf{Y}\|^2$, trong đó $\|.\|$ biểu diễn 2-norm của một vector, $\mathbf{A}$ là ma trận hệ số được định nghĩa bởi

$$\mathbf{A} = \begin{bmatrix} a_1(u_1) & a_2(u_1) & \cdots & a_m(u_1) \\ a_1(u_2) & a_2(u_2) & \cdots & a_m(u_2) \\ \vdots & \vdots & \ddots & \vdots \\ a_1(u_n) & a_2(u_n) & \cdots & a_m(u_n) \end{bmatrix},$$

và $\mathbf{Y} = (D(a_1), D(a_2), \ldots, D(a_m))^{\mathrm{T}}$ là một *vector quyết định*. Để tìm nghiệm tối ưu, đầu tiên giả sử $\mathbf{A}\lambda = \mathbf{Y}$, khi đó nhân cả hai vế với $\mathbf{A}^{\mathrm{T}}$ để phương trình trở thành

$\mathbf{A}^{\mathrm{T}}\mathbf{A}\lambda = \mathbf{A}^{\mathrm{T}}\mathbf{Y}$. Cuối cùng, chúng ta thu được $\lambda = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{Y}$. Tuy nhiên, trong thực tế, bảng quyết định có thể chứa số lượng thuộc tính nhiều hơn so với số lượng đối tượng. Khi đó, ma trận $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ có thể *không khả nghịch*. Để giải quyết vấn đề này, một đại lượng phạt hay còn gọi là một tham số điều chỉnh được bổ sung vào hàm tối ưu. Khi đó, hàm tối ưu trở thành $\mathcal{L}(\lambda) = \|\mathbf{A}\gamma - \mathbf{Y}\|^2 + \|\lambda\|^2$ và là một hàm lồi có giá trị nhỏ nhất khi $\mathcal{L}'(\lambda) = 2\mathbf{A}^{\mathrm{T}}(\mathbf{A}\lambda - \mathbf{Y}) = 0$. Do đó, $\left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \mathbf{E}\right)\lambda = \mathbf{A}^{\mathrm{T}}\mathbf{Y}$, trong đó $\mathbf{E}$ là một ma trận đơn vị cùng chiều với ma trận $\mathbf{A}^{\mathrm{T}}\mathbf{A}$. Từ đây, chúng ta có thể dễ dàng thu được nghiệm tối ưu $\lambda = (\mathbf{A}^{\mathrm{T}}\mathbf{A} + \mathbf{E})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{Y}$.

Dựa trên hệ số phân hoạch của các thuộc tính, trọng số của mỗi thuộc tính $a \in C$ được xác định bởi công thức:

$$\omega(a) = \frac{|C| \cdot |\lambda(a)|}{\sum_{a \in C} |\lambda(a)|} \tag{1.6}$$

trong đó, $|C|$ biểu diễn số lượng các thuộc tính điều kiện trong bảng quyết định và $|\lambda(a)|$ là giá trị tuyệt đối của $\lambda(a)$.

Trong nghiên cứu này, để thể hiện tính tổng quát, công thức khoảng cách trọng số được định nghĩa lại như sau:

$$\Delta_A^{\omega}(u, v) = \sqrt[p]{\sum_{a \in A} (\omega(a) \,.\, |a(u) - a(v)|)^p} \tag{1.7}$$

Qua đó, xét một bảng quyết định $IS = (U, C \cup D)$ và một tập thuộc tính $A \subseteq C$, *quan hệ lân cận trọng số*, ký hiệu là $\mathcal{R}_A^{\delta,\omega}$, được định nghĩa bởi:

$$\mathcal{R}_A^{\delta,\omega} = \{(u, v) \in U \times U : \Delta_A^{\omega}(u, v) \leq \delta\} \tag{1.8}$$

Khi đó, với một đối tượng $u \in U$, một hạt thông tin lân cận của u được tạo bởi trọng số các thuộc tính được gọi là một *hạt thông tin lân cận trọng số thuộc tính*, ký hiệu là $[u]_A^{\delta,\omega}$ và được xác định như sau:

$$[u]_A^{\delta,\omega} = \{v \in U : (u, v) \in \mathcal{R}_A^{\delta,\omega}\} \tag{1.9}$$

Rõ ràng, họ của tất cả các hạt thông tin này sẽ tạo nên một phủ trên U, ký hiệu là $U/\mathcal{R}_A^{\delta,\omega} = \{[u]_A^{\delta,\omega} : u \in U\}$. Theo đó, với một tập đối tượng $X \subseteq U$, *xấp xỉ dưới* và *xấp xỉ trên* của X dựa trên các hạt thông tin lân cận trọng số theo A được xác định tương ứng như sau:

$$\underline{WN_A}(X) = \left\{u \in U : [u]_A^{\delta,\omega} \subseteq X\right\} \tag{1.10}$$

và

$$\overline{\mathrm{WN}_A}(X) = \{u \in U : [u]_A^\omega \cap X \neq \emptyset\} \tag{1.11}$$

Dựa trên các khái niệm đã được trình bày, các tác giả trong [29] đã xây dựng hàm phụ thuộc lân cận nhằm đánh giá độ quan trọng của từng thuộc tính.

$$\gamma(A, D) = \frac{|WPOS_A(D)|}{|U|} \tag{1.12}$$

trong đó $WPOS_A(D) = \bigcup_{d_i \in U/D} \underline{WN_A}(d_i)$ là miền dương quyết định với U/D là một phân hoạch quyết định trên U.

Tiếp nối hướng tiếp cận này, Xie và các cộng sự [30] đã tính toán trọng số thuộc tính thông qua phân cụm dữ liệu và lý thuyết entropy, từ đó đề xuất mô hình tập thô xác suất lân cận có trọng số và giới thiệu một số tính chất quan trọng. Mới đây, Thuy và Wongthanavasu [31] cũng đã đề xuất phương pháp gán trọng số cho các thuộc tính dựa trên độ phân kỳ mờ trong bài toán rút gọn thuộc tính. Mặc dù các kết quả thực nghiệm đã chỉ ra rằng mô hình tập thô lân cận trọng số mang lại hiệu quả cao trong việc cải thiện quá trình rút gọn, nhưng các mô hình này chủ yếu chỉ tập trung vào số lượng đối tượng trong mỗi hạt thông tin. Điều này có thể hiểu là các đối tượng trong hạt thông tin đều được xem như có độ quan trọng như nhau trong việc ra quyết định về đối tượng u. Tuy nhiên, dữ liệu trong thực tế luôn có sự phân bố đa dạng, nghĩa là mỗi đối tượng trong hạt thông tin sẽ có vai trò khác nhau. Ví dụ, những đối tượng gần u có ảnh hưởng lớn hơn so với các đối tượng xa u. Do đó, việc đánh giá chính xác độ quan trọng của các đối tượng trong hạt thông tin lân cận trọng số là rất cần thiết.

Từ vấn đề này, Wang và cộng sự [32] đã đề xuất mô hình tập thô lân cận k-láng giềng trọng số (WKNRS). Phương pháp này đánh giá chất lượng và gán trọng số cho các đối tượng lân cận gần nhất dựa trên độ lệch chuẩn. Gần đây, Thuy và các cộng sự [33] đã giới thiệu một mô hình trọng số tổng quát và chứng minh rằng một số mở rộng của mô hình tập thô lân cận có thể được coi là một trường hợp đặc biệt của mô hình này. Có thể nhận thấy rằng các phương pháp rút gọn thuộc tính theo nhánh mở rộng của mô hình NRS đã mang lại những kết quả khả quan trong việc xử lý các bảng quyết định số. Tuy nhiên, một số mô hình vẫn tồn tại những hạn chế sau:

- Cấu trúc hạt thông tin trong các mô hình thường mang tính chất đơn giản, chưa phản ánh đầy đủ vai trò của từng đối tượng trong hạt, đồng thời không xem xét được sự không chắc chắn và do dự vốn luôn tồn tại trong dữ liệu thực tế. Nói cách

khác, các mô hình này chủ yếu tập trung vào số lượng đối tượng trong mỗi hạt thông tin mà chưa đánh giá một cách chi tiết sự phân bố của chúng bên trong hạt.

- Các mô hình vẫn chưa có sự tích hợp giữa cả trọng số thuộc tính và trọng số đối tượng, mặc dù sự kết hợp này hứa hẹn sẽ mang tới khả năng cải thiện đáng kể về hiệu quả lựa chọn một rút gọn tối ưu.

- Các mô hình thường sử dụng độ phụ thuộc lân cận như một tiêu chuẩn để định nghĩa một rút gọn. Khi đó, nếu bảng quyết định là không nhất quán thì độ đo này chỉ xét đến các đối tượng nằm trong miền dương của bảng quyết định mà bỏ qua rất nhiều đối tượng khác, kể cả những đối tượng nằm trong miền biên. Rõ ràng, những đối tượng này được xem là không thể phân loại một cách chắc chắn và việc loại bỏ chúng trong quá trình tính toán sẽ ảnh hưởng đáng kể đến chất lượng của các rút gọn.

Nhánh nghiên cứu về mô hình tập thô mờ đã được phát triển song song và hạn chế được một số khó khăn của mô hình tập thô lân cận trong việc biểu diễn đặc trưng của các hạt thông tin. Dựa trên sự kết hợp giữa lý thuyết tập thô truyền thống và lý thuyết tập mờ, lý thuyết tập thô mờ (FRS), được đề xuất bởi Dübois và Prade [34], cũng được xem là một hướng nghiên cứu mới để xử lý dữ liệu liên tục hoặc số. Ý tưởng của mô hình này là sử dụng quan hệ nhị phân mờ được thiết lập dựa trên một công thức tính toán mức độ tương đương giữa hai đối tượng trong tập vũ trụ. Khi đó, mỗi thuộc tính điều kiện $a \in C$ sẽ xác định một quan hệ nhị phân mờ $\widetilde{\mathcal{R}}_{\{a\}}$ trên $U \times U$ với $\widetilde{\mathcal{R}}_{\{a\}}(u, v) \in [0, 1]$. Khi đó, $\widetilde{\mathcal{R}}_{\{a\}}$ được gọi là quan hệ tương đương mờ nếu thỏa mãn các tính chất sau mọi đối tượng $u, v \in U$.

1. Tính phản xạ: $\widetilde{\mathcal{R}}_{\{a\}}(u, u) = 1$,
2. Tính đối xứng: $\widetilde{\mathcal{R}}_{\{a\}}(u, v) = \widetilde{\mathcal{R}}_{\{a\}}(v, u)$,
3. Tính bắc cầu: $\widetilde{\mathcal{R}}_{\{a\}}(u, v) \geq \sup\limits_{t \in U} \min \left\{ \widetilde{R}_{\{a\}}(u, t), \widetilde{R}_{\{a\}}(t, v) \right\}$.

Quan hệ tương đương mờ sẽ sinh ra một phân hoạch mờ $U/\widetilde{\mathcal{R}}_{\{a\}} = \left\{ \widetilde{[u]}_{\{a\}} : u \in U \right\}$ trên U, trong đó $\widetilde{[u]}_{\{a\}}$ được gọi là hạt thông tin mờ của đối tượng u. Lưu ý rằng, $\widetilde{[u]}_{\{a\}}$ cũng là một tập mờ và độ thuộc của mỗi đối tượng $v \in U$ được ký hiệu là $\widetilde{[u]}_{\{a\}}(v)$, chính là giá trị của $\widetilde{\mathcal{R}}_{\{a\}}(u, v)$. Qua đó, lực lượng của hạt thông tin $\widetilde{[u]}_{\{a\}}$, ký hiệu là $|\widetilde{[u]}_{\{a\}}|$ được xác định bởi $\left|\widetilde{[u]}_{\{a\}}\right| = \sum\limits_{v \in U} \widetilde{[u]}_{\{a\}}(v)$.

Có thể thấy rằng, cấu trúc của hạt thông tin mờ được biểu diễn chi tiết hơn so với hạt thông tin lân cận. Do đó, các đặc tính của dữ liệu được thể hiện đầy đủ hơn trong

mô hình FRSs. Tiếp theo, luận án sẽ điểm qua một số nghiên cứu được mở rộng từ mô hình FRSs nhằm ứng dụng trong bài toán rút gọn thuộc tính.

Khởi điểm trong hướng tiếp cận này là một phương pháp được đề xuất bởi Jensen và Shen với ứng dụng trong phân loại web [35]. Tiếp theo, Tsang và cộng sự [36] đã sử dụng ma trận phân biệt để định nghĩa các phần tử rút gọn trong môi trường FRS, sau đó giới thiệu một thuật toán để xác định tất cả các phần tử rút gọn và một thuật toán heuristic để tìm một phần tử rút gọn. Cũng dựa trên ma trận phân biệt, nhiều phương pháp rút gọn thuộc tính đã được đề xuất để xử lý dữ liệu liên tục, chẳng hạn như [37, 38, 39, 40]. Từ góc nhìn về lý thuyết thông tin, Dai và Xu [41] đã lựa chọn các thuộc tính dựa trên tỷ lệ thông tin mờ. Bên cạnh đó, các thuật toán rút gọn thuộc tính dựa trên FRS cũng được phát triển với nhiều độ đo khác nhau, ví dụ như khoảng cách mờ [42, 43], miền dương mờ [44, 45, 46], thông tin tương hỗ mờ [47], entropy mờ [48, 49], và hạt thông tin mờ [50]. Kết quả thực nghiệm chỉ ra rằng các thuật toán rút gọn thuộc tính dựa trên FRS mang lại hiệu quả vượt trội so với các thuật toán dựa trên tập thô truyền thống trong việc xử lý các bảng quyết định số hoặc liên tục. Tuy nhiên, qua quá trình khảo sát nhận thấy rằng, mô hình FRS vẫn còn một số hạn chế như sau:

- Nhiều đối tượng nằm trong hạt thông tin mờ với các mức độ thuộc rất nhỏ có thể gây ra nhiễu trong quá trình tính toán. Điều này cũng giống với một số kết quả đã được chỉ ra trong các nghiên cứu [51, 59] khi các thuật toán dựa trên mô hình FRS gặp nhiều khó khăn khi xử lý trên các tập dữ liệu nhiễu không nhất quán. Trong nghiên cứu này, các tập dữ liệu nhiễu được đặc trưng bởi hiệu quả phân lớp thấp và không khả thi khi áp dụng vào việc xây dựng các mô hình học máy.

- Các độ đo đánh giá thuộc tính được xây dựng trên mô hình đều dựa trên quá trình tính toán lực lượng của các hạt thông tin mờ. Do đó, khi xử lý trên các tập dữ liệu có độ chính xác phân lớp ban đầu thấp thường thu được một rút gọn chưa tối ưu về kích thước cũng như hiệu quả phân lớp.

- Các phương pháp theo mô hình FRS hiện chưa áp dụng việc thiết lập trọng số của các thuộc tính điều kiện. Do đó, quá trình rút gọn có thể bỏ qua các đối tượng có ý nghĩa cao. Đây có thể xem như một hạn chế của mô hình FRS so với một số mở rộng của mô hình tập NRS.

Nhìn chung, cả hai hướng mở rộng của lý thuyết tập thô đều đối mặt với không ít khó khăn. Những hạn chế này đã được phân tích chi tiết trong công trình [CT3] thuộc

danh mục nghiên cứu của luận án và là động lực thúc đẩy tác giả đề xuất những mở rộng mới trên cả hai nhánh nghiên cứu. Để giải quyết các vấn đề nêu trên, mô hình tập mờ trực cảm được coi là một giải pháp tiềm năng, vừa giúp mô tả rõ ràng cấu trúc hạt thông tin, vừa giảm nhiễu cho các thuật toán khi xử lý trên các tập dữ liệu có độ chính xác phân lớp ban đầu thấp. Vì vậy, mô hình này được lựa chọn như một nền tảng then chốt trong việc phát triển một số mô hình mở rộng khác, nhằm khắc phục các hạn chế còn tồn tại trong các nhánh của lý thuyết tập thô.

1.3 Mô hình tập thô mờ trực cảm

Trong phần này, luận án sẽ giới thiệu một số khái niệm cơ bản về mô hình tập thô mờ trực cảm. Đồng thời, một số hướng nghiên cứu liên quan đến bài toán rút gọn thuộc tính dựa trên tập thô mờ trực cảm cũng sẽ được trình bày. Trên cơ sở đó, luận án phân tích những ưu điểm và hạn chế của mô hình này, làm tiền đề cho việc xây dựng và phát triển các mô hình mở rộng nhằm nâng cao hiệu quả của các thuật toán rút gọn thuộc tính.

1.3.1 Lý thuyết tập mờ trực cảm

Lý thuyết tập mờ trực cảm [89] là một sự mở rộng tổng quát so với lý thuyết tập mờ nhờ khả năng biểu diễn chi tiết sự mơ hồ, không chắc chắn của thông tin. Trên thực tế, tập mờ trực cảm là một lựa chọn phù hợp trong những trường hợp mà việc biểu diễn mức độ không thuộc trở nên đơn giản hơn so với mức độ thuộc. Khái niệm này xuất phát từ kỳ vọng rằng quá trình ra quyết định của con người cũng như các hoạt động dựa trên chuyên môn và tri thức, vốn thường không chính xác tuyệt đối hoặc thiếu độ tin cậy cao, có thể được mô phỏng hiệu quả thông qua tập mờ cảm. Lý thuyết tập mờ trực cảm đã được ứng dụng thành công trong nhiều lĩnh vực, đặc biệt là trong nhận dạng mẫu và ra quyết định.

Định nghĩa 1.1 (Tập mờ trực cảm [89]). Cho U là tập hữu hạn khác rỗng các đối tượng, một tập mờ trực cảm (IFS) $\ddot{K}$ trên U có dạng $\ddot{K} = \{(u, \gamma_{\ddot{K}}(u), \eta_{\ddot{K}}(u)) \,|u \in U\}$, với $\gamma_{\ddot{K}} : U \to [0, 1]$ là độ thuộc và $\eta_{\ddot{K}} : U \to [0, 1]$ là độ không thuộc của đối tượng u trong $\ddot{K}$ sao cho $0 \leq \gamma_{\ddot{K}}(u) + \eta_{\ddot{K}}(u) \leq 1$, $\forall u \in U$.

Dựa vào khái niệm trên, có thể thấy tập mờ truyền thống được xem như một trường hợp đặc biệt của tập mờ trực cảm và có thể được biểu diễn dưới dạng $\ddot{K} = \{(u, \gamma_{\ddot{K}}(u), 1 - \gamma_{\ddot{K}}(u)) \,|u \in U\}$. Độ do dự của đối tượng u trong tập $\ddot{K}$ được xác định bởi $\pi_{\ddot{K}}(u) = 1 - \gamma_{\ddot{K}}(u) - \eta_{\ddot{K}}(u)$. Khi $\pi_{\ddot{K}}(u) = 0$, $\forall u \in U$ tập mờ trực cảm có thể chuyển

thành tập mờ truyền thống. Lực lượng của $\ddot{K}$ được ký hiệu là $\left|\ddot{K}\right|$ và được tính bởi công thức sau [90]:

$$\left|\ddot{K}\right| = \sum_{u \in U} \frac{1 + \gamma_{\ddot{K}}(u) - \eta_{\ddot{K}}(u)}{2} \tag{1.13}$$

Định nghĩa 1.2 (Các phép toán quan hệ trên IFS [89]). Cho $\ddot{K}$ và $\ddot{P}$ là hai tập mờ trực cảm trên U. Một số phép toán đánh giá mối quan hệ của $\ddot{K}$ và $\ddot{P}$ được trình bày như sau:

1. $\ddot{K} \subseteq \ddot{P}$ nếu và chỉ nếu $\gamma_{\ddot{K}}(u) \leq \gamma_{\ddot{P}}(u)$ và $\eta_{\ddot{K}}(u) \geq \eta_{\ddot{P}}(u)$, với mọi $u \in U$,
2. $\ddot{K} = \ddot{P}$ nếu và chỉ nếu $\ddot{K} \subseteq \ddot{P}$ và $\ddot{P} \subseteq \ddot{K}$,
3. $\ddot{K} \cap \ddot{P} = \{(u, \min(\gamma_{\ddot{K}}(u), \gamma_{\ddot{P}}(u)), \max(\eta_{\ddot{K}}(u), \eta_{\ddot{P}}(u)))\}$, với mọi $u \in U$,
4. $\ddot{K} \cup \ddot{P} = \{(u, \max(\gamma_{\ddot{K}}(u), \gamma_{\ddot{P}}(u)), \min(\eta_{\ddot{K}}(u), \eta_{\ddot{P}}(u)))\}$, với mọi $u \in U$.

Các phép toán cơ bản tổng quát trong tập mờ trực cảm còn được mở rộng cho các phép hợp và giao bằng cách sử dụng một số toán tử $\mathcal{T}_{norm}$ và $\mathcal{T}_{conorm}$ như trình bày trong Bảng 1.2.

Bảng 1.2: Một số phép toán tổng quát

$\mathcal{T}_{norm}$	$\mathcal{T}_{conorm}$
$\mathcal{T}_m(x,y) = \min\{x,y\}$	$\mathcal{S}_m(x,y) = \max\{x,y\}$
$\mathcal{T}_p(x,y) = xy$	$\mathcal{S}_p(x,y) = x + y - xy$
$\mathcal{T}_L(x,y) = \max\{x+y-1, 0\}$	$\mathcal{S}_L(x,y) = \min\{x+y, 1\}$
$\mathcal{T}_{cos}(x,y) =$	$\mathcal{S}_{cos}(x,y) =$
$\max\left\{xy - \sqrt{1-x^2}\sqrt{1-y^2}, 0\right\}$	$\min\left\{x + y - xy + \sqrt{2x-x^2}\sqrt{2y-y^2}, 1\right\}$

1.3.2 Mô hình tập thô mờ trực cảm

Trong phần này, luận án bắt đầu bằng việc đưa ra khái niệm về quan hệ mờ trực cảm làm cơ sở hình thành mô hình tập thô mờ trực cảm. Khái niệm này được xem là một mở rộng của quan hệ mờ trong lý thuyết tập mờ cổ điển bằng việc xét đến độ khác biệt giữa hai đối tượng.

Định nghĩa 1.3 (Quan hệ mờ trực cảm [91]). Cho bảng quyết định $IS = (U, C \cup D)$, mỗi thuộc tính $a \in C$ sẽ xác định một quan hệ nhị phân mờ trực cảm $\ddot{\mathcal{R}}_{\{a\}}$ trên $U \times U$ như sau:

$$\ddot{\mathcal{R}}_{\{a\}} = \left\{\left((u,v), \gamma_{\ddot{\mathcal{R}}_{\{a\}}}(u,v), \eta_{\ddot{\mathcal{R}}_{\{a\}}}(u,v)\right) : (u,v) \in U \times U\right\} \tag{1.14}$$

trong đó, $\gamma_{\ddot{\mathcal{R}}_{\{a\}}}(u,v) \in [0,1]$ và $\eta_{\ddot{\mathcal{R}}_{\{a\}}}(u,v) \in [0,1]$ tương ứng là độ tương tự và độ khác biệt của đối tượng u và v theo $\ddot{\mathcal{R}}_{\{a\}}$ sao cho $0 \leq \gamma_{\ddot{\mathcal{R}}_{\{a\}}}(u,v) + \eta_{\ddot{\mathcal{R}}_{\{a\}}}(u,v) \leq 1$.

Từ Định nghĩa 1.3, một quan hệ $\ddot{\mathcal{R}}_{\{a\}}$ được gọi là quan hệ tương đương mờ trực cảm nếu thỏa mãn các tính chất sau với mọi $u, v \in U$

1. Tính phản xạ: $\gamma_{\ddot{\mathcal{R}}_{\{a\}}}(u,u) = 1$ và $\eta_{\ddot{\mathcal{R}}_{\{a\}}}(u,u) = 0$,
2. Tính đối xứng: $\gamma_{\ddot{\mathcal{R}}_{\{a\}}}(u,v) = \gamma_{\ddot{\mathcal{R}}_{\{a\}}}(v,u)$ và $\eta_{\ddot{\mathcal{R}}_{\{a\}}}(u,v) = \eta_{\ddot{\mathcal{R}}_{\{a\}}}(v,u)$,
3. Tính bắc cầu: $\begin{cases} \gamma_{\ddot{\mathcal{R}}_{\{a\}}}(u,v) \geq \sup\limits_{t\in U}\left\{\min\left(\gamma_{\ddot{\mathcal{R}}_{\{a\}}}(u,t), \gamma_{\ddot{\mathcal{R}}_{\{a\}}}(t,v)\right)\right\} \\ \eta_{\ddot{\mathcal{R}}_{\{a\}}}(u,v) \leq \inf\limits_{t\in U}\left\{\max\left(\eta_{\ddot{\mathcal{R}}_{\{a\}}}(u,t), \eta_{\ddot{\mathcal{R}}_{\{a\}}}(t,v)\right)\right\} \end{cases}$

Mỗi thuộc tính $a \in C$ trên bảng quyết định sẽ xác định một quan hệ tương đương mờ trực cảm $\ddot{\mathcal{R}}_{\{a\}}$. Khi đó, một quan hệ tương đương mờ trực cảm $\ddot{\mathcal{R}}_{\{a\}}$ sẽ sinh ra một phân hoạch mờ trực cảm $U/\ddot{\mathcal{R}}_{\{a\}} = \left\{[\ddot{u}]_{\{a\}} \,|u \in U\right\}$ trên tập vũ trụ U, trong đó $[\ddot{u}]_{\{a\}}$ được gọi là một hạt thông tin mờ trực cảm của đối tượng u theo quan hệ tương đương mờ trực cảm $\ddot{\mathcal{R}}_{\{a\}}$. Rõ ràng, mỗi hạt thông tin mờ trực cảm $[\ddot{u}]_{\{a\}}$ là một tập mờ trực cảm trên U. Khi đó, với mỗi $v \in U$ thì $[\ddot{u}]_{\{a\}}(v) = \left(\gamma_{[\ddot{u}]_{\{a\}}}(v), \eta_{[\ddot{u}]_{\{a\}}}(v)\right) = \left(\gamma_{\ddot{\mathcal{R}}_{\{a\}}}(u,v), \eta_{\ddot{\mathcal{R}}_{\{a\}}}(u,v)\right)$.

Việc định nghĩa một quan hệ trong không gian tập mờ trực cảm phụ thuộc vào cách xây dựng các công thức tính toán độ tương tự và độ khác biệt. Trên cơ sở đó, nhiều nghiên cứu đã tập trung vào việc thiết lập các giá trị này cụ thể trên mỗi tập thuộc tính của bảng quyết định. Cụ thể, Tan và các cộng sự [55] đã áp dụng một phương pháp đơn giản bằng cách tính toán độ tương tự của hai đối tượng $u, v \in U$ theo $a \in C$ như sau:

$$\gamma_{[\ddot{u}]_{\{a\}}}(v) = \max\left(\min\left(\frac{a(u) - a(v) + \sigma_a}{\sigma_a}, \frac{a(u) - a(v) + \sigma_a}{\sigma_a}\right), 0\right) \tag{1.15}$$

trong đó, σ_a là độ lệch chuẩn của thuộc tính a. Độ khác biệt được thiết lập như sau:

$$\eta_{[\ddot{u}]_{\{a\}}}(v) = 1 - \gamma_{[\ddot{u}]_{\{a\}}}(v) \tag{1.16}$$

Bên cạnh đó, Jain và các cộng sự [56] đã xây dựng các công thức tính toán độ tương tự và độ khác biệt dựa trên khoảng cách Euclidean. Để xử lý trên các bảng quyết định hỗn hợp, Tan và các cộng sự [57] đã đưa ra các công thức tính toán độ tương tự giữa hai đối tượng u và v theo thuộc tính $a \in C$ trong hai trường hợp.

1. Nếu a là thuộc tính số/liên tục:

$$\gamma_{[\ddot{u}]_{\{a\}}}(v) = 1 - \frac{|a(u) - a(v)|}{\max a - \min a} \tag{1.17}$$

trong đó $\max(a)$ và $\min(a)$ tương ứng là giá trị lớn nhất và nhỏ nhất của thuộc tính a. Về bản chất, thành phần mẫu số của công thức trên là quá trình chuẩn hóa dữ liệu min-max để đảm bảo các giá trị của bảng quyết định luôn nằm trong khoảng [0,1].

2. Nếu a là thuộc tính rời rạc:

$$\gamma_{[\ddot{u}]_{\{a\}}}(v) = \begin{cases} 1, & a(u) = a(v) \\ 0, & a(u) \neq a(v) \end{cases} \tag{1.18}$$

Dựa trên toán tử $\mathcal{T}_{norm}$ Lukasiewicz, các tác giả trong [57] xây dựng độ khác biệt giữa hai đối tượng u và v từ lân cận chung gần nhất của chúng thuộc các lớp khác nhau.

$$\eta_{[\ddot{u}]_{\{a\}}}(v) = \begin{cases} \max\left(\min\limits_{D(t)\neq D(u)} \gamma_{\ddot{\mathcal{R}}_{\{a\}}}(u,t) + \gamma_{\ddot{\mathcal{R}}_{\{a\}}}(v,t) - 1, 0\right), & D(u) = D(v) \\ 1 - \gamma_{\ddot{\mathcal{R}}_{\{a\}}}(u,v), & D(u) \neq D(v) \end{cases} \tag{1.19}$$

Từ các khái niệm được trình bày, các tác giả trong [94, 95] đã xây dựng tập thô mờ trực cảm thông qua các định nghĩa về tập xấp xỉ trên và xấp xỉ dưới.

Định nghĩa 1.4 (Tập thô mờ trực cảm [94, 95]). Cho bảng quyết định $IS = (U, C \cup D)$, một quan hệ tương đương mờ trực cảm $\ddot{\mathcal{R}}_{\{a\}}$ được xác định trên thuộc tính $a \in C$ và một tập con đối tượng $X \subseteq U$. Khi đó, các tập xấp xỉ dưới và xấp xỉ trên của X theo $\ddot{\mathcal{R}}_{\{a\}}$ được định nghĩa tương ứng như sau:

$$\underline{\ddot{\mathcal{R}}}_{\{a\}}(X) = \left\{\left(u, \inf_{v\in U} \vee\left(\eta_{[\ddot{u}]_{\{a\}}}(v), \gamma_X(v)\right), \sup_{v\in U} \wedge\left(\gamma_{[\ddot{u}]_{\{a\}}}(v), \eta_X(v)\right)\right) : u \in U\right\} \tag{1.20}$$

và

$$\overline{\ddot{\mathcal{R}}}_{\{a\}}(X) = \left\{\left(u, \sup_{v\in U} \wedge\left(\gamma_{[\ddot{u}]_{\{a\}}}(v), \gamma_X(v)\right), \inf_{v\in U} \vee\left(\eta_{[\ddot{u}]_{\{a\}}}(v), \eta_X(v)\right)\right) : u \in U\right\} \tag{1.21}$$

Nếu $\underline{\ddot{\mathcal{R}}}_{\{a\}}(X) = \overline{\ddot{\mathcal{R}}}_{\{a\}}(X)$ thì cặp $\left(\underline{\ddot{\mathcal{R}}}_{\{a\}}(X), \overline{\ddot{\mathcal{R}}}_{\{a\}}(X)\right)$ được gọi là một tập xác định theo nghĩa mờ trực cảm, ngược lại thì cặp $\left(\underline{\ddot{\mathcal{R}}}_{\{a\}}(X), \overline{\ddot{\mathcal{R}}}_{\{a\}}(X)\right)$ được gọi là một tập thô mờ trực cảm (IFRS).

1.4 Các nghiên cứu liên quan đến rút gọn thuộc tính dựa trên tập thô mờ trực cảm

Trong phần này, luận án sẽ trình bày khái quát một số nghiên cứu điển hình về các phương pháp rút gọn thuộc tính dựa trên mô hình IFRS. Các phương pháp này xử lý trên các bảng quyết định cố định và bảng quyết định có sự thay đổi tập đối tượng.

1.4.1 Rút gọn thuộc tính trên bảng quyết định cố định

Trong bảng quyết định, các thuộc tính điều kiện được phân thành ba nhóm chính: thuộc tính lõi, thuộc tính rút gọn và thuộc tính dư thừa. Thuộc tính lõi là những thuộc tính thiết yếu, luôn xuất hiện trong mọi tập rút gọn. Thuộc tính rút gọn là những thuộc

tính có mặt trong ít nhất một tập rút gọn. Thuộc tính dư thừa là những thuộc tính mà việc loại bỏ chúng không làm ảnh hưởng đến kết quả phân lớp dữ liệu. Ngoài ra, cũng có thể xét đến thuộc tính không liên quan, tức những thuộc tính không mang lại giá trị thông tin đáng kể trong quá trình phân tích và ra quyết định. Trong các nghiên cứu về rút gọn thuộc tính, một rút gọn thường được định nghĩa dựa trên một độ đo đánh giá khả năng bảo toàn thông tin so với toàn bộ thuộc tính ban đầu. Ngoài ra, các thuộc tính trong rút gọn đóng góp đáng kể vào hiệu quả phân lớp và tăng khả năng học cho các thuật toán trong quá trình huấn luyện. Trên cơ sở đó, nhiều nghiên cứu đã định nghĩa rút gọn thông qua một số độ đo trên mô hình IFRS.

1) *Miền dương mờ trực cảm*

Cũng giống như ý tưởng mở rộng từ độ đo miền dương trong mô hình cổ điển sang mô hình FRSs, các nhà nghiên cứu đã tập trung vào các tính chất của tập mờ trực cảm trên không gian tập thô để xây dựng khái niệm miền dương mờ trực cảm. Xuất phát từ mô hình IFRSs, Tan và các cộng sự [55] ban đầu phân hoạch dữ liệu thành một họ các lớp quyết định $U/D = \{d_1, d_2, \ldots, d_k\}$ theo các nhãn quyết định khác nhau. Qua đó, với mỗi tập con thuộc tính quyết định $A \subseteq C$, các tác giả đã tính toán lại các xấp xỉ dưới của mỗi lớp quyết định d_i với $1 \leq i \leq k$.

$$\ddot{\underline{\mathcal{R}}}_A(d_i) = \left\{ \left(u, \inf_{v \in U} \vee \left(\eta_{[\ddot{u}]_A}(v), \gamma_{d_i}(v) \right), \sup_{v \in U} \wedge \left(\gamma_{[\ddot{u}]_A}(v), \eta_{d_i}(v) \right) \right) : u \in U \right\} \tag{1.22}$$

trong đó, $\gamma_{[\ddot{u}]_A}(v) = \min_{a \in A} \gamma_{[\ddot{u}]_{\{a\}}}(v)$ và $\eta_{[\ddot{u}]_A}(v) = \max_{a \in A} \eta_{[\ddot{u}]_{\{a\}}}(v)$.

Tiếp theo, các tác giả đã xây dựng độ đo miền dương mờ trực cảm của tập thuộc tính A từ việc kết hợp các xấp xỉ dưới của các lớp quyết định.

$$IFPOS_A(D) = \bigcup_{d_i \in U/D} \ddot{\underline{\mathcal{R}}}_A(d_i) \tag{1.23}$$

Rõ ràng, $IFPOS_A(D)$ là một tập mờ trực cảm được tạo ra bằng cách lấy hợp các xấp xỉ dưới của tất cả các lớp quyết định. Qua đó, nó phản ánh khả năng phân lớp của tập con thuộc tính A, nghĩa là $IFPOS_A(D)$ càng lớn, khả năng phân lớp của A càng mạnh. Từ độ đo miền dương mờ trực cảm, các tác giả trong [55] đã định nghĩa một rút gọn của bảng quyết định. Cụ thể, một tập con $A \subseteq C$ được gọi là một rút gọn của IS nếu thỏa mãn:

1. $IFPOS_A(D) = IFPOS_C(D)$,
2. $IFPOS_{A \setminus \{a\}}(D) \subset IFPOS_A(D)$, $\forall a \in C$.

Trong một số nghiên cứu khác, Tiwari và các cộng sự [53] áp dụng lý thuyết miền dương cho mô hình IFRSs để định nghĩa một rút gọn trên bảng quyết định. Gần đây,

Redman và các cộng sự [96] hoàn thiện một số chứng minh của Tiwari trong [54] và xây dựng lại các xấp xỉ trong việc hình thành tập thô dung sai mờ trực cảm. Qua đó, các tác giả đề xuất một thuật toán rút gọn thuộc tính để tìm kiếm các rút gọn trên bảng quyết định. Với mục tiêu hạn chế nhiễu trong dữ liệu, Jain và các cộng sự [56] cấu trúc lại mô hình IFRSs với đề xuất mới về lớp quyết định mờ trực cảm. Sau đó, các tác giả đã xây dựng hàm phụ thuộc là tỉ lệ giữa lực lượng của miền dương mờ trực cảm và số lượng đối tượng trong bảng quyết định để định nghĩa một rút gọn và thiết kế thuật toán tìm kiếm các thuộc tính tối ưu của bảng quyết định.

2) *Entropy mờ trực cảm*

Entropy là một khái niệm có thể biểu diễn bản chất của tri thức và thông tin dưới nhiều hình thức khác nhau, nhờ đó nó được ứng dụng rộng rãi trong nhiều lĩnh vực, đặc biệt là trong việc định lượng mức độ không chắc chắn của dữ liệu. Dựa trên nền tảng lý thuyết do Shannon đề xuất, ba loại entropy thường được sử dụng bao gồm: entropy thông tin, entropy kết hợp và entropy có điều kiện. Trong các bài toán rút gọn thuộc tính, entropy đóng vai trò là một thước đo hiệu quả để đánh giá mức độ quan trọng của các thuộc tính thông qua khả năng phản ánh sự không chắc chắn mà chúng mang lại. Nhằm bổ sung một số ràng buộc chặt chẽ hơn so với entropy truyền thống, Tan và các cộng sự [57] đã xây dựng độ đo entropy điều kiện từ mô hình IFRSs thông qua việc tính toán trên các hạt thông tin của một tập thuộc tính $A \subseteq C$, cụ thể như sau:

$$IE\left(D\,|A\right) = -\frac{1}{|U|}\sum_{u\in U}\log\frac{|[\ddot{u}]_A \cap [\ddot{u}]_D|}{|[\ddot{u}]_A|} \tag{1.24}$$

Dựa trên độ đo đề xuất, các tác giả đã định nghĩa một rút gọn tương đối của bảng quyết định làm cơ sở cho việc thiết kế một thuật toán rút gọn thuộc tính theo hướng tiếp cận lọc. Cụ thể, một tập con thuộc tính $A \subseteq C$ được gọi là một rút gọn của IS nếu thỏa mãn

1. $IE\left(D|A\right) = IE\left(D|C\right)$,
2. $IE\left(D|A'\right) \neq IE\left(D|A\right)$, với mọi $A' \subset A$.

Có thể thấy rằng, một rút gọn tương đối là một tập con thuộc tính tối thiểu bảo toàn giá trị entropy nguyên thủy của bảng quyết định. Từ độ đo entropy mờ trực cảm, các tác giả sau đó đã chứng minh tính chất đơn điệu của độ đo với kích thước tập thuộc tính và định nghĩa độ quan trọng của thuộc tính để lựa chọn các thuộc tính có ý nghĩa. Cụ thể, với một tập con thuộc tính $A \subseteq C$ và một thuộc tính $a \in C$, độ quan trọng của

a theo A được định nghĩa bởi

$$Sig(a, A) = IE\left(D|A\right) - IE\left(D|A \cup \{a\}\right) \tag{1.25}$$

Bên cạnh công trình [57], Revanasiddappa và cộng sự [58] đã áp dụng độ đo entropy mờ trực cảm để rút gọn thuộc tính cho bài toán phân lớp văn bản. Tiếp theo, Rahimi và các cộng sự [97] đã kết hợp các khái niệm về entropy và tập mờ trực cảm để hình thành một độ đo entropy hiệu quả trong việc lựa chọn và xếp hạng các nhà cung cấp theo các thuộc tính. Gần đây, Liu và các cộng sự [98] đã chỉ ra nhược điểm trong nghiên cứu [57] khi chỉ xét đến tầm quan trọng của từng thuộc tính đối với quyết định một cách riêng lẻ. Từ quan điểm này, các tác giả đã xây dựng độ đo entropy ưu thế dựa trên quan hệ ưu thế mờ trực cảm và thiết kế một thuật toán rút gọn thuộc tính. Nhìn chung, với khả năng đánh giá sự không chắc chắn của thông tin, entropy được xem là một độ đo quan trọng và được ứng dụng hiệu quả trong việc lựa chọn các thuộc tính tối ưu trên bảng quyết định.

3) *Khoảng cách mờ trực cảm*

Độ đo khoảng cách là độ đo quan trọng được sử dụng để đo lường sự khác biệt giữa hai đối tượng. Trong công trình [91], nhóm nghiên cứu đã xây dựng khoảng cách mờ trực cảm để đo lường lượng thông tin của một phân hoạch mờ trực cảm không thuộc vào phân lớp.

$$\mathcal{D}\left(U/\ddot{\mathcal{R}}_A, U/\ddot{\mathcal{R}}_{A\cup D}\right) = \frac{1}{|U|^2}\sum_{u\in U}\left(|[\ddot{u}]_A| - |[\ddot{u}]_A \cap [\ddot{u}]_D|\right) \tag{1.26}$$

Dựa trên công thức này, các tác giả đã chứng minh tính phản đơn điệu của độ đo so với kích thước của tập thuộc tính và định nghĩa một rút gọn mới với các tiêu chí sau

1. $\mathcal{D}\left(U/\ddot{\mathcal{R}}_A, U/\ddot{\mathcal{R}}_{A\cup D}\right) = \mathcal{D}\left(U/\ddot{\mathcal{R}}_C, U/\ddot{\mathcal{R}}_{C\cup D}\right)$,
2. $\forall A^{'} \subseteq A$, $\mathcal{D}\left(U/\ddot{\mathcal{R}}_{A'}, U/\ddot{\mathcal{R}}_{A'\cup D}\right) > \mathcal{D}\left(U/\ddot{\mathcal{R}}_A, U/\ddot{\mathcal{R}}_{A\cup D}\right)$.

Khác với entropy mờ trực cảm, tính chất phản đơn điệu được thể hiện giữa độ đo khoảng cách mờ trực cảm và kích thước của tập thuộc tính. Khi đó, độ quan trọng của thuộc tính cũng được xây dựng để lựa chọn những thuộc tính có ý nghĩa vào rút gọn.

$$Sig(a, D) = \mathcal{D}\left(U/\ddot{\mathcal{R}}_{A\cup\{a\}}, U/\ddot{\mathcal{R}}_{A\cup\{a\}\cup D}\right) - \mathcal{D}\left(U/\ddot{\mathcal{R}}_A, U/\ddot{\mathcal{R}}_{A\cup D}\right) \tag{1.27}$$

Cuối cùng, các tác giả đã xây dựng thuật toán heuristic để tìm rút gọn trên bảng quyết định. Bên cạnh những thuật toán được trình bày, luận án tổng hợp một số phương pháp rút gọn thuộc tính theo các cách tiếp cận và độ đo khác nhau trong Bảng 1.3.

Bảng 1.3: Một số phương pháp rút gọn thuộc tính trên mô hình tập thô mờ trực cảm

TT	Độ đo	Kiểu dữ liệu	Tiêu chuẩn đánh giá	Tài liệu
1	Miền dương mờ trực cảm	Hỗn hợp (mixed)	Lực lượng rút gọn, thời gian xử lý, độ chính xác phân lớp	[53, 54, 55, 56, 99, 100, 101, 102]
2	Entropy mờ trực cảm	Số (numeric)	Lực lượng rút gọn, thời gian xử lý, độ chính xác phân lớp	[57, 58, 97, 98]
3	Khoảng cách mờ trực cảm	Hỗn hợp (mixed)	Lực lượng rút gọn, thời gian xử lý, độ chính xác phân lớp	[59, 91]
4	Hạt tri thức mờ trực cảm	Hỗn hợp (mixed)	Lực lượng rút gọn, thời gian xử lý, độ chính xác phân lớp	[103]

Bây giờ, luận án sẽ trình bày một số ưu điểm của mô hình tập thô mờ trực cảm. Những ưu điểm này là cơ sở vững chắc trong việc phát triển các mô hình mở rộng nhằm khắc phục những hạn chế còn gặp phải trên các nhánh của lý thuyết tập thô.

- So với các mô hình NRS và FRS, mô hình IFRS có cấu trúc hạt tương đối chi tiết, phản ảnh rõ nét sự mơ hồ và do dự vốn tồn tại trong dữ liệu thực tế. Qua đó việc áp dụng các độ đo đánh giá thuộc tính trên các hạt thông tin này sẽ mang lại hiệu quả cao hơn.

- Sự bổ sung của thành phần hàm không thuộc trong hạt thông tin mờ trực cảm giúp điều chỉnh hiệu quả các thông tin bị ảnh hưởng bởi các đối tượng được sinh ra từ nhiễu. Nhờ vậy, các độ đo đánh giá thuộc tính, chủ yếu thông qua quá trình tính toán lực lượng, sẽ hạn chế tối đa sự đóng góp thông tin từ những đối tượng này. Do đó, rút gọn thu được sẽ cải thiện đáng kể về hiệu quả phân lớp.

- Một số kết quả thực nghiệm từ các nghiên cứu đã trình bày cho thấy tập rút gọn thu được bằng các phương pháp tiếp cận theo mô hình IFRSs thường đạt độ chính xác phân lớp cao hơn so với tập rút gọn từ các thuật toán dựa trên các mô hình tập thô mở rộng, đặc biệt trong các bộ dữ liệu có nhiễu và không nhất quán.

Tuy nhiên, qua quá trình khảo sát, luận án cũng thấy rằng mô hình IFRSs vẫn còn một số những hạn chế như sau:

- Chính sự bổ sung của thành phần độ không thuộc trong mô hình cũng khiến cho các thuật toán tốn kém về không gian lưu trữ và có thời gian xử lý chưa tối ưu, đặc biệt là khi áp dụng trên các bộ dữ liệu có số chiều lớn.

- Nhiều đối tượng nằm trong hạt thông tin mờ trực cảm với các mức độ thuộc rất nhỏ và mức độ không thuộc lớn có thể gây ra nhiễu và dư thừa trong quá trình tính toán. Khi đó, thông tin của các đối tượng này vẫn đóng góp vào các độ đo đánh giá thuộc tính mà chưa bị loại bỏ hoàn toàn.

Để hệ thống lại những nội dung đã trình bày, Bảng 1.4 trình bày tổng hợp một số ưu điểm và nhược điểm chính của các mô hình khi áp dụng trong các thuật toán rút gọn thuộc tính.

Như vậy, động lực nghiên cứu của luận án không chỉ dừng lại ở việc kế thừa và khai thác ưu điểm từ các mô hình IFS và IFRS để phát triển những mô hình mới nhằm nâng cao hiệu quả so với các mô hình thuộc hai nhánh của lý thuyết tập thô, mà còn hướng tới việc khắc phục những hạn chế của chính các mô hình IFS và IFRS về thời gian thực thi cũng như không gian lưu trữ phát sinh từ việc bổ sung thành phần độ không thuộc.

Một trong những vấn đề cần được quan tâm đó là số lượng đối tượng trong dữ liệu thường xuyên thay đổi, gây ra nhiều thách thức không chỉ cho các thuật toán rút gọn thuộc tính theo tiếp cận IFRS mà còn đối với các phương pháp dựa trên các mô hình khác. Trước những thách thức này, các phương pháp rút gọn thuộc tính theo hướng tiếp cận gia tăng đã trở thành một hướng nghiên cứu mở rộng và được quan tâm rất lớn.

1.4.2 Rút gọn thuộc tính trên bảng quyết định thay đổi

Dữ liệu trong các kịch bản thực tế thường có sự thay đổi và được cập nhật liên tục theo thời gian. Một ví dụ điển hình là bài toán chẩn đoán bệnh trong y học, nơi bác sĩ cần đánh giá các triệu chứng lâm sàng dựa trên hàng loạt chỉ số xét nghiệm. Khi đó, số lượng bệnh nhân liên tục gia tăng khiến cho việc xây dựng các mô hình phân lớp nhằm hỗ trợ quá trình chẩn đoán trở nên ngày càng phức tạp. Để xây dựng được các mô hình phân lớp hiệu quả, một yêu cầu đặt ra là phải giải quyết bài toán rút gọn thuộc tính trên các bảng quyết định có kích thước lớn và có sự thay đổi liên tục về tập đối tượng. Rõ ràng, đây được xem là hai thách thức lớn đối với các thuật toán rút gọn truyền thống. Trong trường hợp bảng quyết định có kích thước lớn, các thuật toán này

Bảng 1.4: Một số ưu điểm và nhược điểm chính của các mô hình

TT	Mô hình	Ưu điểm	Nhược điểm
1	Tập thô lân cận	Một số mở rộng của mô hình đã xét tới ảnh hưởng của từng thuộc tính đối với quyết định của các đối tượng. Thời gian xử lý tương đối hiệu quả và được áp dụng tốt trên các bộ dữ liệu có số chiều lớn.	Cấu trúc các hạt thông tin còn đơn giản, chưa phản ánh rõ vai trò của từng đối tượng trong hạt. Các độ đo thường chỉ xét số lượng đối tượng trong hạt, chưa xem xét đầy đủ các đối tượng khác.
2	Tập thô mờ	Thời gian tính toán hiệu quả. Các hạt thông tin được đặc trưng bởi độ giá trị độ thuộc của mỗi đối tượng	Chưa xem xét vai trò của các thuộc tính điều kiện. Hiệu quả xử lý còn hạn chế đối với các tập dữ liệu đạt có hiệu quả thấp trên các mô hình phân lớp.
3	Tập thô mờ trực cảm	Cấu trúc hạt thông tin tương đối chi tiết. Có khả năng cải thiện hiệu quả phân lớp trên các tập dữ liệu nhiễu.	Thời gian tính toán cao, chưa loại bỏ hoàn toàn ảnh hưởng của các đối tượng nhiễu.

thường gặp hạn chế nghiêm trọng về không gian lưu trữ và thời gian xử lý. Còn đối với bảng quyết định có sự thay đổi về tập đối tượng, các thuật toán này phải tái tính toán tập rút gọn trên mỗi giai đoạn thay đổi của bảng quyết định. Do đó, thời gian thực thi của các thuật toán là rất lớn và không thể đáp ứng những yêu cầu thực tế.

Để khắc phục những hạn chế trên, các phương pháp rút gọn thuộc tính theo hướng tiếp cận gia tăng đã được đề xuất và nghiên cứu. Với đặc điểm chỉ xử lý trên phần dữ liệu thay đổi thay vì toàn bộ tập dữ liệu, các thuật toán gia tăng giúp rút ngắn đáng kể thời gian thực thi so với các thuật toán theo hướng tiếp cận truyền thống. Trong tình huống bảng quyết định có kích thước lớn, dữ liệu được chia nhỏ thành nhiều phần, rút gọn khi đó được tìm kiếm riêng biệt trên từng phần trước khi tổng hợp để thu được

một rút gọn cuối cùng. Hướng tiếp cận này đã được triển khai trên nhiều mô hình khác nhau và mang lại kết quả ấn tượng, đặc biệt là trong việc tối ưu thời gian xử lý.

Một số phương pháp gia tăng rút gọn thuộc tính cũng đã được phát triển trên mô hình tập thô mờ thay thế cho mô hình tập thô truyền thống để xử lý trên các bảng quyết định chứa thuộc tính có giá trị số, liên tục. Theo đó, các phương pháp này thường tập trung vào việc thiết lập các độ đo nhằm xử lý nhanh trên các bảng quyết định có sự thay đổi tập đối tượng. Trong trường hợp bổ sung tập đối tượng, Liu và các cộng sự [76] đã phân tích một số khái niệm cơ bản của xấp xỉ dưới và miền dương mờ để phát triển một số cơ chế gia tăng. Tiếp theo, Yang và các cộng sự [77] xây dựng cơ chế cập nhật quan hệ phân biệt nhằm đề xuất hai thuật toán IV-FS-FRS-1 và IV-FS-FRS-2 tìm tập rút gọn trong trường hợp bổ sung tập đối tượng. Dựa trên việc quan sát sự thay đổi của quan hệ tương đương mờ khi một hoặc nhiều đối tượng được bổ sung vào bảng quyết định, Yang và các cộng sự [78] đã trình bày một góc nhìn sâu sắc về rút gọn thuộc tính thông qua việc thêm và loại bỏ các thuộc tính vào rút gọn được xác định trong giai đoạn trước. Từ việc biểu diễn các hạt thông tin dựa trên độ phủ thông tin, Zhang và các cộng sự [79] đề xuất thuật toán gia tăng AIFWAR tìm tập rút gọn sử dụng entropy có điều kiện mở rộng. Dựa trên khái niệm tập đối tượng chính, Ni và các cộng sự [80] xây dựng hai thuật toán gia tăng tìm tập rút gọn dựa trên tập đối tượng chính trong trường hợp bảng quyết định bổ sung tập đối tượng: thuật toán DIAR sử dụng hàm thuộc mờ và thuật toán PIAR sử dụng miền dương mờ.

Để giải quyết một cách toàn diện các trường hợp bảng quyết định có sự bổ sung và loại bỏ tập đối tượng, Giang và các cộng sự [81] đã phát triển hai công thức gia tăng, giúp tính toán nhanh chóng khoảng cách phân hoạch mờ cho những trường hợp này. Theo đó, các tác giả đã thiết kế hai thuật toán gia tăng và chứng minh tính hiệu quả của chúng so với các phương pháp truyền thống sử dụng mô hình tập thô. Gần đây, Xia trong [82] đã áp dụng cơ chế tăng tốc kết hợp với độ đo khoảng cách tri thức, từ đó đề xuất các thuật toán gia tăng cho mô hình tập thô mờ. Nhờ vào việc bỏ qua bước tính toán các quan hệ tương đương mờ và sử dụng một độ đo đơn giản, các thuật toán này đã chứng minh được hiệu quả vượt trội về thời gian thực thi. Về lý thuyết, các thuật toán gia tăng tìm rút gọn theo tiếp cận tập thô mờ nêu trên có thời gian thực hiện ngắn hơn đáng kể so với các thuật toán không gia tăng, đồng thời có khả năng thực thi hiệu quả trên các bảng dữ liệu có kích thước lớn. Tuy nhiên, giống như các phương pháp xử lý trên bảng quyết định cố định, các thuật toán gia tăng dựa trên mô hình tập thô và

tập thô mờ gặp khó khăn trong việc xử lý các bảng quyết định có hiệu quả phân lớp ban đầu thấp. Cụ thể, các rút gọn tìm được vẫn chưa tối ưu về kích thước và chưa cải thiện đáng kể hiệu quả phân lớp. Vì vậy, việc phát triển các thuật toán gia tăng theo mô hình tập thô mờ trực cảm đã mở ra một hướng đi đầy tiềm năng, hứa hẹn sẽ mang lại những cải tiến rõ rệt trong hiệu quả xử lý.

Gần đây, Anh và các cộng sự [91] đã mở rộng độ đo khoảng cách phân hoạch mờ trên không gian tập mờ trực cảm và xây dựng một thuật toán gia tăng nhằm tìm kiếm một rút gọn xấp xỉ hiệu quả trên bảng quyết định có sự bổ sung tập đối tượng. Thông qua một số kết quả thực nghiệm, các tác giả đã chứng minh được khả năng vượt trội của thuật toán đề xuất trên các tập dữ liệu có độ chính xác phân lớp ban đầu thấp. Có thể thấy rằng, các thuật toán gia tăng trên mô hình tập thô mờ trực cảm hiện nay vẫn còn ít được nghiên cứu và chưa được khai thác một cách đầy đủ, mặc dù các kết quả bước đầu cho thấy tiềm năng và hiệu quả ứng dụng là rất hứa hẹn. Do đó, việc phát triển các thuật toán trên các bảng quyết định cố định và bảng quyết định thay đổi dựa trên mô hình tập thô mờ trực cảm có thể xem là hướng nghiên cứu quan trọng và cần được khai thác.

1.5 Các phương pháp đánh giá hiệu quả thuật toán

Hiện nay, các thuật toán rút gọn thuộc tính thường được đánh giá dựa trên ba tiêu chí chính: kích thước của tập rút gọn thu được, độ chính xác phân lớp của tập rút gọn và thời gian thực thi của thuật toán. Tập rút gọn có kích thước càng nhỏ thì càng giúp giảm chi phí tính toán và nâng cao hiệu quả trong quá trình xây dựng các mô hình phân lớp. Độ chính xác phân lớp càng cao càng chứng tỏ tính hiệu quả của phương pháp rút gọn thuộc tính cũng như ưu điểm của mô hình phân lớp được áp dụng. Bên cạnh đó, thời gian thực thi càng ngắn thể hiện khả năng ứng dụng thực tế của thuật toán, đặc biệt đối với các tập dữ liệu có kích thước lớn. Do đó, mục tiêu chung của các thuật toán rút gọn thuộc tính là đạt được sự cân bằng tối ưu giữa cả ba tiêu chí nêu trên.

Theo nghiên cứu của Yuan và các cộng sự [92], các mô hình phân lớp thường được sử dụng để đánh giá hiệu quả của các phương pháp rút gọn thuộc tính bao gồm mô hình k láng giềng gần nhất (KNN), mô hình cây quyết định và mô hình máy vectơ hỗ trợ (SVM). Để đảm bảo tính khách quan và công bằng trong quá trình đánh giá, kỹ thuật kiểm định chéo k-fold thường được áp dụng. Kỹ thuật này chia tập dữ liệu một cách ngẫu nhiên thành k phần xấp xỉ bằng nhau; trong đó, một phần được sử dụng làm tập kiểm thử và $k-1$ phần còn lại dùng để huấn luyện mô hình. Quá trình này được lặp

lại k lần với các tập kiểm thử khác nhau, và kết quả phân lớp cuối cùng được xác định dựa trên trung bình các lần đánh giá.

Các bộ dữ liệu phục vụ cho quá trình đánh giá thuật toán chủ yếu được thu thập từ các kho dữ liệu uy tín như OpenML và UCI. Đây là những kho dữ liệu phong phú, bao phủ nhiều lĩnh vực khác nhau và được sử dụng rộng rãi trong các nghiên cứu về khai phá dữ liệu, trí tuệ nhân tạo và xử lý ảnh.

Trong nghiên cứu này, luận án sử dụng mô hình KNN để đánh giá độ chính xác phân lớp của các tập rút gọn, kết hợp với kỹ thuật kiểm định chéo 10-fold. Đồng thời, các bộ dữ liệu được tải từ OpenML và UCI cũng được sử dụng để thực hiện quá trình đánh giá hiệu quả của các thuật toán rút gọn thuộc tính được đề xuất.

1.6 Định hướng nghiên cứu của luận án

Dựa trên các vấn đề đã phân tích, đặc biệt là một số hạn chế còn tồn tại trên cả hai hướng mở rộng của lý thuyết tập thô, có thể thấy rằng các thuật toán dựa trên mô hình tập thô mờ trực cảm đã giải quyết được một số hạn chế từ hai hướng nghiên cứu này. Xuất phát từ sự bổ sung của thành phần độ không thuộc, mô hình này giúp điều chỉnh tính sai lệch của thông tin được tạo ra từ các đối tượng nhiễu để cải thiện hiệu quả trên các độ đo đánh giá thuộc tính. Tuy nhiên, mô hình này vẫn tồn tại một số hạn chế, điển hình như chưa thể loại bỏ hoàn toàn ảnh hưởng của các đối tượng nhiễu, khiến các độ đo vẫn phải thực hiện trên những giá trị được hình thành từ chúng. Hệ quả là các rút gọn thu được đôi khi vẫn chưa tối ưu cả về kích thước lẫn khả năng phân lớp. Bên cạnh đó, việc bổ sung thành phần độ không thuộc cũng làm gia tăng chi phí tính toán, dẫn đến hạn chế về thời gian thực thi của các thuật toán. Bên cạnh đó, để đáp ứng cho các kịch bản thực tế của dữ liệu khi có sự bổ sung và loại bỏ tập đối tượng, các thuật toán gia tăng dựa trên mô hình tập thô mờ và tập thô lân cận mặc dù đã được triển khai rộng rãi, tuy nhiên trên mô hình IFRS thì hướng tiếp cận này còn mới và chưa được biết tới, mặc dù hiệu quả mang lại là rất tiềm năng đối với các lớp dữ liệu nhiễu. Từ những khó khăn được trình bày, định hướng nghiên cứu của luận án bao gồm:

1) Kế thừa những ưu điểm của mô hình tập mờ trực cảm (IFS) để phát triển các mô hình mới, nhằm hạn chế những khó khăn từ các mở rộng theo FRS và NRS. Ngoài ra, các mô hình đề xuất cũng chứng minh được khả năng cải thiện cả những hạn chế của chính mô hình IFS cũng như IFRS.

2) Dựa trên các mô hình được xây dựng, nghiên cứu và phát triển một số thuật toán rút gọn thuộc tính nhằm xử lý trên các bảng quyết định cố định và bảng quyết định có

sự thay đổi tập đối tượng.

1.7 Kết luận Chương 1

Trong Chương 1, luận án đã trình bày khái quát về bài toán rút gọn thuộc tính, với mục tiêu giữ lại các thuộc tính then chốt và loại bỏ những thuộc tính dư thừa nhưng vẫn đảm bảo độ chính xác trong phân loại và dự đoán. Thêm vào đó, Chương 1 cũng đưa ra một số hướng tiếp cận chính của rút gọn thuộc tính và trình bày mốt số khái niệm cơ bản liên quan tới bảng quyết định. Từ một số hướng tiếp cận cho bài toán, một số nghiên cứu liên quan tới các nhánh mở rộng từ lý thuyết tập thô cũng được trình bày thông qua những khó khăn và hạn chế còn tồn tại. Trên cơ sở đó, lý thuyết IFS và mô hình IFRS được xem xét như một sự lựa chọn nền tảng để phát triển một số mô hình tiên tiến, góp phần khắc phục những hạn chế của các thuật toán rút gọn thuộc tính.

Với các động lực nghiên cứu được trình bày, Chương 2 tập trung đề xuất mô hình tập mờ trực cảm mức α, β và xây dựng các thuật toán rút gọn thuộc tính nhằm khắc phục các hạn chế của nhánh mô hình FRS. Cụ thể, bằng cách mở rộng khái niệm lát cắt α, β từ lý thuyết tập mờ, mô hình đề xuất cho phép giảm thiểu ảnh hưởng của các đối tượng có độ thuộc nhỏ và độ không thuộc lớn trong các hạt thông tin mờ trực cảm. Những đối tượng này thường có nguồn gốc từ nhiễu và là nguyên nhân làm suy giảm hiệu quả phân lớp trong các mô hình trước đây. Việc loại bỏ hoặc làm giảm ảnh hưởng của chúng không chỉ giúp cải thiện chất lượng rút gọn mà còn thu hẹp không gian tính toán, từ đó giảm thời gian thực thi của các thuật toán. Dựa trên mô hình tập mờ trực cảm mức α, β, luận án tiếp tục thiết kế các thuật toán rút gọn thuộc tính cho cả bảng quyết định cố định và bảng quyết định thay đổi, hướng tới các kịch bản dữ liệu thực tế. Các thuật toán đề xuất được đánh giá thông qua quá trình thực nghiệm so sánh với nhiều phương pháp theo các hướng tiếp cận khác nhau, nhằm chứng minh các rút gọn thu được đạt hiệu quả vượt trội trên các bộ dữ liệu tiêu chuẩn.

CHƯƠNG 2
ĐỀ XUẤT MỘT SỐ THUẬT TOÁN RÚT GỌN THUỘC TÍNH DỰA TRÊN TẬP MỜ TRỰC CẢM MỨC α, β

2.1 Mở đầu

Chương 1 đã trình bày một số ưu điểm của mô hình tập thô mờ trực cảm khi ứng dụng vào bài toán rút gọn thuộc tính trên các bảng quyết định nhiễu hoặc ko nhất quán, qua đó thu được những kết quả ấn tượng so với các phương pháp dựa trên mô hình tập thô và tập thô mờ. Có thể nói rằng hiệu quả của mô hình này xuất phát từ sự bổ sung của thành phần độ không thuộc, giúp điều chỉnh thông tin từ các đối tượng nhiễu trong dữ liệu, vốn được xem là yếu tố chính làm giảm hiệu quả trên các mô hình phân lớp. Tuy nhiên, cơ chế này chưa thực sự loại bỏ hoàn toàn tác động của các đối tượng nhiễu đến một số độ đo đánh giá thuộc tính. Cụ thể, những đối tượng có độ thuộc nhỏ và độ không thuộc lớn vẫn đóng góp thông tin các độ đo đánh giá thuộc tính, dẫn đến việc các tập rút gọn thu được chưa thực sự tối ưu cả về kích thước lẫn hiệu quả phân lớp. Bên cạnh đó, việc bổ sung thành phần độ không thuộc trong các mô hình tập thô mờ trực cảm làm gia tăng đáng kể không gian lưu trữ cũng như chi phí tính toán, đặc biệt đối với các tập dữ liệu có kích thước lớn hoặc có sự thay đổi theo thời gian. Diều này khiến cho các thuật toán rút gọn thuộc tính gặp rất nhiều khó khăn khi thực thi trên các bộ dữ liệu có kích thước lớn và có sự thay đổi liên tục. Rõ ràng, những hạn chế này tiếp tục đặt ra những thách thức mới cho nhánh mở rộng thứ hai của lý thuyết tập thô. Để khắc phục những hạn chế được trình bày, nội dung của luận án trong Chương 2 sẽ trình bày các đóng góp chính như sau:

Thứ nhất, luận án đề xuất mô hình tập mờ trực cảm mức α, β và phân tích một số tính chất quan trọng của mô hình này. Trên cơ sở đó, luận án xây dựng độ đo khoảng cách giữa hai phân hoạch mờ trực cảm mức α, β nhằm định nghĩa lại một rút gọn mới, đồng thời đề xuất thuật toán rút gọn thuộc tính cho bảng quyết định cố định theo hướng tiếp cận lọc.

Thứ hai, luận án phát triển hai công thức gia tăng dựa trên độ đo khoảng cách phân hoạch mờ trực cảm mức α, β để áp dụng cho các trường hợp bảng quyết định thay đổi. Qua đó, luận án đề xuất hai thuật toán rút gọn thuộc tính theo tiếp cận gia tăng nhằm xử lý trong các trường hợp bảng quyết định có sự bổ sung và loại bỏ tập đối tượng.

Thứ ba, luận án chứng minh hiệu quả các thuật toán đề xuất dựa trên việc so sánh

với một số thuật toán theo mô hình tập thô mờ, tập thô mờ trực cảm.

Kết quả nghiên cứu của chương này được công bố trong các công trình [CT1], [CT2], [CT4] và [CT5] thuộc phần Danh mục các công trình nghiên cứu của luận án.

2.2 Mô hình tập mờ trực cảm mức α, β

Trong phần này, luận án sẽ giới thiệu một mô hình mới được mở rộng từ lý thuyết tập mờ trực cảm được gọi là mô hình tập mờ trực cảm mức α, β. Qua đó, một số tính chất của mô hình cũng được trình bày làm cơ sở đề xuất các thuật toán rút gọn thuộc tính.

2.2.1 Khái niệm về tập mờ trực cảm mức α, β

Luận án bắt đầu bằng việc trình bày một khái niệm quan trọng được mở rộng từ khái niệm tập lát cắt α trong lý thuyết tập mờ, được gọi là tập lát cắt α, β. Theo đó, tập lát cắt α, β được xác định bằng cách giữ lại các đối tượng trong một tập mờ trực cảm cho trước có độ thuộc không nhỏ hơn α và và độ không thuộc không lớn hơn β. Dựa trên ý tưởng này, luận án áp dụng khái niệm tập lát cắt α, β trên các hạt thông tin mờ trực cảm, vốn cũng được xem là một tập mờ trực cảm để hình thành một cấu trúc hạt thông tin mới có khả năng tổng quát hóa tốt hơn.

Định nghĩa 2.1 [93] Xét bảng quyết định $IS = (U, C \cup D)$ với $\ddot{\mathcal{R}}_{\{a\}}$ là một quan hệ tương đương mờ trực cảm được hình thành bởi thuộc tính điều kiện $a \in C$ và $u \in U$ là một đối tượng trong tập vũ trụ. Khi đó, $[\ddot{u}]_{\{a\}}$ là một hạt thông tin mờ trực cảm của đối tượng u theo thuộc tính a. Với α và β là hai số thực nằm trong khoảng $[0, 1]$ thỏa mãn $\alpha + \beta \leq 1$, tập lát cắt α, β của $[\ddot{u}]_{\{a\}}$ là một tập rõ được định nghĩa như sau:

$$[u]^{\alpha,\beta}_{\{a\}} = \left\{v \in U : \gamma_{[\ddot{u}]_{\{a\}}}(v) \geq \alpha \wedge \eta_{[\ddot{u}]_{\{a\}}}(v) \leq \beta\right\} \tag{2.1}$$

Theo Định nghĩa 2.1, có thể nhận thấy rằng tập lát cắt α, β cho phép loại bỏ những đối tượng có độ thuộc thấp và độ không thuộc cao trong các hạt thông tin mờ trực cảm. Đây là các đối tượng được sinh ra bởi nhiễu, có phân bố khác biệt so với phần lớn các đối tượng trong tập vũ trụ và gây ảnh hưởng lớn đến hiệu quả phân lớp. Do đó, việc loại bỏ các đối tượng này giúp các độ đo đánh giá thuộc tính không bị chi phối bởi những thông tin sai lệch, từ đó góp phần lựa chọn được các thuộc tính thực sự có ý nghĩa.

Tiếp theo, một tập mờ trực cảm mức α, β, ký hiệu là $[\ddot{u}]^{\alpha,\beta}_{\{a\}}$ dựa trên mỗi phần tử của tập lát cắt α, β với độ tương tự và độ khác biệt của mỗi đối tượng được xác định

như sau:

$$[\ddot{u}]^{\alpha,\beta}_{\{a\}}(v) = \left(\gamma_{[\ddot{u}]^{\alpha,\beta}_{\{a\}}}(v), \eta_{[\ddot{u}]^{\alpha,\beta}_{\{a\}}}(v)\right) = \begin{cases} \left(\gamma_{[\ddot{u}]_{\{a\}}}(v), \eta_{[\ddot{u}]_{\{a\}}}(v)\right), v \in [u]^{\alpha,\beta}_{\{a\}} \\ (0,1), v \notin [u]^{\alpha,\beta}_{\{a\}} \end{cases} \tag{2.2}$$

Nếu thuộc tính a là kiểu thuộc tính mang giá trị số liên tục, giá trị độ tương tự của đối tượng v trong hạt thông tin mờ trực cảm của đối tượng u được xác định theo Công thức 1.17. Đối với việc tính toán độ khác biệt, luận án sử dụng công thức được đưa ra trong công trình [91].

$$\eta_{[\ddot{u}]_{\{a\}}}(v) = \frac{1 - \eta_{[\ddot{u}]_{\{a\}}}(v)}{1 + \upsilon(a) \times \eta_{[\ddot{u}]_{\{a\}}}(v)} \tag{2.3}$$

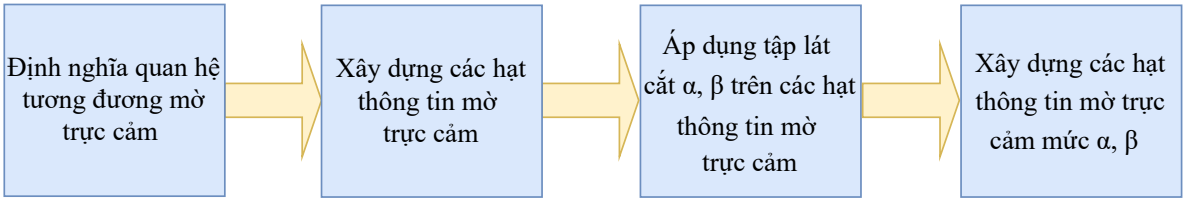
trong đó, $\upsilon(a) \geq 0$. Nếu $\upsilon(a) = 0$ thì toàn bộ các hạt thông tin mờ trực cảm sẽ được biểu diễn như một tập mờ truyền thống. Trong trường hợp còn lại, độ khác biệt và độ tương tự sẽ tỉ lệ nghịch với nhau và thỏa mãn các tính chất của tập mờ trực cảm. Giá trị $\upsilon(a)$ được xác định bởi công thức sau đây:

$$\upsilon(a) = \begin{cases} 1, & \sigma(a) = 0 \\ \dfrac{\left|U/\widetilde{\mathcal{R}}_{\{a\}} \cap U/\widetilde{\mathcal{R}}_D\right|}{\left|U/\widetilde{\mathcal{R}}_D\right| \times \sigma(a)}, & \sigma(a) > 0 \end{cases} \tag{2.4}$$

trong đó, $\sigma(a)$ là độ lệch chuẩn được xác định trên miền giá trị thuộc tính a.

Nếu thuộc tính a mang giá trị rời rạc hoặc có dạng ký tự thì độ tương tự của hai đối tượng u và v được xác định theo Công thức 1.18. Rõ ràng, $[\ddot{u}]^{a,\beta}_{\{a\}}(v)$ được hình thành từ việc điều chỉnh các số mờ trực cảm trong $[\ddot{u}]_{\{a\}}$. Các số mờ được điều chỉnh sẽ có độ tương tự nhỏ hơn hoặc bằng α hoặc độ khác biệt lớn hơn hoặc bằng β. Theo đó, nếu một đối tượng bị loại bỏ thì độ tương tự và độ khác biệt sẽ được biểu diễn trong $[\ddot{u}]^{a,\beta}_{\{a\}}$ dưới dạng một phần tử là (0,1). Trong nghiên cứu này, $[\ddot{u}]^{a,\beta}_{\{a\}}$ được gọi là một hạt thông tin mờ trực cảm mức α, β của đối tượng u và cũng được xem như một tập mờ trực cảm mức α, β (α, β-IFS). Rõ ràng, α, β-IFS giúp loại bỏ thông tin từ các đối tượng không nằm trong tập lát cắt α, β và ngăn các đối tượng này đóng góp thông tin vào các độ đo đánh giá thuộc tính. Hình 2.1 trình bày tổng thể về quá trình xây dựng các hạt thông tin mờ trực cảm mức α, β.

Từ các khái niệm về hạt thông tin mờ trực cảm mức α, β, nếu xét trên toàn bộ các đối tượng trong tập vũ trụ, một họ $\left\{[\ddot{u}]^{a,\beta}_{\{a\}} : u \in U\right\}$ sẽ tạo ra một phân hoạch mờ trực cảm trên U. Để tránh nhầm lẫn, họ này sẽ được ký hiệu là $U/\ddot{\mathcal{R}}^{\alpha,\beta}_{\{a\}}$ và được gọi là một phân hoạch mờ trực cảm mức α, β.



Hình 2.1: Quá trình xây dựng các hạt thông tin mờ trực cảm mức α, β

Giả sử rằng, có hai phân hoạch mờ trực cảm mức α, β là $U/\ddot{\mathcal{R}}_A^{\alpha,\beta}$ và $U/\ddot{\mathcal{R}}_B^{\alpha,\beta}$ được tạo ra bởi các quan hệ tương đương mờ trực cảm từ hai tập con thuộc tính điều kiện $A, B \subseteq C$. Khi đó, chúng ta nói rằng $U/\ddot{\mathcal{R}}_A^{\alpha,\beta}$ mịn hơn $U/\ddot{\mathcal{R}}_B^{\alpha,\beta}$, ký hiệu là $U/\ddot{\mathcal{R}}_A^{\alpha,\beta} \preceq U/\ddot{\mathcal{R}}_B^{\alpha,\beta}$, nếu với mọi đối tượng $u \in U$, $[\ddot{u}]_A^{\alpha,\beta} \subseteq [\ddot{u}]_B^{\alpha,\beta}$.

Trong phần sau, luận án sẽ trình bày một số tính chất thiết yếu của hạt thông tin và phân hoạch mờ trực cảm mức α, β.

Ví dụ 2.1 Cho $\alpha = 0.75$ và $\beta = 0.15$, quan hệ tương đương mờ $\widetilde{\mathcal{R}}_{\{a\}}$ của mỗi thuộc tính $a \in C$ được xác định theo Công thức 1.17.

Xét thuộc tính a_1, chúng ta dễ dàng tính được độ lệch chuẩn $\sigma(a_1) = 0.287$, các phân hoạch mờ trên thuộc tính a_1 và thuộc tính quyết định.

$$U/\widetilde{\mathcal{R}}_D = \begin{bmatrix} 1.00 & 0.00 & 1.00 & 0.00 & 1.00 \\ 0.00 & 1.00 & 0.00 & 1.00 & 0.00 \\ 1.00 & 0.00 & 1.00 & 0.00 & 1.00 \\ 0.00 & 1.00 & 0.00 & 1.00 & 0.00 \\ 1.00 & 0.00 & 1.00 & 0.00 & 1.00 \end{bmatrix}, U/\widetilde{\mathcal{R}}_{\{a_1\}} = \begin{bmatrix} 1.00 & 0.71 & 0.49 & 0.77 & 0.73 \\ 0.71 & 1.00 & 0.78 & 0.48 & 0.98 \\ 0.49 & 0.78 & 1.00 & 0.26 & 0.76 \\ 0.77 & 0.48 & 0.26 & 1.00 & 0.50 \\ 0.73 & 0.98 & 0.76 & 0.50 & 1.00 \end{bmatrix}$$

Theo Công thức 2.4, $\upsilon(a_1) = 2.661$. Xét đối tượng u_1, chúng ta tính được độ khác biệt của các đối tượng trong hạt thông tin mờ trực cảm $[\ddot{u}_1]_{\{a_1\}}$ theo Công thức 2.3:

$$\eta_{[\ddot{u}_1]_{\{a_1\}}}(u_2) = \frac{1 - 0.71}{1 + 2.661 \times 0.71} = 0.10, \eta_{[\ddot{u}_1]_{\{a_1\}}}(u_3) = \frac{1 - 0.49}{1 + 2.661 \times 0.49} = 0.22,$$
$$\eta_{[\ddot{u}_1]_{\{a_1\}}}(u_4) = \frac{1 - 0.77}{1 + 2.661 \times 0.77} = 0.08, \eta_{[\ddot{u}_1]_{\{a_1\}}}(u_5) = \frac{1 - 0.73}{1 + 2.661 \times 0.73} = 0.09.$$

Do đó, hạt thông tin mờ trực cảm của đối tượng u_1 trên thuộc tính a_1 là:

$$[\ddot{u}_1]_{\{a_1\}} = \left\{ \frac{(1.00, 0.00)}{u_1}, \frac{(0.71, 0.10)}{u_2}, \frac{(0.49, 0.22)}{u_3}, \frac{(0.77, 0.08)}{u_4}, \frac{(0.73, 0.09)}{u_5} \right\}.$$

Tập lát cắt α, β của $[\ddot{u}_1]_{\{a_1\}}$ là $[u_1]_{\{a_1\}}^{\alpha,\beta} = \{u_1, u_2, u_4, u_5\}$. Khi đó, chúng ta thu được hạt thông tin mờ trực cảm mức *alpha, beta* của đối tượng u_1 trên thuộc tính a_1:

$$[\ddot{u}_1]_{\{a_1\}}^{\alpha,\beta} = \left\{ \frac{(1.00, 0.00)}{u_1}, \frac{(0.71, 0.10)}{u_2}, \frac{(0.00, 1.00)}{u_3}, \frac{(0.77, 0.08)}{u_4}, \frac{(0.73, 0.09)}{u_5} \right\}$$

Thực hiện tương tự, chúng ta cũng thu được các hạt thông tin mờ trực cảm mức *alpha, beta* còn lại trên thuộc tính a_1.

$$[\ddot{u}_2]^{\alpha,\beta}_{\{a_1\}} = \left\{\frac{(0.71, 0.10)}{u_1}, \frac{(1.00, 0.00)}{u_2}, \frac{(0.78, 0.07)}{u_3}, \frac{(0.00, 1.00)}{u_4}, \frac{(0.98, 0.01)}{u_5}\right\}$$
$$[\ddot{u}_3]^{\alpha,\beta}_{\{a_1\}} = \left\{\frac{(0.00, 1.00)}{u_1}, \frac{(0.78, 0.07)}{u_2}, \frac{(1.00, 0.00)}{u_3}, \frac{(0.00, 1.00)}{u_4}, \frac{(0.76, 0.08)}{u_5}\right\}$$
$$[\ddot{u}_4]^{\alpha,\beta}_{\{a_1\}} = \left\{\frac{(0.77, 0.08)}{u_1}, \frac{(0.00, 1.00)}{u_2}, \frac{(0.00, 1.00)}{u_3}, \frac{(1.00, 0.00)}{u_4}, \frac{(0.00, 1.00)}{u_5}\right\}$$
$$[\ddot{u}_5]^{\alpha,\beta}_{\{a_1\}} = \left\{\frac{(0.73, 0.09)}{u_1}, \frac{(0.98, 0.01)}{u_2}, \frac{(0.76, 0.08)}{u_3}, \frac{(0.00, 1.00)}{u_4}, \frac{(1.00, 0.00)}{u_5}\right\}$$

Tương tự, chúng ta cũng thu được các phân hoạch mờ trực cảm mức *alpha, beta* trên các thuộc tính còn lại.

2.2.2 Các tính chất của hạt thông tin mờ trực cảm mức alpha, beta

Mệnh đề 2.1 Cho bảng quyết định $IS = (U, C \cup D)$

1. Nếu $A \subseteq B \subseteq C$ thì $[u]^{\alpha,\beta}_B \subseteq [u]^{\alpha,\beta}_A \subseteq U$,
2. $\forall A \subseteq C$, $[\ddot{u}]^{\alpha,\beta}_A \subseteq [\ddot{u}]_A$,
3. $\forall A, B \subseteq C$, $[\ddot{u}]^{\alpha,\beta}_{A\cup B} = [\ddot{u}]^{\alpha,\beta}_A \cap [\ddot{u}]^{\alpha,\beta}_B$.

Chứng minh. Các tính chất 1 và 2 có thể dễ dàng chứng minh. Đối với tính chất 3, chúng ta xem xét hai trường hợp sau đây:

Trường hợp đầu tiên, với $v \in [u]^{\alpha,\beta}_{A\cup B}$ thì $\gamma_{[\ddot{u}]_{A\cup B}}(v) \geq \alpha$ và $\eta_{[\ddot{u}]_{A\cup B}}(v) \leq \beta$. Điều này dẫn tới $\min\left\{\gamma_{[\ddot{u}]^{\alpha,\beta}_A}(v), \gamma_{[\ddot{u}]^{\alpha,\beta}_B}(v)\right\} \geq \alpha$ và $\max\left\{\eta_{[\ddot{u}]^{\alpha,\beta}_A}(v), \eta_{[\ddot{u}]^{\alpha,\beta}_B}(v)\right\} \leq \beta$. Do đó, $v \in [u]^{\alpha,\beta}_A \cap [u]^{\alpha,\beta}_B$ (1). Tương tự, chúng ta cũng dễ dàng chứng minh được $v \in [u]^{\alpha,\beta}_{A\cup B}$ với trường hợp $v \in [u]^{\alpha,\beta}_A \cap [u]^{\alpha,\beta}_B$ (2).

Từ (1) và (2), $[u]^{\alpha,\beta}_{A\cup B} = [u]^{\alpha,\beta}_A \cap [u]^{\alpha,\beta}_B$. Do đó, $[\ddot{u}]^{\alpha,\beta}_{A\cup B} = [\ddot{u}]^{\alpha,\beta}_A \cap [\ddot{u}]^{\alpha,\beta}_B$. □

Rõ ràng, $[\ddot{u}]^{0,1}_A = [\ddot{u}]_A$. Do đó, tính chất 2 của Mệnh đề 2.1 chỉ ra rằng $[\ddot{u}]_A$ là một trường hợp đặc biệt của $[\ddot{u}]^{\alpha,\beta}_A$, nghĩa là $[\ddot{u}]^{\alpha,\beta}_A$ mang đầy đủ các tính chất của $[\ddot{u}]_A$. Hạt thông tin $[\ddot{u}]^{\alpha,\beta}_A$ được đặc trưng bởi các đối tượng thỏa mãn các điều kiện nằm trong tập lát cắt α, β. Đối với các đối tượng còn lại sẽ không có sự đóng góp nhiều vào các độ đo nên có thể bị loại bỏ. Đây là những đối tượng được đặc trưng bởi độ tương tự nhỏ và độ khác biệt lớn. Về mặt trực cảm, những đối tượng này được hình thành bởi nhiễu do dữ liệu và chúng có sự phân bố khác biệt so với phần lớn các đối tượng khác trong tập vũ trụ. Do đó, nếu các độ đo phải thực hiện trên các đối tượng này thì việc đánh giá thuộc tính sẽ bị ảnh hưởng và gặp nhiều khó khăn để thu được một rút gọn hiệu quả.

Ví dụ 2.2 Dựa trên Mệnh đề 2.1, với mọi $u \in U$, chúng ta có $[\ddot{u}]_C^{\alpha,\beta} = \cap_{a\in C}[\ddot{u}]_{\{a\}}^{\alpha,\beta}$. Khi đó, các hạt thông tin của phân hoạch mờ trực cảm mức *alpha, beta* theo C bao gồm:

$$[\ddot{u}_1]_C^{\alpha,\beta} = \left\{\frac{(1.00, 0.00)}{u_1}, \frac{(0.71, 0.10)}{u_2}, \frac{(0.00, 1.00)}{u_3}, \frac{(0.00, 1.00)}{u_4}, \frac{(0.00, 1.00)}{u_5}\right\}$$
$$[\ddot{u}_2]_C^{\alpha,\beta} = \left\{\frac{(0.71, 0.10)}{u_1}, \frac{(1.00, 0.00)}{u_2}, \frac{(0.00, 1.00)}{u_3}, \frac{(0.00, 1.00)}{u_4}, \frac{(0.00, 1.00)}{u_5}\right\}$$
$$[\ddot{u}_3]_C^{\alpha,\beta} = \left\{\frac{(0.00, 1.00)}{u_1}, \frac{(0.00, 1.00)}{u_2}, \frac{(1.00, 0.00)}{u_3}, \frac{(0.00, 1.00)}{u_4}, \frac{(0.00, 1.00)}{u_5}\right\}$$
$$[\ddot{u}_4]_C^{\alpha,\beta} = \left\{\frac{(0.00, 1.00)}{u_1}, \frac{(0.00, 1.00)}{u_2}, \frac{(0.00, 1.00)}{u_3}, \frac{(1.00, 0.00)}{u_4}, \frac{(0.00, 1.00)}{u_5}\right\}$$
$$[\ddot{u}_5]_C^{\alpha,\beta} = \left\{\frac{(0.00, 1.00)}{u_1}, \frac{(0.00, 1.00)}{u_2}, \frac{(0.00, 1.00)}{u_3}, \frac{(0.00, 1.00)}{u_4}, \frac{(1.00, 0.00)}{u_5}\right\}$$

Mệnh đề 2.2 Cho bảng quyết định $IS = (U, C \cup D)$ và một tập con thuộc tính điều kiện $A \subseteq C$, nếu $\alpha_1 \leq \alpha_2$ và $\beta_1 \geq \beta_2$, thì $[\ddot{u}]_A^{\alpha_2,\beta_2} \subseteq [\ddot{u}]_A^{\alpha_1,\beta_1}$.

Chứng minh. Dễ thấy rằng với mọi $u \in U$, $\alpha_1 \leq \alpha_2$ và $\beta_1 \geq \beta_2$, chúng ta luôn có $[u]_A^{\alpha_2,\beta_2} \subseteq [u]_A^{\alpha_1,\beta_1}$. Khi đó, chúng ta xem xét ba trường hợp sau đây:

Xét trường hợp $v \in [u]_A^{\alpha_2,\beta_2}$, khi đó $v \in [u]_A^{\alpha_1,\beta_1}$. Chúng ta dễ dàng thu được $\gamma_{[\ddot{u}]_A^{\alpha_1,\beta_1}}(v) = \gamma_{[\ddot{u}]_A^{\alpha_2,\beta_2}}(v) = \gamma_{[\ddot{u}]_A}(v)$ và $\eta_{[\ddot{u}]_A^{\alpha_1,\beta_1}}(v) = \eta_{[\ddot{u}]_A^{\alpha_2,\beta_2}}(v) = \eta_{[\ddot{u}]_A}(v)$.

Xét trường hợp $v \in U \setminus [u]_A^{\alpha_1,\beta_1}$, khi đó $v \notin [u]_A^{\alpha_1,\beta_1}$ và $v \notin [u]_A^{\alpha_2,\beta_2}$. Điều này dẫn đến $\gamma_{[\ddot{u}]_A^{\alpha_1,\beta_1}}(v) = \gamma_{[\ddot{u}]_A^{\alpha_2,\beta_2}}(v) = 0$ và $\eta_{[\ddot{u}]_A^{\alpha_1,\beta_1}}(v) = \eta_{[\ddot{u}]_A^{\alpha_2,\beta_2}}(v) = 1$.

Xét trường hợp $v \in [u]_A^{\alpha_1,\beta_1} \setminus [u]_A^{\alpha_2,\beta_2}$, khi đó $[\ddot{u}]_A^{\alpha_2,\beta_2}(v) = (0, 1)$ và $[\ddot{u}]_A^{\alpha_1,\beta_1}(v) = [\ddot{u}]_A(v) = \left(\gamma_{[\ddot{u}]_A}(v), \eta_{[\ddot{u}]_A}(v)\right)$ với $\gamma_{[\ddot{u}]_A}(v) \geq 0$ và $\eta_{[\ddot{u}]_A}(v) \leq 1$.

Từ ba trường hợp trên, $\gamma_{[\ddot{u}]_A^{\alpha_2,\beta_2}}(v) \leq \gamma_{[\ddot{u}]_A^{\alpha_1,\beta_1}}(v)$ và $\eta_{[\ddot{u}]_A^{\alpha_2,\beta_2}}(v) \geq \eta_{[\ddot{u}]_A^{\alpha_1,\beta_1}}(v)$, với mọi $v \in U$. Do đó, $[\ddot{u}]_A^{\alpha_2,\beta_2} \subseteq [\ddot{u}]_A^{\alpha_1,\beta_1}$, với mọi $u \in U$. Mệnh đề đã được chứng minh. □

Mệnh đề 2.3 Cho bảng quyết định $IS = (U, C \cup D)$ và $[\ddot{u}]_A^{\alpha,\beta}$, $[\ddot{u}]_B^{\alpha,\beta}$, $[\ddot{u}]_E^{\alpha,\beta}$ là các hạt thông tin mờ trực cảm mức *alpha, beta* của đối tượng $u \in U$ được tạo bởi các tập con thuộc tính điều kiện $A, B, E \subseteq C$. Khi đó, chúng ta có một số tính chất sau:
1. Nếu $[\ddot{u}]_A^{\alpha,\beta} \subseteq [\ddot{u}]_B^{\alpha,\beta}$ thì $\left|[\ddot{u}]_B^{\alpha,\beta}\right| - \left|[\ddot{u}]_{B\cup E}^{\alpha,\beta}\right| \geq \left|[\ddot{u}]_A^{\alpha,\beta}\right| - \left|[\ddot{u}]_{A\cup E}^{\alpha,\beta}\right|$,
2. $\left|[\ddot{u}]_A^{\alpha,\beta}\right| - \left|[\ddot{u}]_{A\cup B}^{\alpha,\beta}\right| + \left|[\ddot{u}]_E^{\alpha,\beta}\right| - \left|[\ddot{u}]_{E\cup A}^{\alpha,\beta}\right| \geq \left|[\ddot{u}]_E^{\alpha,\beta}\right| - \left|[\ddot{u}]_{E\cup B}^{\alpha,\beta}\right|$.

Chứng minh. 1. Từ $[\ddot{u}]_A^{\alpha,\beta} \subseteq [\ddot{u}]_B^{\alpha,\beta}$ với mọi $v \in U$, chúng ta luôn có $\gamma_{[\ddot{u}]_B^{\alpha,\beta}}(v) \geq \gamma_{[\ddot{u}]_A^{\alpha,\beta}}(v)$ và $\eta_{[\ddot{u}]_B^{\alpha,\beta}}(v) \leq \eta_{[\ddot{u}]_A^{\alpha,\beta}}(v)$. Điều này dẫn tới $\gamma_{[\ddot{u}]_B^{\alpha,\beta}}(v) - \gamma_{[\ddot{u}]_{B\cup E}^{\alpha,\beta}}(v) \geq \gamma_{[\ddot{u}]_A^{\alpha,\beta}}(v) - \gamma_{[\ddot{u}]_{A\cup E}^{\alpha,\beta}}(v)$ và $\eta_{[\ddot{u}]_A^{\alpha,\beta}}(v) - \eta_{[\ddot{u}]_{A\cup E}^{\alpha,\beta}}(v) \geq \eta_{[\ddot{u}]_B^{\alpha,\beta}}(v) - \eta_{[\ddot{u}]_{B\cup E}^{\alpha,\beta}}(v)$. Do đó, chúng ta dễ dàng thu được:

$$\sum_{v\in U}\gamma_{[\ddot{u}]_B^{\alpha,\beta}}(v) - \sum_{v\in U}\gamma_{[\ddot{u}]_{B\cup E}^{\alpha,\beta}}(v) \geq \sum_{v\in U}\gamma_{[\ddot{u}]_A^{\alpha,\beta}}(v) - \sum_{v\in U}\gamma_{[\ddot{u}]_{A\cup E}^{\alpha,\beta}}(v)$$

$$\sum_{v\in U}\eta_{[\ddot{u}]_A^{\alpha,\beta}}(v)-\sum_{v\in U}\eta_{[\ddot{u}]_{A\cup E}^{\alpha,\beta}}(v)\geq\sum_{v\in U}\eta_{[\ddot{u}]_B^{\alpha,\beta}}(v)-\sum_{v\in U}\eta_{[\ddot{u}]_{B\cup E}^{\alpha,\beta}}(v)$$

Từ kết quả này, tính chất 1 của mệnh đề được chứng minh.

2. Từ tính chất 1, chúng ta có:
$\left|[\ddot{u}]_A^{\alpha,\beta}\right|-\left|[\ddot{u}]_{A\cup B}^{\alpha,\beta}\right|\geq\left|[\ddot{u}]_{A\cup E}^{\alpha,\beta}\right|-\left|[\ddot{u}]_{A\cup E\cup B}^{\alpha,\beta}\right|$ và $\left|[\ddot{u}]_E^{\alpha,\beta}\right|-\left|[\ddot{u}]_{E\cup A\cup B}^{\alpha,\beta}\right|\geq\left|[\ddot{u}]_E^{\alpha,\beta}\right|-\left|[\ddot{u}]_{E\cup B}^{\alpha,\beta}\right|$.
Do đó, tính chất 2 của mệnh đề được chứng minh. □

2.3 Đề xuất thuật toán rút gọn thuộc tính dựa trên tập mờ trực cảm mức α,β

Trong phần trước, luận án đã giới thiệu mô hình tập mờ trực cảm mức $alpha, beta$ cùng một số tính chất then chốt. Trên cơ sở hình thành mô hình này, phần tiếp theo sẽ tập trung trình bày các thuật toán rút gọn thuộc tính và ứng dụng chúng vào một số trường hợp thực tế của bảng quyết định.

2.3.1 Đề xuất thuật toán rút gọn thuộc tính trên bảng quyết định cố định

Đầu tiên, cho bảng quyết định $IS=(U,C\cup D)$, hai hạt thông tin mờ trực cảm mức $alpha, beta$ là $[\ddot{u}]_A^{\alpha,\beta}$ và $[\ddot{u}]_B^{\alpha,\beta}$ được tạo ra bởi các tập con thuộc tính $A,B\subseteq C$, khoảng cách giữa $[\ddot{u}]_A^{\alpha,\beta}$ và $[\ddot{u}]_B^{\alpha,\beta}$ ký hiệu là $\ddot{\mathcal{D}}\left([\ddot{u}]_A^{\alpha,\beta},[\ddot{u}]_B^{\alpha,\beta}\right)$, được xác định theo công thức:

$$\ddot{\mathcal{D}}\left([\ddot{u}]_A^{\alpha,\beta},[\ddot{u}]_B^{\alpha,\beta}\right)=\frac{\left|[\ddot{u}]_A^{\alpha,\beta}\cup[\ddot{u}]_B^{\alpha,\beta}\right|-\left|[\ddot{u}]_A^{\alpha,\beta}\cap[\ddot{u}]_B^{\alpha,\beta}\right|}{|U|}\tag{2.5}$$

Độ đo khoảng cách trong Công thức 2.5 đo lường sự khác biệt về mặt thông tin giữa hai hạt thông tin mờ trực cảm mức $alpha, beta$. Với tính đơn giản, nó có tiềm năng giảm đáng kể thời gian tính toán cho các phương pháp rút gọn thuộc tính.

Mệnh đề 2.4 Cho bảng quyết định $IS=(U,C\cup D)$ với $A,B\subseteq C$ và hai hạt thông tin mờ trực cảm mức *alpha, beta* là $[\ddot{u}]_A^{\alpha,\beta}$, $[\ddot{u}]_B^{\alpha,\beta}$. Khi đó, chúng ta có một số tính chất sau:
1. $\ddot{\mathcal{D}}\left([\ddot{u}]_A^{\alpha,\beta},[\ddot{u}]_B^{\alpha,\beta}\right)\geq 0$, $\ddot{\mathcal{D}}\left([\ddot{u}]_A^{\alpha,\beta},[\ddot{u}]_B^{\alpha,\beta}\right)=0$ nếu và chỉ nếu $[\ddot{u}]_A^{\alpha,\beta}=[\ddot{u}]_B^{\alpha,\beta}$;
2. $\ddot{\mathcal{D}}\left([\ddot{u}]_A^{\alpha,\beta},[\ddot{u}]_B^{\alpha,\beta}\right)=\ddot{\mathcal{D}}\left([\ddot{u}]_B^{\alpha,\beta},[\ddot{u}]_A^{\alpha,\beta}\right)$;
3. $\ddot{\mathcal{D}}\left([\ddot{u}]_A^{\alpha,\beta},[\ddot{u}]_E^{\alpha,\beta}\right)+\ddot{\mathcal{D}}\left([\ddot{u}]_E^{\alpha,\beta},[\ddot{u}]_B^{\alpha,\beta}\right)\geq\ddot{\mathcal{D}}\left([\ddot{u}]_A^{\alpha,\beta},[\ddot{u}]_B^{\alpha,\beta}\right)$.

Chứng minh. Dễ dàng chứng minh được tính chất 1 và 2 của mệnh đề. Để chứng minh tính chất 3, với hai số thực bất kỳ x và y, ta có $\max\{x,y\}=x+y-\min\{x,y\}$. Khi đó, với mọi đối tượng $u,v\in U$, chúng ta thu được $\gamma_{[\ddot{u}]_{A\cap B}^{\alpha,\beta}}(v)=\gamma_{[\ddot{u}]_A^{\alpha,\beta}}(v)+\gamma_{[\ddot{u}]_B^{\alpha,\beta}}(v)-\gamma_{[\ddot{u}]_{A\cup B}^{\alpha,\beta}}(v)$

và $\eta_{[\ddot{u}]^{\alpha,\beta}_{A\cap B}}(v) = \eta_{[\ddot{u}]^{\alpha,\beta}_{A}}(v) + \eta_{[\ddot{u}]^{\alpha,\beta}_{B}}(v) - \eta_{[\ddot{u}]^{\alpha,\beta}_{A\cup B}}(v)$. Điều này dẫn tới

$$\sum_{v\in U}\gamma_{[\ddot{u}]^{\alpha,\beta}_{A\cap B}}(v) = \sum_{v\in U}\gamma_{[\ddot{u}]^{\alpha,\beta}_{A}}(v) + \sum_{v\in U}\gamma_{[\ddot{u}]^{\alpha,\beta}_{B}}(v) - \sum_{v\in U}\gamma_{[\ddot{u}]^{\alpha,\beta}_{A\cup B}}(v)$$
$$\sum_{v\in U}\eta_{[\ddot{u}]^{\alpha,\beta}_{A\cap B}}(v) = \sum_{v\in U}\eta_{[\ddot{u}]^{\alpha,\beta}_{A}}(v) + \sum_{v\in U}\eta_{[\ddot{u}]^{\alpha,\beta}_{B}}(v) - \sum_{v\in U}\eta_{[\ddot{u}]^{\alpha,\beta}_{A\cup B}}(v)$$

Do đó, $\left|[\ddot{u}]^{\alpha,\beta}_{A\cap B}\right| = \left|[\ddot{u}]^{\alpha,\beta}_{A}\right| + \left|[\ddot{u}]^{\alpha,\beta}_{B}\right| - \left|[\ddot{u}]^{\alpha,\beta}_{A\cup B}\right|$ (1).

Ngoài ra, dựa trên tính chất 2 của Mệnh đề 2.3, chúng ta cũng có:

$$\left|[\ddot{u}]^{\alpha,\beta}_{A}\right| - \left|[\ddot{u}]^{\alpha,\beta}_{A\cup B}\right| + \left|[\ddot{u}]^{\alpha,\beta}_{E}\right| - \left|[\ddot{u}]^{\alpha,\beta}_{E\cup A}\right| \geq \left|[\ddot{u}]^{\alpha,\beta}_{E}\right| - \left|[\ddot{u}]^{\alpha,\beta}_{E\cup B}\right|$$
$$\left|[\ddot{u}]^{\alpha,\beta}_{A}\right| - \left|[\ddot{u}]^{\alpha,\beta}_{A\cup E}\right| + \left|[\ddot{u}]^{\alpha,\beta}_{B}\right| - \left|[\ddot{u}]^{\alpha,\beta}_{B\cup A}\right| \geq \left|[\ddot{u}]^{\alpha,\beta}_{B}\right| - \left|[\ddot{u}]^{\alpha,\beta}_{B\cup E}\right|$$

Khi đó, $\left|[\ddot{u}]^{\alpha,\beta}_{A}\right| + \left|[\ddot{u}]^{\alpha,\beta}_{B}\right| - 2\left|[\ddot{u}]^{\alpha,\beta}_{A\cup B}\right| + \left|[\ddot{u}]^{\alpha,\beta}_{A}\right| + \left|[\ddot{u}]^{\alpha,\beta}_{E}\right| - 2\left|[\ddot{u}]^{\alpha,\beta}_{A\cup E}\right| \geq \left|[\ddot{u}]^{\alpha,\beta}_{B}\right| + \left|[\ddot{u}]^{\alpha,\beta}_{E}\right| - 2\left|[\ddot{u}]^{\alpha,\beta}_{B\cup E}\right|$ (2).

Từ (1) và (2), $\left(\left|[\ddot{u}]^{\alpha,\beta}_{A\cap B}\right| - \left|[\ddot{u}]^{\alpha,\beta}_{A\cup B}\right|\right) + \left(\left|[\ddot{u}]^{\alpha,\beta}_{A\cap E}\right| - \left|[\ddot{u}]^{\alpha,\beta}_{A\cup E}\right|\right) \geq \left(\left|[\ddot{u}]^{\alpha,\beta}_{B\cap E}\right| - \left|[\ddot{u}]^{\alpha,\beta}_{B\cup E}\right|\right)$. Do đó, tính chất 3 của mệnh đề đã được chứng minh. □

Từ độ đo khoảng cách được xây dựng trên hai hạt thông tin mờ trực cảm mức $alpha, beta$, nghiên cứu tiếp tục mở rộng độ đo khoảng cách trên hai phân hoạch mờ trực cảm mức $alpha, beta$ như sau

$$\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}^{\alpha,\beta}_{A}, U/\ddot{\mathcal{R}}^{\alpha,\beta}_{B}\right) = \sum_{u\in U}\frac{\left|[\ddot{u}]^{\alpha,\beta}_{A}\cup[\ddot{u}]^{\alpha,\beta}_{B}\right| - \left|[\ddot{u}]^{\alpha,\beta}_{A}\cap[\ddot{u}]^{\alpha,\beta}_{B}\right|}{|U|^2} \tag{2.6}$$

Rất dễ để thấy rằng, $0 \leq \ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}^{\alpha,\beta}_{A}, U/\ddot{\mathcal{R}}^{\alpha,\beta}_{B}\right) \leq \frac{|U|-1}{|U|}$. Độ đo này được xem như một cơ sở quan trọng để xây dựng chiến lược tìm kiếm các thuộc tính then chốt. Tuy nhiên, như đã biết, một rút gọn cần bảo toàn thông tin và duy trì tính nhất quán của bảng quyết định. Nói cách khác, ý nghĩa của một tập thuộc tính phải thể hiện được mức độ thông tin mà nó tạo ra so với lớp quyết định. Do đó, để đáp ứng yêu cầu này, độ đo khoảng cách tiếp tục được mở rộng nhằm đặc trưng hóa lượng thông tin của một phân hoạch mờ trực cảm mức $alpha, beta$ theo thuộc tính quyết định.

Mệnh đề 2.5 Cho bảng quyết định $IS = (U, C\cup D)$, khoảng cách giữa hai phân hoạch mờ trực cảm mức $alpha, beta$ được tạo bởi C và $C\cup D$ trên U được xác định theo công thức:

$$\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}^{\alpha,\beta}_{C}, U/\ddot{\mathcal{R}}^{\alpha,\beta}_{C\cup D}\right) = \sum_{u\in U}\frac{\left(\left|[\ddot{u}]^{\alpha,\beta}_{C}\right| - \left|[\ddot{u}]^{\alpha,\beta}_{C}\cap[\ddot{u}]^{\alpha,\beta}_{D}\right|\right)}{|U|^2} \tag{2.7}$$

Chứng minh. Dựa trên Công thức 2.6:

$$\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right) = \sum_{u\in U} \frac{\left(\left|[\ddot{u}]_C^{\alpha,\beta} \cup [\ddot{u}]_{C\cup D}^{\alpha,\beta}\right| - \left|[\ddot{u}]_C^{\alpha,\beta} \cap [\ddot{u}]_{C\cap D}^{\alpha,\beta}\right|\right)}{|U|^2}$$
$$= \frac{\sum\limits_{u\in U} \left(\left|[\ddot{u}]_C^{\alpha,\beta} \cup \left([\ddot{u}]_C^{\alpha,\beta} \cap [\ddot{u}]_D^{\alpha,\beta}\right)\right| - \left|[\ddot{u}]_C^{\alpha,\beta} \cap \left([\ddot{u}]_C^{\alpha,\beta} \cap [\ddot{u}]_D^{\alpha,\beta}\right)\right|\right)}{|U|^2}$$

Mặt khác, chúng ta luôn có $[\ddot{u}]_C^{\alpha,\beta} \cap [\ddot{u}]_D^{\alpha,\beta} \subseteq [\ddot{u}]_C^{\alpha,\beta} \Rightarrow [\ddot{u}]_C^{\alpha,\beta} \cup \left([\ddot{u}]_C^{\alpha,\beta} \cap [\ddot{u}]_D^{\alpha,\beta}\right) = [\ddot{u}]_C^{\alpha,\beta}$ và $[\ddot{u}]_C^{\alpha,\beta} \subseteq [\ddot{u}]_D^{\alpha,\beta} \Rightarrow [\ddot{u}]_C^{\alpha,\beta} \cap \left([\ddot{u}]_C^{\alpha,\beta} \cap [\ddot{u}]_D^{\alpha,\beta}\right) = [\ddot{u}]_C^{\alpha,\beta} \cap [\ddot{u}]_D^{\alpha,\beta}$. Do đó, chúng ta thu được Công thức 2.7. □

Về bản chất, Công thức 2.7 biểu diễn một phép biến đổi đặc biệt từ Công thức 2.6 nhằm nâng cao tốc độ tính toán cho các thuật toán rút gọn thuộc tính. Ngoài ra, công thức này được hiểu như một độ đo nhằm đo lường lượng thông tin trung bình của một đối tượng u không nằm trong $[\ddot{u}]_D^{\alpha,\beta}$ trên tập thuộc tính C. Qua đó, một số tính chất quan trọng từ độ đo này được rút ra như sau.

Mệnh đề 2.6 Cho $IS = (U, C \cup D)$ là một bảng quyết định. Nếu $\alpha_1 \leq \alpha_2$ và $\beta_1 \geq \beta_2$ thì $\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_C^{\alpha_1,\beta_1}, U/\ddot{\mathcal{R}}_{C\cup D}^{\alpha_1,\beta_1}\right) \geq \ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_C^{\alpha_2,\beta_2}, U/\ddot{\mathcal{R}}_{C\cup D}^{\alpha_2,\beta_2}\right)$.

Chứng minh. Vì D là thuộc tính quyết định, do đó $[\ddot{u}]_D^{\alpha_1,\beta_1} = [\ddot{u}]_D^{\alpha_2,\beta_2} = [\ddot{u}]_D$. Xét $u, v \in U$, chúng ta có:

$$\left|[\ddot{u}]_C^{\alpha_1,\beta_1}\right| - \left|[\ddot{u}]_C^{\alpha_1,\beta_1} \cap [\ddot{u}]_D^{\alpha_1,\beta_1}\right| = \sum_{[\ddot{u}]_D^{\alpha_1,\beta_1}(v)=(0,1)} \frac{\left[\gamma_{[\ddot{u}]_C^{\alpha_1,\beta_1}}(v) - \eta_{[\ddot{u}]_C^{\alpha_1,\beta_1}}(v)\right] - \left[\gamma_{[\ddot{u}]_D^{\alpha_1,\beta_1}}(v) - \eta_{[\ddot{u}]_D}(v)\right]}{2}$$

$$\left|[\ddot{u}]_C^{\alpha_2,\beta_2}\right| - \left|[\ddot{u}]_C^{\alpha_2,\beta_2} \cap [\ddot{u}]_D^{\alpha_2,\beta_2}\right| = \sum_{[\ddot{u}]_D^{\alpha_2,\beta_2}(v)=(0,1)} \frac{\left[\gamma_{[\ddot{u}]_C^{\alpha_2,\beta_2}}(v) - \eta_{[\ddot{u}]_C^{\alpha_2,\beta_2}}(v)\right] - \left[\gamma_{[\ddot{u}]_D^{\alpha_2,\beta_2}}(v) - \eta_{[\ddot{u}]_D}(v)\right]}{2}$$

Từ $\alpha_1 \leq \alpha_2$ và $\beta_1 \geq \beta_2$, $[\ddot{u}]_C^{\alpha_2,\beta_2} \subseteq [\ddot{u}]_C^{\alpha_1,\beta_1}$. Khi đó, $\left|[\ddot{u}]_C^{\alpha_1,\beta_1}\right| - \left|[\ddot{u}]_C^{\alpha_1,\beta_1} \cap [\ddot{u}]_D^{\alpha_1,\beta_1}\right| \geq \left|[\ddot{u}]_C^{\alpha_2,\beta_2}\right| - \left|[\ddot{u}]_C^{\alpha_2,\beta_2} \cap [\ddot{u}]_D^{\alpha_2,\beta_2}\right| \Rightarrow \ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_C^{\alpha_1,\beta_1}, U/\ddot{\mathcal{R}}_{C\cup D}^{\alpha_1,\beta_1}\right) \geq \ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_C^{\alpha_2,\beta_2}, U/\ddot{\mathcal{R}}_{C\cup D}^{\alpha_2,\beta_2}\right)$. □

Mệnh đề 2.6 chứng minh rằng giá trị của khoảng cách mờ trực cảm giữa hai phân hoạch phụ thuộc vào các mức giá trị của *alpha* và *beta*. Khi giá trị *alpha* lớn và *beta* nhỏ, khoảng cách phân hoạch mờ trực cảm mức $alpha, beta$ theo Công thức 2.7 sẽ càng nhỏ và ngược lại. Khi đó, $\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_C^{0,1}, U/\ddot{\mathcal{R}}_{C\cup D}^{0,1}\right) \geq \ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right)$.

Mệnh đề 2.7 Cho bảng quyết định $IS = (U, C \cup D)$ và hai tập con thuộc tính điều kiện $A, B \subseteq C$. Nếu $A \subseteq B$ thì $\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_A^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{A\cup D}^{\alpha,\beta}\right) \geq \ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_B^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{B\cup D}^{\alpha,\beta}\right)$.

Chứng minh. Từ $A \subseteq B$ với mọi $u, v \in U$, chúng ta luôn có $[\ddot{u}]_B^{\alpha,\beta} \subseteq [\ddot{u}]_A^{\alpha,\beta}$. Khi đó, theo tính chất 1 của Mệnh đề 2.3, $\left|[\ddot{u}]_A^{\alpha,\beta}\right| - \left|[\ddot{u}]_A^{\alpha,\beta} \cap [\ddot{u}]_D^{\alpha,\beta}\right| \geq \left|[\ddot{u}]_B^{\alpha,\beta}\right| - \left|[\ddot{u}]_B^{\alpha,\beta} \cap [\ddot{u}]_D^{\alpha,\beta}\right|$. Điều

này dẫn tới $\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{A}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{A\cup D}^{\alpha,\beta}\right) \geq \ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{B}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{B\cup D}^{\alpha,\beta}\right)$. □

Mệnh đề 2.7 cung cấp tính chất phản đơn điệu của kích thước tập thuộc tính điều kiện với khoảng cách phân hoạch mờ trực cảm. Trên cơ sở này, một rút gọn dựa trên khoảng cách phân hoạch mờ trực cảm mức α, β được định nghĩa thông qua một số điều kiện nhất định.

Định nghĩa 2.2 Cho bảng quyết định $IS = (U, C \cup D)$, khi đó một tập con thuộc tính điều kiện $A \subseteq C$ được gọi là một rút gọn của C nếu:
1. $\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{A}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{A\cup D}^{\alpha,\beta}\right) = \ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{C}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right)$,
2. $\forall A' \subseteq A$, $\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{A'}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{A'\cup D}^{\alpha,\beta}\right) \geq \ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{A}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{A\cup D}^{\alpha,\beta}\right)$.

Từ Định nghĩa 2.2, nếu $\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{A\setminus\{a\}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{A\setminus\{a\}\cup D}^{\alpha,\beta}\right) = \ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{A}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{A\cup D}^{\alpha,\beta}\right)$, với mọi $a \in A$ thì a được xem là một thuộc tính dư thừa trong A. Những thuộc tính như vậy thường đóng góp rất ít vào việc cải thiện độ chính xác của các mô hình phân lớp và có thể gây nhiễu trong quá trình huấn luyện. Ngược lại, a được xem là một thuộc tính quan trọng. Một cách dễ hiểu, không tồn tại một tập con thuộc tính $A' \subseteq A$ thỏa mãn $\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{A'}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{A'\cup D}^{\alpha,\beta}\right) \geq \ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{A}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{A\cup D}^{\alpha,\beta}\right)$. Điều này đặc trưng cho tính tối ưu về kích thước của rút gọn. Từ định nghĩa về rút gọn, luận án tiếp tục đưa ra một độ đo nhằm đánh giá tính quan trọng của mỗi thuộc tính trong bảng quyết định.

Định nghĩa 2.3 Cho bảng quyết định $IS = (U, C \cup D)$, một tập con thuộc tính điều kiện $A \subseteq C$ và một thuộc tính $a \in C\setminus A$, độ quan trọng của a theo tập thuộc tính A, ký hiệu là $Sig_P(a, A)$ được xác định theo công thức sau:

$$Sig_P(a, A) = \ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{A}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{A\cup D}^{\alpha,\beta}\right) - \ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{A\cup\{a\}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{A\cup\{a\}\cup D}^{\alpha,\beta}\right) \tag{2.8}$$

Từ tính chất phản đơn điệu của độ đo khoảng cách, $Sig_P(a, A) \geq 0$. Rất dễ thấy rằng độ quan trọng của một thuộc tính bất kỳ theo một tập con các thuộc tính điều kiện cho trước được xác định bằng việc đánh giá sự thay đổi mức độ chắc chắn khi thuộc tính đó được bổ sung vào. Một cách trực cảm, độ đo này cũng đặc trưng sự đóng góp của thuộc tính a đến chất lượng phân lớp của tập con thuộc tính. Do đó, độ quan trọng có thể xem như một tiêu chuẩn để tìm kiếm các thuộc tính quan trọng.

Từ các khái niệm được trình bày, Thuật toán 2.1 được xây dựng với quy trình xử lý được trình bày trong Hình 2.2 nhằm trích xuất một tập con thuộc tính tối ưu trên bảng quyết định sẽ được trình bày với ý tưởng bắt đầu từ một tập rỗng và bổ sung lần lượt các thuộc tính với độ quan trọng cao nhất vào sau mỗi vòng lặp. Quá trình này sẽ liên

Thuật toán 2.1 Rút gọn thuộc tính dựa trên khoảng cách phân hoạch mờ trực cảm (ARPD)

Đầu vào: Bảng quyết định $IS = (U, C \cup D)$ và các mức α, β.

Đầu ra: Một rút gọn $\mathcal{A}$

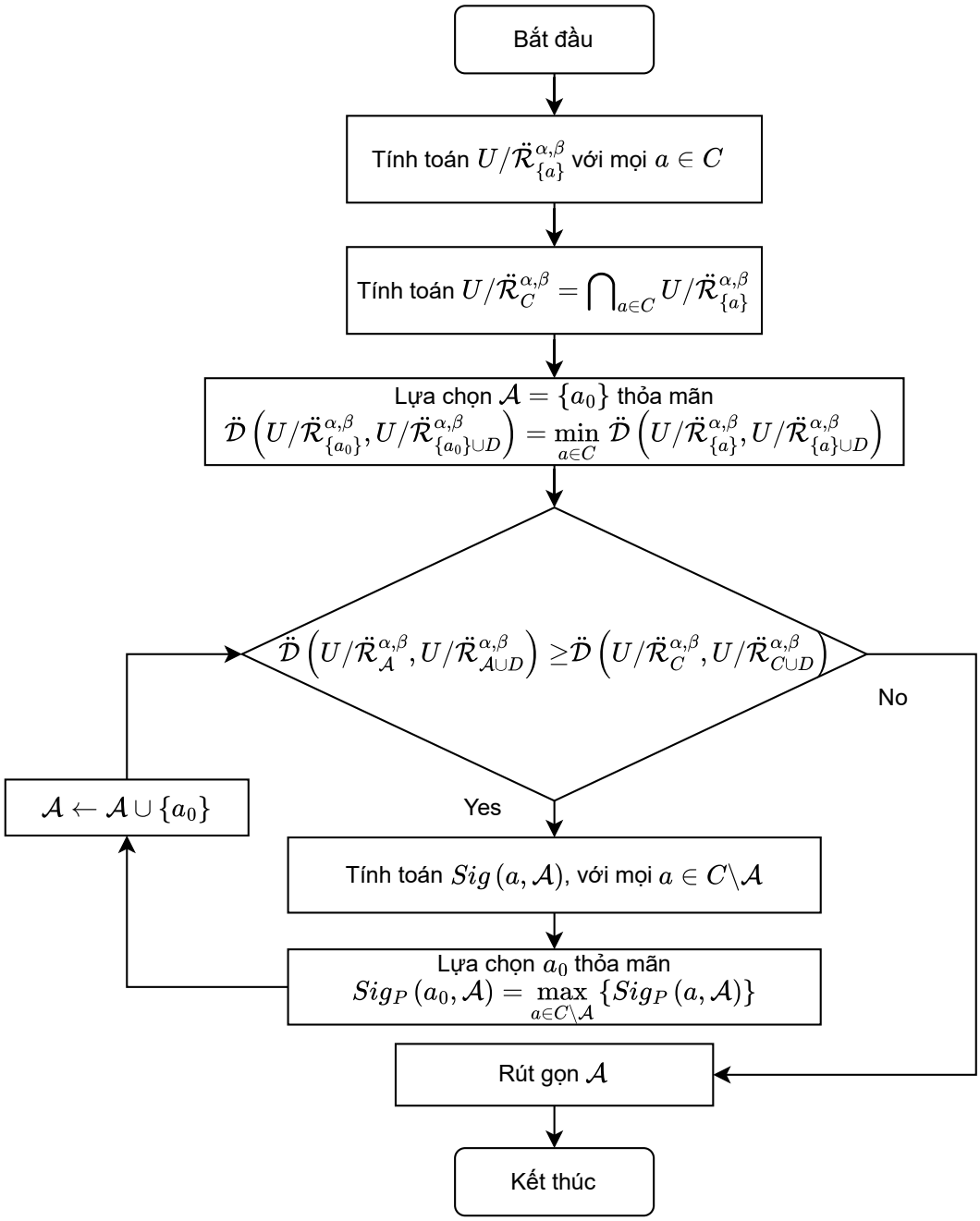
1: tính toán $U/\ddot{\mathcal{R}}_D$.
2: **for** $a \in C$ **do**
3: **if** a là thuộc tính số **then**
4: tính toán $U/\ddot{\mathcal{R}}^{\alpha,\beta}_{\{a\}}$
5: **else**
6: $U/\ddot{\mathcal{R}}^{\alpha,\beta}_{\{a\}} = U/\ddot{\mathcal{R}}_{\{a\}}$
7: **end if**
8: **end for**
9: tính toán $U/\ddot{\mathcal{R}}^{\alpha,\beta}_{C} = \bigcap_{a\in C} U/\ddot{\mathcal{R}}^{\alpha,\beta}_{\{a\}}$
10: $\mathcal{A} = \{a_0\}$ thỏa mãn $\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}^{\alpha,\beta}_{\{a_0\}}, U/\ddot{\mathcal{R}}^{\alpha,\beta}_{\{a_0\}\cup D}\right) = \min\limits_{a\in C} \ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}^{\alpha,\beta}_{\{a\}}, U/\ddot{\mathcal{R}}^{\alpha,\beta}_{\{a\}\cup D}\right)$
11: **while** $\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}^{\alpha,\beta}_{\mathcal{A}}, U/\ddot{\mathcal{R}}^{\alpha,\beta}_{\mathcal{A}\cup D}\right) \geq \ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}^{\alpha,\beta}_{C}, U/\ddot{\mathcal{R}}^{\alpha,\beta}_{C\cup D}\right)$ **do**
12: tính toán $Sig\left(a, \mathcal{A}\right)$, với mọi $a \in C \backslash \mathcal{A}$
13: lựa chọn a_0 thỏa mãn $Sig_P\left(a_0, \mathcal{A}\right) = \max\limits_{a\in C\backslash\mathcal{A}} \left\{Sig_P\left(a, \mathcal{A}\right)\right\}$
14: $\mathcal{A} \leftarrow \mathcal{A} \cup \{a_0\}$
15: **end while**
16: **return** $\mathcal{A}$

tục diễn ra cho tới khi điều kiện dừng xảy ra.

Để đánh giá độ phức tạp của Thuật toán ARPD, giả sử rằng $|C|$, $|U|$ tương ứng là số lượng thuộc tính điều kiện và số lượng đối tượng trên bảng quyết định. Rất dễ để thấy rằng thời gian thực thi của thuật toán sẽ được xử lý trong giai đoạn tính toán các phân hoạch mờ trực cảm mức α, β của mỗi thuộc tính và giai đoạn lọc tìm kiếm thuộc tính có ý nghĩa cao nhất.

Trong giai đoạn đầu tiên, thuật toán ban đầu xác định phân hoạch mờ trực cảm trên thuộc tính quyết định bao gồm $|U|$ hạt thông tin mờ trực cảm. Tại bước này, độ phức tạp của thuật toán là $O\left(|U|^2\right)$. Tiếp theo, thuật toán duyệt trên toàn bộ các thuộc tính trong tập C và thực hiện tính toán các phân hoạch mờ trực cảm tương ứng. Do đó, độ phức tạp tính toán từ dòng 2 đến dòng 8 là $O\left(|C|\,|U|^2\right)$. Tại dòng 9, thuật toán tính phân hoạch mờ trực cảm mức α, β của tập thuộc tính C, với độ phức tạp là $O\left(|C|\right)$.

Giai đoạn lọc của thuật toán được thực hiện từ dòng 11 đến dòng 15. Cụ thể, tại dòng 12, thuật toán tính độ quan trọng của các thuộc tính còn lại với độ phức tạp là $O\left(|C|\,|U|^2\right)$. Sau đó, tại dòng 13 và dòng 14, thuật toán lựa chọn thuộc tính có độ quan trọng lớn nhất để bổ sung vào tập rút gọn. Độ phức tạp của thuật toán trong hai bước này là $O\left(|C|\right)$. Như vậy, độ phức tạp của thuật toán trong vòng lặp while là $O\left(|C|^2|U|^2\right)$. Do đó, thời gian thực thi tổng thể của Thuật toán ARPD được xác định là $O\left(|C|^2|U|^2\right)$.



Hình 2.2: Lưu đồ xử lý của thuật toán ARPD

Ví dụ 2.3 Để tìm rút gọn của Bảng 1.1 dựa trên thuật toán ARPD, đầu tiên chúng ta khởi tạo tập rút gọn $\mathcal{A} = \emptyset$, tính $\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right)$ và $\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{\{a\}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{\{a\}\cup D}^{\alpha,\beta}\right)$, với mọi $a \in C$ từ Công thức 2.7

$$\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right) = \frac{1}{25} \times (0.80 + 0.80 + 0.00 + 0.00 + 0.00) = 0.06$$

Thực hiện tương tự, chúng ta cũng thu được:

$\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{\{a_1\}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{\{a_1\}\cup D}^{\alpha,\beta}\right) = 0.28$, $\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{\{a_2\}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{\{a_2\}\cup D}^{\alpha,\beta}\right) = 0.44$,

$\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{\{a_3\}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{\{a_3\}\cup D}^{\alpha,\beta}\right) = 0.22$, $\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{\{a_4\}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{\{a_4\}\cup D}^{\alpha,\beta}\right) = 0.36$,

$\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{\{a_5\}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{\{a_5\}\cup D}^{\alpha,\beta}\right) = 0.14$. Vì $\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{\{a_5\}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{\{a_5\}\cup D}^{\alpha,\beta}\right) = 0.14$ là nhỏ nhất nên bổ sung thuộc tính a_5 vào rút gọn, lúc này $\mathcal{A} = \{a_5\}$. Nhận thấy $\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{\mathcal{A}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{\mathcal{A}\cup D}^{\alpha,\beta}\right) \geq \ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right)$ nên thuật toán tiếp tục chuyển sang giai đoạn lọc tìm kiếm các thuộc tính còn lại.

$\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{\mathcal{A}\cup\{a_1\}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{\mathcal{A}\cup\{a_1\}\cup D}^{\alpha,\beta}\right) = 0.13 \Rightarrow Sig_P(a_1, \mathcal{A}) = 0.14 - 0.13 = 0.01$,

$\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{\mathcal{A}\cup\{a_2\}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{\mathcal{A}\cup\{a_2\}\cup D}^{\alpha,\beta}\right) = 0.14 \Rightarrow Sig_P(a_2, \mathcal{A}) = 0.14 - 0.14 = 0.00$,

$\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{\mathcal{A}\cup\{a_3\}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{\mathcal{A}\cup\{a_3\}\cup D}^{\alpha,\beta}\right) = 0.07 \Rightarrow Sig_P(a_3, \mathcal{A}) = 0.14 - 0.07 = 0.07$,

$\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{\mathcal{A}\cup\{a_4\}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{\mathcal{A}\cup\{a_4\}\cup D}^{\alpha,\beta}\right) = 0.14 \Rightarrow Sig_P(a_4, \mathcal{A}) = 0.14 - 0.14 = 0.00$.

Vì $Sig_P(a_3, \mathcal{A})$ lớn nhất nên bổ sung tiếp thuộc tính a_3 vào rút gọn, khi đó $\mathcal{A} = \mathcal{A} \cup \{a_3\} = \{a_5, a_3\}$. Thuật toán tiếp tục lọc các thuộc tính còn lại do $\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{\mathcal{A}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{\mathcal{A}\cup D}^{\alpha,\beta}\right) \geq \ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right)$.

$\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{\mathcal{A}\cup\{a_1\}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{\mathcal{A}\cup\{a_1\}\cup D}^{\alpha,\beta}\right) = 0.06 \Rightarrow Sig_P(a_1, \mathcal{A}) = 0.07 - 0.06 = 0.01$,

$\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{\mathcal{A}\cup\{a_2\}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{\mathcal{A}\cup\{a_2\}\cup D}^{\alpha,\beta}\right) = 0.07 \Rightarrow Sig_P(a_2, \mathcal{A}) = 0.07 - 0.07 = 0.00$,

$\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{\mathcal{A}\cup\{a_4\}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{\mathcal{A}\cup\{a_4\}\cup D}^{\alpha,\beta}\right) = 0.07 \Rightarrow Sig_P(a_4, \mathcal{A}) = 0.07 - 0.07 = 0.00$.

Vì $Sig_P(a_1, \mathcal{A})$ có giá trị lớn nhất nên bổ sung thuộc tính a_1 vào rút gọn, khi đó $\mathcal{A} = \mathcal{A} \cup \{a_1\} = \{a_5, a_3, a_1\}$. Lúc này thuật toán kết thúc do điều kiện dừng $\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{\mathcal{A}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{\mathcal{A}\cup D}^{\alpha,\beta}\right) = \ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right)$ xảy ra. Do đó, rút gọn cuối cùng của bảng thu được từ Thuật toán 2.1 là $\mathcal{A} = \{a_5, a_3, a_1\}$.

2.3.2 Đề xuất thuật toán rút gọn thuộc tính trên bảng quyết định thay đổi khi bổ sung tập đối tượng

Như đã đề cập, các thuật toán dựa trên tiếp cận tập mờ trực cảm thường đòi hỏi thời gian thực thi rất lớn, đặc biệt khi bảng quyết định có sự thay đổi do bổ sung hoặc loại bỏ tập đối tượng. Để khắc phục hạn chế này, luận án đề xuất một số thuật toán gia tăng, khai thác khoảng cách giữa hai phân hoạch mờ trực cảm mức α, β nhằm tìm kiếm rút gọn xấp xỉ trên các bảng quyết định động với chi phí tính toán được giảm thiểu.

Trước hết, một công thức tính toán gia tăng dựa trên Mệnh đề 2.5 được xây dựng, cho phép rút ngắn quá trình xác định khoảng cách giữa hai phân hoạch khi có một tập đối tượng mới được bổ sung vào bảng quyết định.

Mệnh đề 2.8 Cho bảng quyết định $IS = (U, C \cup D)$ với $U = \{u_1, u_2, \ldots, u_n\}$, các quan hệ tương đương mờ trực cảm $\ddot{\mathcal{R}}_C$, $\ddot{\mathcal{R}}_D$ và một tập đối tượng mới được bổ sung vào bảng $\Delta U = \{U_{n+1}, U_{n+2}, \ldots, U_{n+z}\}$, trong đó $z \geq 1$. Khi đó, khoảng cách phân hoạch mờ trực cảm giữa $U^+/\ddot{\mathcal{R}}_C^{\alpha,\beta}$ và $U^+/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}$ trên $U^+ = U \cup \Delta U$, ký hiệu là $\ddot{\mathcal{D}}\left(U^+/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right)$, được xác định như sau:

$$\ddot{\mathcal{D}}\left(U^+/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right) = \frac{n^2\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right)}{(n+z)^2} + 2\sum_{i=1}^{z}\frac{\left(\left|[\ddot{u}_{n+i}]_C^{\alpha,\beta}\right| - \left|[\ddot{u}_{n+i}]_C^{\alpha,\beta} \cap [\ddot{u}_{n+i}]_D^{\alpha,\beta}\right| - \vartheta_i\right)}{(n+z)^2} \tag{2.9}$$

trong đó, $\vartheta_1 = 0$ và với mọi $i \geq 2$,

$$\vartheta_i = \sum_{j=1}^{z-1}\left(\gamma_{[\ddot{u}_{n+i}]_C^{\alpha,\beta}}(u_{n+j+1}) - \gamma_{[\ddot{u}_{n+i}]_{C\cup D}^{\alpha,\beta}}(u_{n+j+1}) - \eta_{[\ddot{u}_{n+i}]_C^{\alpha,\beta}}(u_{n+j+1}) + \eta_{[\ddot{u}_{n+i}]_{C\cup D}^{\alpha,\beta}}(u_{n+j+1})\right).$$

Chứng minh. Đặt $\vartheta_{i,j} = \gamma_{[\ddot{u}_i]_C^{\alpha,\beta}}(u_j) - \gamma_{[\ddot{u}_i]_{C\cup D}^{\alpha,\beta}}(u_j) - \eta_{[\ddot{u}_i]_C^{\alpha,\beta}}(u_j) + \eta_{[\ddot{u}_i]_{C\cup D}^{\alpha,\beta}}(u_j)$. Vì có z đối tượng được bổ sung vào U^+ từ U, Công thức 2.7 trở thành:

$$D\left(U^+/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right) = \sum_{i=1}^{n+z}\frac{\left(\left|[\ddot{u}_i]_C^{\alpha,\beta}\right| - \left|[\ddot{u}_i]_C^{\alpha,\beta} \cap [\ddot{u}_i]_D^{\alpha,\beta}\right|\right)}{(n+z)^2}$$

$$= \frac{\sum_{i=1}^{n}\left(\left|[\ddot{u}_i]_C^{\alpha,\beta}\right| - \left|[\ddot{u}_i]_C^{\alpha,\beta} \cap [\ddot{u}_i]_D^{\alpha,\beta}\right|\right) + \sum_{i=1}^{z}\left(\left|[\ddot{u}_{n+i}]_C^{\alpha,\beta}\right| - \left|[\ddot{u}_{n+i}]_C^{\alpha,\beta} \cap [\ddot{u}_{n+i}]_D^{\alpha,\beta}\right|\right)}{(n+z)^2}$$

$$= \frac{\sum_{i=1}^{n}\sum_{j=1}^{z}\vartheta_{i,n+j} + n^2 D\left(U/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right) + \sum_{i=1}^{z}\left(\left|[\ddot{u}_{n+i}]_C^{\alpha,\beta}\right| - \left|[\ddot{u}_{n+i}]_C^{\alpha,\beta} \cap [\ddot{u}_{n+i}]_D^{\alpha,\beta}\right|\right)}{(n+z)^2}$$

$$= \frac{n^2 D\left(U/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right) + 2\sum_{i=1}^{z}\left(\left|[\ddot{u}_{n+i}]_C^{\alpha,\beta}\right| - \left|[\ddot{u}_{n+i}]_C^{\alpha,\beta} \cap [\ddot{u}_{n+i}]_D^{\alpha,\beta}\right|\right) - \sum_{i=1}^{z}\sum_{j=1}^{z}\vartheta_{n+i,n+j}}{(n+z)^2}$$

Mặt khác, dựa trên quan hệ tương đương mờ trực cảm, $\forall i, j$ chúng ta luôn có: $\vartheta_{i,j} = \vartheta_{j,i}$ và $\vartheta_{i,i} = 0$. Do đó, $\sum_{i=1}^{z}\sum_{j=1}^{z}\vartheta_{n+i,n+j} = 2\sum_{i=1}^{z}\sum_{j=1}^{z-1}\vartheta_{n+i,n+j+1} = 2\sum_{i=1}^{z}\vartheta_i$. Từ đây, chúng ta thu được Công thức 2.9. □

Công thức gia tăng 2.9 bao gồm hai thành phần chính. Thành phần đầu tiên tính toán khoảng cách giữa hai phân hoạch mờ trực cảm mức α, β khi chưa bổ sung tập đối

tượng mới. Trên thực tế, thành phần này đã được xác định trong giai đoạn đầu tiên khi bảng quyết định chưa có sự biến động. Do đó, công thức trong Mệnh đề 2.8 chỉ tập trung tính toán ở thành phần thứ hai. Đây là thành phần chứa các hạt thông tin mờ trực cảm mức α, β được hình thành khi bảng quyết định cập nhật thêm tập đối tượng mới. Như vậy, trong trường hợp bảng quyết định có sự bổ sung một tập đối tượng mới, Công thức 2.9 sẽ được thay thế cho Công thức 2.7 để giảm thiểu thời gian tính toán.

Mệnh đề 2.9 Cho bảng quyết định $IS = (U, C \cup D)$ với $U = \{u_1, u_2, \ldots, u_n\}$, giả sử rằng, $A \subseteq C$ là một rút gọn dựa trên khoảng cách phân hoạch mờ trực cảm mức α, β trên tập đối tượng U và $\Delta U = \{u_{n+1}, u_{n+2}, \ldots, u_{n+z}\}$, trong đó $z \geq 1$ là một tập đối tượng bổ sung vào U, khi đó chúng ta có:

1. Nếu tất cả các đối tượng trong ΔU có giá trị thuộc tính quyết định giống nhau thì:

$$\ddot{\mathcal{D}}\left(U^+/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right) = \frac{n^2\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right)}{(n+z)^2} + 2\sum_{i=1}^{z}\frac{\left(\left|[\ddot{u}_{n+i}]_C^{\alpha,\beta}\right| - \left|[\ddot{u}_{n+i}]_C^{\alpha,\beta} \cap [\ddot{u}_{n+i}]_D^{\alpha,\beta}\right|\right)}{(n+z)^2} \tag{2.10}$$

2. Nếu $[\ddot{u}_{n+i}]_A^{\alpha,\beta} \subseteq [\ddot{u}_{n+i}]_D^{\alpha,\beta}$ với $i = 1, 2, \ldots, z$ thì:

$$\ddot{\mathcal{D}}\left(U^+/\ddot{\mathcal{R}}_A^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{A\cup D}^{\alpha,\beta}\right) = \ddot{\mathcal{D}}\left(U^+/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right)$$

Chứng minh. Cho $i = 1, 2, \ldots, z$ và $j = 1, 2, \ldots, z-1$, chúng ta xét:

1. Vì tất cả các đối tượng trong tập ΔU có giá trị thuộc tính quyết định giống nhau nên $\gamma_{[\ddot{u}_{n+i}]_C^{\alpha,\beta}}(u_{n+j+1}) = 1$ và $\eta_{[\ddot{u}_{n+i}]_C^{\alpha,\beta}}(u_{n+j+1}) = 0$. Do đó, chúng ta có thể dễ dàng thu được $\gamma_{[\ddot{u}_{n+i}]_{C\cup D}^{\alpha,\beta}}(u_{n+j+1}) = \gamma_{[\ddot{u}_{n+i}]_C^{\alpha,\beta}}(u_{n+j+1})$ và $\eta_{[\ddot{u}_{n+i}]_{C\cup D}^{\alpha,\beta}}(u_{n+j+1}) = \eta_{[\ddot{u}_{n+i}]_C^{\alpha,\beta}}(u_{n+j+1})$. Kết hợp với Mệnh đề 2.8, tính chất đầu tiên đã được chứng minh.

2. Vì $[\ddot{u}_{n+i}]_A^{\alpha,\beta} \subseteq [\ddot{u}_{n+i}]_D^{\alpha,\beta}$, nên $\left|[\ddot{u}_{n+i}]_{A\cup D}^{\alpha,\beta}\right| = \left|[\ddot{u}_{n+i}]_A^{\alpha,\beta}\right|$ và $\left|[\ddot{u}_{n+i}]_{C\cup D}^{\alpha,\beta}\right| = \left|[\ddot{u}_{n+i}]_C^{\alpha,\beta}\right|$. Thêm vào đó, chúng ta có $\gamma_{[\ddot{u}_{n+i}]_{A\cup D}^{\alpha,\beta}}(u_{n+j+1}) = \gamma_{[\ddot{u}_{n+i}]_A^{\alpha,\beta}}(u_{n+j+1})$ và $\eta_{[\ddot{u}_{n+i}]_{A\cup D}^{\alpha,\beta}}(u_{n+j+1}) = \eta_{[\ddot{u}_{n+i}]_A^{\alpha,\beta}}(u_{n+j+1})$. Từ tính chất 1 được chứng minh và A là một rút gọn của C dựa trên khoảng cách phân hoạch mờ trực cảm, chúng ta thu được tính chất 2 của mệnh đề. □

Mệnh đề 2.9 chỉ ra hai trường hợp đặc biệt của tập đối tượng được bổ sung vào bảng quyết định. Trong hai trường hợp này, Công thức 2.9 được rút ngắn và xử lý nhanh hơn. Qua đó, luận án đề xuất Thuật toán 2.2 với hiệu năng xử lý nhanh hơn để tìm kiếm một rút gọn xấp xỉ trên bảng quyết định khi có sự bổ sung tập đối tượng được xây dựng với ba giai đoạn chính được trình bày trong bảng mã giả dưới đây.

Thuật toán 2.2 Rút gọn thuộc tính gia tăng khi bổ sung tập đối tượng dựa trên khoảng cách phân hoạch mờ trực cảm mức α, β (IARPD-AO)

Đầu vào: Bảng quyết định $IS = (U, C \cup D)$ với $U = \{u_1, u_2, \ldots, u_n\}$, rút gọn $\mathcal{A} \subseteq C$, các phân hoạch $U/\ddot{\mathcal{R}}_{\mathcal{A}}^{\alpha,\beta}$, $U/\ddot{\mathcal{R}}_{C}^{\alpha,\beta}$ và tập đối tượng bổ sung $\Delta U = \{u_{n+1}, u_{n+2}, \ldots, u_{n+z}\}$.

Đầu ra: Rút gọn xấp xỉ $\mathcal{A}^+$ trên $U^+ = U \cup \Delta U$

// kiểm tra tập đối tượng bổ sung

$\mathcal{A}^+ \leftarrow \mathcal{A}$, $Z \leftarrow \Delta U$

for $i = 1$ to z **do**

  if $[\ddot{u}_{n+i}]_{\mathcal{A}}^{\alpha,\beta} \subseteq [\ddot{u}_{n+i}]_{D}^{\alpha,\beta}$ **then**

    $Z \leftarrow Z \backslash \{u_{n+i}\}$

  end if

end for

if $Z = \emptyset$ **then**

  return $\mathcal{A}^+$

end if

$\Delta U \leftarrow Z$, $z \leftarrow |\Delta U|$

cập nhật $U^+/\ddot{\mathcal{R}}_{\mathcal{A}}^{\alpha,\beta}$, $U^+/\ddot{\mathcal{R}}_{C}^{\alpha,\beta}$ và $U^+/\ddot{\mathcal{R}}_{\{a\}}^{\alpha,\beta}$ với mọi $a \in C \backslash \mathcal{A}$

// lọc tìm kiếm rút gọn xấp xỉ

tính toán $\ddot{\mathcal{D}}\left(U^+/\ddot{\mathcal{R}}_{\mathcal{A}^+}^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{\mathcal{A}^+\cup D}^{\alpha,\beta}\right)$ and $\ddot{\mathcal{D}}\left(U^+/\ddot{\mathcal{R}}_{C}^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right)$ từ Công thức 2.9.

while $\ddot{\mathcal{D}}\left(U^+/\ddot{\mathcal{R}}_{\mathcal{A}^+}^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{\mathcal{A}^+\cup D}^{\alpha,\beta}\right) \geq \ddot{\mathcal{D}}\left(U^+/\ddot{\mathcal{R}}_{C}^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right)$ **do**

  tính toán $Sig\left(a, \mathcal{A}^+\right)$, với mọi $a \in C \backslash \mathcal{A}^+$

  lựa chọn a_0 thỏa mãn: $Sig_P\left(a_0, \mathcal{A}^+\right) = \max\limits_{a \in C\backslash\mathcal{A}^+} \left\{Sig\left(a, \mathcal{A}^+\right)\right\}$

  $\mathcal{A}^+ \leftarrow \mathcal{A}^+ \cup \{a_0\}$

end while

// loại bỏ các thuộc tính dư thừa

for $a \in \mathcal{A}^+$ **do**

  tính toán $\ddot{\mathcal{D}}\left(U^+/\ddot{\mathcal{R}}_{\mathcal{A}^+\backslash\{a\}}^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{\mathcal{A}^+\backslash\{a\}\cup D}^{\alpha,\beta}\right)$

  if $\ddot{\mathcal{D}}\left(U^+/\ddot{\mathcal{R}}_{\mathcal{A}^+\backslash\{a\}}^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{\mathcal{A}^+\backslash\{a\}\cup D}^{\alpha,\beta}\right) = \ddot{\mathcal{D}}\left(U^+/\ddot{\mathcal{R}}_{\mathcal{A}^+}^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{\mathcal{A}^+\cup D}^{\alpha,\beta}\right)$ **then**

    $\mathcal{A}^+ \leftarrow \mathcal{A}^+ \backslash \{a_0\}$

  end if

end for

return $\mathcal{A}^+$

Trong giai đoạn đầu tiên, thuật toán tiến hành kiểm tra các đối tượng bổ sung. Nếu các đối tượng này không ảnh hưởng đến kết quả rút gọn theo tính chất 2 của Mệnh đề 2.9, chúng sẽ bị loại bỏ để giảm thiểu không gian tính toán khi cập nhật các phân hoạch mờ trực cảm mức α, β. Giai đoạn thứ hai của thuật toán sẽ duyệt qua các thuộc tính còn lại trong bảng quyết định, những thuộc tính này không nằm trong rút gọn của giai đoạn trước. Thuộc tính có độ quan trọng cao nhất sẽ được bổ sung vào rút gọn tới khi thỏa mãn điều kiện dừng của thuật toán. Cuối cùng, ở giai đoạn thứ ba, thuật toán loại bỏ các thuộc tính dư thừa trong rút gọn thu được từ giai đoạn thứ hai. Đây là những thuộc tính không bảo toàn tính tối ưu theo định nghĩa của rút gọn đã được trình bày.

Để chứng minh hiệu quả trong việc giảm thiểu thời gian thực thi của thuật toán, giả sử rằng, $\left|\mathcal{A}^+\right|$ là số lượng thuộc tính của tập rút gọn thu được từ thuật toán, $|\Delta U|$ biểu diễn số lượng các đối tượng được bổ sung vào bảng quyết định và $\left|U^+\right|$ biểu diễn tổng số đối tượng trên bảng quyết định mới. Ở dòng đầu tiên, tập rút gọn của thuật toán được khởi tạo bằng chính tập rút gọn thu được từ giai đoạn trước. Trong trường hợp đơn giản, toàn bộ quá trình xử lý của thuật toán sẽ kết thúc ở dòng 9. Trong đó, thuật toán sẽ duyệt trên các thuộc tính của rút gọn để xác định $|\Delta U|$ hạt thông tin mờ trực cảm mức α, β gồm $\left|U^+\right|$ đối tượng trong mỗi hạt. Đây là các hạt thông tin được tạo ra bởi tập đối tượng bổ sung. Do đó, độ phức tạp của thuật toán trong trường hợp này là $O\left(\left|\mathcal{A}^+\right|\left|U^+\right||\Delta U|\right)$. Trong trường hợp còn lại, độ phức tạp dòng 11 của thuật toán là $O\left(\left|\mathcal{A}^+\right|\left|U^+\right||\Delta U|\right)$ và trên công thức gia tăng tại dòng 12 là $O\left(\left|U^+\right||\Delta U|\right)$. Độ phức tạp của vòng lặp *while* trong thuật toán cũng được tính tương tự như vòng lặp *while* trong ARPD là $O\left(\left(|C|-\left|\mathcal{A}^+\right|\right)^2\left|U^+\right||\Delta U|\right)$. Từ dòng 18 tới dòng 22, độ phức tạp của vòng lặp for là $O\left(\left|\mathcal{A}^+\right|\left|U^+\right||\Delta U|\right)$. Từ hai trường hợp được phân tích, độ phức tạp cuối cùng của IARPD-AO là $\max\left\{O\left(\left|\mathcal{A}^+\right|\left|U^+\right||\Delta U|\right), O\left(\left(|C|-\left|\mathcal{A}^+\right|\right)^2\left|U^+\right||\Delta U|\right)\right\}$. Do đó, thời gian tính toán của IARPD-AO được cải thiện rõ rệt khi so sánh với thuật toán ARPD.

Trong trường hợp dữ liệu bổ sung có các giá trị thuộc tính quyết định giống nhau, công thức gia tăng được thay thế bằng Công thức 2.10 từ tính chất đầu tiên của Mệnh đề 2.9.

2.3.3 Đề xuất thuật toán rút gọn thuộc tính trên bảng quyết định thay đổi khi loại bỏ tập đối tượng

Trong phần trước, nghiên cứu đã giới thiệu một thuật toán gia tăng để xử lý trường hợp bảng quyết định bổ sung một tập đối tượng. Để giải quyết kịch bản bảng quyết

định có sự loại bỏ đối tượng, nghiên cứu tiếp tục phát triển một công thức gia tăng mới giúp tính toán nhanh chóng khoảng cách phân hoạch mờ trực cảm. Dựa trên đó, một thuật toán gia tăng rút gọn thuộc tính cũng được thiết kế nhằm tìm kiếm tập rút gọn xấp xỉ trong trường hợp này.

Mệnh đề 2.10 Cho bảng quyết định $IS = (U, C \cup D)$ với $U = \{u_1, u_2, \ldots, u_n\}$, các quan hệ tương đương mờ trực cảm $\ddot{\mathcal{R}}_C, \ddot{\mathcal{R}}_D$ và một tập đối tượng $\Delta U = \{u_n, u_{n-1}, \ldots, u_{n-z+1}\}$ được loại bỏ khỏi bảng, trong đó $z \geq 1$. Khi đó, khoảng cách phân hoạch mờ trực cảm mức *alpha, beta* giữa $U^-/\ddot{\mathcal{R}}_C^{\alpha,\beta}$ và $U^-/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}$ trên $U^- = U \setminus \Delta U$, ký hiệu là $\ddot{\mathcal{D}}\left(U^-/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U^-/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right)$, được xác định như sau:

$$\ddot{\mathcal{D}}\left(U^-/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U^-/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right) = \frac{n^2\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right)}{(n-z)^2} - 2\sum_{i=1}^{z}\frac{\left(\left|[\ddot{u}_{n-i+1}]_C^{\alpha,\beta}\right| - \left|[\ddot{u}_{n-i+1}]_C^{\alpha,\beta} \cap [\ddot{u}_{n-i+1}]_D^{\alpha,\beta}\right| - \xi_i\right)}{(n-z)^2} \tag{2.11}$$

trong đó, $\xi_1 = 0$ và với mọi $i \geq 2$ thì

$$\xi_i = \sum_{j=i}^{z}\left(\gamma_{[\ddot{u}_{n-i+1}]_C^{\alpha,\beta}}(u_{n-j+1}) - \gamma_{[\ddot{u}_{n-i+1}]_{C\cup D}^{\alpha,\beta}}(u_{n-j+1}) - \eta_{[\ddot{u}_{n-i+1}]_C^{\alpha,\beta}}(u_{n-j+1}) + \eta_{[\ddot{u}_{n-i+1}]_{C\cup D}^{\alpha,\beta}}(u_{n-j+1})\right)$$

Chứng minh. Đặt: $\xi_{i,j} = \gamma_{[\ddot{u}_i]_C^{\alpha,\beta}}(u_j) - \gamma_{[\ddot{u}_i]_{C\cup D}^{\alpha,\beta}}(u_j) - \eta_{[\ddot{u}_i]_C^{\alpha,\beta}}(u_j) + \eta_{[\ddot{u}_i]_{C\cup D}^{\alpha,\beta}}(u_j)$. Vì U loại bỏ đi z đối tượng, Công thức 2.7 trở thành:

$$\begin{aligned}
&\ddot{\mathcal{D}}\left(U^-/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U^-/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right) \\
&= \frac{n^2\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right) - \sum_{i=1}^{z}\left(\left|[\ddot{u}_{n-i+1}]_C^{\alpha,\beta}\right| - \left|[\ddot{u}_{n-i+1}]_{C\cup D}^{\alpha,\beta}\right|\right) - \sum_{i=1}^{n-z}\sum_{j=1}^{z}\xi_{i,n-j+1}}{(n-z)^2} \\
&= \frac{n^2\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right) - 2\sum_{i=1}^{z}\left(\left|[\ddot{u}_{n-i+1}]_C^{\alpha,\beta}\right| - \left|[\ddot{u}_{n-i+1}]_{C\cup D}^{\alpha,\beta}\right|\right) + \sum_{i=1}^{z}\sum_{j=1}^{z}\xi_{n-i+1,n-j+1}}{(n-z)^2}
\end{aligned}$$

Theo tính chất của quan hệ tương đương mờ trực cảm, $\xi_{i,j} = \xi_{j,i}$ và $\xi_{i,i} = 0$, $\forall i, j$. Do đó, $\sum_{i=1}^{z}\sum_{j=1}^{z}\xi_{n-i+1,n-j+1} = 2\sum_{i=1}^{z}\sum_{j=i}^{z}\xi_{n-i+1,n-j+1} = 2\sum_{i=1}^{z}\xi_i$. Từ đây, có thể dễ dàng thu được Công thức 2.11 và mệnh đề được chứng minh. □

Mệnh đề 2.11 Cho bảng quyết định $IS = (U, C \cup D)$ với $U = \{u_1, u_2, \ldots, u_n\}$, giả sử rằng $A \subseteq C$ là một rút gọn dưa trên khoảng cách phân hoạch mờ trực cảm mức *alpha, beta* trên U và một tập đối tượng $\Delta U = \{u_n, u_{n-1}, \ldots, u_{n-z+1}\}$ được loại bỏ khỏi bảng, trong đó $z \geq 1$ và $U^- = U \setminus \Delta U$ là một tập đối tượng loại bỏ, khi đó:

1. Nếu tất cả các đối tượng trong ΔU có giá trị thuộc tính quyết định giống nhau thì:

$$\ddot{\mathcal{D}}\left(U^{-}/\ddot{\mathcal{R}}_{C}^{\alpha,\beta}, U^{-}/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right) = \frac{n^2\ddot{\mathcal{D}}\left(U/\ddot{\mathcal{R}}_{C}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right)}{(n-z)^2} - 2\sum_{i=1}^{z}\frac{\left(\left|[\ddot{u}_{n-i+1}]_{C}^{\alpha,\beta}\right| - \left|[\ddot{u}_{n-i+1}]_{C}^{\alpha,\beta} \cap [\ddot{u}_{n-i+1}]_{D}^{\alpha,\beta}\right|\right)}{(n-z)^2} \tag{2.12}$$

2. $\ddot{\mathcal{D}}\left(U^{-}/\ddot{\mathcal{R}}_{A}^{\alpha,\beta}, U^{-}/\ddot{\mathcal{R}}_{A\cup D}^{\alpha,\beta}\right) = \ddot{\mathcal{D}}\left(U^{-}/\ddot{\mathcal{R}}_{C}^{\alpha,\beta}, U^{-}/\ddot{\mathcal{R}}_{C\cup D}^{\alpha,\beta}\right)$

Chứng minh. Tính chất 1 trong trong Mệnh đề 2.11 có thể dễ dàng chứng minh tương tự như tính chất 1 của Mệnh đề 2.9. Đối với tính chất 2, vì $A \subseteq C$, nên với mọi $u \in U$, ta có $\left|[\ddot{u}]_{A}^{\alpha,\beta}\right| - \left|[\ddot{u}]_{A\cup D}^{\alpha,\beta}\right| \geq \left|[\ddot{u}]_{C}^{\alpha,\beta}\right| - \left|[\ddot{u}]_{C\cup\mathcal{D}}^{\alpha,\beta}\right|$. Mặt khác, A là một rút gọn của C nên $\sum_{u\in U}\left(\left|[\ddot{u}]_{A}^{\alpha,\beta}\right| - \left|[\ddot{u}]_{A\cup D}^{\alpha,\beta}\right|\right) = \sum_{u\in U}\left(\left|[\ddot{u}]_{C}^{\alpha,\beta}\right| - \left|[\ddot{u}]_{C\cup D}^{\alpha,\beta}\right|\right)$. Khi đó, $\left|[\ddot{u}]_{A}^{\alpha,\beta}\right| - \left|[\ddot{u}]_{A\cup D}^{\alpha,\beta}\right| = \left|[\ddot{u}]_{C}^{\alpha,\beta}\right| - \left|[\ddot{u}]_{C\cup D}^{\alpha,\beta}\right|$ với mọi $u \in U$. Do đó, tính chất 2 của mệnh đề được chứng minh. □

Mệnh đề trên trình bày về một số trường hợp đặc biệt của bảng quyết định khi loại bỏ một tập đối tượng. Tất nhiên, việc kiểm chứng xem các đối tượng được loại bỏ hay bổ sung vào bảng là rất quan trọng. Điều này sẽ giúp cho các thuật toán gia tăng xử lý nhanh và đạt được hiệu quả cao. Mặt khác, tính chất 2 của mệnh đề này cung cấp cơ sở để đề xuất một điều kiện dừng cho thuật toán gia tăng và đảm bảo tính tối ưu của kích thước rút gọn mới trên bảng quyết định.

Trên cơ sở được trình bày, luận án sẽ xây dựng Thuật toán 2.3 nhằm tìm kiếm một rút gọn xấp xỉ khi bảng quyết định loại bỏ các đối tượng. Thuật toán này được thực hiện thông qua hai giai đoạn chính. Trong giai đoạn đầu tiên, thuật toán sẽ cập nhật các phân hoạch mờ trực cảm mức α, β của tập thuộc tính C cùng với các thuộc tính điều kiện trong bảng quyết định mới. Tiếp theo, trong giai đoạn sau, thuật toán sẽ xem xét tập rút gọn đã thu được từ giai đoạn trước và loại bỏ từng thuộc tính không thỏa mãn tính chất của một rút gọn.

Để phân tích độ phức tạp của thuật toán IARPD-RO, giả sử rằng, $|\Delta U|$ biểu diễn số lượng các đối tượng được loại bỏ từ bảng quyết định và $\left|U^{-}\right|$ biểu diễn tổng số đối tượng trên bảng quyết định mới. Độ phức tạp trong dòng 2 của thuật toán đề xuất là $O\left(\left|\mathcal{A}^{-}\right|\left|U^{-}\right||\Delta U|\right)$ và quá trình tính toán gia tăng tại dòng 3 là $O\left(\left|U^{-}\right||\Delta U|\right)$. Tương tự như giai đoạn loại bỏ các thuộc tính dư thừa trong Thuật toán 2.2, độ phức tạp trong vòng lặp for của IARPD-RO là $O\left(\left|\mathcal{A}^{-}\right|\left|U^{-}\right||\Delta U|\right)$. Như vậy, độ phức tạp trên toàn bộ thuật toán IARPD-RO là $O\left(\left|\mathcal{A}^{-}\right|\left|U^{-}\right||\Delta U|\right)$. Với độ phức tạp này, thuật toán gia tăng

Thuật toán 2.3 Rút gọn thuộc tính gia tăng khi loại bỏ tập đối tượng dựa trên khoảng cách phân hoạch mờ trực cảm mức α, β (IARPD-RO)

Đầu vào: Bảng quyết định $IS = (U, C \cup D)$ với $U = \{u_1, u_2, \ldots, u_n\}$, rút gọn $\mathcal{A} \subseteq C$, các phân hoạch $U/\ddot{\mathcal{R}}_{\mathcal{A}}^{\alpha,\beta}$, $U/\ddot{\mathcal{R}}_{C}^{\alpha,\beta}$ và tập đối tượng loại bỏ $\Delta U = \{u_{n+1}, u_{n+2}, \ldots, u_{n+z}\}$.

Đầu ra: Rút gọn xấp xỉ $\mathcal{A}^-$ trên $U^- = U \cup \Delta U$

// khởi tạo và cập nhật

1: $\mathcal{A}^- \leftarrow \mathcal{A}$

2: cập nhật $U^-/\ddot{\mathcal{R}}_{\mathcal{A}}^{\alpha,\beta}$, $U^-/\ddot{\mathcal{R}}_{C}^{\alpha,\beta}$ and $U^-/\ddot{\mathcal{R}}_{\{a\}}^{\alpha,\beta}$, với mọi $a \in C \backslash \mathcal{A}$

3: tính toán $\ddot{\mathcal{D}}\left(U^-/\ddot{\mathcal{R}}_{\mathcal{A}^-}^{\alpha,\beta}, U^-/\ddot{\mathcal{R}}_{\mathcal{A}^- \cup D}^{\alpha,\beta}\right)$ và $\ddot{\mathcal{D}}\left(U^-/\ddot{\mathcal{R}}_{C}^{\alpha,\beta}, U^-/\ddot{\mathcal{R}}_{C \cup D}^{\alpha,\beta}\right)$ bởi Công thức 2.11.

// loại bỏ thuộc tính dư thừa

4: **for** $a \in \mathcal{A}^-$ **do**

5: tính toán $\ddot{\mathcal{D}}\left(U^-/\ddot{\mathcal{R}}_{\mathcal{A}^- \backslash \{a\}}^{\alpha,\beta}, U^-/\ddot{\mathcal{R}}_{\mathcal{A}^- \backslash \{a\} \cup D}^{\alpha,\beta}\right)$ theo Công thức 2.9.

6: **if** $\ddot{\mathcal{D}}\left(U^-/\ddot{\mathcal{R}}_{\mathcal{A}^- \backslash \{a\}}^{\alpha,\beta}, U^-/\ddot{\mathcal{R}}_{\mathcal{A}^- \backslash \{a\} \cup D}^{\alpha,\beta}\right) = \ddot{\mathcal{D}}\left(U^-/\ddot{\mathcal{R}}_{\mathcal{A}^-}^{\alpha,\beta}, U^-/\ddot{\mathcal{R}}_{\mathcal{A}^- \cup D}^{\alpha,\beta}\right)$ **then**

7: $\mathcal{A}^- \leftarrow \mathcal{A}^- \backslash \{a_0\}$

8: **end if**

9: **end for**

10: **return** $\mathcal{A}^-$

IARPD-RO cho thấy thời gian thực thi được giảm thiểu đáng kể khi so sánh với thuật toán ARPD.

2.4 Thử nghiệm và đánh giá các thuật toán đề xuất

Trong phần này, một số thực nghiệm sẽ được trình bày nhằm chứng minh hiệu quả của các thuật toán được đề xuất. Quá trình đánh giá được thực hiện bằng cách so sánh kết quả của các thuật toán dựa trên một số tiêu chí, bao gồm kích thước rút gọn thu được, độ chính xác phân lớp và thời gian thực thi. Toàn bộ các thực nghiệm được thực hiện trên một máy tính với bộ xử lý Intel Xeon CPU E3-1545M v5 2,90 GHz (8 CPUs), 16 GB bộ nhớ, sử dụng ngôn ngữ lập trình Python.

2.4.1 Hiệu năng của thuật toán IARPD-AO

Quá trình thực nghiệm được thực hiện trên một số bộ dữ liệu tiêu chuẩn được thu thập từ hai kho dữ liệu uy tín là UCI và OpenML như trong Bảng 2.1. Đầu tiên, các bộ dữ liệu này được chia thành hai phần xấp xỉ nhau, ký hiệu là U_{ori} và U_{inc}. Trong đó, tập dữ liệu U_{ori} được sử dụng cho các thuật toán ARPD, F-FDAR [81], ARIFPD

[91], IFPR [55], NIFS [82] và FMIFRFS [56] để tìm một rút gọn trên bảng quyết định cố định. Trong đó, các thuật toán theo hướng tiếp cận tập thô mờ bao gồm NIFS và F-FDAR, các thuật toán ARIFPD, IFPR và FMIFRFS là các thuật toán theo hướng tiếp cận tập thô mờ trực cảm.

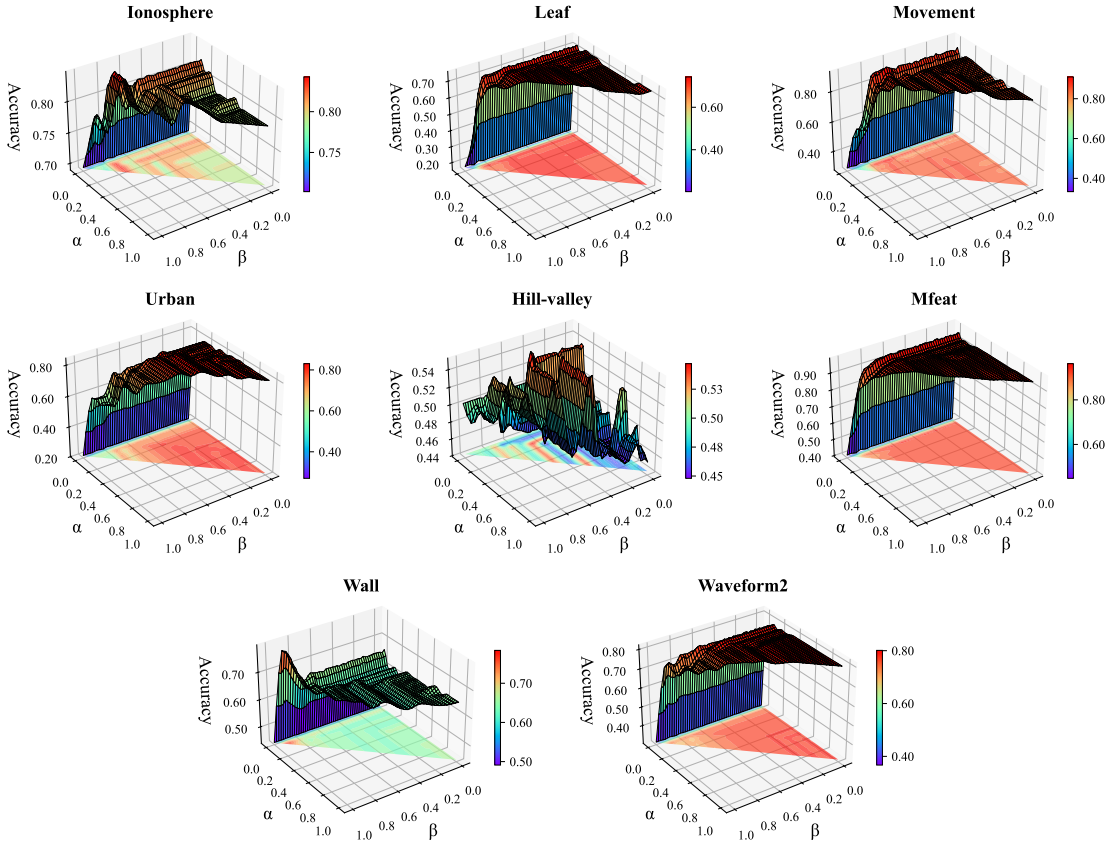
Bảng 2.1: Các tập dữ liệu thử nghiệm cho IARPD-AO và một số thuật toán

TT	Tập dữ liệu	Số đối tượng	$\|U_{ori}\|$	$\|U_{inc}\|$	Số thuộc tính	Số lớp	Nguồn dữ liệu
1	Ionosphere	351	175	176	34	2	UCI
2	Leaf	340	170	170	15	30	UCI
3	Movement	360	180	180	90	15	UCI
4	Urban	675	337	338	147	9	OpenML
5	Hill-valley	1212	606	606	100	2	UCI
6	Mfeat	2000	1000	1000	76	10	OpenML
7	Wall	5456	2728	2728	24	4	OpenML
8	Waveform2	5000	2500	2500	40	3	UCI

Tiếp theo, tập U_{inc} sẽ được chia thành năm phần bằng nhau, ký hiệu lần lượt từ U_1 đến U_5. Những phần này sau đó sẽ được bổ sung từng bước vào U_{ori} để đánh giá tập rút gọn thu được từ các thuật toán gia tăng. Các ký hiệu $|U_{ori}|$ và $|U_{inc}|$ tương ứng là số lượng các đối tượng trong tập U_{ori} và U_{inc}.

Đối với Thuật toán 2.1, các giá trị α và β được xem như các tham số và được duyệt trong khoảng từ 0 tới 1 với bước nhảy là 0.05 để xây dựng các phân hoạch mờ trực cảm mức α, β. Sự thay đổi của các tham số này dẫn đến việc hình thành các hạt thông tin mờ trực cảm khác nhau, từ đó ảnh hưởng trực tiếp đến hiệu quả của thuật toán. Ảnh hưởng này được thể hiện trong Hình 2.3, qua đó cho thấy sự tác động của các tham số đến hiệu quả phân lớp thông qua các rút gọn thu được. Các giá trị α, β sau đó được lưu trữ như các biến nhằm xây dựng các hạt thông tin mới trong các bước gia tăng tiếp theo, bao gồm việc bổ sung hay loại bỏ tập đối tượng. Cần lưu ý rằng, với mỗi giá trị α, khoảng duyệt của β sẽ phải thỏa mãn điều kiện $\alpha + \beta \leq 1$. Qua đó, một rút gọn tối ưu sẽ được lựa chọn và áp dụng làm đầu vào cho các thuật toán gia tăng.

Dựa trên Hình 2.3, với mỗi mức α và β, thuật toán sẽ tìm được một rút gọn khác nhau. Nếu giá trị α lớn và β nhỏ, số lượng đối tượng trong các hạt thông tin mờ trực cảm sẽ bị loại bỏ càng nhiều. Điều này làm giảm các giá trị khoảng cách của mỗi thuộc tính và khiến thuật toán hội tụ nhanh hơn. Do đó, kích thước rút gọn có thể sẽ càng nhỏ và ảnh hưởng tới chất lượng của mô hình phân lớp. Rõ ràng, sẽ rất khó để lựa chọn



Hình 2.3: Độ chính xác phân lớp của ARPD khi duyệt các giá trị tham số trên U_{ori}

một bộ tham số phù hợp cho tất cả các bộ dữ liệu. Tuy nhiên, với những bộ dữ liệu chứa các đối tượng có sự phân bố khác biệt so với phần lớn các đối tượng trong tập vũ trụ, việc lựa chọn một giá trị α lớn và β nhỏ sẽ giúp loại bỏ tốt những đối tượng này. Về mặt trực cảm, những đối tượng này được tạo ra bởi nhiễu và làm giảm độ chính xác phân lớp trên các mô hình học máy.

Kết quả của các thuật toán khi xử lý trên bảng quyết định cố định được biểu diễn trong Bảng 2.2 và 2.3. Bảng 2.2 mô tả kích thước các rút gọn và thời gian thực thi của các thuật toán. Trong đó, cột *time* biểu diễn thời gian thực thi và cột $|\mathcal{A}|$ là kích thước rút gọn thu được từ các thuật toán. Có thể thấy rằng, các thuật toán theo hướng tiếp cận tập thô mờ có thời gian xử lý nhanh hơn so với các thuật toán theo hướng tiếp cận tập thô mờ trực cảm. Thuật toán NIFS có thời gian thực thi nhanh nhất trên tám bộ dữ liệu. Nguyên nhân chính là việc NIFS không cần tính toán các ma trận quan hệ và sử dụng một độ đo đơn giản để tìm kiếm các thuộc tính quan trọng. Nếu chỉ xét trong không gian tập mờ trực cảm, không ngạc nhiên khi thuật toán đề xuất có thời gian thực thi tốt nhất. Thời gian xử lý trung bình của thuật toán đề xuất là 13.590 giây, trong khi

Bảng 2.2: Kích thước rút gọn, thời gian xử lý của ARPD và các thuật toán trên U_{ori}

Tập dữ liệu	ARPD		FMIFRFS		ARIFPD		IFPR		NIFS		F-FDAR	
	Time	$\|\mathcal{A}\|$	*Time*	$\|\mathcal{A}\|$	*Time*	$\|\mathcal{A}\|$	*Time*	$\|\mathcal{A}\|$	*Time*	$\|\mathcal{A}\|$	*Time*	$\|\mathcal{A}\|$
Ionosphere	0.0192	5	0.1257	**4**	0.0846	14	0.0902	9	**0.0033**	9	0.0348	14
Leaf	0.0186	9	0.1242	10	0.0171	9	0.0174	12	**0.0028**	10	0.0049	**8**
Movement	0.1123	**9**	1.1194	21	0.4815	20	0.3383	19	**0.0162**	54	0.0908	21
Urban	3.4501	48	9.1160	47	4.2266	**35**	30.652	43	1.2784	74	**0.6702**	38
Hill-valley	1.2533	**2**	1.3471	3	17.322	5	1.3789	4	**0.0521**	4	3.4982	24
Mfeat	23.382	**22**	36.674	23	56.114	30	30.890	29	**6.8312**	56	6.3366	35
Wall	8.0352	3	15.774	4	38.941	8	5.2763	**2**	**1.0913**	8	5.5972	14
Waveform2	72.449	24	57.093	**15**	119.17	24	78.322	38	**1.6991**	26	14.378	28
Trung bình	13.590	**15.3**	15.172	15.9	29.545	16.3	18.371	19.5	**1.3718**	30.1	3.8263	22.8

với ARIFPD, FMIFRFS và IFPR lần lượt là 29.545 giây, 15.172 giây và 18.371 giây. Do đó, trong không gian tập thô mờ trực cảm, thuật toán đã chứng minh được khả năng giảm thiểu thời gian xử lý hiệu quả.

Từ kết quả Bảng 2.2, có thể thấy rằng các rút gọn thu được từ ARPD và FMIFRFS có kích thước nhỏ nhất trên hầu hết các bộ dữ liệu. Phần lớn, các thuật toán theo hướng tiếp cận tập mờ trực cảm thu được các rút gọn có kích thước nhỏ hơn so với các thuật toán theo hướng tiếp cận tập thô mờ, đặc biệt trên các bộ dữ liệu có độ phân lớp ban đầu thấp như Hill-valley, Mfeat, Urban và Movement. Kích thước trung bình của các rút gọn từ thuật toán đề xuất là nhỏ nhất. Điều này nhấn mạnh rằng thuật toán đề xuất có hiệu quả cao trong việc rút gọn thuộc tính, đặc biệt khi xử lý cho các dữ liệu nhiều chiều.

Để đánh giá hiệu quả về độ chính xác của các rút gọn thu được, mỗi bộ dữ liệu được chia thành mười phần có kích thước xấp xỉ nhau. Một phần ngẫu nhiên sẽ được chọn để kiểm chứng và các phần còn lại được sử dụng để huấn luyện. Quá trình này sẽ được lặp lại mười lần để thu được một kết quả phân lớp cuối cùng được biểu diễn dưới dạng $x \pm \sigma$, trong đó x là giá trị trung bình sau mười lần thực thi và σ là độ lệch chuẩn. Kết quả Bảng 2.3 cho thấy, có 7 trong 8 trường hợp độ chính xác phân lớp từ tập rút gọn của thuật toán đề xuất cao hơn so với các thuật toán khác. Điều này chứng tỏ phương pháp được đề xuất có khả năng chọn lọc hiệu quả các thuộc tính quan trọng. Đặc biệt, đối với những bộ dữ liệu chứa nhiều nhiễu như Hill-valley, Wall và Waveform2, các thuật toán dựa trên tập thô mờ trực cảm tỏ ra ưu thế hơn so với các phương pháp dựa trên tập thô mờ. Tuy nhiên, khi xét riêng trong không gian tập mờ trực cảm, thuật toán ARPD

Bảng 2.3: So sánh độ chính xác phân lớp của ARPD với một số thuật toán trên U_{ori}

Tập dữ liệu	Tập gốc	ARPD	FMIFRFS	ARIFPD	IFPR	NIFS	F-FDAR
Ionosphere	0.766 ± 0.092	**0.841 ± 0.097**	0.840 ± 0.112	0.789 ± 0.121	0.817 ± 0.068	0.822 ± 0.089	0.812 ± 0.113
Leaf	0.718 ± 0.101	**0.741 ± 0.096**	0.724 ± 0.118	**0.741 ± 0.096**	**0.741 ± 0.106**	**0.741 ± 0.124**	0.735 ± 0.121
Movement	0.867 ± 0.090	**0.883 ± 0.091**	**0.883 ± 0.084**	0.856 ± 0.097	0.856 ± 0.114	0.867 ± 0.112	0.867 ± 0.097
Urban	0.801 ± 0.065	**0.828 ± 0.054**	0.813 ± 0.051	0.810 ± 0.059	0.778 ± 0.052	0.804 ± 0.048	0.810 ± 0.066
Hill-valley	0.510 ± 0.037	**0.544 ± 0.040**	0.525 ± 0.055	0.500 ± 0.052	0.517 ± 0.046	0.490 ± 0.049	0.497 ± 0.040
Mfeat	0.925 ± 0.018	**0.949 ± 0.016**	0.945 ± 0.019	0.947 ± 0.017	0.939 ± 0.020	0.910 ± 0.019	0.942 ± 0.013
Wall	0.638 ± 0.094	**0.750 ± 0.045**	0.746 ± 0.064	0.682 ± 0.070	0.724 ± 0.067	0.684 ± 0.089	0.650 ± 0.076
Waveform2	0.782 ± 0.020	0.800 ± 0.022	**0.817 ± 0.013**	0.800 ± 0.022	0.762 ± 0.020	0.783 ± 0.022	0.787 ± 0.017
Trung bình	0.751 ± 0.065	**0.792 ± 0.058**	0.787 ± 0.065	0.766 ± 0.067	0.767 ± 0.062	0.763 ± 0.069	0.763 ± 0.068

vẫn thể hiện khả năng cải thiện độ chính xác phân lớp một cách vượt trội, mặc dù rút gọn thu được có kích thước nhỏ hơn, như được minh họa trong Bảng 2.2.

Tiếp theo, thuật toán IARPD-AO sẽ được so sánh với một số thuật toán gia tăng dựa trên cách tiếp cận tập thô mờ (IF-FDAR-AO [81], AIFSA-FKD [82]) và tập mờ trực cảm (ARIFPD-AO [91]). Trong đó, IF-FDAR-AO là một thuật toán lọc sử dụng khoảng cách phân hoạch mờ với đầu vào là tập rút gọn thu được từ thuật toán F-FDAR. Thuật toán AIFSA-FKD là thuật toán lọc sử dụng khoảng cách tri thức mờ với đầu vào là tập rút gọn thu được từ thuật toán NIFS. Cuối cùng, ARIFPD-AO là thuật toán dựa trên khoảng cách phân hoạch mờ trực cảm sử dụng rút gọn từ ARIFPD làm đầu vào.

Như đã trình bày, quá trình xử lý của các thuật toán gia tăng sẽ diễn ra trong năm giai đoạn bổ sung tập đối tượng. Từ các kết quả của Bảng 2.4, có thể thấy thấy các thuật toán gia tăng có thời gian xử lý nhanh hơn rất nhiều so với các thuật toán thực thi trên bảng quyết định cố định, mặc dù số lượng đối tượng lớn hơn. Tương tự như

khi so sánh với các thuật toán trong Bảng 2.2, thời gian thực thi của các thuật toán gia tăng dựa trên cách tiếp cận tập thô mờ vẫn cho thấy khả năng vượt trội. Cụ thể, thuật toán AIFSA-FKD có thời gian thực thi nhanh nhất trong bốn thuật toán do quá trình trích xuất các thuộc tính quan trọng sử dụng độ đo khoảng cách tri thức mờ bỏ qua bước tính toán các ma trận quan hệ. Thêm vào đó, thuật toán này cũng sử dụng một cơ chế tăng tốc thông qua việc loại bỏ các hạt thông tin dư thừa.

Khi so sánh với các thuật toán theo hướng tiếp cận mờ trực cảm, thuật toán đề xuất cho thấy thời gian thực thi nhanh hơn trên phần lớn các bước gia tăng. Điều này nhấn mạnh tầm ảnh hưởng của tập lát cắt α, β trong việc loại bỏ các đối tượng nhiễu trong các hạt thông tin mờ trực cảm giúp thu hẹp không gian tìm kiếm. Do đó, thuật toán IARPD-AO đảm bảo khả năng xử lý tốt trên các bộ dữ liệu nhiễu và nhiều chiều.

Theo kết quả trong Bảng 2.4, kích thước rút gọn từ các thuật toán dựa trên mô hình tập mờ trực cảm là nhỏ hơn so với các thuật toán dựa trên mô hình tập thô mờ trong hầu hết các bộ dữ liệu. Thuật toán đề xuất thu được kích thước tập rút gọn trung bình nhỏ nhất trong số bốn thuật toán được so sánh. Đặc biệt, đối với các bộ dữ liệu đạt hiệu quả phân lớp ban đầu thấp như Movement, Hill-valley và Wall, số lượng thuộc tính thu được từ IARPD-AO là rất nhỏ. Kết quả này cho thấy IARPD-AO có khả năng loại bỏ hiệu quả các thuộc tính dư thừa trong dữ liệu nhiễu vượt trội hơn so với các thuật toán còn lại. Bên cạnh đó, trong một số bước gia tăng, tập rút gọn của thuật toán không có sự thay đổi. Điều này là do các đối tượng mới được bổ sung không làm biến đổi đặc trưng của dữ liệu. Do đó, tập rút gọn vẫn bảo toàn đầy đủ thông tin của toàn bộ dữ liệu, đồng thời duy trì kích thước và tính ổn định.

Để đảm bảo tính hiệu quả, luận án sẽ đánh giá và so sánh độ chính xác phân lớp giữa các rút gọn thu được của IARPD-AO với các rút gọn thu được từ các thuật toán khác. Các kết quả từ Bảng 2.5 cho thấy trong phần lớn các bước gia tăng, các thuật toán dựa trên mô hình tập mờ trực cảm đạt độ chính xác phân lớp cao hơn so với các thuật toán theo hướng tiếp cận tập thô mờ. Điều này khẳng định vai trò quan trọng của thành phần độ không thuộc trong việc điều chỉnh thông tin từ các đối tượng, giúp các độ đo lựa chọn hiệu quả các thuộc tính cần thiết và nâng cao khả năng phân lớp. Kết quả này đã dẫn tới sự cạnh tranh rõ rệt giữa thuật toán đề xuất và thuật toán ARIFPD-AO. Trong toàn bộ các trường hợp, thuật toán đề xuất đạt độ chính xác phân lớp cao nhất ở 31 trường hợp, trong khi ARIFPD-AO chỉ đạt kết quả tương ứng ở 13 trường hợp.

Phương pháp kiểm định t-tests phụ thuộc được tiến hành với mức tin cậy là 0.95 để

Bảng 2.4: Thời gian xử lý, kích thước rút gọn của IARPD-AO và các thuật toán gia tăng

Tập dữ liệu	Tập dữ liệu bổ sung	IARPD-AO		ARIFPD-AO		IF-FDAR-AO		AIFSA-FKD	
		Time	$\|\mathcal{A}^+\|$	*Time*	$\|\mathcal{A}^+\|$	*Time*	$\|\mathcal{A}^+\|$	*Time*	$\|\mathcal{A}^+\|$
Ionosphere	$\|U_1\| = 210$	0.004	**5**	0.018	14	0.006	19	**0.002**	13
	$\|U_2\| = 245$	0.010	**6**	0.025	15	**0.001**	20	**0.001**	14
	$\|U_3\| = 280$	0.006	**6**	0.038	16	**0.001**	21	**0.001**	15
	$\|U_4\| = 315$	0.007	**6**	0.036	16	0.001	21	**0.001**	16
	$\|U_5\| = 351$	0.007	**6**	0.037	16	0.001	21	**0.001**	16
Leaf	$\|U_1\| = 204$	0.007	10	0.007	10	0.017	10	**0.001**	11
	$\|U_2\| = 238$	0.012	11	0.008	10	0.002	**10**	**0.001**	11
	$\|U_3\| = 272$	0.012	11	0.009	10	0.001	11	**0.001**	12
	$\|U_4\| = 306$	0.013	11	0.011	11	0.001	11	**0.001**	12
	$\|U_5\| = 340$	0.017	12	0.012	11	0.002	12	**0.001**	13
Movement	$\|U_1\| = 216$	0.018	**10**	0.032	20	0.032	29	**0.001**	54
	$\|U_2\| = 252$	0.012	10	0.042	19	0.004	30	**0.002**	54
	$\|U_3\| = 288$	0.054	11	0.066	20	0.004	31	**0.002**	54
	$\|U_4\| = 324$	0.029	11	0.061	19	**0.001**	31	0.002	54
	$\|U_5\| = 360$	0.019	11	0.070	18	**0.001**	31	0.002	54
Urban	$\|U_1\| = 404$	0.950	48	0.612	**35**	0.204	56	0.003	74
	$\|U_2\| = 471$	1.020	48	0.832	**35**	0.039	59	0.004	74
	$\|U_3\| = 538$	1.247	48	1.779	34	0.035	60	0.004	74
	$\|U_4\| = 605$	1.928	48	2.062	33	0.164	64	0.005	74
	$\|U_5\| = 675$	2.752	48	2.805	33	**0.001**	64	0.005	74
Hill-valley	$\|U_1\| = 727$	0.012	**2**	0.062	4	**0.001**	24	0.002	6
	$\|U_2\| = 848$	0.018	**2**	0.065	4	**0.001**	24	0.003	6
	$\|U_3\| = 969$	0.023	**2**	0.084	4	**0.001**	24	0.003	7
	$\|U_4\| = 1090$	0.029	**2**	0.106	4	**0.001**	24	0.004	8
	$\|U_5\| = 1212$	0.037	**2**	0.134	4	**0.001**	24	0.005	8
Mfeat	$\|U_1\| = 1200$	4.216	**23**	8.454	32	1.639	54	**0.070**	57
	$\|U_2\| = 1400$	5.047	**24**	9.573	32	0.313	59	**0.097**	58
	$\|U_3\| = 1600$	6.989	**26**	12.54	32	0.159	60	**0.100**	59
	$\|U_4\| = 1800$	0.816	**26**	15.78	32	**0.001**	60	0.106	60
	$\|U_5\| = 2000$	8.148	**27**	17.52	32	0.184	61	**0.127**	61
Wall	$\|U_1\| = 3273$	0.581	**3**	11.58	11	0.522	17	**0.007**	8
	$\|U_2\| = 3818$	0.669	**3**	13.60	12	0.153	18	**0.010**	8
	$\|U_3\| = 4363$	7.047	**3**	16.08	12	**0.001**	18	0.010	8
	$\|U_4\| = 4908$	9.520	**3**	21.05	13	**0.001**	18	0.034	9
	$\|U_5\| = 5456$	12.13	**3**	23.96	13	**0.001**	18	0.015	9
Waveform2	$\|U_1\| = 3000$	22.98	**24**	24.05	**24**	0.691	33	**0.119**	27
	$\|U_2\| = 3500$	25.44	24	31.35	**23**	**0.001**	33	0.124	28
	$\|U_3\| = 4000$	28.41	24	37.96	**23**	**0.001**	33	0.120	29
	$\|U_4\| = 4500$	30.29	24	47.35	**23**	**0.001**	33	0.133	30
	$\|U_5\| = 5000$	40.53	24	58.74	**23**	**0.001**	33	0.134	31

đánh giá sự khác biệt giữa IARPD-AO, ARIFPD-AO, IF-FDAR-AO và AIFSA-FKD. Kết quả cho thấy các giá trị p-value (two-tailed) lần lượt là 2.815E-05, 4.060E-08 và 1.039E-9. Những kết quả này là cơ sở vững chắc để khẳng định rằng thuật toán đề xuất có hiệu năng vượt trội hơn hẳn các thuật toán so sánh với ý nghĩa thống kê rõ rệt.

Bảng 2.5: So sánh độ chính xác phân lớp của IARPD-AO với một số thuật toán gia tăng

Tập dữ liệu	Tập dữ liệu gốc	IARPD-AO	ARIFPD-AO	IF-FDAR-AO	AIFSA-FKD
Ionosphere	0.776 ± 0.102	**0.848 ± 0.085**	0.810 ± 0.115	0.800 ± 0.108	0.810 ± 0.104
	0.792 ± 0.098	0.824 ± 0.068	**0.830 ± 0.119**	0.821 ± 0.100	0.825 ± 0.106
	0.807 ± 0.070	0.854 ± 0.049	**0.861 ± 0.092**	0.832 ± 0.078	0.821 ± 0.085
	0.832 ± 0.059	0.873 ± 0.051	**0.874 ± 0.080**	0.848 ± 0.058	0.848 ± 0.079
	0.838 ± 0.064	**0.886 ± 0.046**	0.872 ± 0.071	0.852 ± 0.055	0.849 ± 0.082
Leaf	0.683 ± 0.109	**0.692 ± 0.074**	**0.692 ± 0.074**	0.692 ± 0.074	0.663 ± 0.095
	0.643 ± 0.114	**0.677 ± 0.095**	0.643 ± 0.082	0.643 ± 0.082	0.623 ± 0.095
	0.608 ± 0.118	**0.622 ± 0.106**	0.611 ± 0.088	0.622 ± 0.106	0.597 ± 0.100
	0.588 ± 0.074	**0.588 ± 0.069**	**0.588 ± 0.054**	**0.588 ± 0.069**	0.572 ± 0.073
	0.606 ± 0.063	**0.612 ± 0.055**	0.582 ± 0.061	0.612 ± 0.055	0.606 ± 0.051
Movement	0.851 ± 0.105	**0.856 ± 0.082**	0.838 ± 0.106	0.847 ± 0.102	**0.856 ± 0.100**
	0.770 ± 0.101	**0.790 ± 0.090**	0.778 ± 0.099	0.762 ± 0.104	0.770 ± 0.105
	0.767 ± 0.102	0.743 ± 0.095	**0.767 ± 0.108**	0.750 ± 0.100	0.756 ± 0.104
	0.739 ± 0.110	**0.723 ± 0.098**	0.721 ± 0.122	**0.723 ± 0.104**	0.720 ± 0.102
	0.758 ± 0.117	**0.742 ± 0.105**	0.722 ± 0.131	0.736 ± 0.110	0.733 ± 0.111
Urban	0.782 ± 0.060	**0.807 ± 0.067**	0.777 ± 0.070	0.757 ± 0.061	0.777 ± 0.059
	0.771 ± 0.054	**0.783 ± 0.036**	0.771 ± 0.047	0.762 ± 0.043	0.758 ± 0.052
	0.794 ± 0.035	**0.805 ± 0.039**	0.792 ± 0.033	0.782 ± 0.029	0.788 ± 0.030
	0.810 ± 0.048	**0.812 ± 0.037**	0.807 ± 0.033	0.795 ± 0.032	0.797 ± 0.034
	0.800 ± 0.036	0.796 ± 0.039	0.785 ± 0.047	0.784 ± 0.032	**0.798 ± 0.034**
Hill-valley	0.512 ± 0.075	**0.546 ± 0.065**	0.499 ± 0.055	0.481 ± 0.048	0.493 ± 0.067
	0.512 ± 0.042	**0.526 ± 0.025**	0.470 ± 0.051	0.470 ± 0.044	0.468 ± 0.048
	0.525 ± 0.026	**0.547 ± 0.034**	0.514 ± 0.071	0.491 ± 0.041	0.497 ± 0.041
	0.511 ± 0.034	**0.530 ± 0.024**	0.486 ± 0.041	0.494 ± 0.025	0.498 ± 0.030
	0.531 ± 0.032	0.517 ± 0.053	**0.529 ± 0.049**	0.503 ± 0.038	0.521 ± 0.042
Mfeat	0.910 ± 0.019	**0.937 ± 0.014**	0.933 ± 0.022	0.930 ± 0.016	0.898 ± 0.014
	0.906 ± 0.019	**0.940 ± 0.014**	0.937 ± 0.020	0.921 ± 0.025	0.909 ± 0.013
	0.889 ± 0.022	**0.922 ± 0.018**	**0.922 ± 0.019**	0.898 ± 0.026	0.894 ± 0.015
	0.898 ± 0.021	**0.926 ± 0.017**	**0.926 ± 0.018**	0.906 ± 0.024	0.902 ± 0.014
	0.807 ± 0.014	**0.829 ± 0.019**	0.823 ± 0.010	0.814 ± 0.018	0.803 ± 0.011
Wall	0.690 ± 0.048	**0.788 ± 0.057**	0.717 ± 0.036	0.676 ± 0.027	0.716 ± 0.050
	0.729 ± 0.083	**0.802 ± 0.031**	0.757 ± 0.077	0.725 ± 0.063	0.742 ± 0.077
	0.755 ± 0.064	**0.810 ± 0.041**	0.776 ± 0.067	0.747 ± 0.065	0.755 ± 0.063
	0.766 ± 0.071	**0.829 ± 0.049**	0.794 ± 0.075	0.770 ± 0.068	0.772 ± 0.068
	0.773 ± 0.059	**0.830 ± 0.030**	0.802 ± 0.070	0.770 ± 0.066	0.777 ± 0.063
Waveform2	0.785 ± 0.018	**0.799 ± 0.019**	**0.799 ± 0.019**	0.783 ± 0.019	0.786 ± 0.028
	0.783 ± 0.023	**0.811 ± 0.017**	**0.811 ± 0.017**	0.784 ± 0.015	0.797 ± 0.018
	0.795 ± 0.020	**0.811 ± 0.020**	0.808 ± 0.019	0.795 ± 0.025	0.805 ± 0.022
	0.798 ± 0.016	0.819 ± 0.014	**0.821 ± 0.016**	0.805 ± 0.020	0.804 ± 0.021
	0.801 ± 0.015	0.813 ± 0.021	**0.816 ± 0.015**	0.798 ± 0.022	0.809 ± 0.020

2.4.2 Hiệu năng của thuật toán IARPD-RO

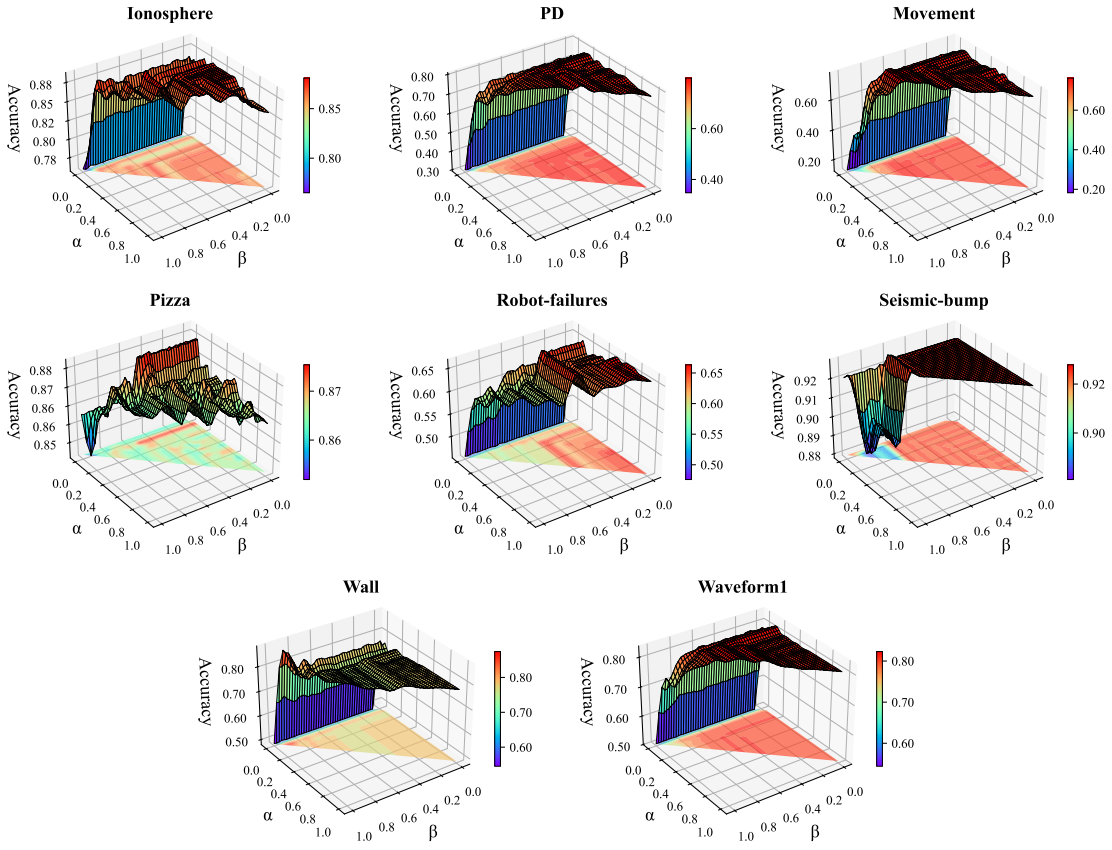
Trong phần này, một số thực nghiệm sẽ được thực hiện nhằm chứng minh hiệu quả của thuật toán gia tăng IARPD-RO trong việc tìm kiếm một rút gọn xấp xỉ trên bảng quyết định có sự loại bỏ tập đối tượng. Các thực nghiệm được tiến hành trên những bộ dữ liệu chuẩn, được mô tả trong Bảng 2.6. Khác với trường hợp áp dụng các thuật toán gia tăng khi bổ sung tập đối tượng, ở đây tập con U_{dec} bao gồm một nửa dữ liệu gốc và được chia thành năm phần gần bằng nhau, ký hiệu từ U_1 đến U_5. Các phần này lần

lượt được loại bỏ khỏi U để áp dụng cho các thuật toán gia tăng.

Bảng 2.6: Các tập dữ liệu thử nghiệm cho IARPD-RO và một số thuật toán

TT	Tập dữ liệu	Số đối tượng	$\|U_{dec}\|$	Số thuộc tính	Số lớp	Nguồn dữ liệu
1	Robot-failures	164	84	90	5	OpenML
2	Ionosphere	351	176	34	2	UCI
3	Movement	360	180	90	15	UCI
4	Pizza	1043	523	35	2	OpenML
5	PD	756	381	754	2	UCI
6	Seismic-bumps	2584	1294	18	2	UCI
7	Waveform1	5000	2500	21	3	UCI
8	Wall	5456	2728	24	4	OpenML

Đầu tiên, các thuật toán ARPD, ARIFPD, IFPR, NIFS và F-FDAR sẽ tìm kiếm rút gọn trên các tập dữ liệu gốc. Riêng với thuật toán ARPD, các giá trị tham số α và β được duyệt trong khoảng từ 0 đến 1 với bước nhảy 0.05 nhằm xác định rút gọn tối ưu. Quá trình duyệt tham số được minh họa trong Hình 2.4 cho thấy ảnh hưởng rõ rệt của các tham số đến kết quả của thuật toán.



Hình 2.4: Độ chính xác phân lớp của ARPD khi duyệt các giá trị tham số trên U

Dựa trên kết quả trong Bảng 2.7, thời gian thực thi của các thuật toán theo tiếp cận tập thô mờ đạt hiệu quả cao hơn so với các thuật toán theo tiếp cận tập mờ trực cảm. Sự chênh lệnh này càng được thể hiện rõ khi xử lý trên các tập dữ liệu lớn.

Tương tự như kết quả trong Bảng 2.2, với ưu điểm khi sử dụng một độ đo đơn giản và cơ chế tăng tốc, thuật toán NIFS có thời gian thực thi nhanh nhất so với các thuật toán khác. Tuy nhiên, nếu chỉ xét trên không gian tập thô mờ trực cảm, thuật toán đề xuất có tốc độ xử lý nhanh hơn ba thuật toán còn lại. Cụ thể, thời gian thực thi trung bình của ARPD trên các bộ dữ liệu là 40.995 giây, trong khi đối với các thuật toàn FMIFRFS, ARIFPD và IFPR thời gian thực thi lần lượt là 52.252 giây, 57.514 giây và 63.478 giây.

Bảng 2.7 trình bày kích thước rút gọn thu được của mỗi phương pháp. Nhìn chung, các thuật toán đều cho thấy khả năng chọn lọc một tập con thuộc tính có kích thước nhỏ hơn đáng kể so với tập dữ liệu ban đầu, đặc biệt trên các bộ dữ liệu như Pizza, PD và Ionosphere. Xét trên toàn bộ các trường hợp, thuật toán ARPD đạt kết quả nổi bật nhất trong bốn trường hợp, với số lượng thuộc tính được trích xuất ít hơn so với các phương pháp còn lại. Đáng chú ý, kích thước rút gọn trung bình của ARPD là 14.0, thấp hơn so với các thuật toán FMIFRFS, ARIFPD, IFPR, NIFS và FDAR, lần lượt là 16.4, 19.0, 17.8, 18.8 và 18.9. Do đó, có thể nhận thấy rằng ARPD có khả năng tìm kiếm các rút gọn tối ưu trên các lớp dữ liệu đa dạng và có tính ổn định trong quá trình xử lý.

Bảng 2.7: Kích thước rút gọn, thời gian xử lý của ARPD và các thuật toán trên U.

Tập dữ liệu	ARPD		FMIFRFS		ARIFPD		IFPR		NIFS		F-FDAR	
	Time	$\|\mathcal{A}\|$	*Time*	$\|\mathcal{A}\|$	*Time*	$\|\mathcal{A}\|$	*Time*	$\|\mathcal{A}\|$	*Time*	$\|\mathcal{A}\|$	*Time*	$\|\mathcal{A}\|$
Robot-failures	0.1280	**15**	0.7683	22	0.1392	16	1.6414	17	**0.0151**	19	0.0711	19
Ionosphere	0.1619	12	0.2356	9	0.1523	15	0.8762	**8**	**0.0084**	14	0.0584	14
Movement	0.4845	**15**	5.5189	18	0.8367	23	2.0156	23	**0.1163**	15	0.2482	20
Pizza	3.3650	**7**	9.6722	19	5.9454	14	5.0071	13	1.2779	22	**1.0891**	15
PD	218.54	41	211.70	**38**	267.13	49	251.56	47	**6.4523**	46	79.811	47
Seismic-bumps	4.1578	3	11.474	9	4.5112	3	1.6542	**2**	**0.1371**	5	1.8836	6
Waveform1	68.391	15	76.427	**9**	72.667	16	107.64	19	**2.9914**	16	19.477	15
Wall	32.730	**4**	102.22	7	108.73	16	137.43	13	**2.8172**	13	23.320	15
Trung bình	40.995	**14.0**	52.252	16.4	57.514	19.0	63.478	17.8	**1.7270**	18.8	15.745	18.9

Tiếp theo, độ chính xác phân lớp của các tập rút gọn thu được từ các thuật toán thực hiện trên bảng quyết định cố định được trình bày trong Bảng 2.8. Có thể nhận

Bảng 2.8: So sánh độ chính xác phân lớp của ARPD với một số thuật toán trên U

Tập dữ liệu	Tập gốc	ARPD	FMIFRFS	ARIFPD	IFPR	NIFS	F-FDAR
Robot-failures	0.535 ± 0.099	**0.658 ± 0.113**	0.622 ± 0.106	**0.658 ± 0.113**	0.609 ± 0.121	0.583 ± 0.133	0.597 ± 0.097
Ionosphere	0.838 ± 0.064	0.880 ± 0.074	0.903 ± 0.053	0.858 ± 0.055	**0.906 ± 0.055**	0.849 ± 0.060	0.860 ± 0.049
Movement	0.758 ± 0.117	0.761 ± 0.130	**0.769 ± 0.119**	0.756 ± 0.114	**0.764 ± 0.115**	0.733 ± 0.116	0.747 ± 0.126
Pizza	0.863 ± 0.022	**0.869 ± 0.027**	**0.869 ± 0.020**	0.865 ± 0.024	0.867 ± 0.020	0.866 ± 0.023	0.866 ± 0.022
PD	0.792 ± 0.061	**0.798 ± 0.036**	0.796 ± 0.046	0.774 ± 0.047	0.780 ± 0.044	0.742 ± 0.070	0.784 ± 0.057
Seismic-bumps	0.908 ± 0.043	0.925 ± 0.028	0.914 ± 0.032	0.925 ± 0.028	**0.932 ± 0.007**	0.913 ± 0.046	0.909 ± 0.036
Waveform1	0.821 ± 0.021	**0.823 ± 0.022**	0.812 ± 0.010	0.816 ± 0.022	0.806 ± 0.023	0.817 ± 0.019	0.814 ± 0.016
Wall	0.773 ± 0.059	**0.831 ± 0.040**	0.816 ± 0.038	0.784 ± 0.062	0.782 ± 0.065	0.773 ± 0.064	0.779 ± 0.074
Trung bình	0.786 ± 0.061	**0.818 ± 0.059**	0.813 ± 0.053	0.805 ± 0.058	0.806 ± 0.056	0.785 ± 0.066	0.795 ± 0.060

thấy rằng hầu hết các thuật toán đều cải thiện đáng kể hiệu quả phân lớp trên phần lớn các tập dữ liệu. Tuy nhiên, đối với các tập dữ liệu có độ chính xác phân lớp ban đầu thấp như Robot-failures, Movement, PD và Wall, các thuật toán dựa trên tiếp cận tập thô mờ cho thấy hiệu quả kém hơn. Điều này phản ánh những thách thức của mô hình tập thô mờ trong việc loại bỏ các thuộc tính chứa thông tin nhiễu.

Trong số sáu thuật toán được so sánh, ARPD đạt độ chính xác phân lớp trung bình cao nhất, mặc dù tập rút gọn thu được là nhỏ nhất. Cụ thể, độ chính xác phân lớp trung bình của ARPD tăng 4,2% so với khi sử dụng toàn bộ tập thuộc tính ban đầu. Nếu chỉ xét trong không gian tập thô mờ trực cảm, ARPD đạt hiệu quả phân lớp cao nhất trong năm trường hợp. Điều này cho thấy tính hiệu quả của thuật toán được đề xuất so với các phương pháp cùng hướng tiếp cận, đặc biệt là trong việc xử lý các tập dữ liệu có nhiễu và độ chính xác phân lớp ban đầu thấp. Cuối cùng, một số phân tích liên quan đến các thuật toán gia tăng trên bảng quyết định có sự loại bỏ tập đối tượng

sẽ được trình bày thông qua việc đánh giá các tiêu chí hiệu năng từ kết quả của ba thuật toán IARPD-DO, IF-FDAR-DO [81] và AIFSD-FKD [82]. Trong đó, IF-FDAR-DO và AIFSD-FKD là các thuật toán gia tăng dựa trên khoảng cách phân hoạch mờ và khoảng cách tri thức mờ. Các thuật toán này sử dụng tập rút gọn đầu vào được sinh ra từ các phương pháp ARPD, IF-FDAR và NIFS tương ứng để thực hiện các bước tính toán gia tăng. Bên cạnh đó, nhằm chứng minh tính hiệu quả của phương pháp đề xuất trong không gian tập thô mờ trực cảm, một thuật toán gia tăng rút gọn thuộc tính dựa trên khoảng cách phân hoạch mờ trực cảm cũng được cài đặt dựa trên nghiên cứu [91]. Thuật toán này được đặt tên là ARIFPD-DO.

Quá trình xử lý của các thuật toán gia tăng loại bỏ được diễn ra trong năm giai đoạn. Các kết quả về thời gian thực thi và kích thước của các rút gọn xấp xỉ được trình bày trong Bảng 2.9. Kết quả cho thấy, các thuật toán gia tăng đạt được tốc độ xử lý nhanh hơn đáng kể so với các thuật toán thực hiện trên bảng quyết định cố định. Ở mỗi bước gia tăng, các thuật toán dựa trên tập thô mờ luôn thể hiện ưu thế rõ rệt về tốc độ so với các thuật toán theo cách tiếp cận tập mờ trực cảm. Đặc biệt, AIFSD-FKD là thuật toán có thời gian thực thi ngắn nhất trong hầu hết các bước, nhờ cơ chế bỏ qua việc tính toán các ma trận quan hệ và sử dụng một độ đo đơn giản để trích chọn thuộc tính.

Khi xem xét các thuật toán trong không gian tập mờ trực cảm, kết quả quan sát cũng hoàn toàn nhất quán với các phân tích trước đó. Trong phần lớn các bước gia tăng, IARPD-DO cho thời gian thực thi nhanh hơn ARIFPD-DO, nhờ đặc tính của tập lát cắt α, β giúp thu hẹp không gian tính toán bằng cách loại bỏ một số đối tượng trong các hạt thông tin mờ trực cảm. Đáng chú ý hơn, ở một số bước gia tăng, thuật toán đề xuất còn đạt thời gian xử lý ngắn nhất. Nguyên nhân xuất phát từ việc nó tạo ra một rút gọn xấp xỉ có kích thước nhỏ nhất, kéo theo số vòng lặp trong quá trình chọn lọc thuộc tính cũng giảm đến mức tối thiểu.

Từ các kết quả trình bày trong Bảng 2.9, có thể nhận thấy thuật toán đề xuất thu được các rút gọn có kích thước nhỏ nhất trong phần lớn các bước loại bỏ tập đối tượng so với các thuật toán còn lại. Chỉ có ba trường hợp rút gọn của thuật toán đề xuất có kích thước lớn hơn so với các thuật toán khác. Trong ba mươi bảy trường hợp còn lại, thuật toán đề xuất đều thu được các rút gọn nhỏ hơn, đặc biệt trong các trường hợp dữ liệu có hiệu quả phân lớp ban đầu thấp.

Cuối cùng, nghiên cứu tiến hành đánh giá hiệu quả phân lớp nhằm làm rõ mức độ

Bảng 2.9: Thời gian chạy, kích thước rút gọn của IARPD-RO và các thuật toán gia tăng

Tập dữ liệu	Tập dữ liệu loại bỏ	IARPD-RO		ARIFPD-DO		IF-FDAR-DO		AIFSD-FKD	
		time	$\lvert\mathcal{A}^-\rvert$	*time*	$\lvert\mathcal{A}^-\rvert$	*time*	$\lvert\mathcal{A}^-\rvert$	*time*	$\lvert\mathcal{A}^-\rvert$
Robot-failures	$\lvert U_1\rvert = 148$	0.007	**14**	0.009	15	0.007	18	**0.004**	18
	$\lvert U_2\rvert = 132$	0.006	**13**	0.009	14	0.005	17	**0.003**	17
	$\lvert U_3\rvert = 116$	0.005	**12**	0.008	13	0.005	17	**0.003**	16
	$\lvert U_4\rvert = 100$	0.005	**11**	0.007	12	0.004	17	**0.002**	15
	$\lvert U_5\rvert = 84$	**0.001**	**11**	0.006	**11**	0.004	16	0.002	14
Ionosphere	$\lvert U_1\rvert = 316$	0.019	**11**	0.031	14	0.011	14	**0.005**	14
	$\lvert U_2\rvert = 281$	**0.002**	**11**	0.025	13	0.011	14	**0.002**	12
	$\lvert U_3\rvert = 246$	0.002	**11**	0.007	13	0.012	13	**0.001**	12
	$\lvert U_4\rvert = 211$	0.002	**11**	0.007	13	0.009	13	**0.001**	12
	$\lvert U_5\rvert = 176$	0.016	**10**	0.006	13	0.009	13	**0.001**	12
Movement	$\lvert U_1\rvert = 324$	0.031	**14**	0.068	22	**0.021**	20	0.006	**14**
	$\lvert U_2\rvert = 288$	0.031	**13**	0.042	21	**0.019**	19	0.005	**13**
	$\lvert U_3\rvert = 252$	0.032	13	0.066	20	**0.019**	18	0.005	**12**
	$\lvert U_4\rvert = 216$	0.026	12	0.052	19	0.018	17	**0.003**	**10**
	$\lvert U_5\rvert = 180$	0.022	11	0.048	18	0.013	16	**0.003**	**9**
Pizza	$\lvert U_1\rvert = 939$	0.002	**7**	0.811	13	**0.001**	15	0.003	22
	$\lvert U_2\rvert = 835$	**0.001**	**7**	0.004	13	**0.001**	15	0.002	22
	$\lvert U_3\rvert = 731$	**0.001**	**7**	0.002	13	**0.001**	15	0.002	22
	$\lvert U_4\rvert = 627$	**0.001**	**7**	**0.001**	13	**0.001**	15	**0.001**	22
	$\lvert U_5\rvert = 523$	**0.001**	**7**	**0.001**	13	**0.001**	15	**0.001**	22
PD	$\lvert U_1\rvert = 681$	3.678	**40**	5.524	48	1.761	46	**0.017**	46
	$\lvert U_2\rvert = 606$	3.514	**39**	5.288	47	1.644	46	**0.014**	46
	$\lvert U_3\rvert = 531$	3.355	**38**	5.012	46	1.525	45	**0.011**	46
	$\lvert U_4\rvert = 456$	3.154	**37**	4.952	45	1.482	44	**0.001**	46
	$\lvert U_5\rvert = 381$	3.110	**36**	4.683	44	1.306	44	**0.001**	46
Seismic-bumps	$\lvert U_1\rvert = 2326$	0.166	**2**	0.172	**2**	**0.001**	6	0.011	4
	$\lvert U_2\rvert = 2068$	0.061	**1**	0.068	**1**	**0.001**	6	0.007	3
	$\lvert U_3\rvert = 1810$	**0.001**	**1**	**0.001**	**1**	**0.001**	6	0.002	2
	$\lvert U_4\rvert = 1552$	**0.001**	**1**	**0.001**	**1**	**0.001**	6	**0.001**	2
	$\lvert U_5\rvert = 1294$	**0.001**	**1**	**0.001**	**1**	**0.001**	6	**0.001**	2
Waveform1	$\lvert U_1\rvert = 4500$	21.45	**14**	24.84	15	7.318	**14**	**0.026**	16
	$\lvert U_2\rvert = 4000$	16.49	**13**	22.37	14	6.479	14	**0.024**	16
	$\lvert U_3\rvert = 3500$	15.87	**12**	18.69	13	4.731	**12**	**0.074**	14
	$\lvert U_4\rvert = 3000$	13.51	**11**	16.73	12	3.914	**11**	**0.058**	13
	$\lvert U_5\rvert = 2500$	11.31	**10**	13.55	11	3.199	**10**	**0.041**	12
Wall	$\lvert U_1\rvert = 4911$	1.450	**3**	29.01	15	6.852	14	**0.135**	13
	$\lvert U_2\rvert = 4366$	1.012	**2**	26.35	14	5.991	13	**0.124**	13
	$\lvert U_3\rvert = 3821$	**0.010**	**2**	22.21	13	5.103	12	0.120	13
	$\lvert U_4\rvert = 3276$	**0.028**	**2**	19.52	12	4.345	11	0.133	12
	$\lvert U_5\rvert = 2731$	**0.015**	**2**	16.23	11	3.551	10	0.135	11

tối ưu đạt được từ các rút gọn. Trong Bảng 2.10, có 13 trong tổng số 40 trường hợp rút gọn từ thuật toán đề xuất có hiệu quả phân lớp kém hơn so với các thuật toán khác. Tuy nhiên, ở các trường hợp còn lại, thuật toán đề xuất chứng minh hiệu quả nổi bật trong việc nâng cao độ chính xác phân lớp. Đặc biệt, đối với các tập dữ liệu không nhất quán và có độ chính xác phân lớp ban đầu thấp như Robot-Failures, Movement và Wall,

độ chính xác phân lớp trên các bước loại bỏ của IARPD-DO cao hơn đáng kể so với các thuật toán so sánh, mặc dù kích thước tập rút gọn thu được nhỏ hơn. Như đã phân tích ở phần trên, cơ chế điều chỉnh nhiễu thông qua tập lát cắt α, β đóng vai trò quan trọng trong việc nâng cao hiệu quả của các độ đo đánh giá thuộc tính. Nhờ đó, các thuộc tính điều kiện được trích xuất bởi IARPD-DO thể hiện khả năng phân lớp vượt trội.

Bảng 2.10: So sánh độ chính xác phân lớp của IARPD-RO với các thuật toán gia tăng

Tập dữ liệu	Tập dữ liệu gốc	IARPD-AO	ARIFPD-AO	IF-FDAR-AO	AIFSA-FKD
Robot -failures	0.573 ± 0.102	**0.669 ± 0.092**	**0.669 ± 0.092**	0.635 ± 0.129	0.560 ± 0.084
	0.575 ± 0.107	**0.659 ± 0.117**	0.636 ± 0.120	0.605 ± 0.068	0.583 ± 0.126
	0.588 ± 0.104	**0.639 ± 0.103**	**0.639 ± 0.103**	0.605 ± 0.092	0.570 ± 0.073
	0.660 ± 0.102	**0.680 ± 0.117**	**0.680 ± 0.117**	0.670 ± 0.078	0.630 ± 0.127
	0.629 ± 0.105	0.640 ± 0.099	0.653 ± 0.088	**0.665 ± 0.110**	0.642 ± 0.144
Ionosphere	0.830 ± 0.057	**0.877 ± 0.066**	0.871 ± 0.059	0.852 ± 0.048	0.842 ± 0.080
	0.808 ± 0.070	0.861 ± 0.091	**0.872 ± 0.075**	0.851 ± 0.065	0.843 ± 0.093
	0.794 ± 0.095	**0.863 ± 0.112**	0.859 ± 0.115	0.846 ± 0.085	0.838 ± 0.081
	0.777 ± 0.109	0.834 ± 0.137	**0.844 ± 0.106**	0.839 ± 0.098	0.810 ± 0.107
	0.773 ± 0.097	0.801 ± 0.123	0.801 ± 0.102	**0.819 ± 0.089**	0.794 ± 0.098
Movement	0.739 ± 0.110	**0.764 ± 0.117**	0.733 ± 0.107	0.727 ± 0.119	0.717 ± 0.096
	0.767 ± 0.102	**0.795 ± 0.119**	0.757 ± 0.097	0.747 ± 0.113	0.771 ± 0.101
	0.770 ± 0.101	**0.814 ± 0.095**	0.758 ± 0.092	0.755 ± 0.111	0.774 ± 0.092
	0.851 ± 0.105	0.838 ± 0.088	0.837 ± 0.098	**0.847 ± 0.088**	0.838 ± 0.063
	0.867 ± 0.090	**0.883 ± 0.072**	0.872 ± 0.086	0.861 ± 0.094	0.861 ± 0.071
Pizza	0.865 ± 0.032	**0.873 ± 0.022**	0.864 ± 0.028	0.864 ± 0.030	0.864 ± 0.031
	0.872 ± 0.033	**0.872 ± 0.026**	0.870 ± 0.032	0.870 ± 0.034	0.870 ± 0.033
	0.866 ± 0.026	**0.866 ± 0.029**	0.863 ± 0.027	**0.866 ± 0.026**	0.863 ± 0.027
	0.857 ± 0.022	**0.874 ± 0.025**	0.852 ± 0.024	0.853 ± 0.025	0.853 ± 0.022
	0.851 ± 0.024	**0.864 ± 0.021**	0.849 ± 0.026	0.853 ± 0.024	0.851 ± 0.024
PD	0.802 ± 0.068	**0.803 ± 0.051**	0.792 ± 0.047	0.789 ± 0.047	0.777 ± 0.056
	0.802 ± 0.069	**0.824 ± 0.064**	0.805 ± 0.062	0.789 ± 0.072	0.802 ± 0.068
	0.801 ± 0.061	0.787 ± 0.081	0.795 ± 0.082	0.791 ± 0.070	**0.801 ± 0.061**
	0.816 ± 0.074	0.812 ± 0.079	0.803 ± 0.099	**0.823 ± 0.088**	0.816 ± 0.074
	0.785 ± 0.079	0.788 ± 0.075	**0.793 ± 0.070**	0.811 ± 0.060	0.785 ± 0.079
Seismic -bumps	0.907 ± 0.042	**0.930 ± 0.002**	0.903 ± 0.081	0.905 ± 0.046	0.912 ± 0.032
	0.900 ± 0.052	**0.926 ± 0.002**	**0.926 ± 0.002**	0.899 ± 0.050	0.902 ± 0.060
	0.889 ± 0.022	**0.920 ± 0.002**	**0.920 ± 0.003**	0.883 ± 0.046	0.919 ± 0.006
	0.878 ± 0.060	**0.918 ± 0.003**	**0.918 ± 0.003**	0.891 ± 0.035	0.910 ± 0.007
	0.871 ± 0.049	**0.906 ± 0.002**	**0.906 ± 0.002**	0.864 ± 0.061	0.893 ± 0.010
Waveform1	0.818 ± 0.022	0.820 ± 0.018	**0.826 ± 0.024**	0.819 ± 0.022	0.820 ± 0.015
	0.820 ± 0.021	0.808 ± 0.029	**0.813 ± 0.024**	0.808 ± 0.029	0.809 ± 0.013
	0.817 ± 0.021	0.815 ± 0.017	**0.816 ± 0.017**	0.815 ± 0.017	0.815 ± 0.022
	0.817 ± 0.024	0.807 ± 0.013	0.802 ± 0.021	0.810 ± 0.023	**0.814 ± 0.016**
	0.812 ± 0.020	0.804 ± 0.017	0.790 ± 0.027	0.794 ± 0.019	**0.808 ± 0.019**
Wall	0.765 ± 0.070	**0.850 ± 0.030**	0.782 ± 0.078	0.782 ± 0.070	0.765 ± 0.063
	0.755 ± 0.065	**0.774 ± 0.034**	0.769 ± 0.068	0.776 ± 0.068	0.746 ± 0.065
	0.728 ± 0.082	**0.764 ± 0.036**	0.744 ± 0.070	0.750 ± 0.076	0.724 ± 0.066
	0.693 ± 0.044	**0.766 ± 0.054**	0.712 ± 0.044	0.715 ± 0.050	0.693 ± 0.042
	0.633 ± 0.096	**0.733 ± 0.069**	0.670 ± 0.075	0.654 ± 0.078	0.644 ± 0.090

Dựa trên độ chính xác phân lớp của các phương pháp trong Bảng 2.10, phương pháp kiểm định t-tests phụ thuộc cũng được tiến hành với mức tin cậy là 0.95 để đánh giá

sự khác biệt giữa IARPD-DO, ARIFPD-DO, IF-FDAR-DO và AIFSD-FKD. Các giá trị tương ứng p-values (two-tailed) được xác định tương ứng là 0.0009, 3.896E-05 và 4.812E-06 đối với bộ phân lớp KNN. Những kết quả này cung cấp cơ sở vững chắc để khẳng định rằng thuật toán đề xuất vượt trội so với các thuật toán được so sánh với ý nghĩa thống kê.

2.5 Kết luận Chương 2

Từ sự mở rộng của khái niệm lát cắt α trong lý thuyết tập mờ, tập lát cắt α, β ban đầu được xây dựng với mục tiêu loại bỏ những đối tượng có độ khác biệt cao và độ tương tự thấp trong các hạt thông tin mờ trực cảm. Những đối tượng này được sinh ra bởi nhiễu và là nguyên nhân chính làm suy giảm hiệu quả của các mô hình phân lớp. Từ việc loại bỏ các đối tượng nhiễu, các hạt thông tin mờ trực cảm mức α, β được hình thành và được xem như một trường hợp tổng quát của các hạt thông tin mờ trực cảm truyền thống. Dựa trên các hạt thông tin này, một phân hoạch mờ trực cảm mức α, β được thiết lập, đóng vai trò nền tảng để xây dựng một độ đo khoảng cách phân hoạch mới. Với đặc trưng của một độ đo khoảng cách tuyệt đối, độ đo đề xuất có khả năng xét đến toàn bộ các đối tượng trong tập vũ trụ. Nhờ đó, việc lựa chọn các thuộc tính quan trọng từ độ đo đã mang lại hiệu quả đáng kể. Trên cơ sở này, nghiên cứu đã định nghĩa lại một rút gọn tối ưu của bảng quyết định và thiết kế một thuật toán rút gọn thuộc tính trên bảng quyết định cố định (ARPD). Rõ ràng, nhờ có cơ chế loại bỏ ảnh hưởng của các đối tượng nhiễu, các hạt thông tin mờ trực cảm mức α, β có ý nghĩa rất cao trong việc biểu diễn thông tin. Do đó, các hạt thông tin này giữ vai trò then chốt trong việc quyết định hiệu quả của thuật toán. Nói cách khác, hiệu quả đạt được của thuật toán xuất phát trực tiếp từ những ưu điểm của mô hình. Một số kết quả thực nghiệm cho thấy thuật toán đạt hiệu năng vượt trội so với các thuật toán dựa trên mô hình tập thô mờ, đặc biệt trên các bộ dữ liệu mà các mô hình phân lớp thường cho hiệu quả thấp. Bên cạnh đó, đối với các bộ dữ liệu có số chiều lớn, thuật toán đề xuất còn cho thấy khả năng cải thiện đáng kể thời gian thực thi so với các thuật toán dựa trên tập thô mờ trực cảm.

Từ các phân tích trên có thể nhận thấy rằng mô hình đề xuất có khả năng áp dụng hiệu quả trên các tập dữ liệu có số chiều lớn, đặc biệt là các tập dữ liệu chịu ảnh hưởng mạnh bởi nhiễu và chứa các đối tượng có sự phân bố khác biệt so với phần lớn các đối tượng trong tập vũ trụ.

Ngoài ra, nhằm đáp ứng các tình huống thực tế của dữ liệu, hai công thức tính

khoảng cách giữa các phân hoạch mờ trực cảm ở mức α, β đã được xây dựng, đóng vai trò nền tảng cho việc đề xuất các thuật toán gia tăng nhằm xử lý các bảng quyết định có sự bổ sung hoặc loại bỏ tập đối tượng. Kết quả thực nghiệm cũng cho thấy các phương pháp này đạt hiệu quả trên phần lớn các bộ dữ liệu nhiễu, thể hiện ở khả năng cải thiện độ chính xác phân lớp cũng như giảm thiểu đáng kể thời gian thực thi trên các bộ dữ liệu có số chiều lớn so với các thuật toán gia tăng tiếp cận dựa trên mô hình tập thô mờ trực cảm.

Từ những kết quả này, một số ưu điểm nổi bật của mô hình đề xuất đã được làm rõ. Ưu điểm đầu tiên đến từ khả năng kế thừa các tính chất quan trọng của tập mờ trực cảm thông qua việc bổ sung thành phần hàm không thuộc giúp điều chỉnh thông tin từ các đối tượng nhiễu trong dữ liệu. Ưu điểm thứ hai đến từ cơ chế loại bỏ hoàn toàn ảnh hưởng của các đối tượng nhiễu thông qua tập lát cắt α, β. Do đó, không gian tính toán được rút gọn và thời gian xử lý được cải thiện tốt hơn. Tuy nhiên, trong thực tế, mỗi thuộc tính điều kiện thường có những ảnh hưởng khác nhau tới quyết định của mỗi đối tượng. Mô hình tập mờ trực cảm mức α, β đã bỏ qua vấn đề này khi chỉ xem xét vai trò của các thuộc tính này như nhau khi thiết lập các hạt thông tin mờ trực cảm. Do đó, trong một số trường hợp các thuật toán không chọn được các thuộc tính thực sự quan trọng. Để giải quyết vấn đề này và một số hạn chế của nhánh mở rộng từ mô hình tập thô lân cận, Chương 3 của luận án sẽ trình bày về một mô hình mới có những ưu điểm nổi trội hơn so với mô hình tập mờ trực cảm mức α, β. Mô hình này có khả năng phản ánh đúng mức độ liên quan giữa từng thuộc tính điều kiện với thuộc tính quyết định và đánh giá chi tiết vai trò của mỗi đối tượng trong một hạt thông tin. Nhờ đó, các thuật toán rút gọn thuộc tính triển khai trên mô hình này được kỳ vọng sẽ mang tới những kết quả ấn tượng hơn.

CHƯƠNG 3
ĐỀ XUẤT MỘT SỐ THUẬT TOÁN RÚT GỌN THUỘC TÍNH DỰA TRÊN TẬP THÔ LÂN CẬN MỜ TRỰC CẢM CÓ TRỌNG SỐ

3.1 Mở đầu

Như đã trình bày trong Chương 2, một số mô hình theo cả hai nhánh mở rộng của lý thuyết tập thô vẫn tồn tại một số hạn chế nhất định. Thứ nhất, một số mô hình thường bỏ qua việc phản ánh mức độ liên quan của từng thuộc tính điều kiện với thuộc tính quyết định. Hệ quả là trong quá trình rút gọn, các thuộc tính có ý nghĩa thực sự đối với quyết định có thể không được ưu tiên lựa chọn, làm suy giảm chất lượng tập rút gọn và hiệu quả phân lớp. Thứ hai, một số mở rộng từ lý thuyết tập thô lân cận có cấu trúc hạt còn đơn giản và chưa biểu diễn chi tiết vai trò của các đối tượng trong một hạt thông tin. Các phương pháp hiện có chủ yếu tập trung vào việc phản ánh sự tương đồng với đối tượng đang xem xét, đồng thời bỏ qua sự không chắc chắn và do dự vốn tồn tại trong dữ liệu thực tế. Thứ ba, một số mô hình tập thô lân cận mở rộng hiện nay chưa tích hợp đồng thời hai cơ chế gán trọng số thuộc tính và trọng số đối tượng, mặc dù sự kết hợp này được kỳ vọng sẽ mang lại những hiệu suất cải tiến đáng kể trong quá trình rút gọn. Nói cách khác, việc thiếu một cơ chế tích hợp thống nhất khiến mô hình khó khai thác hết thông tin của dữ liệu, làm hạn chế khả năng nâng cao hiệu suất rút gọn. Xuất phát từ những hạn chế nêu trên, nội dung của luận án trong Chương 3 sẽ tập trung vào các đóng góp chính như sau:

Thứ nhất, luận án đề xuất mô hình tập thô lân cận mờ trực cảm có trọng số, trong đó mỗi đối tượng được đặc trưng không chỉ bởi độ thuộc mà còn bởi độ không thuộc và độ do dự. Trên cơ sở này, luận án xây dựng khoảng cách giữa hai họ lân cận mờ trực cảm có trọng số nhằm định nghĩa lại một kiểu rút gọn mới làm nền tảng trong việc đề xuất một thuật toán rút gọn thuộc tính trên bảng quyết định cố định theo tiếp cận lọc.

Thứ hai, luận án phát triển hai công thức gia tăng dựa trên độ đo khoảng cách họ lân cận mờ trực cảm có trọng số để áp dụng cho các trường hợp bảng quyết định có sự thay đổi tập đối tượng, đồng thời đề xuất hai thuật toán gia tăng rút gọn thuộc tính trên bảng quyết định thay đổi.

Thứ ba, luận án chứng minh hiệu quả của các thuật toán đề xuất dựa trên việc so

sánh với một số thuật toán theo mô hình tập thô mờ và tập thô lân cận trọng số.

Kết quả nghiên cứu của chương này được công bố trong các công trình [CT6] và [CT7] thuộc phần Danh mục các công trình nghiên cứu của luận án.

3.2 Mô hình tập thô lân cận mờ trực cảm có trọng số

Nhằm khắc phục hạn chế của các mô hình hiện có trong việc phản ánh đúng mức độ liên quan giữa từng thuộc tính điều kiện và thuộc tính quyết định, luận án đề xuất một mô hình mới kết hợp tập thô lân cận có trọng số với tập thô mờ trực cảm. Sự kết hợp này tận dụng ưu điểm của cả hai khung lý thuyết, hứa hẹn nâng cao hiệu quả các thuật toán rút gọn thuộc tính.

Có thể thấy rằng, các mô hình được trình bày chỉ tập trung vào số lượng các đối tượng nằm trong một hạt thông tin. Điều này có thể được hiểu rằng, các đối tượng trong hạt thông tin đều chỉ có một độ quan trọng như nhau trong việc ra quyết định về đối tượng u. Tuy nhiên, dữ liệu luôn có sự phân bố đa dạng trong thực tế, nghĩa là mỗi đối tượng trong một hạt thông tin sẽ đóng một vai trò khác nhau. Ví dụ những đối tượng gần u có vai trò lớn hơn so với các đối tượng nằm xa u. Do đó, việc đưa ra một đánh giá chính xác về độ quan trọng của các đối tượng trong hạt thông tin lân cận trọng số là rất cần thiết. Để giải quyết vấn đề này, luận án sẽ đề xuất một mô hình mới dựa trên tập mờ trực cảm để đánh giá cho các đối tượng trong một lân cận, nghĩa là mỗi một đối tượng sẽ được biểu diễn bởi một độ thuộc và độ không thuộc của nó trong hạt thông tin lân cận trọng số.

3.2.1 Khái niệm về tập thô lân cận mờ trực cảm có trọng số

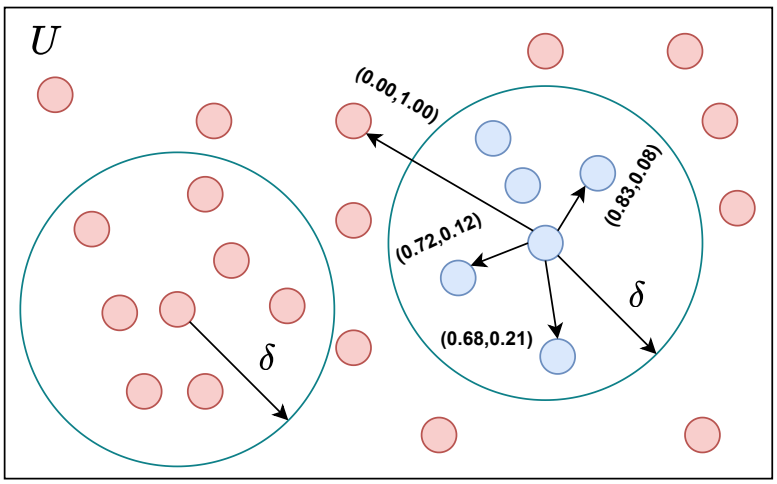
Dựa trên những ưu điểm từ tập thô lân cận trọng số được trình bày, trong phần này, luận án sẽ trình bày về một mô hình mới được gọi là mô hình tập thô lân cận mờ trực cảm có trọng số. Qua đó, một số khái niệm cơ bản của mô hình cũng sẽ được giới thiệu để làm nổi bật những ưu điểm mà mô hình mang lại.

Đầu tiên, cho bảng quyết định $IS = (U, C \cup D)$ và một tập con thuộc tính $A \subseteq C$. Giả sử rằng, $[u]_A^{\delta,\omega} = \{u_1, u_2, \ldots, u_k\}$ với $k \leq |U|$ là một hạt thông tin lân cận trọng số của đối tượng u theo A. Khi đó, các đối tượng trong $[u]_A^{\delta,\omega}$ đều được xem là có mức độ quan trọng như nhau khi đưa ra quyết định cho đối tượng u. Tuy nhiên, vai trò của các đối tượng này là hoàn toàn khác nhau trong việc đánh giá u. Do đó, việc xét tới mức độ quan trọng của mỗi đối tượng trong một hạt thông tin là rất cần thiết. Để giải quyết vấn đề này, nghiên cứu hướng tới việc gán trọng số cho mỗi đối tượng trong một hạt thông tin dựa trên hai thành phần theo đặc trưng của một tập mờ trực cảm. Cụ thể, trọng số

của mỗi đối tượng $u_i \in [u]_A^{\delta,\omega}$ được ký hiệu là $[\ddot{u}]_A^{\delta,\omega}(u_i) = \left(\gamma_{[\ddot{u}]_A^{\delta,\omega}}(u_i), \eta_{[\ddot{u}]_A^{\delta,\omega}}(u_i)\right)$, trong đó, $\gamma_{[\ddot{u}]_A^{\delta,\omega}}(u_i)$ và $\eta_{[\ddot{u}]_A^{\delta,\omega}}(u_i)$ tương ứng là độ thuộc và độ không thuộc của đối tượng u_i trong hạt thông tin $[u]_A^{\delta,\omega}$, được xác định như sau:

- $\gamma_{[\ddot{u}]_A^{\delta,\omega}}(u_i) = 1 - \Delta_A^{\omega}(u, u_i)$, với $\Delta_A^{\omega}(u, u_i)$ là khoảng cách trọng số giữa u và u_i theo Công thức 1.7, trong trường hợp $p = \infty$,

- $\eta_{[\ddot{u}]_A^{\delta,\omega}}(u_i) = \dfrac{1 - \gamma_{[\ddot{u}]_A^{\delta,\omega}}(u_i)}{1 + w^o \times \gamma_{[\ddot{u}]_A^{\delta,\omega}}(u_i)}$, với $w^o = \dfrac{1}{|U|}\sum\limits_{u \in U} \dfrac{\left|[u]_C^{\delta,\omega} \cap [u]_D\right|}{\left|[u]_C^{\delta,\omega}\right|}$ là *tỉ lệ phân lớp trung bình* dựa trên các hạt thông tin lân cận trọng số của C và $[u]_D$ là lớp tương đương của đối tượng u trên D.

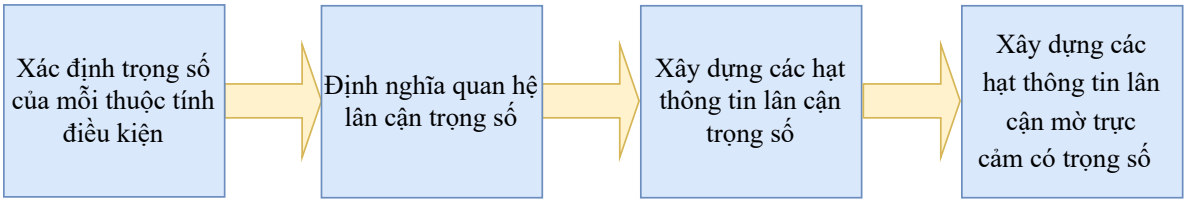


Hình 3.1: Hạt thông tin lân cận mờ trực cảm có trọng số

Rõ ràng, nếu w^o càng lớn thì bảng quyết định được xem là có tính nhất quán càng cao và tầm ảnh hưởng của độ không thuộc sẽ giảm. Ngược lại, nếu bảng quyết định có độ nhất quán thấp thì độ không thuộc sẽ có tầm ảnh hưởng lớn hơn để điều chỉnh các thông tin trên đối tượng nhiễu. Đối với các đối tượng không nằm trong hạt thông tin lân cận trọng số thì $[\ddot{u}]_A^{\delta,\omega}(u_i) = (0, 1)$. Kết hợp toàn bộ các đánh giá này, chúng ta thu được một hạt thông tin mới $[\ddot{u}]_A^{\delta,\omega} = \left\{[\ddot{u}]_A^{\delta,\omega}(u_1), \ldots, [\ddot{u}]_A^{\delta,\omega}\left(u_{|U|}\right)\right\}$, được gọi là hạt thông tin lân cận mờ trực cảm có trọng số và được biểu diễn trong Hình 3.1.

Hình 3.2 trình bày tổng quan về quá trình xây dựng các hạt thông tin lân cận mờ trực cảm có trọng số.

Rất dễ để thấy rằng $0 \leq \gamma_{[\ddot{u}]_A^{\delta,\omega}}(u_i) + \eta_{[\ddot{u}]_A^{\delta,\omega}}(u_i) \leq 1$, với mọi $u_i \in U$ nên $[\ddot{u}]_A^{\delta,\omega}$ có đặc trưng như một tập mờ trực cảm và được gọi là hạt thông tin lân cận mờ trực cảm có trọng số. Khi đó, nếu xét trên toàn bộ các đối tượng trong U, một họ $\mathcal{F}_A^{\delta,\omega} = \left\{[\ddot{u}]_A^{\delta,\omega} : u \in U\right\}$ sẽ được cấu thành và được gọi là họ lân cận mờ trực cảm có trọng số của thuộc tính A.



Hình 3.2: Quá trình xây dựng các hạt thông tin lân cận mờ trực cảm có trọng số

Tiếp theo, luận án sẽ trình bày một số tính chất then chốt của hạt thông tin lân cận mờ trực cảm có trọng số.

Mệnh đề 3.1 Cho bảng quyết định $IS = (U, C \cup D)$ và hai tập con thuộc tính $A, B \subseteq C$. Nếu $p = \infty$, thì $[\ddot{u}]^{\delta,\omega}_{A\cup B} = [\ddot{u}]^{\delta,\omega}_{A} \cap [\ddot{u}]^{\delta,\omega}_{B}$, với mọi $u \in U$.

Từ Mệnh đề 3.1, có thể thấy việc sử dụng khoảng cách trọng số Chebyshev có thể làm tăng vai trò ảnh hưởng của các trọng số thuộc tính điều kiện trong việc thiết lập các hạt thông tin lân cận. Bên cạnh đó, mệnh đề này còn chỉ ra rằng $[\ddot{u}]^{\delta,\omega}_{A} = \cap_{a\in A}[\ddot{u}]^{\delta,\omega}_{\{a\}}$. Đây được xem là một tính chất quan trọng để xây dựng một chiến lược cho việc thiết kế một thuật toán rút gọn thuộc tính trên bảng quyết định cố định.

Ví dụ 3.1 Để xác định các hạt thông tin lân cận mờ trực cảm có trọng số của tập thuộc tính C trong Bảng 1.1 với ngưỡng bán kính $\delta = 0.5$, chúng ta thực hiện các bước sau:

Tính toán trọng số của các thuộc tính điều kiện theo Công thức 1.6 và khoảng cách trọng số giữa các đối tượng trong C.

$\omega(a_1) = 0.13$, $\omega(a_2) = 0.97$, $\omega(a_3) = 1.73$, $\omega(a_4) = 1.42$, $\omega(a_5) = 0.74$.

$\Delta^{\omega}_{C}(u_1, u_2) = \max\{0.13\times0.29, 0.97\times0.01, 1.73\times0.13, 1.42\times0.08, 0.74\times0.12\} = 0.22$.

Tương tự, chúng ta cũng thu được khoảng cách trọng số với các cặp đối tượng còn lại trong bảng.

$\Delta^{\omega}_{C}(u_1, u_3) = 0.36$, $\Delta^{\omega}_{C}(u_1, u_4) = 0.76$, $\Delta^{\omega}_{C}(u_1, u_5) = 0.64$,

$\Delta^{\omega}_{C}(u_2, u_3) = 0.57$, $\Delta^{\omega}_{C}(u_2, u_4) = 0.54$, $\Delta^{\omega}_{C}(u_2, u_5) = 0.61$,

$\Delta^{\omega}_{C}(u_3, u_4) = 1.11$, $\Delta^{\omega}_{C}(u_3, u_5) = 0.99$, $\Delta^{\omega}_{C}(u_4, u_5) = 0.56$.

Xác định các hạt thông tin lân cận trọng số trên tập thuộc tính C:

$[u_1]^{0.5,\omega}_{C} = \{u_1, u_2, u_3\}$, $[u_2]^{0.5,\omega}_{C} = \{u_1, u_2\}$,

$[u_3]^{0.5,\omega}_{C} = \{u_1, u_3\}$, $[u_4]^{0.5,\omega}_{C} = \{u_4\}$, $[u_5]^{0.5,\omega}_{C} = \{u_5\}$.

Tính toán tỉ lệ phân lớp trung bình: $w^o = \frac{1}{5} \times \left(\frac{2}{3} + \frac{1}{2} + 1 + 1 + 1\right) = 0.83$.

Thực hiện tính toán trọng số của các đối tượng trong hạt thông tin lân cận mờ trực cảm có trọng số.

$$\gamma_{[\ddot{u}_1]_C^{0.5,\omega}}(u_2) = 1 - 0.22 = 0.78 \Rightarrow \eta_{[\ddot{u}_1]_C^{0.5,\omega}}(u_2) = \frac{1-0.78}{1+0.83\times 0.78} = 0.13,$$
$$\gamma_{[\ddot{u}_1]_C^{0.5,\omega}}(u_3) = 1 - 0.36 = 0.64 \Rightarrow \eta_{[\ddot{u}_1]_C^{0.5,\omega}}(u_3) = \frac{1-0.64}{1+0.83\times 0.64} = 0.24.$$

Từ đây, chúng ta thu được các hạt thông tin lân cận mờ trực cảm có trọng số của tập thuộc tính C như sau:

$$[\ddot{u}_1]_C^{0,5,\omega} = \left\{\frac{(1.00,0.00)}{u_1}, \frac{(0.78,0.13)}{u_2}, \frac{(0.64,0.24)}{u_3}, \frac{(0.00,1.00)}{u_4}, \frac{(0.00,1.00)}{u_5}\right\}$$
$$[\ddot{u}_2]_C^{0,5,\omega} = \left\{\frac{(0.78,0.13)}{u_1}, \frac{(1.00,0.00)}{u_2}, \frac{(0.00,1.00)}{u_3}, \frac{(0.00,1.00)}{u_4}, \frac{(0.00,1.00)}{u_5}\right\}$$
$$[\ddot{u}_3]_C^{0,5,\omega} = \left\{\frac{(0.64,0.24)}{u_1}, \frac{(0.00,1.00)}{u_2}, \frac{(1.00,0.00)}{u_3}, \frac{(0.00,1.00)}{u_4}, \frac{(0.00,1.00)}{u_5}\right\}$$
$$[\ddot{u}_4]_C^{0,5,\omega} = \left\{\frac{(0.00,1.00)}{u_1}, \frac{(0.00,1.00)}{u_2}, \frac{(0.00,1.00)}{u_3}, \frac{(1.00,0.00)}{u_4}, \frac{(0.00,1.00)}{u_5}\right\}$$
$$[\ddot{u}_5]_C^{0,5,\omega} = \left\{\frac{(0.00,1.00)}{u_1}, \frac{(0.00,1.00)}{u_2}, \frac{(0.00,1.00)}{u_3}, \frac{(0.00,1.00)}{u_4}, \frac{(1.00,0.00)}{u_5}\right\}$$

Mệnh đề 3.2 Cho bảng quyết định $IS = (U, C \cup D)$, tập con thuộc tính điều kiện $A \subseteq C$ và hai bán kính lân cận δ_1 và δ_2. Nếu $\delta_1 \leq \delta_2$ thì $[\ddot{u}]_A^{\delta_1,\omega} \subseteq [\ddot{u}]_A^{\delta_2,\omega}$.

Chứng minh. Từ Công thức 1.7 với $p = \infty$ và $\delta_1 \leq \delta_2$, nếu $\max_{a\in A}\{\omega(a)\cdot|a(u)-a(v)|\} \leq \delta_1$ thì $\max_{a\in A}\{\omega(a)\cdot|a(u)-a(v)|\} \leq \delta_2$, với mọi $u, v \in U$. Do đó, $[u]_A^{\delta_1,\omega} \subseteq [u]_A^{\delta_2,\omega}$ và dễ dàng dẫn đến $[\ddot{u}]_A^{\delta_1,\omega} \subseteq [\ddot{u}]_A^{\delta_2,\omega}$. □

Dựa trên Mệnh đề 3.4, không khó để thấy rằng khi giá trị bán kính lân cận càng nhỏ, số lượng đối tượng trong các hạt thông tin lân cận trọng số không được xét tới sẽ càng ít. Qua đó, việc thay đổi các giá trị bán kính lân cận khác nhau sẽ thu được các hạt thông tin lân cận trọng số mờ trực cảm khác nhau. Nói cách khác, chúng ta hoàn toàn có thể thu hẹp không gian của mỗi hạt thông tin lân cận nhờ việc bỏ qua các đối tượng có sự phân bố khác biệt. Về mặt trực cảm, các đối tượng này được tạo ra bởi nhiễu và ảnh hưởng rất lớn tới kết quả trên các mô hình phân lớp.

Từ các khái niệm về hạt thông tin lân cận trọng số mờ trực cảm, mô hình tập thô lân cận mờ trực cảm có trọng số (IFWNRS) được định nghĩa dựa trên các xấp xỉ trên và xấp xỉ dưới của một tập con đối tượng cho trước. Cụ thể, xét bảng quyết định $IS = (U, C \cup D)$, một tập con thuộc tính $A \subseteq C$ và một tập con $X \subseteq U$. Các xấp xỉ dưới và xấp xỉ trên của X dựa trên hạt thông tin lân cận mờ trực cảm có trọng số theo A, ký hiệu lần lượt là $\underline{IW_A}(X)$ và $\overline{IW_A}(X)$, được xác định tương ứng như sau

$$\underline{IW_A}(X) = \left\{u \in U : \frac{\left|[\ddot{u}]_A^{\delta,\omega} \cap X\right|}{\left|[\ddot{u}]_A^{\delta,\omega}\right|} \geq \ddot{\alpha}\right\} \tag{3.1}$$

và

$$\overline{IW_A}(X) = \left\{ u \in U : \frac{\left|[\ddot{u}]_A^{\delta,\omega} \cap X\right|}{\left|[\ddot{u}]_A^{\delta,\omega}\right|} \geq \ddot{\beta} \right\} \tag{3.2}$$

trong đó, $\ddot{\alpha}$ và $\ddot{\beta}$ là các tham số thỏa mãn $0 \leq \ddot{\beta} \leq \ddot{\alpha} \leq 1$.

Khi đó, miền biên của X theo $\mathcal{F}_A^{\delta,\omega}$, ký hiệu là $IWBN_A(X)$, được định nghĩa như sau:

$$IWBN_A(X) = \overline{IW_A}(X) - \underline{IW_A}(X) \tag{3.3}$$

Miền dương và miên biên lân cận trọng số mờ trực cảm của quyết định D được xác định bởi.

$$IWPOS_A(D) = \bigcup_{X \in U/D} \underline{IW_A}(X) \tag{3.4}$$

và

$$IWBN_A(D) = \overline{IW_A}(D) - \underline{IW_A}(D) \tag{3.5}$$

trong đó, $\overline{IW_A}(D) = \bigcup_{X \in U/D} \overline{IW_A}(X)$ và $\underline{IW_A}(D) = \bigcup_{X \in U/D} \underline{IW_A}(X)$.

Cuối cùng, hàm phụ thuộc lân cận mờ trực cảm có trọng số của quyết định D theo tập con thuộc tính A được xác định như sau:

$$\ddot{\gamma}(A, D) = \frac{|IWPOS_A(D)|}{|U|} \tag{3.6}$$

Từ Công thức 3.6, có thể nói rằng IS là nhất quán nếu $\ddot{\gamma}(C, D) = 1$ và không nhất quán nếu $\ddot{\gamma}(C, D) \neq 1$.

3.2.2 Một số tính chất của IFWNRS

Trong phần này, luận án sẽ trình bày một số tính chất quan trọng của IFWNRS để làm rõ những ưu điểm trong việc ứng dụng vào các thuật toán rút gọn thuộc tính.

Mệnh đề 3.3 Cho bảng quyết định $IS = (U, C \cup D)$ và tập thuộc tính $A \subseteq C$, một số tính chất của IFWNRs được trình bày như sau:

1. Nếu $\ddot{\beta} = 0$ thì $\overline{IW_A}(D) = U$ và $IWBN_A(D) = U - IWPOS_A(D)$;
2. $IWBN_A(D) \cap IWPOS_A(D) = \emptyset$;
3. $IWBN_A(D) \cup IWPOS_A(D) = \overline{IW_A}(D)$.

Chứng minh. 1. Từ Công thức 3.2 và $\ddot{\beta} = 0$ với mọi $v \in X$, chúng ta có $v \in \overline{IW_A}(X)$. Như vậy, $X \subseteq \overline{IW_A}(X)$ với mọi $X \in U/D$. Do đó, $\bigcup_{X \in U/D} X \subseteq \bigcup_{X \in U/D} \overline{IW_A}(X) =$

$\overline{IW_A}(D)$. Bên cạnh đó, $\bigcup_{X\in U/D} X = U$ nên $U \subseteq \overline{IW_A}(D)$. Vì U chứa toàn bộ các đối tượng nên $\overline{IW_A}(D) \subseteq U$. Do đó, $\overline{IW_A}(D) = U$.

Từ Công thức 3.2, chúng ta có $IWBN_A(D) = U - \underline{IW_A}(D)$. Thêm vào đó, $\underline{IW_A}(D) = \bigcup_{X\in U/D} \underline{IW_A}(X) = IWPOS_A(D)$. Do đó, $IWBN_A(D) = U - IWPOS_A(D)$.

2. Từ Công thức 3.5, chúng ta thu được $IWBN_A(D) = \overline{IW_A}(D) - \underline{IW_A}(D) = \overline{IW_A}(D) - IWPOS_A(D)$. Do đó, $IWBN_A(D) \cap IWPOS_A(D) = \emptyset$.

3) Vì $IWBN_A(D) = \overline{IW_A}(D) - \underline{IW_A}(D) = \overline{IW_A}(D) - IWPOS_A(D)$ nên suy ra $IWBN_A(D) \cup IWPOS_A(D) = \overline{IW_A}(D)$. Mệnh đề đã được chứng minh. □

Mệnh đề 3.4 Cho bảng quyết định $IS = (U, C \cup D)$, $A \subseteq C$ và $X \in U/D$. Nếu $p = 2$, $\ddot{\alpha} = 1$ và $\ddot{\beta} = 0$, thì $\underline{IW_A}(X) = \underline{WN_A}(X)$ và $\overline{IW_A}(X) = \overline{WN_A}(X)$.

Chứng minh. Từ Công thức 3.1 và $\ddot{\alpha} = 1$, ta có $\underline{IW_A}(X) = \left\{ u \in U : \frac{\left|[\ddot{u}]_A^{\delta,\omega} \cap X\right|}{\left|[\ddot{u}]_A^{\delta,\omega}\right|} \geq 1 \right\}$. Điều này dẫn đến $\left|[\ddot{u}]_A^{\delta,\omega} \cap X\right| \geq \left|[\ddot{u}]_A^{\delta,\omega}\right|$. Do đó, chúng ta thu được $[\ddot{u}]_A^{\delta,\omega} \subseteq X$ và $\underline{IW_A}(X) = \left\{u \in U : [\ddot{u}]_A^{\delta,\omega} \subseteq X\right\} = \underline{WN_A}(X)$.

Chứng minh tương tự với $\ddot{\beta} = 0$, ta có $\overline{IW_A}(X) = \left\{ u \in U : \frac{\left|[\ddot{u}]_A^{\delta,\omega} \cap X\right|}{\left|[\ddot{u}]_A^{\delta,\omega}\right|} \geq 0 \right\} = \left\{u \in U : [\ddot{u}]_A^{\delta,\omega} \cap X \neq \emptyset\right\} = \overline{WN_A}(X)$. Mệnh đề đã được chứng minh. □

Mệnh đề 3.4 chỉ ra rằng, WNRS là một trường hợp đặc biệt của IFWNRS. Nói cách khác, IFWNRS có tính chất tổng quát hơn so với WNRs và có hiệu quả hơn trong việc xây dựng các phương pháp rút gọn thuộc tính khi xử lý trên các kịch bản dữ liệu khác nhau.

Mệnh đề 3.5 Cho bảng quyết định $IS = (U, C \cup D)$ và hai tập thuộc tính $A, B \subseteq C$. Nếu $A \subseteq B$, thì

1. $\forall u \in U$, $[\ddot{u}]_B^{\delta,\omega} \subseteq [\ddot{u}]_A^{\delta,\omega}$;
2. Nếu $\ddot{\beta} = 0$ thì $\forall X \subseteq U$, $\overline{IW_B}(X) \subseteq \overline{IW_A}(X)$;
3. Nếu $\ddot{\alpha} = 1$ thì $\forall X \subseteq U$, $\underline{IW_A}(X) \subseteq \underline{IW_B}(X)$;
4. $IWPOS_A(D) \subseteq IWPOS_B(D)$ và $\ddot{\gamma}(A, D) \leq \ddot{\gamma}(B, D)$

Chứng minh. 1. Từ $A \subseteq B$ và Công thức 1.7 với $p = \infty$, chúng ta thu được $\max_{a\in A}\{\omega(a) \cdot |a(u) - a(v)|\} \leq \max_{a\in B}\{\omega(a) \cdot |a(u) - a(v)|\}$. Do đó, $[u]_B^{\delta,\omega} \subseteq [u]_A^{\delta,\omega} \Rightarrow [\ddot{u}]_B^{\delta,\omega} \subseteq [\ddot{u}]_A^{\delta,\omega}$, với mọi $u \in U$.

2. Từ $\ddot{\beta} = 0$, với $u \in \overline{IW_B}(X)$ bất kỳ thì $\frac{\left|[\ddot{u}]_B^{\delta,\omega} \cap X\right|}{\left|[\ddot{u}]_B^{\delta,\omega}\right|} \geq 0$, nghĩa là $\left|[\ddot{u}]_B^{\delta,\omega} \cap X\right| \geq 0$. Mặt khác, theo tính chất 1, vì $A \subseteq B$ nên $[\ddot{u}]_B^{\delta,\omega} \subseteq [\ddot{u}]_A^{\delta,\omega}$. Do đó, $\left|[\ddot{u}]_A^{\delta,\omega} \cap X\right| \geq 0$ và $u \in \overline{IW_A}(X)$. Như vậy, $\overline{IW_B}(X) \subseteq \overline{IW_A}(X)$.

3. Từ $\ddot{\alpha} = 1$, với $u \in \underline{IW_A}(X)$ bất kỳ thì $\frac{\left|[\ddot{u}]_A^{\delta,\omega} \cap X\right|}{\left|[\ddot{u}]_A^{\delta,\omega}\right|} \geq 1$, nghĩa là $[\ddot{u}]_A^{\delta,\omega} \subseteq X$. Mặt khác, theo tính chất 1, vì $A \subseteq B$ nên $[\ddot{u}]_B^{\delta,\omega} \subseteq [\ddot{u}]_A^{\delta,\omega}$. Do đó, $[\ddot{u}]_B^{\delta,\omega} \subseteq X$ và $\frac{\left|[\ddot{u}]_B^{\delta,\omega} \cap X\right|}{\left|[\ddot{u}]_B^{\delta,\omega}\right|} = 1$. Như vậy, $u \in \underline{IW_B}(X)$ và $\underline{IW_A}(X) \subseteq \underline{IW_B}(X)$.

4. Dựa trên kết quả thu được từ tính chất 3, chúng ta luôn có: $IWPOS_A(D) = \bigcup_{x \in U/D} \underline{IW_A}(X) \subseteq \bigcup_{x \in U/D} \underline{IW_B}(X) = IWPOS_B(D)$. Do đó, $\ddot{\gamma}(A, D) \leq \ddot{\gamma}(B, D)$. □

3.3 Đề xuất thuật toán rút gọn thuộc tính dựa trên IFWNRS

Phần này định nghĩa một số kiểu rút gọn theo tiếp cận IFWNRS, đồng thời phân tích hiệu quả của chúng để làm cơ sở thiết kế các thuật toán rút gọn thuộc tính trên bảng quyết định cố định.

3.3.1 Đề xuất thuật toán rút gọn thuộc tính trên bảng quyết định cố định

Định nghĩa 3.1 Cho bảng quyết định $IS = (U, C \cup D)$, một tập con thuộc tính $A \subseteq C$ được gọi là một $\ddot{\gamma}$-rút gọn của C nếu và chỉ nếu A thỏa mãn

1. $\ddot{\gamma}(A, D) = \ddot{\gamma}(C, D)$;
2. $\forall a \in A$, $\ddot{\gamma}(A \backslash \{a\}, D) < \ddot{\gamma}(A, D)$.

Dễ thấy rằng một $\ddot{\gamma}$-rút gọn trong Định nghĩa 3.1 là tập con thuộc tính tối thiểu bảo toàn yếu tố nhất quán của tất cả các thuộc tính trong bảng quyết định. Nói cách khác, rút gọn này chỉ chứa các đối tượng trong miền dương của bảng quyết định và bỏ qua các đối tượng khác. Tuy nhiên, trong các trường hợp bảng quyết định không nhất quán, số lượng các đối tượng nằm ngoài miền dương sẽ nhiều hơn. Các đối tượng được xem là không thể phân loại một cách chắc chắn và việc bỏ qua chúng trong quá trình tính toán sẽ có ảnh hưởng không nhỏ tới chất lượng của rút gọn thu dược. Do đó, luận án sẽ định nghĩa một rút gọn khác để xử lý tất cả các đối tượng trong vũ trụ. Trước hết, độ đo khoảng cách giữa hai hạt thông tin lân cận trọng số mờ trực cảm sẽ được trình bày dựa trên định nghĩa dưới đây.

Định nghĩa 3.2 Cho bảng quyết định $IS = (U, C \cup D)$, hai hạt thông tin lân cận trọng số mờ trực cảm $[\ddot{u}]_A^{\delta,\omega}$ và $[\ddot{u}]_B^{\delta,\omega}$ được tạo bởi các tập con thuộc tính $A, B \subseteq C$. Khi đó, khoảng cách giữa $[\ddot{u}]_A^{\delta,\omega}$ và $[\ddot{u}]_B^{\delta,\omega}$ được ký hiệu bởi $\ddot{\mathcal{D}}\left([\ddot{u}]_A^{\delta,\omega}, [\ddot{u}]_B^{\delta,\omega}\right)$ và được xác định như sau:

$$\ddot{\mathcal{D}}\left([\ddot{u}]_A^{\delta,\omega}, [\ddot{u}]_B^{\delta,\omega}\right) = \frac{\left|[\ddot{u}]_A^{\delta,\omega} \cup [\ddot{u}]_B^{\delta,\omega}\right| - \left|[\ddot{u}]_A^{\delta,\omega} \cap [\ddot{u}]_B^{\delta,\omega}\right|}{|U|} \tag{3.7}$$

Từ độ đo khoảng cách được xây dựng trên hai hạt thông tin lân cận trọng số mờ trực cảm, luận án tiếp tục mở rộng độ đo khoảng cách trên hai họ lân cận trọng số mờ trực cảm như sau:

$$\ddot{\mathcal{D}}\left(\mathcal{F}_A^{\delta,\omega}, \mathcal{F}_B^{\delta,\omega}\right) = \sum_{u \in U} \frac{\left(\left|[\ddot{u}]_A^{\delta,\omega} \cup [\ddot{u}]_B^{\delta,\omega}\right| - \left|[\ddot{u}]_A^{\delta,\omega} \cap [\ddot{u}]_B^{\delta,\omega}\right|\right)}{|U|^2} \tag{3.8}$$

Rõ ràng, cấu trúc của một hạt thông tin lân cận mờ trực cảm có trọng số tương tự như cấu trúc của hạt thông tin mờ trực cảm ở mức $alpha, beta$. Vì vậy, độ đo khoảng cách giữa hai họ lân cận mờ trực cảm có trọng số cũng thừa hưởng các đặc tính của khoảng cách phân hoạch mờ trực cảm mức $alpha, beta$. Trên cơ sở đó, một số tính chất của độ đo sẽ được trình bày, làm nền tảng cho việc xây dựng thuật toán rút gọn thuộc tính trên bảng quyết định cố định.

Mệnh đề 3.6 Cho bảng quyết định $IS = (U, C \cup D)$, khi đó khoảng cách giữa hai họ lân cận mờ trực cảm trọng số $\mathcal{F}_C^{\delta,\omega}$ và $\mathcal{F}_{C\cup D}^{\delta,\omega}$ được xác định như sau:

$$\ddot{\mathcal{D}}\left(\mathcal{F}_C^{\delta,\omega}, \mathcal{F}_{C\cup D}^{\delta,\omega}\right) = \sum_{u \in U} \frac{\left(\left|[\ddot{u}]_C^{\delta,\omega}\right| - \left|[\ddot{u}]_C^{\delta,\omega} \cap [\ddot{u}]_D^{\delta,\omega}\right|\right)}{|U|^2} \tag{3.9}$$

Mệnh đề 3.6 biểu diễn một độ đo nhằm đo lường lượng thông tin trung bình của một đối tượng u không nằm trong $[\ddot{u}]_D^{\delta,\omega}$ trên tập thuộc tính C.

Mệnh đề 3.7 Cho bảng quyết định $IS = (U, C \cup D)$ và một tập con thuộc tính điều kiện $A \subseteq C$, nếu $\delta_1 \leq \delta_2$ thì $\ddot{\mathcal{D}}\left(\mathcal{F}_A^{\delta_1,\omega}, \mathcal{F}_{A\cup D}^{\delta_1,\omega}\right) \leq \ddot{\mathcal{D}}\left(\mathcal{F}_A^{\delta_2,\omega}, \mathcal{F}_{A\cup D}^{\delta_2,\omega}\right)$.

Chứng minh. Với mọi $u, v \in U$, chúng ta có:

$$\left|[\ddot{u}]_A^{\delta_1,\omega}\right| - \left|[\ddot{u}]_A^{\delta_1,\omega} \cap [u]_D\right| = \sum_{v \in [u]_D} \frac{1 + \gamma_{[\ddot{u}]_A^{\delta_1,\omega}}(v) - \eta_{[\ddot{u}]_A^{\delta_1,\omega}}(v)}{2}$$

$$\left|[\ddot{u}]_A^{\delta_2,\omega}\right| - \left|[\ddot{u}]_A^{\delta_2,\omega} \cap [u]_D\right| = \sum_{v \in [u]_D} \frac{1 + \gamma_{[\ddot{u}]_A^{\delta_2,\omega}}(v) - \eta_{[\ddot{u}]_A^{\delta_2,\omega}}(v)}{2}.$$

Từ Mệnh đề 3.2 và $\delta_1 \leq \delta_2$, chúng ta có $[\ddot{u}]_A^{\delta_1,\omega} \subseteq [\ddot{u}]_A^{\delta_2,\omega}$ với mọi $u \in U$. Điều này suy ra $\gamma_{[\ddot{u}]_A^{\delta_1,\omega}}(v) - \eta_{[\ddot{u}]_A^{\delta_1,\omega}}(v) \geq \gamma_{[\ddot{u}]_A^{\delta_2,\omega}}(v) - \eta_{[\ddot{u}]_A^{\delta_2,\omega}}(v)$, nghĩa là $\left|[\ddot{u}]_A^{\delta_1,\omega}\right| - \left|[\ddot{u}]_A^{\delta_1,\omega} \cap [\ddot{u}]_D^{\delta_1,\omega}\right| \leq$

$\left|[\ddot{u}]_A^{\delta_2,\omega}\right| - \left|[\ddot{u}]_A^{\delta_2,\omega} \cap [\ddot{u}]_D^{\delta_2,\omega}\right|$. Do đó, chúng ta thu được $\ddot{\mathcal{D}}\left(\mathcal{F}_A^{\delta_1,\omega}, \mathcal{F}_{A\cup D}^{\delta_1,\omega}\right) \leq \ddot{\mathcal{D}}\left(\mathcal{F}_A^{\delta_2,\omega}, \mathcal{F}_{A\cup D}^{\delta_2,\omega}\right)$. Mệnh đề đã được chứng minh. □

Mệnh đề 3.8 Cho bảng quyết định $IS = (U, C \cup D)$ với hai tập con thuộc tính $A, B \subseteq C$. Nếu $A \subseteq B$ thì $\ddot{\mathcal{D}}\left(\mathcal{F}_A^{\delta,\omega}, \mathcal{F}_{A\cup D}^{\delta,\omega}\right) \geq \ddot{\mathcal{D}}\left(\mathcal{F}_B^{\delta,\omega}, \mathcal{F}_{B\cup D}^{\delta,\omega}\right)$.

Chứng minh. Xét trường hợp $v \in [u]_D$, khi đó: $\gamma_{[\ddot{u}]_A^{\delta,\omega}}(v) - \min\left\{\gamma_{[\ddot{u}]_A^{\delta,\omega}}(v), \gamma_{[\ddot{u}]_D^{\delta,\omega}}(v)\right\} = \gamma_{[\ddot{u}]_B^{\delta,\omega}}(v) - \min\left\{\gamma_{[\ddot{u}]_B^{\delta,\omega}}(v), \gamma_{[\ddot{u}]_D^{\delta,\omega}}(v)\right\} = 0$ và $\eta_{[\ddot{u}]_A^{\delta,\omega}}(v) - \max\left\{\eta_{[\ddot{u}]_A^{\delta,\omega}}(v), \eta_{[\ddot{u}]_D^{\delta,\omega}}(v)\right\} = \eta_{[\ddot{u}]_B^{\delta,\omega}}(v) - \max\left\{\eta_{[\ddot{u}]_B^{\delta,\omega}}(v), \eta_{[\ddot{u}]_D^{\delta,\omega}}(v)\right\} = 0$.

Trong trường hợp $v \notin [u]_D$, khi đó: $\gamma_{[\ddot{u}]_A^{\delta,\omega}}(v) - \min\left\{\gamma_{[\ddot{u}]_A^{\delta,\omega}}(v), \gamma_{[\ddot{u}]_D^{\delta,\omega}}(v)\right\} = \gamma_{[\ddot{u}]_A^{\delta,\omega}}(v)$, $\gamma_{[\ddot{u}]_B^{\delta,\omega}}(v) - \min\left\{\gamma_{[\ddot{u}]_B^{\delta,\omega}}(v), \gamma_{[\ddot{u}]_D^{\delta,\omega}}(v)\right\} = \gamma_{[\ddot{u}]_B^{\delta,\omega}}(v)$, $\eta_{[\ddot{u}]_A^{\delta,\omega}}(v) - \max\left\{\eta_{[\ddot{u}]_A^{\delta,\omega}}(v), \eta_{[\ddot{u}]_D^{\delta,\omega}}(v)\right\} = \eta_{[\ddot{u}]_A^{\delta,\omega}}(v) - 1$ và $\eta_{[\ddot{u}]_B^{\delta,\omega}}(v) - \max\left\{\eta_{[\ddot{u}]_B^{\delta,\omega}}(v), \eta_{[\ddot{u}]_D^{\delta,\omega}}(v)\right\} = \eta_{[\ddot{u}]_B^{\delta,\omega}}(v) - 1$.

Vì $A \subseteq B$, nên $\gamma_{[\ddot{u}]_A^{\delta,\omega}}(v) \geq \gamma_{[\ddot{u}]_B^{\delta,\omega}}(v)$ và $\eta_{[\ddot{u}]_A^{\delta,\omega}}(v) \leq \eta_{[\ddot{u}]_B^{\delta,\omega}}(v)$, với mọi $u, v \in U$. Điều này suy ra $\left|[\ddot{u}]_A^{\delta,\omega}\right| - \left|[\ddot{u}]_A^{\delta,\omega} \cap [\ddot{u}]_D^{\delta,\omega}\right| \geq \left|[\ddot{u}]_B^{\delta,\omega}\right| - \left|[\ddot{u}]_B^{\delta,\omega} \cap [\ddot{u}]_D^{\delta,\omega}\right|$.

Do đó, $\ddot{\mathcal{D}}\left(\mathcal{F}_A^{\delta,\omega}, \mathcal{F}_{A\cup D}^{\delta,\omega}\right) \geq \ddot{\mathcal{D}}\left(\mathcal{F}_B^{\delta,\omega}, \mathcal{F}_{B\cup D}^{\delta,\omega}\right)$. Mệnh đề đã được chứng minh. □

Định nghĩa 3.3 Cho bảng quyết định $IS = (U, C \cup D)$, một tập con thuộc tính $A \subseteq C$ được gọi là một $\ddot{\mathcal{D}}$-rút gọn của C nếu thỏa mãn

1. $\ddot{\mathcal{D}}\left(\mathcal{F}_A^{\delta,\omega}, \mathcal{F}_{A\cup D}^{\delta,\omega}\right) = \ddot{\mathcal{D}}\left(\mathcal{F}_C^{\delta,\omega}, \mathcal{F}_{C\cup D}^{\delta,\omega}\right)$;
2. $\ddot{\mathcal{D}}\left(\mathcal{F}_A^{\delta,\omega}, \mathcal{F}_{A\cup D}^{\delta,\omega}\right) > \ddot{\mathcal{D}}\left(\mathcal{F}_{A\setminus\{a\}}^{\delta,\omega}, \mathcal{F}_{A\setminus\{a\}\cup D}^{\delta,\omega}\right)$.

Rõ ràng, $\ddot{\mathcal{D}}$-rút gọn xét tới tất cả các đối tượng trọng bảng quyết định và có tính tổng quát tốt hơn so với $\ddot{\gamma}$-rút gọn.

Định nghĩa 3.4 Cho bảng quyết định $IS = (U, C \cup D)$, một tập con thuộc tính $A \subseteq C$ và một thuộc tính $a \subseteq C \setminus A$, độ quan trọng của a theo A dựa trên khoảng cách họ lân cận mờ trực cảm có trọng số, ký hiệu là $Sig_F(a, A)$ được xác định như sau

$$Sig_F(a, A) = \ddot{\mathcal{D}}\left(\mathcal{F}_{A\setminus\{a\}}^{\delta,\omega}, \mathcal{F}_{A\setminus\{a\}\cup D}^{\delta,\omega}\right) - \ddot{\mathcal{D}}\left(\mathcal{F}_A^{\delta,\omega}, \mathcal{F}_{A\cup D}^{\delta,\omega}\right) \tag{3.10}$$

Dựa trên các khái niệm được trình bày, luận án tiếp tục xây dựng Thuật toán 3.1 quả nhằm tìm kiếm một rút gọn trên bảng quyết định.

Rất dễ thấy rằng độ phức tạp của thuật toán ARIFW khi tính toán các trọng số của từng thuộc tính điều kiện tại Bước 1 là $O(|C||U|)$. Tại Bước 2, thuật toán cần tính toán các giá trị khoảng cách của toàn bộ đối tượng trong U theo tập thuộc tính C để xác định họ lân cận mờ trực cảm có trọng số của C. Do đó, độ phức tạp tại bước này là $O\left(|C||U|^2\right)$. Đây cũng là độ phức tạp được thực hiện trong vòng lặp *for* từ Bước 3

Thuật toán 3.1 Rút gọn thuộc tính dựa trên IFWNRS (ARIFW)

Đầu vào: Bảng quyết định $IS = (U, C \cup D)$ và bán kính lân cận δ.

Đầu ra: Một rút gọn $\mathcal{B}$

1: tính trọng số của mỗi thuộc tính theo Công thức 1.6
2: tính toán $\ddot{\mathcal{D}}\left(\mathcal{F}_{C}^{\delta,\omega}, \mathcal{F}_{C\cup D}^{\delta,\omega}\right)$
3: **for** $a \in C$ **do**
4: tính các họ lân cận mờ trực cảm có trọng số $\mathcal{F}_{\{a\}}^{\delta,\omega}$
5: tính toán $\ddot{\mathcal{D}}\left(\mathcal{F}_{\{a\}}^{\delta,\omega}, \mathcal{F}_{\{a\}\cup D}^{\delta,\omega}\right)$
6: **end for**
7: $\mathcal{B} = \{a_0\}$ thỏa mãn $\ddot{\mathcal{D}}\left(\mathcal{F}_{\{a_0\}}^{\delta,\omega}, \mathcal{F}_{\{a_0\}\cup D}^{\delta,\omega}\right) = \min\limits_{a\in C} \ddot{\mathcal{D}}\left(\mathcal{F}_{\{a\}}^{\delta,\omega}, \mathcal{F}_{\{a\}\cup D}^{\delta,\omega}\right)$
8: **while** $\ddot{\mathcal{D}}\left(\mathcal{F}_{\mathcal{B}}^{\delta,\omega}, \mathcal{F}_{\mathcal{B}\cup D}^{\delta,\omega}\right) > \ddot{\mathcal{D}}\left(\mathcal{F}_{C}^{\delta,\omega}, \mathcal{F}_{C\cup D}^{\delta,\omega}\right)$ **do**
9: tính toán $Sig_F(a, \mathcal{B})$, với mọi $a \in C \backslash \mathcal{B}$
10: lựa chọn a_0 thỏa mãn $Sig_F(a_0, \mathcal{B}) = \max\limits_{a\in C\backslash\mathcal{B}} \{Sig_F(a, \mathcal{B})\}$
11: $\mathcal{B} \leftarrow \mathcal{B} \cup \{a_0\}$
12: **end while**
13: **return** $\mathcal{B}$

đến 6. Tại bước 7, thuật toán lựa chọn thuộc tính có giá trị khoảng cách họ lân cận mờ trực cảm có trọng số nhỏ nhất để bổ sung vào tập rút gọn. Độ phức tạp trong Bước này là $O(|C|)$.

Trong vòng lặp *while*, với mỗi lần thuộc tính được bổ sung không thỏa mãn tính chất của một rút gọn, các thuộc tính còn lại sẽ được xét tiếp để lựa chọn thuộc tính độ quan trọng lớn nhất. Khi đó, độ phức tạp từ Bước 8 đến 12 là $O\left(|C|^2|U|^2\right)$. Do đó, độ phức tạp của toàn bộ thuật toán là $O\left(|C|^2|U|^2\right)$.

Ví dụ 3.2 Để tìm rút gọn của Bảng 1.1 dựa trên thuật toán ARIFW, đầu tiên chúng ta khởi tạo tập rút gọn $\mathcal{B} = \emptyset$, tính $\ddot{\mathcal{D}}\left(\mathcal{F}_{C}^{0.5,\omega}, \mathcal{F}_{C\cup D}^{0.5,\omega}\right)$ và $\ddot{\mathcal{D}}\left(\mathcal{F}_{\{a\}}^{0.5,\omega}, \mathcal{F}_{\{a\}\cup D}^{0.5,\omega}\right)$, với mọi $a \in C$ từ Công thức 3.9

$$\ddot{\mathcal{D}}\left(\mathcal{F}_{C}^{0.5,\omega}, \mathcal{F}_{C\cup D}^{0.5,\omega}\right) = \frac{1}{25} \times (0.82 + 0.82 + 0.00 + 0.00 + 0.00) = 0.07$$

Thực hiện tương tự, chúng ta thu được:
$\ddot{\mathcal{D}}\left(\mathcal{F}_{\{a_1\}}^{0.5,\omega}, \mathcal{F}_{\{a_1\}\cup D}^{0.5,\omega}\right) = 0.46$, $\ddot{\mathcal{D}}\left(\mathcal{F}_{\{a_2\}}^{0.5,\omega}, \mathcal{F}_{\{a_2\}\cup D}^{0.5,\omega}\right) = 0.43$, $\ddot{\mathcal{D}}\left(\mathcal{F}_{\{a_3\}}^{0.5,\omega}, \mathcal{F}_{\{a_3\}\cup D}^{0.5,\omega}\right) = 0.19$,
$\ddot{\mathcal{D}}\left(\mathcal{F}_{\{a_4\}}^{0.5,\omega}, \mathcal{F}_{\{a_4\}\cup D}^{0.5,\omega}\right) = 0.33$, $\ddot{\mathcal{D}}\left(\mathcal{F}_{\{a_5\}}^{0.5,\omega}, \mathcal{F}_{\{a_5\}\cup D}^{0.5,\omega}\right) = 0.27$. Có thể nhận thấy rằng $\ddot{\mathcal{D}}\left(\mathcal{F}_{\{a_3\}}^{0.5,\omega}, \mathcal{F}_{\{a_3\}\cup D}^{0.5,\omega}\right)$ nhỏ nhất nên bổ sung thuộc tính a_3 vào tập rút gọn $\mathcal{B}$, khi đó

$\mathcal{B} = \{a_3\}$. Do $\ddot{\mathcal{D}}\left(\mathcal{F}_{\mathcal{B}}^{\delta,\omega}, \mathcal{F}_{\mathcal{B}\cup D}^{\delta,\omega}\right) > \ddot{\mathcal{D}}\left(\mathcal{F}_{C}^{\delta,\omega}, \mathcal{F}_{C\cup D}^{\delta,\omega}\right)$, thuật toán tiếp tục lọc tìm kiếm các thuộc tính còn lại.

$\ddot{\mathcal{D}}\left(\mathcal{F}_{\mathcal{B}\cup\{a_1\}}^{\delta,\omega}, \mathcal{F}_{\mathcal{B}\cup\{a_1\}\cup D}^{\delta,\omega}\right) = 0.19 \Rightarrow Sig_F\left(a_1, \mathcal{B}\right) = 0.19 - 0.19 = 0.00,$

$\ddot{\mathcal{D}}\left(\mathcal{F}_{\mathcal{B}\cup\{a_2\}}^{\delta,\omega}, \mathcal{F}_{\mathcal{B}\cup\{a_2\}\cup D}^{\delta,\omega}\right) = 0.19 \Rightarrow Sig_F\left(a_2, \mathcal{B}\right) = 0.19 - 0.19 = 0.00,$

$\ddot{\mathcal{D}}\left(\mathcal{F}_{\mathcal{B}\cup\{a_4\}}^{\delta,\omega}, \mathcal{F}_{\mathcal{B}\cup\{a_4\}\cup D}^{\delta,\omega}\right) = 0.14 \Rightarrow Sig_F\left(a_4, \mathcal{B}\right) = 0.19 - 0.14 = 0.05,$

$\ddot{\mathcal{D}}\left(\mathcal{F}_{\mathcal{B}\cup\{a_5\}}^{\delta,\omega}, \mathcal{F}_{\mathcal{B}\cup\{a_5\}\cup D}^{\delta,\omega}\right) = 0.12 \Rightarrow Sig_F\left(a_5, \mathcal{B}\right) = 0.19 - 0.12 = 0.07.$

Vì $Sig_F\left(a_5, \mathcal{B}\right)$ lớn nhất nên bổ sung thuộc tính a_5 vào rút gọn, khi đó $\mathcal{B} = \mathcal{B}\cup\{a_5\} = \{a_3, a_5\}$. Do $\ddot{\mathcal{D}}\left(\mathcal{F}_{\mathcal{B}}^{\delta,\omega}, \mathcal{F}_{\mathcal{B}\cup D}^{\delta,\omega}\right) > \ddot{\mathcal{D}}\left(\mathcal{F}_{C}^{\delta,\omega}, \mathcal{F}_{C\cup D}^{\delta,\omega}\right)$, thuật toán tiếp tục lọc tìm kiếm các thuộc tính còn lại.

$\ddot{\mathcal{D}}\left(\mathcal{F}_{\mathcal{B}\cup\{a_1\}}^{\delta,\omega}, \mathcal{F}_{\mathcal{B}\cup\{a_1\}\cup D}^{\delta,\omega}\right) = 0.12 \Rightarrow Sig_F\left(a_1, \mathcal{B}\right) = 0.12 - 0.12 = 0.00,$

$\ddot{\mathcal{D}}\left(\mathcal{F}_{\mathcal{B}\cup\{a_2\}}^{\delta,\omega}, \mathcal{F}_{\mathcal{B}\cup\{a_2\}\cup D}^{\delta,\omega}\right) = 0.12 \Rightarrow Sig_F\left(a_2, \mathcal{B}\right) = 0.12 - 0.12 = 0.00,$

$\ddot{\mathcal{D}}\left(\mathcal{F}_{\mathcal{B}\cup\{a_4\}}^{\delta,\omega}, \mathcal{F}_{\mathcal{B}\cup\{a_4\}\cup D}^{\delta,\omega}\right) = 0.07 \Rightarrow Sig_F\left(a_4, \mathcal{B}\right) = 0.12 - 0.07 = 0.05$

Vì $Sig_F\left(a_4, \mathcal{B}\right)$ lớn nhất nên bổ sung thuộc tính a_4 vào rút gọn, khi đó $\mathcal{B} = \mathcal{B}\cup\{a_4\} = \{a_3, a_5, a_4\}$. Lúc này $\ddot{\mathcal{D}}\left(\mathcal{F}_{\mathcal{B}}^{\delta,\omega}, \mathcal{F}_{\mathcal{B}\cup D}^{\delta,\omega}\right) = \ddot{\mathcal{D}}\left(\mathcal{F}_{C}^{\delta,\omega}, \mathcal{F}_{C\cup D}^{\delta,\omega}\right)$ nên thuật toán kết thúc. Rút gọn cuối cùng của Bảng 1.1 được xác định bởi Thuật toán ARIFW là $\mathcal{B} = \{a_3, a_5, a_4\}$.

3.3.2 Đề xuất thuật toán rút gọn thuộc tính trên bảng quyết định thay đổi khi bổ sung tập đối tượng

Đầu tiên, có thể thấy rằng, khi một tập đối tượng mới được bổ sung vào bảng quyết định, các hạt thông tin lân cận trọng số sẽ có sự thay đổi. Do đó, việc đưa ra một chiến lược cập nhập giá trị của các hạt thông tin lân cận trọng số mờ trực cảm là một vấn đề quan trọng. Để giải quyết vấn đề này, luận án ban đầu sẽ trình bày về một quá trình cập nhật trọng số khi bảng quyết định có sự bổ sung một tập đối tượng mới.

Cho bảng quyết định $IS = (U, C \cup D)$ với $U = (u_1, u_2, \ldots, u_n)$. Giả sử rằng một tập đối tượng mới $\Delta U = (u_{n+1}, u_{n+2}, \ldots, u_{n+z})$ được bổ sung vào U, khi đó, quá trình cập nhật trọng số của một thuộc tính $a \in C$ với các đối tượng $u_j \in U^+ = U \cup \Delta U$ trên một hạt thông tin lân cận trọng số bất kỳ $[u_i]_{\{a\}}^{\delta,\omega}$ được trình bày như sau

$$\omega^+ = \begin{cases} \omega & \text{nếu } 1 \leq i, j \leq n \\ \omega^{\Delta} & \text{nếu ngược lại} \end{cases} \tag{3.11}$$

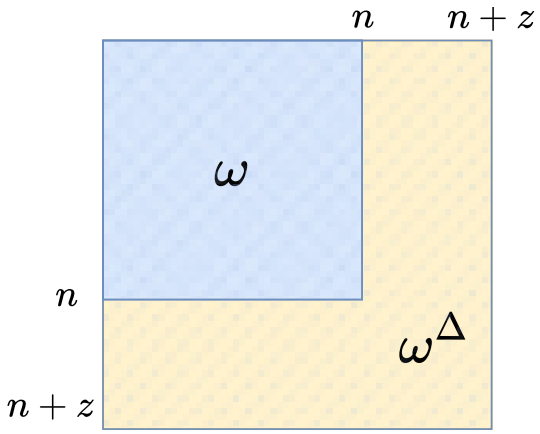
trong đó, ω là trọng số gốc được tạo bởi tập đối tượng ban đầu và ω^{Δ} là trọng số mới được áp dụng trên tập đối tượng ΔU. Để xác định ω^{Δ}, chúng ta có thể dựa trên công thức cập nhật ma trận nghịch đảo được trình bày trong [104] nhằm tính toán vector hệ số phân hoạch của các thuộc tính trên bảng quyết định mới. Tuy nhiên, chúng ta cũng

có thể sử dụng hàm "np.linalg.solve" trong thư viện numpy để tính toán giá trị này một cách nhanh chóng trên tập đối tượng ΔU.

Dựa trên quá trình cập nhật trọng số, tiếp theo để xác định các hạt thông tin lân cận mờ trực cảm có trọng số, luận án đưa ra một cơ chế cập nhật hạt dựa trên công thức sau đây.

$$[\ddot{u}_i]_{\{a\}}^{\delta,\omega^+}(u_j) = \begin{cases} [\ddot{u}_i]_a^{\delta,\omega}(u_j) & \text{if } 1 \le i,j \le n \\ [\ddot{u}_i]_a^{\delta,\omega^\Delta}(u_j) & \text{if } u_j \in [u_i]_{\{a}^{\delta,\omega^\Delta}, (n+1 \le i \le n+z) \vee (n+1 \le j \le n+z) \\ (0,1) & \text{trường hợp còn lại} \end{cases} \tag{3.12}$$

Rõ ràng, khi các hạt thông tin lân cận trọng số mờ trực cảm được hình thành, chúng ta cũng thu được một họ lân cận trọng số mờ trực cảm mới trên thuộc tính a, ký hiệu là $\mathcal{F}_{\{a\}}^{\delta,\omega^+}$. Hình 3.3 biểu diễn cơ chế cập nhật các hạt thông tin lân cận trọng số mờ trực cảm của $\mathcal{F}_{\{a\}}^{\delta,\omega^+}$ dựa trên trên Công thức 3.12.



Hình 3.3: Quá trình cập nhật trọng số trên họ lân cận mờ trực cảm có trọng số

Mệnh đề 3.9 Cho bảng quyết định $IS = (U, C \cup D)$ với $U = (u_1, u_2, \ldots, u_n)$ và tập đối tượng mới $\Delta U = (u_{n+1}, u_{n+2}, \ldots, u_{n+z})$, trong đó $z \ge 1$. Khi đó, khoảng cách giữa hai họ lân cận mờ trực cảm có trọng số $\mathcal{F}_C^{\delta,\omega^+}$ và $\mathcal{F}_{C\cup D}^{\delta,\omega^+}$ trên tập đối tượng $U^+ = U \cup \Delta U$, ký hiệu là $\ddot{\mathcal{D}}\left(\mathcal{F}_C^{\delta,\omega^+}, \mathcal{F}_{C\cup D}^{\delta,\omega^+}\right)$ được tính như sau:

$$\ddot{\mathcal{D}}\left(\mathcal{F}_C^{\delta,\omega^+}, \mathcal{F}_{C\cup D}^{\delta,\omega^+}\right) = \frac{n^2 \ddot{\mathcal{D}}\left(\mathcal{F}_C^{\delta,\omega}, \mathcal{F}_{C\cup D}^{\delta,\omega}\right)}{(n+z)^2} + \frac{2\sum_{i=1}^{z}\left(\left|[\ddot{u}_{n+i}]_C^{\delta,\omega^+}\right| - \left|[\ddot{u}_{n+i}]_C^{\delta,\omega^+} \cap [\ddot{u}_{n+i}]_D^{\delta,\omega^+}\right| - \theta_i\right)}{(n+z)^2} \tag{3.13}$$

trong đó, $\theta_1 = 0$ và với mọi $i \geq 2$,

$$\theta_i = \sum_{j=1}^{z-1} \left(\gamma_{[\ddot{u}_{n+i}]_C^{\delta,\omega^+}}(u_{n+j+1}) - \gamma_{[\ddot{u}_{n+i}]_{C\cup D}^{\delta,\omega^+}}(u_{n+j+1}) - \eta_{[\ddot{u}_{n+i}]_C^{\delta,\omega^+}}(u_{n+j+1}) + \eta_{[\ddot{u}_{n+i}]_{C\cup D}^{\delta,\omega^+}}(u_{n+j+1}) \right).$$

Mệnh đề 3.10 Cho bảng quyết định $IS = (U, C \cup D)$ với $U = (u_1, u_2, \ldots, u_n)$, $A \subseteq C$ là một rút gọn dựa trên khoảng cách giữa hai họ lân cận mờ trực cảm có trọng số được xác định trên U và $\Delta U = (u_{n+1}, u_{n+2}, \ldots, u_{n+z})$, trong đó $z \geq 1$ là tập đối tượng bổ sung vào U, khi đó chúng ta có:

1. Nếu tất cả các đối tượng trong ΔU có giá trị thuộc tính quyết định giống nhau thì:

$$\ddot{\mathcal{D}}\left(\mathcal{F}_C^{\delta,\omega^+}, \mathcal{F}_{C\cup D}^{\delta,\omega^+}\right) = \frac{n^2 \ddot{\mathcal{D}}\left(\mathcal{F}_C^{\delta,\omega}, \mathcal{F}_{C\cup D}^{\delta,\omega}\right) + 2\sum_{i=1}^{z}\left(\left|[\ddot{u}_{n+i}]_C^{\delta,\omega^+}\right| - \left|[\ddot{u}_{n+i}]_C^{\delta,\omega^+} \cap [\ddot{u}_{n+i}]_D^{\delta,\omega^+}\right|\right)}{(n+z)^2} \tag{3.14}$$

2. Nếu $[\ddot{u}_{n+i}]_B^{\delta,\omega^+} \subseteq [\ddot{u}_{n+i}]_D^{\delta,\omega^+}$ với $i = 1, 2, \ldots, z$ thì:

$$\ddot{\mathcal{D}}\left(\mathcal{F}_A^{\delta,\omega^+}, \mathcal{F}_{A\cup D}^{\delta,\omega^+}\right) = \ddot{\mathcal{D}}\left(\mathcal{F}_C^{\delta,\omega^+}, \mathcal{F}_{C\cup D}^{\delta,\omega^+}\right)$$

Mệnh đề 3.10 chỉ ra hai trường hợp đặc biệt của tập đối tượng được bổ sung vào bảng quyết định. Rõ ràng, trong hai trường hợp này, Công thức 3.13 được rút ngắn và xử lý nhanh hơn. Từ đây, Thuật toán gia tăng 3.2 được đề xuất để tìm kiếm một rút gọn xấp xỉ trên bảng quyết định có sự bổ sung tập đối tượng sẽ được đề xuất. Thuật toán bao gồm bốn thành phần chính được trình bày trong bảng mã giả sau:

Tiếp theo, để chứng minh hiệu quả trong việc giảm thiểu thời gian thực thi của thuật toán, giả sử $\left|\mathcal{B}^+\right|$ là số lượng thuộc tính của tập rút gọn thu được từ thuật toán, $|\Delta U|$ biểu diễn số lượng các đối tượng được bổ sung vào bảng quyết định và $\left|U^+\right|$ biểu diễn tổng số đối tượng trên bảng quyết định mới. Từ dòng 1 đến dòng 2, độ phức tạp của thuật toán khi xác định các họ lân cận mờ trực cảm có trọng số mới trên mỗi thuộc tính điều kiện là $O\left(\left|\mathcal{B}^+\right|\left|U^+\right||\Delta U|\right)$. Tương tự như Thuật toán 2.2, trong trường hợp đơn giản, toàn bộ quá trình xử lý của thuật toán sẽ kết thúc ở dòng 7, dẫn tới độ phức tạp của thuật toán trong trường hợp này là $O\left(\left|\mathcal{B}^+\right|\left|U^+\right||\Delta U|\right)$.

Đối với trường hợp còn lại, độ phức tạp trên công thức gia tăng tại dòng 9 là $O\left(\left|U^+\right||\Delta U|\right)$. Độ phức tạp của vòng lặp *while* trong thuật toán cũng được tính tương tự như vòng lặp *while* trong IARIF-AO là $O\left(\left(|C| - \left|\mathcal{B}^+\right|\right)^2\left|U^+\right||\Delta U|\right)$. Từ dòng 15 tới dòng 18, độ phức tạp của vòng lặp *for* là $O\left(\left|\mathcal{B}^+\right|\left|U^+\right||\Delta U|\right)$. Như vậy, độ phức tạp cuối cùng của IARIF-AO là $\max\left\{O\left(\left|\mathcal{B}^+\right|\left|U^+\right||\Delta U|\right), O\left(\left(|C| - \left|\mathcal{B}^+\right|\right)^2\left|U^+\right||\Delta U|\right)\right\}$. Do đó, thời gian tính toán của IARIF-AO được giảm thiểu đáng kể so với thuật toán

Thuật toán 3.2 Rút gọn thuộc tính gia tăng khi bổ sung tập đối tượng dựa trên khoảng cách họ lân cận mờ trực cảm có trọng số (IARIF-AO)

Đầu vào: Bảng quyết định $IS = (U, C \cup D)$ với $U = \{u_1, u_2, \ldots, u_n\}$, rút gọn $\mathcal{B} \subseteq C$, $\mathcal{F}_{\mathcal{B}}^{\delta,\omega}$, $\mathcal{F}_{C}^{\delta,\omega}$ và tập đối tượng bổ sung $\Delta U = \{u_{n+1}, u_{n+2}, \ldots, u_{n+z}\}$.

Đầu ra: Rút gọn xấp xỉ $\mathcal{B}^+$ trên $U^+ = U \cup \Delta U$

// khởi tạo và cập nhật

$\mathcal{B}^+ \leftarrow \mathcal{B}$

cập nhật ω^+ từ Công thức 3.11 và $\mathcal{F}_{\{a\}}^{\delta,\omega^+}$ trên U^+ với mỗi $a \in C \backslash \mathcal{B}^+$ từ Công thức 3.12.

// kiểm tra tập đối tượng bổ sung

$Z \leftarrow \Delta U$

for $i = 1$ to z **do**

  if $[\ddot{u}_{n+i}]_{\mathcal{B}^+}^{\delta,\omega^+} \subseteq [u_{n+i}]_D$ **then** $Z \leftarrow Z \backslash \{u_{n+i}\}$

end for

if $Z = \emptyset$ **then** return $\mathcal{B}^+$

set $\Delta U \leftarrow Z$, $z \leftarrow |\Delta U|$

// tìm kiếm rút gọn

tính toán $\ddot{\mathcal{D}}\left(\mathcal{F}_{\mathcal{B}^+}^{\delta,\omega^+}, \mathcal{F}_{\mathcal{B}^+\cup D}^{\delta,\omega^+}\right)$ và $\ddot{\mathcal{D}}\left(\mathcal{F}_{C}^{\delta,\omega^+}, \mathcal{F}_{C\cup D}^{\delta,\omega^+}\right)$ theo Công thức 3.13.

while $\ddot{\mathcal{D}}\left(\mathcal{F}_{\mathcal{B}^+}^{\delta,\omega^+}, \mathcal{F}_{\mathcal{B}^+\cup D}^{\delta,\omega^+}\right) > \ddot{\mathcal{D}}\left(\mathcal{F}_{C}^{\delta,\omega^+}, \mathcal{F}_{C\cup D}^{\delta,\omega^+}\right)$ **do**

  tính toán $Sig_F\left(a, \mathcal{B}^+\right)$, với mọi $a \in C \backslash \mathcal{B}^+$ theo Công thức 3.10.

  lựa chọn thuộc tính a_0 thỏa mãn: $Sig_F\left(a_0, \mathcal{B}^+\right) = \max\limits_{a \in C\backslash\mathcal{B}^+} \left\{Sig_F\left(a, \mathcal{B}^+\right)\right\}$

  $\mathcal{B}^+ \leftarrow \mathcal{B}^+ \cup \{a_0\}$

end while

// loại bỏ các thuộc tính dư thừa

for $a \in \mathcal{B}^+$ **do**

  tính toán $\ddot{\mathcal{D}}\left(\mathcal{F}_{\mathcal{B}^+\backslash\{a\}}^{\delta,\omega^+}, \mathcal{F}_{\mathcal{B}^+\backslash\{a\}\cup D}^{\delta,\omega^+}\right)$ theo Công thức 3.9.

  if $\ddot{\mathcal{D}}\left(\mathcal{F}_{\mathcal{B}^+\backslash\{a\}}^{\delta,\omega^+}, \mathcal{F}_{\mathcal{B}^+\backslash\{a\}\cup D}^{\delta,\omega^+}\right) = \ddot{\mathcal{D}}\left(\mathcal{F}_{\mathcal{B}^+}^{\delta,\omega^+}, \mathcal{F}_{\mathcal{B}^+\cup D}^{\delta,\omega^+}\right)$ **then** $\mathcal{B}^+ \leftarrow \mathcal{B}^+ \backslash \{a\}$

end for

return $\mathcal{B}^+$

ARIFW. Trong trường hợp dữ liệu được bổ sung có các giá trị thuộc tính quyết định giống nhau, công thức gia tăng được thay thế bằng Công thức 3.14 từ tính chất đầu tiên của Mệnh đề 3.10.

3.3.3 Đề xuất thuật toán rút gọn thuộc tính trên bảng quyết định thay đổi khi loại bỏ tập đối tượng

Trước hết, có thể thấy rằng khi một tập đối tượng bị loại bỏ khỏi bảng quyết định, các hạt thông tin lân cận trọng số sẽ có sự biến động. Điều này sẽ dẫn đến sự thay đổi trên các hạt thông tin lân cận trọng số mờ trực cảm. Do đó, việc xây dựng một cơ chế cập nhật các hạt thông tin này trở thành một yếu tố quan trọng. Tuy nhiên, trong nghiên cứu này, quá trình cập nhật trọng số và cơ chế cập nhật các hạt thông tin lân cận trọng số mờ trực cảm đối với các đối tượng không bị thay đổi sẽ được giữ nguyên, bất kể bảng quyết định có sự bổ sung hay loại bỏ tập đối tượng. Do đó, nghiên cứu này thiết lập $\omega^- = \omega$ với ω^- là trọng số của các thuộc tính điều kiện được thiết lập khi bảng quyết định có sự loại bỏ tập đối tượng.

Từ đây, xét bảng quyết định $IS = (U, C \cup D)$ với $U = \{u_1, u_2, \ldots, u_n\}$ và một tập đối tượng $\Delta U = \{u_n, u_{n-1}, \ldots, u_{n-z+1}\}$ được loại bỏ khỏi bảng, trong đó $z \geq 1$. Khi đó, với một thuộc tính $a \in C$, $[\ddot{u}_i]^{\delta,\omega^-}_{\{a\}}$ với $1 \leq i \leq n-z$ biểu diễn một hạt thông tin lân cận mới của thuộc tính a trên bảng quyết định có sự loại bỏ tập đối tượng. Dĩ nhiên, hạt thông tin này được đặc trưng bởi các giá trị độ thuộc và độ không thuộc của $n-z$ đối tượng còn lại trong bảng. Qua đó, nếu xét trên toàn bộ các đối tượng, chúng ta sẽ thu được một họ lân cận trọng số mờ trực cảm mới trên $U^- = U \backslash \Delta U$, ký hiệu là $\mathcal{F}^{\delta,\omega^-}_{\{a\}}$.

Mệnh đề 3.11 Cho bảng quyết định $IS = (U, C \cup D)$ với $U = \{u_1, u_2, \ldots, u_n\}$ và một tập đối tượng $\Delta U = \{u_n, u_{n-1}, \ldots, u_{n-z+1}\}$ được loại bỏ khỏi bảng, trong đó $z \geq 1$. Khi đó, khoảng cách giữa hai họ lân cận trọng số mờ trực cảm trên tập đối tượng $U^- = U \backslash \Delta U$, ký hiệu là $\ddot{\mathcal{D}}\left(\mathcal{F}^{\delta,\omega^-}_{C}, \mathcal{F}^{\delta,\omega^-}_{C\cup D}\right)$, được xác định như sau:

$$\mathcal{D}\left(\mathcal{F}^{\delta,\omega^-}_{C}, \mathcal{F}^{\delta,\omega^-}_{C\cup D}\right) = \frac{n^2\mathcal{D}\left(\mathcal{F}^{\delta,\omega}_{C}, \mathcal{F}^{\delta,\omega}_{C\cup D}\right)}{(n-z)^2} - \frac{2\sum_{i=1}^{z}\left(\left|[\ddot{u}_{n-i+1}]^{\delta,\omega}_{C}\right| - \left|[\ddot{u}_{n-i+1}]^{\delta,\omega}_{C} \cap [\ddot{u}_{n-i+1}]^{\delta,\omega}_{D}\right| - \zeta_i\right)}{(n-z)^2} \tag{3.15}$$

trong đó, $\zeta_1 = 0$ và với mọi $i \geq 2$ thì

$$\zeta_i = \sum_{j=i}^{z}\left(\gamma_{[\ddot{u}_{n-i+1}]^{\delta,\omega}_{C}}(u_{n-j+1}) - \gamma_{[\ddot{u}_{n-i+1}]^{\delta,\omega}_{C\cup D}}(u_{n-j+1}) - \eta_{[\ddot{u}_{n-i+1}]^{\delta,\omega}_{C}}(u_{n-j+1}) + \eta_{[\ddot{u}_{n-i+1}]^{\delta,\omega}_{C\cup D}}(u_{n-j+1})\right)$$

Mệnh đề 3.12 Cho bảng quyết định $IS = (U, C \cup D)$ với $U = \{u_1, u_2, \ldots, u_n\}$, $A \subseteq C$ là một rút gọn dựa vào khoảng cách giữa hai họ lân cận trọng số mờ trực cảm trên U và một tập đối tượng $\Delta U = \{u_n, u_{n-1}, \ldots, u_{n-z+1}\}$ được loại bỏ khỏi bảng, trong đó $z \geq 1$

và $U^- = U \backslash \Delta U$ là một tập đối tượng còn lại, khi đó:

1. Nếu tất cả các đối tượng trong ΔU có giá trị thuộc tính quyết định giống nhau thì:

$$\ddot{\mathcal{D}}\left(\mathcal{F}_C^{\delta,\omega^-}, \mathcal{F}_{C\cup D}^{\delta,\omega^-}\right) = \frac{n^2\mathcal{D}\left(\mathcal{F}_C^{\delta,\omega}, \mathcal{F}_{C\cup D}^{\delta,\omega}\right) - 2\sum_{i=1}^{z}\left(\left|[\ddot{u}_{n-i+1}]_C^{\delta,\omega}\right| - \left|[\ddot{u}_{n-i+1}]_C^{\delta,\omega} \cap [\ddot{u}_{n-i+1}]_D^{\delta,\omega}\right|\right)}{(n-z)^2} \tag{3.16}$$

2. $\ddot{\mathcal{D}}\left(\mathcal{F}_A^{\delta,\omega^-}, \mathcal{F}_{A\cup D}^{\delta,\omega^-}\right) = \ddot{\mathcal{D}}\left(\mathcal{F}_C^{\delta,\omega^-}, \mathcal{F}_{C\cup D}^{\delta,\omega^-}\right)$

Tính chất 2 của Mệnh đề 3.12 cung cấp cơ sở để đề xuất một điều kiện dừng cho thuật toán gia tăng và đảm bảo tính tối ưu của kích thước rút gọn mới trên bảng quyết định sau khi loại bỏ tập con đối tượng. Dựa trên các khái niệm được trình bày, Thuật toán gia tăng 3.3 sẽ được thiết kế nhằm tìm kiếm một rút gọn xấp xỉ khi bảng quyết định loại bỏ các đối tượng. Thuật toán này bao gồm hai bước chính và được trình bày theo bảng mã giả dưới đây.

Thuật toán 3.3 Rút gọn thuộc tính gia tăng khi loại bỏ tập đối tượng dựa trên khoảng cách họ lân cận mờ trực cảm có trọng số (IARIF-DO)

Đầu vào: Bảng quyết định $IS = (U, C \cup D)$ với $U = \{u_1, u_2, \ldots, u_n\}$, rút gọn $\mathcal{B} \subseteq C$, $\mathcal{F}_{\mathcal{B}}^{\delta,\omega}$ và $\mathcal{F}_C^{\delta,\omega}$, tập đối tượng bị loại bỏ $\Delta U = \{u_n, u_{n-1}, \ldots, u_{n-z+1}\}$

Đầu ra: Một rút gọn red^- on $U^- = U \backslash \Delta U$

1: khởi tạo $\mathcal{B}^- \leftarrow \mathcal{B}$

2: cập nhật $\mathcal{F}_C^{\delta,\omega^-}$

3: tính toán $\ddot{\mathcal{D}}\left(\mathcal{F}_C^{\delta,\omega^-}, \mathcal{F}_{C\cup D}^{\delta,\omega^-}\right)$ theo Công thức 3.15.

4: $\ddot{\mathcal{D}}\left(\mathcal{F}_{\mathcal{B}^-}^{\delta,\omega^-}, \mathcal{F}_{\mathcal{B}^-\cup D}^{\delta,\omega^-}\right) \leftarrow \ddot{\mathcal{D}}\left(\mathcal{F}_C^{\delta,\omega^-}, \mathcal{F}_{C\cup D}^{\delta,\omega^-}\right)$

5: **for** $a \in \mathcal{B}^-$ **do**

6: tính toán $\ddot{\mathcal{D}}\left(\mathcal{F}_{\mathcal{B}^-\backslash\{a\}}^{\delta,\omega^-}, \mathcal{F}_{\mathcal{B}^-\backslash\{a\}\cup D}^{\delta,\omega^-}\right)$ theo Công thức 3.9.

7: **if** $\ddot{\mathcal{D}}\left(\mathcal{F}_{\mathcal{B}^-\backslash\{a\}}^{\delta,\omega^-}, \mathcal{F}_{\mathcal{B}^-\backslash\{a\}\cup D}^{\delta,\omega^-}\right) = \ddot{\mathcal{D}}\left(\mathcal{F}_{\mathcal{B}^-}^{\delta,\omega^-}, \mathcal{F}_{\mathcal{B}^-\cup D}^{\delta,\omega^-}\right)$ **then** $\mathcal{B}^- \leftarrow \mathcal{B}^- \backslash \{a\}$

8: **end for**

9: **return** $\mathcal{B}^-$

Giả sử $|\Delta U|$ biểu diễn số lượng các đối tượng được loại bỏ từ bảng quyết định và $\left|U^-\right|$ biểu diễn tổng số đối tượng trên bảng quyết định còn lại. Độ phức tạp trong dòng 2 của thuật toán đề xuất là $O\left(\left|\mathcal{B}^-\right|\left|U^-\right||\Delta U|\right)$ và quá trình tính toán gia tăng tại dòng 3 là $O\left(\left|U^-\right||\Delta U|\right)$. Tương tự như quá trình loại bỏ các thuộc tính dư thừa trong tập rút gọn của Thuật toán 3.2, độ phức tạp trong vòng lặp *for* của IARIF-DO là $O\left(\left|\mathcal{B}^-\right|\left|U^-\right||\Delta U|\right)$.

Như vậy, độ phức tạp trên toàn bộ thuật toán IARIF-DO là $O\left(\left|\mathcal{B}^{-}\right|\left|U^{-}\right|\left|\Delta U\right|\right)$. Với độ phức tạp này, thời gian thực thi của thuật toán gia tăng đề xuất IARIF-DO giảm thiểu đáng kể so với ARIFW.

3.4 Thử nghiệm và đánh giá các thuật toán đề xuất

3.4.1 Hiệu năng của thuật toán IARIF-AO

Để đánh giá hiệu quả của thuật toán được đề xuất, trước tiên các tập dữ liệu trong Bảng 3.1 được chia thành hai phần xấp xỉ nhau, ký hiệu là U_{ori} và U_{inc}. Cụ thể, U_{ori} sẽ được sử dụng cho các thuật toán rút gọn thuộc tính trên bảng quyết định cố định, bao gồm các thuật toán: NIFS [82], WNRS [29], W-MGMN [86], FMIFRFS [56], ARPD và ARIFW. Trong đó, NIFS là thuật toán dựa trên tập thô mờ, WNRS và W-MGMN là các thuật toán dựa trên tập thô lân cận, trong khi FMIFRFS và ARPD là các thuật toán dựa trên tập thô mờ trực cảm. Tiếp theo, tập U_{inc} sẽ được chia nhỏ thành năm phần xấp xỉ bằng nhau. Mỗi phần sau đó sẽ được bổ sung lần lượt vào U_{ori} để đánh giá hiệu quả của các thuật toán gia tăng.

Bảng 3.1: Các tập dữ liệu thử nghiệm cho IARIF-AO và một số thuật toán

TT	Tập dữ liệu	Số đối tượng	$\|U_{ori}\|$	$\|U_{inc}\|$	Số thuộc tính	Số lớp	Nguồn
1	Leaf	340	170	170	15	30	UCI
2	Ionosphere	351	175	176	34	2	UCI
3	Vehicle	846	423	423	18	4	OpenML
4	Vowel	990	495	495	12	11	OpenML
5	German	1000	500	500	20	2	UCI
6	LSVT	126	63	63	310	2	OpenML
7	Leukemia	72	36	36	7129	2	UCI
8	Mfeat	2000	1000	1000	76	10	OpenML
9	PD	756	378	378	754	2	UCI
10	Segmentation	2310	1155	1155	19	7	OpenML
11	Wall	5456	2728	2728	24	4	OpenML
12	Waveform2	5000	2500	2500	40	3	OpenML

Như đã mô tả, với mỗi tập dữ liệu, các thuật toán sẽ tiến hành duyệt giá trị tham số từ 0 đến 1 để tìm ra tập con thuộc tính mang lại độ chính xác phân lớp cao nhất. Đối với thuật toán được đề xuất, tham số δ được sử dụng để xây dựng các hạt thông tin lân cận mờ trực cảm có trọng số. Các giá trị khác nhau của δ có thể dẫn đến việc hình thành các hạt thông tin khác nhau và có tác động đến hiệu quả của thuật toán. Hình 3.4 cho thấy tham số δ ảnh hưởng đáng kể đến hiệu quả phân lớp khi sử dụng KNN. Trong một số tập dữ liệu, việc điều chỉnh tham số này dẫn đến xu hướng tăng hoặc giảm rõ rệt về độ chính xác phân lớp như trên các tập dữ liệu LSVT, German, Ionosphere, PD

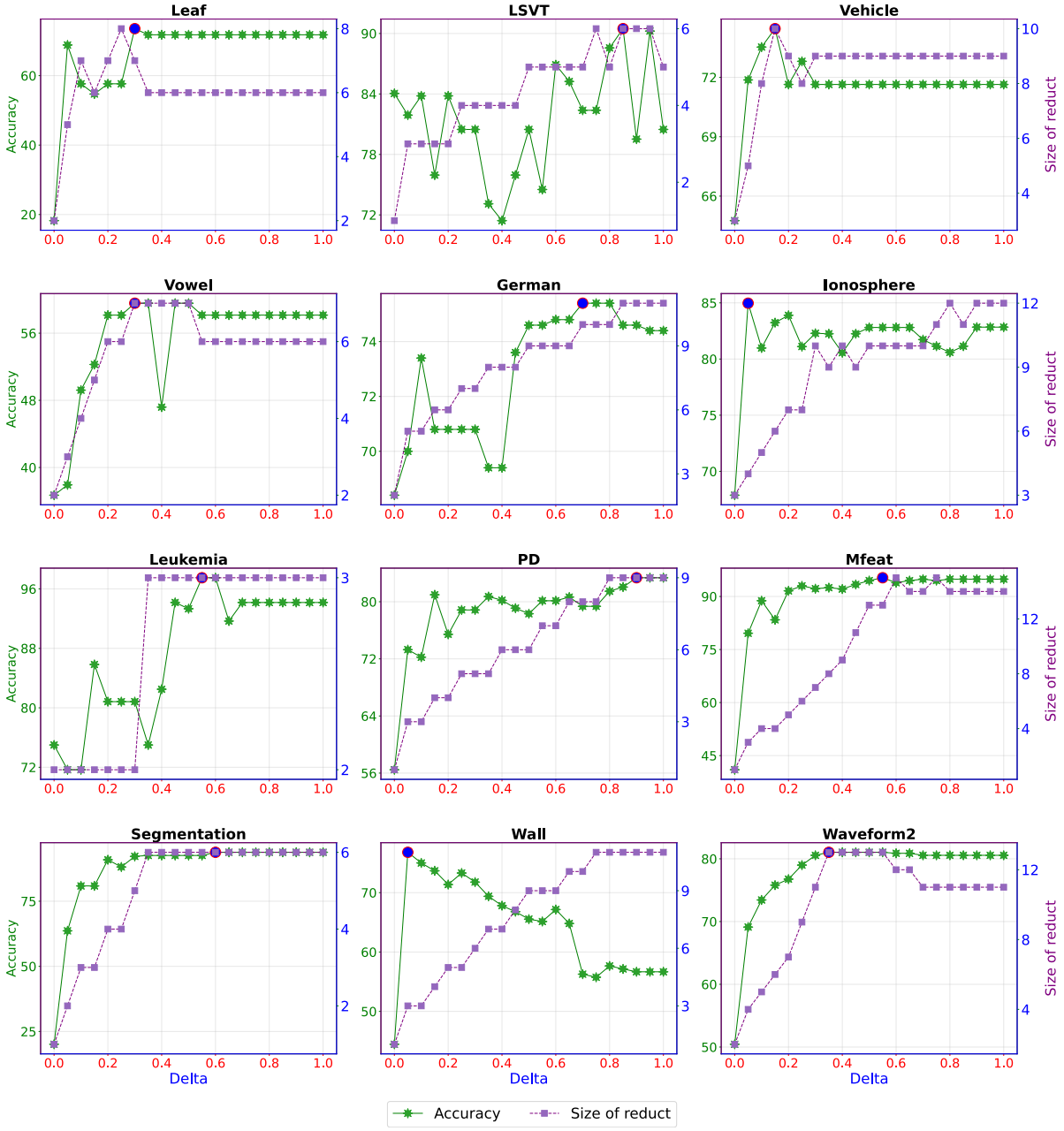
và Wall. Tuy nhiên, ở một số tập khác như Leaf, Vehicle, Vowel và Segmentation, khi giá trị δ đạt đến một ngưỡng nhất định, tác động của nó đến độ chính xác phân lớp có sự suy giảm.

Bảng 3.2: Kích thước rút gọn và tham số của các thuật toán với KNN trên U_{ori}

Tập dữ liệu	NIFS		WNRS		W-MGMN		FMIFRFS		ARPD		ARIFW	
	k	$\|\mathcal{B}\|$	*delta*	$\|\mathcal{B}\|$	*delta*	$\|\mathcal{B}\|$	*epsilon*	$\|\mathcal{B}\|$	*alpha*	$\|\mathcal{B}\|$	*delta*	$\|\mathcal{B}\|$
Leaf	0.002	10	0.15	9	0.05	9	0.75	10	0.4	9	0.3	7
Ionosphere	0.001	9	0.1	6	0.1	9	1.0	4	0.75	5	0.05	4
Vehicle	0.001	13	0.1	14	0.0	10	0.7	17	0.7	8	0.15	10
Vowel	0.002	6	0.1	9	0.05	8	0.35	8	0.5	7	0.3	7
German	0.002	6	0.05	12	0.0	9	0.55	12	0.05	6	0.7	10
LSVT	0.001	6	0.6	10	0.3	18	0.05	18	0.8	7	0.85	6
Leukemia	0.001	2	0.05	2	0.1	10	0.85	3	0.4	7	0.55	3
Mfeat	0.003	56	0.35	27	0.05	21	0.5	23	0.6	22	0.55	13
PD	0.001	25	0.65	21	0.1	6	0.7	31	0.5	41	0.9	9
Segmentation	0.001	14	0.05	9	0.05	13	0.9	10	0.8	6	0.6	6
Wall	0.006	8	0.05	14	0.2	15	1.0	4	0.9	3	0.05	3
Waveform2	0.003	26	0.05	8	0.05	31	0.45	15	0.2	24	0.35	13
Trung bình		16.08		11.75		13.25		12.92		12.08		7.58

Bảng 3.2 trình bày kích thước các rút gọn và các tham số tương ứng của từng thuật toán. Rõ ràng, các phương pháp đã giảm số lượng thuộc tính so với dữ liệu gốc một cách hiệu quả. Tuy nhiên, cần lưu ý rằng đối với các tập dữ liệu có độ chính xác phân lớp ban đầu thấp như Leaf, Ionosphere, Vehicle, Vowel, German, LSVT, Wall và Waveform2, các thuật toán NIFS, WNRS và W-MGMN đôi khi cho kết quả kém hiệu quả hơn. Ngược lại, các thuật toán FMIFRFS, ARPD và ARIFW thường trích xuất được các rút gọn nhỏ hơn so với các thuật toán khác. Khi xem xét tất cả các trường hợp, ARPD và ARIFW cũng thường xuyên thu được các tập rút gọn nhỏ nhất trên hầu hết các tập dữ liệu. Cụ thể, ARPD đạt được tập rút gọn nhỏ nhất ở 4 tập dữ liệu, trong khi ARIFW đạt được ở 6 tập dữ liệu. Ngoài ra, kích thước rút gọn trung bình thu được bởi ARIFW cũng là nhỏ nhất. Do đó, có thể khẳng định rằng thuật toán được đề xuất có hiệu quả cao trong việc rút gọn thuộc tính, đặc biệt là đối với các tập dữ liệu có số chiều lớn.

Bảng 3.3 trình bày thời gian tính toán của các thuật toán rút gọn thuộc tính khi áp dụng trên tập dữ liệu U_{ori}. Không ngạc nhiên khi các thuật toán dựa trên tập thô lân cận có trọng số và tập thô mờ có thời gian chạy nhanh hơn so với các thuật toán dựa trên tập thô mờ trực cảm. Nguyên nhân là do thuật toán WNRS thực hiện các phép tính trên tập rõ, vốn có quy trình xử lý nhanh và đơn giản. Mặc dù thuật toán NIFS dựa trên phương pháp tập thô mờ, nhưng nó không yêu cầu tính toán ma trận quan hệ và sử dụng một độ đo đơn giản để lựa chọn các thuộc tính quan trọng. Ngược lại,



Hình 3.4: Độ chính xác phân lớp và kích thước rút gọn khi duyệt δ với bộ phân lớp KNN

WNRS phải trải qua quá trình tính trọng số cho các thuộc tính, nên NIFS thể hiện tốc độ xử lý tốt nhất trong số các thuật toán.

Đáng chú ý, mặc dù W-MGMN thực hiện trên tập rõ, nhưng lại có thời gian thực thi chậm nhất trong tất cả các thuật toán. Nguyên nhân là do thuật toán này không chỉ tính toán trọng số cho từng thuộc tính, mà còn sử dụng thuật toán DBSCAN để nhóm các hạt tri thức có cùng ngưỡng trọng số thành một hạt duy nhất. Bên cạnh đó, thuật toán này còn sử dụng một độ đo phức tạp để đánh giá ý nghĩa của thuộc tính. Khi xét riêng các thuật toán trên khung tập thô mờ trực cảm, thuật toán được đề xuất đạt tốc độ thực thi nhanh nhất. Cụ thể, thời gian xử lý trung bình của thuật toán là 28.168

Bảng 3.3: Thời gian thực thi của ARIFW và các thuật toán trên U_{ori}

TT	Tập dữ liệu	NIFS	WNRS	W-MGMN	FMIFRFS	ARPD	ARIFW
1	Leaf	0.0028	0.0534	1.4969	0.1242	0.0186	0.0152
2	Ionosphere	0.0033	0.1806	3.4146	0.1257	0.0192	0.0186
3	Vehicle	0.0074	0.2942	14.757	0.7542	0.3798	0.3759
4	Vowel	0.0051	0.2253	22.159	1.2874	0.2784	0.2519
5	German	0.0048	0.6966	21.973	0.6743	0.5128	0.5236
6	LSVT	0.0743	0.3816	2.7744	1.9096	0.7639	0.7483
7	Leukemia	13.433	6.6504	38.491	231.83	229.15	209.56
8	Mfeat	6.8312	33.086	60.789	36.674	23.382	17.947
9	PD	2.9826	61.321	307.52	103.65	108.68	28.183
10	Segmentation	1.0652	4.0692	214.63	10.687	5.7058	4.6415
11	Wall	1.0913	29.779	598.78	15.774	8.0352	7.8904
12	Waveform2	1.6991	56.158	626.06	57.093	72.449	43.865
Trung bình		2.2667	16.075	109.40	38.382	37.448	26.168

giây, trong khi FMIFRFS mất 38.382 giây và ARPD mất 37.448 giây. Do đó, thuật toán đề xuất thể hiện hiệu quả xử lý mạnh mẽ trong việc giảm thiểu thời gian thực thi, đặc biệt trong không gian tập thô mờ trực cảm.

Kết quả so sánh độ chính xác phân lớp của các thuật toán rút gọn thuộc tính trên bảng quyết định cố định được trình bày trong Bảng 3.4. Trong 12 trường hợp, tập rút gọn của thuật toán đề xuất đạt hiệu quả phân lớp vượt trội so với dữ liệu gốc. Hơn nữa, độ chính xác phân lớp trung bình của thuật toán đề xuất khi sử dụng KNN là 82.18%, cao hơn rõ rệt so với giá trị của dữ liệu gốc là 74.54%. Trên các tập dữ liệu có độ chính xác phân lớp ban đầu thấp và số lượng đối tượng lớn, chẳng hạn như German, Wall và Waveform2, các thuật toán dựa trên tập thô mờ trực cảm luôn thu được các rút gọn có hiệu quả phân lớp vượt trội so với các phương pháp dựa trên tập thô mờ và tập thô lân cận có trọng số. Điều này làm nổi bật lợi thế của tập thô mờ trực cảm, trong đó việc kết hợp hàm không thuộc giúp mô hình giảm thiểu ảnh hưởng của các đối tượng nhiễu trong tập dữ liệu.

Từ Bảng 3.4, tỷ lệ Thắng/Hòa/Thua giữa thuật toán đề xuất ARIFW và các thuật toán khác lần lượt là: 10/0/2, 8/1/3. 11/0/1, 9/1/2 và 10/0/2. Có thể thấy rõ rằng trên hầu hết các tập dữ liệu, các tập rút gọn được tạo ra bởi thuật toán ARIFW mang lại kết quả phân lớp tốt hơn so với các thuật toán khác, mặc dù số lượng thuộc tính được chọn bởi ARIFW ít hơn. Hơn nữa, thuật toán đề xuất đã đạt độ chính xác phân lớp cao nhất trong 8 trên 12 trường hợp, điều này chứng minh rằng thuật toán đề xuất đã lựa chọn hiệu quả các thuộc tính quan trọng phù hợp với nhiều tập dữ liệu khác nhau.

Tiếp theo, luận án tiến hành so sánh hiệu suất của thuật toán IARIF-AO với các

thuật toán AIFSA-FKD [82], W-MGMA [86] và IARPD-AO. Trong đó, AIFSA-FKD là thuật toán gia tăng dựa trên tập thô mờ sử dụng độ đo khoảng cách tri thức mờ với đầu vào là tập rút gọn từ thuật toán NIFS. Thuật toán gia tăng W-MGMA dựa trên tập thô lân cận có trọng số, sử dụng độ đo entropy, với đầu vào là tập rút gọn được xác định từ thuật toán W-MGMN. Trong khi đó, thuật toán IARPD-AO dựa trên tập thô mờ trực cảm, sử dụng độ đo khoảng cách phân hoạch lấy rút gọn từ thuật toán ARPD làm đầu vào và thuật toán đề xuất IARIF-AO sử dụng tập rút gọn từ thuật toán IARFW làm đầu vào. Các thuật toán này sẽ xử lý trên tập dữ liệu khi bổ sung lần lượt các tập đối tượng từ U_1 đến U_5 vào U_{ori}.

Bảng 3.4: So sánh độ chính xác phân lớp của ARIFW với các thuật toán trên U_{ori}.

Tập dữ liệu	Tập gốc	NIFS	WNRS	W-MGMN	FMIFRFS	ARPD	ARIFW
Leaf	71.76 ± 10.12	74.12 ± 12.39	75.88 ± 8.5	**79.41 ± 9.58**	72.35 ± 11.78	74.12 ± 9.56	73.53 ± 9.58
Ionosphere	76.57 ± 9.22	82.19 ± 8.94	80.59 ± 8.46	82.81 ± 5.21	84.02 ± 11.23	84.05 ± 9.73	**84.97 ± 9.94**
Vehicle	67.85 ± 4.23	67.40 ± 5.82	67.60 ± 4.61	71.87 ± 7.05	68.08 ± 4	67.84 ± 5.51	**74.46 ± 6.91**
Vowel	48.81 ± 12.36	61.53 ± 11.32	**65.44 ± 5.71**	59.14 ± 8.42	50.43 ± 13.99	52.43 ± 15.45	59.57 ± 9.55
German	73.20 ± 4.21	72.00 ± 6.13	72.80 ± 4.02	74.80 ± 4.12	75.20 ± 3.12	73.60 ± 3.2	**75.40 ± 4.9**
LSVT	74.30 ± 14.9	79.29 ± 11.67	**90.48 ± 10.81**	79.05 ± 12.77	86.90 ± 9.95	88.81 ± 10.49	**90.48 ± 7.82**
Leukemia	86.67 ± 13.54	91.67 ± 12.91	96.67 ± 10	90.00 ± 22.91	**97.50 ± 7.5**	96.67 ± 10	**97.50 ± 7.5**
Mfeat	92.50 ± 1.8	91.00 ± 1.9	92.00 ± 2.86	93.30 ± 1.62	94.50 ± 1.91	94.90 ± 1.64	**95.30 ± 1.73**
PD	79.11 ± 8.42	75.41 ± 5.97	**86.54 ± 7.38**	79.11 ± 8.42	83.88 ± 8.9	83.88 ± 6.98	83.36 ± 5.47
Segmentation	93.77 ± 2.45	92.82 ± 2.37	93.00 ± 3.08	93.51 ± 2.11	93.69 ± 2.37	93.43 ± 2.11	**93.78 ± 2.33**
Wall	63.78 ± 9.39	68.37 ± 8.91	64.59 ± 8.29	71.30 ± 7.82	74.63 ± 6.44	75.00 ± 4.47	**76.80 ± 5.15**
Waveform2	78.20 ± 2.02	78.28 ± 2.21	72.96 ± 2.63	79.40 ± 1.92	**81.72 ± 1.3**	79.96 ± 2.19	81.04 ± 2.8
Trung bình	74.54 ± 7.72	77.84 ± 7.55	79.88 ± 6.36	79.48 ± 7.66	80.24 ± 6.87	80.39 ± 6.78	**82.18 ± 6.14**
Win/Draw/Loss	12/0/0	10/0/2	8/1/3	11/0/1	9/1/2	10/0/2	*best*

Bảng 3.5 trình bày thời gian thực thi của thuật toán IARIF-AO so với các thuật toán AIFSA-FKD, W-MGMA và IARPD-AO. Có thể thấy rằng các thuật toán gia tăng cho kết quả vượt trội hơn so với các thuật toán tương ứng của chúng khi xử lý trên bảng quyết định cố định, mặc dù phải thực hiện tính toán trên các tập dữ liệu có số lượng đối tượng lớn hơn.

Tương tự như kết quả so sánh trong Bảng 3.3, các thuật toán AIFSA-FKD có thời gian thực thi nhanh hơn so với các thuật toán dựa trên phương pháp tập thô mờ trực cảm. Khi xem xét các thuật toán dựa trên tập thô mờ trực cảm, có thể thấy thuật toán đề xuất có thời gian xử lý nhanh hơn trong hầu hết các bước gia tăng. Điều này là do đặc điểm của quan hệ lân cận có trọng số, giúp thu hẹp không gian tính toán bằng cách chỉ tập trung vào các đối tượng nằm trong một bán kính nhất định. Thêm vào đó, thuật toán này cũng chọn lọc được ít thuộc tính hơn. Do đó, thuật toán kết thúc sớm và dễ

dàng thu được một rút gọn tối ưu.

Bảng 3.5: So sánh thời gian thực thi của IARIF-AO với các thuật toán gia tăng

Thuật toán	Tập dữ liệu	Tập dữ liệu bổ sung					Tập dữ liệu	Tập dữ liệu bổ sung				
		U_1	U_2	U_3	U_4	U_5		U_1	U_2	U_3	U_4	U_5
AIFSA-FKD	Leaf	0.0008	0.0006	0.0009	0.0007	0.0015	Leukemia	0.0019	0.0026	0.0037	0.0086	0.0001
W-MGMA		0.0252	0.0189	0.0211	0.0756	0.0210		9.3544	9.5742	12.772	10.423	16.558
IARPD-AO		0.0068	0.0121	0.0122	0.0125	0.0165		0.3872	0.0022	2.3345	1.8313	1.9353
IARIF-AO		0.0072	0.0065	0.0058	0.0075	0.0061		0.3456	0.0016	0.3834	0.5223	0.5457
AIFSA-FKD	Ionosphere	0.0019	0.0010	0.0012	0.0014	0.0008	Mfeat	0.0704	0.0968	0.0997	0.1059	0.1272
W-MGMA		0.0560	0.0821	0.2074	0.2627	0.3234		5.3126	5.8170	6.5846	7.1969	9.2799
IARPD-AO		0.0042	0.0102	0.0061	0.0066	0.0071		4.2163	5.0467	6.9885	0.8157	8.1484
IARIF-AO		0.0025	0.0028	0.0031	0.0033	0.0034		0.6138	2.2373	4.3382	0.8016	5.7476
AIFSA-FKD	Vehicle	0.0013	0.0018	0.0013	0.0014	0.0022	PD	0.4048	0.0095	0.0107	0.3561	0.0197
W-MGMA		0.1276	0.1788	0.2452	0.5774	0.8665		9.0822	10.201	14.163	23.494	63.348
IARPD-AO		0.0825	0.0635	0.0654	0.0681	0.0692		2.6713	3.1246	2.7535	2.9667	3.0134
IARIF-AO		0.0454	0.0663	0.0486	0.0499	0.0539		0.0599	0.0615	1.2382	0.0644	3.8178
AIFSA-FKD	Vowel	0.0015	0.0036	0.0017	0.0019	0.0021	Segmentation	0.0167	0.0040	0.0057	0.0066	0.0075
W-MGMA		0.0827	0.1150	0.1482	0.1892	0.4946		0.8096	1.2081	1.4557	1.8673	5.9377
IARPD-AO		0.0884	0.0898	0.0913	0.0939	0.1127		0.8562	0.2748	0.3091	0.3263	0.3594
IARIF-AO		0.0685	0.0693	0.0714	0.0728	0.0819		0.7522	0.2335	0.2557	0.2785	0.2993
AIFSA-FKD	German	0.0020	0.0014	0.0015	0.0017	0.0018	Wall	0.0069	0.0097	0.0103	0.0344	0.0152
W-MGMA		0.2318	0.3126	0.7743	0.4638	0.5761		6.8664	9.4415	12.396	15.715	19.285
IARPD-AO		0.0891	0.0908	0.0926	0.0945	0.0973		0.5813	0.6694	1.7047	1.9520	2.1213
IARIF-AO		0.0714	0.1023	0.0725	0.0744	0.0769		0.5748	0.6592	0.9121	1.1385	1.3192
AIFSA-FKD	LSVT	0.0016	0.0024	0.0081	0.0001	0.0143	Waveform2	0.1189	0.1242	0.1207	0.1328	0.1344
W-MGMA		0.0937	0.1210	0.1523	0.5508	0.2355		5.6647	7.7862	10.142	12.881	15.870
IARPD-AO		0.0297	0.0031	0.0316	0.0328	0.0353		22.983	25.441	28.407	30.286	40.532
IARIF-AO		0.0492	0.0283	0.0298	0.0037	0.0315		3.8428	5.1245	8.6328	11.096	13.991

Độ chính xác phân lớp và kích thước rút gọn thu được từ các thuật toán được biểu diễn trong Bảng 3.6, các thuật toán dựa trên mô hình tập thô mờ trực cảm thường trích chọn được các rút gọn nhỏ hơn so với các thuật toán dựa trên tập thô lân cận có trọng số hoặc tập thô mờ trong phần lớn các bước gia tăng. Đáng chú ý, với các tập dữ liệu có độ chính xác phân lớp ban đầu thấp như Leaf, Ionosphere, Vehicle, Vowel, Wall và Waveform2, số lượng thuộc tính được trả về bởi IARPD-AO và IARIF-AO nhỏ hơn đáng kể, giúp chúng có tính cạnh tranh cao về kích thước rút gọn. Từ Bảng 3.6, trong tổng số 60 trường hợp, các rút gọn thu được từ IARPD-AO và IARIF-AO có kích thước nhỏ nhất trong 20 và 43 trường hợp. Điều này càng nhấn mạnh hiệu quả của phương pháp đề xuất trong việc loại bỏ các thuộc tính dư thừa khỏi bảng quyết định trong các bước gia tăng, đặc biệt trong các trường hợp bảng quyết định chứa các thuộc tính làm suy giảm hiệu suất phân lớp. Tương tự như kết quả của Thuật toán 2.2, tại một số bước gia tăng, rút gọn thu được không có sự thay đổi. Nguyên nhân là do kích thước của dữ liệu được bổ sung là không đáng kể, không làm biến đổi các đặc trưng của dữ liệu ban đầu cũng như không làm thay đổi sự phân bố và số lượng đối tượng trong các hạt thông tin lân cận trọng số. Do đó, tập rút gọn vẫn bảo toàn đầy đủ thông tin của dữ liệu, duy trì kích thước ổn định và đảm bảo hiệu quả phân lớp trên tập dữ liệu mới.

Bảng 3.6: Kích thước rút gọn, độ chính xác phân lớp của IARIF-AO và các thuật toán gia tăng

TT	Tập bổ sung	Tập gốc	AIFSA-FKD		W-MGMA		IARPD-AO		IARIF-AO	
			$\|\mathcal{B}^+\|$	*Accuracy*	$\|\mathcal{B}^+\|$	*Accuracy*	$\|\mathcal{B}^+\|$	*Accuracy*	$\|\mathcal{B}^+\|$	*Accuracy*
1	$\|U_1\| = 204$	68.31±10.89	11	66.31±9.46	**10**	**80.9±7.77**	**10**	69.21±7.41	**10**	76.95±6.23
	$\|U_2\| = 238$	64.33±11.37	11	62.28±9.47	**10**	76.09±7.86	11	67.68±9.5	11	**76.94±8.85**
	$\|U_3\| = 272$	60.75±11.77	12	59.66±10.03	**10**	70.25±5.6	11	62.22±10.6	11	**72.04±6.08**
	$\|U_4\| = 306$	58.78±7.39	12	57.19±7.3	13	60.13±6.76	**11**	58.82±6.86	12	**70.25±5.03**
	$\|U_5\| = 340$	60.59±6.34	13	60.59±5.13	15	60.59±6.34	**12**	61.18±5.55	**12**	**70.59±4.92**
2	$\|U_1\| = 210$	77.62±10.22	13	80.95±10.43	9	83.81±10.03	5	**84.76±8.46**	**4**	**84.76±8.73**
	$\|U_2\| = 245$	79.22±9.77	14	82.47±10.6	9	**85.38±8.16**	6	82.40±6.81	**4**	84.43±6.12
	$\|U_3\| = 280$	80.71±7	15	82.14±8.45	10	86.07±6.86	6	85.36±4.91	**4**	**87.14±7**
	$\|U_4\| = 315$	83.24±5.88	16	84.85±7.86	11	86.76±8.77	6	87.33±5.1	**4**	**88.27±5.45**
	$\|U_5\| = 351$	83.76±6.39	16	84.90±8.18	12	87.19±5.11	6	88.60±4.61	**4**	**89.72±4.66**
3	$\|U_1\| = 507$	67.27±5.5	13	67.08±4.43	**10**	71.20±5.79	11	66.09±4.9	**10**	**71.59±3.5**
	$\|U_2\| = 591$	68.52±5.62	14	69.03±3.43	**10**	69.71±5.91	11	68.35±4.85	11	**70.94±3.78**
	$\|U_3\| = 675$	67.72±5.57	14	66.97±4.75	**10**	68.87±5.04	11	67.26±4.29	11	**69.64±3.35**
	$\|U_4\| = 759$	69.03±2.86	14	67.72±3.29	**11**	**73.26±4.57**	**11**	68.11±3.02	**11**	69.17±2.93
	$\|U_5\| = 846$	70.45±1.85	15	69.28±3.46	13	**74.71±2.52**	**11**	68.80±3.76	**11**	70.34±3
4	$\|U_1\| = 594$	35.86±14.33	**6**	**63.96±7.8**	8	53.88±7.92	10	35.69±15.36	7	52.23±10.88
	$\|U_2\| = 693$	45.81±14.38	11	43.94±14.86	8	**58.78±12.42**	10	45.37±14.21	**7**	58.37±14.04
	$\|U_3\| = 792$	38.12±6.99	11	37.37±8.17	8	53.29±6.91	10	37.62±6.64	**7**	**53.80±8.92**
	$\|U_4\| = 891$	40.62±11.9	11	39.73±12.38	8	53.88±8.44	10	39.94±11.57	**7**	**56.35±9.93**
	$\|U_5\| = 990$	51.52±12.38	11	49.29±11.67	10	**68.38±4.65**	11	52.02±11.3	**8**	63.33±11.06
5	$\|U_1\| = 600$	71.17±4.89	7	72.00±6.13	9	74.00±3.27	**6**	**74.33±5.69**	10	73.50±5.84
	$\|U_2\| = 700$	72.57±4.46	7	71.71±4.81	9	**72.14±4.34**	**6**	71.00±3.78	11	**72.14±3.9**
	$\|U_3\| = 800$	71.12±4.27	7	71.25±4.64	10	**71.62±2.91**	**6**	69.62±3.4	11	71.12±4.16
	$\|U_4\| = 900$	72.67±3.15	7	70.56±3.27	10	**72.89±2.12**	**6**	69.56±2.18	11	71.78±4.19
	$\|U_5\| = 1000$	73.00±2.79	7	70.50±7.47	10	71.30±4.73	**6**	71.00±5.1	11	**73.10±4.18**
6	$\|U_1\| = 75$	76.07±16.23	10	81.25±10.36	18	82.32±10.67	**8**	86.61±9.08	**8**	**88.04±6.92**
	$\|U_2\| = 87$	80.69±13.21	14	79.17±10.28	18	84.86±11.05	**8**	86.11±9.92	9	**90.69±7.3**
	$\|U_3\| = 99$	82.56±11.1	23	82.89±9.98	18	82.78±10.14	**9**	**92.00±9.8**	10	91.00±8.31
	$\|U_4\| = 111$	83.71±12.11	23	84.55±11.54	20	82.80±9.56	**10**	87.27±9.27	**10**	**90.00±10.33**
	$\|U_5\| = 126$	83.97±14.31	36	81.60±11.47	20	81.86±8.78	**11**	85.64±7.29	**11**	**89.68±4.99**
7	$\|U_1\| = 43$	84.50±13.68	**3**	87.00±14	19	92.50±11.46	8	95.00±10	4	**97.50±7.5**
	$\|U_2\| = 50$	90.00±13.42	**4**	92.00±13.27	26	92.00±9.8	8	**96.00±8**	**4**	**96.00±8**
	$\|U_3\| = 57$	85.67±10.44	**5**	**94.67±8.19**	36	89.67±11.3	13	93.00±8.62	**5**	**94.67±8.19**
	$\|U_4\| = 64$	80.95±12.78	7	89.29±7.08	41	90.71±9.9	16	93.81±7.62	**6**	**96.67±6.67**
	$\|U_5\| = 72$	84.46±13.51	**7**	86.07±12.8	55	88.93±7.9	19	88.75±10.66	**7**	**92.86±11.52**
8	$\|U_1\| = 1200$	91.00±1.93	57	89.83±1.38	21	92.17±1.67	23	93.67±1.35	**13**	**94.08±1.84**
	$\|U_2\| = 1400$	90.64±1.87	58	90.86±1.27	21	92.57±1.36	24	94.00±1.36	**16**	**94.07±1.69**
	$\|U_3\| = 1600$	88.88±2.22	59	89.38±1.53	**21**	90.81±1.58	26	92.19±1.77	**21**	**92.69±2.22**
	$\|U_4\| = 1800$	89.78±2.11	60	90.17±1.45	**21**	91.61±1.48	26	92.56±1.71	**21**	**93.22±2.29**
	$\|U_5\| = 2000$	80.70±1.36	61	80.30±1.05	**22**	82.60±2.11	27	82.90±1.92	25	**82.95±1.27**
9	$\|U_1\| = 453$	81.94±7.02	29	79.30±7.99	**6**	79.50±8.88	41	83.94±7.36	9	**82.60±6.51**
	$\|U_2\| = 528$	80.11±6.08	29	76.51±6.11	**7**	79.72±7.2	43	78.96±6.49	9	**82.39±5.07**
	$\|U_3\| = 603$	80.28±7.41	29	76.47±6.42	**10**	75.62±3.57	43	80.10±8.41	**10**	**81.43±4.82**
	$\|U_4\| = 678$	79.64±6.1	32	76.83±6.4	19	78.16±4.88	43	**79.20±5.4**	**10**	78.89±3.34
	$\|U_5\| = 756$	79.22±6.07	32	71.93±9.22	75	81.07±6.41	43	77.24±5.26	**12**	**81.77±2.06**
10	$\|U_1\| = 1386$	94.88±2.3	15	94.45±2.27	13	94.74±2.63	11	94.66±2.04	**10**	**95.60±2.62**
	$\|U_2\| = 1617$	94.56±2.28	15	94.50±2.21	13	94.93±2.09	11	95.06±2.26	**10**	**95.86±1.83**
	$\|U_3\| = 1848$	94.91±2.11	15	94.86±2.04	13	95.24±2.21	11	95.18±2.1	**10**	**95.73±2.03**
	$\|U_4\| = 2079$	95.05±1.49	15	95.29±1.56	13	95.82±1.36	11	95.53±1.84	**10**	**95.96±1.26**
	$\|U_5\| = 2310$	95.28±1.52	15	94.94±1.3	15	**96.32±1.06**	11	95.15±1.88	**10**	96.28±1.2
11	$\|U_1\| = 3273$	69.02±4.76	8	71.56±4.99	15	75.50±3.75	**3**	78.77±5.74	**3**	**81.06±4.79**
	$\|U_2\| = 3818$	72.90±8.31	8	74.18±7.67	15	77.35±6.44	**3**	80.20±3.14	**3**	**80.78±4.56**
	$\|U_3\| = 4363$	75.50±6.37	8	75.53±6.31	15	80.34±5.33	**3**	**81.05±4.12**	**3**	80.34±4.71
	$\|U_4\| = 4908$	76.59±7.05	9	77.16±6.81	15	81.95±5.56	**3**	82.87±4.89	**3**	**83.01±4.25**
	$\|U_5\| = 5456$	77.26±5.91	9	77.7±6.35	15	82.24±4.9	**3**	83.03±3.04	**3**	**83.32±4.56**
12	$\|U_1\| = 3000$	78.53±1.84	27	78.57±2.79	31	78.90±2.22	24	79.90±1.91	**15**	**80.00±1.8**
	$\|U_2\| = 3500$	78.29±2.26	28	79.66±1.83	31	78.83±1.83	24	**81.11±1.71**	**15**	80.91 ± 2.42
	$\|U_3\| = 4000$	79.52±2.01	29	80.48±2.15	31	80.20±2.58	24	81.12±1.96	**15**	**81.50±2.02**
	$\|U_4\| = 4500$	79.80±1.63	30	80.38±2.08	31	79.78±1.95	24	**81.89±1.35**	**15**	81.40±2.08
	$\|U_5\| = 5000$	80.08±1.46	31	80.90±2.03	31	80.24±1.45	24	81.34±2.1	**15**	**82.32±1.83**

Để đánh giá toàn diện hơn hiệu quả của các rút gọn thu được, luận án tiếp tục phân tích hiệu quả phân lớp dựa trên các rút gọn này. Có thể thấy trong phần lớn các bước gia tăng, các rút gọn xấp xỉ trả về từ các thuật toán thường có độ chính xác cao hơn hoặc tương đương so với dữ liệu gốc. Tuy nhiên, ở các bước gia tăng mà tập dữ liệu có sự thay đổi đáng kể về độ chính xác phân lớp, thuật toán AIFSA-FKD thường gặp khó khăn. Ví dụ, với các tập dữ liệu như Leaf, Vowel, LSVT và PD, độ chính xác phân lớp của các rút gọn từ AIFSA-FKD trong một số bước gia tăng thường thấp hơn so với ban đầu, trong khi các rút gọn từ các thuật toán khác vẫn tiếp tục cải thiện hiệu quả. Nguyên nhân là do các thuật toán W-MGMA, ARPD-AO và IARIF-AO có cơ chế cập nhật hạt thông tin khi một tập đối tượng mới được thêm vào.

Rõ ràng, các thuật toán dựa trên phương pháp tập thô mờ trực cảm thường cho kết quả vượt trội so với các thuật toán khác trên các tập dữ liệu có độ chính xác phân loại ban đầu thấp và số lượng đối tượng lớn, chẳng hạn như Wall và Waveform2. Về mặt trực cảm, các tập dữ liệu này thường chứa các đối tượng có phân bố khác biệt so với phần lớn các đối tượng còn lại, điều này làm giảm hiệu quả của mô hình phân lớp. Khi sử dụng ngưỡng bán kính lớn, các hạt thông tin hàng xóm có trọng số của thuật toán W-MGMA có thể bao gồm các đối tượng này và coi chúng là quan trọng. Ngược lại, các đối tượng này sẽ được xem là có vai trò rất nhỏ bởi thuật toán IARIF-AO và bị loại bỏ hoàn toàn thông qua tập lát cắt α, β của thuật toán IARPD-AO. Thêm vào đó, mức độ không thuộc trong các thuật toán dựa trên tập thô mờ trực cảm có thể làm giảm đáng kể ảnh hưởng của các đối tượng nhiễu trong dữ liệu. Do đó, trong các trường hợp này, rõ ràng các rút gọn thu được từ các thuật toán dựa trên phương pháp tập thô mờ trực cảm mang lại hiệu suất tốt hơn các thuật toán khác.

Theo Bảng 3.6, thuật toán IARIF-AO vượt trội hơn các thuật toán còn lại bằng cách đạt được độ chính xác cao nhất trong 44 trên tổng số 60 trường hợp gia tăng, trong khi AIFSA-FKD, W-MGMA và IARPD-AO lần lượt đạt cao nhất trong 2, 11 và 8 trường hợp. Những kết quả này cho thấy phương pháp IARIF-AO có độ chính xác phân lớp vượt trội đáng kể so với tất cả các phương pháp được so sánh. Phương pháp kiểm định t-test phụ thuộc cũng được thực hiện với mức tin cậy 0.95 để đánh giá sự khác biệt giữa thuật toán đề xuất với các thuật toán AIFSA-FKD, W-MGMA và IARPD-AO. Các giá trị p-values (two-tailed) tương ứng là 1.243E-10, 1.714E-05 và 1.547E-06 đối với bộ phân lớp KNN. Những kết quả này chỉ ra rằng phương pháp đề xuất đạt được độ chính xác phân loại cao hơn đáng kể so với tất cả các phương pháp so sánh.

3.4.2 Hiệu năng của thuật toán IARIF-DO

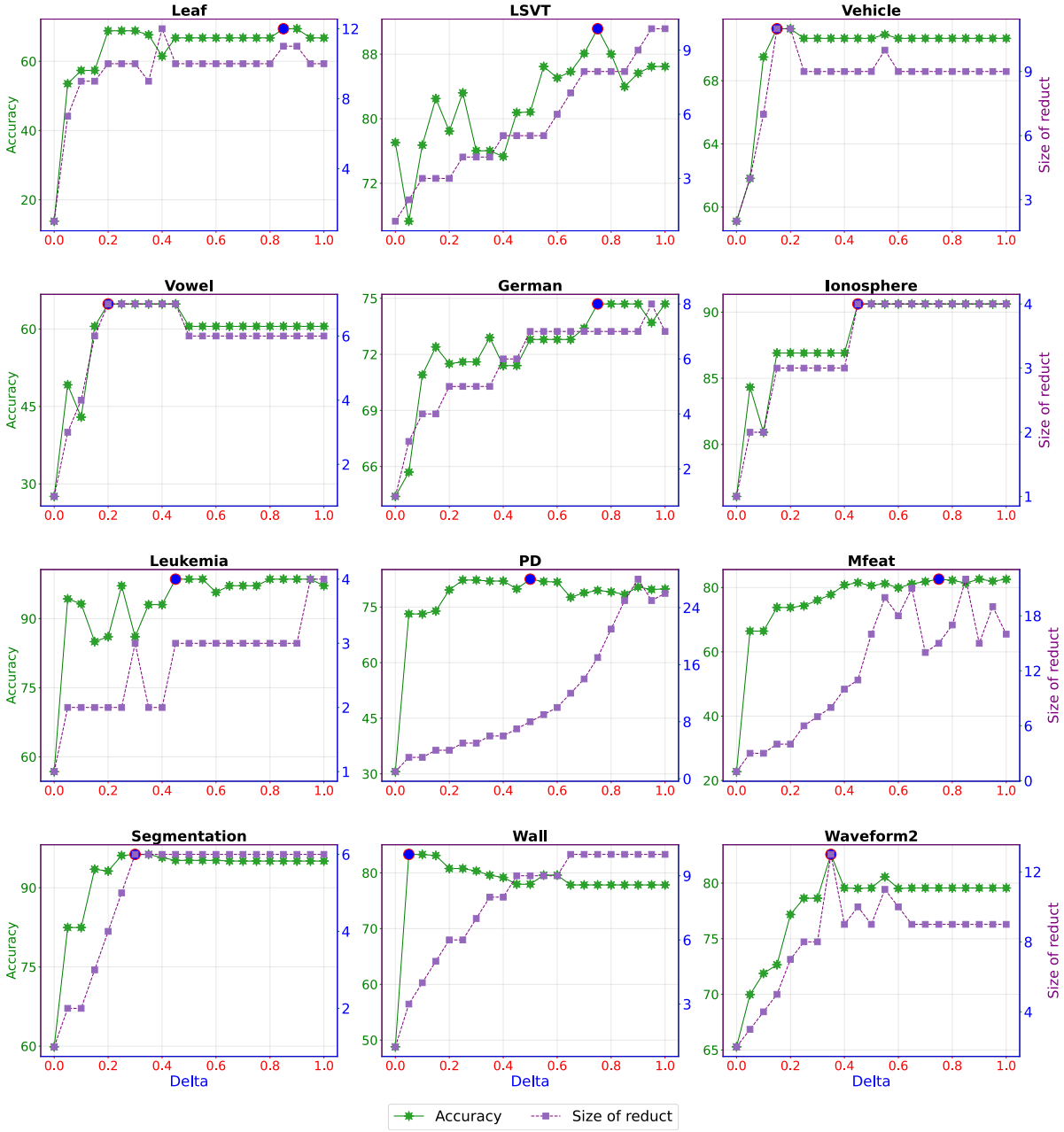
Trong phần này, luận án sẽ tiến hành một số thực nghiệm để đánh giá hiệu quả của thuật toán gia tăng trên các bảng quyết định có sự loại bỏ tập đối tượng. Quá trình thực nghiệm được thực hiện trên các bộ dữ liệu tiêu chuẩn chuẩn được mô tả trong Bảng 3.7.

Bảng 3.7: Các tập dữ liệu thử nghiệm cho IARIF-DO và một số thuật toán

TT	Tập dữ liệu	Số đối tượng	$\|U_{dec}\|$	Số thuộc tính	Số lớp	Nguồn
1	Leaf	340	170	15	30	UCI
2	Ionosphere	351	176	34	2	UCI
3	Vehicle	846	423	18	4	OpenML
4	Vowel	990	495	12	11	OpenML
5	German	1000	500	20	2	UCI
6	LSVT	126	63	310	2	OpenML
7	Leukemia	72	36	7129	2	UCI
8	Mfeat	2000	1000	76	10	OpenML
9	PD	756	378	754	2	UCI
10	Segmentation	2310	1155	19	7	OpenML
11	Wall	5456	2728	24	4	OpenML
12	Waveform2	5000	2500	40	3	OpenML

Đầu tiên, luận án tiến hành so sánh các thuật toán NIFS [82], WNRS [29], W-MGMN [86], FMIFRFS [56], ARPD và ARIFW thông qua việc trích xuất một rút gọn tối ưu trên các bộ dữ liệu gốc. Cụ thể, rút gọn thu được của mỗi thuật toán là một tập con các thuộc tính có kích thước nhỏ nhất và đạt độ chính xác phân lớp cao nhất, được xác định thông qua quá trình điều chỉnh các tham số của mỗi thuật toán trong khoảng từ 0 đến 1 với bước nhảy là 0.05. Đối với thuật toán đề xuất, quá trình duyệt tham số được thể hiện trong Hình 3.5. Rõ ràng, việc điều chỉnh giá trị của tham số δ trong mô hình có tác động rất lớn tới kết quả rút gọn thu được. Khi đó, các giá trị khác nhau của δ có thể làm cho độ chính xác phân lớp cũng như kích thước rút gọn có xu hướng tăng hoặc giảm. Điều này phản ánh rõ nét sự thay đổi của các hạt thông tin lân cận mờ trực cảm có trọng số trong mô hình.

Bảng 3.8 trình bày kích thước các rút gọn thu được và các tham số tương ứng của từng thuật toán. Rõ ràng, các phương pháp đều cho thấy khả năng trích lọc một tập con thuộc tính có kích thước nhỏ hơn rất nhiều so với dữ liệu gốc. Tuy nhiên, trên phần lớn các tập dữ liệu có độ chính xác phân lớp thấp như Ionosphere, Vehicle, Vowel, German, LSVT, Wall và Waveform2, các thuật toán NIFS, WNRS và W-MGMN vẫn thu được các rút gọn có kích thước rất lớn. Trong khi đó, các thuật toán theo tiếp cận mô hình tập thô mờ trực cảm thường trích xuất được các rút gọn nhỏ hơn. Khi xem xét tất cả



Hình 3.5: Độ chính xác phân lớp và kích thước rút gọn khi duyệt δ với bộ phân lớp KNN

các trường hợp, hai thuật toán ARPD và ARIFW thường thu được các tập rút gọn có kích thước nhỏ nhất. Cụ thể, ARPD thu được tập rút gọn nhỏ nhất ở 2 tập dữ liệu và ARIFW thu được tập rút gọn nhỏ nhất ở 7 tập dữ liệu. Đây cũng là hai phương pháp có kích thước rút gọn trung bình thấp nhất trong các thuật toán. Tuy nhiên, kích thước rút gọn trung bình thu được bởi ARIFW là nhỏ hơn so với ARPD. Điều này chỉ ra hiệu quả của thuật toán đề xuất có khả năng xử lý ổn định trên các tập dữ liệu có kích thước lớn hơn.

Bảng 3.9 trình bày thời gian thực thi của các thuật toán rút gọn thuộc tính trên bảng quyết định cố định. Tương tự như kết quả trong Bảng 3.5, các thuật toán NIFS

Bảng 3.8: Kích thước rút gọn và tham số của các thuật toán với KNN trên U

Tập dữ liệu	NIFS		WNRS		W-MGMN		FMIFRFS		ARPD		ARIFW	
	k	$\|red\|$	$delta$	$\|red\|$	$delta$	$\|red\|$	$epsilon$	$\|red\|$	$alpha$	$\|red\|$	$delta$	$\|red\|$
Leaf	0.001	12	0.05	9	0.95	11	0.7	13	0.55	13	0.85	11
Ionosphere	0.001	14	0.05	6	0.1	13	0.9	9	0.55	12	0.45	4
Vehicle	0.001	14	0.1	13	0.2	15	0.65	12	0.1	14	0.15	11
Vowel	0.008	10	0.05	7	0.55	11	0.4	8	0.9	6	0.3	7
German	0.005	10	0.1	19	0.15	14	0.05	15	0.9	8	0.75	7
LSVT	0.002	9	0.8	22	0.45	34	0.65	35	0.75	6	0.75	8
Leukemia	0.001	4	0.9	5	0.5	19	0.7	8	0.6	5	0.45	3
Mfeat	0.002	54	0.45	59	0.05	21	0.55	35	0.45	33	0.75	15
PD	0.003	46	0.45	18	0.15	30	0.3	38	0.35	41	0.5	8
Segmentation	0.004	10	0.05	9	0.85	18	0.95	9	0.85	7	0.3	6
Wall	0.005	13	0.1	20	0.2	15	0.05	7	0.9	4	0.05	3
Waveform2	0.002	27	0.05	9	0.45	15	0.5	17	0.2	25	0.35	13
Trung bình		18.58		16.33		18.00		17.17		14.50		8.00

và WNRS có thời gian thực thi tốt hơn so với các thuật toán dựa trên tập thô mờ trực cảm. Trong khi, W-MGMN có thời gian thực thi lớn nhất do phải trải qua quá trình nhóm các hạt tri thức có cùng ngưỡng trọng số và sử dụng một độ đo phức tạp để đánh giá thuộc tính. Rõ ràng, khi bảng quyết định có nhiều đối tượng, thuật toán đề xuất càng thể hiện được khả năng xử lý vượt trội so với các thuật toán theo hướng tiếp cận tập mờ trực cảm. Cụ thể, thời gian xử lý trung bình của thuật toán là 56.450 giây, trong khi FMIFRFS mất 97.748 giây và ARPD mất 92.472 giây. Do đó, có thể nói rằng thuật toán đề xuất đã chứng minh được khả năng giảm thiểu thời gian so với các thuật toán theo hướng tiếp cận tập thô mờ trực cảm, đặc biệt khi bảng quyết định có số lượng đối tượng lớn.

Bảng 3.9: Thời gian thực thi của ARIFW với một số thuật toán khác trên U.

TT	Tập dữ liệu	NIFS	WNRS	W-MGMN	FMIFRFS	ARPD	ARIFW
1	Leaf	0.0079	0.2311	5.5813	0.6643	0.0756	0.0430
2	Ionosphere	0.0084	0.7941	8.1018	0.2356	0.1619	0.0507
3	Vehicle	0.0248	1.6584	35.647	3.0443	1.7041	1.5401
4	Vowel	0.0186	1.3825	56.821	4.7608	0.7494	0.9378
5	German	0.0204	2.3336	59.884	3.2374	2.8125	2.3473
6	LSVT	0.2605	2.7798	19.387	1.6951	1.1382	1.4905
7	Leukemia	28.880	37.931	92.566	380.93	472.54	372.56
8	Mfeat	16.722	95.189	185.40	136.63	112.73	59.473
9	PD	6.4523	211.73	714.05	211.70	218.54	47.528
10	Segmentation	1.9302	18.528	617.08	28.053	8.2013	7.3485
11	Wall	2.8172	63.196	1328.4	102.22	32.730	27.505
12	Waveform2	3.7391	236.16	1421.8	299.80	258.29	156.58
Trung bình		5.0776	55.993	378.73	97.748	92.472	56.450

Kết quả so sánh độ chính xác phân lớp của các thuật toán khi xử lý trên tập dữ liệu

U được trình bày trong Bảng 3.10. Dễ dàng nhận thấy rằng thuật toán ARIFW được đề xuất đã chọn lọc hiệu quả một tập con thuộc tính quan trọng trên tất cả các tập dữ liệu. Cụ thể, trong toàn bộ 12 trường hợp, tập rút gọn của thuật toán luôn đạt hiệu suất phân lớp vượt trội so với dữ liệu gốc. Hơn nữa, độ chính xác phân loại trung bình của thuật toán đề xuất khi sử dụng bộ phân loại KNN là 82.33%, cao hơn rõ rệt so với giá trị của dữ liệu gốc là 76.69%. Luận án tiếp tục so sánh hiệu suất phân lớp giữa các thuật toán. Trên các tập dữ liệu có độ chính xác phân lớp ban đầu thấp và số lượng đối tượng lớn, chẳng hạn như German, Wall và Waveform2, các thuật toán dựa trên tập thô mờ trực cảm luôn thu được các rút gọn có hiệu quả phân lớp vượt trội so với các phương pháp dựa trên tập thô mờ và tập thô lân cận có trọng số. Điều này làm nổi bật lợi thế của tập thô mờ trực cảm, trong đó việc kết hợp hàm không thuộc giúp mô hình giảm thiểu ảnh hưởng của các đối tượng nhiễu trong tập dữ liệu.

Bảng 3.10: So sánh độ chính xác phân lớp của ARIFW với các thuật toán khác trên U

Tập dữ liệu	Tập gốc	NIFS	WNRS	W-MGMN	FMIFRFS	ARPD	ARIFW
Leaf	60.59 ± 6.34	59.41 ± 5.55	69.12 ± 3.01	67.94 ± 4.64	61.47 ± 5.33	61.47 ± 6.76	**69.41 ± 4.78**
Ionosphere	83.76 ± 6.39	84.89 ± 6.02	88.90 ± 6.89	86.04 ± 5.34	90.33 ± 5.27	88.32 ± 7.61	**90.62 ± 6.08**
Vehicle	70.45 ± 1.85	69.39 ± 3.04	69.86 ± 2.02	70.57 ± 2.75	70.10 ± 2.91	70.22 ± 2.83	**71.29 ± 3.19**
Vowel	51.52 ± 12.38	53.03 ± 11.65	**64.85 ± 5.8**	52.02 ± 11.3	53.54 ± 6.94	52.12 ± 6.94	**64.85 ± 5.8**
German	73.00 ± 2.79	69.40 ± 4.43	69.60 ± 3.9	69.10 ± 2.88	73.50 ± 2.94	71.50 ± 2.77	**74.70 ± 2.97**
LSVT	83.97 ± 14.31	80.06 ± 9.07	89.68 ± 6.39	86.47 ± 5.07	86.90 ± 9.95	90.58 ± 7.5	**91.15 ± 9.39**
Leukemia	84.46 ± 13.51	90.36 ± 6.35	**98.57 ± 10**	91.43 ± 9.48	**98.57 ± 4.29**	**98.57 ± 4.29**	**98.57 ± 4.29**
Mfeat	80.70 ± 1.36	81.50 ± 1.99	80.15 ± 1.52	82.15 ± 1.64	81.65 ± 1.99	**83.15 ± 1.69**	82.60 ± 1.84
PD	79.22 ± 6.07	74.18 ± 7.02	**84.25 ± 5.89**	78.44 ± 4.34	79.63 ± 4.58	79.76 ± 3.61	82.52 ± 3.98
Segmentation	95.28 ± 1.52	95.41 ± 1.5	95.71 ± 1.88	95.19 ± 1.32	95.63 ± 1.67	95.71 ± 1.53	**96.32 ± 1.44**
Wall	77.26 ± 5.91	77.27 ± 6.36	78.41 ± 5.35	75.81 ± 5.88	81.56 ± 3.78	83.14 ± 4.01	**83.32 ± 4.56**
Waveform2	80.08 ± 1.46	80.44 ± 1.51	78.38 ± 1.69	**83.24 ± 1.18**	82.86 ± 1.44	81.82 ± 1.54	82.58 ± 1.06
Trung bình	76.69 ± 6.16	76.28 ± 5.37	80.62 ± 4.53	79.06 ± 7.87	78.20 ± 4.65	79.70 ± 4.26	**82.33 ± 4.12**
Win/Draw/Loss	12/0/0	12/0/0	9/2/1	11/0/1	10/1/1	10/1/1	*best*

Từ Bảng 3.10, tỷ lệ Thắng/Hòa/Thua (Win/Draw/Loss) giữa thuật toán đề xuất ARIFW và các thuật toán khác lần lượt là: 12/0/0, 9/2/1, 11/0/1, 10/1/1 và 10/1/1. Có thể thấy rằng trên hầu hết các tập dữ liệu, các tập rút gọn được tạo ra bởi thuật toán ARIFW mang lại kết quả phân loại tốt hơn so với các thuật toán khác, mặc dù số lượng thuộc tính được chọn bởi ARIFW ít hơn. Hơn nữa, phương pháp đề xuất cũng đạt độ chính xác phân lớp cao nhất trong 9 trên 12 trường hợp, điều này chứng minh rằng thuật toán đề xuất có khả năng lựa chọn hiệu quả các thuộc tính quan trọng và có khả năng xử lý trên các tập dữ liệu với đặc trưng đa dạng.

Cuối cùng, luận án trình bày một số so sánh về hiệu quả xử lý của thuật toán IARIF-

DO với các thuật toán gia tăng AIFSD-FKD [82], W-MGMD [86] và IARPD-RO. Trong đó, AIFSD-FKD là thuật toán gia tăng dựa trên tập thô mờ sử dụng độ đo khoảng cách tri thức mờ với đầu vào là tập rút gọn từ thuật toán NIFS. Thuật toán gia tăng W-MGMD dựa trên tập thô lân cận có trọng số, sử dụng độ đo entropy, với đầu vào là tập rút gọn từ thuật toán W-MGMN. Trong khi đó, thuật toán IARPD-RO dựa trên tập thô mờ trực cảm, sử dụng độ đo khoảng cách phân hoạch lấy rút gọn từ thuật toán ARPD làm đầu vào và thuật toán đề xuất IARIF-DO sử dụng tập rút gọn từ thuật toán IARFW làm đầu vào. Các thuật toán sẽ xử lý trên tập dữ liệu bằng cách loại bỏ lần lượt các tập đối tượng từ U_1 đến U_5 từ U.

Bảng 3.11: So sánh thời gian thực thi của IARIF-DO với các thuật toán gia tăng

Thuật toán	Dữ liệu	Tập dữ liệu loại bỏ					Dữ liệu	Tập dữ liệu loại bỏ				
		U_1	U_2	U_3	U_4	U_5		U_1	U_2	U_3	U_4	U_5
AIFSD-FKD	Leaf	0.0014	0.0012	0.0011	0.0009	0.0019	Leukemia	0.0003	0.0003	0.0002	0.0001	0.0001
W-MGMD		0.0936	0.0727	0.0542	0.0410	0.0261		0.0137	0.0090	0.0084	0.0078	0.0066
IARPD-RO		0.0199	0.0019	0.0018	0.0017	0.0015		0.0016	0.0008	0.0007	0.0007	0.0005
IARIF-DO		0.0017	0.0015	0.0013	0.0010	0.0148		0.0010	0.0007	0.0006	0.0004	0.0003
AIFSD-FKD	Ionosphere	0.0045	0.0024	0.0012	0.0011	0.0010	Mfeat	0.3399	0.3125	0.2973	0.2854	0.2722
W-MGMD		0.1092	0.0647	0.0481	0.0354	0.0243		7.6094	5.9371	4.5768	3.6133	2.7868
IARPD-RO		0.0193	0.0017	0.0016	0.0015	0.0158		0.1167	0.1089	0.1048	3.1072	0.0976
IARIF-DO		0.0055	0.0031	0.0011	0.0009	0.0008		0.0513	0.0479	0.1003	0.0932	0.0868
AIFSD-FKD	Vehicle	0.0027	0.0035	0.0024	0.0022	0.0031	PD	0.0173	0.0137	0.0114	0.0009	0.0005
W-MGMD		0.8093	0.5392	0.3373	0.2365	0.1610		1.1643	0.7671	0.4681	0.2533	0.1237
IARPD-RO		0.0057	0.0054	0.0051	0.0047	0.0308		3.6784	3.5140	3.3552	3.1538	3.1100
IARIF-DO		0.0201	0.0186	0.0163	0.0144	0.0125		0.0626	0.0604	0.0584	0.0562	0.0531
AIFSD-FKD	Vowel	0.0020	0.0019	0.0017	0.0028	0.0028	Segmentation	0.0068	0.0055	0.0041	0.0035	0.0139
W-MGMD		0.7238	0.4566	0.2463	0.1781	0.1214		7.1595	4.9197	3.2126	2.3421	1.6481
IARPD-RO		0.0021	0.0075	0.0017	0.0015	0.0013		0.0054	0.0048	0.0158	0.0033	0.0027
IARIF-DO		0.0032	0.0030	0.0188	0.0156	0.0135		0.0050	0.0043	0.0358	0.0029	0.0024
AIFSD-FKD	German	0.0032	0.0029	0.0025	0.0022	0.0020	Wall	0.1345	0.1242	0.1198	0.1325	0.1347
W-MGMD		0.8169	0.6301	0.4103	0.3124	0.2160		38.134	29.808	22.805	16.854	11.612
IARPD-RO		0.0028	0.0025	0.0023	0.0020	0.0017		1.4515	1.0124	0.0103	0.0275	0.0151
IARIF-DO		0.0213	0.0176	0.0019	0.0017	0.0015		0.0259	0.0218	0.0185	0.0162	0.0134
AIFSD-FKD	LSVT	0.0004	0.0004	0.0003	0.0014	0.0012	Waveform2	0.2132	0.1928	0.1739	0.1890	0.1306
W-MGMD		0.0446	0.0233	0.0157	0.0091	0.0088		29.932	21.917	15.427	10.415	6.5485
IARPD-RO		0.0028	0.0026	0.0023	0.0014	0.0013		27.452	25.898	24.204	22.682	21.147
IARIF-DO		0.0032	0.0017	0.0026	0.0016	0.0014		16.207	13.301	10.384	12.628	4.9480

Bảng 3.11 trình bày thời gian thực thi của thuật toán IARIF-DO so với các thuật toán AIFSD-FKD, W-MGMD và IARPD-RO. Mặc dù phải thực hiện tính toán trên các tập dữ liệu có số lượng đối tượng lớn hơn, tuy nhiên các thuật toán gia tăng vẫn chứng minh được khả năng vượt trội về thời gian thực thi so với các thuật toán tương ứng của chúng khi xử lý trên bảng quyết định cố định. Thuật toán AIFSD-FKD vẫn chứng minh được khả năng xử lý vượt trội so với các thuật toán khác. Khi xem xét các thuật toán dựa trên tập thô mờ trực cảm, có thể thấy thuật toán đề xuất có thời gian xử lý nhanh hơn trong hầu hết các bước gia tăng. Điều này là do đặc điểm của quan hệ lân cận có trọng số, giúp thu hẹp không gian tính toán bằng cách chỉ tập trung vào các đối

tượng nằm trong bán kính. Thêm vào đó, thuật toán này cũng chọn lọc được ít thuộc tính hơn. Do đó, thuật toán kết thúc sớm và dễ dàng thu được một rút gọn.

Cuối cùng, luận án trình bày kết về quả độ chính xác phân lớp và kích thước rút gọn thu được từ các thuật toán gia tăng. Qua các số liệu trong Bảng 3.12, có thể thấy rằng các thuật toán dựa trên mô hình tập thô mờ trực cảm thường thu được các rút gọn có kích thước nhỏ hơn so với các thuật toán dựa trên tập thô lân cận có trọng số và tập thô mờ trong hầu hết các bước loại bỏ. Đáng chú ý hơn, với các tập dữ liệu có độ chính xác phân lớp ban đầu thấp như Leaf, Ionosphere, Vehicle, Vowel, Wall và Waveform2, số lượng thuộc tính được trả về bởi IARPD-RO và IARIF-DO nhỏ hơn đáng kể. Cụ thể, với 60 trường hợp loại bỏ, các rút gọn thu được bởi IARPD-RO và IARIF-DO có kích thước nhỏ nhất lần lượt trong 18 và 45 trường hợp. Điều này nhấn mạnh hiệu quả của phương pháp đề xuất trong việc loại bỏ các thuộc tính dư thừa khỏi bảng quyết định trong các bước gia tăng, đặc biệt trong các trường hợp bảng quyết định chứa các thuộc tính làm suy giảm hiệu quả phân lớp. Để đánh giá toàn diện hơn hiệu quả của các rút gọn thu được, luận án tiếp tục phân tích hiệu quả phân lớp dựa trên các rút gọn này.

Độ chính xác phân lớp của các rút gọn thu được từ các phương pháp được trình bày trong Bảng 3.12. Trong phần lớn các bước loại bỏ, các rút gọn xấp xỉ được trả về từ các thuật toán thường có độ chính xác cao hơn hoặc tương đương so với dữ liệu gốc. Các thuật toán dựa trên mô hình tập thô mờ trực cảm vẫn chứng minh được hiệu quả vượt trội so với các thuật toán khác trên các tập dữ liệu có độ chính xác phân lớp ban đầu thấp cũng như có số lượng đối tượng lớn, chẳng hạn như Wall và Waveform2.

Theo Bảng 3.12, thuật toán đề xuất vượt trội hơn các thuật toán khác khi đạt được độ chính xác cao nhất trong 47 trên tổng số 60 trường hợp loại bỏ, trong khi AIFSA-FKD, W-MGMA và IARPD-AO lần lượt đạt cao nhất trong 1, 6 và 8 trường hợp. Phương pháp kiểm định t-test phụ thuộc cũng được thực hiện với mức tin cậy 0.95 để đánh giá sự khác biệt giữa thuật toán đề xuất với các thuật toán AIFSD-FKD, W-MGMD và IARPD-RO. Các giá trị p-values (two-tailed) tương ứng là 1.371E-09, 1.253E-11 và 3.332E-06 đối với bộ phân lớp KNN.

Những kết quả được trình bày đã cho thấy phương pháp đề xuất đạt được độ chính xác phân loại cao hơn đáng kể so với tất cả các phương pháp so sánh.

Bảng 3.12: Kích thước rút gọn, độ chính xác phân lớp của IARIF-DO và các thuật toán gia tăng

TT	Tập loại bỏ	Tập gốc	AIFSD-FKD		W-MGMD		IARPD-RO		IARIF-DO	
			$\|\mathcal{B}^-\|$	*Accuracy*	$\|\mathcal{B}^-\|$	*Accuracy*	$\|\mathcal{B}^-\|$	*Accuracy*	$\|\mathcal{B}^-\|$	*Accuracy*
1	$\|U_1\| = 306$	58.78±7.39	12	58.15±5.55	**11**	58.15±5.2	12	59.46±6.28	**11**	**67.65±4.52**
	$\|U_2\| = 272$	60.75±11.77	12	60.03±9.75	**11**	59.18±4.79	12	62.22±10.6	**11**	**69.13±4**
	$\|U_3\| = 238$	64.33±11.77	12	63.51±11.55	**11**	66.85±6.95	12	67.68±9.5	**11**	**74.82±7.67**
	$\|U_4\| = 204$	68.31±10.89	12	67.38±13.08	**11**	74.12±8.81	12	72.69±11.12	**11**	**79.90 ± 7.83**
	$\|U_5\| = 170$	71.76±10.12	11	80.59±8.74	11	75.29±8.65	12	75.88±10.34	**10**	**77.65±9.41**
2	$\|U_1\| = 316$	82.98±5.68	14	84.22±7.98	11	86.13±6.85	11	87.73±6.6	**3**	**89.93±4.78**
	$\|U_2\| = 281$	80.79±6.98	12	84.33±9.35	9	85.75±7.34	11	86.12±9.1	**2**	**87.91±7.34**
	$\|U_3\| = 246$	79.37±9.52	12	83.85±8.11	9	83.00±10.09	11	**86.33±11.18**	**2**	86.28±9.42
	$\|U_4\| = 211$	77.73±10.86	12	81.02±10.68	9	79.09±10.31	11	83.40±13.69	**2**	**84.81±8.5**
	$\|U_5\| = 176$	77.29±9.74	12	79.44±9.84	9	80.49±9.29	10	80.10±12.26	**2**	**81.21±6.84**
3	$\|U_1\| = 762$	69.03±2.65	14	**70.21±2.41**	13	68.90±2.02	14	69.42±3.31	**10**	69.69±2.6
	$\|U_2\| = 678$	67.85±5.75	13	66.07±4.1	11	65.93±5.19	14	67.84±5.25	**9**	**70.35±4.87**
	$\|U_3\| = 594$	68.70±5.16	13	66.84±5.04	9	69.89±4.86	14	68.69±5.08	**8**	**72.24±6.22**
	$\|U_4\| = 510$	67.84±4.04	13	70.00±4.02	9	70.98±5.17	14	68.04±5.82	**7**	**73.73±4.57**
	$\|U_5\| = 426$	67.61±4.36	12	67.34±4.08	9	70.19±6.98	13	66.42±5.93	**6**	**70.90±4.97**
4	$\|U_1\| = 891$	40.62±11.90	10	42.30±12.71	9	60.29±9.69	**6**	39.39±8.72	7	**64.32±8.22**
	$\|U_2\| = 792$	38.12±6.99	10	37.88±6.13	7	53.55±8.89	**5**	36.11±7.09	7	**64.77±7.29**
	$\|U_3\| = 693$	45.81±14.38	10	45.37±13.26	**5**	46.67±13.55	**5**	39.46±14.22	6	**60.79±10.71**
	$\|U_4\| = 594$	35.86±14.33	9	63.30±8.67	**5**	43.16±14.52	**5**	35.55±13.54	**5**	**59.44±7.4**
	$\|U_5\| = 495$	48.81±12.36	8	60.59±10.76	5	51.72±8.87	5	49.43±14.71	**4**	**60.34±12.87**
5	$\|U_1\| = 900$	72.67±3.15	10	70.44±5.22	13	69.22±4.5	8	72.33±3.93	**6**	**73.00±3.62**
	$\|U_2\| = 800$	71.12±4.27	10	69.75±2.95	12	66.00±4.34	8	68.62±3.56	**5**	**72.25±4.77**
	$\|U_3\| = 700$	72.57±4.46	10	68.71±5.01	11	67.00±4.01	8	71.00±3.44	**5**	**71.14±4.55**
	$\|U_4\| = 600$	71.17±4.89	10	68.33±3.5	11	65.50±4.41	8	71.83±5.02	**5**	**72.00±5.26**
	$\|U_5\| = 500$	73.20±4.21	10	71.20±4.49	11	68.00±5.14	8	**72.80±3.12**	**5**	72.00±5.93
6	$\|U_1\| = 114$	85.76±12.47	9	73.48±15.78	29	83.33±11.45	**5**	87.73±5.88	7	**90.23±8.45**
	$\|U_2\| = 102$	82.27±12.53	9	75.45±14.32	24	81.36±6.73	**4**	83.27±7.84	7	**88.00±8.72**
	$\|U_3\| = 90$	84.44±8.89	9	78.89±7.78	19	85.56±10	**3**	83.33±10.24	6	**88.89±8.61**
	$\|U_4\| = 78$	81.07±14.99	8	81.07±11.44	15	83.39±12.48	**3**	76.96±7.47	6	**84.46±9.75**
	$\|U_5\| = 66$	75.95±14.42	7	72.14±13	15	86.43±12	**3**	81.43±11.7	6	**89.05±7.24**
7	$\|U_1\| = 65$	79.76±13.05	4	91.19±7.23	13	93.81±7.62	4	**98.57±4.29**	**3**	**98.57±4.29**
	$\|U_2\| = 58$	84.33±11.84	4	88.00±10.87	10	94.67±8.19	**3**	91.67±8.33	**3**	**96.67±6.67**
	$\|U_3\| = 51$	90.33±9.71	4	88.33±13.1	9	**96.00±8**	**3**	90.33±9.71	**3**	**96.00±8**
	$\|U_4\| = 44$	82.50±15.33	4	88.5±11.63	8	**95.50±9.07**	**3**	91.00±11.14	**3**	93.00±10.77
	$\|U_5\| = 37$	86.67±13.54	4	83.33±17.87	8	92.50±11.46	**3**	91.67±12.91	**3**	**96.67±10**
8	$\|U_1\| = 1800$	89.78±2.11	53	90.50±2.1	20	91.50±1.49	33	**92.11±1.77**	**15**	92.00±2.47
	$\|U_2\| = 1600$	88.88±2.22	52	89.81±2.5	19	90.00±1.79	33	91.50±1.94	**15**	**91.56±2.41**
	$\|U_3\| = 1400$	90.64±1.87	51	91.79±1.64	18	91.07±2.54	33	**93.21±1.51**	**14**	92.79±1.87
	$\|U_4\| = 1200$	91.00±1.93	50	92.17±2.01	18	90.75±2.72	31	**92.83±1.45**	**13**	92.50±2.01
	$\|U_5\| = 1000$	92.50±1.8	49	92.50±2.2	18	91.50±3.2	31	**93.80±1.17**	**12**	92.60±2.65
9	$\|U_1\| = 681$	80.19±6.83	46	77.68±5.6	26	79.15±5.34	40	80.33±5.05	**8**	**83.70±3.87**
	$\|U_2\| = 606$	80.22±6.92	46	80.17±5.24	22	78.21±7.3	39	82.37±6.45	**8**	**83.17±4.3**
	$\|U_3\| = 531$	80.06±6.06	46	80.80±6.6	18	77.77±6.14	38	78.74±8.12	**8**	**83.44±4.71**
	$\|U_4\| = 456$	81.61±7.38	46	82.92±5.97	14	80.74±4.09	37	81.20±7.89	**8**	**86.86±4.54**
	$\|U_5\| = 381$	78.52±7.88	46	77.98±6.24	10	78.50±4.35	36	78.78±7.48	**8**	**83.22±4.01**
10	$\|U_1\| = 2079$	95.05±1.49	10	95.29±1.25	16	94.95±1.4	7	95.53±1.65	**6**	**95.96±1.46**
	$\|U_2\| = 1848$	94.91±2.11	10	94.86±1.85	14	93.45±2.07	7	95.40±2.23	**6**	**95.89±1.95**
	$\|U_3\| = 1617$	94.56±2.28	10	94.38±2.25	12	92.46±2.15	6	94.25±2.11	**5**	**95.18±2.02**
	$\|U_4\| = 1386$	94.88±2.3	10	94.67±2.63	12	91.49±2.31	6	94.02±2.65	**5**	**94.52±2.59**
	$\|U_5\| = 1155$	93.77±2.45	9	93.00±3.08	12	90.74±1.92	6	92.30±1.34	**5**	**93.25±2.3**
11	$\|U_1\| = 4911$	76.50±7.02	13	76.52±6.32	15	74.10±6.63	**3**	**84.95±2.97**	**3**	83.02±4.21
	$\|U_2\| = 4366$	75.52 ± 6.51	13	74.63±6.49	15	72.17±5.98	**2**	77.35±3.4	3	**80.17±5.12**
	$\|U_3\| = 3821$	72.84 ± 8.24	13	72.42±6.63	15	70.30±7.11	**2**	76.45±3.62	3	**80.84±4.57**
	$\|U_4\| = 3276$	69.29 ± 4.42	12	69.26±4.15	15	67.19±4.39	**2**	76.56±5.42	3	**81.50±4.85**
	$\|U_5\| = 2731$	63.31 ± 9.56	11	64.41±9.04	15	61.70±9.85	**2**	73.31±6.9	3	**76.64±5.18**
12	$\|U_1\| = 4500$	79.80±1.63	27	80.00±1.72	14	**83.00±1.6**	25	82.02±1.52	**13**	81.71±1.36
	$\|U_2\| = 4000$	79.52±2.01	27	79.72±1.59	**13**	**82.80±2.61**	25	81.32±1.71	**13**	81.68±1.64
	$\|U_3\| = 3500$	78.29±2.26	27	78.69±1.14	**12**	**81.54±1.54**	25	81.29±1.61	13	81.29±1.66
	$\|U_4\| = 3000$	78.53±1.84	25	78.87±1.77	**11**	**80.67±1.66**	25	80.13±1.72	12	79.73±2.01
	$\|U_5\| = 2500$	78.20±2.02	25	77.28±2.1	**10**	78.92±1.13	25	79.20±1.71	12	**80.64±1.89**

3.5 Kết luận Chương 3

Trong chương này, luận án đã giới thiệu một dạng hạt mới, được hình thành dựa trên trọng số của các thuộc tính điều kiện và các đối tượng trong hạt thông tin. Từ cơ sở đó, mô hình tập thô lân cận mờ trực cảm có trọng số đã được đề xuất cùng với một số tính chất quan trọng. Kế thừa ưu điểm của mô hình tập thô lân cận, mô hình mới giúp rút ngắn đáng kể thời gian tính toán trong các phương pháp rút gọn thuộc tính khi chỉ tập trung vào các đối tượng bên trong hạt thông tin. Đồng thời, nhờ khả năng đánh giá chi tiết mức độ ảnh hưởng của từng thuộc tính điều kiện đến khả năng phân lớp của mỗi đối tượng thông qua trọng số, mô hình cho phép lựa chọn hiệu quả các thuộc tính có ý nghĩa cao. Hơn nữa, việc thiết lập trọng số cho các đối tượng trong cùng một hạt thông tin dựa trên hai thành phần đã tạo nên một cấu trúc hạt đặc trưng và chi tiết hơn so với nhiều mô hình tập thô lân cận mở rộng khác. Nhờ đó, các thuật toán rút gọn thuộc tính được hỗ trợ tốt hơn trong quá trình tìm kiếm một rút gọn tối ưu.

Với đặc trưng của một tập mờ trực cảm, mô hình đề xuất cũng giảm thiểu ảnh hưởng của các đối tượng nhiễu trong dữ liệu. Từ mô hình này, luận án phát triển độ đo khoảng cách họ lân cận mờ trực cảm có trọng số và thiết kế một thuật toán giảm thuộc tính cho các bảng quyết định cố định. Để giải quyết các kịch bản dữ liệu thực tế, luận án xây dựng các cơ chế cập nhật hạt thông tin và đề xuất hai thuật toán gia tăng để xử lý nhanh trên các bảng quyết định động. Kết quả thực nghiệm cho thấy các phương pháp được đề xuất đạt độ chính xác phân lớp vượt trội trên các bộ dữ liệu nhiễu và không nhất quán so với các phương pháp dựa trên tập thô mờ và tập thô lân cận có trọng số. Đối với các bộ dữ liệu này, hiệu quả của các mô hình phân lớp truyền thống thường bị suy giảm do sự tồn tại của một số đối tượng có phân bố khác biệt đáng kể so với phần lớn các đối tượng trong tập vũ trụ. Bên cạnh đó, các thuật toán đề xuất còn cho thấy khả năng giảm thiểu đáng kể thời gian xử lý so với các phương pháp dựa trên tập thô mờ trực cảm.

Rõ ràng, mô hình tập thô lân cận mờ trực cảm có trọng số đã khắc phục được nhiều hạn chế còn tồn tại trong một số mô hình thuộc nhánh mở rộng thứ nhất của lý thuyết tập thô. Trên cơ sở đó, mô hình được xem như một công cụ hiệu quả để giải quyết bài toán rút gọn thuộc tính trên nhiều loại dữ liệu khác nhau, đặc biệt là các tập dữ liệu nhiễu và có mật độ phân bố không đồng đều. Tuy nhiên, trong trường hợp dữ liệu có mật độ phân bố quá dày đặc, số lượng đối tượng thuộc các hạt thông tin có thể gia tăng đáng kể, từ đó làm tăng chi phí tính toán và thời gian xử lý của các thuật toán.

KẾT LUẬN VÀ KIẾN NGHỊ

KẾT LUẬN

Luận án tập trung nghiên cứu bài toán rút gọn thuộc tính trên bảng quyết định với mục tiêu nâng cao hiệu quả cho các mô hình học máy, đồng thời giảm thiểu tính phức tạp của tập luật. Lý thuyết tập thô được xem là một công cụ nền tảng quan trọng cho sự phát triển của các phương pháp rút gọn thuộc tính. Tuy vậy, các phương pháp dựa trên mô hình tập thô truyền thống vẫn còn nhiều hạn chế khi áp dụng cho các bảng quyết định dạng số và liên tục. Thực tiễn này đã dẫn đến sự hình thành hai hướng nghiên cứu mở rộng nhằm khắc phục những bất cập trên. Trên cơ sở khảo sát, phân tích và đối chiếu, luận án đã chỉ ra một số nhược điểm của các phương pháp mở rộng hiện có, đồng thời đề xuất một số đóng góp chính như sau:

Thứ nhất, trước những hạn chế của hướng mở rộng tập thô mờ trong việc xử lý dữ liệu nhiễu cũng như sự kém hiệu quả về thời gian của mô hình tập thô mờ trực cảm, luận án đã hướng tới việc đề xuất mô hình tập mờ trực cảm mức *alpha, beta* như một sự tổng quát hóa của tập mờ trực cảm. Qua đó, một số tính chất then chốt của mô hình cũng được trình bày để làm rõ những ưu điểm mà mô hình mang lại. Trên cơ sở này, luận án đưa ra một số thuật toán rút gọn thuộc tính với những đóng góp chính như sau:

- Xây dựng độ đo khoảng cách phân hoạch mờ trực cảm mức *alpha, beta* làm cơ sở trong việc định nghĩa lại một rút gọn hiệu quả trên bảng quyết định và xây dựng độ quan trọng của thuộc tính nhằm lựa chọn các thuộc tính có ý nghĩa cao. Qua đó, đề xuất thuật toán rút gọn thuộc tính trên bảng quyết định cố định (ARPD) theo hướng tiếp cận lọc với độ phức tạp đa thức.

- Mở rộng công thức tính khoảng cách phân hoạch mờ trực cảm mức *alpha, beta* để xử lý nhanh cho các trường hợp bảng quyết định có sự bổ sung và loại bỏ tập đối tượng, hướng tới việc đề xuất hai thuật toán gia tăng trên bảng quyết định thay đổi, ứng dụng cho các kịch bản thực tế của dữ liệu.

Thứ hai, nhằm khắc phục những hạn chế từ một số mở rộng theo nhánh tập thô lân cận, luận án đã đề xuất mô hình tập thô lân cận mờ trực cảm có trọng số nhằm đánh giá ảnh hưởng của các thuộc tính điều kiện đến quyết định của các đối tượng và đặc trưng chi tiết vai trò của các đối tượng trong một hạt thông tin. Từ mô hình này, luận án cũng đưa ra một số thuật toán rút gọn thuộc tính với những đóng góp nổi bật sau:

- Xây dựng độ đo khoảng cách giữa hai họ lân cận mờ trực cảm có trọng số làm cơ

sở trong việc định nghĩa một rút gọn hiệu quả trên bảng quyết định và xây dựng độ quan trọng của thuộc tính. Qua đó, đề xuất một thuật toán trích chọn các thuộc tinh cần thiết trên bảng quyết định cố định (ARIFW) với thời gian đa thức.

- Mở rộng công thức tính khoảng cách giữa hai họ lân cận mờ trực cảm có trọng số để xử lý cho các trường hợp bảng quyết định có sự bổ sung và loại bỏ tập đối tượng hướng tới việc đề xuất hai thuật toán gia tăng trên bảng quyết định thay đổi tập đối tượng.

KIẾN NGHỊ

Phạm vi của luận án hiện tại chủ yếu tập trung vào các bảng quyết định thay đổi trong trường hợp bổ sung và loại bỏ tập đối tượng. Trong tương lai, hướng nghiên cứu sẽ tiếp tục được mở rộng sang các phương pháp rút gọn thuộc tính đối với những bảng có sự thay đổi ở tập thuộc tính. Đây là một kịch bản phổ biến trong thực tế, khi dữ liệu luôn được cập nhật liên tục theo thời gian.

Thứ hai, trong các mô hình được đề xuất, các lớp quyết định về bản chất vẫn được biểu diễn dưới dạng tập rõ. Điều này khiến việc đo lường lượng thông tin của từng lớp gặp nhiều hạn chế. Để khắc phục vấn đề này, luận án trong tương lai sẽ tập trung phát triển một mô hình cho phép chuyển đổi các đối tượng trong lớp quyết định thành các số mờ trực cảm. Cách tiếp cận này hứa hẹn mang lại các độ đo linh hoạt và hiệu quả hơn trong việc tìm kiếm các thuộc tính quan trọng.

Thứ ba, trong mô hình tập thô lân cận mờ trực cảm có trọng số, việc sử dụng một ngưỡng bán kính cố định đôi khi gặp khó khăn trong bối cảnh dữ liệu có mật độ phân bố lớn. Khi đó, một hạt thông tin có thể bao hàm quá nhiều đối tượng, gây ảnh hưởng đến kết quả rút gọn. Để khắc phục, trong giai đoạn nghiên cứu tiếp theo, luận án sẽ hướng đến việc đề xuất một mô hình mới nhằm hạn chế số lượng đối tượng có vai trò lớn trong từng hạt thông tin, qua đó tạo cơ sở cho việc xây dựng các thuật toán rút gọn thuộc tính hiệu quả hơn.

DANH MỤC CÔNG TRÌNH CÔNG BỐ LIÊN QUAN ĐẾN LUẬN ÁN

A. Các công trình đã công bố

[CT1] **Pham Viet Anh**, Nguyen Ngoc Thuy, Vu Duc Thi, Nguyen Long Giang, "On distance-based attribute reduction with α, β-level intuitionistic fuzzy sets", IEEE Access, vol. 11, pp. 138095–138107, 2023. (SCIE Q1 IF 3.4).

[CT2] **Pham Viet Anh**, Nguyen Ngoc Thuy, Le Hoang Son, Tran Hung Cuong, Nguyen Long Giang, "Incremental attribute reduction with α, β-level intuitionistic fuzzy sets", International Journal of Approximate Reasoning, vol. 176, p. 109326, 2025. (SCIE Q1 IF 3.2).

[CT3] Nguyen Long Giang, **Pham Viet Anh**, Janos Demetrovics, Vu Duc Thi, "Attribute reduction based on rough set theory and its extensions: A review", Journal of Computer Science and Cybernetics, vol. 41, no. 3, pp. 105-123, 2025.

[CT4] **Phạm Việt Anh**, Nguyễn Long Giang, Nguyễn Ngọc Thủy, Nguyễn Thế Thủy, and Phạm Đình Khánh, "Về một thuật toán gia tăng tìm tập rút gọn trên bảng quyết định khi loại bỏ tập đối tượng", Các công trình nghiên cứu và phát triển CNTT và truyền thông, Hà Nội, số 2, tr. 58-65, 2023.

[CT5] **Phạm Việt Anh**, Nguyễn Long Giang, Nguyễn Ngọc Thủy, Cao Chính Nghĩa, Vũ Đức Thi, "Rút gọn thuộc tính dựa trên mô hình tập thô mờ trực cảm sử dụng độ đo khoảng cách mở rộng và lát cắt α", Kỷ yếu Hội nghị Khoa học Công nghệ Quốc Gia lần thứ XV: Nghiên cứu cơ bản và ứng dụng công nghệ thông tin, Hà Nội, 11/2022, tr. 320-331, 2023.

[CT6] **Pham Viet Anh**, Nguyen Long Giang, Nguyen Ngoc Thuy, Le Van Dung, Phung Hong Quan, "Hybrid filter-wrapper attribute reduction method with the uncertainty classification degree", in Advances in Data Science and Optimization of Complex Systems - Proceedings of the International Conference on Applied Mathematics and Computer Science - ICAMCS, vol. 1569, pp. 298-309, 2025.

B. Các công trình chờ phản biện

[CT7] **Pham Viet Anh**, Nguyen Ngoc Thuy, Nguyen Long Giang, "Incremental attribute reduction on dynamic decision tables with intuitionistic fuzzy weighted neighborhood rough sets", Đang chờ phản biện.

TÀI LIỆU THAM KHẢO

[1] Nguyễn Văn Thiện. *Một số phương pháp kết hợp trong rút gọn thuộc tính theo tiếp cận tập thô mờ*, Luận án Tiến sĩ Máy tính, Học viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam, 2018.

[2] Hồ Thị Phượng. *Phương pháp gia tăng rút gọn thuộc tính trong bảng quyết định thay đổi theo tiếp cận tập thô mờ*, Luận án Tiến sĩ Máy tính, Học viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam, 2021.

[3] Trần Thanh Đại. *Phương pháp gia tăng rút gọn thuộc tính trong bảng quyết định thay đổi theo tiếp cận tập thô mờ*, Luận án Tiến sĩ Máy tính, Học viện Khoa học và Công nghệ - Viện Hàn lâm Khoa học và Công nghệ Việt Nam, 2021.

[4] Z. Pawlak. "Rough sets". In: *International Journal of Computer & Information Sciences* 11.5 (1982), pp. 341-356.

[5] Y. Yao and Y. Zhao. "Discernibility matrix simplification for constructing attribute reducts". In: *Information Sciences* 179.7 (2009), pp. 867-882.

[6] W. Wei, X. Wu, J. Liang, J. Cui, and Y. Sun. "Discernibility matrix based incremental attribute reduction for dynamic data". In: *Knowledge-Based Systems* 140 (2018), pp. 142-157.

[7] Y. Gonzalez-Diaz, J. F. Martınez-Trinidad, J. A. Carrasco-Ochoa, and M. S. Lazo-Cortes. "Algorithm for computing all the shortest reducts based on a new pruning strategy". In: *Information Sciences* 585 (2022), pp. 113-126.

[8] Y. Qian, J. Liang, W. Pedrycz, and C. Dang. "Positive approximation: An accelerator for attribute reduction in rough set theory". In: *Artificial Intelligence* 174.9 (2010), pp. 597-618.

[9] J. Qian, D. Miao, Z. Zhang, and W. Li. "Hybrid approaches to attribute reduction based on indiscernibility and discernibility relation". In: *International Journal of Approximate Reasoning* 52.2 (2011), pp. 212-230.

[10] F. Jiang, Y. Sui, and L. Zhou. "A relative decision entropy-based feature selection approach". In: *Pattern Recognition* 48.7 (2015), pp. 2151-2163.

[11] F. Li, Z. Zhang, and C. Jin. “Feature selection with partition differentiation entropy for large-scale data sets”. In: *Information Sciences* 329 (2016), pp. 690-700.

[12] M. S. Raza and U. Qamar. “A heuristic based dependency calculation technique for rough set theory”. In: *Pattern Recognition* 81 (2018), pp. 309-325.

[13] Q. Hu, J. Liu, and D. Yu. “Mixed feature selection based on granulation and approximation”. In: *Knowledge-Based Systems* 21 (2008), pp. 137-150.

[14] Q. Hu, D. Yu, J. Liu, and C. Wu. “Neighborhood rough set based heterogeneous feature subset selection”. In: *Information Sciences* 178 (2008), pp. 3577-3594.

[15] Q. Hu, W. Pedrycz, D. Yu, and J. Lang. “Selecting discrete and continuous features based on neighborhood decision error minimization”. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B Cybernetics* 40 (2010), pp. 137-150.

[16] C. Wang, Q. Hu, X. Wang, D. Chen, Y. Qian, and Z. Dong. “Feature selection based on neighborhood discrimination index”. In: *IEEE Transactions on Neural Networks and Learning Systems* 29 (2018), pp. 2986-2999.

[17] C. Wang, Y. Huang, M. Shao, Q. Hu, and D. Chen. “Feature Selection Based on Neighborhood Self-Information”. In: *IEEE Transactions on Cybernetics* 50.9 (2020), pp. 4031-4042.

[18] L. Sun, L. Wang, Y. Qian, J. Xu, and S. Zhang. “Feature selection using lebesgue and entropy measures for incomplete neighborhood decision systems”. In: *Knowledge-Based Systems* 186 (2019), p. 104942.

[19] L. Sun, L. Wang, W. Ding, Y. Qian, and J. Xu. “Feature Selection Based on Neighborhood Self-Information”. In: *IEEE Transactions on Fuzzy Systems* 29.1 (2021), pp. 19-33.

[20] P. Zhang, T. Li, Z. Yuan, C. Luo, K. Liu, and X. Yang. “Heterogeneous Feature Selection Based on Neighborhood Combination Entropy”. In: *IEEE Transactions on Neural Networks and Learning Systems* 35.3 (2022), pp. 1-14.

[21] X. Yang, H. Chen, T. Li, J. Wan, and B. Sang. “Neighborhood rough sets with distance metric learning for feature selection”. In: *Knowledge-Based Systems* 224 (2021), p. 107076.

[22] X. Chen, Z. Yuan, S. Feng. "Anomaly detection based on improved k-nearest neighbor rough sets". In: *International Journal of Approximate Reasoning* 176 (2025), p. 109323.

[23] P. Xu, P. Zhu. "Attribute reduction based on weighted neighborhood distribution entropy". In: *International Journal of Approximate Reasoning* 187 (2025), p. 109539.

[24] D. Chen, W. Li, X. Zhang and, S. Kwong. "Evidence-theory-based numerical algorithms of attribute reduction with neighborhood covering rough sets". In: *International Journal of Approximate Reasoning* 55.3 (2014), pp. 908-923.

[25] Q. Wang, Y. Qian, X. Liang, Q. Guo, and J. Liang. "Local neighborhood rough set". In: *Knowledge-Based Systems* 153 (2018), pp. 53–64.

[26] X. Fan, W. Zhao, C. Wang, and Y. Huang. "Attribute reduction based on max-decision neighborhood rough set model". In: *Knowledge-Based Systems* 151 (2018), pp. 16–23.

[27] S. An, X. Guo, C. Wang, G. Guo, and J. Dai. "A soft neighborhood rough set model and its applications". In: *Information Sciences* 624 (2023), pp. 185–199.

[28] L. Chen, J. Chen, Y. Lin. "Feature selection considering synergy between features based on soft neighborhood rough sets". In: *Engineering Applications of Artificial Intelligence* 150 (2025), p. 110553.

[29] M. Hu, E. C. C. Tsang, Y. Guo, D. Chen, and W. Xu. "A novel approach to attribute reduction based on weighted neighborhood rough sets". In: *Knowledge-Based Systems* 220 (2021), p. 106908.

[30] J. Xie, B. Q. Hu, and H. Jiang. "A novel method to attribute reduction based on weighted neighborhood probabilistic rough sets". In: *International Journal of Approximate Reasoning* 144 (2022), pp. 1-17.

[31] N. N. Thuy and S. Wongthanavasu. "Attribute reduction with fuzzy divergence-based weighted neighborhood rough sets". In: *International Journal of Approximate Reasoning* 173 (2024), p. 109256.

[32] N. Wang and E. Zhao. “A new method for feature selection based on weighted k-nearest neighborhood rough set”. In: *Expert Systems With Applications* 238 (2024), p. 122324.

[33] N. N. Thuy, T. D. Anh, and L. M. Thanh. “Generalized weighted neighborhood rough sets”. In: *Information Sciences* 707 (2025), p. 122020.

[34] D. Dübois and H. Prade. “Rough fuzzy sets and fuzzy rough sets”. In: *International Journal of General Systems* 17 (1990), pp. 191-209.

[35] R. Jensen and Q. Shen. “Fuzzy-rough attribute reduction with application to web categorization”. In: *Fuzzy Sets and Systems* 141 (2004), pp. 469–485.

[36] E. C. C. Tsang, D. Chen, D. S. Yeung, X. Z. Wang, and J. W. T. Lee. “Attributes reduction using fuzzy rough sets”. In: *IEEE Transactions on Fuzzy Systems* 16.5 (2008), pp. 1130–1141.

[37] R. Jensen and Q. Shen. “New approaches to fuzzy-rough feature selection”. In: *IEEE Transactions on Fuzzy Systems* 17.4 (2009), pp. 824–838.

[38] D. Chen, L. Zhang, S. Zhao, Q. Hu, and P. Zhu. “A novel algorithm for finding reducts with fuzzy rough sets”. In: *IEEE Transactions on Fuzzy Systems* 20.2 (2012), pp. 385-389.

[39] S. Zhao, H. Chen, C. Li, M. Zhai, and X. Du. “RFRR: Robust fuzzy rough reduction”. In: *IEEE Transactions on Fuzzy Systems* 21.5 (2013), pp. 825-841.

[40] D. Chen and Y. Yang. “Attribute reduction for heterogeneous data based on the combination of classical and fuzzy rough set models”. In: *IEEE Transactions on Fuzzy Systems* 22.5 (2014), pp. 1325–1334.

[41] J. Dai and Q. Xu. “Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification”. In: *Applied Soft Computing* 13.1 (2013), pp. 211–221.

[42] C. Wang, Y. Qi, and Q. He. “Attribute reduction using distance-based fuzzy rough sets”. In: *International Conference on Machine Learning and Cybernetics (ICMLC)* (2015), pp. 860-865.

[43] C. Wang, Y. Huang, M. Shao, and X. Fan. "Fuzzy rough set-based attribute reduction using distance measures". In: *Knowledge-Based Systems* 164 (2019), pp. 205-212.

[44] Y. Lin, Y. Li, C. Wang, and J. Chen, "Attribute reduction for multi-label learning with fuzzy rough set". In: *Knowledge-Based Systems* 152 (2018), pp. 51-61.

[45] T. K. Sheeja, and A. S. Kuriakose. "A novel feature selection method using fuzzy rough sets". In: *Computers in Industry* 97 (2018), pp. 111-116.

[46] Y. Qian, Q. Wang, H. Cheng, J. Liang, and C. Dang. "Fuzzy-rough feature selection accelerator". In: *Fuzzy Sets and Systems* 258 (2015), pp. 61-78.

[47] Y. Lin, Q. Hu, J. Liu, J. Li, and X. Wu. "Streaming Feature Selection for Multilabel Learning Based on Fuzzy Mutual Information". In: *IEEE Transactions on Fuzzy Systems* 25 (2017), pp. 1491-1507.

[48] Q. Hu, D. Yu, Z. Xie, and J. Liu. "Fuzzy probabilistic approximation spaces and their information measures", *IEEE Transactions on Fuzzy Systems* 14 (2006), pp. 191-201.

[49] J. Dai, Y. Yan, Z. Li, and B. Liao. "Dominance-based fuzzy rough set approach for incomplete interval-valued data". In: *Journal of Intelligent & Fuzzy Systems* 34 (2018), pp. 423-436.

[50] Y. Qian, J. Liang, W. Z. Wu, and C. Dang. "Information Granularity in Fuzzy Binary GrC Model". In: *IEEE Transactions on Fuzzy Systems* 19 (2011), pp. 253-264.

[51] W. L. Hung and M. S. Yang. "Similarity measures of intuitionistic fuzzy sets based on lp metric". In: *International Journal of Approximate Reasoning* 46 (2007), pp. 120-136.

[52] B. Huang, C. X. Guo, Y. L. Zhuang, H. X. Li, and X. Z. Zhou. "Intuitionistic fuzzy multigranulation rough sets". In: *Information Sciences* 277 (2014), pp. 299-320.

[53] A. K. Tiwari, S. Shreevastava, T. Som, and K. K. Shukla. "Tolerance-based intuitionistic fuzzy-rough set approach for attribute reduction". In: *Expert Systems With Applications* 101 (2018), pp. 205-212.

[54] A. K. Tiwari, S. Shreevastava, K. Subbiah, and T. Som. "An intuitionistic fuzzy-rough set model and its application to feature selection". In: *Journal of Intelligent & Fuzzy Systems* 36 (2019), pp. 4969-4979.

[55] A. Tan, W. Z. Wu, , Y. Qian, J. Liang, J. Chen, and J. Li. "Intuitionistic fuzzy rough set-based granular structures and attribute subset selection". In: *IEEE Transactions on Fuzzy Systems* 27.3 (2019), pp. 527-539.

[56] P. Jain, A. K. Tiwari, and T. Som. "A fitting model based intuitionistic fuzzy rough feature selection". In: *Engineering Applications of Artificial Intelligence* 89 (2020), p. 103421.

[57] A. Tan, S. Shi, W. Z. Wu, J. Li, and W. Pedrycz. "Granularity and Entropy of Intuitionistic Fuzzy Information and Their Applications". In: *IEEE Transactions on Cybernetics* 52 (2022), pp. 192-204.

[58] M. B. Revanasiddappa and B. S. Harish. "A new feature selection method based on intuitionistic fuzzy entropy to categorize text documents". In: *International Journal of Interactive Multimedia and Artificial Intelligence* 5.3 (2018), pp. 106-117.

[59] N. T. Thang, N. L. Giang, T. T. Dai, N. T. Tuan, N. Q. Huy, P. V. Anh, and V. D. Thi. "A Novel Filter-Wrapper Algorithm on Intuitionistic Fuzzy Set for Attribute Reduction from Decision Tables". In: *International Journal of Data Warehousing and Mining* 17 (2021), pp. 67-100.

[60] Z. Liu. "An incremental arithmetic for the smallest reduction of attributes". In: *Chinese Journal of Electronics* 27.11 (1999), pp. 96-98.

[61] J. Wang and J. Wang. "Reduction algorithms based on discernibility matrix: The ordered attributes method". In: *Journal of Computer Science and Technology* 16 (2001), pp. 489-504.

[62] Y. Xu, L. Wang, and R. Zhang. "A dynamic attribute reduction algorithm based on 0-1 integer programming". In: *Knowledge-Based Systems* 24 (2011), pp. 1341-1347.

[63] M. Yang. "An incremental updating algorithm of the computation of a core based on the improved discernibility matrix". In: *Chinese Journal of Computers* 29.3 (2006), pp. 407-413.

[64] M. Yang. “An incremental updating algorithm for attribute reduction based on improved discernibility matrix”. In: *Chinese Journal of Computers* 30.5 (2007), pp. 815-822.

[65] W. Shu, W. Qian, and Y. Xie. “Incremental approaches for feature selection from dynamic data with the variation of multiple objects”. In: *Knowledge-Based Systems* 163 (2019), pp. 320-331.

[66] A. K. Das, S. Sengupta, and S. Bhattacharyya. “A group incremental feature selection for classification using rough set theory based genetic algorithm”. In: *Applied Soft Computing* 65 (2018), pp. 400-411.

[67] G. Lang, M. Cai, H. Fujita, and Q. Xiao. “Related families-based attribute reduction of dynamic covering decision information systems”. In: *Knowledge-Based Systems* 162 (2018), pp. 161-173.

[68] G. Hao, L. Longshu, Y. Chuanjian, and D. Jian. “Incremental reduction algorithm with acceleration strategy based on conflict region”. In: *Artificial Intelligence Review* 51.4 (2019), pp. 507-536.

[69] J. Y. Liang, F. Fang, C. Y. Dang, and Y. H. Qian. “A group incremental approach to feature selection applying rough set technique”. In: *IEEE Transactions on Knowledge and Data Engineering* 2 (2014), pp. 294-308.

[70] W. Shu, W. Qian, and Y. Xie. “Incremental feature selection for dynamic hybrid data using neighborhood rough set”. In: *Knowledge-Based Systems* 194 (2020), p. 105516.

[71] Y. G. Jing, T. R. Li, J. F. Huang, H.M. Chen, and S. J. Horng. “A Group Incremental Reduction Algorithm with Varying Data Values”. In: *International Journal of Intelligent Systems* 32.9 (2017), pp. 900-925.

[72] Y. G. Jing, T. R. Li, H. Fujita, Z. Yu, B. Wang. “An incremental attribute reduction approach based on knowledge granularity with a multi-granulation view”. In: *Information Sciences* 411 (2017), pp. 23-38.

[73] C. Zhang, J. Dai, and J. Chen. “Knowledge granularity based incremental attribute reduction for incomplete decision systems”. In: *International Journal of Machine Learning and Cybernetics* 11 (2020), pp. 1141-1157.

[74] M. Cai, G. Lang, H. Fujita, Z. Li, and T. Yang. "Incremental approaches to updating reducts under dynamic covering granularity". In: *Knowledge-Based Systems* 172 (2019), pp. 130-140.

[75] C. Zhang and J. Dai. "An incremental attribute reduction approach based on knowledge granularity for incomplete decision systems". In: *Granular Computing* 5 (2019), pp. 545-559.

[76] Y. M. Liu, S. Y. Zhao, H. Chen, C. P. Li, Y. M. Lu. "Fuzzy Rough Incremental Attribute Reduction Applying Dependency Measures". In: *APWeb-WAIM 2017: Web and Big Data* (2017), pp. 484-492.

[77] Y. Y. Yang, D. G. Chen, H. Wang, and X. H. Wang. "Incremental perspective for feature selection based on fuzzy rough sets". In: *IEEE Transactions on Fuzzy Systems* 26.3 (2017), pp. 1257-1273.

[78] Y. Y. Yang, D. G. Chen, H. Wang, E. C.C.Tsang, and D. L. Zhang. "Fuzzy rough set based incremental attribute reduction from dynamic data with sample arriving". In: *Fuzzy Sets and Systems* 312 (2017), pp. 66-86.

[79] X. Zhang, C. L Mei, D. G Chen, Y. Y Yang, and J. H. Li. "Active Incremental Feature Selection Using a Fuzzy-Rough-Set-Based Information Entropy". In: *IEEE Transactions on Fuzzy Systems* 28.5 (2020), pp. 901-915.

[80] P. Ni, S. Y. Zhao, X. H. Wang, H. Chen, C. P Li, and E. C. C Tsang. "Incremental Feature Selection Based on Fuzzy Rough Sets". In: *Information Sciences* 536 (2020), pp. 185-204.

[81] N. L. Giang, L. H. Son, T. T. Ngan, T. M. Tuan, H. T. Phuong, M. A. Basset, A. R. L. de Macêdo, and V. C. de Albuquerque. "Novel Incremental Algorithms for Attribute Reduction from DynamicDecision Tables using Hybrid Filter-Wrapper with Fuzzy Partition Distance". In: *IEEE Transactions on Fuzzy Systems* 28.5 (2020), pp. 858-873.

[82] D. Xia, G. Wang, Q. Zhang, J. Yang, S. Li, and M. Gao. "Incremental Approximation Feature Selection With Accelerator for Rough Fuzzy Sets by Knowledge Distance". In: *IEEE Transactions on Fuzzy Systems* 31.11 (2023), pp. 3959-3973.

[83] P. Dhal, C. Azad. “A comprehensive survey on feature selection in the various fields of machine learning”. In: *Applied Intelligence* 52 (2022), pp. 4543-4581.

[84] H. Kuang, L. Chen, L. L. H. Chan, R. C. C. Cheung, H. Yan. “Feature Selection Based on Tensor Decomposition and Object Proposal for Night-Time Multiclass Vehicle Detection”. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 49 (2019), pp. 71-80.

[85] B. Chandra, M. Gupta. “An efficient statistical feature selection approach for classification of gene expression data”. In: *Journal of Biomedical Informatics* 44 (2011), pp. 529-535.

[86] W. Xu and Q. Bu. “Matrix-based incremental feature selection method using weight-partitioned multigranulation rough set”. In: *Information Sciences* 681 (2024), p. 121219.

[87] J. C. Ang, A. Mirzal, H. Haron, H. N. A. Hamed. “Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 13 (2016), pp. 971-989.

[88] J. Demetrovics, N. L. Giang, and V. D.Thi. “On finding all reducts of consistent decision tables”. In: *Cybernetics and Information Technologies* 14.4 (2014), pp. 3-10.

[89] K. Atanassov. “Intuitionistic fuzzy sets”. In: *Fuzzy sets and Systems* 20.1 (1986), pp. 87-96.

[90] I. Iancu. “Intuitionistic fuzzy similarity measures based on Frank t-norms family”. In: *Pattern Recognition Letters* 42 (2014), pp. 128-136.

[91] P. V. Anh, N. N. Thuy, N. L. Giang, P. D. Khanh, and N. T. Thuy. “A Novel Incremental Attribute Reduction Algorithm Based on Intuitionistic Fuzzy Partition Distance”. In: *Computer Systems Science and Engineering* 47.3 (2023), pp. 2971-2988.

[92] Z. Yuan, H. Chen, P. Xie, P. Zhang, J. Liu, and T. Li. “Attribute reduction methods in fuzzy rough set theory: An overview, comparative experiments, and new directions”. In: *Applied Soft Computing* 107 (2021), p. 107353.

[93] D. K. Basnet. "(α, β)-cut of intuitionistic fuzzy ideals". In: *Applied Soft Computing* 4.27 (2010), pp. 1329-1334

[94] L. Zhou and W. Wu. "On generalized intuitionistic fuzzy rough approximation operators". In: *Information Sciences* 178 (2008), pp. 2448-2465.

[95] L. Zhou, W. Wu, and W. Zhang. "On characterization of intuitionistic fuzzy rough sets based on intuitionistic fuzzy implicators". In: *Information Sciences* 179 (2009), pp. 883-898.

[96] N. Rehman, A. Ali, and K. Hila. "Note on 'Tolerance-based intuitionistic fuzzy-rough set approach for attribute reduction"'. In: *Expert Systems With Applications* 175 (2021), p. 114869.

[97] M. Rahimi, P. Kumar, B. Moomivand, and G. Yari. "An intuitionistic fuzzy entropy approach for supplier selection". In: *Complex & Intelligent Systems* 7.4 (2021), pp. 1869-1876.

[98] X. Liu, H. Mo, and J. Dai. "Attribute reduction based on intuitionistic fuzzy dominance mutual information in intuitionistic fuzzy information systems". In: *Information Sciences* 676 (2024), p. 120851.

[99] X. Zhang, B. Zhou, and P. Li. "A general frame for intuitionistic fuzzy rough sets". In: *Information Sciences* 216 (2012), pp. 34-49.

[100] Z. Zhang. "Attributes reduction based on intuitionistic fuzzy rough sets". In: *Journal of Intelligent and Fuzzy Systems* 30.2 (2016), pp. 1127-1137.

[101] S. Shreevastava, A. K. Tiwari, and T. Som. "Intuitionistic fuzzy neighborhood rough set model for feature selection". In: *International Journal of Fuzzy System Applications* 7.2 (2018), pp. 75-84.

[102] Z. Zhang and J. Tian. "On attribute reduction with intuitionistic fuzzy rough sets". In: *International Journal of Uncertainty, Fuzziness and Knowlege-Based Systems* 20.1 (2012), pp. 59-76.

[103] B. B. Shang, X. Y. Zhang, and W. H. Xu. "Attribute reduction of relative knowledge granularity in intuitionistic fuzzy ordered decision table". In: *Filomat* 32.5 (2018), pp. 1727-1736.

[104] H. V. Henderson and S. R. Searle. “On Deriving the Inverse of a Sum of Matrices”. In: *SIAM Review* 23.1 (1981), pp. 53-60.