

**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**

TÓM TẮT LUẬN ÁN TIẾN SĨ MÁY TÍNH

Ngành: Hệ thống thông tin

Mã số: 9 48 01 04

Hà Nội - 2026

Công trình được hoàn thành tại: Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

Người hướng dẫn khoa học:

1. Người hướng dẫn khoa học 1: PGS.TS. Nguyễn Long Giang
2. Người hướng dẫn khoa học 2: TS. Nguyễn Ngọc Thủy

Phản biện 1:

Phản biện 2:

Luận án được bảo vệ trước Hội đồng đánh giá luận án tiến sĩ cấp Học viện họp tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam vào hồi giờ, ngày tháng năm

Có thể tìm hiểu luận án tại:

1. Thư viện Học viện Khoa học và Công nghệ
2. Thư viện Quốc gia Việt Nam

Mở đầu

1. Tính cấp thiết của luận án

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]. 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. 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 được mở rộng từ lý thuyết tập thô. 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. 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 nhiều có độ chính xác phân loại thấp.

Trong những năm gần đây, các phương pháp rút gọn thuộc tính theo *mô hình tập thô mờ trực cảm* đang nhận được nhiều sự quan tâm. Ưu điểm của mô hình này là sự bổ sung cần thiết của thành phần hàm không thuộc 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. 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 tồn tại một số nhược điểm. Thứ nhất, việc bổ sung thành phần hàm không thuộc khiến cho các thuật toán theo hướng tiếp cận này tiêu tốn nhiều không gian lưu trữ và có độ phức tạp tính toán cao hơn so với các phương pháp tập thô mờ truyền thống. Thứ hai, những đối tượng có sự phân bố khác biệt so với phần lớn các đối tượng trong tập vũ trụ sẽ tạo ra nhiễu trong tính toán.

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. Điều này đã mang tới những động lực trong việc nghiên cứu và phát triển các thuật toán gia tăng theo tiếp cận mô hình tập thô mờ trực cảm.

2. Mục tiêu nghiên cứu

1) *Đề xuất một số mô hình được mở rộng từ mô hình tập thô 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 thô 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.

3. Đối tượng nghiên cứu

Luận án tập trung nghiên cứu về các khái niệm cơ bản như bảng quyết định, tập rút gọn và một số phương pháp rút gọn thuộc tính trên bảng quyết định thông qua một số nhánh mở rộng của mô hình tập thô bao gồm:

- 1) Khảo sát một số mở rộng theo nhánh mô hình tập thô lân cận và mô hình tập thô mờ cùng các phương pháp rút gọn thuộc tính.
- 2) Khảo sát mô hình tập thô mờ trực cảm, một số độ đo đánh giá ý nghĩa của thuộc tính và các phương pháp rút gọn thuộc tính theo tiếp cận Heuristic.

4. Phạm vi nghiên cứu

Phạm vi nghiên cứu của luận án tập trung vào các hạn chế và hướng cải tiến trong một số mở rộng của mô hình tập thô và các phương pháp rút gọn thuộc tính được áp dụng trên bảng quyết định cố định và bảng quyết định có sự thay đổi tập đối tượng, cụ thể:

- 1) Nghiên cứu một số mô hình được mở rộng từ nền tảng mô hình tập thô mờ trực cảm để xây dựng các thuật toán rút gọn thuộc tính trong bảng quyết định cố định.
- 2) Nghiên cứu các thuật toán gia tăng thông qua các mô hình đề xuất nhằm tìm kiếm rút gọn trên bảng quyết định có sự thay đổi tập đối tượng.

5. Phương pháp nghiên cứu

- 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 UCI¹ và OpenML² nhằm đánh giá tính hiệu quả của các thuật toán đề xuất.

6. 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 thô mờ trực cảm, là 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 sẽ 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 có sự bổ sung và loại bỏ tập đối tượng theo tiếp cận mô hình tập thô mờ trực cảm mức α, β .

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.

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, hướng phát triển trong tương lai và những vấn đề quan tâm của tác giả.

¹<https://archive.ics.uci.edu/datasets>

²<https://openml.org/search?type=data&status=active&sort=runs>.

Chương 1: Tổng quan về bài toán rút gọn thuộc tính trên bảng quyết định

Chương 1 sẽ trình bày một số đóng góp chính như sau:

- (1) Giới thiệu tổng quan về bài toán rút gọn thuộc tính và số hướng tiếp cận điển hình.
- (2) 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.
- (3) 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.

1.1. Tổng quan về rút gọn thuộc tính

1.1.1 Một số khái niệm cơ bản

Rút gọn thuộc tính là bài toán quan trọng bước tiền xử lý dữ liệu với mục tiêu chính là 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. Hiện nay, các bài toán rút gọn thuộc tính thường được xử lý trên các bảng quyết định.

Bảng quyết định 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 $c : U \rightarrow V_c$ 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 $c(u)$. 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].

1.1.2 Một số mô hình trong rút gọn thuộc tính

Tập thô lân cận (NRSs) và tập thô k -lân cận gần nhất (KNNRSs) được Hu và cộng sự lần đầu giới thiệu vào năm 2008 [13]. 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. 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 (WNRSs) sử dụng khoảng cách dựa trên trọng số các thuộc tính:

$$\Delta_A^\omega(u, v) = \sqrt{\sum_{a \in A} \omega^2(a) \cdot (a(u) - a(v))^2} \quad (1.1)$$

trong đó, $\omega(a)$ là trọng số của thuộc tính $a \in A$, được xác định như sau. Cho bảng quyết định $IS = (U, C \cup D)$, $U = \{u_1, u_2, \dots, u_n\}$ là tập các đối tượng, $C = \{a_1, a_2, \dots, a_m\}$ 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 $\lambda = (\lambda(a_1), \lambda(a_2), \dots, \lambda(a_m))^T$ được xác định dựa trên nghiệm $\lambda = (\mathbf{A}^T \mathbf{A} + \mathbf{E})^{-1} \mathbf{A}^T \mathbf{Y}$, trong đó, $\mathbf{A}$ 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 $\mathbf{Y} = (D(a_1), D(a_2), \dots, D(a_m))^T$ là một *vector quyết định*.

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:

$$\omega(a) = \frac{|C| \cdot |\lambda(a)|}{\sum_{a \in C} |\lambda(a)|} \quad (1.2)$$

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:

$$\Delta_A^\omega(u, v) = \sqrt[p]{\sum_{a \in A} (\omega(a) \cdot |a(u) - a(v)|)^p} \quad (1.3)$$

Quan hệ lân cận trọng số, ký hiệu là $\mathcal{R}_A^{\delta, \omega}$, được định nghĩa bởi:

$$\mathcal{R}_A^{\delta, \omega} = \{(u, v) \in U \times U : \Delta_A^\omega(u, v) \leq \delta\} \quad (1.4)$$

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:

$$[u]_A^{\delta, \omega} = \{v \in U : (u, v) \in \mathcal{R}_A^{\delta, \omega}\} \quad (1.5)$$

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:

$$\underline{WN}_A(X) = \{u \in U : [u]_A^{\delta, \omega} \subseteq X\} \quad (1.6)$$

$$\overline{WN}_A(X) = \{u \in U : [u]_A^{\delta, \omega} \cap X \neq \emptyset\} \quad (1.7)$$

Từ vấn đề này, các công trình [32] và [33] đã đề xuất các mô hình có gán trọng số cho các đối tượng và chứng minh được hiệu quả của các thuật toán rút gọn thuộc tính. 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.

- 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.

Nhánh nghiên cứu về mô hình tập thô mờ (FRSs) 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. Cụ thể, 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: $\widetilde{\mathcal{R}}_{\{a\}}(u, v) \geq \sup_{t \in U} \min \{ \widetilde{\mathcal{R}}_{\{a\}}(u, t), \widetilde{\mathcal{R}}_{\{a\}}(t, v) \}$.

Quan hệ tương đương mờ sẽ sinh ra một phân hoạch mờ $U/\widetilde{\mathcal{R}}_{\{a\}} = \{ \widetilde{[u]}_{\{a\}} : u \in U \}$ 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 $|\widetilde{[u]}_{\{a\}}| = \sum_{v \in U} \widetilde{[u]}_{\{a\}}(v)$.

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. Tuy nhiên, mô hình FRSs 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.
- 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 hiệu quả.
- Các phương pháp theo mô hình FRSs hiện chưa áp dụng việc thiết lập trọng số của các thuộc tính điều kiện.

Nhìn chung, cả hai hướng mở rộng của lý thuyết tập thô đều đối mặt với không ít khó khăn. Những hạn chế này 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 thô mờ trực cảm được coi là một giải pháp tiềm năng trong việc phát triển các mô hình mới.

1.2. Mô hình tập thô mờ trực cảm

Định nghĩa 1.1. (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:

$$\ddot{\mathcal{R}}_{\{a\}} = \left\{ \left((u, v), \gamma_{\ddot{\mathcal{R}}_{\{a\}}}(u, v), \eta_{\ddot{\mathcal{R}}_{\{a\}}}(u, v) \right) : (u, v) \in U \times U \right\} \quad (1.8)$$

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

Từ Định nghĩa 1.1, một quan hệ $\check{\mathcal{R}}_{\{a\}}$ 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: $\gamma_{\check{\mathcal{R}}_{\{a\}}}(u, u) = 1$ và $\eta_{\check{\mathcal{R}}_{\{a\}}}(u, u) = 0$,
2. Tính đối xứng: $\gamma_{\check{\mathcal{R}}_{\{a\}}}(u, v) = \gamma_{\check{\mathcal{R}}_{\{a\}}}(v, u)$ và $\eta_{\check{\mathcal{R}}_{\{a\}}}(u, v) = \eta_{\check{\mathcal{R}}_{\{a\}}}(v, u)$,
3. Tính bắc cầu:
$$\begin{cases} \gamma_{\check{\mathcal{R}}_{\{a\}}}(u, v) \geq \sup_{t \in U} \left\{ \min \left(\gamma_{\check{\mathcal{R}}_{\{a\}}}(u, t), \gamma_{\check{\mathcal{R}}_{\{a\}}}(t, v) \right) \right\} \\ \eta_{\check{\mathcal{R}}_{\{a\}}}(u, v) \leq \inf_{t \in U} \left\{ \max \left(\eta_{\check{\mathcal{R}}_{\{a\}}}(u, t), \eta_{\check{\mathcal{R}}_{\{a\}}}(t, v) \right) \right\} \end{cases}$$

Mỗi thuộc tính $a \in C$ trên bảng quyết định sẽ xác định một quan hệ tương đương mờ trực cảm $\check{\mathcal{R}}_{\{a\}}$. Khi đó, một quan hệ tương đương mờ trực cảm $\check{\mathcal{R}}_{\{a\}}$ sẽ sinh ra một phân hoạch mờ trực cảm $U/\check{\mathcal{R}}_{\{a\}} = \{[u]_{\check{\mathcal{R}}_{\{a\}}} \mid u \in U\}$ trên tập vũ trụ U , trong đó $[u]_{\check{\mathcal{R}}_{\{a\}}}$ được gọi là một hạt thông tin mờ trực cảm của đối tượng u theo quan hệ $\check{\mathcal{R}}_{\{a\}}$.

Từ mô hình này, đã có nhiều phương pháp rút gọn thuộc tính được phát triển theo nhiều độ đo khác nhau như [53, 96, 54, 56], entropy mờ trực cảm [57, 58, 97, 98] và khoảng cách mờ trực cảm [59, 91]. Các phương pháp này đều đã chứng minh được hiệu quả của mô hình IFRSs trong việc cải thiện hiệu quả phân lớp trên các bộ dữ liệu có độ chính xác phân lớp ban đầu thấp. Ưu điểm của IFRSs là cơ sở vững chắc trong việc phát triển các mô hình mở rộng nhằm khắc phục những hạn chế trên các nhánh của mô hình RSs.

- So với các mô hình NRSs và FRSSs, mô hình IFRSs có cấu trúc hạt chi tiết hơn, phản ánh rõ nét sự mơ hồ và do dự vốn tồn tại trong dữ liệu thực tế.
- 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 nhiều.

Tuy nhiên, qua quá trình khảo sát, mô hình IFRSs vẫn còn một số những hạn chế 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.
- 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.

Như vậy, động lực nghiên cứu của luận án không chỉ dừng lại ở việc sử dụng mô hình IFRSs như một sự kế thừa để phát triển các mô hình mới, mà còn tập trung vào việc giải quyết những hạn chế của chính mô hình IFRSs. 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 IFRSs 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.

Chương 2: Đề xuất một số thuật toán rút gọn thuộc tính dựa trên tập mờ trực cảm mức alpha, beta

2.1. Mở đầu

Chương 2 sẽ trình bày một số đóng góp chính như sau: (1) Đề xuất mô hình tập mờ trực cảm mức *alpha*, *beta* và trình bày các tính chất quan trọng của mô hình, (2) Đề xuất thuật toán rút gọn thuộc tính trên bảng quyết định cố định, (3) Đề xuất hai thuật toán gia tăng rút gọn thuộc tính trên bảng quyết định thay đổi, (4) Chứng minh hiệu quả các thuật toán đề xuất so với một số thuật toán theo mô hình tập mờ, tập 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.

2.1.1 Khái niệm về tập mờ trực cảm mức alpha, beta

Định nghĩa 2.1. Cho $[\ddot{u}]_{\{a\}}$ 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 *alpha*, *beta* của $[\ddot{u}]_{\{a\}}$ là một tập rõ được định nghĩa như sau:

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

Một tập mờ trực cảm mức *alpha*, *beta*, ký hiệu là $[\ddot{u}]_{\{a\}}^{\alpha, \beta}$ 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:

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

Nếu a là kiểu thuộc tính số, giá trị độ tương tự và độ khác biệt của v trong hạt thông tin mờ trực cảm của u được xác định theo các công thức trong [91]. Trong nghiên cứu này, $[\ddot{u}]_{\{a\}}^{\alpha, \beta}$ được gọi là một hạt thông tin mờ trực cảm mức *alpha*, *beta* của u . Qua đó, một họ $U/\ddot{\mathcal{R}}_{\{a\}}^{\alpha, \beta} = \left\{ [\ddot{u}]_{\{a\}}^{\alpha, \beta} : u \in U \right\}$ được gọi là một phân hoạch mờ trực cảm mức *alpha*, *beta*.

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

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

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

Mệnh đề 2.2. Cho $IS = (U, C \cup D)$ và $A \subseteq C$, nếu $\alpha_1 \leq \alpha_2$ và $\beta_1 \geq \beta_2$, thì $[\ddot{u}]_A^{\alpha_2, \beta_2} \subseteq [\ddot{u}]_A^{\alpha_1, \beta_1}$.

2.2. Đề xuất thuật toán rút gọn thuộc tính trên bảng quyết định

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

Mệnh đề 2.3. 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:

$$\ddot{D} \left(U/\ddot{\mathcal{R}}_C^{\alpha, \beta}, U/\ddot{\mathcal{R}}_{C \cup D}^{\alpha, \beta} \right) = \sum_{u \in U} \frac{\left(\left| [\ddot{u}]_C^{\alpha, \beta} \right| - \left| [\ddot{u}]_C^{\alpha, \beta} \cap [\ddot{u}]_D^{\alpha, \beta} \right| \right)}{|U|^2} \quad (2.3)$$

Mệnh đề 2.4. Cho $A, B \subseteq C$. Nếu $A \subseteq B$ thì $\ddot{D}(U/\ddot{\mathcal{R}}_A^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{A \cup D}^{\alpha,\beta}) \geq \ddot{D}(U/\ddot{\mathcal{R}}_B^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{B \cup D}^{\alpha,\beta})$.

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. $\ddot{D}(U/\ddot{\mathcal{R}}_A^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{A \cup D}^{\alpha,\beta}) = \ddot{D}(U/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C \cup D}^{\alpha,\beta})$,
2. $\forall A' \subseteq A, \ddot{D}(U/\ddot{\mathcal{R}}_{A'}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{A' \cup D}^{\alpha,\beta}) \geq \ddot{D}(U/\ddot{\mathcal{R}}_A^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{A \cup D}^{\alpha,\beta})$.

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:

$$Sig_P(a, A) = \ddot{D}(U/\ddot{\mathcal{R}}_A^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{A \cup D}^{\alpha,\beta}) - \ddot{D}(U/\ddot{\mathcal{R}}_{A \cup \{a\}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{A \cup \{a\} \cup D}^{\alpha,\beta}) \quad (2.4)$$

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 $U/\ddot{\mathcal{R}}_{\{a\}}^{\alpha,\beta}$
 - 5: **else**
 - 6: $U/\ddot{\mathcal{R}}_{\{a\}}^{\alpha,\beta} = U/\ddot{\mathcal{R}}_{\{a\}}$
 - 7: **end if**
 - 8: **end for**
 - 9: tính toán $U/\ddot{\mathcal{R}}_C^{\alpha,\beta} = \bigcap_{a \in C} U/\ddot{\mathcal{R}}_{\{a\}}^{\alpha,\beta}$
 - 10: $\mathcal{A} = \{a_0\}$ thỏa mãn $\ddot{D}(U/\ddot{\mathcal{R}}_{\{a_0\}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{\{a_0\} \cup D}^{\alpha,\beta}) = \min_{a \in C} \ddot{D}(U/\ddot{\mathcal{R}}_{\{a\}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{\{a\} \cup D}^{\alpha,\beta})$
 - 11: **while** $\ddot{D}(U/\ddot{\mathcal{R}}_{\mathcal{A}}^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{\mathcal{A} \cup D}^{\alpha,\beta}) \geq \ddot{D}(U/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C \cup D}^{\alpha,\beta})$ **do**
 - 12: tính toán $Sig(a, \mathcal{A})$, với mọi $a \in C \setminus \mathcal{A}$
 - 13: lựa chọn a_0 thỏa mãn $Sig_P(a_0, \mathcal{A}) = \max_{a \in C \setminus \mathcal{A}} \{Sig_P(a, \mathcal{A})\}$
 - 14: $\mathcal{A} \leftarrow \mathcal{A} \cup \{a_0\}$
 - 15: **end while**
 - 16: **return** $\mathcal{A}$
-

Thời gian thực thi của Thuật toán 2.1 là $O(|C|^2|U|^2)$. Trong đó, $|C|, |U|$ tương ứng là số lượng thuộc tính điều kiện và số lượng đối tượng trên bảng quyết định.

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

Mệnh đề 2.5. Cho $IS = (U, C \cup D)$ với $U = \{u_1, u_2, \dots, u_n\}$, các quan hệ tương đương mờ trực cảm $\ddot{\mathcal{R}}_C, \ddot{\mathcal{R}}_D$ và $\Delta U = \{u_{n+1}, u_{n+2}, \dots, u_{n+z}\}$ là một tập đối tượng mới được bổ sung vào bảng, 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 $\ddot{D}(U^+/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{C \cup D}^{\alpha,\beta})$, được xác định như sau:

$$\begin{aligned} \ddot{D}(U^+/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{C \cup D}^{\alpha,\beta}) &= \frac{n^2 \ddot{D}(U/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C \cup D}^{\alpha,\beta})}{(n+z)^2} \\ &\quad + 2 \sum_{i=1}^z \left(\left| [\ddot{u}_{n+i}]_C^{\alpha,\beta} \right| - \left| [\ddot{u}_{n+i}]_C^{\alpha,\beta} \cap [\ddot{u}_{n+i}]_D^{\alpha,\beta} \right| - \vartheta_i \right) \end{aligned} \quad (2.5)$$

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

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

Mệnh đề 2.6. 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 $\Delta U = \{U_{n+1}, U_{n+2}, \dots, U_{n+z}\}$, 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:

$$\ddot{D}(U^+/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{C \cup D}^{\alpha,\beta}) = \frac{n^2 \ddot{D}(U/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C \cup D}^{\alpha,\beta}) + 2 \sum_{i=1}^z \left(\left| [\ddot{u}_{n+i}]_C^{\alpha,\beta} \right| - \left| [\ddot{u}_{n+i}]_C^{\alpha,\beta} \cap [\ddot{u}_{n+i}]_D^{\alpha,\beta} \right| \right)}{(n+z)^2} \quad (2.6)$$

2. $\ddot{D}(U^+/\ddot{\mathcal{R}}_A^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{A \cup D}^{\alpha,\beta}) = \ddot{D}(U^+/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{C \cup D}^{\alpha,\beta})$ nếu $[\ddot{u}_{n+i}]_A^{\alpha,\beta} \subseteq [\ddot{u}_{n+i}]_D^{\alpha,\beta}$ với $i = 1, \dots, z$.

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 $U = \{u_1, u_2, \dots, u_n\}$, rút gọn $\mathcal{A} \subseteq C$, các phân hoạch $U/\ddot{\mathcal{R}}_A^{\alpha,\beta}$, $U/\ddot{\mathcal{R}}_C^{\alpha,\beta}$ và tập đối tượng bổ sung $\Delta U = \{u_{n+1}, u_{n+2}, \dots, u_{n+z}\}$.

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

// kiểm tra tập đối tượng bổ sung

```

1:  $\mathcal{A}^+ \leftarrow \mathcal{A}$ ,  $Z \leftarrow \Delta U$ 
2: for  $i = 1$  to  $z$  do
3:   if  $[\ddot{u}_{n+i}]_{\mathcal{A}}^{\alpha,\beta} \subseteq [\ddot{u}_{n+i}]_D^{\alpha,\beta}$  then
4:      $Z \leftarrow Z \setminus \{u_{n+i}\}$ 
5:   end if
6: end for
7: if  $Z = \emptyset$  then
8:   return  $\mathcal{A}^+$ 
9: end if
10:  $\Delta U \leftarrow Z$ ,  $z \leftarrow |\Delta U|$ 
11: 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  $a \in C \setminus \mathcal{A}$ 
// lọc tìm kiếm rút gọn xấp xỉ
12: tính toán  $\ddot{D}(U^+/\ddot{\mathcal{R}}_{\mathcal{A}^+}^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{\mathcal{A}^+ \cup D}^{\alpha,\beta})$  and  $\ddot{D}(U^+/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{C \cup D}^{\alpha,\beta})$  từ Công thức 2.5.
13: while  $\ddot{D}(U^+/\ddot{\mathcal{R}}_{\mathcal{A}^+}^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{\mathcal{A}^+ \cup D}^{\alpha,\beta}) \geq \ddot{D}(U^+/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{C \cup D}^{\alpha,\beta})$  do
14:   tính toán  $Sig(a, \mathcal{A}^+)$ , với mọi  $a \in C \setminus \mathcal{A}^+$ 
15:   lựa chọn  $a_0$  thỏa mãn:  $Sig_P(a_0, \mathcal{A}^+) = \max_{a \in C \setminus \mathcal{A}^+} \{Sig(a, \mathcal{A}^+)\}$ 
16:    $\mathcal{A}^+ \leftarrow \mathcal{A}^+ \cup \{a_0\}$ 
17: end while
// loại bỏ các thuộc tính dư thừa
18: for  $a \in \mathcal{A}^+$  do
19:   tính toán  $\ddot{D}(U^+/\ddot{\mathcal{R}}_{\mathcal{A}^+ \setminus \{a\}}^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{\mathcal{A}^+ \setminus \{a\} \cup D}^{\alpha,\beta})$ 
20:   if  $\ddot{D}(U^+/\ddot{\mathcal{R}}_{\mathcal{A}^+ \setminus \{a\}}^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{\mathcal{A}^+ \setminus \{a\} \cup D}^{\alpha,\beta}) = \ddot{D}(U^+/\ddot{\mathcal{R}}_{\mathcal{A}^+}^{\alpha,\beta}, U^+/\ddot{\mathcal{R}}_{\mathcal{A}^+ \cup D}^{\alpha,\beta})$  then
21:      $\mathcal{A}^+ \leftarrow \mathcal{A}^+ \setminus \{a_0\}$ 
22:   end if
23: end for
24: return  $\mathcal{A}^+$ 

```

Độ phức tạp của IARPD-AO là $\max \left\{ O(|\mathcal{A}^+| |U^+| |\Delta U|), O\left((|C| - |\mathcal{A}^+|)^2 |U^+| |\Delta U|\right) \right\}$ được thực hiện qua ba giai đoạn chính. Trong giai đoạn đầu tiên, thuật toán tiến hành

kiểm tra các đối tượng bổ sung. 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. Ở giai đoạn thứ ba, thuật toán loại bỏ các thuộc tính dư thừa trong rút gọn thu được.

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

Mệnh đề 2.7. 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 $U^- = U \setminus \Delta U$, với $\Delta U = \{u_n, u_{n-1}, \dots, u_{n-z+1}\}$ là tập đối tượng được loại bỏ khỏi bảng, trong đó $z \geq 1$, được xác định như sau:

$$\ddot{D}(U^-/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U^-/\ddot{\mathcal{R}}_{C \cup D}^{\alpha,\beta}) = \frac{n^2 \ddot{D}(U/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C \cup D}^{\alpha,\beta})}{(n-z)^2} - 2 \sum_{i=1}^z \frac{\left(\left| [\ddot{u}_{n-i+1}]_C^{\alpha,\beta} \right| - \left| [\ddot{u}_{n-i+1}]_C^{\alpha,\beta} \cap [\ddot{u}_{n-i+1}]_D^{\alpha,\beta} \right| - \xi_i \right)}{(n-z)^2} \quad (2.7)$$

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

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

Mệnh đề 2.8. Giả sử rằng $A \subseteq C$ là một rút gọn trên U và một tập đối tượng $\Delta U = \{u_n, u_{n-1}, \dots, u_{n-z+1}\}$ được loại bỏ khỏi bảng, trong đó $z \geq 1$ và $U^- = U \setminus \Delta U$, khi đó:

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

$$\ddot{D}(U^-/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U^-/\ddot{\mathcal{R}}_{C \cup D}^{\alpha,\beta}) = \frac{n^2 \ddot{D}(U/\ddot{\mathcal{R}}_C^{\alpha,\beta}, U/\ddot{\mathcal{R}}_{C \cup D}^{\alpha,\beta})}{(n-z)^2} - 2 \sum_{i=1}^z \frac{\left(\left| [\ddot{u}_{n-i+1}]_C^{\alpha,\beta} \right| - \left| [\ddot{u}_{n-i+1}]_C^{\alpha,\beta} \cap [\ddot{u}_{n-i+1}]_D^{\alpha,\beta} \right| \right)}{(n-z)^2} \quad (2.8)$$

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

Trên các cơ sở được trình bày, luận án đề xuất Thuật toán IARPD-RO với độ phức tạp là $O(|A^-| |U^-| |\Delta U|)$, bao gồm hai giai đoạn chính. Trong giai đoạn đầu tiên, thuật toán sẽ cập nhật các phân hoạch mờ trực cảm mức *alpha*, *beta* của tập thuộc tính C cùng với các thuộc tính điều kiện trong bảng quyết định mới. 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.

2.3. Thử nghiệm và đánh giá các thuật toán đề xuất

2.3.1 Hiệu năng của thuật toán IARPD-AO

Quá trình thực nghiệm được thực hiện trên một số bộ dữ liệu được mô tả trong Bảng 2.1. 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.

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 $U = \{u_1, u_2, \dots, u_n\}$, rút gọn $\mathcal{A} \subseteq C$, các phân hoạch $U/\tilde{\mathcal{R}}_{\mathcal{A}}^{\alpha, \beta}$, $U/\tilde{\mathcal{R}}_C^{\alpha, \beta}$ và tập đối tượng loại bỏ $\Delta U = \{u_{n+1}, u_{n+2}, \dots, u_{n+z}\}$.

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 $U^-/\tilde{\mathcal{R}}_{\mathcal{A}}^{\alpha, \beta}$, $U^-/\tilde{\mathcal{R}}_C^{\alpha, \beta}$ and $U^-/\tilde{\mathcal{R}}_{\{a\}}^{\alpha, \beta}$, với mọi $a \in C \setminus \mathcal{A}$
- 3: tính toán $\tilde{D}(U^-/\tilde{\mathcal{R}}_{\mathcal{A}^-}^{\alpha, \beta}, U^-/\tilde{\mathcal{R}}_{\mathcal{A}^- \cup D}^{\alpha, \beta})$ và $\tilde{D}(U^-/\tilde{\mathcal{R}}_C^{\alpha, \beta}, U^-/\tilde{\mathcal{R}}_{C \cup D}^{\alpha, \beta})$ bởi Công thức 2.7.

// tìm kiếm rút gọn

- 4: **for** $a \in \mathcal{A}^-$ **do**
- 5: tính toán $\tilde{D}(U^-/\tilde{\mathcal{R}}_{\mathcal{A}^- \setminus \{a\}}^{\alpha, \beta}, U^-/\tilde{\mathcal{R}}_{\mathcal{A}^- \setminus \{a\} \cup D}^{\alpha, \beta})$ theo Công thức 2.5.
- 6: **if** $\tilde{D}(U^-/\tilde{\mathcal{R}}_{\mathcal{A}^- \setminus \{a\}}^{\alpha, \beta}, U^-/\tilde{\mathcal{R}}_{\mathcal{A}^- \setminus \{a\} \cup D}^{\alpha, \beta}) = \tilde{D}(U^-/\tilde{\mathcal{R}}_{\mathcal{A}^-}^{\alpha, \beta}, U^-/\tilde{\mathcal{R}}_{\mathcal{A}^- \cup D}^{\alpha, \beta})$ **then**
- 7: $\mathcal{A}^- \leftarrow \mathcal{A}^- \setminus \{a_0\}$
- 8: **end if**
- 9: **end for**
- 10: **return** $\mathcal{A}^-$

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

Kích thước rút gọn và thời gian xử lý của các thuật toán khi xử lý trên U_{ori} được biểu diễn trong Bảng 2.2. 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.

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	
	thời gian	$ \mathcal{A} $	thời gian	$ \mathcal{A} $	thời gian	$ \mathcal{A} $	time	$ \mathcal{A} $	thời gian	$ \mathcal{A} $	thời gian	$ \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

Tuy nhiên, nếu xét trong không gian tập mờ trực cảm, 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 với ARIFPD, FMIFRFS và IFPR lần lượt là 29.545 giây, 15.172 giây và 18.371 giây. Các thuật toán theo mô hình tập mờ trực cảm thường thu được các rút gọn có kích thước nhỏ hơn so với các thuật toán theo mô hình tập thô mờ. Kích thước trung bình của các rút gọn từ thuật toán đề xuất là nhỏ nhất. Từ Bảng 2.3, hiệu quả phân lớp của ARPD cũng được xem là vượt trội so với các thuật toán còn lại.

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 lần lượt vào U_{ori} để đánh giá tập rút gọn thu

được từ các thuật toán gia tăng. Cụ thể, thuật toán IARPD-AO sẽ được so sánh với một số thuật toán gia tăng dựa trên tập thô mờ (IF-FDAR-AO [81], AIFSA-FKD [82]) và tập mờ trực cảm (ARIFPD-AO [91]). Các kết quả được phân tích dựa trên quá trình xử lý các thuật toán trong năm giai đoạn bổ sung tập đối tượng.

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

- *Thời gian xử lý*: 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 kết quả 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. 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 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.

- *Kích thước rút gọn*: 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 rút gọn trung bình là nhỏ nhất trong cả bốn thuật toán. Đặc biệt, trên các bộ dữ liệu có độ chính xác phân lớp ban đầu thấp, chẳng hạn như Movement, Hill-valley và Wall, số lượng thuộc tính thu được từ IARPD-AO là rất nhỏ. Điều này chỉ ra rằng, IARPD-AO có khả năng loại bỏ các thuộc tính dư thừa từ dữ liệu nhiễu tốt hơn so với các thuật toán khác.

- *Độ chính xác phân lớp*: Các rút gọn thu được từ các thuật toán dựa trên mô hình tập mờ trực cảm đạt độ chính xác cao hơn so với các thuật toán theo tiếp cận mô hình tập thô mờ trên phần lớn các bước gia tăng. Điều này nhấn mạnh tầm quan trọng của thành phần độ không thuộc trong việc điều chỉnh các độ đo. Qua đó, các thuật toán có thể lựa chọn các thuộc tính cần thiết và đạt hiệu quả phân lớp cao. Kết quả này cũng dẫn tới sự cạnh tranh rõ rệt giữa thuật toán đề xuất và thuật toán ARIFPD-AO. Cụ thể, trong toàn bộ các trường hợp, thuật toán đề xuất đạt độ chính xác cao nhất ở 31 trường hợp, trong khi ARIFPD-AO chỉ đạt 13 trường hợp.

2.3.2 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. Quá trình này được tiến hành trên những bộ dữ liệu tiêu chuẩn, được mô tả trong Bảng 2.4. 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 , lần lượt được loại bỏ khỏi U . Theo Bảng 2.5, 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.

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 toán toàn FMIFRFS, ARIFPD và IFPR thời gian thực thi lần lượt là 52.252 giây, 57.514 giây và 63.478 giây.

Bảng 2.4: 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

Bảng 2.5: Kích thước rút gọn, thời gian thực thi của ARPD và một số thuật toán trên U.

Tập dữ liệu	ARPD		FMIFRFS		ARIFPD		IFPR		NIFS		F-FDAR	
	<i>Time</i>	$ \mathcal{A} $	<i>Time</i>	$ \mathcal{A} $	<i>Time</i>	$ \mathcal{A} $	<i>Time</i>	$ \mathcal{A} $	<i>Time</i>	$ \mathcal{A} $	<i>Time</i>	$ \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

Bảng 2.6: 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

Từ Bảng 2.6, ARPD đạt độ chính xác phân lớp trung bình cao nhất và 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. 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 hiệu năng từ kết quả của ba thuật toán IARPD-DO, IF-FDAR-DO [81] và AIFSD-FKD [82]. Các kết quả rút gọn cũng tương tự như với thuật toán IARPD-AO khi thuật toán đề xuất đều chứng minh được hiệu quả về thời gian thực thi, kích thước rút gọn và độ chính xác phân lớp trên các bước loại bỏ.

2.4. Kết luận Chương 2

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). Một số kết quả thực nghiệm đã cho thấy hiệu năng của thuật toán đề xuất là vượt trội so với các thuật toán dựa trên mô hình tập thô mờ và tập thô mờ trực cảm. Bên cạnh đó, để xử lý các tình huống thực tế của dữ liệu, hai công thức tính khoảng cách phân hoạch mờ trực cảm mức α , β cũng được phát triển, làm nền tảng cho việc đề xuất các thuật toán gia tăng xử lý trên các bảng quyết định có sự bổ sung hoặc loại bỏ tập đối tượng.

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

Chương 3 tập trung đề xuất một số giải pháp mới với các đóng góp chính như sau: (1) 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, (2) Xây dựng độ đo khoảng cách họ lân cận mờ trực cảm có trọng số và định nghĩa lại một rút gọn hiệu quả, (3) Xây dựng độ quan trọng của thuộc tính làm cơ sở cho việc lựa chọn các thuộc tính có ý nghĩa cao, (4) Đề xuất thuật toán rút gọn thuộc tính trên bảng quyết định cố định theo hướng tiếp cận lọc, (5) Phát triển hai công thức tính toán gia tăng dựa trên độ đo khoảng cách họ lân cận mờ trực cảm có trọng số cho các trường hợp bảng quyết định thay đổi, (6) Đề xuất hai thuật toán gia tăng rút gọn thuộc tính trên bảng quyết định thay đổi, (7) 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, 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

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

Nghiên cứu ban đầu 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 1.3, trong trường hợp $p = \infty$,

$$- \eta_{[\ddot{u}]_A^{\delta, \omega}}(u_i) = \frac{1 - \gamma_{[\ddot{u}]_A^{\delta, \omega}}(u_i)}{1 + w^o \times \gamma_{[\ddot{u}]_A^{\delta, \omega}}(u_i)}, \text{ với } w^o = \frac{1}{|U|} \sum_{u \in U} \frac{|[u]_C^{\delta, \omega} \cap [u]_D|}{|[u]_C^{\delta, \omega}|} \text{ là tỉ lệ phân lớp trung}$$

bình dựa trên các hạt thông tin lân cận trọng số của C .

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 $[\ddot{u}]_A^{\delta, \omega} = \left\{ [\ddot{u}]_A^{\delta, \omega}(u_1), \dots, [\ddot{u}]_A^{\delta, \omega}(u_{|U|}) \right\}$, 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 .

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

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

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

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

Mệnh đề 3.1. Cho $IS = (U, C \cup D)$ và $A, B \subseteq C$. Nếu $p = \infty$, thì $[\ddot{u}]_{A \cup B}^{\delta, \omega} = [\ddot{u}]_A^{\delta, \omega} \cap [\ddot{u}]_B^{\delta, \omega}$, với mọi $u \in U$.

3.2.2 Một số tính chất của IFWNRs

Mệnh đề 3.2. 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 $\overline{IW}_A(X) = \overline{WN}_A(X)$.

Mệnh đề 3.2 chỉ ra rằng, WNRs là một trường hợp đặc biệt của IFWNRs.

Mệnh đề 3.3. 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 $\forall X \subseteq U, \underline{IW}_A(X) \subseteq \underline{IW}_B(X)$.

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

Mệnh đề 3.4. 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:

$$\ddot{D}(\mathcal{F}_C^{\delta, \omega}, \mathcal{F}_{C \cup D}^{\delta, \omega}) = \sum_{u \in U} \frac{\left(\left| [\ddot{u}]_C^{\delta, \omega} \right| - \left| [\ddot{u}]_C^{\delta, \omega} \cap [\ddot{u}]_D^{\delta, \omega} \right| \right)}{|U|^2} \quad (3.3)$$

Mệnh đề 3.5. 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 $\ddot{D}(\mathcal{F}_A^{\delta_1, \omega}, \mathcal{F}_{A \cup D}^{\delta_1, \omega}) \leq \ddot{D}(\mathcal{F}_A^{\delta_2, \omega}, \mathcal{F}_{A \cup D}^{\delta_2, \omega})$.

Mệnh đề 3.6. Cho bảng quyết định $IS = (U, C \cup D)$ với hai tập con thuộc tính $A, B \subseteq C$. Nếu $A \subseteq B$ thì $\ddot{D}(\mathcal{F}_A^{\delta, \omega}, \mathcal{F}_{A \cup D}^{\delta, \omega}) \geq \ddot{D}(\mathcal{F}_B^{\delta, \omega}, \mathcal{F}_{B \cup D}^{\delta, \omega})$.

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 $\ddot{D}$ -rút gọn của C nếu thỏa mãn

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

Định nghĩa 3.2. 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

$$Sig_F(a, A) = \ddot{D}(\mathcal{F}_{A \setminus \{a\}}^{\delta, \omega}, \mathcal{F}_{A \setminus \{a\} \cup D}^{\delta, \omega}) - \ddot{D}(\mathcal{F}_A^{\delta, \omega}, \mathcal{F}_{A \cup D}^{\delta, \omega}) \quad (3.4)$$

Trên cơ sở được trình bày, luận án đề xuất Thuật toán ARIFW trong bảng mã giả 3.1 nhằm tìm kiếm một rút gọn trên bảng quyết định với độ phức tạp là $O(|C|^2|U|^2)$.

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 1.2
 - 2: tính toán $\ddot{D}(\mathcal{F}_C^{\delta, \omega}, \mathcal{F}_{C \cup D}^{\delta, \omega})$
 - 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 $\ddot{D}(\mathcal{F}_{\{a\}}^{\delta, \omega}, \mathcal{F}_{\{a\} \cup D}^{\delta, \omega})$
 - 6: **end for**
 - 7: $\mathcal{B} = \{a_0\}$ thỏa mãn $\ddot{D}(\mathcal{F}_{\{a_0\}}^{\delta, \omega}, \mathcal{F}_{\{a_0\} \cup D}^{\delta, \omega}) = \min_{a \in C} \ddot{D}(\mathcal{F}_{\{a\}}^{\delta, \omega}, \mathcal{F}_{\{a\} \cup D}^{\delta, \omega})$
 - 8: **while** $\ddot{D}(\mathcal{F}_{\mathcal{B}}^{\delta, \omega}, \mathcal{F}_{\mathcal{B} \cup D}^{\delta, \omega}) > \ddot{D}(\mathcal{F}_C^{\delta, \omega}, \mathcal{F}_{C \cup D}^{\delta, \omega})$ **do**
 - 9: tính toán $Sig_F(a, \mathcal{B})$, với mọi $a \in C \setminus \mathcal{B}$
 - 10: lựa chọn a_0 thỏa mãn $Sig_F(a_0, \mathcal{B}) = \max_{a \in C \setminus \mathcal{B}} \{Sig_F(a, \mathcal{B})\}$
 - 11: $\mathcal{B} \leftarrow \mathcal{B} \cup \{a_0\}$
 - 12: **end while**
 - 13: **return** $\mathcal{B}$
-

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

Giả sử rằng một tập đối tượng mới $\Delta U = (u_{n+1}, u_{n+2}, \dots, u_{n+z})$ 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 $[u_i]_{\{a\}}^{\delta, \omega}$ được trình bày như sau

$$\omega^+ = \begin{cases} \omega & \text{nếu } 1 \leq i, j \leq n \\ \omega^\Delta & \text{nếu ngược lại} \end{cases} \quad (3.5)$$

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 ΔU . Khi đó, cơ chế cập nhật hạt thông tin lân cận mờ trực cảm có trọng số dựa trên công thức sau đây.

$$[\ddot{u}_i]_{\{a\}}^{\delta, \omega^+}(u_j) = \begin{cases} [\ddot{u}_i]_a^{\delta, \omega}(u_j) & \text{if } 1 \leq i, j \leq n \\ [\ddot{u}_i]_a^{\delta, \omega^\Delta}(u_j) & \text{if } u_j \in [u_i]_{\{a\}}^{\delta, \omega^\Delta} \quad (n+1 \leq i \leq n+z) \vee (n+1 \leq j \leq n+z) \\ (0, 1) & \text{trường hợp còn lại} \end{cases} \quad (3.6)$$

Khi các hạt thông tin lân cận mờ trực cảm có trọng số được hình thành sẽ thu được một họ lân cận mờ trực cảm có trọng số mới trên thuộc tính a , ký hiệu là $\mathcal{F}_{\{a\}}^{\delta, \omega^+}$.

Mệnh đề 3.7. Cho bảng quyết định $IS = (U, C \cup D)$ với $U = (u_1, u_2, \dots, u_n)$ và tập đối tượng mới $\Delta U = (u_{n+1}, u_{n+2}, \dots, u_{n+z})$, trong đó $z \geq 1$. 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 $\ddot{D}(\mathcal{F}_C^{\delta, \omega^+}, \mathcal{F}_{C \cup D}^{\delta, \omega^+})$ được tính bởi

$$\ddot{D}(\mathcal{F}_C^{\delta, \omega^+}, \mathcal{F}_{C \cup D}^{\delta, \omega^+}) = \frac{n^2 \ddot{D}(\mathcal{F}_C^{\delta, \omega}, \mathcal{F}_{C \cup D}^{\delta, \omega}) + 2 \sum_{i=1}^z \left(\left| [\ddot{u}_{n+i}]_C^{\delta, \omega^+} \right| - \left| [\ddot{u}_{n+i}]_C^{\delta, \omega^+} \cap [\ddot{u}_{n+i}]_D^{\delta, \omega^+} \right| - \theta_i \right)}{(n+z)^2} \quad (3.7)$$

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 3.8. Cho bảng quyết định $IS = (U, C \cup D)$ với $U = (u_1, u_2, \dots, u_n)$, $A \subseteq C$ 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 $\Delta U = (u_{n+1}, u_{n+2}, \dots, u_{n+z})$, 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:

$$\ddot{D}(\mathcal{F}_C^{\delta, \omega^+}, \mathcal{F}_{C \cup D}^{\delta, \omega^+}) = \frac{n^2 \ddot{D}(\mathcal{F}_C^{\delta, \omega}, \mathcal{F}_{C \cup D}^{\delta, \omega}) + 2 \sum_{i=1}^z \left(\left| [\ddot{u}_{n+i}]_C^{\delta, \omega^+} \right| - \left| [\ddot{u}_{n+i}]_C^{\delta, \omega^+} \cap [\ddot{u}_{n+i}]_D^{\delta, \omega^+} \right| \right)}{(n+z)^2} \quad (3.8)$$

2. Nếu $[\ddot{u}_{n+i}]_B^{\delta, \omega^+} \subseteq [\ddot{u}_{n+i}]_D^{\delta, \omega^+}$ với $i = 1, 2, \dots, z$ thì $\ddot{D}(\mathcal{F}_A^{\delta, \omega^+}, \mathcal{F}_{A \cup D}^{\delta, \omega^+}) = \ddot{D}(\mathcal{F}_C^{\delta, \omega^+}, \mathcal{F}_{C \cup D}^{\delta, \omega^+})$.

Độ phức tạp của IARIF-AO là $\max \left\{ O(|\mathcal{B}^+| |U^+| |\Delta U|), O((|C| - |\mathcal{B}^+|)^2 |U^+| |\Delta U|) \right\}$

với bốn giai đoạn chính là khởi tạo và cập nhật, kiểm tra tập đối tượng bổ sung, lọc tìm kiếm rút gọn và loại bỏ các thuộc tính dư thừa.

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

Mệnh đề 3.9. Cho bảng quyết định $IS = (U, C \cup D)$ với $U = \{u_1, u_2, \dots, u_n\}$ và $\Delta U = \{u_n, u_{n-1}, \dots, u_{n-z+1}\}$ là một tập đối tượng 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 $U^- = U \setminus \Delta U$, ký hiệu là $\mathcal{D}(\mathcal{F}_C^{\delta, \omega^-}, \mathcal{F}_{C \cup D}^{\delta, \omega^-})$, được xác định như sau:

$$\mathcal{D}(\mathcal{F}_C^{\delta, \omega^-}, \mathcal{F}_{C \cup D}^{\delta, \omega^-}) = \frac{n^2 \mathcal{D}(\mathcal{F}_C^{\delta, \omega}, \mathcal{F}_{C \cup D}^{\delta, \omega}) - 2 \sum_{i=1}^z \left(\left| [\ddot{u}_{n-i+1}]_C^{\delta, \omega} \right| - \left| [\ddot{u}_{n-i+1}]_C^{\delta, \omega} \cap [\ddot{u}_{n-i+1}]_D^{\delta, \omega} \right| - \zeta_i \right)}{(n-z)^2} \quad (3.9)$$

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

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

Mệnh đề 3.10. Cho bảng quyết định $IS = (U, C \cup D)$ với $U = \{u_1, u_2, \dots, u_n\}$, $A \subseteq C$ là một rút gọn dựa vào khoảng cách giữa hai họ lân cận trọng số mờ trực cảm trên U và

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 $U = \{u_1, u_2, \dots, u_n\}$, rút gọn $\mathcal{B} \subseteq C$, $\mathcal{F}_B^{\delta, \omega}$, $\mathcal{F}_C^{\delta, \omega}$ và tập đối tượng bổ sung $\Delta U = \{u_{n+1}, u_{n+2}, \dots, u_{n+z}\}$.

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

// khởi tạo và cập nhật

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

2: cập nhật ω^+ từ Công thức 3.5 và $\mathcal{F}_{\{a\}}^{\delta, \omega^+}$ trên U^+ với mỗi $a \in C \setminus \mathcal{B}^+$ từ Công thức 3.6.

// kiểm tra tập đối tượng bổ sung

3: $Z \leftarrow \Delta U$

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

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

6: **end for**

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

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

// tìm kiếm rút gọn

9: tính toán $\ddot{D}(\mathcal{F}_{\mathcal{B}^+}^{\delta, \omega^+}, \mathcal{F}_{\mathcal{B}^+ \cup D}^{\delta, \omega^+})$ và $\ddot{D}(\mathcal{F}_C^{\delta, \omega^+}, \mathcal{F}_{C \cup D}^{\delta, \omega^+})$ theo Công thức 3.2.

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

11: tính toán $Sig_F(a, \mathcal{B}^+)$, với mọi $a \in C \setminus \mathcal{B}^+$ theo Công thức 3.4.

12: lựa chọn thuộc tính a_0 thỏa mãn: $Sig_F(a_0, \mathcal{B}^+) = \max_{a \in C \setminus \mathcal{B}^+} \{Sig_F(a, \mathcal{B}^+)\}$

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

14: **end while**

// loại bỏ các thuộc tính dư thừa

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

16: tính toán $\ddot{D}(\mathcal{F}_{\mathcal{B}^+ \setminus \{a\}}^{\delta, \omega^+}, \mathcal{F}_{\mathcal{B}^+ \setminus \{a\} \cup D}^{\delta, \omega^+})$ theo Công thức 3.3.

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

18: **end for**

19: **return** $\mathcal{B}^+$

một tập đối tượng $\Delta U = \{u_n, u_{n-1}, \dots, u_{n-z+1}\}$ được loại bỏ khỏi bảng, trong đó $z \geq 1$ và $U^- = U \setminus \Delta U$ là một tập đối tượng còn lại, khi đó:

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

$$\ddot{D}(\mathcal{F}_C^{\delta, \omega^-}, \mathcal{F}_{C \cup D}^{\delta, \omega^-}) = \frac{n^2 \mathcal{D}(\mathcal{F}_C^{\delta, \omega^-}, \mathcal{F}_{C \cup D}^{\delta, \omega^-}) - 2 \sum_{i=1}^z \left(\left| [\ddot{u}_{n-i+1}]_C^{\delta, \omega^-} \right| - \left| [\ddot{u}_{n-i+1}]_C^{\delta, \omega^-} \cap [\ddot{u}_{n-i+1}]_D^{\delta, \omega^-} \right| \right)}{(n-z)^2} \quad (3.10)$$

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

Độ phức tạp của Thuật toán IDARIFW trong bảng mã giả 3.3 là $O(|\mathcal{B}^-| |U^-| |\Delta U|)$. 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

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 $U = \{u_1, u_2, \dots, u_n\}$, rút gọn $\mathcal{B} \subseteq C$, $\mathcal{F}_B^{\delta, \omega}$ và $\mathcal{F}_C^{\delta, \omega}$, tập đối tượng bị loại bỏ $\Delta U = \{u_n, u_{n-1}, \dots, u_{n-z+1}\}$

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

- 1: khởi tạo $\mathcal{B}^- \leftarrow \mathcal{B}$
- 2: cập nhật $\mathcal{F}_C^{\delta, \omega^-}$
- 3: tính toán $\ddot{D}(\mathcal{F}_C^{\delta, \omega^-}, \mathcal{F}_{C \cup D}^{\delta, \omega^-})$ theo Công thức 3.9.
- 4: $\ddot{D}(\mathcal{F}_{\mathcal{B}^-}^{\delta, \omega^-}, \mathcal{F}_{\mathcal{B}^- \cup D}^{\delta, \omega^-}) \leftarrow \ddot{D}(\mathcal{F}_C^{\delta, \omega^-}, \mathcal{F}_{C \cup D}^{\delta, \omega^-})$
- 5: **for** $a \in \mathcal{B}^-$ **do**
- 6: tính toán $\ddot{D}(\mathcal{F}_{\mathcal{B}^- \setminus \{a\}}^{\delta, \omega^-}, \mathcal{F}_{\mathcal{B}^- \setminus \{a\} \cup D}^{\delta, \omega^-})$ theo Công thức 3.3.
- 7: **if** $\ddot{D}(\mathcal{F}_{\mathcal{B}^- \setminus \{a\}}^{\delta, \omega^-}, \mathcal{F}_{\mathcal{B}^- \setminus \{a\} \cup D}^{\delta, \omega^-}) = \ddot{D}(\mathcal{F}_{\mathcal{B}^-}^{\delta, \omega^-}, \mathcal{F}_{\mathcal{B}^- \cup D}^{\delta, \omega^-})$ **then** $\mathcal{B}^- \leftarrow \mathcal{B}^- \setminus \{a\}$
- 8: **end for**
- 9: **return** $\mathcal{B}^-$

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. 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.

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

Bảng 3.2 trình bày kích thước các rút gọn và các tham số tương ứng của từng thuật toán. 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

đơn giản để lựa chọn các thuộc tính quan trọng. Ngược lại, 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.

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} $	δ	$ \mathcal{B} $	δ	$ \mathcal{B} $	ϵ	$ \mathcal{B} $	α	$ \mathcal{B} $	δ	$ \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 3.3: 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

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, 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. Có thể thấy 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.

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	FMIFRS	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	<i>best</i>

Tiếp theo, thuật toán IARIF-AO sẽ được so sánh với các thuật toán AIFSA-FKD [82], W-MGMA [86] và IARPD-AO. 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} . Một số kết quả được phân tích như sau:

- *Thời gian thực thi*: 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 lớn hơn. Khi xem xét các thuật toán dựa trên mô hình IFRSs, 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 trọng số, giúp thu hẹp không gian tính toán.

- *Kích thước rút gọn*: Trong các trường hợp, các rút gọn thu được bởi IARPD-AO và IARIF-AO có kích thước nhỏ nhất lần lượt 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.

- *Độ chính xác phân lớp*: 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. 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 và làm giảm hiệu quả của mô hình phân lớp. 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.

3.4.2 Hiệu năng của thuật toán IARIF-DO

Phần này 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 3.5.

Bảng 3.5: 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.

Bảng 3.6 trình bày kích thước các rút gọn thu được và các tham số tương ứng của từng thuật toán. 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 các trường hợp, hai thuật toán ARPD và ARIFW thường thu được các tập rút gọn có kích thước nhỏ nhất. Tuy nhiên, kích thước rút gọn trung bình của ARIFW là nhỏ hơn so với ARPD.

Bảng 3.6: Kích thước rút gọn và tham số của các thuật toán với KNN trên U

Tập dữ liệu	NIFS		WNRS		W-MGMN		FMIFRFS		ARPĐ		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

Bảng 3.7 trình bày thời gian thực thi của các thuật toán rút gọn thuộc tính trên bảng quyết định cố định. 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.

Bảng 3.7: Thời gian thực thi của ARIFW với một số thuật toán khác trên U.

TT	Tập dữ liệu	NIFS	WNRS	W-MGMN	FMIFRFS	ARPĐ	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

Từ Bảng 3.8, 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.

Bảng 3.8: 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 $\pm$ 6.34	59.41 $\pm$ 5.55	69.12 $\pm$ 3.01	67.94 $\pm$ 4.64	61.47 $\pm$ 5.33	61.47 $\pm$ 6.76	69.41 $\pm$ 4.78
Ionosphere	83.76 $\pm$ 6.39	84.89 $\pm$ 6.02	88.90 $\pm$ 6.89	86.04 $\pm$ 5.34	90.33 $\pm$ 5.27	88.32 $\pm$ 7.61	90.62 $\pm$ 6.08
Vehicle	70.45 $\pm$ 1.85	69.39 $\pm$ 3.04	69.86 $\pm$ 2.02	70.57 $\pm$ 2.75	70.10 $\pm$ 2.91	70.22 $\pm$ 2.83	71.29 $\pm$ 3.19
Vowel	51.52 $\pm$ 12.38	53.03 $\pm$ 11.65	64.85 $\pm$ 5.8	52.02 $\pm$ 11.3	53.54 $\pm$ 6.94	52.12 $\pm$ 6.94	64.85 $\pm$ 5.8
German	73.00 $\pm$ 2.79	69.40 $\pm$ 4.43	69.60 $\pm$ 3.9	69.10 $\pm$ 2.88	73.50 $\pm$ 2.94	71.50 $\pm$ 2.77	74.70 $\pm$ 2.97
LSVT	83.97 $\pm$ 14.31	80.06 $\pm$ 9.07	89.68 $\pm$ 6.39	86.47 $\pm$ 5.07	86.90 $\pm$ 9.95	90.58 $\pm$ 7.5	91.15 $\pm$ 9.39
Leukemia	84.46 $\pm$ 13.51	90.36 $\pm$ 6.35	98.57 $\pm$ 10	91.43 $\pm$ 9.48	98.57 $\pm$ 4.29	98.57 $\pm$ 4.29	98.57 $\pm$ 4.29
Mfeat	80.70 $\pm$ 1.36	81.50 $\pm$ 1.99	80.15 $\pm$ 1.52	82.15 $\pm$ 1.64	81.65 $\pm$ 1.99	83.15 $\pm$ 1.69	82.60 $\pm$ 1.84
PD	79.22 $\pm$ 6.07	74.18 $\pm$ 7.02	84.25 $\pm$ 5.89	78.44 $\pm$ 4.34	79.63 $\pm$ 4.58	79.76 $\pm$ 3.61	82.52 $\pm$ 3.98
Segmentation	95.28 $\pm$ 1.52	95.41 $\pm$ 1.5	95.71 $\pm$ 1.88	95.19 $\pm$ 1.32	95.63 $\pm$ 1.67	95.71 $\pm$ 1.53	96.32 $\pm$ 1.44
Wall	77.26 $\pm$ 5.91	77.27 $\pm$ 6.36	78.41 $\pm$ 5.35	75.81 $\pm$ 5.88	81.56 $\pm$ 3.78	83.14 $\pm$ 4.01	83.32 $\pm$ 4.56
Waveform2	80.08 $\pm$ 1.46	80.44 $\pm$ 1.51	78.38 $\pm$ 1.69	83.24 $\pm$ 1.18	82.86 $\pm$ 1.44	81.82 $\pm$ 1.54	82.58 $\pm$ 1.06
Trung bình	76.69 $\pm$ 6.16	76.28 $\pm$ 5.37	80.62 $\pm$ 4.53	79.06 $\pm$ 7.87	78.20 $\pm$ 4.65	79.70 $\pm$ 4.26	82.33 $\pm$ 4.12
Win/Draw/Loss	12/0/0	12/0/0	9/2/1	11/0/1	10/1/1	10/1/1	<i>best</i>

Cuối cùng, luận án trình bày một số so sánh về hiệu quả của thuật toán IARIF-DO với các thuật toán gia tăng AIFSD-FKD [82], W-MGMD [86] và IARPD-RO. Các thuật toán sẽ xử lý trên tập dữ liệu khi loại bỏ lần lượt các tập đối tượng từ U_1 đến U_5 từ U . Các kết quả cũng tương tự như đối với thuật toán IARIF-AO khi thuật toán đề xuất cho thấy khả năng tìm kiếm các rút gọn có kích thước nhỏ với độ chính xác phân lớp vượt trội.

3.5. Kết luận Chương 3

Chương này trình bày mô hình tập thô lân cận mờ trực cảm có trọng số (IFWNRSs) cùng một số tính chất quan trọng của mô hình. Từ mô hình này, độ đo khoảng cách giữa hai họ lân cận mờ trực cảm có trọng số đã được xây dựng làm cơ sở cho việc thiết kế một thuật toán rút gọn thuộc tính trên các bảng quyết định cố định. Để giải quyết các kịch bản dữ liệu thực tế, Chương 3 cũng 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 đề xuất đạt được độ chính xác phân lớp vượt trội trên phần lớn các bộ dữ liệu so với các phương pháp dựa trên tập thô mờ và tập thô lân cận trọng số, đồng thời cũng giảm thời gian xử lý so với các phương pháp dựa trên tập thô mờ trực cảm.

Có thể nhận thấy, IFWNRSs đã khắc phục được những hạn chế còn tồn tại trong một số mô hình thuộc nhánh mở rộng thứ nhất của lý thuyết tập thô. Trên nền tảng đó, mô hình đề xuất được xem như một công cụ hữu hiệu, giúp giải quyết hiệu quả bài toán rút gọn thuộc tính trên nhiều loại dữ liệu khác nhau, đồng thời có tính khả thi cao khi ứng dụng trong thực tiễn, đặc biệt trong bối cảnh dữ liệu lớn.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

A. Những kết quả chính của luận án

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.
- 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 tính 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.

B. Hướng phát triển tiếp theo của luận án

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 tương lai, 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 CÁC BÀI BÁO ĐÃ XUẤT BẢN LIÊN QUAN ĐẾN LUẬN ÁN

A. Các công trình đã công bố

1. Pham Viet Anh, Nguyen Ngoc Thuy, Vu Duc Thi, and 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).
2. Pham Viet Anh, Nguyen Ngoc Thuy, Le Hoang Son, Tran Hung Cuong, and 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).
3. Nguyen Long Giang, Pham Viet Anh, Janos Demetrovics, and 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.
4. Phạm Việt Anh, Nguyễn Long Giang, Nguyễn Ngọc Thủy, Nguyễn Thế Thủy, 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.
5. 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.
6. Pham Viet Anh, Nguyen Long Giang, Nguyen Ngoc Thuy, Le Van Dung, and 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

7. 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.