BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Phạm Minh Ngọc Hà

NGHIÊN CỨU MỘT SỐ PHƯƠNG PHÁP NÂNG CAO HIỆU QUẢ TÍNH TOÁN TẬP RÚT GỌN TRÊN KHÔNG GIAN XẤP XỈ MỜ

TÓM TẮT LUẬN ÁN TIẾN SĨ HỆ THỐNG THÔNG TIN

Mã số: 9 48 01 04

Hà Nội – 2024

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 1: PGS.TS Nguyễn Long Giang, Viện Công nghệ Thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, Hà Nội, Việt Nam
2. Người hướng dẫn 2: TS Nguyễn Mạnh Hùng, Học viện Kĩ thuật Quân sự, Hà Nội, Việt Nam

Phản biện 1: .....

Phản biện 2: .....

Phản biện 3:.....

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

Tính cấp thiết của đề tài luận án

Chọn lọc thuộc tính, còn được gọi là chọn lọc đặc trưng, là một bước quan trọng trong phân tích dữ liệu và học máy thống kê. Quá trình này bao gồm việc lựa chọn một tập con các thuộc tính có liên quan từ tập thuộc tính ban đầu, sao cho thông tin quan trọng được giữ lại một cách tối đa. Chọn lọc thuộc tính mang lại nhiều lợi ích đáng kể: 1) giảm độ phức tạp tính toán, 2) cải thiện khả năng diễn giải mô hình, và 3) nâng cao khả năng dự đoán. Mục tiêu chính là tìm ra một tập con các đặc trưng, từ tập đặc trưng ban đầu, mà vẫn đảm bảo bảo toàn thông tin hoặc khả năng đưa ra quyết định chính xác. Các ứng dụng quan trọng của chọn lọc thuộc tính xuất hiện rộng rãi trong các lĩnh vực như nhận dạng mẫu và khai thác dữ liệu, bao gồm phân loại văn bản [1], [2], xử lý ảnh [3]–[5], và xử lý tiếng nói [6]–[9].

Năm 1982, Pawlak giới thiệu mô hình lý thuyết tập thô (Rough Set - RS) [10], được cộng đồng khoa học đánh giá cao về khả năng phân tích dữ liệu trong các tình huống không đầy đủ và thiếu nhất quán. Nhờ khả năng này, chọn lọc thuộc tính theo tiếp cận RS đã thu hút sự quan tâm của nhiều nhà nghiên cứu trong lĩnh vực lý thuyết tập thô trong nhiều năm qua [4], [11]–[16]. Dựa trên khái niệm không gian xấp xỉ (Approximation Space - AP) của RS, nhiều độ đo đã được đề xuất để định nghĩa reduct và hỗ trợ chọn lọc thuộc tính. Các nghiên cứu gần đây cho thấy rằng, các phương pháp rút gọn thuộc tính theo tiếp cận tập thô truyền thống mang lại nhiều kết quả đáng chú ý trong việc giảm số lượng thuộc tính mà vẫn bảo toàn khả năng phân lớp của bảng quyết định [1], [8], [14]–[17].

Tuy nhiên, tập thô truyền thống chủ yếu phù hợp với các bảng quyết định có miền giá trị rời rạc [18]. Do đó, cần phải rời rạc hóa dữ liệu của các bảng quyết định số (miền giá trị liên tục) trước khi thực hiện chọn lọc thuộc tính. Quá trình này phát sinh thêm chi phí tính toán, có thể làm mất đi tính tự nhiên của dữ liệu, và tiềm ẩn nguy cơ làm mất thông tin quan trọng. Để khắc phục những hạn chế này, các nhà nghiên cứu đã đề xuất mở rộng RS trên không gian xấp xỉ mờ, tạo ra mô hình tập thô mờ (Fuzzy Rough Set - FRS) [19]–[21], và mở rộng RS trên không gian xấp xỉ mờ trực cảm, tạo ra mô hình tập thô mờ trực cảm (Intuitionistic Fuzzy Rough Set - IFRS) [22]. Các mô hình này cho phép xây dựng các phương pháp chọn lọc thuộc tính trực tiếp trên bảng quyết định gốc mà không cần rời rạc hóa.

Không gian xấp xỉ mờ của FRS sử dụng khái niệm quan hệ tương tự (similarity relation) thay cho quan hệ tương đương (equivalence relation) để xây dựng không gian quan hệ giữa các đối tượng. Nhờ đó, quan hệ giữa các đối tượng trong không gian xấp xỉ mờ trở nên mềm dẻo hơn so với quan hệ tương đương truyền thống. Mức độ quan hệ giữa các đối tượng được biểu diễn bằng các giá trị trong khoảng $[0, 1]$ thay vì chỉ 0 hoặc 1 như trong tập thô truyền thống. Hiện nay, việc phát triển các phương pháp chọn lọc thuộc tính dựa trên việc xây dựng độ đo trên không gian xấp xỉ mờ diễn ra rất sôi động, với nhiều độ đo điển hình đã được đề xuất, bao gồm độ đo FPOS [20], [21], [23]–[28], độ đo FIE [4], [29]–[32], và độ đo FD [6], [11], [33]–[35]. Tại Việt Nam, luận án của TS. Nguyễn Văn Thiện đã mở rộng một số độ đo trên không gian xấp xỉ mờ, rút gọn thuộc tính cho bảng quyết định số.

Tuy các nghiên cứu đã chỉ ra rằng các phương pháp chọn lọc thuộc tính theo tiếp cận xây dựng độ đo trên không gian xấp xỉ mờ của FRS hoạt động hiệu quả với các bộ dữ liệu có miền giá trị số, nhưng hiệu quả của chúng có thể giảm khi áp dụng cho các bộ dữ liệu số có độ nhạy cao về tỉ lệ phân nhầm lớp (tức là có nhiều nhiễu). Do đó, mô hình tập thô mờ độ chính xác thay đổi (Variable Precision Fuzzy Rough Sets - VPFRS) [19], [31], [36]–[50] và các độ đo xây dựng trên không gian xấp xỉ mờ trực cảm của mô hình IFRS [13], [20], [22], [41], [46], [51]–[54] đã được đề xuất để giải

quyết vấn đề này.

Theo tiếp cận VPFRS, việc điều chỉnh thành phần trong tập xấp xỉ dưới sẽ ảnh hưởng đến miền dương của thuộc tính, dẫn đến độ phụ thuộc của thuộc tính sẽ thay đổi. Khác với tiếp cận VPFRS, các độ đo xây dựng theo tiếp cận IFRS hoàn toàn phụ thuộc vào không gian xấp xỉ mờ trực cảm. Trong đó, mỗi phần tử của không gian xấp xỉ mờ trực cảm biểu diễn mức độ tương tự và không tương tự giữa hai đối tượng được xét. Do đó, không gian xấp xỉ mờ trực cảm mô tả mối quan hệ giữa các đối tượng đa chiều hơn so với không gian xấp xỉ mờ của FRS [52]. Các công trình nghiên cứu [51] cho thấy tiếp cận IFRS có thể cải thiện chất lượng reduct trên các bộ dữ liệu nhiễu, tuy nhiên thời gian tính toán còn nhiều hạn chế (chi phí tính toán gấp đôi tiếp cận FRS). Tại Việt Nam, luận án TS của tác giả Trần Thanh Đại đề xuất độ đo khoảng cách giữa các phân hoạch mờ trực cảm (Intuitionistic Fuzzy Distance - IFD), chọn lọc thuộc tính cho các bảng quyết định số có chứa nhiễu. Tuy nhiên thời gian tính toán còn hạn chế do việc xác định công thức tính độ thành viên và không thành viên cho AP. Do đó, *mục tiêu nghiên cứu thứ nhất* của luận án là nghiên cứu mở rộng mô hình VPFRS sao cho thời gian tính toán các tập xấp xỉ hiệu quả hơn mô hình VPFRS hiện có. Mục tiêu nghiên cứu thứ nhất thuộc *nhóm các phương pháp chọn lọc thuộc tính cho bảng quyết định tĩnh*, nghĩa là các bảng quyết định có nội dung không thay đổi theo thời gian.

Trong thực tế, các ứng dụng học máy thường xuyên phải cập nhật mô hình để thích ứng với những thay đổi của dữ liệu theo thời gian. Do đó, việc phát triển các phương pháp chọn lọc thuộc tính hiệu quả khi dữ liệu được cập nhật là một yêu cầu cấp thiết [55]. Đến nay, đã có nhiều phương pháp tính toán gia tăng được đề xuất nhằm cập nhật tập rút gọn một cách hiệu quả [3], [6], [55]–[92]. Kỹ thuật tính toán gia tăng này chỉ đánh giá các thông tin mới và kết hợp chúng với kết quả trước đó để cập nhật reduct. Có ba kịch bản thay đổi dữ liệu chính: thay đổi tập thuộc tính [55], [56], [58], [60], [61], thay đổi tập đối tượng [3], [6], [64], [69], [70], [74], [78]–[80], và thay đổi nội dung của đối tượg [3].

Tại Việt Nam, luận án tiến sĩ của Nguyễn Bá Quảng đã đề xuất một phương pháp tính toán gia tăng dựa trên độ đo khoảng cách, được xây dựng dựa trên tính đơn điệu của các phép hợp và giao giữa hai tập hợp. Gần đây, Yang và cộng sự đã đề xuất một phương pháp tính toán gia tăng theo tiếp cận độ đo hạt thông tin tri thức [70]. Không giống như độ đo khoảng cách, độ đo hạt thông tin tri thức dựa trên độ thô và độ mịn của các phân hoạch, giúp công thức xây dựng đơn giản hơn và thời gian tính toán nhanh hơn. Tuy nhiên, nghiên cứu của Zhang và cộng sự mới chỉ phát triển độ đo hạt thông tin tri thức trên không gian xấp xỉ rõ (crisp approximation space), chứ chưa mở rộng nó trên không gian xấp xỉ mờ.

Do đó, *mục tiêu thứ hai* của luận án này là nghiên cứu và mở rộng độ đo hạt thông tin tri thức trên không gian xấp xỉ mờ, ứng dụng vào việc xây dựng một phương pháp cập nhật thuộc tính cho bảng quyết định số có sự thay đổi về đối tượng. Mục tiêu nghiên cứu thứ hai này thuộc *nhóm các phương pháp chọn lọc thuộc tính cho bảng quyết định có sự thay đổi về đối tượng*.

Mục tiêu nghiên cứu: Hạn chế chính của các phương pháp rút gọn thuộc tính hiện tại, áp dụng cho các bảng quyết định số có chứa nhiễu và các bảng quyết định động, là chi phí thời gian tính toán cao. Do đó, luận án này tập trung vào mục tiêu cải thiện thời gian tính toán tập rút gọn trên cả hai loại bảng quyết định này. Cụ thể, mục tiêu này được chia thành hai hướng nghiên cứu chính:

1. *Cải thiện thời gian tính toán tập rút gọn trên bảng quyết định số có chứa nhiễu*:

   Để đạt được mục tiêu này, luận án sẽ giải quyết các vấn đề nghiên cứu sau:

   - **Vấn đề 1:** Nghiên cứu tổng quan các phương pháp chọn lọc thuộc tính hiện có nhằm giảm thiểu ảnh hưởng của nhiễu. Phân tích ưu và nhược điểm của từng phương pháp, và lý giải tại sao luận án lựa chọn tiếp cận VPFRS (Variable Precision Fuzzy Rough Sets) để phát triển.

- **Vấn đề 2:** Phát triển và tối ưu hóa các phép toán cơ bản, nhằm cải thiện hiệu quả thời gian tính toán các tập xấp xỉ trong VPFRS.
- **Vấn đề 3:** Đề xuất một phương pháp chọn lọc thuộc tính mới, dựa trên tiếp cận VPFRS đã được mở rộng và tối ưu hóa.

2. *Cải thiện thời gian tính toán tập rút gọn trên các bảng quyết định động*:

   Để cải thiện thời gian tính toán tập rút gọn trên các bảng quyết định động nói chung, và đặc biệt là trên các bảng quyết định số có sự thay đổi về tập đối tượng, luận án sẽ tập trung vào các vấn đề sau:

   - **Vấn đề 1:** Nghiên cứu và đánh giá các phương pháp chọn lọc thuộc tính gia tăng hiện có, dựa trên tiếp cận tính toán hạt (granular computing). Xác định các khoảng trống nghiên cứu trong lĩnh vực này và lý giải tại sao luận án lựa chọn tiếp cận hạt thông tin tri thức (information-theoretic granular measure) để mở rộng.
   - **Vấn đề 2:** Mở rộng khái niệm hạt thông tin tri thức trên không gian xấp xỉ mờ và xây dựng một độ đo mới dựa trên tiếp cận tính toán hạt.
   - **Vấn đề 3:** Xây dựng các công thức tính toán gia tăng tương ứng với các trường hợp bổ sung và loại bỏ đối tượng trong bảng quyết định số.
   - **Vấn đề 4:** Đề xuất các phương pháp chọn lọc thuộc tính gia tăng mới, tương ứng với các công thức tính toán gia tăng đã được xây dựng.

Đối tượng nghiên cứu: Luận án tập trung vào việc rút gọn các bảng quyết định đầy đủ có miền giá trị số, thường được gọi chung là bảng quyết định số, thông qua hai nhóm phương pháp sau:

- Nhóm phương pháp chọn lọc thuộc tính nhằm cải thiện độ chính xác trong các bảng quyết định số có chứa nhiễu.
- Nhóm phương pháp chọn lọc thuộc tính gia tăng áp dụng cho các bảng quyết định số.

Để thực hiện nghiên cứu này, luận án dựa trên các kiến thức nền tảng sau:

- Tổng quan về các khái niệm cơ bản liên quan đến bảng quyết định số và định nghĩa reduct trong ngữ cảnh chọn lọc thuộc tính.
- Khảo sát lý thuyết tập thô (Rough Set) và các mở rộng của nó, cùng với các phương pháp chọn lọc thuộc tính dựa trên các lý thuyết này.
- Nghiên cứu quy trình chung để xây dựng reduct thông qua phương pháp chọn lọc thuộc tính.
- Tìm hiểu về các công cụ chuẩn hóa dữ liệu, cũng như các phương pháp đo lường và đánh giá hiệu quả của mô hình phân lớp dữ liệu.
- Nghiên cứu các bộ dữ liệu số đầy đủ (complete numerical datasets) có sẵn từ kho dữ liệu học máy UCI [93].

Phạm vi nghiên cứu: Luận án tập trung vào nghiên cứu các phương pháp chọn lọc thuộc tính dựa trên các biến thể của các độ đo được xây dựng trên không gian xấp xỉ mờ, với ứng dụng chính là chọn lọc thuộc tính cho hai trường hợp dữ liệu sau:

- **Phương pháp chọn lọc thuộc tính cải thiện độ chính xác** cho các bảng quyết định số có chứa nhiễu.
- **Phương pháp chọn lọc thuộc tính gia tăng** trên các bảng quyết định số có sự thay đổi về tập đối tượng.

Phương pháp nghiên cứu: Để đạt được kết quả nghiên cứu của luận án, nghiên cứu lý thuyết và nghiên cứu thực nghiệm được kết hợp nhằm đảm bảo tính chặt chẽ về cơ sở toán học và tính ứng dụng trên các bộ dữ liệu thực tiễn. Trong đó:

- *Về phương diện lý thuyết*: nghiên cứu và xây dựng các các định nghĩa, các mệnh đề được trình bày và chứng minh rõ ràng, chặt chẽ, đầy đủ dựa trên cơ sở lý thuyết vững chắc của tập thô và tập thô mở rộng.

- *Về góc độ thực nghiệm*: sử dụng các bộ dữ liệu mẫu từ UCI [93] là nguyên liệu đầu vào cho các thuật toán. Sử dụng các mô hình phân lớp dữ liệu tiêu chuẩn cho dữ liệu số, kết hợp với các độ đo và loại hình đánh giá nhằm kết luận khả năng phân lớp của reduct thu được.

Kết quả nghiên cứu: Hai kết quả nghiên cứu chính của luận án đã đạt được gồm có:

1. *Đề xuất phương pháp rút gọt thuộc tính theo tiếp cận mở rộng tập thô mờ độ chính xác thay đổi trong không gian xấp xỉ mờ*

   Nội dung chính của kết quả này được trình bày trong Chương 2 của luận án, với một số các điểm đóng góp chính như sau:

   - Đề xuất phương pháp mở rộng mô hình tập thô mờ độ chính xác thay đổi VPOFRS.
   - Đề xuất thuật toán rút gọn thuộc tính theo tiếp cận VPOFRS.

   Các kết quả phân tích và thực nghiệm cho thấy, thuật toán rút gọn thuộc tính đề xuất trên mô hình VPOFRS cho thời gian thực hiện nhanh hơn so với thuật toán rút gọn thuộc tính theo tiếp cận IFRS và VPFRS truyền thống. **Đáng chú ý, tập rút gọn thu được có kích thước nhỏ hơn và độ chính xác phân lớp cao hơn so với một số phương pháp hiện có trên một số bộ dữ liệu.** Các kết quả nghiên cứu đã được công bố trên các công trình [CT2, CT3] của luận án.

2. *Đề xuất phương pháp tính toán tập rút gọn gia tăng trên không gian xấp xỉ mờ*

   Nội dung chính của kết quả này được trình bày trong Chương 3 của luận án, với một số các điểm đóng góp chính như sau:

   - Đề xuất khái niệm hạt thông tin mờ và độ đo hạt thông tin mờ.
   - Đề xuất mô hình và thuật toán rút gọn thuộc tính theo tiếp cận hạt thông tin mờ.
   - Xây dựng công thức gia tăng khi bổ sung tập đối tượng và thuật toán cập nhật tập rút gọn tương ứng.
   - Xây dựng công thức gia tăng khi loại bỏ tập đối tượng và thuật toán cập nhật tập rút gọn tương ứng.

   Các kết quả phân tích và thực nghiệm cho thấy, thuật toán rút gọn thuộc tính đề xuất theo tiếp cận độ đo hạt thông tin mờ cho thời gian thực hiện nhanh hơn so với thuật toán rút gọn thuộc tính theo tiếp cận khoảng cách mờ truyền thống. **Hơn nữa, tập rút gọn thu được có kích thước nhỏ hơn và độ chính xác phân lớp cao hơn so với một số phương pháp khác trên một số bộ dữ liệu**. Các kết quả nghiên cứu đã được công bố trên các công trình [CT1, CT4] của luận án.

Cấu trúc của luận án: Ngoài phần mở đầu và phần kết luận, luận án xây dựng 03 chương nghiên cứu với các nội dung chính như sau:

Chương 1. Nghiên cứu tổng quan về phương pháp rút gọn thuộc tính trên không gian xấp xỉ mờ. Chương này giới thiệu tổng quát về bài toán rút gọn thuộc tính, trình bày các kiến thức cơ sở của lý thuyết tập thô và tập thô mở rộng trên không gian xấp xỉ mờ, rút gọn thuộc tính cho bảng quyết định số.

Chương 2. Rút gọn thuộc tính theo tiếp cận mở rộng tập thô mờ độ chính xác thây đổi trên không gian xấp xỉ mờ. Chương này trình bày các đóng góp quan trọng trong việc xây dựng độ đo độ phụ thuộc mở rộng, có thời gian tính toán hiệu quả. Đáng chú ý, một số tập rút gọn trên các bảng quyết định số có chứa nhiễu có khả năng cải thiện độ chính xác phân lớp tốt hơn so với các phương pháp hiện có.

Chương 3 Phương pháp tính toán gia tăng tập rút gọn theo tiếp cận tính toán hạt trên không gian xấp xỉ mờ. Chương này trình bày các đóng góp chính trong việc xây dựng phương pháp tính toán gia tăng tập rút gọn hiệu quả trên không gian xấp xỉ mở. Hơn nữa, phương pháp đề xuất cũng trình bày một số kết quả thực nghiệm đáng chú ý về khả năng cải thiện độ chính xác phân lớp tốt trên một số bộ dữ liệu có kích thước lớn.

CHƯƠNG 1. TTỔNG QUAN VỀ PHƯƠNG PHÁP RÚT GỌN THUỘC TÍNH TRÊN KHÔNG GIAN XẤP XỈ MỜ.

Chương 1 của luận án tập trung vào việc cung cấp một cái nhìn tổng quan về bài toán chọn lọc thuộc tính (còn được gọi là rút gọn thuộc tính hay chọn lọc đặc trưng) và các kiến thức nền tảng cần thiết cho các chương tiếp theo [1]. Chương này trình bày tổng quan về các phương pháp chọn lọc thuộc tính trên không gian xấp xỉ mờ, tập trung vào hai hướng tiếp cận chính: cải thiện độ chính xác phân lớp trên các bảng quyết định số có chứa nhiễu và tính toán gia tăng trên các bảng quyết định số cập nhật tập các đối tượng [1, 2].

1.1. Mở đầu

Phần mở đầu giới thiệu tính cấp thiết của đề tài luận án, khẳng định rằng chọn lọc thuộc tính là một bước quan trọng trong tiền xử lý dữ liệu cho các tác vụ phân tích dữ liệu và học máy, đặc biệt trong bối cảnh dữ liệu ngày càng lớn và phức tạp [3, 4]. Quá trình này giúp chọn ra một tập con các thuộc tính có liên quan từ tập thuộc tính ban đầu, bảo toàn thông tin và khả năng phân lớp của dữ liệu gốc, đồng thời giảm độ phức tạp tính toán, cải thiện khả năng diễn giải mô hình và nâng cao khả năng dự đoán [4, 5].

1.1.1. Khái quát bài toán chọn lọc thuộc tính theo tiếp cận tập thô

Luận án nhắc lại việc Pawlak giới thiệu mô hình lý thuyết tập thô (Rough Set - RS) vào năm 1982, một công cụ mạnh mẽ để phân tích dữ liệu không đầy đủ và thiếu nhất quán [6, 7]. Dựa trên khả năng này, chọn lọc thuộc tính theo tiếp cận RS đã thu hút sự quan tâm rộng rãi [5, 6]. Không giống như các phương pháp giảm chiều dữ liệu như PCA hay SVD, phương pháp này chọn lọc trực tiếp các thuộc tính mà không làm thay đổi tính tự nhiên của dữ liệu sau khi rút gọn [5].

Bài toán chọn lọc thuộc tính theo tiếp cận tập thô bao gồm các bước chính [8]:

- Xác định không gian tìm kiếm: Tập hợp tất cả các tập con có thể của tập thuộc tính ban đầu.
- Xây dựng hàm đánh giá thuộc tính: Sử dụng các độ đo dựa trên khái niệm không gian xấp xỉ

của RS để đánh giá mức độ quan trọng của từng thuộc tính hoặc tập thuộc tính. Các độ đo phổ biến bao gồm độ đo miền dương, độ đo phụ thuộc, và độ đo độ phân biệt.

- Sử dụng chiến lược tìm kiếm: Áp dụng các thuật toán tìm kiếm khác nhau (ví dụ: tìm kiếm vét cạn, tìm kiếm tham lam, thuật toán di truyền) để tìm ra tập con thuộc tính tối ưu theo hàm đánh giá đã xây dựng.

- Xác thực tính đúng đắn của tập con: Đảm bảo tập thuộc tính đã chọn vẫn bảo toàn thông tin và khả năng phân lớp so với dữ liệu gốc.

Luận án cũng đề cập đến ba tiếp cận chính trong chọn lọc thuộc tính [8]:

- Tiếp cận filter thuộc tính: Dựa trên các đặc tính vốn có của dữ liệu (ví dụ: thông tin, tương quan) để đánh giá và chọn lọc thuộc tính một cách độc lập với bất kỳ thuật toán học máy cụ thể nào [8, 9]. Các chiến lược thường dùng bao gồm bổ sung thuộc tính quan trọng (Hình 1.1 [10]) hoặc loại bỏ thuộc tính dư thừa (Hình 1.2 [11]).

- Tiếp cận wrapper thuộc tính: Sử dụng một thuật toán học máy cụ thể làm hàm đánh giá. Quá trình tìm kiếm tập thuộc tính tối ưu được thực hiện thông qua việc huấn luyện và đánh giá hiệu suất của thuật toán học máy trên các tập con thuộc tính khác nhau (Hình 1.3 [12]). Tiếp cận này thường cho độ chính xác cao hơn nhưng có độ phức tạp tính toán lớn hơn, đặc biệt với chiến lược tìm kiếm vét cạn [9].

- Tiếp cận lai ghép filter-wrapper: Kết hợp ưu điểm của cả hai phương pháp filter (tính toán nhanh) và wrapper (độ chính xác cao) [13].

1.1.2. Một số ứng dụng của bài toán chọn lọc thuộc tính

Phần này trình bày tầm quan trọng và tính ứng dụng rộng rãi của chọn lọc thuộc tính trong nhiều lĩnh vực khác nhau [13]:

- Phân lớp dữ liệu: Cải thiện hiệu suất của các mô hình phân loại bằng cách chọn các thuộc tính có tính phân biệt cao nhất và giảm nguy cơ overfitting [14].

- Hồi quy dữ liệu: Xác định các yếu tố dự đoán có ảnh hưởng nhất và đơn giản hóa mô hình trong khi vẫn duy trì khả năng dự đoán [14].

- Phân cụm dữ liệu: Nâng cao hiệu quả của thuật toán phân cụm bằng cách loại bỏ các thuộc tính không liên quan hoặc dư thừa, dẫn đến các cụm có ý nghĩa hơn [14].

- Bảo mật thông tin: Sàng lọc các đặc trưng gây hại hoặc rủi ro nhất, ví dụ trong phát hiện Spam Web (giảm thời gian và chi phí tính toán [15, 16]) và phát hiện Malware (chọn lọc hàm API quan trọng [16]).

1.2. Các độ đo độ phụ thuộc của thuộc tính trên không gian xấp xỉ mờ

Để giải quyết hạn chế của tập thô truyền thống với dữ liệu có miền giá trị liên tục, các nhà nghiên cứu đã phát triển mô hình tập thô mờ (Fuzzy Rough Set - FRS) trên không gian xấp xỉ mờ [17, 18]. Trong FRS, mức độ quan hệ giữa các đối tượng được biểu diễn bằng các giá trị trong khoảng [19], cho phép xử lý tốt hơn sự mơ hồ và không chắc chắn trong dữ liệu số [20]. Nhiều độ đo đã được đề xuất để định nghĩa reduct và hỗ trợ chọn lọc thuộc tính trong không gian xấp xỉ mờ [20, 21]. Các độ đo này thường được xây dựng dựa trên các biến thể mở rộng từ mô hình tập thô truyền thống [21].

1.2.1. Độ đo miền dương mờ

Độ đo miền dương là một độ đo dựa trên các giá trị thống kê của các tập xấp xỉ dưới trong lý thuyết tập thô [21]. Độ đo miền dương mờ là sự mở rộng của độ đo này cho không gian xấp xỉ mờ, thể hiện mức độ phụ thuộc của thuộc tính quyết định vào các thuộc tính điều kiện thông qua các giá trị khẳng định của miền dương thống kê được [21, 22]. Độ đo này cho phép đo lường độ nhất quán của bảng quyết định (bảng được gọi là nhất quán nếu độ phụ thuộc của thuộc tính quyết định là lớn nhất, bằng 1) [22]. Bảng 1.1 [23] thống kê một số nghiên cứu liên quan đến việc xây dựng độ đo miền dương mờ cho các loại dữ liệu và mô hình không gian khác nhau [22, 24].

1.2.2. Độ đo Entropy mờ

Độ đo Entropy của Shannon được sử dụng rộng rãi trong học máy, đặc biệt với cây quyết định, để xác định lượng thông tin cần thiết để phân loại mỗi phần tử [24]. Trong không gian xấp xỉ mờ, độ đo Entropy cũng được mở rộng để đánh giá độ lợi thông tin của từng thuộc tính, làm tiêu chuẩn để xếp hạng mức độ quan trọng của chúng trong bảng quyết định [25]. Bảng 1.2 [26] liệt kê một số nghiên cứu về độ đo Entropy mờ trên các kiểu dữ liệu và không gian xấp xỉ khác nhau [25, 27].

1.2.3. Độ đo hạt thông tin

Độ đo hạt thông tin là một khái niệm quan trọng trong lĩnh vực tính toán hạt [27]. Trong tập thô truyền thống, độ đo hạt được xây dựng dựa trên các giá trị thống kê của một phân hoạch hoặc phủ, thể hiện mức độ thô và mịn của phân hoạch [27]. Độ đo độ phụ thuộc theo tiếp cận tính toán hạt được sử dụng để đánh giá sự phụ thuộc của thuộc tính trong bảng quyết định thông qua mức độ thô mịn của hạt thông tin phụ thuộc [27, 28]. Độ đo này cũng được mở rộng cho không gian xấp xỉ mờ để phục vụ bài toán chọn lọc thuộc tính trên dữ liệu số [28]. Bảng 1.3 [26] tổng hợp các nghiên cứu liên quan đến độ đo hạt thông tin mờ [28, 29].

1.3. Phương pháp chọn lọc thuộc tính cải thiện độ chính xác trên không gian xấp xỉ mờ

Sự phát triển của dữ liệu lớn đã làm gia tăng các trường hợp dữ liệu thiếu, không đầy đủ và thiếu nhất quán, thường chứa nhiều thuộc tính nhiễu [29]. Việc xác định và loại bỏ các thuộc tính này là rất quan trọng để giảm thời gian huấn luyện và cải thiện độ chính xác của mô hình [29, 30].

1.3.1. Các nghiên cứu liên quan

Đối với bảng quyết định có miền giá trị rời rạc và không đầy đủ, Tập Thô Độ Chính Xác Thay Đổi (Variable Precision Rough Set - VPRS) đã được đề xuất để điều chỉnh thành phần trong tập xấp xỉ dựa trên mức độ thành viên, giúp xử lý nhiễu tốt hơn [30, 31].

Với bảng quyết định số, mô hình VPRS được mở rộng thành Tập Thô Mờ Độ Chính Xác Thay Đổi (Variable Precision Fuzzy Rough Set - VPFRS) trên không gian xấp xỉ mờ [31]. Trong VPFRS, các phép toán xấp xỉ dưới và trên có thể được điều chỉnh để thay đổi độ chính xác của các tập xấp xỉ, ảnh hưởng đến độ phụ thuộc của thuộc tính [31, 32]. Tuy nhiên, phương pháp này phụ thuộc nhiều vào việc lựa chọn giá trị ngưỡng điều chỉnh (β), thường khác nhau giữa các bộ dữ liệu và đòi hỏi nhiều kinh nghiệm [32].

Một tiếp cận khác là sử dụng không gian xấp xỉ mờ trực cảm (Intuitionistic Fuzzy Rough Set - IFRS) [32, 33]. IFRS có các ràng buộc chặt chẽ hơn dựa trên hàm thành viên và hàm không thành viên, mô tả mối quan hệ giữa các đối tượng đa chiều hơn FRS [33-35]. IFRS được chứng minh là

có thể cải thiện chất lượng reduct trên dữ liệu nhiễu, nhưng có chi phí tính toán cao hơn [35]. Việc xây dựng các hàm thành viên và không thành viên độc lập cũng là một thách thức [33].

Bảng 1.4 [17] thống kê các nghiên cứu liên quan đến phương pháp chọn lọc thuộc tính cải thiện độ chính xác phân lớp, bao gồm cả các tiếp cận VPRS, VPFRS và IFRS [36, 37].

1.3.2. Vấn đề còn tồn tại

Các nghiên cứu gần đây về phương pháp chọn lọc thuộc tính cải thiện độ chính xác phân lớp, đặc biệt là dựa trên VPFRS, còn hạn chế về hiệu quả thời gian tính toán [38, 39]. Việc xem xét toàn bộ các đối tượng trong bảng quyết định khi điều chỉnh độ chính xác dẫn đến chi phí tính toán lớn [39]. Luận án đặt mục tiêu cải thiện thời gian tính toán bằng cách tối ưu hóa các phép toán xấp xỉ dựa trên các tính chất cơ bản của tập thô mờ với độ chính xác thay đổi [36, 40].

1.4. Phương pháp chọn lọc thuộc tính gia tăng trên không gian xấp xỉ mờ

Trong nhiều ứng dụng thực tế, dữ liệu liên tục được cập nhật và thay đổi theo thời gian, đòi hỏi mô hình phân lớp phải được điều chỉnh tương ứng [40]. Chọn lọc thuộc tính gia tăng đóng vai trò quan trọng trong việc giảm thời gian tiền xử lý dữ liệu khi có sự thay đổi, từ đó tăng hiệu quả cho quá trình cập nhật mô hình [40, 41].

1.4.1. Các nghiên cứu liên quan

Các phương pháp chọn lọc thuộc tính dựa trên độ đo xây dựng từ tập xấp xỉ dưới thường phù hợp với bảng quyết định tĩnh [41]. Tuy nhiên, trong thực tế, bảng quyết định thường xuyên được cập nhật, đòi hỏi khả năng tính toán hiệu quả chỉ dựa trên các thông tin mới [41, 42].

Các phương pháp tính toán gia tăng đã được đề xuất để cập nhật tập rút gọn một cách hiệu quả khi có sự thay đổi dữ liệu [43]. Có ba kịch bản thay đổi dữ liệu chính [42, 43]:

- Thay đổi về tập đối tượng: Bổ sung hoặc loại bỏ các mẫu dữ liệu [42, 44]. Nhiều nghiên cứu đã tập trung vào việc xây dựng các công thức gia tăng và thuật toán cập nhật reduct trong trường hợp này, sử dụng các độ đo như hạt thông tin [44], entropy thông tin [44], và phụ thuộc hàm [44]. Bảng 1.5 [20] thống kê các phương pháp gia tăng khi thay đổi tập đối tượng [42, 44].
- Thay đổi về tập thuộc tính: Bổ sung hoặc loại bỏ các thuộc tính [42, 45]. Các nghiên cứu trong lĩnh vực này cũng tập trung vào việc phát triển các phương pháp cập nhật reduct hiệu quả khi tập thuộc tính thay đổi, thường sử dụng độ đo hạt thông tin [45]. Bảng 1.6 [46] liệt kê các phương pháp gia tăng khi thay đổi tập thuộc tính [45].
- Thay đổi về giá trị thuộc tính.
- Thay đổi hỗn hợp: Kết hợp nhiều loại thay đổi [42, 45].

1.4.2. Vấn đề còn tồn tại

Các phương pháp tính toán gia tăng dựa trên độ đo khoảng cách tri thức thường có công thức phức tạp, khó chứng minh và xác nhận tính đúng đắn [45, 47]. Ngược lại, các công thức tính toán gia tăng dựa trên độ đo hạt thông tin tri thức có cấu trúc đơn giản hơn, dễ dàng chứng minh trong các trường hợp bổ sung hoặc loại bỏ đối tượng [47]. Tuy nhiên, độ đo hạt thông tin tri thức hiện tại chủ yếu được phát triển trên không gian xấp xỉ rõ (crisp approximation space) và chưa được mở rộng trên không gian xấp xỉ mờ [48]. Luận án đặt mục tiêu nghiên cứu và mở rộng độ đo hạt thông

tin tri thức trên không gian xấp xỉ mờ, ứng dụng vào việc xây dựng phương pháp cập nhật thuộc tính cho bảng quyết định số có sự thay đổi về đối tượng [48, 49].

1.5. Các kiến thức nền tảng

Phần này cung cấp các định nghĩa và khái niệm cơ bản làm nền tảng cho các chương tiếp theo [7].

1.5.1. Tập thô truyền thống

Mô hình tập thô của Pawlak là một phương pháp toán học để xử lý dữ liệu không đầy đủ, không chính xác và mơ hồ, dựa trên khái niệm xấp xỉ tập hợp [47].

- Hệ thống thông tin (Information System - IS): Được định nghĩa là một bộ $S = (U, A)$, trong đó U là một tập hữu hạn các đối tượng và A là một tập hữu hạn các thuộc tính [50]. Mỗi thuộc tính $a \in A$ là một hàm $a : U \to V_a$, gán một giá trị từ tập giá trị V_a cho mỗi đối tượng.
- Bảng quyết định (Decision Table - DT): Là một hệ thống thông tin $NDT = (U, C \cup D)$, trong đó C là tập thuộc tính điều kiện và D là tập thuộc tính quyết định [51].
- Quan hệ không phân biệt (Indiscernibility Relation): Với một tập con thuộc tính $B \subseteq A$, quan hệ không phân biệt $IND(B)$ là một quan hệ tương đương trên U được định nghĩa như sau: $(x, y) \in IND(B)$ nếu và chỉ nếu $a(x) = a(y)$ cho mọi $a \in B$. Quan hệ này tạo ra một phân hoạch của U thành các lớp tương đương $[x]_B = \{y \in U | (x, y) \in IND(B)\}$ [50].
- Xấp xỉ dưới (Lower Approximation): Cho $X \subseteq U$. Xấp xỉ dưới của X theo B, ký hiệu là $\underline{B}X$, là tập hợp tất cả các đối tượng chắc chắn thuộc về X dựa trên thông tin trong B: $\underline{B}X = \{x \in U | [x]_B \subseteq X\}$ [50].
- Xấp xỉ trên (Upper Approximation): Cho $X \subseteq U$. Xấp xỉ trên của X theo B, ký hiệu là $\overline{B}X$, là tập hợp tất cả các đối tượng có thể thuộc về X dựa trên thông tin trong B: $\overline{B}X = \{x \in U | [x]_B \cap X \neq \emptyset\}$ [50, 52].
- Vùng biên (Boundary Region): Vùng biên của X theo B, ký hiệu là $BN_B(X)$, chứa các đối tượng không thể xác định chắc chắn là thuộc hay không thuộc về X: $BN_B(X) = \overline{B}X - \underline{B}X$ [52].
- Tập hợp thô (Rough Set): Một tập hợp X được gọi là tập hợp thô nếu $\underline{B}X \neq \overline{B}X$, tức là vùng biên không rỗng [52]. Mô hình lý thuyết tập thô được minh họa trong Hình 1.4 [11, 52].

1.5.2. Tập mờ truyền thống

Lý thuyết tập mờ được giới thiệu bởi Zadeh để xử lý sự mơ hồ trong dữ liệu [53].

- Tập mờ (Fuzzy Set): Một tập mờ A trong vũ trụ U được định nghĩa bởi một hàm thành viên $\mu_A : U \to [19]$, trong đó $\mu_A(x)$ biểu thị mức độ thuộc của phần tử x vào tập mờ A [43, 53].
- Các toán tử cơ bản của tập mờ: Bảng 1.8 [3] mô tả các toán tử cơ bản như bù (complement), hợp (union), và giao (intersection) của các tập mờ, với các phép toán T-norm (ví dụ: min, tích) và T-conorm (ví dụ: max, tổng đại số) khác nhau [43].

1.5.3. Không gian xấp xỉ mờ

Không gian xấp xỉ mờ là sự kết hợp giữa không gian đối tượng và một quan hệ tương đương mờ [43].

- Quan hệ tương đương mờ (Fuzzy Equivalence Relation): Một quan hệ mờ $R : U \times U \to [19]$ được gọi là quan hệ tương đương mờ nếu nó thỏa mãn ba tính chất [49, 54, 55]:
  1. Tính phản xạ (Reflexivity): $R(x,x) = 1$ cho mọi $x \in U$.
  2. Tính đối xứng (Symmetry): $R(x,y) = R(y,x)$ cho mọi $x,y \in U$.
  3. Tính bắc cầu - min (Transitivity - min): $R(x,z) \geq \min\{R(x,y), R(y,z)\}$ cho mọi $x,y,z \in U$.
- Ma trận quan hệ tương đương mờ của một thuộc tính: Cho bảng quyết định $NDT = (U,C,D)$, ma trận quan hệ $S(a,U)$ của thuộc tính $a \in C$ thể hiện mức độ tương tự giữa các đối tượng dựa trên thuộc tính a. Ví dụ về cách xây dựng ma trận quan hệ từ dữ liệu số được trình bày trong [55].
- Ma trận quan hệ tương đương mờ của một tập thuộc tính: Với $P \subseteq C$, ma trận quan hệ $S(P,U)$ là giao của các ma trận quan hệ của từng thuộc tính trong P, thường sử dụng phép toán min: $s_{P_{ij}} = \min\{s_{a_{ij}} : a \in P\}$ [55].
- Phân hoạch mờ (Fuzzy Partition): Cho quan hệ tương đương mờ R trên U, phân hoạch mờ của R trên U là $U/R = \{[x]_R : x \in U\}$, trong đó $[x]_R$ là lớp tương đương mờ của x [56]. Tương tự, phân hoạch mờ của U theo một thuộc tính $a \in C$ là U/R_a [56], và theo một tập thuộc tính $P \subseteq C$ là U/R_P [56, 57].

1.5.4. Tập thô mờ

Tập thô mờ là sự mở rộng của tập thô truyền thống trên không gian xấp xỉ mờ [18].

- Cho bảng quyết định $NDT = (U,C,D,f)$ và quan hệ tương đương mờ R trên U. Với mọi tập mờ $A \in F(U)$, độ thuộc của đối tượng $x \in U$ vào tập xấp xỉ dưới mờ $\underline{R}A(x)$ và tập xấp xỉ trên mờ $\overline{R}A(x)$ của A theo R lần lượt được xác định như sau [58]:
  - $\underline{R}A(x) = \inf_{y \in U} I(R(x,y), A(y))$
  - $\overline{R}A(x) = \sup_{y \in U} \min(R(x,y), A(y))$

  trong đó I là một hàm implication mờ. Ví dụ về cách tính tập xấp xỉ mờ được minh họa [58].

1.6. Quy trình xây dựng và đánh giá thuật toán chọn lọc đặc trưng

Phần này mô tả quy trình chung để xây dựng và đánh giá các thuật toán chọn lọc đặc trưng, áp dụng cho cả tập thô truyền thống và tập thô mở rộng [59].

1.6.1. Trình tự các bước xây dựng phương pháp chọn lọc đặc trưng

Các bước cơ bản bao gồm [59]:

- Xây dựng phương pháp đánh giá, phân loại và chọn lọc thuộc tính: Sử dụng tiếp cận ma trận phân biệt (phân loại thuộc tính theo mức độ phân biệt) hoặc tiếp cận độ đo (đánh giá mức độ phụ thuộc, lượng thông tin, kích thước hạt thông tin) [59].

- Xây dựng tập rút gọn (Reduct): Dựa trên các điều kiện cần (tính bảo toàn khả năng phân lớp) và điều kiện đủ (tính tối thiểu số lượng thuộc tính) đã được định nghĩa [59].
- Thiết kế thuật toán: Sử dụng các phương pháp filter (chiến lược bổ sung hoặc loại bỏ thuộc tính), wrapper (kết hợp với thuật toán tiến hóa), hoặc lai ghép để tìm kiếm tập rút gọn [60]. Thuật toán 1.1 [60] mô tả mô hình tổng quát cho hầu hết các thuật toán rút gọn.

1.6.2. Dữ liệu và chuẩn hóa dữ liệu

Khi xây dựng mô hình học máy, chuẩn hóa dữ liệu là một bước quan trọng để đảm bảo hiệu suất và độ chính xác của mô hình [61, 62]. Các tác vụ chuẩn hóa cơ bản bao gồm [61, 62]:

- Chuẩn hóa đơn vị: Chuyển đổi đơn vị về cùng miền giá trị hoặc đơn vị đo lường [61].
- Xử lý giá trị thiếu: Điền hoặc loại bỏ các giá trị thiếu [61].
- Chuẩn hóa thời gian: Đồng bộ hóa thông tin thời gian [62].
- Chuẩn hóa tỷ lệ: Đảm bảo các biến số có cùng tỷ lệ để tránh ảnh hưởng không cân xứng đến mô hình [62].

1.6.3. Mô hình phân lớp và phương pháp đánh giá

Luận án đề cập đến việc sử dụng các mô hình phân lớp dữ liệu tiêu chuẩn cho dữ liệu số để đánh giá khả năng phân lớp của tập rút gọn thu được [63]. Một số mô hình phổ biến bao gồm [64]:

- K-Nearest Neighbors (k-NN)
- Support Vector Machines (SVM)

Để đánh giá hiệu suất của mô hình phân lớp, các độ đo thường được sử dụng bao gồm [63, 64]:

- Độ chính xác (Accuracy): Tỷ lệ các trường hợp được phân loại đúng.
- Độ chính xác (Precision): Tỷ lệ các trường hợp được dự đoán là dương tính và thực tế là dương tính.
- Độrecall (Recall): Tỷ lệ các trường hợp dương tính thực tế được dự đoán là dương tính.
- F1-score: Trung bình điều hòa của Precision và Recall.

Bảng 1.9 [3] minh họa ma trận nhầm lẫn nhị phân, cơ sở để tính toán các độ đo này [63].

Phương pháp đánh giá chéo k-fold (k-fold cross-validation) là một kỹ thuật quan trọng để đánh giá mô hình một cách khách quan và chính xác [63, 65, 66]. Quá trình này bao gồm chia dữ liệu thành k phần, lặp lại quá trình huấn luyện và đánh giá k lần, mỗi lần sử dụng một phần làm tập kiểm thử và k-1 phần còn lại làm tập huấn luyện [66, 67]. Độ chính xác trung bình được tính toán sau k lần lặp [67]. Phương pháp này giúp giảm thiểu hiện tượng overfitting và underfitting [68]. Luận án đề cập cụ thể đến phương pháp đánh giá chéo 10-fold [68].

1.7. Kết luận Chương 1

Chương 1 đã cung cấp một cái nhìn tổng quan về bài toán chọn lọc đặc trưng, ý nghĩa và vai trò của nó trong học máy và khai thác dữ liệu [69]. Nghiên cứu tổng quan các nghiên cứu liên quan về hai nhóm phương pháp chính: cải thiện độ chính xác phân lớp và chọn lọc đặc trưng gia tăng [69].

Phân tích các nghiên cứu gần đây chỉ ra rằng các phương pháp cải thiện độ chính xác phân lớp dựa trên VPFRS gặp hạn chế về thời gian tính toán do phải xét toàn bộ đối tượng [69, 70]. Tiếp cận IFRS có độ phức tạp cao hơn và việc xây dựng hàm thành viên/không thành viên phù hợp là khó khăn [70, 71]. Do đó, Luận án định hướng Chương 2 sẽ nghiên cứu mở rộng mô hình VPFRS để cải thiện độ chính xác phân lớp và hiệu quả thời gian [71].

Đối với phương pháp chọn lọc đặc trưng gia tăng khi bảng quyết định thay đổi về đối tượng, độ đo hạt thông tin tri thức có công thức đơn giản hơn so với độ đo khoảng cách [72]. Tuy nhiên, độ đo này chưa được mở rộng trên không gian xấp xỉ mờ [72, 73]. Vì vậy, Luận án định hướng Chương 3 sẽ nghiên cứu mở rộng độ đo hạt thông tin tri thức trên không gian xấp xỉ mờ để xây dựng các công thức tính toán gia tăng [73].

CHƯƠNG 2. RÚT GỌN THUỘC TÍNH THEO TIẾP CẬN MỞ RỘNG TẬP THÔ MỜ ĐỘ CHÍNH XÁC THAY ĐỔI TRÊN KHÔNG GIAN XẤP XỈ MỜ

Chương 2 của luận án tập trung vào bài toán **chọn lọc thuộc tính cải thiện độ chính xác phân lớp cho bảng quyết định số có chứa nhiễu** [1]. Xuất phát từ **hạn chế về thời gian tính toán của các phương pháp dựa trên mô hình tập thô mờ điều chỉnh (VPFRS)** và sự phức tạp của không gian xấp xỉ mờ trực cảm (IFRS), chương này đề xuất một **phương pháp tiếp cận mới bằng cách mở rộng mô hình VPFRS trên không gian xấp xỉ mờ** [1, 2]. Mục tiêu chính là **cải thiện hiệu quả thời gian tính toán trong quá trình rút gọn thuộc tính**, đồng thời **nâng cao độ chính xác phân lớp trên các bộ dữ liệu nhiễu** [2]. Phương pháp đề xuất, được gọi là VPOFRS (Variable Precision Optimized Fuzzy Rough Set), hứa hẹn giảm đáng kể thời gian tính toán so với các phương pháp trước đó [2]. Các kết quả nghiên cứu đã được công bố trong các công trình [CT2, CT3] [2].

2.1. Các kiến thức cơ sở

Phần này nhắc lại một số kiến thức quan trọng làm nền tảng cho các đề xuất trong chương. Đầu tiên là mô hình **Tập Thô Độ Chính Xác Thay Đổi (VPRS)** do Ziarko đề xuất năm 1993, được ứng dụng để rút gọn thuộc tính cải thiện độ chính xác phân lớp cho bảng quyết định phân loại không đầy đủ [3]. Tiếp theo là mô hình **Tập Thô Mờ Độ Chính Xác Thay Đổi (VPFRS)** được Zhao giới thiệu năm 2009 [4], là sự mở rộng của VPRS trên không gian xấp xỉ mờ, và đã được chứng minh hiệu quả trong rút gọn thuộc tính trên các bộ dữ liệu số [3]. Tuy nhiên, các công thức tính toán tập xấp xỉ trong VPFRS (công thức 2.8 và 2.9 trong luận án) còn chưa tối ưu do phải xét toàn bộ các đối tượng [5].

2.2. Đề xuất phương pháp cải tiến mô hình tập thô mờ điều chỉnh

Nhằm khắc phục những hạn chế của các công thức tính tập xấp xỉ hiện có trong VPFRS, phần này trình bày phương pháp cải tiến các phép toán liên quan đến độ thuộc của đối tượng và cách xác định giới hạn của các đối tượng [6].

2.2.1. Không gian xấp xỉ mờ đề xuất

Định nghĩa 2.7 (Quan hệ tương đương mờ): Cho U là tập không rỗng các đối tượng. Một quan hệ mờ $R : U \times U \to [7]$ được gọi là quan hệ tương đương mờ nếu nó thỏa mãn ba tính chất:

- **Tính phản xạ (Reflexivity)**: $R(x,x) = 1, \forall x \in U$.
- **Tính đối xứng (Symmetry)**: $R(x,y) = R(y,x), \forall x,y \in U$.
- **Tính bắc cầu theo T-chuẩn (T-transitivity)**: $T(R(x,y),R(y,z)) \leq R(x,z), \forall x,y,z \in U$, trong đó T là một T-chuẩn. Thông thường, $T(a,b) = \min(a,b)$ được sử dụng [5].

Định nghĩa 2.8 (Không gian xấp xỉ mờ): Một cặp (U,R) được gọi là không gian xấp xỉ mờ, trong đó U là tập không rỗng các đối tượng và R là một quan hệ tương đương mờ trên U [5, 8].

Định nghĩa 2.9 (Phân hoạch mờ): Cho bảng quyết định $NDT = (U,C,D)$, và R là một quan hệ tương đương mờ trên U. Khi đó, phân hoạch mờ của R trên U được xác định như sau:

$$U/R = \{[x]_R : x \in U\}$$

trong đó $[x]_R$ là lớp tương đương mờ chứa x, được định nghĩa bởi $[x]_R(y) = R(x,y), \forall y \in U$ [8].

Định nghĩa 2.10 (Phân hoạch mờ của một thuộc tính): Cho bảng quyết định $NDT = (U,C,D)$, và R là một quan hệ tương đương mờ trên U. Khi đó, phân hoạch mờ của U theo thuộc tính $a \in C$ trên R được kí hiệu bởi U/R_a, và được xác định theo công thức (2.12) [8].

Định nghĩa 2.11 (Phân hoạch mờ của tập thuộc tính): Cho bảng quyết định $NDT = (U,C,D)$ và $P \subseteq C$ là một tập con các thuộc tính điều kiện. Quan hệ tương đương mờ R_P được định nghĩa là giao của các quan hệ tương đương mờ tương ứng với mỗi thuộc tính trong P:

$$R_P(x,y) = \min_{a \in P}\{R_a(x,y)\}, \forall x,y \in U$$

Phân hoạch mờ của U theo tập thuộc tính P trên R được kí hiệu bởi U/R_P, và được xác định bởi $U/R_P = \{[x]_{R_P} : x \in U\}$ [8]. Mệnh đề 2.1 và 2.2 trong luận án chứng minh các tính chất liên quan đến quan hệ giữa các phân hoạch mờ khi có sự bao hàm giữa các tập thuộc tính [9].

Định nghĩa 2.12 (Phân hoạch mịn nhất): Cho bảng quyết định $NDT = (U,C,D)$ và P là hai phân hoạch mờ của U với $P \subseteq C$. Khi đó P được gọi là phân hoạch mịn nhất nếu mọi $x \in U, [x]_P = \{x\}$ [9].

Định nghĩa 2.13 (Phân hoạch thô nhất): Cho bảng quyết định $NDT = (U,C,D)$ và Q là hai phân hoạch mờ của U với $Q \subseteq C$. Khi đó Q được gọi là phân hoạch thô nhất nếu mọi $x \in U, [x]_Q = U$ [9].

2.2.2. Mô hình tập thô mờ điều chỉnh đề xuất

Dựa trên những vấn đề còn tồn tại của các công thức tính tập xấp xỉ hiện có (2.8) và (2.9), phần này trình bày phương pháp cải tiến cách tính độ thuộc của $x \in U$ và cách xác định giới hạn của $y \in U$ [6].

Định nghĩa 2.14 (Cải tiến độ thuộc của $x \in U$): Cho U là tập không rỗng các đối tượng, với mọi $x,y \in U, A \subseteq U$.

- β-thuộc của x với tập xấp xỉ dưới của A theo quan hệ R trên U được xác định như sau:

$$R_\beta A(x) = \min\{\inf_{y \in U} I(R(x,y),A(y)),\beta\}$$

trong đó $I(a,b) = 1 - a + a \cdot b$ là hàm ý nghĩa Lukasiewicz.

- β-thuộc của x với tập xấp xỉ trên của A theo quan hệ R trên U được xác định như sau:

$$\overline{R}_\beta A(x) = \max\{\sup_{y\in U} \min(R(x,y), A(y)), 1-\beta\}$$

Các ngưỡng $\beta \in [7]$ cho phép điều chỉnh mức độ bi quan/lạc quan trong việc xác định các tập xấp xỉ [6].

Mệnh đề 2.3 (Cải tiến giới hạn của $y \in U$): Cho bảng quyết định $NDT = (U, C, D)$. Với một tập thuộc tính $B \subseteq C$ và một tập quyết định D_k (một lớp quyết định), tập xấp xỉ dưới và trên mờ độ chính xác β của D_k có thể được tính toán hiệu quả hơn bằng cách chỉ xét các đối tượng y thuộc chính lớp quyết định D_k hoặc các đối tượng có quan hệ mờ với các đối tượng trong D_k vượt quá một ngưỡng nhất định [6, 10].

Định nghĩa 2.15 (Các miền của tập thô mờ độ chính xác thay đổi mở rộng): Cho không gian xấp xỉ mờ (U, R) và một tập mờ $A \in F(U)$, ngưỡng $\beta \in [7]$.

- β-miền dương (miền chắc chắn thuộc) của A được xác định bởi:

$$POS_\beta A = \{x \in U : R_\beta A(x) > 1-\beta\}$$

- β-miền biên (miền không chắc chắn) của A được xác định bởi:

$$BND_\beta A = \{x \in U : 1-\beta \geq R_\beta A(x) > \beta\}$$

  hoặc tương đương $BND_\beta A = \{x \in U : \overline{R}_\beta A(x) \geq 1-\beta \text{ và } R_\beta A(x) \leq 1-\beta\}$. Miền biên thể hiện sự chồng lấn giữa xấp xỉ dưới và xấp xỉ trên với tỉ lệ lỗi β [10].

- β-miền âm (miền phủ định) của A được xác định bởi:

$$NEG_\beta A = \{x \in U : R_\beta A(x) \leq \beta\}$$

2.2.3. Độ đo độ phụ thuộc đề xuất

Dựa trên mô hình VPOFRS được mở rộng, độ đo độ phụ thuộc được xây dựng để ứng dụng trong phân tích dữ liệu của bảng quyết định.

Định nghĩa 2.16 (β-miền dương của bảng quyết định): Cho bảng quyết định $NDT = (U, C, D)$. β-miền dương của tập thuộc tính điều kiện $P \subseteq C$ đối với tập thuộc tính quyết định D được định nghĩa là hợp của các β-miền dương của mỗi lớp quyết định $D_k \in D$:

$$POS_\beta(P, D) = \bigcup_{D_k \in D} POS_\beta(R_P, D_k)$$

trong đó $POS_\beta(R_P, D_k) = \{x \in U : R_{P\beta} D_k(x) > 1-\beta\}$ [10].

Định nghĩa 2.17 (β-độ phụ thuộc của thuộc tính quyết định vào thuộc tính điều kiện): β-độ phụ thuộc của tập thuộc tính quyết định D vào tập thuộc tính điều kiện $P \subseteq C$ được định nghĩa là tỷ lệ các đối tượng thuộc β-miền dương của P đối với D:

$$\gamma_\beta(P, D) = \frac{|POS_\beta(P, D)|}{|U|}$$

Giá trị $\gamma_\beta(P, D) \in [7]$ thể hiện mức độ nhất quán của bảng quyết định dưới góc độ tập thô mờ độ chính xác β [10, 11]. **Mệnh đề 2.6** chứng minh tính đơn điệu của độ đo độ phụ thuộc này: nếu $P \subseteq Q \subseteq C$, thì $\gamma_\beta(P, D) \leq \gamma_\beta(Q, D)$ [11].

2.3. Đề xuất phương pháp rút gọn thuộc tính cải thiện độ chính xác phân lớp

Dựa trên độ đo độ nhất quán của bảng quyết định được đề xuất, phần này trình bày phương pháp rút gọn thuộc tính cải thiện nhiễu theo tiếp cận VPOFRS.

2.3.1. Mô hình đề xuất

Dựa trên tính chất đơn điệu của độ đo độ nhất quán γ_β đã được chứng minh, độ đo đánh giá độ quan trọng của thuộc tính được xây dựng như sau:

Định nghĩa 2.18 (Độ quan trọng của thuộc tính): Cho bảng quyết định $NDT = (U, C, D)$ và một tập con các thuộc tính $B \subseteq C$. Độ quan trọng của một thuộc tính $c \in C \setminus B$ đối với B ở mức β được định nghĩa là sự thay đổi trong độ phụ thuộc khi thêm c vào B:

$$Sig_\beta(c, B) = \gamma_\beta(B \cup \{c\}, D) - \gamma_\beta(B, D)$$

Một thuộc tính $c \in C \setminus B$ được gọi là không quan trọng với B nếu $Sig_\beta(c, B) = 0$, tức là việc thêm c vào B không làm thay đổi độ phụ thuộc [12].

Định nghĩa 2.20 (β-tập rút gọn): Cho bảng quyết định $NDT = (U, C, D)$. Một tập con $B \subseteq C$ được gọi là β-tập rút gọn nếu các điều kiện sau đây được thỏa mãn:

1. **Điều kiện cần (Tính bảo toàn)**: $\gamma_\beta(B, D) = \gamma_\beta(C, D)$. Tập thuộc tính rút gọn phải bảo toàn độ nhất quán so với tập thuộc tính gốc.
2. **Điều kiện đủ (Tính tối thiểu)**: Với mọi $b \in B, \gamma_\beta(B \setminus \{b\}, D) \neq \gamma_\beta(B, D)$. Không có thuộc tính nào trong tập rút gọn là dư thừa.

[12, 13].

2.3.2. Thuật toán đề xuất

Thuật toán VPOFRS_ARD (Variable Precision Optimized Fuzzy Rough Set based Attribute Reduction for improving classification accuracy) được đề xuất theo phương pháp filter thuộc tính với chiến lược tìm tập rút gọn qua các bước sau [13, 14]:

1. Khởi tạo tập rút gọn $B = \emptyset$.
2. Tính độ phụ thuộc ban đầu $\gamma_\beta(C, D)$.
3. Lặp cho đến khi $\gamma_\beta(B, D) = \gamma_\beta(C, D)$:
   (a) Với mỗi thuộc tính $c \in C \setminus B$, tính độ quan trọng $Sig_\beta(c, B) = \gamma_\beta(B \cup \{c\}, D) - \gamma_\beta(B, D)$.
   (b) Chọn thuộc tính $b \in C \setminus B$ có độ quan trọng lớn nhất. Nếu có nhiều thuộc tính có độ quan trọng bằng nhau, chọn một cách ngẫu nhiên.
   (c) Bổ sung thuộc tính b vào tập rút gọn $B = B \cup \{b\}$.
4. Trả về tập rút gọn B.

Phân tích độ phức tạp cho thấy độ phức tạp của thuật toán VPOFRS_ARD là $O(|D||D_k||U||C|^2)$, trong đó $|U|$ là số lượng đối tượng, $|C|$ là số lượng thuộc tính điều kiện, $|D|$ là số lượng lớp quyết định và $|D_k|$ là số đối tượng của lớp thứ k [15]. Về mặt lý thuyết, do $|D_k| \leq |U|$, thuật toán VPOFRS_ARD có thời gian thực hiện nhanh hơn thuật toán VPFRS truyền thống [16].

2.3.3. Ví dụ số minh họa thuật toán

Phần này trình bày một ví dụ số để minh họa các bước thực hiện của thuật toán VPOFRS_ARD [16].

2.4. Thực nghiệm và đánh giá

Phần này trình bày kết quả thực nghiệm nhằm đánh giá hiệu quả về thời gian rút gọn thuộc tính, độ chính xác phân lớp và kích thước của tập rút gọn thu được từ thuật toán đề xuất [17].

2.4.1. Kịch bản và môi trường thực nghiệm

Kịch bản thực nghiệm bao gồm [17, 18]:

- Khảo sát vùng giá trị β tối ưu cho từng bộ dữ liệu để đạt được sự cân bằng tốt nhất giữa kích thước và độ chính xác phân lớp của tập rút gọn.
- So sánh thuật toán VPOFRS_ARD với các thuật toán rút gọn thuộc tính cải thiện độ chính xác hiện có là VPFRS [4] và IFD [19] dựa trên ba tiêu chí: thời gian rút gọn thuộc tính, độ chính xác phân lớp và kích thước tập rút gọn.

Thực nghiệm được thực hiện trên các bộ dữ liệu có độ chính xác phân lớp thấp (cho thấy sự tồn tại của nhiễu) [18] sử dụng ngôn ngữ lập trình Python trên nền tảng Windows 10 với bộ xử lý i5 và 8GB RAM [20].

2.4.2. Đánh giá thuật toán đề xuất

Phần này đánh giá mối quan hệ giữa ngưỡng β, kích thước tập rút gọn và độ chính xác phân lớp (sử dụng mô hình SVM và k-NN) [20-28]. Kết quả cho thấy kích thước tập rút gọn có xu hướng giảm khi β tăng, tuy nhiên có sự khác biệt giữa các bộ dữ liệu [23, 24]. Đáng chú ý, trên các bộ dữ liệu UFDC và SHDC, độ chính xác phân lớp không giảm, thậm chí còn tăng lên khi β thay đổi, cho thấy khả năng loại bỏ hiệu quả các thuộc tính nhiễu của thuật toán [25-28].

2.4.3. So sánh các thuật toán rút gọn thuộc tính cải thiện độ chính xác phân lớp

2.4.3.1. So sánh kích thước của tập rút gọn thu được từ các thuật toán

Hình 2.5 so sánh kích thước tập rút gọn giữa VPOFRS_ARD, VPFRS và IFD [29-31]. IFD có xu hướng tạo ra tập rút gọn nhỏ nhất, tuy nhiên VPOFRS_ARD lại vượt trội trên các bộ dữ liệu UFDC và UFDD (hai bộ có độ chính xác ban đầu thấp nhất) [29, 30]. Điều này cho thấy VPOFRS_ARD có khả năng xác định và loại bỏ thuộc tính nhiễu hiệu quả hơn [30].

2.4.3.2. So sánh độ chính xác phân lớp của tập rút gọn thu được từ các thuật toán

Hình 2.6 (SVM) và Hình 2.7 (k-NN) so sánh độ chính xác phân lớp của các tập rút gọn [31-35]. Nhìn chung, độ chính xác giữa các thuật toán không khác biệt nhiều, tuy nhiên trên bộ dữ liệu UFDC, VPOFRS_ARD đạt độ chính xác cao hơn đáng kể so với IFD và VPFRS, đồng thời giảm kích thước tập rút gọn [32-35]. Điều này củng cố thêm khả năng xử lý dữ liệu nhiễu của VPOFRS_ARD.

2.4.3.3. So sánh thời gian thực hiện của các thuật toán

Hình 2.8 so sánh thời gian thực hiện của VPOFRS_ARD và VPFRS [35-38]. VPOFRS_ARD cho thấy sự vượt trội về hiệu quả thời gian trên mọi bộ dữ liệu, đặc biệt là trên các bộ dữ liệu lớn [36, 37]. Sự cải thiện này có thể là do các tối ưu hóa trong mô hình VPOFRS, giúp giảm số lượng phép tính cần thiết [37].

2.5. Kết luận Chương 2

Chương 2 đã trình bày một phương pháp rút gọn thuộc tính mới dựa trên việc mở rộng mô hình tập thô mờ độ chính xác thay đổi (VPOFRS) trên không gian xấp xỉ mờ [38, 39]. Phương pháp đề xuất xây dựng một độ đo độ quan trọng của thuộc tính dựa trên độ phụ thuộc theo tiếp cận VPOFRS, cho phép chọn ra các thuộc tính quan trọng nhất và loại bỏ các thuộc tính ít liên quan hoặc gây nhiễu [39].

Kết quả thực nghiệm cho thấy thuật toán VPOFRS_ARD có thời gian thực hiện nhanh hơn so với VPFRS và IFRS, đồng thời tạo ra các tập rút gọn có kích thước nhỏ hơn và độ chính xác phân lớp cao hơn trên một số bộ dữ liệu nhiễu, đặc biệt là UFDC [2, 38]. Điều này khẳng định tính hiệu quả của phương pháp đề xuất trong việc giảm thiểu ảnh hưởng của nhiễu và cải thiện hiệu suất tính toán [2].

Ngoài ứng dụng trong bài toán rút gọn thuộc tính, mô hình tập thô VPOFRS còn có tiềm năng ứng dụng rộng rãi trong các bài toán phân tích dữ liệu khác liên quan đến dữ liệu mờ và không chắc chắn [40].

CHƯƠNG 3. PHƯƠNG PHÁP TÍNH TOÁN GIA TĂNG TẬP RÚT GỌN THEO TIẾP CẬN TÍNH TOÁN HẠT THÔNG TIN TRÊN KHÔNG GIAN XẤP XỈ MỜ

Chương 3 của luận án tập trung vào việc giải quyết bài toán **chọn lọc thuộc tính gia tăng** trong môi trường dữ liệu động, cụ thể là khi bảng quyết định có sự thay đổi về tập đối tượng [1]. Chương này xuất phát từ những hạn chế của các phương pháp truyền thống, chủ yếu dựa trên độ đo khoảng cách giữa các phân hoạch, trong việc xử lý hiệu quả dữ liệu biến động. Luận án đề xuất một hướng tiếp cận mới dựa trên việc **mở rộng khái niệm hạt thông tin tri thức để xây dựng độ đo trên không gian xấp xỉ mờ** [1].

Điểm nhấn chính của chương là việc xây dựng các **công thức tính toán gia tăng** để cập nhật tập thuộc tính rút gọn một cách hiệu quả khi có sự thay đổi về tập đối tượng (thêm hoặc bớt đối tượng) mà không cần tính toán lại từ đầu [1]. Chương cũng giới thiệu các **thuật toán tương ứng** cho việc cập nhật này, được đánh giá thông qua các thí nghiệm trên các bộ dữ liệu thực tế [1, 2]. Kết quả cho thấy phương pháp đề xuất không chỉ hiệu quả về mặt tính toán mà còn có khả năng cải thiện chất lượng của tập thuộc tính rút gọn [2].

3.1. Các kiến thức cơ sở

Phần này nhắc lại một số kiến thức cơ bản về **ma trận quan hệ tương đương mờ** của một thuộc tính và ma trận quan hệ tương đương mờ của tập thuộc tính [3]. Các ma trận quan hệ này đóng vai trò là nền tảng để xây dựng các công cụ tính toán trong các phần tiếp theo của chương [4]. Chi tiết về các kiến thức cơ sở này có thể tham khảo trong các công trình [5], [6].

Định nghĩa 3.1 (Quan hệ tương tự của hai đối tượng trên một thuộc tính) [4]: Cho bảng quyết định $NDT = \langle U, C, D \rangle$, quan hệ tương tự của hai đối tượng $i, j \in U$ theo thuộc tính $a \in C$ được xác định bởi công thức:

$$sa_{ij} = 1 - |a(i) - a(j)| \quad (3.1) \tag{3.1}$$

Trong đó, $|a(i) - a(j)|$ thường được chuẩn hóa về khoảng [7].

Từ quan hệ tương tự của hai đối tượng trên một thuộc tính, có thể xây dựng ma trận quan hệ tương đương mờ cho từng thuộc tính và cho một tập thuộc tính bằng cách lấy giao của các quan hệ tương đương mờ tương ứng [4].

3.2. Độ đo hạt thông tin mờ

Phần này giới thiệu cách mở rộng độ đo hạt thông tin từ không gian xấp xỉ rõ sang độ đo hạt thông tin trên không gian xấp xỉ mờ, gọi tắt là **độ đo hạt thông tin mờ (Fuzzy Information Granule - FIG)** [8]. Đầu tiên, khái niệm hạt thông tin mờ của thuộc tính được định nghĩa, sau đó các độ đo trên hạt thông tin mờ được xây dựng và cuối cùng là thuật toán rút gọn thuộc tính theo tiếp cận độ đo hạt thông tin mờ cho bảng quyết định NDT được đề xuất [8].

3.2.1. Hạt thông tin mờ

Định nghĩa 3.4 (Hạt thông tin mờ) [8]: Cho bảng quyết định $NDT = \langle U, C, D \rangle$ và tập thuộc tính $P \subseteq C$, khi đó hạt thông tin mờ của P trên U được xác định bởi:

$$[i]_P = \{(j, \mu_{[i]_P}(j)) | j \in U\} \tag{3.2}$$

Trong đó, $\mu_{[i]_P}(j) = \mu_P(i, j)$ là độ tương tự mờ giữa đối tượng i và j dựa trên tập thuộc tính P. Độ tương tự mờ $\mu_P(i, j)$ thường được tính bằng cách lấy giá trị nhỏ nhất (hoặc một toán tử T-chuẩn khác) của độ tương tự trên từng thuộc tính trong P:

$$\mu_P(i, j) = \min_{a \in P} \{sa_{ij}\} \tag{3.3}$$

Kích thước của hạt thông tin mờ $[i]_P$ có thể được định nghĩa là tổng độ thuộc của tất cả các đối tượng $j \in U$ vào hạt $[i]_P$:

$$|[i]_P| = \sum_{j \in U} \mu_{[i]_P}(j) \tag{3.4}$$

3.2.2. Độ đo hạt thông tin mờ

Dựa trên khái niệm hạt thông tin mờ, **độ đo hạt thông tin mờ (FIG)** của một tập thuộc tính P trên U khi xem xét thuộc tính quyết định D được định nghĩa như sau [9]:

$$FIG(P \cup D, U) = 1 - \frac{1}{|U|^2} \sum_{i=1}^{|U|} |[i]_{P \cup D}| \tag{3.5}$$

Độ đo này thể hiện mức độ thô của việc phân hoạch tập đối tượng U dựa trên tập thuộc tính $P \cup D$. Giá trị của $FIG(P \cup D, U)$ càng nhỏ, sự phân hoạch càng mịn, và do đó, thông tin càng nhiều.

Tương tự, độ đo hạt thông tin mờ của P trên U khi chỉ xem xét tập thuộc tính điều kiện P là:

$$FIG(P,U) = 1 - \frac{1}{|U|^2}\sum_{i=1}^{|U|}|[i]_P| \tag{3.6}$$

Độ phụ thuộc của thuộc tính quyết định D vào tập thuộc tính điều kiện P dựa trên độ đo hạt thông tin mờ được định nghĩa là:

$$FIG(D|P,U) = FIG(P,U) - FIG(P\cup D,U) = \frac{1}{|U|^2}\sum_{i=1}^{|U|}(|[i]_P| - |[i]_{P\cup D}|) \tag{3.7}$$

Độ đo này thể hiện mức độ mà việc biết thông tin từ P giúp làm giảm sự thô của việc phân hoạch khi xem xét cả P và D. Giá trị $FIG(D|P,U)$ càng lớn, sự phụ thuộc càng cao.

Mệnh đề 3.2 [9] chứng minh tính đơn điệu của độ đo hạt thông tin mờ: Nếu $P \subseteq Q \subseteq C$, thì $FIG(P\cup D,U) \leq FIG(Q\cup D,U)$. Điều này có nghĩa là khi thêm thuộc tính vào tập điều kiện, độ thô của phân hoạch (khi xem xét cả điều kiện và quyết định) sẽ giảm hoặc giữ nguyên.

3.2.3. Xây dựng thuật toán rút gọn thuộc tính

Dựa trên tính đơn điệu của độ đo hạt thông tin mờ, **Định nghĩa 3.6** [9] định nghĩa độ quan trọng của một thuộc tính $a \in C \setminus S$ đối với tập thuộc tính $S \subseteq C$ là:

$$Sig_U(a,S) = FIG(D|S\cup\{a\},U) - FIG(D|S,U) \tag{3.8}$$

Thuộc tính có độ quan trọng lớn nhất sẽ được ưu tiên thêm vào tập rút gọn.

Định nghĩa 3.7 [10] định nghĩa một tập thuộc tính $R \subseteq C$ là một tập rút gọn nếu nó thỏa mãn hai điều kiện:

- $FIG(D|R,U) = FIG(D|C,U)$ (tính bảo toàn độ phụ thuộc).
- Với mọi $a \in R$, $FIG(D|R\setminus\{a\},U) < FIG(D|C,U)$ (tính tối thiểu).

Thuật toán **FIG (Fuzzy Information Granule)** [10] được đề xuất để tìm tập rút gọn khi bảng quyết định NDT không thay đổi. Thuật toán này là một thuật toán **Greedy** theo phương pháp **filter thuộc tính** với chiến lược **bổ sung thuộc tính**. Các bước chính của thuật toán bao gồm:

1. Khởi tạo tập rút gọn $S = \emptyset$.
2. Tính $FIG(D|C,U)$ (độ phụ thuộc của quyết định vào toàn bộ thuộc tính điều kiện).
3. Lặp cho đến khi đạt được điều kiện dừng:
   (a) Tính độ quan trọng $Sig_U(a,S)$ cho mỗi thuộc tính $a \in C \setminus S$.
   (b) Chọn thuộc tính $s \in C \setminus S$ có độ quan trọng lớn nhất.
   (c) Thêm s vào tập rút gọn $S = S \cup \{s\}$.
   (d) Kiểm tra điều kiện dừng (ví dụ: $FIG(D|S,U) \approx FIG(D|C,U)$ hoặc đạt được số lượng thuộc tính mong muốn).
4. Trả về tập rút gọn S.

Độ phức tạp của thuật toán FIG được phân tích là $O(|U|^2|C|^2)$ [11]. Để tối ưu kích thước tập rút gọn, có thể sử dụng một ngưỡng chênh lệch δ (ví dụ, 0.5%) giữa $FIG(D|S,U)$ và $FIG(D|C,U)$ làm điều kiện dừng [11].

3.3. Xây dựng thuật toán cập nhật tập rút gọn khi bổ sung tập đối tượng

Phần này trình bày phương pháp xây dựng thuật toán cập nhật tập rút gọn khi bảng quyết định NDT có thêm một tập đối tượng $\Delta u = \{u_1, u_2, ..., u_k\}$ [12]. Các bước chính bao gồm: xây dựng công thức tính toán gia tăng cho độ đo hạt thông tin mờ khi thêm một đối tượng, mở rộng cho trường hợp thêm một tập đối tượng và cuối cùng đề xuất thuật toán cập nhật tập rút gọn tương ứng (**FIGAMO - Fuzzy Information Granule based Attribute reduction for Adding Multiple Objects**) [12].

3.3.1. Công thức gia tăng khi bổ sung thêm một đối tượng

Mệnh đề 3.3 (Hạt thông tin mờ gia tăng khi bổ sung một đối tượng) [12]: Cho bảng quyết định $NDT = \langle U, C, D\rangle$ và một đối tượng mới u_{new} được thêm vào, tạo thành tập đối tượng $U' = U \cup \{u_{new}\}$. Khi đó, hạt thông tin mờ của tập thuộc tính $P \subseteq C$ trên U' được tính như sau:

$$FIG(P,U') = 1 - \frac{1}{|U'|^2}\left(\sum_{i\in U}|[i]_P| + |[u_{new}]_P| + \sum_{i\in U}\mu_P(i,u_{new}) + \sum_{i\in U}\mu_P(u_{new},i) + \mu_P(u_{new},u_{new})\right) \tag{3.9}$$

Tương tự, công thức gia tăng cho $FIG(P\cup D, U')$ cũng được xây dựng [12]. Từ đó, công thức gia tăng cho độ phụ thuộc $FIG(D|P,U')$ khi thêm một đối tượng có thể được suy ra.

3.3.2. Công thức gia tăng khi bổ sung thêm một tập đối tượng

Mệnh đề 3.4 (Hạt thông tin mờ gia tăng khi bổ sung tập đối tượng) [13, 14]: Cho bảng quyết định $NDT = \langle U, C, D\rangle$ và tập đối tượng cần được bổ sung $\Delta u = \{u_1, u_2, ..., u_k\}$. Hạt thông tin mờ của P trên $U \cup \Delta u$ được tính bằng công thức gia tăng dựa trên $FIG(P,U)$ và độ tương tự giữa các đối tượng cũ và các đối tượng mới [14].

Mệnh đề 3.5 (Hạt thông tin mờ phụ thuộc gia tăng khi bổ sung tập đối tượng) [14, 15]: Công thức gia tăng cho $FIG(D|P, U\cup\Delta u)$ được xây dựng dựa trên $FIG(D|P,U)$ và các độ tương tự liên quan đến các đối tượng mới [15].

3.3.3. Đề xuất thuật toán gia tăng khi bổ sung tập đối tượng

Mệnh đề 3.6 [15, 16] đưa ra công thức tính độ quan trọng của thuộc tính $a \in C \setminus B$ với tập thuộc tính B trên $U \cup \Delta u$ dựa trên độ quan trọng trên U và độ tương tự liên quan đến các đối tượng mới [16].

Thuật toán **FIGAMO** [16, 17] được đề xuất để cập nhật tập rút gọn khi bổ sung tập đối tượng Δu. Các bước chính của thuật toán:

1. Cho RU là tập rút gọn thu được trên bảng quyết định với tập đối tượng ban đầu U. Khởi tạo tập rút gọn mới $S = RU$.
2. Tính $FIG(D|S, U\cup\Delta u)$ (Fs) và $FIG(D|C, U\cup\Delta u)$ (Fc) bằng các công thức gia tăng.
3. Nếu độ chênh lệch giữa Fs và Fc nhỏ hơn một ngưỡng ε (ví dụ, 0.5%), trả về S.
4. Ngược lại, lặp lại quá trình chọn thuộc tính có độ quan trọng lớn nhất (tính bằng công thức gia tăng) từ $C \setminus S$ và thêm vào S cho đến khi độ chênh lệch giữa $FIG(D|S, U\cup\Delta u)$ và $FIG(D|C, U\cup\Delta u)$ đạt ngưỡng cho phép.

5. Loại bỏ các thuộc tính dư thừa từ S bằng cách kiểm tra nếu việc loại bỏ một thuộc tính làm giảm đáng kể độ phụ thuộc.

6. Trả về tập rút gọn S được cập nhật.

Độ phức tạp của thuật toán FIGAMO được phân tích là xấp xỉ $O(|\Delta u|^2|C|)$ trong trường hợp lý tưởng khi tập rút gọn không thay đổi nhiều [17].

3.4. Xây dựng thuật toán cập nhật tập rút gọn khi loại bỏ tập đối tượng

Phần này trình bày phương pháp xây dựng thuật toán cập nhật tập rút gọn khi bảng quyết định NDT loại bỏ một tập đối tượng Δu [18]. Tương tự như trường hợp bổ sung, các bước chính bao gồm: xây dựng công thức tính toán gia tăng cho độ đo hạt thông tin mờ khi loại bỏ một đối tượng, mở rộng cho trường hợp loại bỏ một tập đối tượng và cuối cùng đề xuất thuật toán cập nhật tập rút gọn tương ứng (**FIGDMO - Fuzzy Information Granule based Attribute reduction for Deleting Multiple Objects**) [18].

3.4.1. Công thức gia tăng khi loại bỏ một đối tượng

Mệnh đề 3.7 (Hạt thông tin mờ gia tăng khi loại bỏ một đối tượng) [18]: Cho bảng quyết định $NDT = \langle U, C, D\rangle$ và một đối tượng $u_{removed}$ bị loại bỏ, tạo thành tập đối tượng $U' = U \setminus \{u_{removed}\}$. Khi đó, hạt thông tin mờ của tập thuộc tính $P \subseteq C$ trên U' được tính bằng cách trừ đóng góp của $u_{removed}$ khỏi tổng [18]. Tương tự, công thức gia tăng cho $FIG(P \cup D, U')$ và độ phụ thuộc $FIG(D|P, U')$ khi loại bỏ một đối tượng cũng được xây dựng.

3.4.2. Công thức gia tăng khi loại bỏ một tập đối tượng

Mệnh đề 3.8 (Hạt thông tin mờ gia tăng khi loại bỏ tập đối tượng) [19]: Cho bảng quyết định $NDT = \langle U, C, D\rangle$ và tập đối tượng cần được loại bỏ $\Delta u = \{u_1, u_2, ..., u_k\}$. Hạt thông tin mờ của P trên $U \setminus \Delta u$ được tính bằng công thức gia tăng dựa trên $FIG(P, U)$ và độ tương tự liên quan đến các đối tượng bị loại bỏ [20].

Mệnh đề 3.9 (Hạt thông tin mờ phụ thuộc gia tăng khi loại bỏ tập đối tượng) [20]: Công thức gia tăng cho $FIG(D|P, U \setminus \Delta u)$ được xây dựng dựa trên $FIG(D|P, U)$ và các độ tương tự liên quan đến các đối tượng bị loại bỏ.

3.4.3. Đề xuất thuật toán gia tăng khi loại bỏ tập đối tượng

Mệnh đề 3.10 [20] đưa ra công thức tính độ quan trọng của thuộc tính $a \in C \setminus B$ với tập thuộc tính B trên $U \setminus \Delta u$ dựa trên độ quan trọng trên U và độ tương tự liên quan đến các đối tượng bị loại bỏ.

Thuật toán **FIGDMO** [21] được đề xuất để cập nhật tập rút gọn khi loại bỏ tập đối tượng Δu. Tương tự như FIGAMO, thuật toán này sử dụng các công thức gia tăng để tính toán lại độ phụ thuộc và độ quan trọng của thuộc tính, từ đó cập nhật tập rút gọn một cách hiệu quả [21].

3.5. Thực nghiệm và đánh giá các thuật toán

Phần này trình bày kết quả thực nghiệm để chứng minh tính hiệu quả về thời gian và chất lượng của các thuật toán rút gọn thuộc tính gia tăng FIGAMO và FIGDMO [22]. Tuy nhiên, phần thực

nghiệm chủ yếu tập trung vào FIG và FIGAMO [22, 23].

3.5.1. Kịch bản và môi trường thực nghiệm

Các thuật toán được cài đặt bằng ngôn ngữ Python trên nền tảng Windows 10 với bộ xử lý i5 và 8GB RAM [24]. Các bộ dữ liệu chuẩn từ UCI Repository được sử dụng để đánh giá hiệu năng của các thuật toán [24, 25]. Quá trình đánh giá bao gồm so sánh thời gian thực hiện, kích thước của tập rút gọn và độ chính xác phân lớp (sử dụng các mô hình SVM và k-NN với đánh giá chéo 10-fold) [24].

3.5.2. Đánh giá thuật toán FIG

Thuật toán FIG được so sánh với thuật toán FD (Fuzzy Dependency based attribute reduction) [6] trên các bộ dữ liệu tĩnh [24].

Thời gian thực hiện

Hình 3.3 [26] cho thấy **thuật toán FIG có thời gian thực hiện tốt hơn thuật toán FD trên mọi bộ dữ liệu thử nghiệm** [26]. Điều này chứng tỏ điều kiện dừng (ngưỡng chênh lệch về độ đo hạt thông tin) được thiết kế trong thuật toán FIG giúp giảm thời gian tính toán [26].

Kích thước tập rút gọn

Hình 3.4 [27] cho thấy **kích thước tập rút gọn của thuật toán FIG thường nhỏ hơn hoặc tương đương so với thuật toán FD** [27]. Điều này khẳng định khả năng giảm thuộc tính hiệu quả của thuật toán FIG [27]. Sự đánh đổi giữa thời gian thực hiện và kích thước tập rút gọn được quan sát khi thay đổi ngưỡng chênh lệch [28].

Độ chính xác phân lớp

Hình 3.5 và 3.6 [28-31] so sánh độ chính xác phân lớp (sử dụng SVM và k-NN) của tập rút gọn thu được từ FIG và FD so với tập thuộc tính gốc. Kết quả cho thấy:

- Trên các bộ dữ liệu nhỏ, độ chính xác phân lớp không có sự khác biệt đáng kể [29].
- Trên các bộ dữ liệu lớn (CMSC, BCWD), **tập thuộc tính rút gọn từ FIG thường đạt độ chính xác phân lớp cao hơn so với FD** [29, 30].

Những kết quả này cho thấy thuật toán FIG hiệu quả trong việc chọn lọc các thuộc tính quan trọng và duy trì hoặc cải thiện khả năng phân lớp [31].

3.5.3. Đánh giá thuật toán FIGAMO

Thuật toán FIGAMO được so sánh với thuật toán FDMAO (incremental FD) [6] khi bổ sung tập đối tượng [32].

Thời gian thực hiện

Hình 3.7 [33] cho thấy **thuật toán FIGAMO có thời gian thực hiện tốt hơn thuật toán FDMAO trên mọi bộ dữ liệu thử nghiệm khi bổ sung đối tượng** [33]. Điều kiện dừng trong FIGAMO tiếp tục đóng vai trò quan trọng trong việc cải thiện hiệu suất thời gian [33]. Sự khác biệt về thời gian thực hiện giữa các mức bổ sung khác nhau thường không lớn, nhưng FIGAMO cho thấy sự ổn định hơn so với FDMAO trên các bộ dữ liệu lớn [34].

Kích thước tập rút gọn

Hình 3.8 [35] cho thấy **kích thước tập rút gọn của thuật toán FIGAMO thường nhỏ hơn hoặc tương đương so với thuật toán FDMAO sau khi bổ sung đối tượng** [35]. Điều này cho thấy FIGAMO duy trì khả năng nén dữ liệu hiệu quả trong môi trường dữ liệu động [36].

Độ chính xác phân lớp

Hình 3.9 và 3.10 [37-40] so sánh độ chính xác phân lớp (sử dụng SVM và k-NN) của tập rút gọn thu được từ FIGAMO và FDMAO so với tập thuộc tính gốc sau khi bổ sung đối tượng. Kết quả cho thấy:

- Trên nhiều bộ dữ liệu, **FIGAMO duy trì độ chính xác phân lớp tương đương hoặc cao hơn so với tập thuộc tính gốc và thường tốt hơn FDMAO** [38, 39].
- Tuy nhiên, hiệu quả của việc rút gọn thuộc tính có thể phụ thuộc vào từng bộ dữ liệu cụ thể [40, 41].

Nhìn chung, FIGAMO cho thấy khả năng duy trì hoặc cải thiện độ chính xác phân lớp sau khi rút gọn thuộc tính trong môi trường dữ liệu có sự thay đổi về đối tượng [41].

3.6. Kết luận Chương 3

Chương 3 của luận án đã trình bày một **phương pháp rút gọn thuộc tính gia tăng hiệu quả** sử dụng tiếp cận tính toán hạt trên không gian xấp xỉ mờ [41, 42]. Phương pháp này được thiết kế để xử lý các bộ dữ liệu lớn một cách hiệu quả, đặc biệt khi dữ liệu liên tục được cập nhật hoặc thay đổi [42].

Điểm nổi bật của phương pháp là việc **mở rộng độ đo hạt thông tin trên không gian xấp xỉ mờ**, giúp đơn giản hóa công thức tính toán gia tăng và giảm đáng kể thời gian cần thiết để cập nhật tập rút gọn [42]. Kết quả thực nghiệm đã chứng minh rằng thuật toán đề xuất (**FIG và FIGAMO**) không chỉ có thời gian tính toán hiệu quả hơn so với các phương pháp khác mà còn tạo ra các tập thuộc tính rút gọn có kích thước nhỏ hơn và cải thiện độ chính xác phân lớp trên một số bộ dữ liệu lớn [43].

Tuy nhiên, luận án cũng chỉ ra rằng việc **lựa chọn ngưỡng điều chỉnh về độ lệch** là quan trọng để cân bằng giữa thời gian tính toán gia tăng và độ chính xác phân lớp của tập rút gọn cuối cùng [43, 44]. Sự cân bằng này cần được xem xét kỹ lưỡng trong các ứng dụng thực tế [44].

KẾT LUẬN

A. Những kết quả chính của luận án

Luận án tham gia vào luồng nghiên cứu về vấn đề tiền xử lý dữ liệu, ứng dụng cho các lĩnh vực khai thác dữ liệu và học máy với các lớp bài toán chọn lọc thuộc tính cho bảng quyết định miền giá trị số. Trước tiên luận án tập trung khảo sát các khoảng trống nghiên cứu của các phương pháp hiện có và đưa ra một số các đóng góp chính như sau:

Thứ nhất, luận án đề xuất *phương pháp chọn lọc thuộc tính cho bảng quyết định theo tiếp cận mở rộng tập thô mờ độ chính xác thay đổi trên không gian xấp xỉ mờ.*

Điểm cốt lõi của phương pháp nằm ở việc xây dựng một độ đo để đánh giá mức độ quan trọng của từng thuộc tính. Độ đo này dựa trên độ phụ thuộc, được định nghĩa theo cách tiếp cận VPOFRS, cho phép xác định chính xác mức độ ảnh hưởng của mỗi thuộc tính đến quá trình phân loại dữ liệu. Điều này giúp chọn ra những thuộc tính quan trọng nhất, đồng thời loại bỏ những thuộc tính ít liên quan hoặc gây nhiễu.

Để kiểm tra hiệu quả của phương pháp đề xuất, luận án đã tiến hành các thử nghiệm trên nhiều tập dữ liệu khác nhau. Kết quả cho thấy rằng thuật toán rút gọn thuộc tính mới không chỉ cải thiện đáng kể thời gian tính toán mà còn tạo ra các tập rút gọn có độ chính xác cao hơn và kích thước nhỏ hơn so với các thuật toán truyền thống. Điều này cho thấy phương pháp mới không chỉ nhanh hơn mà còn hiệu quả hơn trong việc đơn giản hóa mô hình và cải thiện khả năng tổng quát hóa của mô hình.

Ngoài việc áp dụng cho bài toán rút gọn thuộc tính, luận án cũng chỉ ra rằng mô hình tập thô VPOFRS có tiềm năng ứng dụng rộng rãi trong các bài toán phân tích dữ liệu khác. Với khả năng xử lý dữ liệu mờ và không chắc chắn, VPOFRS có thể trở thành một công cụ mạnh mẽ trong nhiều lĩnh vực khác nhau, từ nhận dạng mẫu đến dự báo và ra quyết định, mở ra nhiều hướng nghiên cứu và ứng dụng trong tương lai.

Thứ hai, luận án đề xuất *phương pháp tính toán gia tăng tập rút gọn sử dụng phương pháp tính toán hạt thông tin trên không gian xấp xỉ mờ.*

Điểm nổi bật của phương pháp là việc mở rộng độ đo hạt thông tin trên không gian xấp xỉ mờ. Bằng cách tận dụng khái niệm hạt thông tin, công thức tính toán được đơn giản hóa, giúp giảm thiểu đáng kể thời gian cần thiết để tính toán gia tăng và cập nhật tập rút gọn. Điều này đặc biệt quan trọng khi xử lý dữ liệu lớn, nơi mà thời gian tính toán trở thành một yếu tố then chốt.

Mặc dù thuật toán đề xuất hứa hẹn trong việc rút gọn thuộc tính cho dữ liệu lớn, nghiên cứu cũng chỉ ra rằng cần phải khảo sát và điều chỉnh giá trị ngưỡng điều chỉnh về độ lệch trước khi áp dụng. Việc này nhằm đảm bảo sự cân bằng giữa thời gian tính toán gia tăng tập rút gọn và độ chính xác phân lớp của tập rút gọn cuối cùng. Sự cân bằng này là rất quan trọng để đạt được hiệu quả tốt nhất trong các ứng dụng thực tế.

B. Hướng phát triển tiếp theo của luận án

Các kết quả nghiên cứu hiện tại của luận án được xây dựng dựa trên không gian xấp xỉ mờ, nơi các bảng quyết định số được sử dụng làm dữ liệu đầu vào. Tuy nhiên, một hạn chế của cách tiếp cận này là trong thực tế, nhiều bài toán phân tích dữ liệu liên quan đến các bảng quyết định phức tạp hơn, không chỉ chứa các thuộc tính số (liên tục) mà còn bao gồm các thuộc tính phân loại xen kẽ. Sự hiện diện của cả hai loại thuộc tính này trong cùng một bảng quyết định tạo ra những thách thức riêng trong quá trình rút gọn thuộc tính.

Do đó, một hướng phát triển tiềm năng cho các kết quả nghiên cứu của luận án là mở rộng sang không gian lân cận. Việc mở rộng này sẽ cho phép áp dụng các phương pháp rút gọn thuộc tính đã phát triển cho các bảng quyết định hỗn hợp, bao gồm cả thuộc tính số và thuộc tính phân loại.

Bằng cách mở rộng sang không gian lân cận, luận án có thể giải quyết được một phạm vi lớn hơn các bài toán thực tế. Điều này sẽ làm tăng tính ứng dụng và giá trị thực tiễn của các kết quả nghiên cứu, vì dữ liệu trong thế giới thực thường không thuần nhất và chứa nhiều loại thuộc tính khác nhau.

DANH MỤC CÁC CÔNG TRÌNH NGHIÊN CỨU

[CT1] **Phạm Minh Ngọc Hà**, Nguyễn Long Giang, Nguyễn Văn Thiện, Nguyễn Bá Quảng, "Về một thuật toán gia tăng tìm tập rút gọn của bảng quyết định không đầy đủ", *Chuyên san các công trình nghiên cứu phát triển CNTT&TT, Tạp chí Công nghệ thông tin và truyền thông - Bộ TT&TT*, Tập 2019, Số 1, Tháng 9, Tr. 11-18.

[CT2] **Phạm Minh Ngọc Hà**, Nguyễn Long Giang, Trần Thanh Đại, Trần Văn Sinh, "Rút gọn thuộc tính trong bảng quyết định đầy đủ theo tiếp cận xác suất phân lớp của hạt thông tin lân cận mờ", *Hội thảo quốc gia lần thứ XXV Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông*, Hà Nội, 2022.

[CT3] **Pham Minh Ngoc Ha**, Tran Thanh Dai, Nguyen Manh Hung, Hoang Tuan Dung, "A Novel Extension Method of VPFRS Mode for Attribute Reduction Problem in Numerial Decision Tables," *J. Comput. Sci. Cybern.*, no.1 Mar, 2024.

[CT4] **Pham Minh Ngoc Ha**, Nguyen Long Giang, Nguyen Manh Hung, Tran Thanh Dai, "Incremental Attribute Reduction in Rough Set for Fuzzy Decision Tables", *Vietnam Journal of Science and Technology*, 2024.