BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Trần Thanh Đại

RÚT GỌN THUỘC TÍNH TRONG BẢNG QUYẾT ĐỊNH THEO TIẾP CẬN TẬP THÔ MỜ TRỰC CẢM VÀ TÔPÔ SUY RỘNG

TÓM TẮT LUẬN ÁN TIẾN SĨ HỆ THỐNG THÔNG TIN

Mã số: 9 48 01 04

Hà Nội – 2023

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.
2. Người hướng dẫn 2: GS. TS Vũ Đức Thi, Viện Công nghệ Thông tin, Đại học Quốc Gia Hà Nội.

Phản biện 1: PGS. TS Hoàng Việt Long .....

Phản biện 2: PGS.TS Nguyễn Hà Nam.....

Phản biện 3: TS Lê Quang Minh.....

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 9 giờ , ngày 21 tháng 11 năm 2023.

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

Rút gọn thuộc tính [1]–[3] hay chọn lọc thuộc tính là bước tiền xử lý dữ liệu quan trọng, được ứng dụng rỗng rãi trong các lĩnh vực liên quan đến nhận dạng mẫu và khai thác dữ liệu gồm có: phân lớp dữ liệu [4], [5], nhận dạng chữ viết tay [6], [7], nhận dạng tiếng nói [8], [9], phát hiện và phân loại spam [10], [11] và hỗ trợ ra quyết định [12], [13]. Rút gọn thuộc tính nhằm xác định và chọn lọc tập con của tập thuộc tính ban đầu có liên quan nhiều nhất hoặc loại bỏ các thuộc tính dư thừa ít liên quan nhất tới việc ra quyết định của bài toán. Rút gọn thuộc tính thường được thực hiện để mô hình đạt được một số mục tiêu gồm có: tăng tính dễ hiểu của luật, cải thiện hiệu năng, giảm chi phí tính toán.

Mô hình lý thuyết tập thô (Rough Set - RS) được Pawlack giới thiệu vào năm 1982 là công cụ toán học mạnh mẽ, hiệu quả cho các trường hợp dữ liệu không chắc chắn, không đầy đủ và thiếu nhất quán [14]. Rút gọn thuộc tính là một trong những ứng dụng quan trọng của mô hình lý thuyết tập thô, đã và đang nhận được sự quan tâm của các nhà nghiên cứu [15]–[17]. Dựa trên khái niệm lớp tương đương và các phép toán xấp xỉ trong mô hình lý thuyết tập thô, nhiều phương pháp đo lường độ quan trọng của thuộc tính được đề xuất để tìm tập rút gọn. Bên cạnh đó, không gian tôpô cũng là một khái niệm quan trọng trong mô hình lý thuyết tập thô [18], [19]. Khái niệm tôpô theo tiếp cận tập thô cũng được Pawlack giới thiệu vào năm 1988 và nhận được nhiều quan tâm của các nhà nghiên cứu [4], [20].

Hơn ba thập kỉ vừa qua, hướng rút gọn thuộc tính theo tiếp cận tập thô [14] đã và đang thu hút được sự quan tâm của nhiều nhà nghiên cứu. Các kết quả nghiên cứu cho thấy phương pháp rút gọn thuộc tính theo tiếp cận tập thô hiệu quả trên các bảng quyết định có thuộc tính giá trị rời rạc. Tuy nhiên, với các bảng quyết định có thuộc tính giá trị liên tục (bảng quyết định số) cần phải thực hiện bước biến đổi miền giá trị liên tục về miền giá trị rời rạc trước khi rút gọn thuộc tính. Bước biến đổi này phát sinh chi phí thực hiện và có thể làm mất dữ liệu trong quá trình biến đổi. Do đó, các nhà nghiên cứu đề xuất phương pháp rút gọn thuộc tính trực tiếp trên các bảng quyết định gốc mà không phải qua quá trình rời rạc hóa dữ liệu.

Để rút gọn thuộc tính trực tiếp trên bảng quyết định gốc, các nhà nghiên cứu đã mở rộng mô hình lý thuyết tập thô truyền thống trên nền các tập mờ (Fuzzy Set - FS) và tập mờ trực cảm (Intuitionistic Fuzzy Set - IFS) gồm có:

1. Tập thô mờ (Fuzzy Rough Set - FRS)

Mô hình tập thô mờ [21], [22] sử dụng khái niệm tương tự thay cho khái niệm không phân biệt được trong mô hình lý thuyết tập thô truyền thống. Do đó, chúng ta không cần phải rời rạc hóa dữ liệu mà vẫn đánh giá chính xác mối quan hệ của các đối tượng trong một tập. Cho đến nay, các hướng nghiên cứu rút gọn thuộc tính theo tiếp cận tập thô mờ diễn ra khá sôi động với các đề xuất mới về độ đo gồm có: độ đo miền dương mờ (Fuzzy POS - FPOS) [17], [23]–[29], độ đo entropy thông tin mờ (Fuzzy Information Entropy - FIE) [13], [30]–[32], độ đo khoảng cách mờ (Fuzzy Distance - FD) [33].

2. Tập thô mờ trực cảm (Intuitionistic Fuzzy Rough Set - IFRS)

Theo định nghĩa của IFRS, mỗi phần tử trong một tập mờ trực cảm được biểu diễn bởi hai thành phần gồm có: hàm thuộc và hàm không thuộc. Việc đánh giá mỗi quan hệ của hai đối tượng dựa trên hai thành phần này được cho là chặt hơn so với tập mờ truyền thống [34], [35]. Do đó, các nhà nghiên cứu nhận định thuật toán rút gọn thuộc được xây dựng theo tiếp cận IFRS có khả năng cải thiện độ chính xác phân lớp cho các tập rút gọn tốt hơn so với tiếp cận FRS trong các trường hợp tập dữ liệu nhiễu. Trong đó các tập dữ liệu nhiễu là các tập dữ

liệu có độ chính xác phân lớp ban đầu thấp. Gần đây, các công bố điển hình về rút gọn thuộc tính theo tiếp cận IFRS gồm có: phương pháp rút gọn thuộc tính theo tiếp cận miền dương mờ trực cảm (Intuitionistic Fuzzy POS) [36], theo tiếp cận entropy thông tin mờ trực cảm (Intuitionistic Fuzzy Information Entropy - IFIE) [15].

Tại Việt Nam, đã có một số luận án tiến sĩ nghiên cứu phương pháp rút gọn thuộc tính trực tiếp trên bảng quyết định số gồm có: luận án tiến sĩ của tác giả Cao Chính Nghĩa [3] nghiên cứu rút gọn thuộc tính và sinh luật quyết định trên các bảng dữ liệu số, có miền xác định đầy đủ sử dụng độ đo miền dương mờ. Luận án tiến sĩ của tác giả Nguyễn Văn Thiện [2] đề xuất độ đo khoảng cách mờ và xây dựng một số thuật toán tìm tập rút gọn theo phương pháp filter và phương pháp filter wrapper. Luận án tiến sĩ của tác giả Hồ Thị Phượng [1] đề xuất một số thuật toán gia tăng tìm tập rút gọn trong các bảng quyết định động sử dụng độ đo khoảng cách mờ.

Từ các kết quả khảo sát bên trên cho thấy, các phương pháp rút gọn thuộc tính trực tiếp trên bảng quyết định số tại Việt Nam hiện nay chỉ dựa trên tiếp cận FRS. Các kết quả thực nghiệm cho thấy tập rút gon thu được theo tiếp cận FRS còn chưa hiệu quả về kích thước và độ chính xác phân lớp trên các bộ dữ liệu nhiễu do không gian xấp xỉ mờ là chưa đủ để mô tả mối quan hệ của các đối tượng trong một tập. Đối với phương pháp rút gọn thuộc tính theo tiếp cận IFRS [15], [36] trên thế giới hiện nay còn chưa hiệu quả về kích thước của tập rút gọn và thời gian thực hiện của thuật toán do cách thức xây dựng không gian xấp xỉ mờ trực cảm các tác giả đề xuất chưa phản ánh đầy đủ thông tin quan hệ của một đối tượng và độ đo đánh giá độ quan trọng của thuộc tính còn quá phức tạp. Do đó, *mục tiêu nghiên cứu thứ nhất* của luận án là xây dựng phương pháp rút gọn thuộc tính theo tiếp cận IFRS hiệu quả về thời gian, kích thước, cải thiện độ chính xác phân lớp đối với các tập dữ liệu nhiễu.

Bên cạnh các phương pháp rút gọn thuộc tính theo tiếp cận tập thô và tập thô mở rộng như đã được trình bày bên trên. Phương pháp rút gọn thuộc tính theo tiếp cận tôpô cũng được các nhà nghiên cứu quan tâm và đề xuất trong những năm gần đây do các tính chất hoạt động của tôpô khá tương đồng với mô hình lý thuyết tập thô [37], [38].

Theo tiếp cận tôpô, khái niệm tập rút gọn theo cấu trúc tôpô lần đầu tiên được giới thiệu bởi Lashin và các công sự [37]. Để rút gọn thuộc tính cho bảng quyết định theo tiếp cận tôpô, trước tiên cần phải đưa ra các phương pháp xây dựng cấu trúc tôpô dựa trên các thông tin đã có trong bảng quyết định. Đây là một thách thức lớn, đã và đang thu hút được sự quan tâm của nhiều nhà nghiên cứu [37]–[39]. Hiện nay có hai phương pháp xây dựng tôpô theo tiếp cận tập thô gồm có, các phương pháp xây dựng tôpô từ không gian xấp xỉ của tập thô [38], [40]–[42], các phương pháp xây dựng tôpô từ các phép toán xấp xỉ của tập thô [43]. Bên cạnh đó, mối quan hệ của mô hình lý thuyết tôpô và tập thô cũng thu hút được sự chú ý của các nhà nghiên cứu [38], [43]–[47]. Trong đó, các nghiên cứu về sự tương đồng giữa các phép toán xấp xỉ của mô hình lý thuyết tập thô với các phép toán định miền của mô hình lý thuyết tôpô [48]. Trên cơ sở đó, nhiều cấu trúc tôpô được đề suất dựa trên việc xây dựng lại các phép toán xấp xỉ của lý thuyết tập thô [20], [45], [49]. Hơn nữa, dựa trên mối quan hệ này, một số phương pháp cấu trúc lại mô hình tập thô dựa trên cấu trúc tôpô cũng được đề xuất [44], [50], [51].

Tuy nhiên, hầu hết các nghiên cứu được trình bày bên trên chỉ là các nghiên cứu khái quát về mặt lý thuyết và cách tiếp cận xây dựng tôpô từ tập thô và tập thô từ tôpô nhằm nhấn mạnh mối quan hệ lý thuyết chặt chẽ của hai mô hình này. Gần đây, Xie và các công sự [52] đã đề xuất phương pháp rút gọn thuộc tính theo tiếp cận ma trận phân biệt tôpô. Tuy nhiên các kết quả nghiên cứu vẫn còn hạn chế về khung nền tảng lý thuyết và khả năng ứng dụng trong các bộ dữ liệu thực tiễn. Do đó, *mục tiêu nghiên cứu thứ hai của luận án* là nghiên cứu phương pháp rút gọn thuộc tính cho bảng quyết định theo tiếp cận tôpô đại số nhằm xây dựng nền

tảng lý thuyết tôpô đại số, ứng dụng cho bài toán rút gọn thuộc tính.

Mục tiêu nghiên cứu

Xuất phát từ những vấn đề còn tồn tại của các phương pháp rút gọn thuộc tính hiện nay, luận án đặt ra 02 mục tiêu nghiên cứu gồm có: 1) nghiên cứu phương pháp rút gọn thuộc tính theo tiếp cận *tập thô mờ trực cảm*; 2) nghiên cứu phương pháp rút gọn thuộc tính theo tiếp cận *tôpô đại số*.

- *Mục tiêu nghiên cứu thứ nhất*: Với phương pháp rút gọn thuộc tính theo tiếp cận tập thô mờ trực cảm, *vấn đề nghiên cứu trước tiên* là cần tìm hiểu cách thức mô tả mối quan hệ của các đối tượng hiệu quả trên nền tập mờ trực cảm, cụ thể là xây dựng các hàm đánh giá độ thuộc và độ không thuộc cho không gian xấp xỉ mờ trực cảm. Trên cơ sở đó, *vấn đề nghiên cứu tiếp theo* là cần xây dựng độ đo đánh giá độ quan trọng của thuộc tính hiệu quả về mặt thời gian, ứng dụng xây dựng thuật toán rút gọn thuộc tính hiệu quả trên các bộ dữ liệu nhiễu và có số chiều lớn trong thực tiễn.

- *Mục tiêu nghiên cứu thứ hai*: Với phương pháp rút gọn thuộc tính theo tiếp cận tôpô đai số, *vấn đề nghiên cứu trước tiên* là cần tìm hiểu các phương pháp xây dựng cấu trúc tôpô, tìm hiểu các tính chất cơ sở của tôpô sao cho có thể đánh giá tôpô trong một không gian nhỏ hơn để tiết kiệm chi phí tính toán. Trên cơ sở đó, *vấn đề nghiên cứu tiếp theo* là nghiên cứu các phép toán cơ bản trên cấu trúc tôpô nhằm xây dựng các phương pháp đánh giá, nhận diện độ quan trọng của thuộc tính, định nghĩa tập rút gọn thông qua cấu trúc tôpô, ứng dụng xây dựng thuật toán rút gọn thuộc tính hiệu quả trên các bộ dữ liệu có số chiều lớn trong thực tiễn.

Đối tượng nghiên cứu

Luận án tập trung nghiên cứu phương pháp rút gọn thuộc tính trên các bảng quyết định đầy đủ có miền giá trị số, các bảng quyết định nhiễu có số lượng mẫu và chiều từ trung bình đến lớn.

Luận án tập trung nghiên cứu các phương pháp rút gọn thuộc tính trong bảng quyết định theo tiếp cận tập thô và tôpô đại số gồm có:

- Khảo sát các khái niệm cơ bản về tập thô, các độ đo được sử dụng để đánh giá độ quan trọng của thuộc tính và các phương pháp xây dựng thuật toán rút gọn thuộc tính theo tiếp cận Heuristic.

- Khảo sát các khái niệm cơ bản về tôpô theo tiếp cận tập thô, tôpô thu từ không gian xấp xỉ, tôpô thu từ quan hệ của các phép toán xấp xỉ, tính khả li trong không gian tôpô và tôpô rút gọn.

Phạm vi nghiên cứu

Luận án tập trung nghiên cứu các biến thể dựa trên các tiếp cận của tập thô và tôpô đại số trên nền tập mờ và tập mờ trực cảm gồm có:

- Nghiên cứu các mô hình tập thô mở rộng trên nền tập mờ và tập mờ trực cảm, ứng dụng xây dựng thuật toán rút gọn thuộc tính trong bảng quyết định số.

- Nghiên cứu cấu trúc tôpô theo tiếp cận tập thô và một số tính chất khả li của không gian tôpô trên nền tập mờ và tập mờ trực cảm, ứng dụng xây dựng thuật toán rút gọn thuộc tính trong bảng quyết định số.

Phương pháp nghiên cứu:

Các kết quả nghiên cứu của luận án được đánh giá trên hai góc độ nghiên cứu gồm có:

- *Góc độ nghiên cứu lý thuyết*: các định nghĩa được trình bày rõ ràng, các mệnh đề được chứng minh chặt chẽ dựa vào nền tảng cơ bản của lý thuyết tập hợp, độ đo, tập thô, tập mờ, tập mờ trực cảm và entropy Shanon.

- *Góc độ nghiên cứu thực nghiệm*: các thuật toán được cài đặt và thực nghiệm trên các bộ dữ liệu từ UCI[1]. Sử dụng các mô hình phân lớp dữ liệu phù hợp với dữ liệu và các độ đo đánh giá, phương pháp đánh giá nhằm đánh giá chất lượng của tập rút gọn. So sánh chất lượng tập rút gọn từ thuật toán đề xuất với các thuật toán khác nhằm củng cố giả thiết nghiên cứu của luận án là hoàn toàn hợp lý.

Cấu trúc của luận án:

Ngoài phần mở đầu và kết luận, luận án có 04 chương nội dung nghiên cứu như sau:

Chương 1. Luận án giới thiệu và định nghĩa bài toán rút gọn thuộc tính, trình bày một số kiến thức cơ bản và các nghiên cứu liên quan. Các đóng góp chính của luận án được trình bày trong các chương 2, chương 3, và chương 4.

Chương 2. Luận án trình bày phương pháp rút gọn thuộc tính theo tiếp cận tập thô mờ trực cảm.

Chương 3. Luận án trình bày phương pháp rút gọn thuộc tính theo tiếp cận tôpô mờ trực cảm.

Chương 4. Luận án trình bày phương pháp rút gọn thuộc tính theo tiếp cận tôpô Hausdorff.

Cuối cùng, phần kết luận nêu những kết quả đã đạt được của luận án, hướng phát triển trong tương lai và những vấn đề quan tâm của tác giả.

CHƯƠNG 1. TỔNG QUAN BÀI TOÁN RÚT GỌN THUỘC TÍNH THEO TIẾP CẬN TẬP THÔ VÀ TÔPÔ

1.1. Mở đầu

Rút gọn thuộc tính (attribute reduction) hay còn được gọi lựa chọn đặc trưng (feature selection) là một trong những bước tiền xử lý dữ liệu quan trọng trong các lĩnh vực nhận dạng (pattern recognition), học máy (machine learning) và khai thác dữ liệu (data mining). Đối với các tập dữ liệu dành cho các bài toán học không giám sát (unsupervised - learning), rút gọn thuộc tính nhằm lựa chọn một tập con của tập thuộc tính ban đầu bảo toàn thông tin của tập thuộc tính gốc. Đối với các tập dữ liệu cho các bài toán học có giám sát (supervised - learning), rút gọn thuộc tính nhằm chọn ra một tập con của tập thuộc tính ban đầu bảo toàn khả năng phân lớp hay dự báo so với tập thuộc tính gốc.

1.2. Các khái niệm cơ bản

1.2.1. Tập thô truyền thống

Định nghĩa 1.1 (Hệ thông tin [14])**.** Hệ thông tin là một bộ tứ $IS = (U, A, V, f)$ trong đó U là tập hữu hạn khác rỗng các đối tượng, A là tập hữu hạn khác rỗng các thuộc tính, $V = \bigcup_{a \in A} V_a$ với V_a là tập giá trị của thuộc tính $a \in A$ và $f : U \times A \to V_a$ là hàm thông tin, $\forall a \in A,\ u \in U$ ta có $f(u,a) \in V_a$.

Định nghĩa 1.2 (Phân hoạch của thuộc tính [18], [53])**.** Cho bảng quyết định $DT = (U, C, D, f)$ và $P, Q \subseteq C$. Khi đó:

1) Phân hoạch U/P và phân hoạch U/Q được gọi là như nhau hay $U/P = U/Q$, khi và chỉ khi $\forall u \in U,\ [u]_P = [u]_Q$.

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

2) Phân hoạch U/P được gọi là mịn hơn phân hoạch U/P hay $U/P \preceq U/Q$ khi và chỉ khi $\forall u \in U,\ [u]_P \subseteq [u]_Q$

Định nghĩa 1.3 (Mô hình tập thô truyền thống [14], [18], [53])**.** Trong mô hình lý thuyết tập thô truyền thống, để biểu diễn tập $X \subseteq U$ trên cơ sở tri thức của tập thuộc tính B theo khái niệm tập thô, Pawlack sử dụng hai phép toán dựa trên các lớp tương đương của U/B. Các phép toán này được gọi là B-xấp xỉ dưới và B-xấp xỉ trên của X trên U/B, ký hiệu lần lượt là $\underline{B}(X)$ và $\overline{B}(X)$. Trong đó:

$$\underline{B}(X) = \{u \in U \,|[u]_B \subseteq X\} \tag{1.1}$$

$$\overline{B}(X) = \{u \in U \,|[u]_B \cap X \neq \emptyset\} \tag{1.2}$$

1.2.2. Tập thô mờ trực cảm

Định nghĩa 1.4 (Tập mờ trực cảm [54])**.** Cho U là tập không rỗng các đối tượng, tập mờ trực cảm X trên U được xác định bởi:

$$X = \{\langle x, \mu_X(x), \nu_X(x)\rangle \,|x \in U\} \tag{1.3}$$

Trong đó, $\mu_X(x) \in [0,1]$ là mức độ thành viên của $x \in U$ với X và $\nu_A(x) \in [0,1]$ là mức độ không thành viên của $x \in U$ với X sao cho $0 \leq \mu_X(x) + \nu_X(x) \leq 1 \forall x \in U$.

Khi đó, với mỗi tập mờ Y truyền thống, tập mờ trực cảm X có thể được xác định bởi:

$$X = \{\langle x, \mu_Y(x), 1 - \mu_Y(x)\rangle \,|x \in U\} \tag{1.4}$$

Nếu $0 \leq \mu_X(x) + \nu_X(x) < 1$ thì $\pi_X(x) = 1 - \mu_X(x) - \nu_X(x)$ được gọi là độ do dự thành viên của $x \in U$ với X.

Định nghĩa 1.5 (Mô hình tập thô mờ trực cảm [36])**.** Cho bảng quyết định $DT = (U, C, D, f)$, R là quan hệ tương đương mờ xác định trên U và $A \subseteq U$, ta có:

$$\underline{A}(x) = \bigwedge_{y \in U} I(R(x,y), A(y)) \tag{1.5}$$

$$\bar{A}(x) = \bigvee_{y \in U} T(R(x,y), A(y)) \tag{1.6}$$

1.2.3. Không gian tôpô

Không gian tôpô [37] được kí hiệu bởi cặp (U, τ), trong đó U là tập không rỗng các đối tượng và τ là họ các tập con của U thỏa mãn các điều kiện sau:

(T1) $\Phi \in \tau$ and $U \in \tau$.

(T2) τ có tính đóng dưới phép toán hợp bất kì.

(T3) τ có tính đóng dưới phép toán giao hữu hạn.

Cặp (U, τ) được gọi là không gian tôpô xác định trên U với các phần tử là các tập mở và là tập con của U, phần bù của các tập mở được gọi là các tập đóng.

Định nghĩa 1.6 (Tập cơ sở [55])**.** Cho U là tập không rỗng các đối tượng. Khi đó cơ sở (base) của tôpô τ trên U là họ các tập con của C kí hiệu là B sao cho:

(1) Với mỗi $x \in U$, tồn tại $G \subseteq U$ sao cho $x \in G$.

(2) Vỡi mọi $G_1, G_2 \in B$, nếu $x \in G_1 \cap G_2$, thì tồn tại $G_3 \in B$ sao cho $x \in G_3$.

Định nghĩa 1.7 (Tập cơ sở con [55])**.** Cho không gian tôpô (U,τ). Khi đó $S \subseteq \tau$ được gọi là cơ sở con (subbase) của tôpô τ nếu giao hữu hạn các tập con của S tạo thành cơ sở B của tôpô τ.

Định nghĩa 1.8 (Tôpô Hausdorff [37])**.** Cho không gian xấp xỉ (U,τ), tôpô $\tau_H \in (U,\tau)$ được gọi là tôpô Hausdorff nếu mọi $x \neq y \in (U,\tau)$ luôn tồn tại hai lân cận mở $V_x, V_y \in \tau_H$ sao cho $V_x \cap V_y = \emptyset$.

1.2.4. Tập rút gọn

Trong bảng quyết định, các thuộc tính điều kiện được phân thành ba nhóm: thuộc tính lõi (core attribute), thuộc tính rút gọn (reductive attribute) và thuộc tính dư thừa (redundant attribute). Thuộc tính lõi là thuộc tính không thể thiếu trong việc phân lớp chính xác tập dữ liệu. Thuộc tính lõi xuất hiện trong tất cả các tập rút gọn của bảng quyết định. Thuộc tính dư thừa là những thuộc tính mà việc loại bỏ chúng không ảnh hưởng đến việc phân lớp tập dữ liệu, thuộc tính dư thừa không xuất hiện trong bất kỳ tập rút gọn nào của bảng quyết định. Thuộc tính rút gọn là thuộc tính xuất hiện trong một tập rút gọn nào đó của bảng quyết định.

1.3. Một số công thức tính toán độ thành viên

1.3.1. Chuẩn hóa dữ liệu

(1) Min-max normalization:

$$F\left(f_{c_k}(x_i)\right) = \frac{f_{c_k}(x_i) - min_{c_k}}{max_{c_k} - min_{c_k}} \left(max_{c_k}^{'} - min_{c_k}^{'}\right) + min_{c_k}^{'} \tag{1.7}$$

Trong đó max_{c_k} và min_{c_k} là các giá trị nhỏ nhất và lớn nhất của thuộc tính c_k. Sau khi chuẩn hóa, các giá trị của thuộc tính được đưa về đoạn mới $[min_{c_k}^{'}, max_{c_k}^{'}]$.

(2) z-score normalization:

$$F\left(f_{c_k}(x_i)\right) = \frac{f_{c_k}(x_i) - \overline{c_k}}{\sigma_{c_k}} \tag{1.8}$$

Trong đó, $\overline{c_k}$ và σ_{c_k} kí hiệu là giá trị trung bình và độ lệch chuẩn của thuộc tính c_k.

Trong đó I là số nguyên nhỏ nhất sao cho $\max\left(\left|F\left(f_{c_k}(x_i)\right)\right|\right) < 1$.

1.3.2. Độ đo độ tương tự

Đối với các thuộc tính có giá trị rời rạc, độ thành viên $r_{ij}^{c_{k_l}}$ được xác định như sau:

$$r_{ij}^{c_{k_l}} = \begin{cases} 1, & \text{if } f_{c_{k_l}}(x_i) = f_{c_{k_l}}\left(x_j\right) \\ 0, & \text{otherwise.} \end{cases} \tag{1.9}$$

Đối với các thuộc tính có giá trị số, $r_{ij}^{c_{k_l}}$ có thể được xác đinh bởi hàm F như sau:

$$r_{ij}^{c_{k_l}} = F\left(x_i, x_j\right) \tag{1.10}$$

Trong đó, F thỏa mãn $F(x_i,x_i) = 1, F\left(x_i,x_j\right) = F\left(x_j,x_i\right)$, và $F\left(x_i,x_j\right) \in [0,1]$.

Trong đó, $\sigma_{c_{k_l}}$ được gọi là độ lệch chuẩn.

1.4. Phương pháp đánh giá tập rút gọn

1.4.1. Các tiêu chí đánh giá

Các thuật toán rút gọn thuộc tính theo tiếp cận độ đo hiện nay thường được đánh giá dựa trên ba tiêu chí gồm có: *kích thước* của tập rút gọn thu được, *độ chính xác phân lớp* của tập rút gọn trên mô hình được huấn luyện và *thời gian thực hiện* của thuật toán.

Tập rút gọn thu được từ thuật toán có kích thước càng nhỏ thì càng hiệu quả về thời gian xây dựng mô hình. Độ chính xác càng cao thì càng khẳng định được phương pháp chọn lọc thuộc tính và cấu trúc tập rút gọn thu được hiệu quả. Thời gian thực hiện càng nhanh thà khả năng rút gọn dữ liệu trên các tập dữ liệu lớn càng hiệu quả.

Mục tiêu chung của các thuật toán rút gọn thuộc tính là cố gắng đạt được cả ba tiêu chí trên, tuy nhiên trong thực tế với các bộ dữ liệu nhiễu và phức tạp. Tiêu chí kích thước và độ chính xác phân lớp của tập rút gọn được nhiều nhà nghiên cứu quan tâm. Sau đây là một số độ đo đánh giá khả năng phân lớp chính xác của mô hình trên các tập rút gọn.

1.4.2. Mô hình và dữ liệu đánh giá

Theo khảo sát của [56] cho thấy, Các thuật toán phân lớp được sử dụng phổ biến trong đánh giá độ chính xác phân lớp của các tập dữ liệu trước và sau khi rút gọn gồm có: mô hình cây quyết định C.45, cây phân lớp và hồi quy CART, máy vector hỗ trợ SVM và mô hình phân lớp lân cận k-NN. Đối với các bảng quyết định có thuộc tính miền giá trị số, mô hình phân lớp k-NN và SVM được sử dụng nhiều hơn các mô hình phân lớp còn lại.

Hầu hết các thuật toán rút gọn thuộc tính được nghiên cứu và đánh giá dựa trên các tập dữ liệu được tải về từ UCI. Đây là kho dữ liệu đa dạng các chủ đề, đáng tin cậy. Được nhiều chuyên gia và các nhà nghiên cứu sử dụng.

1.4.3. Chỉ số đánh giá

(1) Accuracy:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}. \tag{1.11}$$

(2) Error:

$$Error = \frac{FP+FN}{TP+TN+FP+FN} \tag{1.12}$$

1.5. Một số phương pháp rút gọn thuộc tính

1.5.1. Phương pháp rút gọn thuộc tính theo tiếp cận ma trận phân biệt

Vào năm 1992, Skowron và Rauszer lần đầu tiên giới thiệu phương pháp rút gọn thuộc tính theo tiếp cận ma trận phân biệt trên nền tập thô [57]. Khi đó ma trận phân biệt có kích thước $n \times n$ với $n = |U|$, kí hiệu là $M(DS) = \left(c_{ij}\right)_{n\times n}$ được xác định bởi:

$$c_{ij} = \begin{cases} \left\{c \in C \mid c\left(x_i\right) \neq c\left(x_j\right)\right\}, \omega\left(x_i, x_j\right) \\ \emptyset, \text{ otherwise.} \end{cases} \tag{1.13}$$

Trong đó: $\omega\left(x_i, x_j\right)$ thỏa mãn một trong các điều kiện sau đây:

(1) $x_i \in \mathrm{POS}_C(D) \wedge x_j \notin \mathrm{POS}_C(D)$;

(2) $x_i \notin POS_C(D) \wedge x_j \in POS_C(D)$;

(3) $x_i, x_j \in POP_C(D) \wedge \left(x_i, x_j\right) \notin \mathrm{ind}(D)$.

Hàm phân biệt của ma trận phân biệt $f(C,D)$ là một hàm Boolean được xác định như sau:

$$f(C,D) = \wedge\left\{\vee c_{ij} \mid c_{ij} \neq \emptyset\right\} \tag{1.14}$$

Khi đó tập thuộc tính lõi được xác định bởi:

$$\text{core}_C(D) = \left\{c \mid c_{ij} = \{c\}\right\} \tag{1.15}$$

Cho đến nay, có khá nhiều phương pháp rút gọn thuộc tính theo tiếp cận ma trận phân biệt được đề xuất trong các công trình [58]–[61].

1.5.2. Phương pháp rút gọn thuộc tính theo tiếp cận độ đo

1.5.2.1. độ đo độ phụ thuộc

Độ đo độ phụ thuộc được giới thiệu bởi [39] nhận được nhiều quan tâm của các nhà nghiên cứu, cơ sở của độ đo này dựa trên khái niệm miền dương (POS) của tập thô. Cho bảng quyết định $DT = (U,C,D,f)$ với $B \subseteq C$, $X \subseteq U$ và R là quan hệ tương đương trên U. Khi đó miền dương của D theo B được xác định như sau:

$$\text{POS}_B(D) = \bigcup_{X_i \in U/D} \underline{R_B} X_i \tag{1.16}$$

Khi đó, độ phụ thuộc của D vào B được xác định bởi:

$$\gamma_B(D) = \frac{|P_B(D)|}{|U|} = \frac{\sum_{x \in U} P_B S_B(D)(x)}{|U|} \tag{1.17}$$

Trên cơ sở đó, độ quan trọng của thuộc tính theo tiếp cận POS được xác định dựa trên hai công thức chính sau đây:

$$\text{Sig}_1(a,B,D) = \gamma_B(D) - \gamma_{B-a}(D) \tag{1.18}$$

$$\text{Sig}_2(a,B,D) = \gamma_{B \cup a}(D) - \gamma_B(D) \tag{1.19}$$

Trong đó công thức 1.18 phù hợp với kỹ thuật tìm kiếm tham lam lùi còn công thức 1.19 phù hợp với kỹ thuật tìm kiếm tham lam tiến.

Trên cơ sở đó, các phương pháp rút gọn thuộc tính theo tiếp cận độ phụ thuộc được phát triển dựa trên mở rộng các độ đo này. Chi tiết các phương pháp được trình bày trong Bảng 1.1.

1.5.2.2. độ đo độ chắc chắn

Dựa trên khái niệm Entropy thông tin của Shanon, ba loại độ đo được mở rộng để đánh giá độ chắc chắn thông tin gồm có:

- entropy thông tin:

$$FE(B) = -\frac{1}{|U|}\sum_{i=1}^{|U|} \log_2 \frac{\left|[x_i]_{R_B}\right|}{|U|} \tag{1.20}$$

- entropy kết hợp:

$$FE(B,E) = -\frac{1}{|U|}\sum_{i=1}^{|U|} \log_2 \frac{\left|[x_i]_{R_B} \cap [x_i]_{R_E}\right|}{|U|} \tag{1.21}$$

Bảng 1.1: *Tổng hợp phương pháp rút gọn thuộc tính theo độ phụ thuộc*

STT	Tài liệu tham chiếu	Kiểu dữ liệu	Tiếp cận	Tập nền	Tiêu chuẩn đánh giá
1	[62]–[73]	Hybrid	NRS	Classical	accuracy, size, compuation time
2	[27], [32], [74]–[79]	Number	NRS	FS	accuracy, size, compuation time
3	[80]	Number	NRS	IFS	accuracy, size, compuation time
4	[81]	Hybrid	PRS	Classical	accuracy, size, compuation time
5	[17], [22]–[29], [74], [76], [82]–[89]	Number	FRS	FS	accuracy, size, compuation time
6	[34]–[36], [80], [90]–[96]	Number	IFRS	FS	accuracy, size, compuation time

- entropy có điều kiện:

$$FE(E \mid B) = -\frac{1}{|U|}\sum_{i=1}^{n} \log_2 \frac{\left|[x_i]_{R_E} \cap [x_i]_{R_B}\right|}{\left|[x_i]_{R_B}\right|} \tag{1.22}$$

Khi đó $\forall a \in C - B, B \subseteq C$, hai phương pháp tính độ quan trọng của thuộc tính a với tập thuộc tính B được xác định như sau:

$$\text{Sig}(a,B) = FE(B) - FE(B - \{a\}) \tag{1.23}$$

$$\text{Sig}(a,B,D) = FE(D \mid B - \{a\}) - FE(D \mid B) \tag{1.24}$$

Trên cơ sở đó, các phương pháp rút gọn thuộc tính theo tiếp cận độ chắc chắn được phát triển dựa trên mở rộng các độ đo này. Chi tiết các phương pháp được trình bày trong Bảng 1.2.

Bảng 1.2: *Tổng hợp phương pháp rút gọn thuộc tính theo độ không chắc chắn*

STT	Tài liệu tham chiếu	Kiểu dữ liệu	Tiếp cận	Tập nền	Tiêu chuẩn đánh giá
1	[15], [97], [98]	Number	Entropy thông tin	IFS	accuracy, size, compuation time
2	[31], [75], [99]	Number	Entropy điều kiện	FS	accuracy, size, compuation time
3	[100]	Hybrid	Entropy kết hợp	Classical	accuracy, size, compuation time
4	[101]	Number	Entropy bù	FS	accuracy, size, compuation time

1.5.2.3. độ đo khoảng cách

Cho bảng quyết định $DT = (U, C, D, f)$. Với mọi $P, Q \subseteq C$, với các tri thức tương ứng được kí hiệu bởi $K(P)$ và $K(Q)$. Trong đó $K(P) = \{[u_i]_P | u_i \in U\}$ và $K(Q) = \left\{[u_i]_Q | u_i \in U\right\}$. Khi đó, khoảng cách tri thức giữa P và Q theo tiếp cận Jacard được xác định như sau:

$$d_J(K(P), K(Q)) = 1 - \frac{1}{|U|}\sum_{i=1}^{|U|}\frac{\left|[u_i]_P \cap [u_i]_Q\right|}{\left|[u_i]_P \cup [u_i]_Q\right|} \tag{1.25}$$

Khi đó $\forall a \in C - B, B \subseteq C$, độ quan trọng của thuộc tính a với tập thuộc tính B được xác định như sau:

$$SIG_B(a) = d_J(K(B), K(B \cup D)) - d_J(K(B \cup \{a\}), K(B \cup \{a\} \cup D)) \tag{1.26}$$

Trên cơ sở đó, các phương pháp rút gọn thuộc tính theo tiếp cận độ đo khoảng cách được phát triển dựa trên mở rộng các độ đo này. Chi tiết các phương pháp được trình bày trong Bảng **??**.

Bảng 1.3: *Tổng hợp phương pháp rút gọn thuộc tính theo khoảng cách*

STT	Tài liệu tham chiếu	Kiểu dữ liệu	Tiếp cận	Tập nền	Tiêu chuẩn đánh giá
1	[24], [33], [68], [102], [103]	Hybrid	KD	Classical, FS, IFS	accuracy, size, compuation time
2	[29], [104], [105]	Number	GD	FS	accuracy, size, compuation time
3	[29]	Number	PD	FRS	accuracy, size, compuation time

1.5.3. Phương pháp rút gọn thuộc tính theo tiếp cận tôpô

Định nghĩa 1.9 (Tập rút gọn theo tiếp cận tôpô [37])**.** Cho bảng quyết định $DT = (U, C, D, f)$, với $B \subseteq C$ và $r \in B$. Khi đó r được gọi là quan hệ không cần có trong B nếu: $\beta_B = \beta_{(B-\{r\})}$. Khi đó: B được gọi là tập rút gọn của C khi và chỉ khi:

(i) $\beta_C = \beta_{(B)}$.

(ii) $\beta_C \neq \beta_{(B-\{\mathrm{r}\})}, \forall r \in C - B$.

Dựa trên định nghĩa về cấu trúc tôpô rút gọn, một số các nghiên cứu liên quan đến phương pháp xây dựng tôpô theo tiếp cận tập thô được trình bày trong Bảng **??**

Bảng 1.4: *Tổng hợp phương pháp xây dựng tôpô theo tiếp cận tập thô*

STT	Tài liệu tham chiếu	Cơ sở tính toán
1	[18], [20], [37], [39], [41], [106], [107]	Không gian xấp xỉ
2	[37]–[39], [41], [47], [48], [106]–[109]	Tập xấp xỉ trên và tập xấp xỉ dưới
3	[20], [39], [45], [47], [55], [85], [108], [110], [111]	Không gian mẫu và quan hệ của các phép toán

1.6. Kết luận Chương 1

Chương 1 đã giới thiệu khái quát về bài toán rút gọn thuộc tính và phân loại phương pháp rút gọn thuộc tính. Trình bày các cơ sở lý thuyết quan trọng để thực hiện trong các Chương nghiên cứu tiếp theo của luận án.

CHƯƠNG 2. PHƯƠNG PHÁP RÚT GỌN THUỘC TÍNH TRONG BẢNG QUYẾT ĐỊNH THEO TIẾP CẬN TẬP THÔ MỜ TRỰC CẢM

2.1. Mở đầu

Trong chương này, luận án trình bày phương pháp rút gọn thuộc tính theo tiếp cận độ đo khoảng cách mờ trực cảm. Trước tiên luận án đề xuất độ đo khoảng cách giữa hai phân hoạch mờ trực cảm, trên cơ sở đó luận án xây dựng độ đo đánh giá độ quan trọng của thuộc tính. Tiếp theo luận án đề xuất thuật toán Heuristic tìm tập rút gọn trên cơ sở đề xuất cấu trúc tập rút gọn theo tiếp cận độ tương tự δ - equal. Cuối cùng là thực nghiệm và so sánh thuật toán đề xuất với các thuật toán của A.Tan [36], [112] trên các bộ dữ liệu được tải về từ UCI.

Các kết quả nghiên cứu đã được công bố trên các công trình nghiên cứu [CT3, CT4].

2.2. Xây dựng độ đo khoảng cách mờ trực cảm

2.2.1. Khoảng cách giữa hai tập mờ trực cảm

Bổ đề 2.1 [Số mờ trực cảm]**.** *Cho ba số thực* $a, b, c \in [0,1]$. *Khi đó:*

1) Nếu $a \geq b$ *then* $a - b \geq \min(a,\ c) - \min(b,\ c)$

2) Nếu $a \leq b$ *then* $a - b \leq max(a,\ c) - max(b,\ c)$

Mệnh đề 2.1 (Quan hệ của các IFS)**.** *Cho* $\widetilde{\widetilde{X}}, \widetilde{\widetilde{Y}}, \widetilde{\widetilde{Z}}$ *là các tập mờ trực cảm xác định trên U, với U là tập không rỗng các đối tượng. Khi đó:*

1) Nếu $\widetilde{\widetilde{X}} \subseteq \widetilde{\widetilde{Y}}$ *thì* $\left|\widetilde{\widetilde{Y}}\right| - \left|\widetilde{\widetilde{Y}} \cap \widetilde{\widetilde{Z}}\right| \geq \left|\widetilde{\widetilde{X}}\right| - \left|\widetilde{\widetilde{X}} \cap \widetilde{\widetilde{Z}}\right|$

2) Nếu $\widetilde{\widetilde{X}} \subseteq \widetilde{\widetilde{Y}}$ *thì* $\left|\widetilde{\widetilde{Z}}\right| - \left|\widetilde{\widetilde{Z}} \cap \widetilde{\widetilde{X}}\right| \geq \left|\widetilde{\widetilde{Z}}\right| - \left|\widetilde{\widetilde{Z}} \cap \widetilde{\widetilde{Y}}\right|$

3) $\left|\widetilde{\widetilde{X}}\right| - \left|\widetilde{\widetilde{X}} \cap \widetilde{\widetilde{Y}}\right| + \left|\widetilde{\widetilde{Z}}\right| - \left|\widetilde{\widetilde{Z}} \cap \widetilde{\widetilde{X}}\right| \geq \left|\widetilde{\widetilde{Z}}\right| - \left|\widetilde{\widetilde{Z}} \cap \widetilde{\widetilde{Y}}\right|$

Mệnh đề 2.2 (Khoảng cách giữa hai IFS)**.** *Cho hai tập mờ trực cảm* $\widetilde{\widetilde{X}}, \widetilde{\widetilde{Y}}$ *xác định trên U, với U là tập không rỗng các đối tượng. Khi đó* $\widetilde{\widetilde{d}}\left(\widetilde{\widetilde{X}}, \widetilde{\widetilde{Y}}\right) = \left|\widetilde{\widetilde{X}} \cup \widetilde{\widetilde{Y}}\right| - \left|\widetilde{\widetilde{X}} \cap \widetilde{\widetilde{Y}}\right|$ *là khoảng cách giữa hai tập mờ trực cảm* $\widetilde{\widetilde{X}}, \widetilde{\widetilde{Y}}$.

2.2.2. Khoảng cách giữa hai phân hoạch mờ trực cảm

Định nghĩa 2.1 (Khoảng cách phân hạt mờ trực cảm)**.** Cho bảng quyết định $DT = (U, C, D, f)$ và $[\widetilde{\widetilde{X}}]$, $[X \cup \widetilde{\widetilde{D}}]$ tương ứng là các phân hoạch của X và $X \cup D$ với $X \subseteq C$. Khi đó khoảng cách

giữa $[\overset{\approx}{X}]$, $[X \overset{\approx}{\cup} D]$ được xác định bởi:

$$\overset{\approx}{d}\left([\overset{\approx}{X}],[X \overset{\approx}{\cup} D]\right) = \frac{1}{|U|^2}\sum_{i=1}^{|U|}\left(\left|[u_i]_{[\overset{\approx}{X}]}\right| - \left|[u_i]_{[\overset{\approx}{X}]} \cap [u_i]_{[\overset{\approx}{D}]}\right|\right) \tag{2.1}$$

Mệnh đề 2.3 (Độ đo phân hạt)**.** *Cho bảng quyết định $DT = (U, C, D, f)$ và $[\overset{\approx}{X}]$, $[X \overset{\approx}{\cup} D]$ tương ứng là các phân hoạch của X và $X \cup D$ với $X \subseteq C$. Khi đó $\overset{\approx}{d}\left([\overset{\approx}{X}],[X \overset{\approx}{\cup} D]\right)$ là một độ đo khoảng cách.*

Thuật toán 2.1 Thuật toán filter - wrapper hai giai đoạn sử dụng khoảng cách mờ trực cảm (IFD)

Input: $DT = (U, C, D, f)$, mô hình phân lớp $Model$, $\Delta = \{0.1, 0.2, ..., 0.9\}$
Output: Tập rút gọn R

1: $R_W^A \leftarrow \emptyset$;
2: $R_W^\delta \leftarrow \emptyset$;
3: **for all** $c \in C$ **do**
4: computation $[\overset{\approx}{c}]$;
5: **end for**
6: **for all** $\delta \in \Delta$ **do**
7: $R_F^\delta \leftarrow \emptyset$;
8: **while** $[R_F^\delta \overset{\approx}{\cup} D] \overset{\delta}{\neq} [C \overset{\approx}{\cup} D]$ **do**
9: $c_m \in C - R_F^\delta | SIG_{R_F^\delta}(c_m) = \underset{c \in C - R_F^\delta}{Max}\left\{SIG_{R_F^\delta}(c)\right\}$; {Giai đoạn filter}
10: $R_F^\delta := R_F^\delta \cup \{c_m\}$;
11: **end while**
12: **if** $ACC(Model, R_F^\delta) > ACC(Model, R_W^\delta)$ **then**
13: $R_W^\delta = R_F^\delta$; {Giai đoạn wrapper delta (W_δ)}
14: **end if**
15: **end for**
16: **for** $(i = 1; i < |R_W^\delta|; i++)$ **do**
17: **if** $ACC(Model, R_W^\delta[0:i]) > ACC(Model, R_W^A)$ **then**
18: $R_W^A = R_W^\delta[0:i]$; {Giai đoạn wrapper attribute (W_A)}
19: **end if**
20: **end for**
21: **return** R_W^A;

2.3. Rút gọn thuộc tính trong bảng quyết định sử dụng độ đo khoảng cách mờ trực cảm

2.3.1. Đề xuất thuật toán tìm tập rút gọn theo phương pháp lai ghép filter - wrapper, sử dụng độ đo khoảng cách mờ trực cảm

Định nghĩa 2.2 (Ma trận δ equal)**.** Cho hai ma trận quan hệ mờ trực cảm $\overset{\approx}{M}_B = [b_{ij}]_{n\times n}$ và $\overset{\approx}{M}_C = [c_{ij}]_{n\times n}$ với $n = |U|$. Khi đó $\overset{\approx}{M}_B$ và $\overset{\approx}{M}_C$ được gọi là δ equal khi và chỉ khi:

1) $\sup_{i,j=1}^{n} \left|\mu(b_{ij}) - \mu(c_{ij})\right| \leq 1-\delta$

2) $\sup_{i,j=1}^{n} \left|\nu(b_{ij}) - \nu(c_{ij})\right| \leq 1-\delta$

Trong đó $\sup_{i,j=1}^{n}$ cho biết sự khác biệt lớn nhất của hai ma trận quan hệ mờ trực cảm đạt được tại vị trí i, j, với $\delta \in [0,1]$. Ta kí hiệu $\tilde{\tilde{M}}_B \overset{\delta}{=} \tilde{\tilde{M}}_C$.

Định nghĩa 2.3 (Độ quan trọng của thuộc tính)**.** Cho bảng quyết định $DT = (U, C, D, f)$ và tập thuộc tính $B \subseteq C$. Khi đó độ quan trọng của thuộc tính $a \in C - B$ với tập thuộc tính B được xác định bởi:

$$SIG_B(a) = \tilde{\tilde{d}}\left([\tilde{\tilde{B}}], [B \tilde{\tilde{\cup}} D]\right) - \tilde{\tilde{d}}\left([B \tilde{\tilde{\cup}} a], [B \cup \tilde{\tilde{a}} \cup D]\right) \tag{2.2}$$

Định nghĩa 2.4 (Tập rút gọn)**.** Cho bảng quyết định $DT = (U, C, D, f)$ và tập thuộc tính $B \subseteq C$. Khi đó tập thuộc tính B được gọi là tập rút gọn nếu:

1) $[B \tilde{\tilde{\cup}} D] \overset{\delta}{=} [C \tilde{\tilde{\cup}} D]$;

2) $\forall b \in B,\ [B - \{\tilde{\tilde{b}}\} \cup D] \overset{\delta}{\neq} [C \tilde{\tilde{\cup}} D]$.

Thuật toán có độ phức tạp: $O\left(|C|\,|U|^2\right) + O(\mathbb{T}|\Delta||C|^2|U|^2) + O(\mathbb{T}|R_W^{\delta}|)$.

Bảng 2.1: *Mô tả kích thước thu được của tập rút gọn thu được từ các thuật toán*

ID	Dataset	|C|	|R|			
			IFD-SVM	IFD-KNN	IFPOS[36]	IFIE[15]
1	heart	13	7	9	13	10
2	CMSC	20	11	11	20	20
3	PDS	22	9	7	8	10
4	BCWP	32	25	21	12	12
5	IS	34	16	5	11	19
6	UFDC	43	26	29	8	11
7	UFDD	43	27	25	6	8
8	SHDC	44	2	9	10	14
9	UFDB	51	2	2	5	11
10	DPDS	54	5	7	15	24
11	sonar	60	11	31	17	25
12	VRB	310	11	12	18	35

2.3.2. Thực nghiệm và đánh giá thuật toán

Chương này sử dụng hai thuật toán của A. Tan [15], [36] để so sánh và đánh giá thuật toán đề xuất IFD. Trong đó thuật toán [36] là thuật toán sử dụng độ đo miền dương mờ trực cảm và thuật toán [15] sử dụng độ đo Entropy mờ trực cảm(Intuitionistic Fuzzy Information Entropy - IFIE). Bảng 2.1 so sánh kích thước của tập rút gọn thu được từ các thuật toán, trong khi đó các Bảng 2.2 và 2.3 so sánh độ chính xác phân lớp của các tập rút gọn thu được từ các thuật toán trên hai mô hình phân lớp SVM và KNN.

Bảng 2.2: *So sánh độ chính xác phân lớp của các tập rút gọn trên mô hình phân lớp SVM*

ID	Dataset	|U|	Accuracy			
			Raw	IFD-SVM	IFPOS[36]	IFIE[15]
1	heart	270	84±0.7	84±0	84±0.6	82±0.7
2	CMSC	540	95±0.2	95±0.9	95±0.9	95±0.2
3	PDS	195	84±0.5	85±0.1	85±0.1	84±0.7
4	BCWP	198	77±0.2	77±0.1	76±0.7	76±0.5
5	IS	351	88±0	89±0.9	87±0.6	87±0.6
6	**UFDC**	181	44±0.1	52±0	49±1	49±0.3
7	UFDD	180	68±0.9	68±1	64±0.8	63±0.8
8	SHDC	267	79±0.6	79±0.5	79±0.8	79±0.9
9	UFDB	92	100.0	100.0	100.0	92±0.4
10	DPDS	170	98±0.5	98±0.5	98±0.7	98±0.3
11	**sonar**	208	65±0.3	70±0.5	70±0.2	64±0.7
12	VRB	126	83±0.7	88±0.7	91±0.2	80±0.5

Bảng 2.3: *So sánh độ chính xác phân lớp của các tập rút gọn trên mô hình phân lớp KNN*

ID	Dataset	|U|	Accuracy			
			Raw	IFD-KNN	IFPOS[36]	IFIE[15]
1	heart	270	77±0.4	78±0.2	77±0.6	76±0.8
2	CMSC	540	84±0.9	86±0.9	84±0.4	84±0.6
3	PDS	195	85±0.5	87±0.8	87±0.3	84±0.3
4	BCWP	198	78±0.7	79±0.8	79±0.1	79±0.1
5	IS	351	85±0.3	92±0.5	88±0.6	88±0.6
6	**UFDC**	181	82±0.7	86±0.8	74±0.5	78±0.3
7	UFDD	180	81±0.8	84±0.2	77±0	82±0.1
8	**SHDC**	267	66±0.3	72±0.4	69±0.8	67±0.2
9	UFDB	92	99.0	100.0	100.0	98±0.8
10	DPDS	170	98±1	97±0.2	98±0.5	96±0.8
11	**sonar**	208	68±0.8	69±0.1	62±0.9	60±0.9
12	VRB	126	68±0.6	82±0.7	81±0.7	65±0.2

2.4. Kết luận Chương 2

Chương 2 đã trình bày về phương pháp rút gọn thuộc tính theo tiếp cận tập thô mờ trực cảm trên cơ sở mở rộng độ đo khoảng cách giữa các phân hoạch. Các kết quả thực nghiệm cho thấy thuật toán đề xuất có khả năng cải thiện tốt độ chính xác phân lớp trên một số bộ dữ liệu nhiễu.

CHƯƠNG 3. PHƯƠNG PHÁP RÚT GỌN THUỘC TÍNH TRONG BẢNG QUYẾT ĐỊNH THEO TIẾP CẬN TÔPÔ MỜ TRỰC CẢM

3.1. Mở đầu

Chương này đề xuất phương pháp rút gọn thuộc tính theo tiếp cận tôpô mờ trực cảm. Trước tiên là đề xuất cấu trúc tôpô dựa trên quan hệ ưu tiên mờ trực cảm, trên cơ sở đó nghiên cứu một số tính chất của IF-base và IF-subbase để xây dựng phương pháp đánh giá sự tương đồng giữa hai tôpô mờ trực cảm. Tiếp theo là đề xuất một số thuật toán rút gọn thuộc tính trên cơ sở tính chất tương đồng của hai tôpô thông và định nghĩa tập rút gọn theo cấu trúc tôpô đơn vị.

Các kết quả nghiên cứu trong Chương này được công bố trên các công trình nghiên cứu [CT2] và [CT6] đang chờ phản biện vòng 2.

3.2. Đề xuất cấu trúc tôpô mờ trực cảm

Định nghĩa 3.1 (Công thức quan hệ mờ trực cảm)**.** Cho bảng quyết định $DT = (U, C, D, f)$, với mọi $(x, y) \in U$ và $\delta \in [0.5, 1]$, Khi đó $IFR_a^{\geq}(x, y) = \langle y, \mu_y, \nu_y \rangle$ với $a \in C$ được xác định bởi:

$$\begin{aligned} \mu_y &= \begin{cases} 1 - |a(x) - a(y)| \ if \ p_a(x, y) \geq \delta \\ 0 \ if \ other \end{cases} \\ \nu_y &= 1 - \mu_y \end{aligned} \tag{3.1}$$

Trong đó $p_a(x, y) = \frac{a(x) - a(y) + 1}{2}$. Khi đó, giá trị p_a luôn thuộc đoạn $[0.5, 1]$. Khi giá trị $\delta = 0.5$, quan hệ ưu tiên này có tính chất phản xạ và bắc cầu, khi $\delta > 0.5$ quan hệ ưu tiên này chỉ có tính bắc cầu.

Định nghĩa 3.2 (Cở sở con IF-subbase)**.** Cho bảng quyết định $DT = (U, C, D, f)$. Khi đó IF-subbase của $a \in C$ được định nghĩa bởi:

$$S_a = \{S_a^L, S_a^R\} \tag{3.2}$$

Trong đó S_a^L và S_a^R lần lượt là IF-subbase trái tương ứng với ma trận quan hệ $M_a^{\geq}$ và IF-subbase phải tương ứng với ma trận quan hệ $\left(M_a^{\geq}\right)^T$ trên thuộc tính $a \in C$, với $\left(M_a^{\geq}\right)^T$ là ma trận chuyển vị của ma trận $M_a^{\geq}$.

Định nghĩa 3.3 (Giao hai IF-subbase)**.** Cho bảng quyết định $DT = (U, C, D, f)$ và hai IF-subbases S_p, S_q tương ứng với $p, q \in C$. Khi đó, phép toán giao của hai IF-subbase được định nghĩa bởi:

$$S_p \cap S_q = \{S_p^L \cap S_q^L, S_p^R \cap S_q^R\} \tag{3.3}$$

Định nghĩa 3.4 (Hợp hai IF-subbase)**.** Cho bảng quyết định $DT = (U, C, D, f)$ và hai IF-subbases S_p, S_q tương ứng với $p, q \in C$. Khi đó, phép toán hợp của hai IF-subbase được định nghĩa bởi:

$$S_p \cup S_q = \{S_p^L \cup S_q^L, S_p^R \cup S_q^R\} \tag{3.4}$$

Định nghĩa 3.5 (Cơ sở IF-base)**.** Cho bảng quyết định $DT = (U, C, D, f)$ và IF-subbase $S_a = \{S_a^L, S_a^R\}$ tương ứng với $a \in C$, trong đó S_a^L được gọi là IF-subbase trái và S_a^R được gọi là IF-subbase phải. Khi đó IF-base B_a được định nghĩa bởi:

$$B_a = {S_a}^L \cap {S_a}^R \tag{3.5}$$

Mệnh đề 3.1 (IFT từ IF-base)**.** *Cho bảng quyết định $DT = (U,C,D,f)$ và B_a là một IF-base được xác định bởi công thức 3.5. Khi đó, B_a là một cơ sở của $\mathscr{T}_a$.*

Định nghĩa 3.6 (IF-subbase của tập thuộc tính)**.** Cho bảng quyết định $DT = (U,C,D,f)$, với mọi $p,q \in C$. Khi đó IF-subbase của $\{p\} \cup \{q\}$ được định nghĩa bởi:

$$S_{\{p\}\cup\{q\}} = S_p \cap S_q \tag{3.6}$$

Định nghĩa 3.7 (IF-base mịn nhất)**.** Cho bảng quyết định $DT = (U,C,D,f)$ và IF-base B_a tương với $a \in C$. Khi đó B_a được gọi là IF-base mịn nhất (smoothest) nếu: $B_a[i,j] = \begin{cases} 1_{IF} \ if \ i = j \\ 0_{IF} \ if other \end{cases}$

Trong đó $1_{IF} = (1,0)$ và $0_{IF} = (0,1)$. Kí hiệu IF-base mịn nhất là B_I là cơ sở của tôpô đơn vị mờ trực cảm.

3.3. Đề xuất độ đo tương đồng của hai tôpô mờ trực cảm

Mệnh đề 3.2 (Độ khác biệt giữa hai IF-subbase)**.** *Cho bảng quyết định $DT = (U,C,D,f)$ và hai IF-subbases S_p, S_q tương ứng với $p,q \in C$. Khi đó:*

$$\zeta\left(S_p,S_q\right) = \frac{2}{|U|^2}\sum_{i=1}^{|U|}\left(\left|S_p^L[i] \cup S_q^L[i]\right| - \left|S_p^L[i] \cap S_q^L[i]\right|\right) \tag{3.7}$$

Là độ khác biệt giữa S_p và S_q

Mệnh đề 3.3 (Độ phụ thuộc của thuộc tính theo IF-subbase)**.** *Cho bảng quyết định $DT = (U,C,D,f)$ và hai IF-subbases S_C and $S_{C\cup D}$ tương ứng với C và $C \cup D$. Khi đó:*

$$\zeta\left(S_C,S_{C\cup D}\right) = \frac{2}{|U|^2}\sum_{i=1}^{|U|}\left(\left|S_D^L[i] - S_D^L[i] \cap S_C^L[i]\right|\right) \tag{3.8}$$

Là độ phụ thuộc của thuộc tính D với thuộc tính C.

Mệnh đề 3.4 (Tính chất phản đơn điệu của độ đo tương đồng)**.** *Cho bảng quyết định $DT = (U,C,D,f)$ và hai IF-subbases S_B, S_C tương ứng với B và C. Khi đó, nếu $B \subseteq C$ thì $\zeta\left(S_D,S_{D\cup C}\right) \leq \zeta\left(S_D,S_{D\cup B}\right)$:*

3.4. Rút gọn thuộc tính trong bảng quyết định theo tiếp cận tôpô mờ trực cảm

3.4.1. Đề xuất thuật toán tìm tập rút gọn trong bảng quyết định theo phương pháp filter

Định nghĩa 3.8 (Độ quan trọng của thuộc tính)**.** Cho bảng quyết định $DT = (U,C,D,f)$ và tập thuộc tính $R \subseteq C$. Khi đó, độ quan trọng của thuộc tính $a \in C - R$ với tập thuộc tính R được định nghĩa bởi:

$$Sig_R(a) = \zeta\left(S_D,S_{D\cup R\cup a}\right) - \zeta\left(S_D,S_{D\cup R}\right) \tag{3.9}$$

Mệnh đề 3.5 (Tính tồn tại của tập rút gọn)**.** *Cho bảng quyết định $DT = (U,C,D,f)$ và hai IF-bases B_R và B_C tướng ứng với $R \subseteq C$. Khi đó, nếu $B_R = B_I$ thì $B_C = B_I$.*

Dựa trên mệnh đề 3.5 ta có thể khẳng định, nếu một bảng quyết định tồn tại một tập con

Thuật toán 3.1 Rút gọn thuộc tính theo phương pháp filter sử dụng tiếp cận tôpô mờ trực cảm (F_IFT)

Input: Bảng quyết định $DT = (U, C, D, f)$ và $\delta = 0.5$
Output: Tập rút gọn R

$R \leftarrow \emptyset$;
B_R là cơ sở mờ trực cảm thô nhất;
B_I là cơ sở mờ trực cảm mịn nhất;
for all $c \in C \cup D$ **do**
  calculate S_c; {theo công thức 3.1 và 3.2}
end for
while $B_R \neq B_I$ **do**
  for all $c \in C - R$ **do**
    calculate $Sig_R(c)$; {theo công thức 3.9}
  end for
  select $c_m \in C - R : Sig_R(c_m) = \underset{c \in C-R}{Max} \{Sig_R(c)\}$;
  $R \leftarrow R \cup \{c_m\}$;
  update B_R; {theo công thức 3.5}
end while
return R;

R của tập thuộc tính ban đầu C mà B_R là cơ sở mịn nhất thì chắc chắn B_C cũng là cơ sở mịn nhất. Nghĩa là $B_R = B_C = B_I$. Khi đó, tập rút gọn có thể được định nghĩa như sau:

Định nghĩa 3.9 (Tập rút gọn theo tiếp cận tôpô đơn vị)**.** Cho bảng quyết định $DT = (U, C, D, f)$ và $R \subseteq C$. Khi đó R được gọi là một tập rút gọn của C khi và chỉ khi

(1) $B_R = B_I$

(2) $B_{R-c} \neq B_I$ với mọi $c \in R$

Để đảm bảo tính tồn tại của B_I, quan hệ ưu tiên mờ trực cảm đề xuất phải có tính chất phản xạ, do đó giá trị δ được chọn mặc định là 0.5 cho toàn bộ các ví dụ minh họa và thực nghiệm các thuật toán.

Thuật toán F_IFT có độ phức tạp: $\mathcal{O}\left(|R||C-R||U|^2\right)$ và thuật toán FW_IFT có độ phức tạp: $\mathcal{O}\left(|ST||C-R_F||U|^2\right) + \mathcal{O}\left(|R_W||T|\right)$.

3.4.2. Thực nghiệm và đánh giá các thuật toán

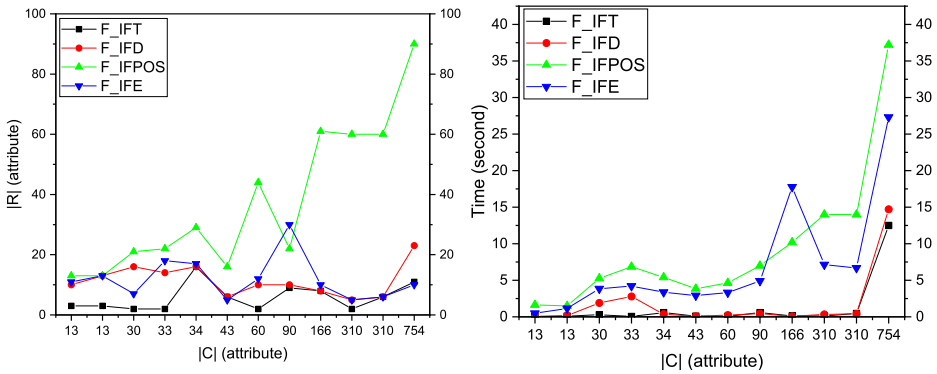
Phần này sẽ trình bày một số kết quả thực nghiệm của hai thuật toán đề suất. Trong đó thuật toán F_IFT sẽ được so sánh với các thuật toán của A. Tan [15], [36] và Thang [102]. Thuật toán FW_IFT sẽ được so sánh với thuật toán FW_IFD [102].

Hình 3.1 cho thấy ưu điểm về thời gian thực hiện của thuật toán F_IFT cũng như mối quan hệ về thời gian và kích thước tập rút gọn thu được của thuật toán. Trong khi đó, các Bảng 3.1 và 3.2 cho thấy những ưu điểm về kích thước và độ chính xác phân lớp của tập rút gọn thu được từ thuật toán FW_IFT.

Thuật toán 3.2 Phương pháp rút gọn thuộc tính lai ghép filter - wrapper sử dụng tiếp cận tôpô mờ trực cảm (FW_IFT)

Input: Bảng quyết định $DT = (U, C, D, f)$ và $\delta = 0.5$, mô hình phân lớp $Model$
Output: Tập rút gọn R

1: $ST \leftarrow \emptyset$; $R_W \leftarrow \emptyset$; $R_F \leftarrow \emptyset$; $R \leftarrow \emptyset$;
2: B_{R_F} là cơ sở mờ trực cảm thô nhất;
3: B_I là cơ sở mờ trực cảm mịn nhất;
4: **for all** $c \in C \cup D$ **do**
5: **calculate** S_c; {theo công thức 3.1 và 3.2}
6: **end for**
7: **for all** $c \in C - R_F$ **do**
8: calculate $Sig_{R_F}(c)$; {theo công thức 3.9}
9: **end for**
10: **for all** $c_m \in \{ \underset{c \in C - R_F}{Max} \{Sig_{R_F}(c)\} \}$ **do**
11: $ST.PUSH(R_F \cup \{c_m\})$; {Đẩy c_m vào Stack}
12: **end for**
13: **while** $ST \neq \emptyset$ **do**
14: $R_F = ST.POP$; {giai đoạn filter}
15: **update** B_{R_F}
16: **if** $B_{R_F} = B_I$ **then**
17: $R_W = R_W \cup \{R_F\}$; {Đưa tập rút gọn ứng viên vào danh sách}
18: **else**
19: quay lại bước 10;
20: **end if**
21: **end while**
22: **for all** $r \in R_W$ **do**
23: **if** $ACC(Model, r) > ACC(Model, R)$ **then**
24: $R = r$; {Giai đoạn wrapper}
25: **end if**
26: **end for**
27: **return** R;



Hình 3.1: *Biểu đồ đánh giá mối quan hệ về kích thước tập rút gọn (trái) và thời gian thực hiện (phải) với số lượng thuộc tính ban đầu của thuật toán F_IFT so với các thuật toán khác*

3.5. Kết luận Chương 3

Chương 3 đã trình bày phương pháp rút gọn thuộc tính theo tiếp cận tôpô mờ trực cảm và đề xuất hai thuật toán. Trong đó thuật toán F_IFT cho tập rút gọn có kích thước và thời

Bảng 3.1: *So sánh kích thước của các tập rút gọn thu được từ các thuật toán theo tiếp cận filter - wrapper trên mô hình phân lớp SVM và KNN*

ID	Dataset	FW_IFT		FW_IFD		$\|C\|$
		SVM	KNN	SVM	KNN	
1	Wine	5	4	10	10	13
2	Heart	6	5	11	11	13
3	Wdbc	3	5	16	16	30
4	Wpbc	3	2	2	2	33
5	Iono	7	5	12	12	34
6	UFDC	8	6	5	5	43
7	Sona	3	2	9	9	60
8	Libras	18	13	7	14	90
9	Musk	5	5	3	3	166
10	LVB	6	2	2	2	310
11	LVG	7	5	5	5	310
12	PD	9	11	17	23	754

Bảng 3.2: *So sánh độ chính xác phân lớp của các tập rút gọn thu được từ các thuật toán theo tiếp cận filter - wrapper trên mô hình phân lớp SVM và KNN*

ID	Data	FW_IFT		FW_IFD		$\|C\|$	
		SVM	KNN	SVM	KNN	SVM	KNN
1	Wine	94.24	91.25	97.87	94.74	98.16	96.25
2	Heart	86.43	78.85	84.65	76.74	84.5	77.44
3	Wdbc	97.15	95.42	97.99	95.02	98.33	95.45
4	Wpbc	77.79	76.12	76.14	78.34	78.02	77.18
5	Iono	87.1	92.05	85.46	89.14	88.37	86.04
6	**UFDC**	68.16	90.9	50.95	69.14	43.49	79.13
7	**Sona**	77.21	68.35	67.35	61	65.45	68.16
8	Libras	70.9	77.59	64.79	78.02	71.41	75.23
9	Musk	73.17	75.13	62.51	64.41	75.54	77.37
10	LVB	85.29	77.19	77.71	76.31	83.24	67.8
11	LVG	90.22	78.64	70.18	66.93	89.05	69.22
12	PD	84.47	84.79	84.8	65.53	81.26	81.8

gian thực hiện hiệu quả, trong khi đó thuật toán FW_IFT cho tập rút gọn có kích thước và độ chính xác phân lớp hiệu quả.

CHƯƠNG 4. PHƯƠNG PHÁP RÚT GỌN THUỘC TÍNH TRONG BẢNG QUYẾT ĐỊNH THEO TIẾP CẬN TÔPÔ HAUSDORFF

4.1. Mở đầu

Phần này trình bày tóm tắt các kết quả nghiên cứu về phương pháp rút gọn thuộc tính theo tiếp cận tôpô Hausdorff. Trong đó, dề xuất cấu trúc tôpô theo tiếp cận tập thô trên không gian

xấp xỉ mờ ngưỡng β và đề xuất cấu trúc tôpô Hausdorff; Đề xuất khái niệm đồng cấu trúc phụ thuộc trong không gian tôpô Hausdorff; Đề xuất thuật toán tìm tập rút gọn dựa trên định nghĩa mới về thuộc tính quan trọng theo tiếp cận tôpô Hausdorff và góm nhóm các thuộc tính theo khái niệm đồng cấu trúc phụ thuộc của tôpô Hausdorff. Các kết quả nghiên cứu trong Chương này được công bố trên các công trình nghiên cứu [CT1], [CT5]

4.2. Đề xuất cấu trúc tôpô từ không gian xấp xỉ mờ ngưỡng β

Định nghĩa 4.1 (Công thức quan hệ mờ ngưỡng β)**.** Quan hệ tương đương mờ ngưỡng β của $u_i, u_j \in U$ được xác định như sau:

$$R^{\beta}\left(u_i,u_j\right)=\begin{cases} 1-\left|u_i-u_j\right| : if\ 1-\left|u_i-u_j\right| \geq \beta \\ 0 : if\ 1-\left|u_i-u_j\right| < \beta. \end{cases} \tag{4.1}$$

Mệnh đề 4.1 (Cấu trúc tôpô theo tiếp cận tập thô)**.** *Cho không gian xấp xỉ* (U,R^{β}) *và* R^{β} *là một quan hệ tương đương mờ. Khi đó* $\mathscr{T}=\left\{X\subseteq U|\underline{R^{\beta}}(X)=\overline{R^{\beta}}(X)\right\}$ *là một tôpô xác định trên* U*.*

4.3. Đề xuất cấu trúc tôpô Hausdorff

Định nghĩa 4.2 (Tính khả li của quan hệ mờ ngưỡng β)**.** Cho không gian xấp xỉ (U,R^{β}) trong đó R^{β} là quan hệ tương đương mờ β. Khi đó R^{β} được gọi là phân biệt được nếu với mọi $u_i \in U$ tồn tại $u_j \neq u_i \in U$ sao cho $[u_i]_{R^{\beta}} \cap [u_j]_{R^{\beta}} = \emptyset$. Kí hiệu quan hệ này là R_H^{β} .

Mệnh đề 4.2 (Tôpô Hausdorff từ quan hệ R_H^{β})**.** *Cho tôpô* $\mathscr{T}_H=\{X\subseteq U|\underline{R^{\beta}}(X)=\overline{R^{\beta}}(X)\}$ *xác định trên* U*. Khi đó,* $\mathscr{T}_H$ *được gọi là tôpô Hausdorff nếu* R^{β} *là một* R_H^{β} *.*

Mệnh đề 4.3 (Xác định thuộc tính có quan hệ R_H^{β})**.** *Cho bảng quyết định* $DT=(U,C,D,f)$ *và* $c \in C$*. Khi đó* c *được gọi là thuộc tính có quan hệ* R_H^{β} *nếu* $max_1(V_c)-max_2(V_c)>\beta$*. Trong đó* V_c *là tập giá trị của thuộc tính* c*.*

4.4. Rút gọn thuộc tính trong bảng quyết định theo tiếp cận tôpô Hausdorff

4.4.1. Đề xuất thuật toán tìm tập rút gọn trong bảng quyết định theo phương pháp lai ghép filter - wrapper, sử dụng cấu trúc tôpô Hausdorff

Định nghĩa 4.3 (Thuộc tính quan trọng theo tiếp cận tôpô Hausdorff)**.** Cho bảng quyết định $DT=(U,C,D,f)$ và $c \in C$. Khi đó c được gọi là thuộc tính quan trọng với D nếu $\mathscr{T}_c$ là một tôpô Hausdorff.

Định nghĩa 4.4 (Đồng cấu trúc phụ thuộc)**.** Cho bảng quyết định $DT=(U,C,D,f)$ và hai tôpô $\mathscr{T}_p$, $\mathscr{T}_q$ xác định trên U tương ứng với $p,q \in C$. Khi đó $\mathscr{T}_p$ được gọi là đồng cấu trúc phụ thuộc với $\mathscr{T}_q$ nếu $\mathscr{T}_p \cup \mathscr{T}_D = \mathscr{T}_q \cup \mathscr{T}_D$.

Tiếp theo sẽ là phần đánh giá độ phức tạp của thuật toán CFW. Kí hiệu $|U|$ là số các đối tượng, $|C|$ là số các thuộc tính, $|H^{\beta}|$ là số các thuộc tính Hausdorff, và $|CH^{\beta}|$ là số các nhóm thuộc tính Hausdorff có cùng cấu trúc phụ thuộc. Khi đó độ phức tạp từ 6-10 là $\mathscr{O}(2|U||C|)$, độ phức tạp từ 11-22 là $\mathscr{O}(|U|^2|H^{\beta}|^2)$. Giả sử $\mathbb{T}$ là thời gian thực hiện của

Thuật toán 4.1 Thuật toán rút gọn thuộc tính theo tiếp cận filter - wrapper các cụm thuộc tính (CFW).

Input Bảng quyết định $DT = (U, C, D)$ với $\Delta = \{0.1, 0.2, ..., 0.8, 0.9\}$ và mô hình phân lớp $Model$

Output Tập rút gọn R

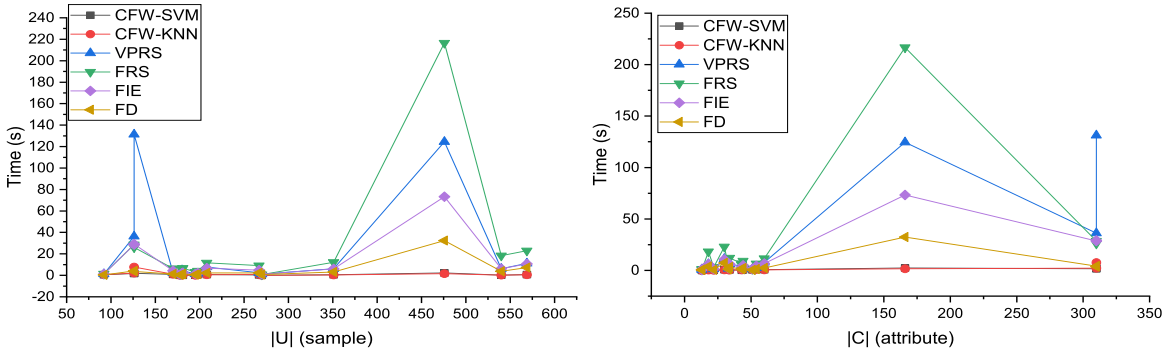
$R = \emptyset$;
for all $\beta \in \Delta$ **do**
  $H^\beta \leftarrow \emptyset$;
  $CH^\beta \leftarrow \emptyset$;
  $R^\beta \leftarrow \emptyset$;
  for all $c \in C$ **do**
    if $max_1(V_c) - max_2(V_c) > \beta$ **then**
      $H^\beta = H^\beta \cup \{c\}$; {Filter các thuộc tính Hausdorff}
    end if
  end for
  for all $p \in \{H^\beta - CH^\beta\}$ **do**
    $U_p = \emptyset$;
    for all $q \in \{H^\beta - CH^\beta - p\}$ **do**
      if $\mathscr{T}_p \cup \mathscr{T}_D = \mathscr{T}_q \cup \mathscr{T}_D$ **then**
        $U_p = U_p \cup \{q\}$; {Phân cụm thuộc tính Hausdorff}
      end if
    end for
    $CH^\beta = CH^\beta \cup U_p$;
    if $ACC_{U_p}^{Model} > ACC_{R^\beta}^{Model}$ **then**
      $R^\beta = U_p$; {Wrapper các cụm thuộc tính Hausdorff}
    end if
  end for
  if $ACC_{R^\beta}^{Model} > ACC_{R}^{Model}$ **then**
    $R = R^\beta$; {Wrapper các tập rút gọn ứng viên β}
  end if
end for
return R;

mô hình phân lớp $Model$. Với số lượng Δ rất nhỏ, khi đó độ phức tạp của thuật toán là $\mathscr{O}(2|U||C|) + \mathscr{O}(|U|^2|H^\beta| + |CH^\beta|\mathbb{T})$.

4.4.2. Thực nghiệm và đánh giá thuật toán

Thuật toán đề xuất được so sánh và đánh giá với các thuật toán rút gọn thuộc tính điển hình trên tiếp cận độ đo gồm có: (1) thuật toán rút gọn thuộc tính theo tiếp cận tập thô với độ chính xác điều chỉnh (VPRS) [113]; (2) thuật toán rút gọn thuộc tính theo tiếp cận tập thô mờ (FRS) [114]; (3) thuật toán rút gọn thuộc tính theo tiếp cận Entropy thông tin mờ (IFE) [87]; (4) thuật toán rút gọn thuộc tính theo tiếp cận khoảng cách mờ (FD) [33].

Bảng 4.1 so sánh và đánh giá kích thước tập rút gọn thu được của các thuật toán, bên cạnh đó mối quan hệ về kích thước của tập dữ liệu với thời gian thực hiện của các thuật toán cũng được mô tả trong hình 4.1. Các Bảng 4.2 và 4.3 so sánh độ chính xác phân lớp của tập rút gọn thu được từ các thuật toán trên hai mô hình phân lớp k-NN và SVM.



Hình 4.1: *Biểu đồ phân tích mối quan hệ giữa thời gian thực hiện của thuật toán và $|U|$ (left), giữa thời gian thực hiện của thuật toán và $|C|$ (right).*

Bảng 4.1: *So sánh kích thước của tập rút gọn thu được từ các thuật toán*

ID	Dataset	|R|					
		$\|C\|$	CFW-SVM	CFW-kNN	VPRS	FRS	FIE	FD
1	wine	13	10.8	7.6	11.8	10.4	10.6	7.1
2	heart	13	6.7	5.5	11.5	13.9	10.2	6.7
3	CMSC	20	8.2	8.7	9.5	20.3	20.1	3.5
4	PDS	22	5.2	4.4	9.4	8.5	10.8	4.3
5	BCWD	30	3.2	3.6	14.8	7.6	12.1	4.1
6	BCWP	32	2.9	2.2	8.9	12.6	12.4	5.8
7	IS	34	2.1	2.1	20.9	11.3	19.6	6.1
8	UFDC	43	13.9	4.3	15.3	8.7	11.7	5.2
9	UFDD	43	5.1	3.6	19.9	6.6	8.3	3.3
10	SHDC	44	3.1	2.2	44.3	10.3	14.7	5.9
11	UFDB	51	4.1	3.4	8.9	5.8	11.9	5.2
12	DPDS	54	2.5	1.6	8.4	15.7	24.4	4.4
13	sonar	60	4.6	7.4	44.3	17.5	25.2	7.6
14	musk	166	5.7	11.4	86.6	23.9	29.5	8.8
15	VRB	310	9.1	4.3	56.6	18.9	35.8	7.5
16	VRG	310	9.6	2.1	72.4	16.5	36.4	10.6

4.5. Kết luận Chương 4

Chương 4 đã trình bày phương pháp rút gọn thuộc tính theo tiếp cận tôpô Hausdorff. Các kết quả thực nghiệm cho thấy thuật toán đề xuất là hoàn toàn vượt trội so với các phương pháp khác.

Bảng 4.2: *So sánh độ chính xác phân lớp của tập rút gọn thu được từ các thuật toán trên mô hình phân lớp SVM*

ID	Dataset	Classification Accuracy (%)					
		Rawset	CFW-SVM	VPRS	FRS	FIE	FD
1	wine	98±0.7	96±0.9	99±0.6	99±0.3	93±0.1	96±0.8
2	heart	84±0.8	86±0.6	84±0.3	84±0.3	82±0.9	80±0.7
3	CMSC	95±0.8	95±0.4	92±0.4	95±0.1	95±0.8	92±0.6
4	PDS	84±0.7	86±0.6	84±0.7	85±0.9	84±0.7	75±0.8
5	BCWD	98±0.6	94±0.7	94±0.2	96±0	96±0.8	94±0.7
6	BCWP	77±0.3	76±0.3	76±0.6	76±0.2	76±0.8	76±0
7	IS	88±0.5	82±1	88±0.9	87±0.5	87±0.3	89±0.6
8	UFDC	44±0.8	**59±0.7**	45±0.5	49±0.1	49±0.6	50±1
9	UFDD	68±0.8	63±0.5	68±0.1	64±1	63±0.7	62±0.5
10	SHDC	79±0.5	79±1	79±0	79±0	79±0.6	79±0.3
11	UFDB	100±0.4	96±0.9	100±0.6	100±0.2	92±0.8	100±0.2
12	DPDS	98±0.6	98±0.3	98±0.3	98±0.6	98±0.4	98±0.5
13	sonar	65±0.8	**73±0.2**	65±0.2	70±0.7	64±0	58±0
14	musk	75±0.3	72±0.2	74±0.8	61±0.4	61±0.1	55±0.4
15	VRB	83±0.1	83±0.2	88±0.6	91±0.4	80±0.8	86±1
16	VRG	85±0.9	80±0.2	91±0.7	82±0.5	67±0.2	68±0.4

Bảng 4.3: *So sánh độ chính xác phân lớp của tập rút gọn thu được từ các thuật toán trên mô hình phân lớp KNN*

ID	Dataset	Classification Accuracy (%)					
		Rawset	CFW-kNN	VPRS	FRS	FIE	FD
1	wine	96±0.2	94±0.1	94±0.1	96±0.9	91±0.4	94±0.6
2	heart	77±0.5	78±0.1	77±0.3	77±0.3	76±0.2	69±0.7
3	CMSC	84±0.1	92±0.1	86±0.2	84±0.6	84±0.9	71±0.1
4	PDS	85±0.7	85±0.3	88±0.9	87±0.1	84±0.3	74±0.5
5	BCWD	95±0.2	93±0.1	93±0.9	93±0.9	94±0.7	93±0.7
6	BCWP	78±0.8	81±0.9	74±0.6	79±0.6	79±0.6	75±0.6
7	IS	85±0.6	88±0.6	86±0.9	88±0.7	88±0.4	89±0.4
8	UFDC	82±0.1	96±0.2	82±0.1	74±0.9	78±0.1	76±0.2
9	UFDD	81±0.5	81±0.9	77±0.9	77±0.5	82±0.6	72±0.7
10	SHDC	66±0.1	75±0.7	66±0.5	69±0.8	67±1	72±0.6
11	UFDB	99±0.8	100	100	100	98±0	99±0.5
12	DPDS	98±0.4	98±0	98±0.3	98±0.4	96±0.9	98±0.2
13	sonar	68±0.3	71±0.3	64±0.5	62±0.7	60±0.7	55±0.3
14	musk	77±0.5	76±0.7	77±0.1	75±1	69±0.4	64±0.3
15	VRB	68±0.3	76±0.6	77±0.1	81±0.3	65±0.1	73±0.1
16	VRG	70±0.8	96±0.4	75±0.8	76±0.9	61±1	60±0.9

KẾT LUẬN

A. Những kết quả chính của luận án

Trên cơ sở các mục tiêu đề ra như đã được trình bày trong phần mở đầu của luận án, các kết quả chính của luận án gồm có: 1) Xây dựng thuật toán rút gọn thuộc tính theo tiếp cận lai ghép filter - wrapper, sử dụng độ đo khoảng cách mờ trực cảm (IFD). 2) Xây dựng thuật toán rút gọn thuộc tính theo tiếp cận filter (F_IFT) và thuật toán lai ghép filter - wrapper (FW_IFT), sử dụng cấu trúc tôpô mờ trực cảm. 3) Xây dựng thuật toán rút gọn thuộc tính theo tiếp cận lai ghép filter - wrapper cụm (CFW), sử dụng cấu trúc tôpô Hausdorff.

Kết quả thực nghiệm trên các bộ dữ liệu tải về từ UCI cho thấy:

- Thuật toán IFD có khả năng cải thiện nhiễu khá tốt, tuy nhiên kích thước và độ chính xác phân lớp của tập rút gọn chưa hiệu quả hơn so với các thuật toán được so sánh.

- Thuật toán F_IFT có thời gian thực hiện hiệu quả và kích thước tập rút gọn thu được tốt nhưng độ chính xác phân lớp còn hạn chế so với các thuật toán được so sánh.

- Thuật toán FW_IFT cho tập rút gọn có kích thước và độ chính xác phân lớp hiệu quả, tuy nhiên thời gian thực hiện của thuật toán còn hạn chế so với các thuật toán được so sánh.

- Thuật toán CFW là hoàn toàn vượt trội về thời gian thực hiện, kích thước và độ chính xác phân lớp của tập rút gọn thu được cũng trội hơn so với các thuật toán tốt nhất được so sánh.

B. Những đóng góp mới của luận án

Các kết quả nghiên cứu của luận án đã đóng góp 03 phương pháp rút gọn thuộc tính gồm có:

- Phương pháp rút gọn thuộc tính theo tiếp cận tập thô mờ trực cảm dựa trên các đề xuất mới về độ đo khoảng cách mờ trực cảm.

- Phương pháp rút gọn thuộc theo tiếp cận tôpô mờ trực cảm dựa trên các đề xuất mới về IF-subbase, IF-base và tôpô đơn vị.

- Phương pháp rút gọn thuộc tính theo tiếp cận tôpô Hausdorff dựa trên các đề xuất mới về tính chất khả li trên không gian xấp xỉ mờ ngưỡng β.

C. Hướng phát triển tiếp theo của luận án

Hiện nay, các bảng quyết định không đầy đủ, thiếu giá trị xuất hiện khá phổ biến trong các lĩnh vực khai thác dữ liệu và học máy. Đã có nhiều phương pháp rút gọn thuộc tính trong bảng quyết định không đầy đủ theo tiếp cận mô hình tập thô mở rộng, tuy nhiên các kết quả nghiên cứu vẫn còn hạn chế về kích thước và độ chính xác phân lớp trên các tập rút gọn thu được. Do đó, hướng nghiên cứu tương lai của luận án sẽ nhắm tới rút gọn thuộc tính cho bảng quyết định không đầy đủ thông qua một số các hướng mở rộng cấu trúc tôpô theo tiếp cận tập thô như sau:

1) Mở rộng cấu trúc tôpô dựa trên không gian xấp xỉ của mô hình tập thô dung sai, nghiên cứu một số tính chất khả li nhằm tìm ra tiêu chuẩn chọn lọc thuộc tính và xây dựng điều kiện dừng của thuật toán. 2) Mở rộng cấu trúc tôpô dựa trên mối quan hệ của các phép toán xấp xỉ của mô hình tập thô dung sai, nghiên cứu một số tính chất khả li nhằm tìm ra tiêu chuẩn chọn lọc thuộc tính và xây dựng điều kiện dừng của thuật toán. 3) Phát triển một số phép toán tính toán gia tăng trên không gian tôpô cho các trường hợp dữ liệu động. 4) Phát triển cấu trúc đại số của tôpô với các định nghĩa mới về toán tử hợp-k và giao-k thuộc tính nhằm tăng tốc quá trình tìm kiếm tập rút gọn.

DANH MỤC CÁC CÔNG TRÌNH NGHIÊN CỨU

A. Published

[CT1] **Trần Thanh Đại**, Nguyễn Long Giang, Trần Thị Ngân, Hoàng Thị Minh Châu, "Rút gọn thuộc tính cho bảng quyết định đầy đủ theo tiếp cận Topo mờ", *Hội thảo quốc gia lần thứ XXIV: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông*, Thái Nguyên, 12/2021 pp. 318-325, 2021.

[CT2] **Trần Thanh Đại**, Nguyễn Long Giang, Trần Thị Ngân, Hoàng Thị Minh Châu, Vũ Thu Uyên, Vương Trung Hiếu, "Về một phương pháp rút gọn thuộc tính cho bảng quyết định theo tiếp cận topo mờ trực cảm", *Các công trình nghiên cứu và phát triển CNTT và truyền thông*, Hà Nội, số 2, tr. 57-64, 2022.

[CT3] Nguyen Truong Thang, Nguyen Long Giang, **Tran Thanh Dai**, Nguyen Trung Tuan, Nguyen Quang Huy, Pham Viet Anh, Vu Duc Thi, "A Novel Filter-Wrapper Algorithm on Intuitionistic Fuzzy Set for Attribute Reduction from Decision Tables", *International Journal of Data Warehousing and Mining (IJDWM)* , số 17(4), tr. 67-100, 2021. (SCIE Q4 IF 0.78).

[CT4] **Trần Thanh Đại**, Nguyễn Long Giang, Hoàng Thị Minh Châu, Trần Thị Ngân, "Rút gọn thuộc tính cho bảng quyết định theo tiếp cận tập thô mờ trực cảm", *Kỷ yếu Hội nghị Khoa học Công nghệ Quốc Gia lần thứ XIII: Nghiên cứu cơ bản và ứng dụng công nghệ thông tin*, Nha Trang, 10/2020, tr. 516-524, 2020.

[CT5] **Trần Thanh Đại**, Nguyễn Long Giang, Vũ Đức Thi, Phan Đăng Hưng,"Về một phương pháp rút gọn thuộc tính theo tiếp cận tôpô Hausdorff", *Hội thảo quốc gia lần thứ XXVI: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông*, Bắc Ninh, 10/2023,tr. 416-523, 2023.

B. Waiting review

[CT6] **Tran Thanh Dai**, Nguyen Long Giang, Vu Duc Thi, Tran Thi Ngan, Hoang Thi Minh Chau, Le Hoang Son "A New Approach for Attribute Reduction from Decision Table based on Intuitionistic Fuzzy Topology", *Soft Computing*. (SCIE Q2 IF 3.8). Đang chờ phản biện vòng 2.